ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«ΠΛΗΡΟΦΟΡΙΚΗ»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΣΤΗΝ ΕΚΤΙΜΗΣΗ ΠΙΣΤΩΤΙΚΟΥ ΚΙΝΔΥΝΟΥ** **COMPARISON OF MACHINE LEARNING ALGORITHMS IN CREDIT RISK ASSESSMENT**
Ονοματεπώνυμο Φοιτητή	**ΗΛΙΑΣ ΣΥΡΜΟΣ**
Πατρώνυμο	**ΠΕΤΡΟΣ**
Αριθμός Μητρώου	**ΜΠΠΛ19055**
Επιβλέπων	**ΔΙΟΝΥΣΙΟΣ ΣΩΤΗΡΟΠΟΥΛΟΣ, ΕΠΙΚ. ΚΑΘΗΓΗΤΗΣ**

Ημερομηνία Παράδοσης **Οκτώβριος 2024**

Τριμελής Εξεταστική Επιτροπή

Διονύσιος Σωτηρόπουλος	Κωνσταντίνα Χρυσαφιάδη	Ευάγγελος Σακκόπουλος
Επίκουρος Καθηγητής	Επίκουρος Καθηγητής	Αναπληρωτής Καθηγητής

Περίληψη

Η αξιολόγηση πιστωτικού κινδύνου αποτελεί κρίσιμο ζήτημα για τα χρηματοπιστωτικά ιδρύματα, καθώς η ακριβής εκτίμηση της πιθανότητας αθέτησης ενός δανείου είναι απαραίτητη για την αποφυγή οικονομικών απωλειών. Στην παρούσα μελέτη, η αξιολόγηση πιστωτικού κινδύνου αντιμετωπίζεται ως ένα πρόβλημα δυαδικής ταξινόμησης μεταξύ των κατηγοριών (α) δανειοληπτών που αθετούν και (β) που δεν αθετούν. Χρησιμοποιείται το σύνολο πραγματικών δεδομένων "Home Credit Default Risk" της πλατοφόρμας Kaggle με στόχο την εκπαίδευση και τη σύγκριση της απόδοσης των μοντέλων μηχανικής μάθησης, Logistic Regression, Random Forest, XGBoost και LightGBM, για την εκτίμηση της πιστοληπτικής ικανότητας των δανειοληπτών. Προηγήθηκε διερευνητική ανάλυση και επεξεργασία των δεδομένων όπως και δημιουργία χαρακτηριστικών για να εμπλουτιστούν τα δεδομένα και να βελτιωθεί η απόδοση των μοντέλων. Παράλληλα, χρησιμοποιήθηκαν τεχνικές επιλογής χαρακτηριστικών, με βάση τη σημαντικότητα τους μέσω του LightGBM, ενώ εξετάστηκε και η εφαρμογή της PCA για τη μείωση της διάστασης των χαρακτηριστικών.

Στο σύνολο των δεδομένων παρατηρήθηκε μεγάλη ανισορροπία μεταξύ των δυο κατηγοριών, με την πλειοψηφία των δανειοληπτών να μην αθετούν, κάτι που οδήγησε στην δοκιμή τεχνικών εξισορρόπησης όπως η SMOTE και η SMOTEENN, προκειμένου να βελτιωθεί η ικανότητα των μοντέλων να αναγνωρίσουν αυτούς που αθετούν. Για την επικύρωση των αποτελεσμάτων χρησιμοποιήθηκε η μέθοδος Stratified KFold ενώ οι επιδόσεις των μοντέλων αξιολογήθηκαν με την χρήση του Confusion Matrix και των παραμέτρων αξιολόγησης ROC-AUC, F1-Score, Precision και Recall, με το μοντέλο LightGBM να αποδεικνύεται ως το πιο αποδοτικό στην πλειονότητα των δοκιμών επιτυγχάνοντας ακρίβεια πρόβλεψης (ROC-AUC=0.7865) και την ανίχνευση περιπτώσεων αθέτησης (Recall=0.67). Ωστόσο, παρά το γεγονός ότι το LightGBM απέδωσε καλύτερα σε σύγκριση με τα υπόλοιπα μοντέλα, τα συνολικά αποτελέσματα παραμένουν μη ικανοποιητικά, καθώς τα μοντέλα δεν είναι ικανά να διαχωρίσουν επαρκώς τις κατηγορίες, λόγω της έντονης ανισορροπίας. Η δυσκολία αυτή υπογραμμίζει την ανάγκη για διερεύνηση πιο εξειδικευμένων τεχνικών εξισορρόπησης και τη χρήση μεθόδων μηχανικής μάθησης, ικανών να αντιμετωπίσουν καλύτερα προβλήματα ανισορροπίας. Η υιοθέτηση αυτόματων τεχνικών δημιουργίας χαρακτηριστικών και η βελτιστοποίηση των υπερπαραμέτρων των μοντέλων θα μπορούσαν να οδηγήσουν σε σημαντικές βελτιώσεις στην απόδοση τους, με αποτέλεσμα τη βελτίωση των προβλέψεων και την καλύτερη διαχείριση του πιστωτικού κινδύνου.

Λέξεις Κλειδιά: μηχανική μάθηση, πιστωτικός κίνδυνος, LighGBM, XGBoost, Logistic Regression, Random Forest, ανισορροπία κατηγοριών

Abstract

Credit risk assessment is a critical issue for financial institutions, as an accurate assessment of the probability of default on a loan is essential to avoid financial losses. In this study, credit risk assessment is treated as a binary classification problem between the categories of (a) defaulting and (b) non-defaulting borrowers. The dataset "Home Credit Default Risk" which contains real data from the Kaggle platform is used to train and compare the performance of the machine learning models, Logistic Regression, Random Forest, XGBoost and LightGBM, in order to assess the creditworthiness of borrowers. This was preceded by exploratory data analysis and data preprocessing as well as feature generation to enrich the data and improve the performance of the models. In addition, feature selection techniques were used based on the feature importance through the use of LightGBM, and the application of PCA to reduce the dimensionality of the features was also considered.

In the dataset there was a large class imbalance between the two categories, with the majority of borrowers not defaulting, which led to the testing of balancing techniques such as SMOTE and SMOTEENN to improve the models' ability to identify those who default. For the validation of the results we used the Stratified KFold method while the performance of the models was evaluated using the Confusion Matrix and other metrics as the ROC-AUC, F1-Score, Precision and Recall, with the LightGBM model proving to be the most efficient in the majority of the tests achieving prediction accuracy (ROC-AUC=0. 7865) and detecting instances of default (Recall=0.67). However, although LightGBM performed better compared to the other models, the overall results remain unsatisfactory as the models are not able to adequately identify the categories due to the strong imbalance. This difficulty highlights the need to explore more specialized balancing techniques and the use of machine learning methods capable of better addressing class imbalance problems. The adoption of automatic feature generation techniques and optimization of model hyperparameters could lead to significant improvements in model performance, resulting in improved forecasting and better credit risk management.

Keywords:machine learning, credit risk, LightGBM, XGBoost, Logistic Regression, Random Forest, Class Imbalance

Περιεχόμενα

Περιεχόμενα Εικόνων

Περιεχόμενα Πινάκων

1 Εισαγωγή

Η αξιολόγηση του πιστωτικού κινδύνου (credit risk assessment), αποτελεί μια από τις πλέον απαραίτητες υποστηρικτικές διαδικασίες για το χρηματοοικονομικό σύστημα. Τόσο οι τράπεζες κατά τον προσδιορισμό τραπεζικών πολιτικών, όσο και οι επιχειρήσεις στη λήψη επιχειρηματικών αποφάσεων, οφείλουν να έχουν εγκαθιδρύσει αξιόπιστες μεθοδολογίες υπολογισμού του στα συστήματά τους. Η αξιολόγηση του χρηματοοικονομικού κινδύνου χαρακτηρίζεται από διασυνδεσιμότητα, αλληλεξάρτηση και μεγάλη πολυπλοκότητα (Wu et al., 2014), ακολουθεί δε αυστηρές προδιαγραφές που διέπονται από τους κανόνες εποπτείας του τραπεζικού συστήματος.

Με τη χρηματοπιστωτική κρίση του 2008, αυτού του είδους η ανάγκη έγινε ακόμη μεγαλύτερη, αφού η βαρύτητα του σκορ πιστωτικού κινδύνου (credit risk score) αυξήθηκε τόσο πολύ, που αποτέλεσε μια από τις κυριότερες μεταβλητές αξιολόγησης στη διαχείριση πιστωτικού κινδύνου. Συγκεκριμένα, βασικός λόγος ύπαρξης της είναι η υποστήριξη της αγοράς στη λήψη αποφάσεων σχετικά με την έγκριση δανειοδότησης ενός υποψήφιου, λαμβάνοντας υπόψη πολλούς διαφορετικούς παραγόντες. Υπάρχουν δε σημαντικά οικονομικά οφέλη από την ύπαρξη ενός ισχυρού credit score (Blöchlinger and Leippold, 2006).

Ωστόσο, τα τελευταία χρόνια, αυτού του είδους οι παραδοσιακές υπηρεσίες, οι οποίες εποπτεύονται σε συστηματική βάση, έχουν διαταραχθεί από νέους ανταγωνιστές οι οποίοι προσφέρουν υπηρεσίες διαμεσολάβησης σε εικονικό διαδικτυακό περιβάλλον, γνωστοί και ως πλατφόρμες κοινωνικού δανεισμού, στις οποίες δανειστές και δανειζόμενοι μπορούν να αλληλοεπιδράσουν αναμεταξύ τους, δίχως την εμπλοκή του αυστηρά εποπτευόμενου χρηματοοικονομικού συστήματος.

Μια τέτοια περίπτωση μη εποπτευόμενου χρηματοδοτικού οργανισμού-πλατφόρμας, αποτελεί και η Home Credit, η εταιρεία από την οποία αντλήθηκαν τα δεδομένα της παρούσας εργασίας. Σε αυτού του είδους τις πλατφόρμες, τόσο η έλλειψη εμπειρίας των δανειστών και η έλλειψη πληροφοριών ή ακόμα χειρότερα οι αβέβαιες ή περιορισμένες πληροφορίες σχετικά με το πιστωτικό προφίλ και την ιστορία των δανειζόμενων, αυξάνει σημαντικά τους κινδύνους χρηματοδότησης.

Για όλους αυτούς τους λόγους, την έλλειψη δηλαδή εποπτείας, τη μη επάρκεια δεδομένων, αλλά και τη μειωμένη εμπειρία των εμπλεκομένων, κρίνεται επιτακτική η ανάγκη δημιουργίας αξιόπιστων μεθόδων υπολογισμού του πιστωτικού σκορ των δανειζόμενων στις διαδικτυακές πλατφόρμες.

Οι αλγόριθμοι μηχανικής μάθησης έχουν τεράστιες δυνατότητες στον τομέα της αξιολόγησης του πιστωτικού κινδύνου λόγω των εξαιρετικών δυνατοτήτων πρόβλεψης και της ταχείας επεξεργασίας τους. Οι δανειστές μπορούν να προσδιορίσουν εάν ένας δανειολήπτης θα αθετήσει την πληρωμή ενός δανείου και να εκτιμήσουν την πιθανότητα αθέτησης με τη χρήση τους.

Σε αυτή όμως της περίπτωση, είναι εξαιρετικά δύσκολος ο σχεδιασμός ενός υποδείγματος πρόβλεψης πιστωτικού κινδύνου εξαιτίας του ότι σε πραγματικά δεδομένα, υπάρχουν συχνά προβλήματα, όπως η ανισορροπία των δεδομένων (class imbalance), οι πολλές διαστάσεις (dimensionality) και ο μεγάλος βαθμός ελλείψεων δεδομένων (missing data). Στο σύνολο των δεδομένων που συλλέγονται, ο αριθμός των αιτούντων με καλή πιστοληπτική ικανότητα είναι συνήθως πολύ μεγαλύτερος από τον αριθμό των κακοπληρωτών και λόγω της έλλειψης επαρκών δεδομένων, η ικανότητα του ταξινομητή να αναγνωρίσει τα μικρά δείγματα είναι ανεπαρκής και είναι δύσκολο να ταξινομηθούν αποτελεσματικά. Δεδομένου ότι η αξιολόγηση της πίστωσης απαιτεί μια ορισμένη κατανόηση των προσωπικών πληροφοριών, θα υπάρχει μεγάλος αριθμός κατηγορικών χαρακτηριστικών, όπως το επάγγελμα, η οικογενειακή κατάσταση. Τέλος στα πραγματικά δεδομένα, είναι αδύνατο να διασφαλιστεί ότι κάθε δείγμα έχει πλήρη δεδομένα. Τα ελλιπή δεδομένα είναι αναπόφευκτα, και η κατάλληλη διαχείριση τους θα έχει αντίκτυπο στην τελική ακρίβεια του μοντέλου.

Σκοπός της μελέτης μας, είναι η διερεύνηση της αποτελεσματικότητας τέτοιων αλγορίθμων στη πρόβλεψη της πιθανότητας αθέτησης των δανειακών υποχρεώσεων των δανειζόμενων (πιστωτικό γεγονός – credit event) μέσα στο χαλαρό μη εποπτευόμενο περιβάλλον των πλατφόρμων εξ αποστάσεως χρηματοδοτήσεων. Θα συγκρίνουμε την απόδοση παραδοσιακών μοντέλων όπως της Logistic Regression και Random Forest και μερικών από τα νέα, πιο εξελιγμένα μοντέλα μηχανικής μάθησης στην πρόβλεψη πιστωτικού κινδύνου και υπολογισμού πιστωτικού σκορ. Τα μοντέλα αυτά ,LightGBM και XGBoost, τα οποία έχουν αυξημένη

προβλεπτική ικανότητα, υπολογιστική αποτελεσματικότητα και μειώνουν αισθητά τη πιθανότητα overfitting, είναι ικανά να υπολογίζουν προβλέψεις χωρίς επίβλεψη, κάτι που μετά την κατάλληλη εκπαίδευση των υποδειγμάτων, τα καθιστά ιδανικά για το offline περιβάλλον λειτουργίας των fintech εταιρειών (Ma, et al. 2018).

2 Βιβλιογραφική διερεύνηση

2.1 Μεθοδολογίες μηχανικής μάθησης για πρόβλεψη πιστωτικού κινδύνου – Επισκόπηση της Βιβλιογραφίας

Μεθοδολογικά, υπάρχουν πληθώρα προσεγγίσεων και μεθόδων βελτιστοποίησης, βάσει των οποίων αξιολογείται ο πιστωτικός κίνδυνος, προσπαθώντας ο καθένας να ισορροπήσει ανάμεσα σε δυο σημαντικές προκλήσεις. Στην ελαχιστοποίηση της πολυπλοκότητας της εκτίμησης και στη βελτιστοποίηση της αξιοπιστίας της μεθοδολογίας.

Στη βιβλιογραφία, ακολουθούνται τρία διακριτά στάδια: α. η προετοιμασία των δεδομένων, β. η επιλογή χαρακτηριστικών (feature selection) και γ. η επιλογή αλγορίθμων (model selection), στο καθένα από τα οποία υπάρχει μεγάλος βαθμός παραμετροποίησης και πληθώρα τεχνικών.

Τα νέα υποδείγματα υπολογισμού της πιθανότητας πτώχευσης, τα οποία χρησιμοποιούν μεθόδους μηχανικής μάθησης και διευρυμένες πηγές δεδομένων, παρουσιάζουν μια μικρή μα αισθητή βελτίωση. Βελτίωση η οποία στην ανάλυση και την αξιολόγηση πιστωτικού κινδύνου, δεν μπορεί να θεωρηθεί διόλου αμελητέα (Montevechi et al, 2024), δεδομένου ότι η παραμικρή αύξηση της προβλεπτικής ικανότητας ενός μοντέλου πιστωτικού κινδύνου (CSM) μπορεί να οδηγήσει σε σημαντική εξοικονόμηση (Baesens et al., 2003; West, 2000). Έχοντας αυτό υπόψη μας, παρά τους ρυθμιστικούς κανονισμούς και την έλλειψη επεξηγησιμότητας των ταξινομητών μηχανικής μάθησης (ML), αυτοί μπορούν επίσης να είναι χρήσιμοι σε διαφορετικά στάδια ανάπτυξης και προετοιμασίας μοντέλων. Συνεπώς, η απόδοση ενός CSM μπορεί να βελτιωθεί μέσω διαφορετικών εργασιών μοντελοποίησης, πέρα από την επιλογή του ταξινομητή και η ML μπορεί να αξιοποιηθεί στη μοντελοποίηση πιστωτικού κινδύνου συμπληρώνοντας παραδοσιακά μοντέλα με διάφορους τρόπους, όπως:

- **Επικύρωση μοντέλων (validation):** δημιουργία πολύπλοκων μοντέλων που λειτουργούν ως σύγκριση με τα πρότυπα μοντέλα.
- **Ποιότητα/ενίσχυση δεδομένων:** βελτίωση και ενίσχυση των συνόλων δεδομένων.
- **Επιλογή χαρακτηριστικών (Feature Selection):** εύρεση επεξηγηματικών χαρακτηριστικών και συνδυασμών αυτών σε μεγάλα σύνολα δεδομένων.
- **Συντονισμός υπερπαραμέτρων (HPT):** αποτελεσματική βοήθεια στη ρύθμιση των παραμέτρων ενός ταξινομητή.

Ως αποτέλεσμα, οι προσεγγίσεις με τη χρήση ML έχουν γίνει αναπόσπαστο κομμάτι της πιστωτικής βαθμολόγησης (Giri et al., 2021) και, την τελευταία δεκαετία, ενσωματώνονται σταδιακά από τράπεζες και τμήματα διαχείρισης κινδύνων εταιριών fintech (Dumitrescu et al., 2021; Hurlin και Pérignon, 2019). Ως απάντηση, έχουν προταθεί πολλές προσεγγίσεις για την αύξηση της απόδοσης πρόβλεψης των CSMs.

Λαμβάνοντας υπόψη την προσέγγιση της μηχανικής μάθησης, η βιβλιογραφία σχετικά με την πιστωτική βαθμολόγηση καταναλωτών (credit scoring) επικεντρώνεται κυρίως σε άρθρα που συγκρίνουν πολυάριθμα CSMs. Αν και τα περισσότερα περιέχουν σύντομες ανασκοπήσεις για τα κύρια θέματα, δεν βρέθηκαν πολλά μεθοδολογικά και αντικειμενικά άρθρα επισκόπησης που να καλύπτουν εκτενώς τη διαδικασία μοντελοποίησης πιστωτικού κινδύνου.

Οι Lin et al. (2011) δημοσίευσαν μια βιβλιογραφική ανασκόπηση για τις μεθόδους ML στην πιστωτική βαθμολόγηση για την πρόβλεψη οικονομικής κρίσης, λαμβάνοντας υπόψη τα πιο διαδεδομένα σύνολα δεδομένων και τα κύρια χαρακτηριστικά τους, καθώς και τις μεθόδους ταξινόμησης και επιλογής χαρακτηριστικών. Οι Lessmann et al. (2015) διεξήγαγαν έρευνα για να συγκρίνουν διάφορους αλγόριθμους ταξινόμησης τελευταίας τεχνολογίας, παρέχοντας μια βιβλιογραφική ανασκόπηση που καλύπτει ταξινομητές, σύνολα δεδομένων και μεθόδους αξιολόγησης. Οι Marqués et al. (2013) δημοσίευσαν μια ανασκόπηση της βιβλιογραφίας σχετικά

με την εφαρμογή της εξελικτικής υπολογιστικής στην πιστωτική βαθμολόγηση. Συγκεκριμένα, οι συγγραφείς επικεντρώνονται στη χρήση της εξελικτικής υπολογιστικής για την ταξινόμηση, την επιλογή χαρακτηριστικών και τον συντονισμό υπερπαραμέτρων. Η εργασία τους καλύπτει επίσης σύνολα δεδομένων, κριτήρια αξιολόγησης απόδοσης και το πρόβλημα της ανισορροπίας των κατηγοριών σε άρθρα που δημοσιεύθηκαν μεταξύ 2000 και 2012. Οι Louzada et al. (2016) παρέχουν μια συστηματική επισκόπηση των ταξινομητών CSM, καλύπτοντας σύνολα δεδομένων και μεθόδους αξιολόγησης. Τέλος, οι Dastile et al. (2020) παρέχουν επίσης μια συστηματική επισκόπηση της βιβλιογραφίας, χρησιμοποιώντας 74 μελέτες σχετικά με σύνολα δεδομένων, μεθόδους ταξινόμησης και επιλογής χαρακτηριστικών, ανισορροπία κατηγοριών, επεξηγησιμότητα και μεθόδους αξιολόγησης. Οι συγγραφείς σημειώνουν ότι, από τις 7 άλλες παρόμοιες ανασκοπήσεις που βρέθηκαν, καμία δεν κάλυπτε την πιστωτική βαθμολόγηση σε αυτόν τον βαθμό, συχνά αγνοώντας τη διαφάνεια του μοντέλου και την ανισορροπία των κατηγοριών, εστιάζοντας κυρίως στη σύγκριση της απόδοσης των μοντέλων.

Οι τρέχουσες απαιτήσεις σε θεσμικό επίπεδο τονίζουν ακόμα περισσότερο την άμεση αναγκαιότητα προτυποποίησης της γνώσης σχετικά από το θέμα, καθώς μεγάλο μέρος του κανονιστικού πλαισίου προέρχεται από τοπικές νομοθετικές πρωτοβουλίες (προσέγγιση από κάτω προς τα πάνω), και όχι από την Επιτροπή της Βασιλείας, η οποία μόλις πρόσφατα ενίσχυσε περαιτέρω την προτεραιοποίηση της αντιμετώπισης του φαινομένου της ψηφιοποίησης της χρηματοοικονομίας και της ανάδυσης των μεθόδων μηχανικής μάθησης (ML).

2.2 Εισαγωγή στη Μηχανική Μάθηση

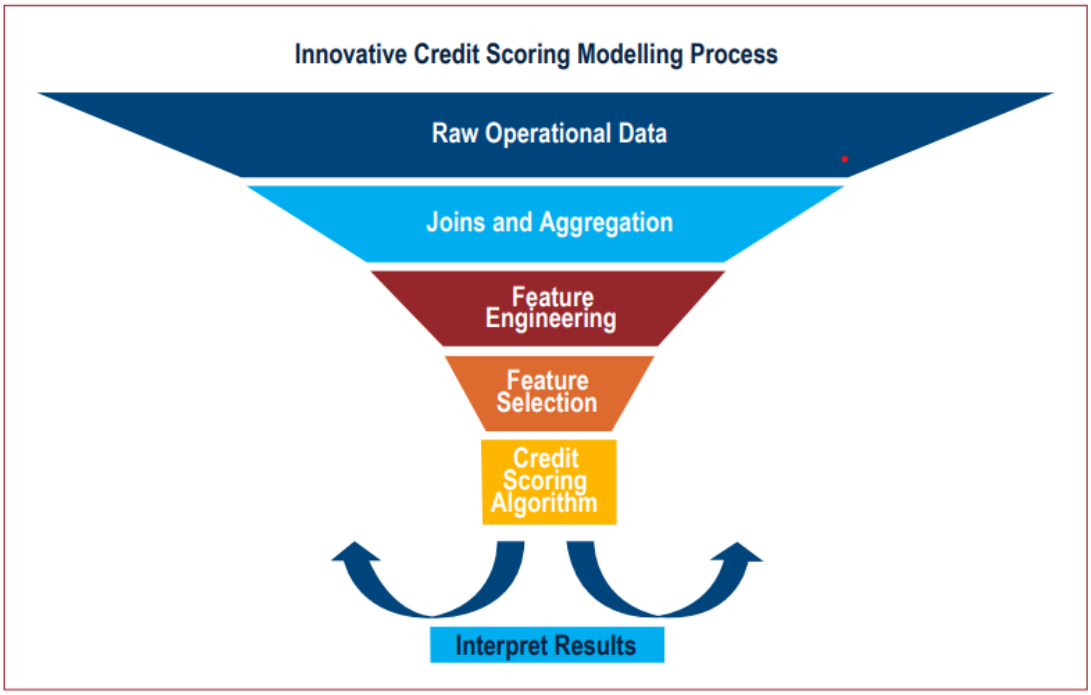
Με βάση την απόδοση της Παγκόσμιας Τράπεζας (2019), με τον όρο Τεχνητή Νοημοσύνη (Artificial Intelligence – AI), αναφερόμαστε στην εφαρμογή υπολογιστικών εργαλείων στην υλοποίηση εργασιών που παραδοσιακά απαιτούν ανθρώπινη εμπειρία και εργασία (SAS, 2019). Η AI επιτρέπει στους υπολογιστές να μάθουν από την εμπειρία, να προσαρμοστούν σε νέα δεδομένα και να εκτελούν διεργασίες που μέχρι τώρα πραγματοποιούνταν μόνο από ανθρώπους (FSB, 2017). Τα περισσότερα δημοφιλή παραδείγματα AI σήμερα - από αυτόνομες μηχανές μέχρι υπεράνθρωπους γιατρούς- εξαρτώνται αποκλειστικά στις μεθόδους βαθιάς μάθησης (deep learning) και επεξεργασίας φυσικής γλώσσας (NPL). Αυτές οι τεχνικές εκμεταλλεύονται την ικανότητα των υπολογιστών να εκτελούν διεργασίες, όπως η υπολογιστική όραση και τα chatbot, μαθαίνοντας από της εμπειρίες τους. Στις μέρες μας η εξέλιξη της τεχνητής νοημοσύνης γίνεται δυνατή βάση της ραγδαίας εξέλιξης σε βασικές τεχνολογίες όπως η υπολογιστική ισχύς, η αποθήκευση και μεταφορά πληροφορίας και πρωτοποριακών αλγορίθμων. Με την χρήση αυτών των τεχνολογιών οι υπολογιστές μπορούν να εκπαιδευτούν στο να εκτελούν συγκεκριμένες διεργασίες μέσω της επεξεργασίας και της αναγνώρισης μοτίβων σε δεδομένα διαφορετικού τύπου και προέλευσης. Η τεχνητή νοημοσύνη είναι ένα πολύ μεγάλο γνωστικό πεδίο και η μηχανική μάθηση αποτελεί μια υποκατηγορία του.

Με τον όρο μηχανική μάθηση ορίζουμε μια μέθοδο σχεδιασμού μιας σειράς ενεργειών με σκοπό την επίλυση ενός προβλήματος, γνωστή ως αλγόριθμος, ο οποίος βελτιστοποιείται αυτόματα μέσω της απόκτησης εμπειρίας με ελάχιστη ανθρώπινη παρέμβαση (SAS, 2019). Αυτές οι τεχνικές χρησιμοποιούνται για την εύρεση πολύπλοκων μοτίβων σε ένα μεγάλο όγκο δεδομένων που αντλούνται από όλο και αυξανόμενα ποικίλες και καινοτόμες πηγές (SAS, 2019).

Σε ένα γενικό επίπεδο, η εφαρμογή αλγορίθμων μηχανικής μάθησης για την πιστωτική βαθμολόγηση περιλαμβάνει την ακόλουθη διαδικασία υψηλού επιπέδου, η οποία μπορεί να διασπαστεί περαιτέρω σε πολλαπλές αποφάσεις:

- Πρόσβαση στα ακατέργαστα δεδομένα.
- Συνδυασμός, σύνδεση και συγκέντρωση των δεδομένων .
- Δημιουργία χαρακτηριστικών, είτε χειροκίνητα από ειδικούς του γνωστικού πεδίου είτε με αυτόματες προσεγγίσεις.
- Επιλογή χρήσιμων χαρακτηριστικών.
- Εφαρμογή του αλγορίθμου μηχανικής μάθησης στο σύνολο δεδομένων εκπαίδευσης.
- Ερμηνεία και αξιολόγηση των αποτελεσμάτων.

Οι καινοτόμες μέθοδοι πιστωτικής βαθμολόγησης περιλαμβάνουν τεχνικές εποπτευόμενης, μη εποπτευόμενης και ημι-εποπτευόμενης μάθησης.



Εικόνα 2-1: Μέθοδος μοντελοποίησης πιστωτικού σκορ (Πηγή: Παγκόσμια Τράπεζα, 2019)

Στην **εποπτευόμενη μάθηση**, στην οποία και θα επικεντρωθούμε, ο αλγόριθμος αναπτύσσεται χρησιμοποιώντας δεδομένα που περιέχουν μια ετικέτα/στόχο (εξαρτημένη μεταβλητή) και τα χαρακτηριστικά (ανεξάρτητες μεταβλητές). Ο αλγόριθμος στη συνέχεια προβλέπει τις μελλοντικές ή άγνωστες τιμές της ετικέτας, χρησιμοποιώντας αυτά τα χαρακτηριστικά . Για παράδειγμα, ένα σύνολο δεδομένων με αντισυμβαλλόμενους μπορεί να περιέχει ετικέτες σε ορισμένα δεδομένα που να υποδεικνύουν ποιοι είναι αυτοί που έχουν αθετήσει την πληρωμή τους και ποιοι όχι. Ο αλγόριθμος θα μάθει έναν γενικό κανόνα ταξινόμησης που θα χρησιμοποιήσει για να προβλέψει τις ετικέτες για τις υπόλοιπες παρατηρήσεις στο σύνολο δεδομένων. Ορισμένες από τις εποπτευόμενες τεχνικές περιλαμβάνουν την παλινδρόμηση(regression), τα τυχαία δάση (random forests) ,τα δέντρα απόφασης (decision trees), το gradient boosting και τα βαθιά νευρωνικά δίκτυα (DNN).

2.3 Σχετικά με τον Τύπο των Δεδομένων

Η ανάδυση μη τραπεζικών οντοτήτων η οποίες παρέχουν μικροπιστώσεις εκτός τραπεζικής εποπτείας, δημιούργησε την ανάγκη συλλογής και χρησιμοποίησης μη συμβατικών τύπων δεδομένων. Με βάση τη κατηγοριοποίηση της Παγκόσμιας Τράπεζας (2019), την οποία και ακολουθεί γενικότερα η βιβλιογραφία (Montevechi et al.,2024), υπάρχουν δύο τύποι δεδομένων, τα παραδοσιακά, καθώς και τα εναλλακτικά.

Οι τύποι των δεδομένων που χρησιμοποιούνται για τη πιστωτική βαθμολόγηση προέρχονται από ποικίλες και πολυδιάστατες πηγές (Εικόνα 2-2). Κατά παράδοση χρησιμοποιούνται πιστωτικά δεδομένα όπως το ποσό του δανείου, η διάρκεια του και το είδος του, καθώς και οι εγγυήσεις, το ιστορικό των πληρωμών όπως οι καθυστερημένες πληρωμές, τα υπόλοιπα των οφειλών, η διάρκεια του πιστωτικού ιστορικού, οι νέες πιστώσεις και τα είδη των πιστώσεων. Όλα αυτά τα δεδομένα λαμβάνονται υπόψιν στην πιστωτική βαθμολόγηση και αποτελούν τους δείκτες πρόθεσης και ικανότητας αποπληρωμής (Márquez, 2008). Αυτά τα συμβατικά σύνολα δεδομένων συνήθως είναι ιδιοκτησία των παρόχων πιστωτικών υπηρεσιών (CSPs) (Trujillo et al., 2015).

Data category	Data type	Credit scoring application
Traditional	Bank transactional data	Records of late payments on current and past credit, current loan amounts and loan purpose, credit history
Traditional	Credit bureau checks	Number of credit inquiries
Traditional	Commercial data	Financial statements, number of working capital loans, and others
Alternative	Utilities data	Steady records of on-time payments as possible consideration as an indicator of creditworthiness
Alternative	Social media	Social media data with possible insights on consumer's lifestyle
Alternative	Mobile applications	Mobile payment systems with possible view on the consumer's behavior
Alternative	Online transactions	Granular transactional data with possible detailed insights on spending patterns
Alternative	Behavioral data	Psychometrics, form filling

Εικόνα 2-2: Τύποι παραδοσιακών και εναλλακτικών δεδομένων.(Πηγή: Παγκόσμια Τράπεζα, 2019)

Η Global Partnership for Financial Inclusion ορίζει τα εναλλακτικά δεδομένα ως έναν ευρύ όρο που περιγράφει τον τεράστιο όγκο δεδομένων που παράγεται από τη συνεχώς αυξανόμενη χρήση ψηφιακών εργαλείων και πληροφοριακών συστημάτων (GPFI, 2018). Εναλλακτικές πηγές μπορεί να περιλαμβάνουν δεδομένα συναλλαγών σε πραγματικό χρόνο, δεδομένα από κινητά και εφαρμογές, μέσα κοινωνικής δικτύωσης και υπηρεσίες κοινής ωφέλειας. Επιπλέον, βιομετρικά και ψυχομετρικά δεδομένα, δεδομένα περιήγησης και αξιολογήσεις στο διαδίκτυο, δεδομένα ακινήτων η προμηθευτών και μεταφορικών μπορεί να παρέχουν μια πληθώρα πληροφοριών. Αυτές οι πηγές συχνά αναφέρονται ως "εναλλακτικά δεδομένα" (ICCR, 2018).

Τα δομημένα δεδομένα συνήθως αποθηκεύονται σε παραδοσιακές βάσεις δεδομένων. Για παράδειγμα, μια δομημένη βάση δεδομένων μπορεί να καταγράφει τις καθημερινές συναλλαγές μιας εταιρείας. Τα μη δομημένα δεδομένα συνήθως δεν έχουν μια προκαθορισμένη σειρά όπως ελεύθερο κείμενο, εικόνες, βίντεο, ήχο. Τα ημιδομημένα δεδομένα δεν ακολουθούν την μορφή των δομημένων γιατί μπορεί να περιέχουν ετικέτες ή δείκτες. Η χρησιμότητα, η αντικειμενικότητα και η ποιότητα των μη δομημένων δεδομένων δεν έχει ακόμη αποδειχθεί πλήρως ότι συμβάλουν στη βελτίωση των μεθόδων πιστωτικής βαθμολόγησης (CGFS και FSB, 2017). Η νόμιμη χρήση των μη δομημένων δεδομένων έχει προοπτικές για περαιτέρω ανάλυση. Πολλοί ρυθμιστικοί φορείς έχουν δημιουργήσει ελεγχόμενα απομονωμένα περιβάλλοντα δοκιμών(sandbox) για να υποστηρίξουν την εξέλιξη καινοτομιών στην επιστήμη των δεδομένων, όπως στην Αυστραλία, τη Σιγκαπούρη, το Ηνωμένο Βασίλειο και στην Ταιβάν (FCA, 2019).

Τα σύγχρονα συστήματα πιστωτικής βαθμολόγησης έχουν την δυνατότητα να συλλέξουν δεδομένα από μια πληθώρα πηγών σε δομημένη, μη δομημένη και ημιδομημένη μορφή. Οι νέες πηγές δεδομένων που χρησιμοποιούνται περιλαμβάνουν τα αναλυτικά δεδομένα των συναλλαγών, δεδομένα από τη χρήση κινητών, γεωεντοπισμού και στοιχεία πληρωμών από άλλες πήγες όπως υπηρεσίες κοινής ωφέλειας. Συχνοί τύποι εναλλακτικών δεδομένων που χρησιμοποιούνται στις μεθόδους πιστωτικής βαθμολόγησης μπορεί να προέρχονται από τα αναλυτικά δεδομένα των συναλλαγών, δεδομένα μέσων κοινωνικής δικτύωσης, δεδομένα κινητής τηλεφωνίας και δεδομένα συμπεριφοράς.

2.3.1 Αναλυτικά δεδομένα των συναλλαγών (granular transactional data)

Τα αναλυτικά δεδομένα των συναλλαγών καταναλωτών και επιχειρήσεων μπορεί να περιλαμβάνει τις αναλυτικές κινήσεις λογαριασμών η πιστωτικών καρτών. Συνήθως παρέχουν μια οργανωμένη και καθαρή εικόνα (given their operational nature) λόγω της λειτουργικής χρήσης

τους και πληροφορίες για τη συμπεριφορά τους μέσω του ιστορικού πληρωμών (Siddiqi 2017). Άλλοι τύποι δεδομένων συναλλαγών περιλαμβάνουν δεδομένα ηλεκτρονικού εμπορίου και δεδομένα από λογιστικά συστήματα. Για τις μικρομεσαίες επιχειρήσεις (SME) και τις εταιρικές οντότητες, τα πιο λεπτομερή δεδομένα συναλλαγών έχουν αποδειχθεί χρήσιμα στο πλαίσιο της πιστωτικής βαθμολόγησης. Αρκετοί CSPs έχουν αναπτύξει εργαλεία για την επεξεργασία συναλλαγών σε πραγματικό χρόνο από τους λογαριασμούς διαχείρισης σε λεπτομερή στοιχεία εσόδων και εξόδων, σε συνδυασμό με ανάλυση για τη δημιουργία, για παράδειγμα, απλουστευμένων οικονομικών καταστάσεων και δεικτών οικονομικής βιωσιμότητας. Τα δεδομένα συναλλαγών μπορούν να προσφέρουν σημαντικά πλουσιότερες και πιο επίκαιρες πληροφορίες σχετικά με τις επιδόσεις των εταιρειών από ό,τι οι ετήσιοι λογαριασμοί (Barasch 2017), και η ίδια λογική ισχύει και για την πιστοληπτική ικανότητα ενός ατόμου.

2.3.2 Mobile Data - Δεδομένα κινητής τηλεφωνίας

Η σημαντική αύξηση της χρήσης των έξυπνων τηλεφώνων έχει δημιουργήσει μια μεγάλη ποικιλία δομημένων και μη δομημένων δεδομένων που μπορούν να χρησιμοποιηθούν για την αξιολόγηση των προτύπων συμπεριφοράς των χρηστών κινητών τηλεφώνων. Οι εφαρμογές κινητών τηλεφώνων μπορούν να συλλέγουν δεδομένα, όπως μετακινήσεις , γεωεντοπισμό και δεδομένα συναλλαγών. Τα αυτά δεδομένα μπορούν να επιτρέψουν στις εφαρμογές κινητών τηλεφώνων να εκτελούν τους απαιτούμενους πιστωτικούς ελέγχους, τους οποίους οι παραδοσιακοί πάροχοι υπηρεσιών CSP βρίσκουν ως πρόκληση (Grab, 2018). Οι χρήστες αυτών ενδέχεται να μην γνωρίζουν ότι τα προσωπικά τους δεδομένα χρησιμοποιούνται για τη βαθμολόγηση της πιστοληπτικής ικανότητας.

2.3.3 Social Media Data - Δεδομένα κοινωνικής δικτύωσης

Ορισμένες ερευνητικές μελέτες έχουν υποδείξει ότι ο αριθμός των αναρτήσεων στα μέσα κοινωνικής δικτύωσης και η συχνότητά τους μπορεί να οδηγήσει σε καλύτερη κατανόηση του τρόπου ζωής των καταναλωτών, των δαπανών τους και της προθυμίας τους να αποπληρώσουν το χρέος (Blazquez and Domenech, 2018). Μια περαιτέρω προέκταση της ανάλυσης των δεδομένων των μέσων κοινωνικής δικτύωσης είναι η δυνατότητα ανάλυσης του δικτύου και των συνδέσεων ενός καταναλωτή. Το δίκτυο και η δραστηριότητα μεταξύ των συνδέσεων έχουν αναφερθεί ότι παρέχουν χρήσιμες πληροφορίες στην περίπτωση που ο υποψήφιος δεν έχει πιστωτικό ιστορικό (Rusli, 2013). Η ποιότητα και ο γενικός χαρακτήρας των δεδομένων των μέσων κοινωνικής δικτύωσης δημιουργούν ανησυχίες σχετικά με τη προστασία των προσωπικών δεδομένων και το ενδεχόμενο απάτης, επειδή τα δεδομένα δεν έχουν ληφθεί απευθείας από τον καταναλωτή.

2.3.4 Λογαριασμοί Κοινής Ωφελείας

Μια άλλη χρήσιμη πηγή για τη βαθμολόγηση της πιστοληπτικής ικανότητας είναι η ανάλυση του ιστορικού πληρωμών λογαριασμών κοινής ωφέλειας. Η πρακτική αυτή βασίζεται στην παραδοχή ότι η ιστορική συμπεριφορά πληρωμών παρέχει πληροφορίες σχετικά με την ικανότητα του καταναλωτή να εξοφλεί τις οφειλές του(World Bank, 2019).

2.3.5 Αποθήκευση και Διαχείριση Δεδομένων

Οι καινοτομίες στην τεχνολογία των υπολογιστών και η ζήτηση δεδομένων οδηγούν σε συνεχείς βελτιώσεις στη συλλογή, προετοιμασία, αποθήκευση, ανάλυση και διανομή δεδομένων. Όταν τα δεδομένα καθαρίζονται και μετασχηματίζονται, συνδυάζονται με άλλες πηγές και διατηρούνται ιστορικά, μπορούν να γίνουν πολύ ισχυρά στοιχεία για ανάλυση. Οι εξελίξεις στις τρέχουσες τεχνολογίες επεξεργασίας δεδομένων και πληροφορικής έχουν οδηγήσει σε δραστική μείωση του κόστους αποθήκευσης και επεξεργασίας, επιτρέποντας πιο αποδοτικά μέσα συλλογής, διαχείρισης και ανάλυσης εξαιρετικά μεγάλων συνόλων δεδομένων. Τα δεδομένα μπορούν να παράγονται από συστήματα και αλληλεπιδράσεις μεταξύ ανθρώπων και συστημάτων για επιχειρησιακούς/Operational σκοπούς. Αν και τα νέα δεδομένα παράγονται για πολλούς λόγους, ο τρόπος με τον οποίο παράγονται και αποθηκεύονται έχει σημαντικές ηθικές και νομικές

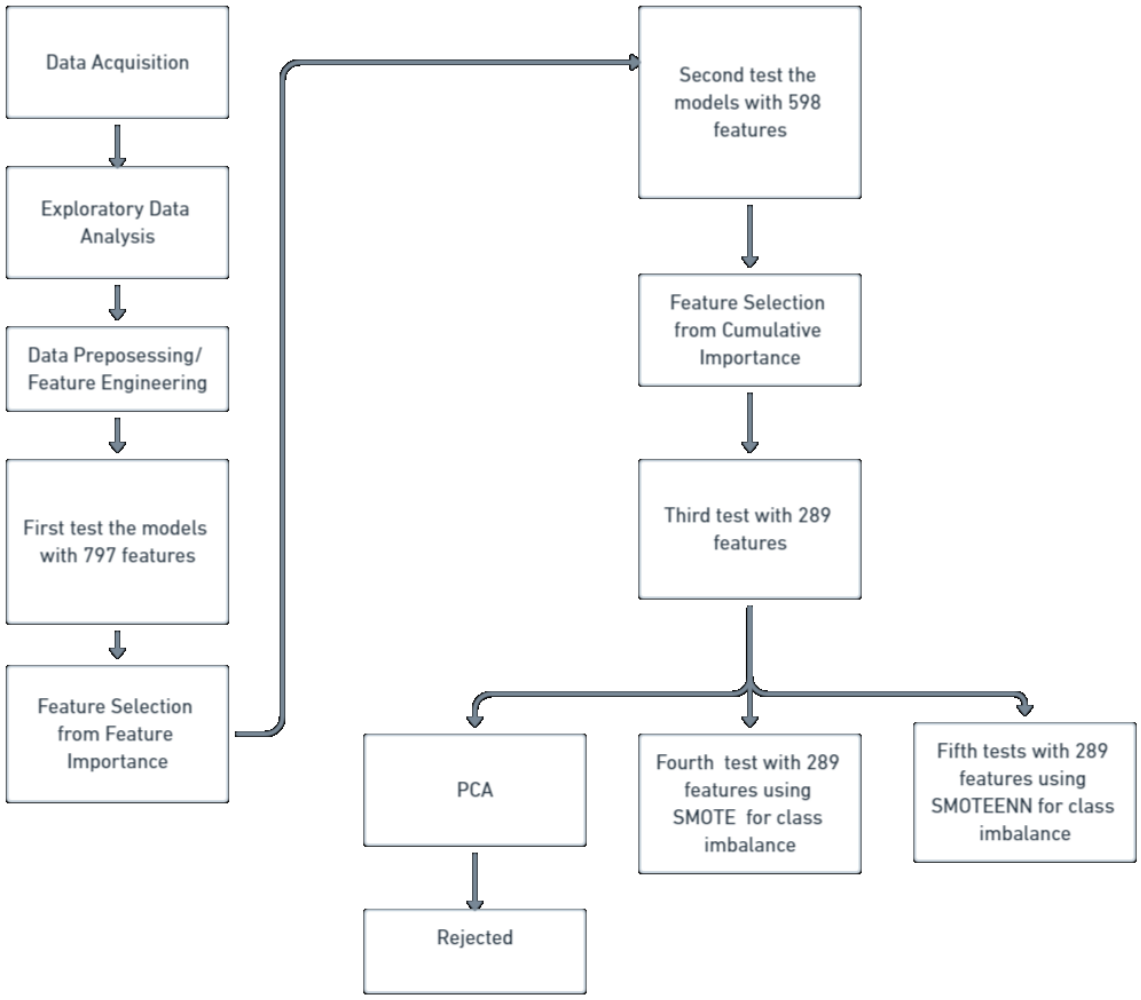
επιπτώσεις. Για παράδειγμα, τα δεδομένα από μια οικονομική συναλλαγή δεν μπορούν να χρησιμοποιηθούν για τους ίδιους σκοπούς με τα προσωπικά δεδομένα από ένα προφίλ σε μια πλατφόρμα κοινωνικής δικτύωσης. Τα δεδομένα που ανήκουν στον δημόσιο τομέα θεωρούνται ιστορικά μη προστατευόμενα. Ο αυξανόμενος όγκος δεδομένων δημιουργεί ευκαιρίες για τρίτους να παρέχουν υπηρεσίες διαχείρισης δεδομένων, όπως η συλλογή, ο καθαρισμός και ο συνδυασμός δεδομένων(World Bank, 2019).

Τα πρόσθετα επίπεδα νέων οργανισμών στην αλυσίδα αξίας των δεδομένων μπορεί να δημιουργήσουν προκλήσεις σχετικά με το ποιος είναι υπεύθυνος και υπόλογος για την ακρίβεια και την ποιότητα των δεδομένων. Μπορεί να καταστεί πρόκληση η αντιστοίχιση των δεδομένων που χρησιμοποιούνται για την τελική απόφαση με την πηγή των δεδομένων. Σε ορισμένες δικαιοδοσίες υπάρχουν κανονιστικές απαιτήσεις σχετικά με την προστασία των προσωπικών δεδομένων (World Bank, 2019). Οι πάροχοι υπηρεσιών CSP που κατέχουν δεδομένα προσωπικού χαρακτήρα θα πρέπει να διαθέτουν πολιτικές και ισχυρές δυνατότητες για να διασφαλίζουν την τήρηση των κανονιστικών απαιτήσεων. Επιπλέον, η παροχή δυνατότητας στους κατόχους δεδομένων να μοιράζονται προσωπικά δεδομένα από τους παρόχους υπηρεσιών στους παρόχους CSP μπορεί να ωφελήσει τα πλαίσια κινδύνου.

Κρίνεται αναγκαίο να επισημάνουμε πως στο σύγχρονο διεθνοποιημένο περιβάλλον, ο εμπλουτισμός των παραδοσιακών βάσεων δεδομένων με εναλλακτικά, με βάση τη βιβλιογραφία παρέχει μοναδικές ιδιότητες (Djeundje et al., 2021), διευρύνοντας την αποτελεσματικότητα των αλγορίθμων μηχανικής μάθησης.

3 Προτεινόμενη Μεθοδολογία

Σκοπός της διπλωματικής εργασίας είναι η σύγκριση της απόδοσης πολλαπλών αλγορίθμων μηχανικής μάθησης για την πρόβλεψη κινδύνου αθέτησης δανείου, αντιμετωπίζοντας προκλήσεις όπως η επιλογή κατάλληλων χαρακτηριστικών και τη μεγάλη ανισορροπία των κατηγοριών. Στην ενότητα αυτή παρουσιάζονται οι τεχνικές και οι μέθοδοι που χρησιμοποιήθηκαν κατά την διάρκεια της μελέτης.



Εικόνα 3-1: Προτεινόμενη μεθοδολογία.

3.1 Απόκτηση Δεδομένων - Data Acquisition

Το πρώτο βήμα της μελέτης είναι η εύρεση του κατάλληλου συνόλου δεδομένων που θα χρησιμοποιηθεί στην εκπαίδευση των μοντέλων. Στην περίπτωση μας τα δεδομένα που χρησιμοποιήθηκαν στη μελέτη προέρχονται από το διαγωνισμό **Home Credit Default Risk**, που δημοσιεύτηκε στην πλατφόρμα Kaggle. Το σύνολο δεδομένων περιλαμβάνει πληθώρα

πληροφοριών για δανειολήπτες, όπως δημογραφικά στοιχεία, ιστορικό συναλλαγών, προηγούμενα δάνεια, υπόλοιπα πιστωτικών καρτών, καθώς και μεταβλητές στόχου που υποδηλώνουν εάν ένας δανειολήπτης κατέληξε σε χρεοκοπία. Αρχικά θα πρέπει να αποκτηθεί εξοικείωση με τη δομή του συνόλου δεδομένων το οποίο περιλαμβάνει την κατανόηση των χαρακτηριστικών που μας δίνονται και του στόχου πρόβλεψης κάτι που θα μας βοηθήσει στην συνέχεια στην περαιτέρω ανάλυση και επεξεργασία των δεδομένων.

3.2 Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA)

Η διερευνητική ανάλυση δεδομένων (EDA) είναι ένα βασικό βήμα για την κατανόηση της κατανομής των δεδομένων , τον εντοπισμό τυχόν ασυνήθιστων ή ακραίων τιμών και τις συσχετίσεις μεταξύ των χαρακτηριστικών. Σε αυτό το στάδιο θα δημιουργηθούν γραφήματα και στατιστικά στοιχεία για να αναλυθούν τα χαρακτηριστικά καθώς και η σχέση τους με τον στόχο πρόβλεψης. Επιπλέον, θα αναλυθεί η ύπαρξη και η κατανομή ελλειπόντων και ακραίων τιμών στα δεδομένα. Σε αυτό το στάδιο, θα ληφθούν αποφάσεις σχετικά με τον χειρισμό αυτών των ελλειπόντων τιμών, όπως η αντικατάστασή τους με μέσες τιμές, η διαγραφή τους ή η χρήση προηγμένων τεχνικών ενίσχυσης δεδομένων (imputation).

3.3 Προεπεξεργασία Δεδομένων και Δημιουργία Χαρακτηριστικών

Αφού ολοκληρωθεί η διερευνητική ανάλυση των δεδομένων και αναγνωριστούν οι βασικές ιδιότητες και τα προβλήματα που υπάρχουν, ξεκινάμε τη διαδικασία της προεπεξεργασίας τους. Η προκαταρκτική επεξεργασία δεδομένων περιλαμβάνει την προετοιμασία των δεδομένων σε κατάλληλη μορφή για την εκπαίδευση των μοντέλων μηχανικής μάθησης. Αυτό αποτελεί ένα κρίσιμο στάδιο της διαδικασίας, καθώς η ποιότητα των δεδομένων που θα εισαχθούν στα μοντέλα επηρεάζει άμεσα την ακρίβεια και την απόδοσή τους.

Αυτό το στάδιο περιλαμβάνει :

3.3.1 Μετατροπή Κατηγορικών Χαρακτηριστικών σε Αριθμητικά (Categorical Encoding)

Τα κατηγορικά χαρακτηριστικά, που περιλαμβάνουν τιμές όπως το φύλο, την ιδιοκτησία αυτοκινήτου ή ακινήτου πρέπει να μετατραπούν σε αριθμητικά. Αυτή η διαδικασία είναι απαραίτητη καθώς τα μοντέλα μηχανικής μάθησης Logistic Regression, Random Forest και XGBoost δεν μπορούν να διαχειριστούν κατηγορικά δεδομένα απευθείας. Οι μέθοδοι που θα χρησιμοποιηθούν είναι η **Label-encoding** με την οποία θα αντικατασταθεί η κάθε κατηγορία η οποία έχει μόνο δυο τιμές με 0 ,1 ενώ για τα υπόλοιπα κατηγορικά που έχουν περισσότερες από δύο κατηγορίες, θα εφαρμόσουμε **One-Hot Encoding**. Αυτή η τεχνική δημιουργεί μια στήλη για κάθε τιμή της κατηγορικής μεταβλητής, με τιμές 0 ή 1, ώστε να μπορεί να χρησιμοποιηθεί σε μοντέλα μηχανικής μάθησης.

3.3.2 Διαχείριση Ακραίων Τιμών (Outliers)

Αρκετά χαρακτηριστικά, όπως το εισόδημα (AMT_INCOME_TOTAL) και οι ημέρες εργασίας (DAYS_EMPLOYED), παρουσίαζαν ακραίες τιμές ,όπως η τιμή 365243 στις μεταβλητές που σχετίζονται με τις ημέρες εργασίας. Αυτές οι τιμές αντικαταστάθηκαν με την κενή τιμή NaN (Not a Number) .

Επίσης στη μεταβλητή CODE_GENDER, παρατηρήθηκε η ύπαρξη της τιμής **XNA**, η οποία πιθανόν υποδηλώνει ότι δεν ήταν γνωστό το φύλο κάποιων πελατών. Η παρουσία της τιμής αυτής εντοπίστηκε σε 4 εγγραφές δεδομένων οι οποίες αποφασίστηκε να αφαιρεθούν από το σύνολο δεδομένων.

3.3.3 Δημιουργία Χαρακτηριστικών (Feature Engineering)

Το **Feature Engineering** είναι μία από τις πιο σημαντικές διαδικασίες στην προεπεξεργασία δεδομένων, καθώς βελτιώνει την απόδοση των μοντέλων με τη δημιουργία νέων χαρακτηριστικών, τα οποία εμπεριέχουν πληροφορίες που μπορεί να μην ήταν προφανείς από τα αρχικά δεδομένα. Στην παρούσα εργασία, θα δημιουργηθούν μερικά νέα χαρακτηριστικά συνδυάζοντας τα ήδη υπάρχοντα και μέσω υπολογισμού συναθροίσεων (aggregations) των αριθμητικών χαρακτηριστικών όπως το άθροισμα, ο μέσος όρος και οι ελάχιστες και μέγιστες τιμές.

Νέα Χαρακτηριστικά

Για να εκφραστούν πιο σύνθετες σχέσεις μεταξύ των μεταβλητών, δημιουργήθηκαν νέες αριθμητικές μεταβλητές. Παραδείγματα τέτοιων μεταβλητών είναι:

- **INCOME_CREDIT_PERC**: Αναλογία του συνολικού εισοδήματος (AMT_INCOME_TOTAL) προς το ποσό του δανείου (AMT_CREDIT). Αυτή η μεταβλητή δίνει πληροφορίες για το πόσο εύκολα μπορεί ένας πελάτης να αποπληρώσει το δάνειό του βάσει του εισοδήματός του.

- **DAYS_EMPLOYED_PERC**: Ποσοστό των ημερών απασχόλησης (DAYS_EMPLOYED) ως προς την ηλικία του πελάτη (DAYS_BIRTH). Αυτή η μεταβλητή εκφράζει την εμπειρία του πελάτη στην αγορά εργασίας σε σχέση με την ηλικία του.

- **PAYMENT_RATE**: Αναλογία της ετήσιας δόσης (AMT_ANNUITY) προς το συνολικό ποσό του δανείου (AMT_CREDIT). Αυτή η μεταβλητή δείχνει το ποσοστό του δανείου που πρέπει να αποπληρώνει ο πελάτης κάθε χρόνο.

Συναθροίσεις (Aggregations)

Για να αξιοποιηθούν πληροφορίες από τα δεδομένα πολλαπλών εγγραφών, όπως αυτά από τα αρχεία bureau.csv και previous_application.csv, χρησιμοποιήθηκαν συναρτήσεις συνάθροισης. Αυτές οι συναθροίσεις επιτρέπουν τη σύγκριση των χαρακτηριστικών πελατών που έχουν υποβάλει πολλαπλές αιτήσεις ή έχουν πολλαπλά δάνεια.

- **Μέσες τιμές (mean)**: Για κάθε πελάτη, υπολογίστηκαν οι μέσες τιμές σε μεταβλητές όπως το ποσό του δανείου (AMT_CREDIT), τον αριθμό πληρωμών (CNT_PAYMENT), και τις ημέρες καθυστέρησης πληρωμών (DAYS_LATE). Οι μέσες τιμές αυτές παρέχουν μια συνολική εικόνα της συμπεριφοράς του πελάτη σε σχέση με δάνεια ή αιτήσεις που έχει υποβάλει στο παρελθόν.

- **Ελάχιστες και μέγιστες τιμές (min, max)**: Εκτός από τις μέσες τιμές, υπολογίστηκαν οι ελάχιστες και μέγιστες τιμές για κάθε πελάτη, όπως ο μέγιστος αριθμός ημερών καθυστέρησης (MAX_DAYS_LATE) ή το ελάχιστο ποσό που δανείστηκε (MIN_AMT_CREDIT).

Συνδυασμός Συναθροίσεων και Νέων Χαρακτηριστικών

Ο συνδυασμός των συναθροίσεων με τα νέα χαρακτηριστικά δίνει μια πιο ολοκληρωμένη εικόνα του κάθε πελάτη και των οικονομικών του συνθηκών. Τα νέα χαρακτηριστικά και οι συναθροίσεις θα ενσωματωθούν στο αντίστοιχο σύνολο δεδομένων τους, έτσι ώστε τα μοντέλα μηχανικής μάθησης να μπορέσουν να λάβουν υπόψη αυτές τις πρόσθετες πληροφορίες κατά την εκπαίδευση.

3.3.4 Clipping και Αντικατάσταση Απεριόριστων Τιμών (Inf)

Μετά τη δημιουργία νέων χαρακτηριστικών από το **Feature Engineering** και τις συναθροίσεις, προέκυψαν πολυάριθμα αριθμητικά χαρακτηριστικά που αφορούσαν τις μέσες, ελάχιστες και μέγιστες τιμές διάφορων μεταβλητών. Αυτά τα νέα χαρακτηριστικά πολλές φορές περιείχαν ακραίες τιμές, που είτε ήταν εξαιρετικά υψηλές είτε άπειρες (Inf).

Η αντιμετώπιση αυτών των τιμών έγινε με δύο βασικές μεθόδους:

- **Clipping**: Ορίστηκε ένα ανώτατο όριο (π.χ., 1 εκατομμύριο) για να περιοριστούν οι εξαιρετικά μεγάλες τιμές που προκύψαν από τις συναθροίσεις. Το clipping εφαρμόστηκε για να εξασφαλιστεί ότι οι ακραίες τιμές δεν θα επηρεάσουν δυσανάλογα τα μοντέλα, ενώ ταυτόχρονα διατηρήθηκαν οι σχετικές πληροφορίες για τους πελάτες.

- **Αντικατάσταση Απεριορίστων Τιμών (Inf) με NaN**: Σε ορισμένες περιπτώσεις, οι συναθροίσεις δημιούργησαν απεριόριστες (Inf) ή αρνητικά απεριόριστες (-Inf) τιμές. Αυτές οι τιμές αντικαταστάθηκαν με NaN, ώστε να μπορέσουμε να τις διαχειριστούμε κατάλληλα στις επόμενες φάσεις της προεπεξεργασίας, όπως το **imputation**.

Ειδικά για τις μεταβλητές που προέκυψαν από συναθροίσεις, όπως οι μέγιστες και ελάχιστες τιμές δανείων και καθυστερήσεων πληρωμών, το clipping ήταν απαραίτητο, διότι ενσωμάτωναν συχνά ακραίες περιπτώσεις που έπρεπε να περιοριστούν για να μη δημιουργηθούν προβλήματα στην απόδοση των μοντέλων.

3.3.5 Συνένωση Δεδομένων (Merging)

Μετά τη δημιουργία των νέων χαρακτηριστικών και την επεξεργασία των τιμών, τα επιμέρους σύνολα δεδομένων συγχωνεύτηκαν με το βασικό σύνολο δεδομένων (application_train.csv) με βάση το κοινό αναγνωριστικό πελάτη (SK_ID_CURR). Αυτή η ένωση επέτρεψε την ενσωμάτωση όλων των σχετικών πληροφοριών από διαφορετικά σύνολα δεδομένων σε ένα ενιαίο σύνολο 797 χαρακτηριστικών που θα χρησιμοποιήσουμε για την εκπαίδευση.

Κατά τη διαδικασία της συνένωσης των δεδομένων και της δημιουργίας νέων χαρακτηριστικών μέσω των συναθροίσεων, προέκυψαν ονόματα στηλών με ειδικούς χαρακτήρες (π.χ., {}, [], :) που μπορούσαν να προκαλέσουν προβλήματα κατά την επεξεργασία και ανάλυση των δεδομένων. Για να διασφαλιστεί η συμβατότητα των δεδομένων, εφαρμόστηκε μια διαδικασία αντικατάστασης των χαρακτήρων αυτών των χαρακτήρων με την κάτω παύλα (_).

Αυτό το βήμα διασφάλισε ότι τα δεδομένα από τις διάφορες πηγές έχουν ενσωματωθεί σωστά και ότι τα ονόματα των στηλών είναι συμβατά, χωρίς ειδικούς χαρακτήρες που θα μπορούσαν να προκαλέσουν σφάλματα.

3.3.6 Αφαίρεση Χαρακτηριστικών με Υψηλά Ποσοστά Κενών Τιμών

Μετά τη συνένωση των δεδομένων από τα διάφορα αρχεία (π.χ., bureau.csv, previous_application.csv) με το βασικό σύνολο δεδομένων (application_train.csv), ακολούθησε η διαδικασία αφαίρεσης των χαρακτηριστικών που περιείχαν μεγάλο ποσοστό κενών τιμών με σκοπό τη μείωση του αριθμού των χαρακτηριστικών που τελικά θα χρησιμοποιηθούν στα μοντέλα.

Αφού εντοπίστηκαν, αποφασίστηκε τα χαρακτηριστικά τα οποία έχουν πάνω από 65% κενές τιμές, να αφαιρεθούν από το σύνολο δεδομένων. Ωστόσο, πριν ληφθεί η τελική απόφαση για την αφαίρεση, χρησιμοποιήθηκαν τα μοντέλα LightGBM και XGBoost , τα οποία μπορούν να μας επιστρέψουν τον βαθμό σημαντικότητας των χαρακτηριστικών που χρησιμοποιήθηκαν (feature importance). Η ανάλυση των αποτελεσμάτων έδειξε ότι τα χαρακτηριστικά με πάνω από 65% κενά δεν συνεισέφεραν στη απόδοση των μοντέλων, καθώς είχαν πολύ χαμηλή ή μηδενική σημασία.

Με βάση αυτή την αξιολόγηση, αποφασίστηκε η αφαίρεση των χαρακτηριστικών με υψηλά ποσοστά κενών τιμών, διότι δεν αναμενόταν να συμβάλουν στην ακρίβεια ή την απόδοση των μοντέλων.

3.3.7 Αντικατάσταση Κενών Τιμών και Κανονικοποίηση (Imputation & Scaling)

Αφού αφαιρέθηκαν τα χαρακτηριστικά με υψηλά ποσοστά κενών τιμών, ακολούθησε η διαδικασία της αντικατάστασης (imputation) των υπολειπόμενων κενών τιμών και της κανονικοποίησης (scaling) των χαρακτηριστικών.

Ορισμένα από τα μοντέλα που χρησιμοποιούμε, όπως η Logistic Regression και το Random Forest, δεν μπορούν να διαχειριστούν κενές τιμές. Για αυτόν τον λόγο, οι κενές τιμές αντικαταστάθηκαν με τη διάμεσο (median) κάθε χαρακτηριστικού. Εναλλακτικά, μοντέλα όπως το LightGBM και το XGBoost μπορούν να διαχειριστούν κενές τιμές αυτόματα, χωρίς την ανάγκη αντικατάστασής τους. Παρ' όλα αυτά, για λόγους συνέπειας και ομοιομορφίας στα δεδομένα, έγινε η επιλογή να εφαρμόσουμε imputation σε όλα τα μοντέλα.

Τέλος επειδή το μοντέλο της Logistic Regression είναι ευαίσθητο στις μεγάλες διακυμάνσεις τιμών προβήκαμε στην κανονικοποίηση τους. Για να διασφαλιστεί η σωστή εκπαίδευση αυτών των μοντέλων θα εφαρμοστεί Min-Max Scaling, το οποίο κλιμακώνει όλα τα αριθμητικά χαρακτηριστικά στην περιοχή [0, 1]. Αυτό επιτρέπει στα μοντέλα να λειτουργούν καλύτερα, αποφεύγοντας προβλήματα όπου χαρακτηριστικά με υψηλές τιμές κυριαρχούν στα υπόλοιπα. Μοντέλα όπως το LightGBM και το XGBoost δεν απαιτούν κανονικοποίηση, καθώς βασίζονται σε δέντρα αποφάσεων, τα οποία δεν επηρεάζονται από τη διαφορετική κλίμακα των χαρακτηριστικών. Παρ' όλα αυτά, εφαρμόστηκε κανονικοποίηση για λόγους συνοχής στα δεδομένα. Σε αυτό το σημείο το σύνολο των δεδομένων μας είναι έτοιμο για την εκπαίδευση των μοντέλων.

3.3.8 Περεταίρω Μείωση Χαρακτηριστικών Βάση Cumulative Feature Importance του LightGBM

Μετά την πρώτη δοκιμή των τεσσάρων μοντέλων (Logistic Regression, Random Forest, XGBoost, και LightGBM) με τα 598 χαρακτηριστικά, το LightGBM παρουσίασε τα καλύτερα αποτελέσματα από τα υπόλοιπα μοντέλα. Βασιζόμενοι σε αυτή την παρατήρηση, αποφασίσαμε το χρησιμοποιήσουμε ως οδηγό για την περαιτέρω μείωση των χαρακτηριστικών, ώστε να βελτιώσουμε την αποδοτικότητα και την απόδοση των μοντέλων.

Για την επιλογή των πιο σημαντικών χαρακτηριστικών, χρησιμοποιήθηκε η έννοια της αθροιστικής σημαντικότητας των χαρακτηριστικών (cumulative feature importance). Ο βαθμός σημαντικότητας των χαρακτηριστικών (feature importance) από το LightGBM αποδίδει έναν αριθμητικό δείκτη σε κάθε χαρακτηριστικό, εκφράζοντας το πόσο συνεισφέρει στη βελτίωση των προβλέψεων του μοντέλου. Αφού υπολογίσαμε την αθροιστική σημαντικότητα των χαρακτηριστικών, καθορίστηκε ένα όριο (threshold) στο 95% της συνολικής σημαντικότητας και διατηρήσαμε τα χαρακτηριστικά που συνεισέφεραν σε αυτό το όριο, ενώ αφαιρέθηκαν τα υπόλοιπα χαρακτηριστικά που είχαν μικρή ή μηδενική συνεισφορά.

Με αυτόν το τρόπο ο συνολικός αριθμός χαρακτηριστικών μειώθηκε στα 289, διατηρώντας μόνο τα πιο σημαντικά χαρακτηριστικά. Αυτό το βήμα είχε ως στόχο τη βελτίωση της αποδοτικότητας των μοντέλων, μειώνοντας τα περιττά χαρακτηριστικά και εστιάζοντας σε εκείνα που είχαν ουσιαστική συνεισφορά στις προβλέψεις.

3.3.9 Μείωση Χαρακτηριστικών μέσω της Ανάλυσης Κύριων Συνιστωσών (PCA)

Μετά τη δεύτερη δοκιμή των μοντέλων με τα 289 χαρακτηριστικά, αποφασίστηκε να εφαρμοστεί η **Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA)** με στόχο τη μείωση των χαρακτηριστικών και τη συμπίεση της πληροφορίας σε λιγότερες διαστάσεις, χωρίς σημαντική απώλεια πληροφορίας. Η επιλογή της PCA έγινε με την ελπίδα ότι η μείωση των διαστάσεων θα βελτίωνε την αποδοτικότητα των μοντέλων.

Η PCA εφαρμόστηκε διατηρώντας τις συνιστώσες που αντιπροσώπευαν το 95% της συνολικής πληροφορίας. Αυτή η μέθοδος δημιουργεί νέες συνιστώσες, γραμμικούς συνδυασμούς των αρχικών χαρακτηριστικών, με στόχο τη διατήρηση της μέγιστης δυνατής πληροφορίας σε λιγότερες διαστάσεις.

Παρά την αναμενόμενη βελτίωση, τα αποτελέσματα έδειξαν ότι η χρήση της δεν οδήγησε σε σημαντική βελτίωση της απόδοσης των μοντέλων. Αντίθετα, παρατηρήθηκε ότι η απόδοση, ειδικά σε όρους AUC και F1-score, μειώθηκε σε σχέση με τη δεύτερη δοκιμή χωρίς PCA. Αυτό πιθανώς οφείλεται στο ότι οι νέες συνιστώσες που δημιουργήθηκαν από την PCA δεν διατήρησαν αρκετά καλά την πληροφορία που σχετίζεται με την ταξινόμηση. Ως εκ τούτου, αυτή η μέθοδος δεν θεωρήθηκε χρήσιμη για τη συγκεκριμένη εφαρμογή, οπότε συνεχίσαμε με το προηγούμενο σύνολο δεδομένων.

3.4 Επιλογή Μοντέλων Μηχανικής Μάθησης

Η επιλογή των κατάλληλων αλγορίθμων μηχανικής μάθησης είναι ένα από τα πιο σημαντικά στάδια της μελέτης καθώς υπάρχει μια πληθώρα αλγορίθμων που χρησιμοποιούνται και καθένας από αυτούς είναι κατάλληλος για διαφορετικά προβλήματα. Το συγκεκριμένο πρόβλημα που καλούμαστε να λύσουμε αποτελεί ένα πρόβλημα δυαδικής ταξινόμησης με μεγάλη ανισορροπία των κατηγοριών.

Τα μοντέλα που θα χρησιμοποιήσουμε, πέρα από το απλό μοντέλο της λογιστικής παλινδρόμησης (logistic regression) και των τυχαίων δέντρων (random forest), είναι δυο νέοι καινοτόμοι αλγόριθμοι μηχανικής μάθησης, τα μοντέλα XGBoost και LightGBM, όπως έχουν εφαρμοσθεί από τους Xiaojun et al. (2018) με σκοπό την πρόβλεψη της πιθανότητας αθέτησης των υποχρεώσεων τους από δανειστές βασισμένοι σε μια βάση δεδομένων πραγματικών peer-to-peer συναλλαγών από την πλατφόρμα Lending Club, και είναι ιδιαίτερα διαδεδομένοι σε διαγωνισμούς στη πλατφόρμα Kaggle. Οι αλγόριθμοι αυτοί έχουν μια εξαιρετικά εύκολη πρακτική εφαρμογή, πολύ καλή απόδοση και μειωμένο overfitting. Το υπόδειγμα LightGBM φαίνεται πως υπερτερεί του XGBoost, με ακρίβεια κοντά στο 80% και σφάλμα της τάξεως του 20%. Και οι δυο αλγόριθμοι βασίζονται πάνω στη θεωρία του GBDT (Gradient Boosting Decision Tree).

3.5 Χρήση Stratified 10-Fold Cross Validation και Τεχνικών Αντιμετώπισης Ανισορροπίας

Λόγω της έντονης ανισορροπίας των κατηγοριών στο σύνολο δεδομένων (91,93% μη-αθέτηση και 8,07% αθέτηση), η επιλογή κατάλληλων τεχνικών αξιολόγησης και διαχείρισης της ανισορροπίας ήταν απαραίτητη. Για την αντιμετώπιση αυτών των προκλήσεων, χρησιμοποιήθηκαν οι εξής μέθοδοι:

3.5.1 Stratified 10-Fold Cross Validation

Η μέθοδος **Stratified 10-Fold Cross Validation** επιλέχθηκε για την αξιολόγηση των μοντέλων, καθώς εξασφαλίζει ότι η αναλογία των κατηγοριών παραμένει σταθερή σε κάθε τμήμα (fold) της επικύρωσης. Με αυτή τη μέθοδο, το σύνολο των δεδομένων χωρίζεται σε 10 διαφορετικά τμήματα. Σε κάθε επανάληψη, ένα τμήμα χρησιμοποιείται ως σύνολο δοκιμών και τα υπόλοιπα 9 τμήματα χρησιμοποιούνται για την εκπαίδευση του μοντέλου. Η Stratified K-Fold διασφαλίζει ότι κάθε τμήμα περιέχει την ίδια αναλογία δειγμάτων των κατηγοριών. Αυτό είναι ιδιαίτερα σημαντικό όταν υπάρχει έντονη ανισορροπία, όπως στην περίπτωσή μας.

Παρέχει μια πιο αντικειμενική εκτίμηση της απόδοσης του μοντέλου, καθώς η απόδοση μετριέται σε διαφορετικά υποσύνολα του συνόλου δεδομένων. Η μέθοδος εφαρμόστηκε σε όλα τα μοντέλα και υπολογίστηκαν παράμετροι αξιολόγησης όπως η Accuracy, η καμπύλη ROC-AUC, η F1-score, η Precision, και η Recall σε κάθε fold. Η τελική απόδοση υπολογίστηκε ως ο μέσος όρος των αποτελεσμάτων από τα 10 folds, εξασφαλίζοντας μια αξιόπιστη εκτίμηση.

3.5.2 SMOTE (Synthetic Minority Oversampling Technique)

Για την αντιμετώπιση της ανισορροπίας στις κατηγορίες, εφαρμόστηκε η τεχνική **SMOTE** (Synthetic Minority Oversampling Technique). Η SMOTE δημιουργεί συνθετικά δείγματα για την μικρότερη κατηγορία, βασισμένα στα υπάρχοντα δεδομένα, ώστε να εξισορροπήσει την αναλογία μεταξύ των κατηγοριών. Αντί να αντιγράφει τα υπάρχοντα δεδομένα της μικρότερης κατηγορίας, η SMOTE δημιουργεί νέα συνθετικά δείγματα υπολογίζοντας τη διαφορά μεταξύ ενός υπάρχοντος δείγματος και των κοντινότερων γειτόνων του. Με βάση αυτή τη διαφορά, παράγει νέα δεδομένα που δεν είναι απλώς αντίγραφα, αλλά βασισμένα στη γεωμετρία των δεδομένων. Η SMOTE χρησιμοποιήθηκε σε συνδυασμό με τη μέθοδο Stratified K-Fold Cross Validation. Εφαρμόστηκε μέσα σε κάθε fold, ώστε το μοντέλο να εκπαιδευτεί σε ισορροπημένα δεδομένα και να αξιολογηθεί σε μη-τροποποιημένα δεδομένα. Αυτό διασφαλίζει ότι η εκπαίδευση του μοντέλου γίνεται σε ένα ισορροπημένο δείγμα, ενώ η αξιολόγηση γίνεται στα αρχικά δεδομένα.

3.5.3 SMOTEENN (SMOTE + Edited Nearest Neighbors)

Για την περαιτέρω βελτίωση της απόδοσης στα δεδομένα με έντονη ανισορροπία, εφαρμόστηκε η τεχνική **SMOTEENN**, η οποία συνδυάζει τη δημιουργία συνθετικών δειγμάτων από τη μέθοδο **SMOTE** με τη διαδικασία αφαίρεσης δεδομένων χρησιμοποιώντας τον αλγόριθμο Edited Nearest Neighbors (ENN). Η SMOTE χρησιμοποιείται αρχικά για τη δημιουργία νέων δειγμάτων για τη μικρότερη κατηγορία, όπως περιεγράφηκε προηγουμένως και στη συνέχεια, η ENN αφαιρεί δείγματα από την πλειοψηφική κατηγορία (μη-αθέτηση δανείου). Με αυτόν τον τρόπο, μειώνονται τα δεδομένα της πλειοψηφικής κατηγορίας, βελτιώνοντας την απόδοση των μοντέλων. Η SMOTEENN εφαρμόστηκε επίσης μέσα σε κάθε fold του Stratified K-Fold Cross Validation. Αυτό επέτρεψε στο μοντέλο να εκπαιδευτεί σε δεδομένα που είναι όχι μόνο ισορροπημένα, αλλά και "καθαρά", δηλαδή απαλλαγμένα από θορυβώδη και κακώς ταξινομημένα δείγματα.

3.6 Παράμετροι Αξιολόγησης

Η αξιολόγηση των μοντέλων μηχανικής μάθησης αποτελεί κρίσιμο στάδιο για την κατανόηση της αποδοτικότητας τους και την επιλογή του κατάλληλου αλγορίθμου για το πρόβλημα που μελετάται. Στην παρούσα μελέτη χρησιμοποιήθηκαν διάφορες παράμετροι αξιολόγησης για την ποσοτικοποίηση της απόδοσης των μοντέλων, οι οποίες περιγράφονται παρακάτω.

3.6.1 Confusion Matrix (Πίνακας Σύγχυσης)

Ο πίνακας σύγχυσης παρέχει λεπτομερή εικόνα της απόδοσης του μοντέλου, απεικονίζοντας τις προβλέψεις του μοντέλου για κάθε κατηγορία σε σχέση με τις πραγματικές τιμές.

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Εικόνα 3-2: Πίνακας Σύγχυσης

Περιλαμβάνει τις εξής τιμές:

- **TP (True Positives)**: Σωστές προβλέψεις της κατηγορίας 0 (μη αθέτηση).
- **TN (True Negatives)**: Σωστές προβλέψεις της κατηγορίας 1 (αθέτηση).
- **FP (False Positives)**: Λανθασμένες προβλέψεις της κατηγορίας 0, ενώ το πραγματικό δείγμα είναι της κατηγορίας 1.
- **FN (False Negatives)**: Λανθασμένες προβλέψεις της κατηγορίας 1, ενώ το πραγματικό δείγμα είναι της κατηγορίας 0.

Ο πίνακας σύγχυσης βοηθά στην ερμηνεία της απόδοσης του μοντέλου και παρέχει πληροφορίες για το πού το μοντέλο κάνει λάθος.

3.6.2 Precision (Ευστοχία)

Η **Precision** μετρά το ποσοστό των σωστών θετικών προβλέψεων από όλες τις θετικές προβλέψεις που έκανε το μοντέλο. Είναι ιδιαίτερα χρήσιμη όταν το κόστος των **False Positives** είναι υψηλό.

Τύπος:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Possitives}$$

3.6.3 Recall (Ανάκληση)

Η Recall μετρά το ποσοστό των πραγματικών θετικών προβλέψεων που αναγνωρίζονται σωστά από το μοντέλο.

Τύπος:

$$Recall = \frac{True\ Positives}{True\ Possitives + False\ Negatives}$$

3.6.4 F1-Score

Η F1-Score είναι μια σημαντική παράμετρος αξιολόγησης σε προβλήματα με ανισορροπία κατηγοριών, καθώς συνδυάζει την Precision και την Recall, δίνοντας μια πιο ισορροπημένη εικόνα της απόδοσης του μοντέλου. Η F1-Score είναι ιδιαίτερα χρήσιμη όταν οι False Positives και οι False Negatives είναι εξίσου σημαντικοί.

Τύπος:

$$F1 = 2 * \frac{Precison * Recall}{Precision + Recall}$$

3.6.5 Accuracy (Ακρίβεια)

Η ακρίβεια (Accuracy) είναι μια από τις πιο κοινές παραμέτρους αξιολόγησης ταξινομητών. Υπολογίζει το ποσοστό των σωστών προβλέψεων σε σχέση με το σύνολο των προβλέψεων. Ωστόσο, η Accuracy στην περίπτωση μας δεν είναι επαρκής λόγο της με έντονης ανισορροπία των κατηγοριών.

- **Τύπος**:

$$Accuracy = \frac{True\ Positive + True\ Negative}{TP + FP + TN + FN}$$

3.6.6 ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

Το ROC (Receiver Operating Characteristic) και το AUC (Area Under the Curve) είναι δύο βασικά εργαλεία που χρησιμοποιούμε για να αξιολογήσουμε την απόδοση των μοντέλων μας, ιδίως όταν έχουμε να κάνουμε με ανισορροπία στις κατηγορίες του συνόλου δεδομένων. Το ROC μας δίνει μια γραφική αναπαράσταση της απόδοσης του μοντέλου σε όλα τα δυνατά thresholds, ενώ το AUC μετρά την πιθανότητα το μοντέλο να κατατάξει σωστά ένα τυχαία επιλεγμένο θετικό δείγμα υψηλότερα από ένα αρνητικό.

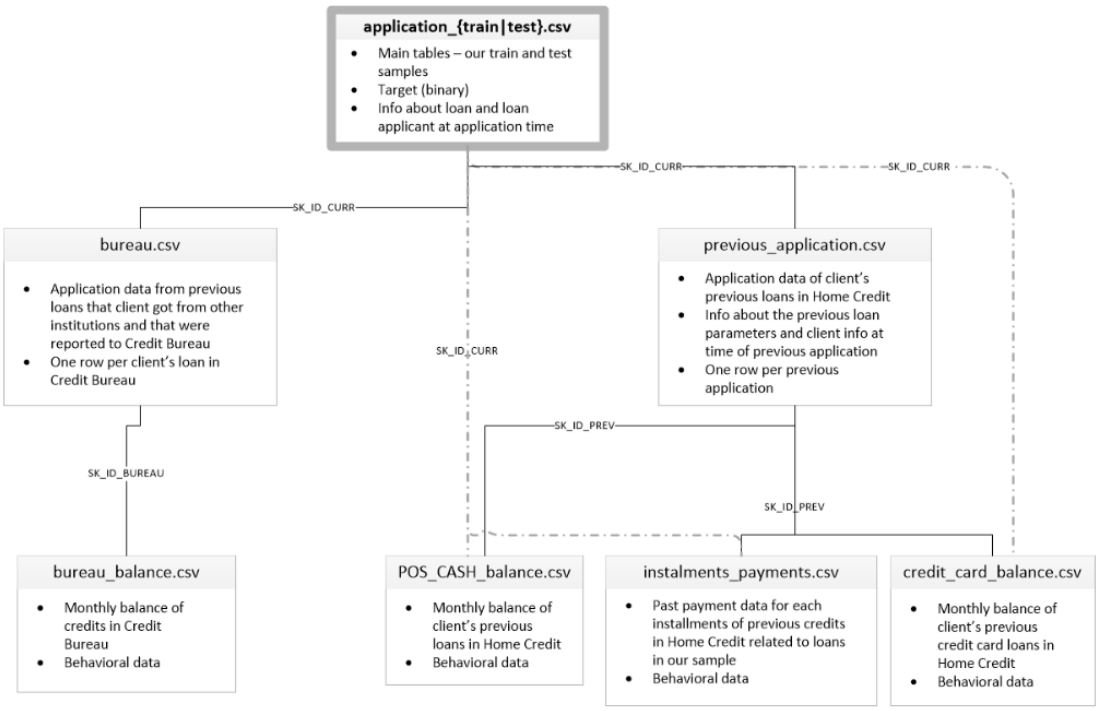
Ένα μοντέλο με AUC=1.0 είναι το ιδανικό, καθώς σημαίνει ότι για κάθε threshold καταφέρνει να εντοπίσει σωστά τα θετικά και τα αρνητικά δείγματα, ενώ ένα μοντέλο με AUC=0.5 είναι ισοδύναμο με μια τυχαία πρόβλεψη.

4 Δεδομένα

Για το σκοπό της μελέτης αυτής θα χρησιμοποιηθεί το σύνολο δεδομένων της χρηματοπιστωτικής πλατφόρμας Home Credit, όπως αυτά έχουν διατεθεί δημόσια στην κοινότητα του Kaggle στα πλαίσια ενός διαγωνισμού (Montoya et al., 2018). Η Kaggle αποτελεί την μεγαλύτερη διαδικτυακή κοινότητα επιστημόνων πληροφορίας και μηχανικής μάθησης. Παρέχει την δυνατότητα στους χρήστες της να βρουν μεγάλα σύνολα δεδομένων δωρεάν με σκοπό την εκπαίδευση τους και την εξέλιξη των ικανοτήτων τους. Επίσης επιτρέπει σε τρίτους, όπως στην περίπτωση της Home Credit, την δημιουργία διαγωνισμών με χρηματικό έπαθλο καλώντας τους χρήστες να δοκιμάσουν τις δυνατότητες τους και να βρουν λύση στο εκάστοτε πρόβλημα (kaggle.com).

Η Home Credit είναι μια διεθνής χρηματοπιστωτική εταιρεία που ειδικεύεται στην παροχή καταναλωτικών δανείων, κυρίως σε άτομα με περιορισμένη ή μηδενική πιστοληπτική ικανότητα. Με έμφαση στη χρήση προηγμένων αναλυτικών μεθόδων και τεχνολογίας, η Home Credit στοχεύει στη βελτίωση της χρηματοοικονομικής συμπερίληψης και την ενίσχυση της πιστοληπτικής ικανότητας των πελατών της.

Το σύνολο δεδομένων Home Credit Default Risk στο Kaggle αποτελεί μια προσπάθεια να ενισχυθεί η ικανότητα πρόβλεψης του κινδύνου αθέτησης δανείων, μέσω της χρήσης καινοτόμων αλγορίθμων μηχανικής μάθησης (homecredit.com).



Εικόνα 4-1: Σχεσιακό μοντέλο των συνόλων δεδομένων (Πηγή:https://kaggle.com/competitions/home-credit-default-risk)

Το σύνολο δεδομένων που παρέχει η Home Credit για την μελέτη, αποτελείται από οχτώ (8) αρχεία CSV:

1. application_train.csv
2. application_test.csv
3. bureau.csv
4. bureau_balance.csv
5. POS_CASH_balance.csv
6. credit_card_balance.csv

7. previous_application.csv
8. installments_payments.csv

4.1 Train/Test Dataset

Τα βασικά σύνολα δεδομένων για την εκπαίδευση και δοκιμή των μοντέλων τα οποία περιέχουν πληροφορίες για την κάθε αίτηση δανείου στη Home Credit. Το κάθε δάνειο προσδιορίζεται από το μοναδικό του χαρακτηριστικό SK_ID_CURR. Το application_train επίσης περιλαμβάνει την ετικέτα TARGET η οποία προσδιορίζει με μηδέν (0): το δάνειο έχει αποπληρωθεί και ένα (1): δεν έχει αποπληρωθεί.

```
train_data = pd.read_csv(r'Data\application_train.csv')
print('Number of rows : ', train_data.shape[0])
print('Number of features : ', train_data.shape[1])
train_data.head()
```

✓ 3.8s Python

```
Number of rows :  307511
Number of features :  122
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_C
0	100002	1	Cash loans	M	N	Y	
1	100003	0	Cash loans	F	N	N	
2	100004	0	Revolving loans	M	Y	Y	
3	100006	0	Cash loans	F	N	Y	
4	100007	0	Cash loans	M	N	Y	

5 rows × 122 columns

Εικόνα 4-2: Συνοπτική παρουσίαση του application_train.csv

```
test_data = pd.read_csv(r'Data\application_test.csv')
print('Number of rows : ', test_data.shape[0])
print('Number of features : ', test_data.shape[1])
test_data.head()
```

✓ 0.7s Python

```
Number of rows :  48744
Number of features :  121
```

	SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN
0	100001	Cash loans	F	N	Y	0
1	100005	Cash loans	M	N	Y	0
2	100013	Cash loans	M	Y	Y	0
3	100028	Cash loans	F	N	Y	2
4	100038	Cash loans	M	Y	N	1

5 rows × 121 columns

Εικόνα 4-3: Συνοπτική παρουσίαση του application_test.csv

Εδώ παρατηρούμε ότι έχουμε 307.511 σειρές δανείων και 122 στήλες χαρακτηριστικών στο application_train περιλαμβανομένης και της ετικέτας TARGET η οποία είναι και ο στόχος μας. Επίσης στο application_test έχουμε 48.744 σειρές με 121 στήλες χαρακτηριστικών. Τα χαρακτηριστικά και στα δυο σύνολα δεδομένων είναι τα ίδια, πλην της στήλης του στόχου.

4.2 Bureau Dataset

Αφορά προηγούμενες πιστώσεις από τρίτα χρηματοπιστωτικά ιδρύματα. Κάθε προηγούμενη πίστωση έχει τη δική της σειρά στο bureau αλλά κάθε δάνειο στο application μπορεί να έχει πολλαπλές προηγούμενες πιστώσεις.

```
bureau_data = pd.read_csv(r'Data\bureau.csv')
print('Number of rows : ', bureau_data.shape[0])
print('Number of features : ', bureau_data.shape[1])
bureau_data.head()
```

✓ 3.6s Python

```
Number of rows :  1716428
Number of features :  17
```

	SK_ID_CURR	SK_ID_BUREAU	CREDIT_ACTIVE	CREDIT_CURRENCY	DAYS_CREDIT	CREDIT_DAY_OVERDUE	DAY
0	215354	5714462	Closed	currency 1	-497	0	
1	215354	5714463	Active	currency 1	-208	0	
2	215354	5714464	Active	currency 1	-203	0	
3	215354	5714465	Active	currency 1	-203	0	
4	215354	5714466	Active	currency 1	-629	0	

Εικόνα 4-4: Συνοπτική παρουσίαση του burreau.csv

4.3 Bureau Balance

Περιέχει μηνιαίες πληροφορίες για της προηγούμενες πιστώσεις του bureau. Κάθε σειρά αφορά ένα μήνα προηγούμενης πίστωσης και κάθε πίστωση μπορεί να έχει πολλαπλές σειρές όσοι και οι μήνες που διαρκεί.

```
bureau_balance = pd.read_csv(r'Data\bureau_balance.csv')
print('Number of rows : ', bureau_balance.shape[0])
print('Number of features : ', bureau_balance.shape[1])
bureau_balance.head()
```

✓ 7.5s Python

```
Number of rows :  27299925
Number of features :  3
```

	SK_ID_BUREAU	MONTHS_BALANCE	STATUS
0	5715448	0	C
1	5715448	-1	C
2	5715448	-2	C
3	5715448	-3	C
4	5715448	-4	C

Εικόνα 4-5: Συνοπτική παρουσίαση του bureau_balance.csv

4.4 Previous Application

Το previous_application.csv έχει καταχωρημένες πληροφορίες για κάθε προηγούμενη αίτηση δανείου από πελάτες της Home Credit που έχουν ήδη δάνειο σε αυτή όπου κάθε σειρά αφορά μία προηγούμενη αίτηση που σχετίζεται με δάνεια του δείγματός μας.

```
previous_app = pd.read_csv(r'Data\previous_application.csv')
print('Number of rows : ', previous_app.shape[0])
print('Number of features : ', previous_app.shape[1])
previous_app.head()
```

✓ 9.2s Python

```
Number of rows :  1670214
Number of features :  37
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_I
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	

Εικόνα 4-6: Συνοπτική παρουσίαση του previous_application.csv

4.5 POS Cash Balance

Μηνιαίο ιστορικό προηγούμενων καταναλωτικών δανείων (POS sales, Cash Loans) στη Home Credit. Έχουμε μια σειρά του πίνακα για κάθε μήνα ιστορικού κάθε προηγούμενης πίστωσης που σχετίζεται με δάνεια του δείγματός μας.

```
cash_balance = pd.read_csv(r'Data\POS_CASH_balance.csv')
print('Number of rows : ', cash_balance.shape[0])
print('Number of features : ', cash_balance.shape[1])
cash_balance.head()
```

✓ 7.3s Python

```
Number of rows :  10001358
Number of features :  8
```

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	CNT_INSTALMENT	CNT_INSTALMENT_FUTURE	NAME_CONTRA
0	1803195	182943	-31	48.0	45.0	
1	1715348	367990	-33	36.0	35.0	
2	1784872	397406	-32	12.0	9.0	
3	1903291	269225	-35	48.0	42.0	
4	2341044	334279	-35	36.0	35.0	

Εικόνα 4-7: Συνοπτική παρουσίαση του POS_CASH_balance.csv

4.6 Installments Payments

Στο αρχείο installments_payments.csv παρέχονται πληροφορίες για Ιστορικό αποπληρωμής για τις πιστώσεις που έχουν εκταμιευθεί προηγουμένως από τη Home Credit, και σχετίζονται με τα δάνεια του δείγματός μας. Κάθε σειρά αντιστοιχεί σε μία πληρωμή μίας δόσης είτε τρέχοντος είτε προηγούμενου δανείου.

```
install_payments = pd.read_csv(r'Data\installments_payments.csv')
print('Number of rows : ', install_payments.shape[0])
print('Number of features : ', install_payments.shape[1])
install_payments.head()
```

✓ 11.9s Python

```
Number of rows :  13605401
Number of features :  8
```

	SK_ID_PREV	SK_ID_CURR	NUM_INSTALMENT_VERSION	NUM_INSTALMENT_NUMBER	DAYS_INSTAL
0	1054186	161674	1.0	6	-
1	1330831	151639	0.0	34	-
2	2085231	193053	2.0	1	
3	2452527	199697	1.0	3	-
4	2714724	167756	1.0	2	-

Εικόνα 4-8: Συνοπτική παρουσίαση του installments_payments.csv

4.7 Credit Card Balance

Μηνιαία στιγμιότυπα υπολοίπων, προηγούμενων πιστωτικών καρτών που έχει ο αιτών με την Home Credit. Αυτός ο πίνακας έχει μία σειρά για κάθε μήνα ιστορικού κάθε προηγούμενης πίστωσης που σχετίζεται με τα δάνεια του δείγματος. Για κάθε προηγούμενο μήνα δίνονται στοιχεία για τα ποσά που κινήθηκαν είτε αυτά αφορούν πληρωμές δόσεων είτε αγορές αγαθών.

```
credit_card_balance = pd.read_csv(r'Data\credit_card_balance.csv')
print('Number of rows : ', credit_card_balance.shape[0])
print('Number of features : ', credit_card_balance.shape[1])
credit_card_balance.head()
```

✓ 8.9s Python

```
Number of rows :  3840312
Number of features :  23
```

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	AMT_BALANCE	AMT_CREDIT_LIMIT_ACTUAL	AMT_I
0	2562384	378907	-6	56.970	135000	
1	2582071	363914	-1	63975.555	45000	
2	1740877	371185	-7	31815.225	450000	
3	1389973	337855	-4	236572.110	225000	
4	1891521	126868	-1	453919.455	450000	

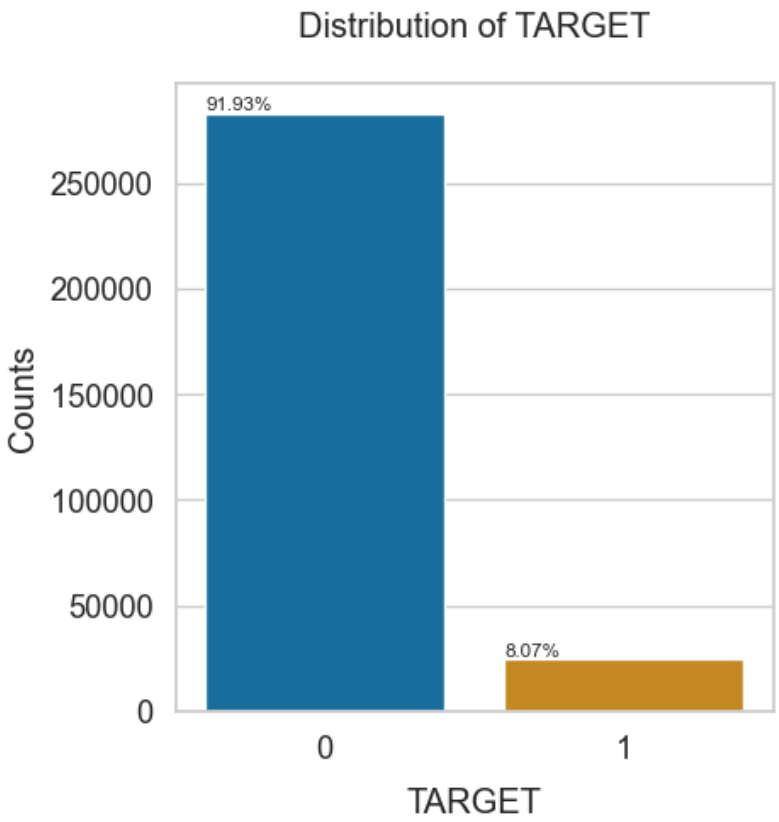
Εικόνα 4-9: Συνοπτική παρουσίαση του credit_card_balance.csv

5Διερευνητική Ανάλυση Δεδομένων – EDA

5.1 Application_train EDA

5.1.1 Ανάλυση Κατανομής Μεταβλητής Στόχου

Η μεταβλητή στόχος, η οποία δείχνει αν ένας πελάτης χρεοκόπησε σε ένα δάνειο (1) ή το αποπλήρωσε (0), αναλύεται ως προς την κατανομή της. Το διάγραμμα παρέχει μια σαφή εικόνα της ανισορροπίας ανάμεσα στις δύο κατηγορίες.



Εικόνα 5-1: Κατανομή της ετικέτας στόχου

Από την ανάλυση της στήλης TARGET παρατηρούμε ότι μόνο το 8.07% των εγγραφών αναφέρονται σε μη εξυπηρετούμενα δάνεια (1) και το υπόλοιπα 91.93% αφορά εξυπηρετούμενα (0). Αυτό σημαίνει ότι υπάρχει μεγάλη ανισορροπία μεταξύ των δύο τάξεων (Data Imbalance) το οποίο θα δούμε στη συνέχεια πως θα αντιμετωπίσουμε.

- **Μοναδικές Κατηγορίες:** Υπάρχουν δύο μοναδικές κατηγορίες στη μεταβλητή στόχου (0 και 1).

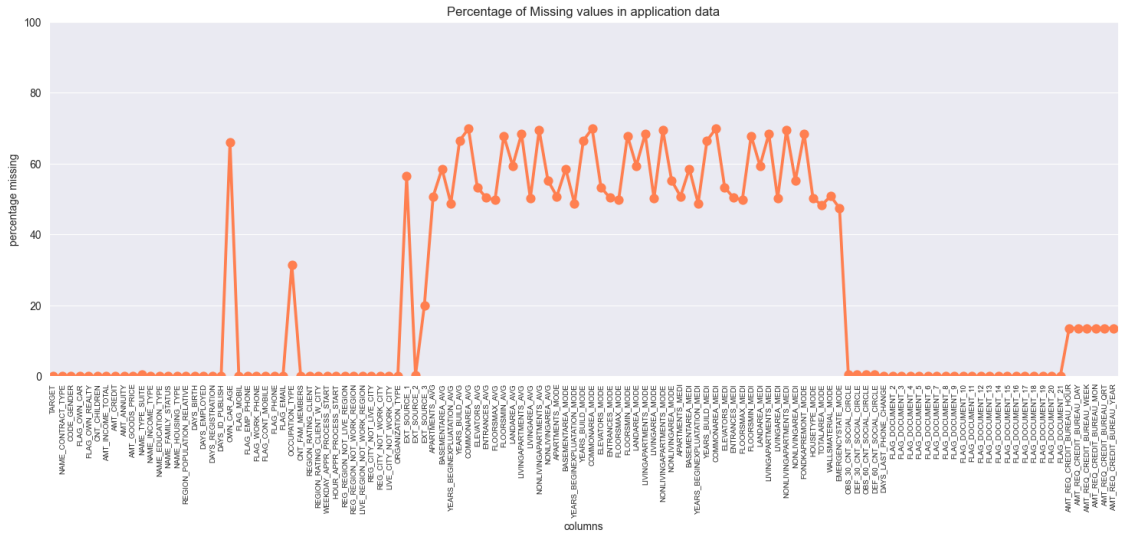
- **Ανισορροπία Κατηγοριών:** Από το διάγραμμα με ράβδους, παρατηρούμε σημαντική ανισορροπία κατηγοριών (Data Imbalance). :
  - Το 91.93% των περιπτώσεων ανήκει στην κατηγορία 0 (πελάτες που αποπλήρωσαν το δάνειο).
  - Το 8.07% των περιπτώσεων ανήκει στην κατηγορία 1 (πελάτες που χρεοκόπησαν).

Συμπεράσματα

Αυτή η ανισορροπία είναι συνηθισμένη σε πραγματικά οικονομικά δεδομένα, και η αντιμετώπισή της είναι κρίσιμη για τη δημιουργία αξιόπιστων μοντέλων μηχανικής μάθησης. Τεχνικές όπως η επαναδειγματοληψία (SMOTE, undersampling), ή η χρήση αλγορίθμων που έχουν σχεδιαστεί για ανισόρροπα σύνολα δεδομένων θα είναι απαραίτητες για την αποτελεσματική μοντελοποίηση και την αποφυγή του overfitting.

5.1.2 Ανάλυση Κενών Τιμών στα Σύνολα Δεδομένων

Σε αυτό το βήμα της Διερευνητικής Ανάλυσης Δεδομένων (EDA), εξετάζουμε το ποσοστό των ελλείψεων (missing values) στα χαρακτηριστικά του συνόλου δεδομένων. Η κατανόηση του πόσο συχνές είναι οι απουσίες στα δεδομένα είναι κρίσιμη για τη λήψη αποφάσεων σχετικά με τον τρόπο χειρισμού τους.

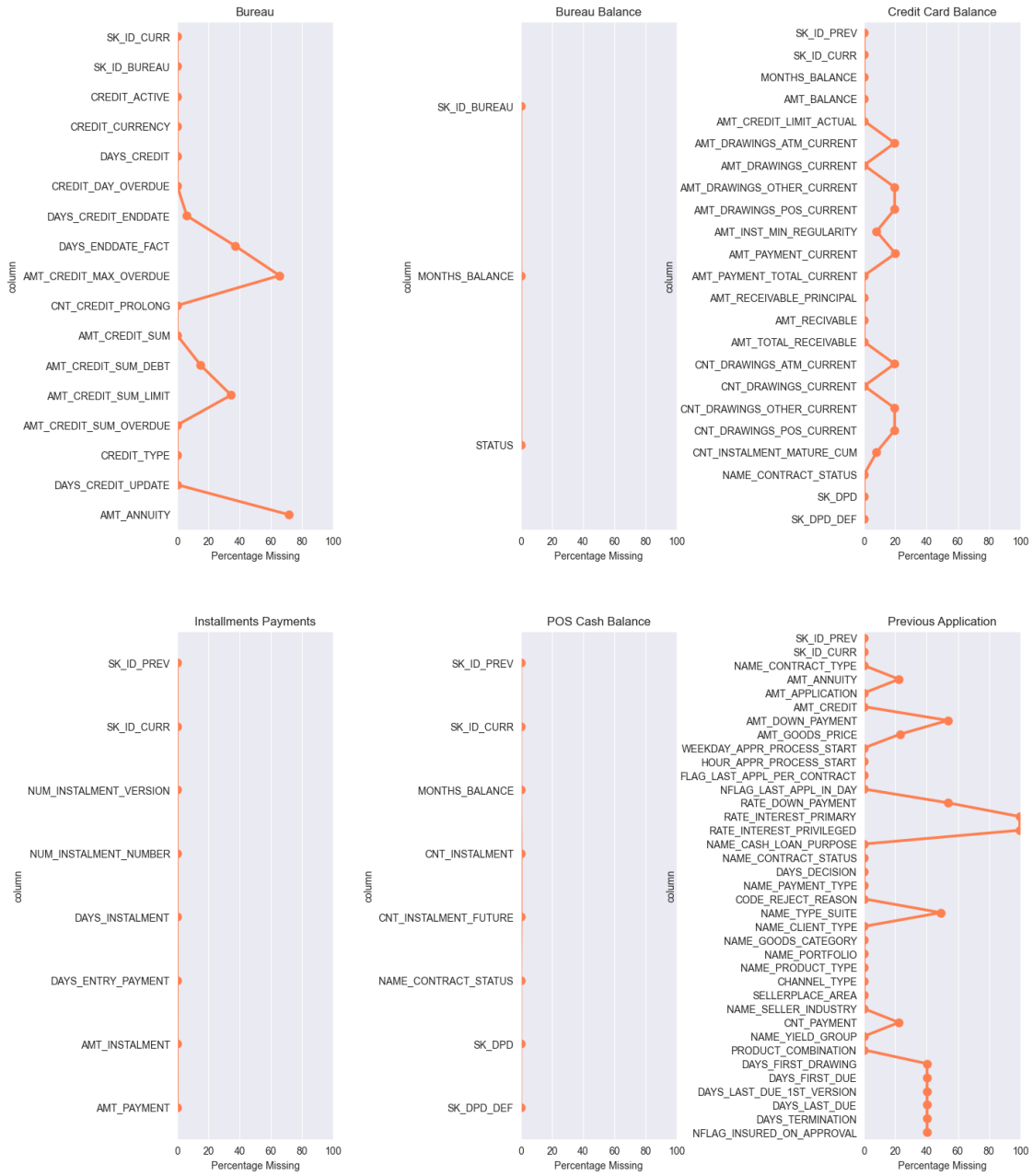


Εικόνα 5-2: Κατανομή κενών τιμών στο application_train.csv

- **Ποσοστό Κενών Τιμών:** Τα παρακάτω διαγράμματα δείχνουν το ποσοστό των ελλείψεων για κάθε στήλη των συνόλων δεδομένων. Τα χαρακτηριστικά με υψηλά ποσοστά απουσιών μπορεί να παρουσιάζουν προβλήματα αξιοπιστίας και ενδέχεται να χρειάζονται ιδιαίτερη προσοχή στη διαχείρισή τους.

- **Αντιμετώπιση των Απουσιών:** Αφού εντοπιστούν οι στήλες με ελλείψεις, μπορούμε να εξετάσουμε διάφορες μεθόδους αντιμετώπισης, όπως η εισαγωγή τιμών (imputation) με τον μέσο όρο ,τη συχνότερη τιμή ή ακόμα και την αφαίρεση ορισμένων στηλών στις οποίες οι απουσίες είναι εκτεταμένες. Η απόφαση για το πώς θα διαχειριστούμε αυτές τις απουσίες εξαρτάται από το πόσο σημαντικά θεωρούνται τα δεδομένα αυτών των χαρακτηριστικών.

Η σωστή αντιμετώπιση της απουσίας δεδομένων θα μας επιτρέψει να διασφαλίσουμε την ακρίβεια και την ποιότητα του συνόλου δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση των μοντέλων.



Εικόνα 5-3: Κατανομή κενών τιμών στα υπόλοιπα σύνολα δεδομένων

5.1.3Ανάλυση για τη Μεταβλητή NAME_CONTRACT_TYPE

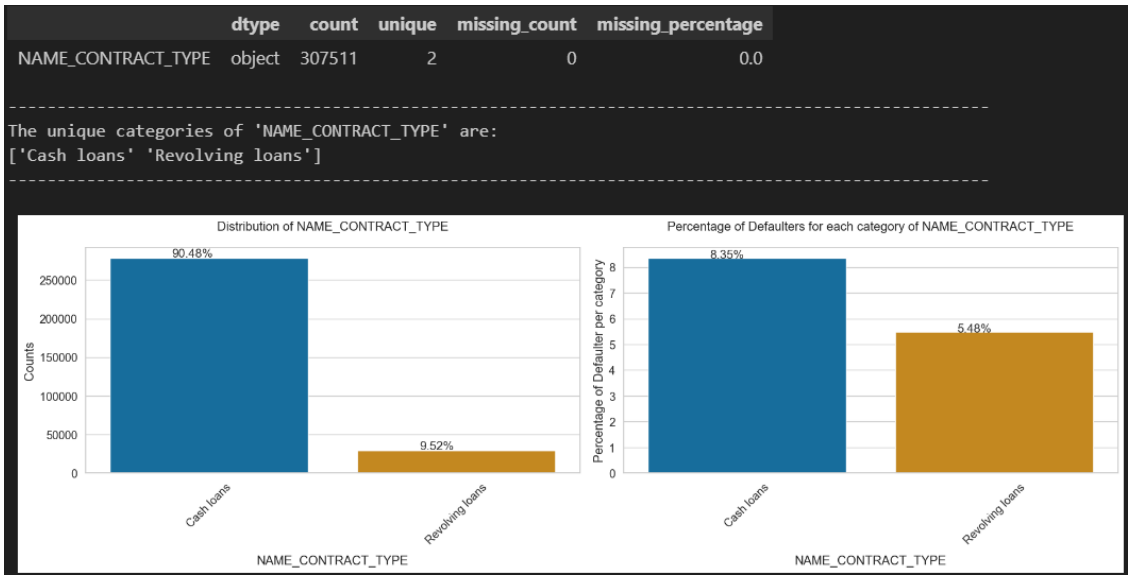
Η μεταβλητή NAME_CONTRACT_TYPE αναφέρεται στον τύπο του δανείου που έχει ο πελάτης.

Από την ανάλυση προκύπτουν τα εξής:

1. **Τύπος Δεδομένων και Πλήθος Τιμών**
   - **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
   - **Συνολικός αριθμός μη κενών τιμών:** Δεν υπάρχουν ελλείψεις (missing values) σε αυτή τη στήλη.
   - **Μοναδικές τιμές:** Υπάρχουν 2 μοναδικοί τύποι συμβάσεων δανείου "Cash loans" και "Revolving loans".

2. **Κατανομή των Δανείων:**
   - Το 90.48% των πελατών έχει συνάψει συμβάσεις τύπου "Cash loans".
   - Το 9.52% των πελατών έχει συνάψει συμβάσεις τύπου "Revolving loans".

3. **Πιθανότητες Αθέτησης ανά Τύπο Σύμβασης Δανείου:**
   - Οι πελάτες με συμβάσεις "Cash loans" έχουν ποσοστό αθέτησης 8.35%.
   - Οι πελάτες με συμβάσεις "Revolving loans" έχουν ποσοστό αθέτησης 5.48%.



Εικόνα 5-4: Κατανομή της NAME_CONTRACT_TYPE

Συμπεράσματα

Από το χαρακτηριστικό, τύποι δανείων (NAME_CONTRACT_TYPE) παρατηρούμε ότι λαμβάνονται περισσότερα δάνεια σε μετρητά απ' ότι κυλιόμενα δάνεια και ότι είναι τα περισσότερα που δεν αποπληρώνονται. Η διαφορά στα ποσοστά αθέτησης δείχνει ότι ο τύπος δανείου μπορεί να επηρεάσει το ρίσκο.

5.1.4 Ανάλυση για τη Μεταβλητή CODE_GENDER

Η μεταβλητή CODE_GENDER αναφέρεται στο φύλο των πελατών.

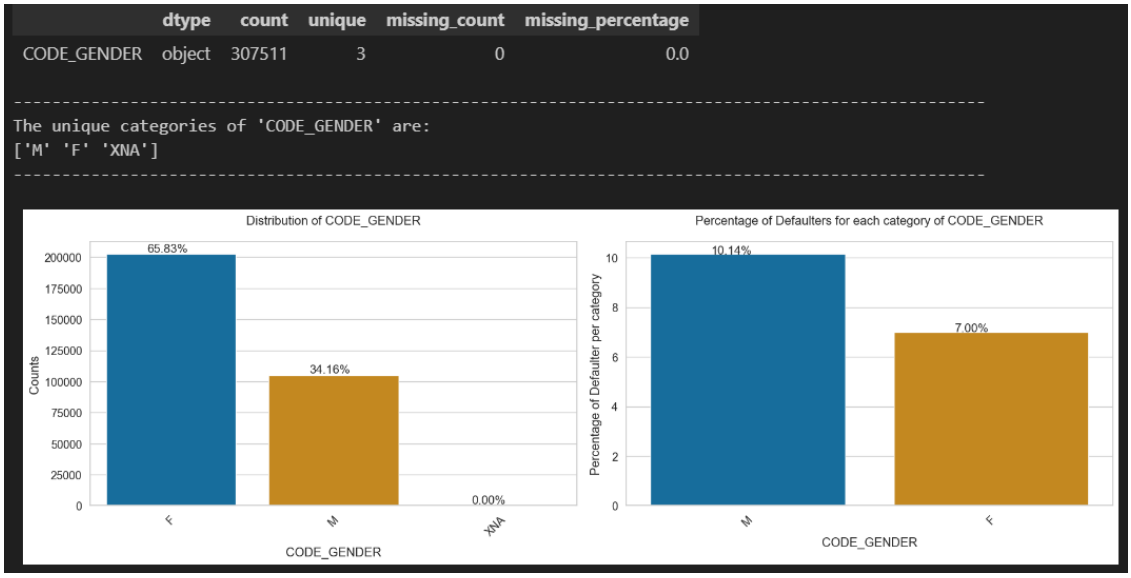
Από την ανάλυση προκύπτουν τα εξής:

1. **Τύπος Δεδομένων και Πλήθος Τιμών**
   - **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
   - **Αριθμός κενών τιμών:** Δεν υπάρχουν.
   - **Μοναδικές τιμές:** Υπάρχουν δύο κύριες κατηγορίες, "F" (γυναίκα) και "M" (άνδρας), καθώς και οι 4 περιπτώσεις "XNA".

2. **Κατανομή Φύλου:**
   - Το 65.83% των πελατών είναι γυναίκες.
   - Το 34.16% των πελατών είναι άνδρες.

- Υπάρχουν 4 τιμές "XNA".

3. **Πιθανότητες Αθέτησης ανά Φύλο:**
   - Οι άνδρες έχουν ποσοστό αθέτησης 10.14%.
   - Οι γυναίκες έχουν ποσοστό αθέτησης 7.00%.



Εικόνα 5-5: Κατανομή της CODE_GENDER

Συμπεράσματα

Από την κατανομή του φύλου παρατηρούμε ότι οι γυναίκες είναι η πλειοψηφία των δανειοληπτών κάτι τι οποίο το μοντέλα πρέπει να λαμβάνουν υπόψη την για να αποφύγουν την υποεκπροσώπηση των ανδρών στις προβλέψεις. Επίσης Από τη διαφορά στα ποσοστά αθέτησης προκύπτει ότι οι άνδρες παρουσιάζουν μεγαλύτερο ρίσκο αθέτησης του δανείου σε σχέση με τις γυναίκες. Έτσι, το φύλο μπορεί να αποδειχθεί ένας ισχυρός παράγοντας πρόβλεψης της πιθανότητας χρεοκοπίας. Τέλος, θα πρέπει να χειριστούμε κάπως τις τιμές "XNA" κατά την προεπεξεργασία των δεδομένων το οποίο πιθανότατα λόγο το μικρού πλήθους τους δεν θα επηρεάσει τα αποτελέσματα .

5.1.5 Ανάλυση για τη Μεταβλητή FLAG_OWN_CAR

Η μεταβλητή FLAG_OWN_CAR δείχνει αν ο πελάτης είναι κάτοχος αυτοκινήτου.

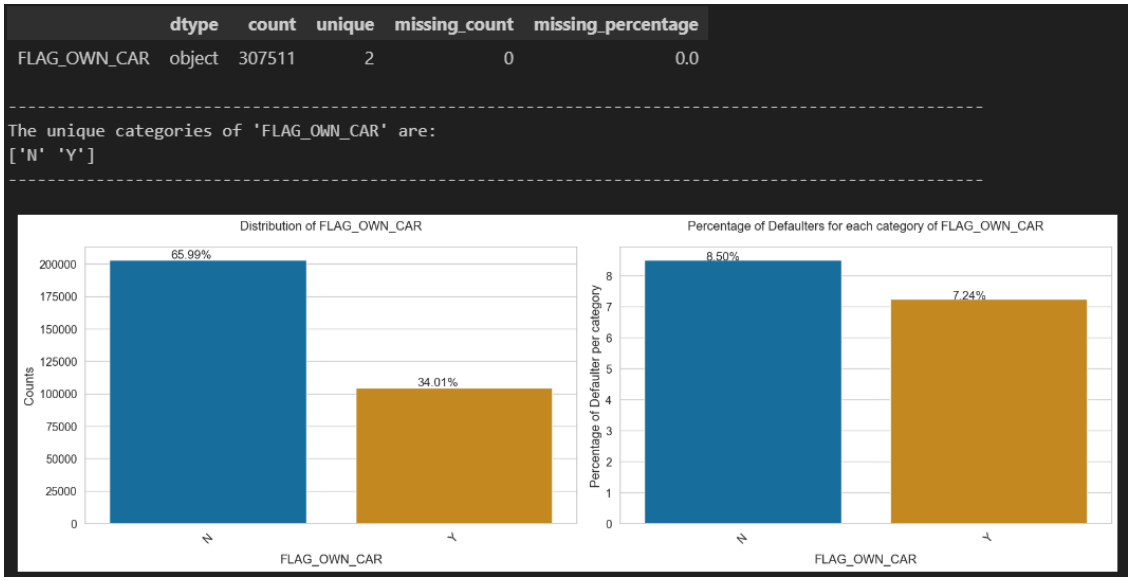
Από την ανάλυση προκύπτουν τα εξής:

1. **Τύπος Δεδομένων και Πλήθος Τιμών**
   - **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
   - **Αριθμός κενών τιμών:** Δεν υπάρχουν.
   - **Μοναδικές τιμές:** Οι δύο κατηγορίες είναι "Yes" (έχει αυτοκίνητο) και "No" (δεν έχει αυτοκίνητο).

2. **Κατανομή Κατόχων Αυτοκινήτου:**
   - Το 65.99% των πελατών δεν έχει αυτοκίνητο.
   - Το 34.01% των πελατών έχει αυτοκίνητο.

3. Πιθανότητες Αθέτησης ανά Κατηγορία:

- Οι πελάτες χωρίς αυτοκίνητο παρουσιάζουν ποσοστό αθέτησης 8.50%.
- Οι πελάτες που έχουν αυτοκίνητο παρουσιάζουν ποσοστό αθέτησης 7.24%.



Εικόνα 5-6: Κατανομή της FLAG_OWN_CAR

Συμπεράσματα

Από το χαρακτηριστικό για το αν ο πελάτης είναι κάτοχος αυτοκινήτου παρατηρούμε ότι η πλειοψηφία των πελατών δεν διαθέτει αυτοκίνητο, όπως άλλωστε και η πλειοψηφία των μη εξυπηρετούμενων. Αυτό μπορεί να αποτελεί σημαντική πληροφορία για την πρόβλεψη της οικονομικής τους συμπεριφοράς, καθώς η ιδιοκτησία αυτοκινήτου μπορεί να υποδηλώνει διαφορετικές οικονομικές συνήθειες και δυναμική.

5.1.6 Ανάλυση για τη Μεταβλητή FLAG_OWN_REALTY

Η μεταβλητή FLAG_OWN_REALTY δείχνει αν ο πελάτης έχει ακίνητη περιουσία.

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα
- **Αριθμός κενών τιμών:** Δεν υπάρχουν σε αυτή τη στήλη.
- **Μοναδικές τιμές:** Οι δύο κατηγορίες είναι "Yes" (έχει ακίνητη περιουσία) και "No" (δεν έχει ακίνητη περιουσία).

2. Κατανομή των Τιμών:

- Το 69.37% των πελατών έχει ακίνητη περιουσία.
- Το 30.63% των πελατών δεν έχει ακίνητη περιουσία.

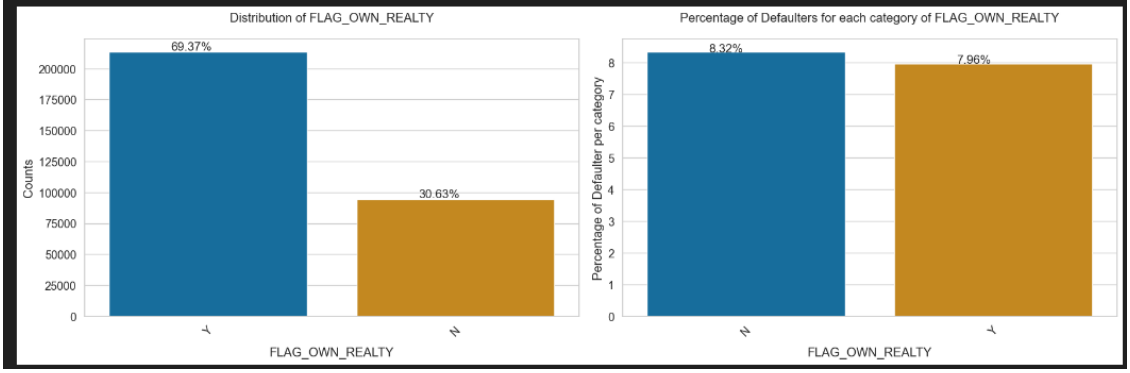
3. Πιθανότητα Αθέτησης ανά Κατηγορία

- Οι πελάτες χωρίς ακίνητη περιουσία παρουσιάζουν ποσοστό χρεοκοπίας 8.32%.

- Οι πελάτες που έχουν ακίνητη περιουσία παρουσιάζουν ποσοστό χρεοκοπίας 7.96%.

	dtype	count	unique	missing_count	missing_percentage
FLAG_OWN_REALTY	object	307511	2	0	0.0

```
----------------------------------------------------------------------------------------------------
The unique categories of 'FLAG_OWN_REALTY' are:
['Y' 'N']
----------------------------------------------------------------------------------------------------
```



Εικόνα 5-7: Κατανομή της FLAG_OWN_REALTY

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα
- **Αριθμός κενών τιμών:** Δεν υπάρχουν σε αυτή τη στήλη.
- **Μοναδικές τιμές:** Οι δύο κατηγορίες είναι "Yes" (έχει ακίνητη περιουσία) και "No" (δεν έχει ακίνητη περιουσία).

2. Κατανομή των Τιμών:

- Το 69.37% των πελατών έχει ακίνητη περιουσία.
- Το 30.63% των πελατών δεν έχει ακίνητη περιουσία.

3. Πιθανότητα Αθέτησης ανά Κατηγορία

- Οι πελάτες χωρίς ακίνητη περιουσία παρουσιάζουν ποσοστό χρεοκοπίας 8.32%.
- Οι πελάτες που έχουν ακίνητη περιουσία παρουσιάζουν ποσοστό χρεοκοπίας 7.96%.

Συμπεράσματα

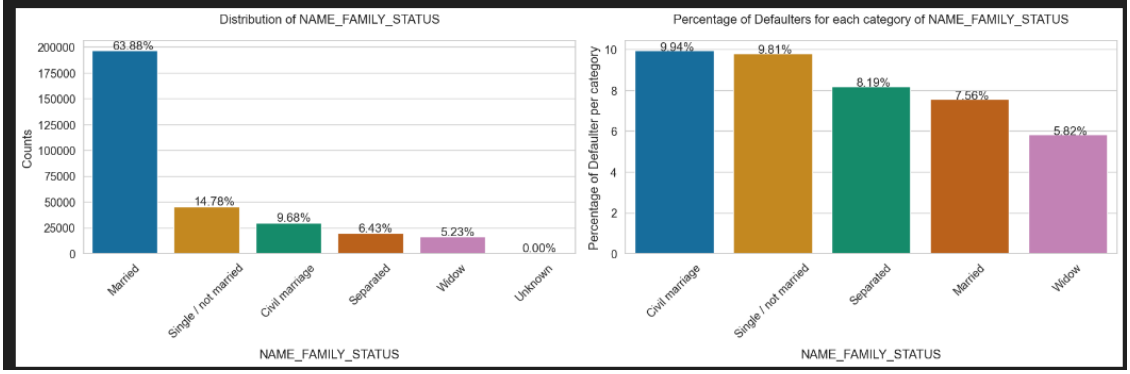
Η πλειοψηφία των πελατών διαθέτει ακίνητο. Αυτό μπορεί να αποτελεί σημαντικό παράγοντα για την εκπαίδευση μοντέλων, καθώς η ιδιοκτησία ακινήτου μπορεί να σχετίζεται με διαφορετικές οικονομικές δυνατότητες. Η μικρή διαφορά στα ποσοστά χρεοκοπίας υποδηλώνει ότι η ιδιοκτησία ακινήτου δεν επηρεάζει δραστικά την πιθανότητα χρεοκοπίας, αν και αυτοί που δεν έχουν ακίνητη περιουσία φαίνεται να έχουν ελαφρώς μεγαλύτερη πιθανότητα αθέτησης πληρωμής. Αυτό μπορεί να οφείλεται στο γεγονός ότι η ακίνητη περιουσία λειτουργεί ως ένδειξη μεγαλύτερης οικονομικής σταθερότητας.

5.1.7 Ανάλυση για τη Μεταβλητή NAME_FAMILY_STATUS

Η μεταβλητή NAME_FAMILY_STATUS αναφέρεται στην οικογενειακή κατάσταση του πελάτη.

	dtype	count	unique	missing_count	missing_percentage
NAME_FAMILY_STATUS	object	307511	6	0	0.0

```
The unique categories of 'NAME_FAMILY_STATUS' are:
['Single / not married' 'Married' 'Civil marriage' 'Widow' 'Separated'
 'Unknown']
```



Εικόνα 5-8: Κατανομή της NAME_FAMILY_STATUS

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
- **Αριθμός κενών τιμών:** Δεν υπάρχουν.
- **Μοναδικές τιμές:** Υπάρχουν 5 κύριες κατηγορίες ("Married", "Single / Not married", "Civil marriage", "Widow", "Separated") και μία κατηγορία "Unknown".

2. Κατανομή Οικογενειακής Κατάστασης:

- Το 63.88% των πελατών είναι παντρεμένοι.
- Το 14.78% είναι ανύπαντροι.
- Το 9.68% βρίσκεται σε πολιτικό γάμο.
- Το 6.43% είναι χήροι.
- Το 5.23% είναι διαζευγμένοι.
- "Unknown" δεν περιέχει παρατηρήσεις.

3. Πιθανότητα Αθέτησης ανά Οικογενειακή Κατάσταση:

- Οι παντρεμένοι με πολιτικό γάμο έχουν ποσοστό χρεοκοπίας 9.94%.
- Οι ανύπαντροι πελάτες έχουν ποσοστό χρεοκοπίας 9.81%.
- Οι διαζευγμένοι πελάτες έχουν ποσοστό χρεοκοπίας 8.19%.
- Οι παντρεμένοι πελάτες έχουν ποσοστό χρεοκοπίας 7.56%.
- Οι χήροι έχουν ποσοστό χρεοκοπίας 5.82%.

Συμπεράσματα

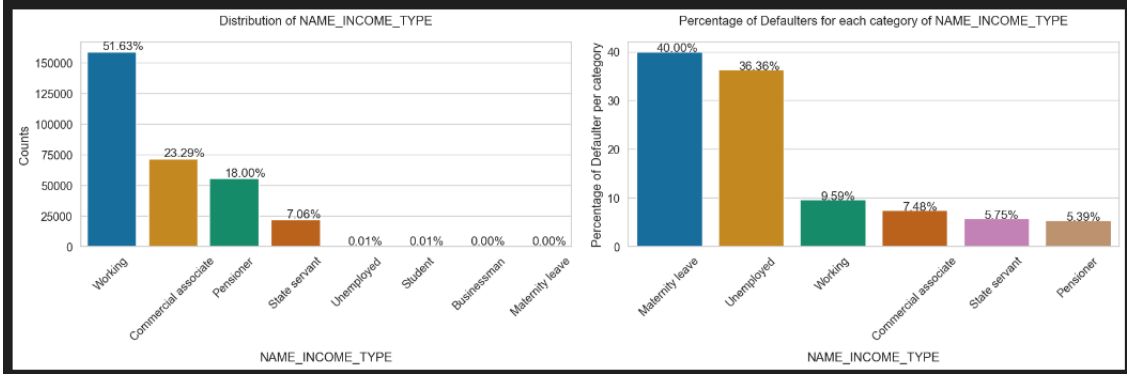
Η οικογενειακή κατάσταση συχνά επηρεάζει τις οικονομικές συνήθειες, κάτι που πρέπει να ληφθεί υπόψη κατά την εκπαίδευση των μοντέλων. Οι παντρεμένοι πελάτες παρουσιάζουν τη χαμηλότερη πιθανότητα αθέτησης δανείου, γεγονός που πιθανότατα σχετίζεται με μεγαλύτερη οικονομική σταθερότητα και υποχρεώσεις. Αυτό είναι σύνηθες σε οικονομικές μελέτες που συνδέουν τον γάμο με καλύτερη οικονομική διαχείριση. Οι ανύπαντροι και οι πελάτες σε πολιτικό γάμο έχουν αυξημένο ρίσκο αθέτησης σε σύγκριση με άλλες κατηγορίες, κάτι που συνδέεται με μικρότερη οικονομική σταθερότητα ή διαφορετικές οικονομικές προτεραιότητες. Οι χήροι εμφανίζουν επίσης χαμηλή πιθανότητα αθέτησης.

5.1.8 Ανάλυση για τη Μεταβλητή NAME_INCOME_TYPE

Η μεταβλητή NAME_INCOME_TYPE αναφέρεται στο είδος του εισοδήματος του πελάτη.

	dtype	count	unique	missing_count	missing_percentage
NAME_INCOME_TYPE	object	307511	8	0	0.0

```
The unique categories of 'NAME_INCOME_TYPE' are:
['Working' 'State servant' 'Commercial associate' 'Pensioner' 'Unemployed'
 'Student' 'Businessman' 'Maternity leave']
```



Εικόνα 5-9: Κατανομή της NAME_INCOME_TYPE

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
- **Αριθμός κενών τιμών:** Δεν υπάρχουν απουσίες σε αυτή τη στήλη.
- **Μοναδικές τιμές:** Υπάρχουν 8 κατηγορίες ("Working", "Commercial associate", "Pensioner", "State servant", "Unemployed", "Student", "Businessman", "Maternity leave").

2. Κατανομή Τύπου Εισοδήματος:

- Το 51.63% των πελατών είναι εργαζόμενοι.
- Το 23.29% είναι εμπορικοί συνεργάτες.
- Το 18.00% είναι συνταξιούχοι.
- Το 7.06% είναι δημόσιοι υπάλληλοι.
- Πολύ μικρά ποσοστά παρατηρούνται σε άνεργους (0.01%, 22 περιπτώσεις), φοιτητές (0.01%, 18 περιπτώσεις), επιχειρηματίες (0.00%, 10 περιπτώσεις), και όσους βρίσκονται σε άδεια μητρότητας (0.00%, 5 περιπτώσεις).

3. Πιθανότητα Αθέτησης ανά Τύπο Εισοδήματος:

- Οι πελάτες σε άδεια μητρότητας έχουν ποσοστό χρεοκοπίας 40.00% (5 περιπτώσεις).
- Οι άνεργοι έχουν ποσοστό χρεοκοπίας 36.36% (22 περιπτώσεις).
- Οι εργαζόμενοι έχουν ποσοστό χρεοκοπίας 9.59%.
- Οι εμπορικοί συνεργάτες έχουν ποσοστό χρεοκοπίας 7.48%.
- Οι δημόσιοι υπάλληλοι έχουν ποσοστό χρεοκοπίας 5.75%.
- Οι συνταξιούχοι έχουν ποσοστό χρεοκοπίας 5.39%.

Συμπεράσματα

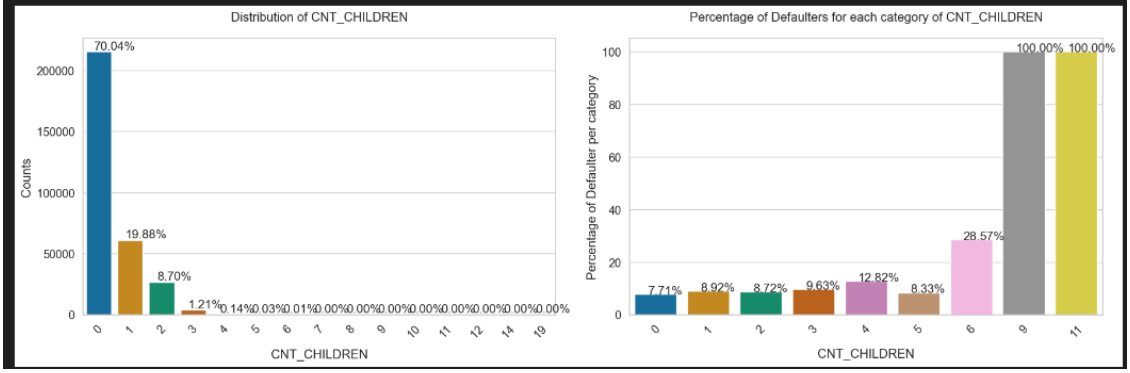
Η πλειοψηφία των πελατών είναι εργαζόμενοι, κάτι που αναμενόμενα αντικατοπτρίζει την εργασιακή κατανομή της κοινωνίας. Οι κατηγορίες όπως οι συνταξιούχοι και οι δημόσιοι υπάλληλοι είναι επίσης σημαντικές και ενδέχεται να σχετίζονται με συγκεκριμένες οικονομικές συμπεριφορές. Οι πολύ υψηλές πιθανότητες χρεοκοπίας για τους πελάτες σε άδεια μητρότητας και τους άνεργους δείχνουν ότι αυτοί οι πελάτες αντιμετωπίζουν μεγαλύτερο οικονομικό ρίσκο. Ωστόσο, αυτά τα ποσοστά πρέπει να ερμηνευτούν με προσοχή, καθώς ο πολύ μικρός αριθμός περιπτώσεων σε αυτές τις κατηγορίες (5 και 22 περιπτώσεις, αντίστοιχα) μπορεί να επηρεάσει δυσανάλογα τα αποτελέσματα και να μην είναι αντιπροσωπευτικός της πραγματικής τους τάσης.

5.1.9 Ανάλυση για τη Μεταβλητή CNT_CHILDREN

Η μεταβλητή CNT_CHILDREN αναφέρεται στον αριθμό των παιδιών που έχουν οι πελάτες.

	dtype	count	unique	missing_count	missing_percentage
CNT_CHILDREN	int64	307511	15	0	0.0

```
The unique categories of 'CNT_CHILDREN' are:
[ 0  1  2  3  4  7  5  6  8  9 11 12 10 19 14]
```



Εικόνα 5-10: Κατανομή της CNT_CHILDREN

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** int, δηλαδή αριθμητικά δεδομένα
- **Αριθμός κενών τιμών:** Δεν υπάρχουν.
- **Μοναδικές τιμές:** Ο αριθμός των παιδιών κυμαίνεται από 0 έως 19.

2. Κατανομή Αριθμού Παιδιών:

- Το 70.04% των πελατών δεν έχει παιδιά .

- Το 19.88% έχει 1 παιδί .
- Το 8.7% έχει 2 παιδιά .
- Μικρά ποσοστά παρατηρούνται για 4 παιδιά (0.14%, 429 περιπτώσεις), 5 παιδιά (0.03%, 84 περιπτώσεις), και 6 παιδιά (0.01%, 21 περιπτώσεις). Πολύ λίγοι πελάτες έχουν 7 ή περισσότερα παιδιά.

3. Πιθανότητα Αθέτησης ανά Αριθμό Παιδιών:

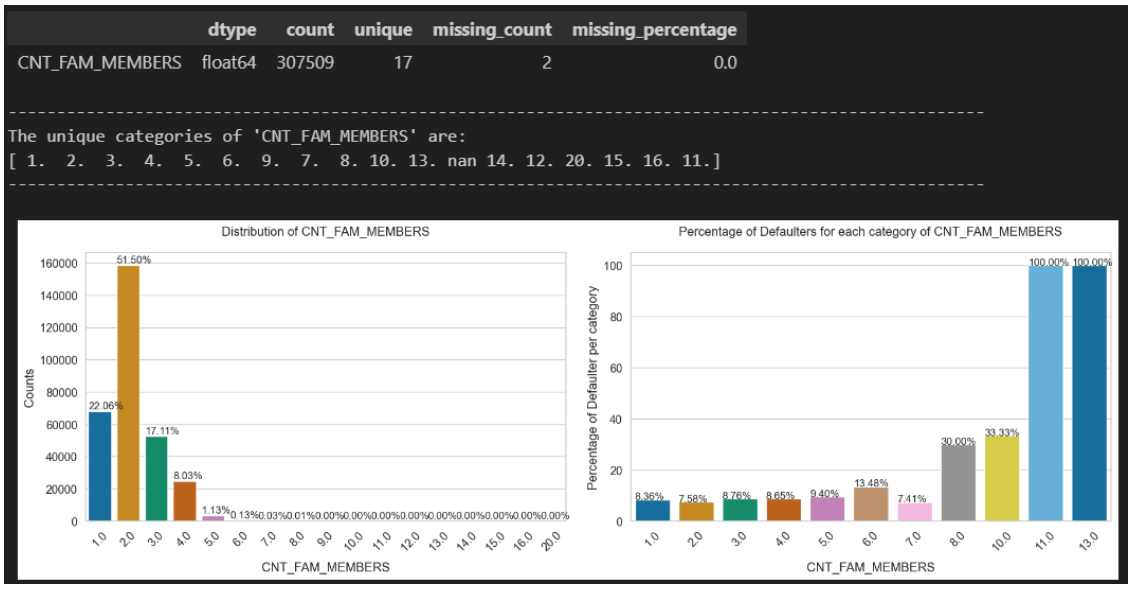
- Οι πελάτες χωρίς παιδιά έχουν ποσοστό χρεοκοπίας 7.71%.
- Οι πελάτες με 1 παιδί έχουν ποσοστό χρεοκοπίας 8.89%.
- Οι πελάτες με 2 παιδιά έχουν ποσοστό χρεοκοπίας 8.72%.
- Οι πελάτες με 3 παιδιά έχουν ποσοστό χρεοκοπίας 9.63%.
- Οι πελάτες με 4 παιδιά έχουν ποσοστό χρεοκοπίας 12.82%.
- Οι πελάτες με 5 παιδιά έχουν ποσοστό χρεοκοπίας 8.33%.
- Οι πελάτες με 6 παιδιά έχουν ποσοστό χρεοκοπίας 28.57%.
- Οι πελάτες με 9 και 11 παιδιά έχουν 100% ποσοστό χρεοκοπίας (πολύ μικρός αριθμός περιπτώσεων).

Συμπεράσματα

Η πλειοψηφία των πελατών δεν έχει παιδιά ή έχει το πολύ 1-2 παιδιά, κάτι που αντικατοπτρίζει την κατανομή των οικογενειών στο δείγμα. Παρατηρούμε μια τάση όπου το ποσοστό χρεοκοπίας αυξάνεται με τον αριθμό των παιδιών, με τα υψηλότερα ποσοστά να παρατηρούνται σε πελάτες με 4 ή περισσότερα παιδιά. Ωστόσο, το γεγονός ότι οι περιπτώσεις με πολύ μεγάλα ποσοστά (6 παιδιά και πάνω) αφορούν πολύ μικρό αριθμό δειγμάτων (π.χ. 2 ή 1 περιπτώσεις) θα πρέπει να ληφθεί υπόψη, καθώς μπορεί να επηρεάζει τα αποτελέσματα.

5.1.10 Ανάλυση για τη Μεταβλητή CNT_FAM_MEMBERS

Η μεταβλητή CNT_FAM_MEMBERS αναφέρεται στον συνολικό αριθμό μελών της οικογένειας του πελάτη.



Εικόνα 5-11: Κατανομή της CNT_FAMILY_MEMBERS

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** float, δηλαδή αριθμητικά δεδομένα
- **Αριθμός κενών τιμών:** Δεν υπάρχουν.
- **Μοναδικές τιμές:** Ο αριθμός των μελών κυμαίνεται από 1 έως 20.

2. Κατανομή Αριθμού Μελών Οικογένειας:

- Το 22.06% των πελατών έχει 1 μέλος στην οικογένεια.
- Το 51.50% έχει 2 μέλη.
- Το 17.11% έχει 3 μέλη.
- Το 8.03% έχει 4 μέλη .
- Μικρά ποσοστά παρατηρούνται για οικογένειες με 5 μέλη (1.13%, 3,478 περιπτώσεις), 6 μέλη (0.13%, 408 περιπτώσεις), και ακόμα μικρότερα ποσοστά για περισσότερα από 6 μέλη.

3.Πιθανότητα Αθέτησης ανά Αριθμό Μελών Οικογένειας:

- Οι πελάτες με 1 μέλος έχουν ποσοστό χρεοκοπίας 8.36%.
- Οι πελάτες με 2 μέλη έχουν ποσοστό χρεοκοπίας 7.58%.
- Οι πελάτες με 3 μέλη έχουν ποσοστό χρεοκοπίας 8.36%.
- Οι πελάτες με 4 μέλη έχουν ποσοστό χρεοκοπίας 8.65%.
- Οι πελάτες με 5 μέλη έχουν ποσοστό χρεοκοπίας 9.4%.
- Οι πελάτες με 6 μέλη έχουν ποσοστό χρεοκοπίας 13.48%.
- Οι πελάτες με 7 μέλη έχουν ποσοστό χρεοκοπίας 7.41%.
- Τα ποσοστά χρεοκοπίας για οικογένειες με 8 μέλη και πάνω είναι σημαντικά αυξημένα, με το ποσοστό να φτάνει το 100% για 11 και 13 μέλη, αν και αυτές οι περιπτώσεις είναι εξαιρετικά σπάνιες.

Συμπεράσματα

Η κατανομή δείχνει ότι η πλειοψηφία των οικογενειών αποτελείται από 2 μέλη, κάτι που πιθανώς περιλαμβάνει τον πελάτη και έναν ακόμα ενήλικα. Η τάση δείχνει ότι όσο μεγαλώνει το μέγεθος της οικογένειας, η πιθανότητα αθέτησης αυξάνεται, παρόμοια με τη μεταβλητή CNT_CHILDREN. Αυτό πιθανώς συνδέεται με το αυξημένο οικονομικό βάρος που φέρει η συντήρηση μιας μεγάλης οικογένειας. Ωστόσο, οι πολύ υψηλές πιθανότητες χρεοκοπίας για οικογένειες με 8 ή περισσότερα μέλη πρέπει να ερμηνευτούν με προσοχή, λόγω του πολύ μικρού αριθμού περιπτώσεων.

5.1.11 Ανάλυση για τη Μεταβλητή OCCUPATION_TYPE

Η μεταβλητή OCCUPATION_TYPE αναφέρεται στο επάγγελμα του πελάτη.

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
- **Αριθμός κενών τιμών:** Δεν υπάρχουν.
- **Μοναδικές τιμές:** Υπάρχουν διάφορα επαγγέλματα, όπως "Laborers", "Sales staff", "Core staff", "Managers", κλπ.

2. Κατανομή Επαγγελμάτων:

- Το 26.14% των πελατών είναι εργάτες (55,186 περιπτώσεις).

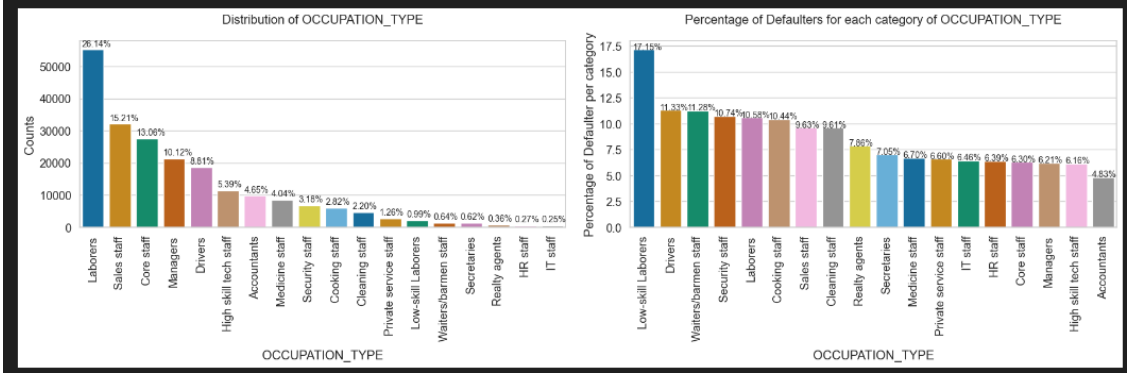
- Το 15.21% είναι υπάλληλοι πωλήσεων (32,102 περιπτώσεις).
- Το 13.06% είναι βασικό προσωπικό (27,570 περιπτώσεις).
- Το 10.12% είναι διευθυντές (21,371 περιπτώσεις).
- Υπάρχουν επίσης μικρότερα ποσοστά για άλλες επαγγελματικές κατηγορίες, όπως οδηγοί (8.81%, 18,603 περιπτώσεις), τεχνικό προσωπικό υψηλών δεξιοτήτων (5.39%, 11,380 περιπτώσεις), και ιατρικό προσωπικό (4.04%, 8,537 περιπτώσεις).

3. Πιθανότητα Αθέτησης ανά Επάγγελμα:

- Οι ανειδίκευτοι εργάτες έχουν ποσοστό χρεοκοπίας 17.15%.
- Οι οδηγοί έχουν ποσοστό χρεοκοπίας 11.33%.
- Οι σερβιτόροι/μπάρμαν έχουν ποσοστό χρεοκοπίας 11.28%.
- Οι υπάλληλοι ασφάλειας έχουν ποσοστό χρεοκοπίας 10.74%.
- Οι εργάτες έχουν ποσοστό χρεοκοπίας 10.58%.
- Οι υπάλληλοι πωλήσεων έχουν ποσοστό χρεοκοπίας 9.63%.
- Οι λογιστές έχουν το χαμηλότερο ποσοστό χρεοκοπίας, 4.83%.

	dtype	count	unique	missing_count	missing_percentage
OCCUPATION_TYPE	object	211120	18	96391	31.35

```
The unique categories of 'OCCUPATION_TYPE' are:
['Laborers' 'Core staff' 'Accountants' 'Managers' nan 'Drivers'
 'Sales staff' 'Cleaning staff' 'Cooking staff' 'Private service staff'
 'Medicine staff' 'Security staff' 'High skill tech staff'
 'Waiters/barmen staff' 'Low-skill Laborers' 'Realty agents' 'Secretaries'
 'IT staff' 'HR staff']
```



Εικόνα 5-12: Κατανομή της OCCUPATION_TYPE

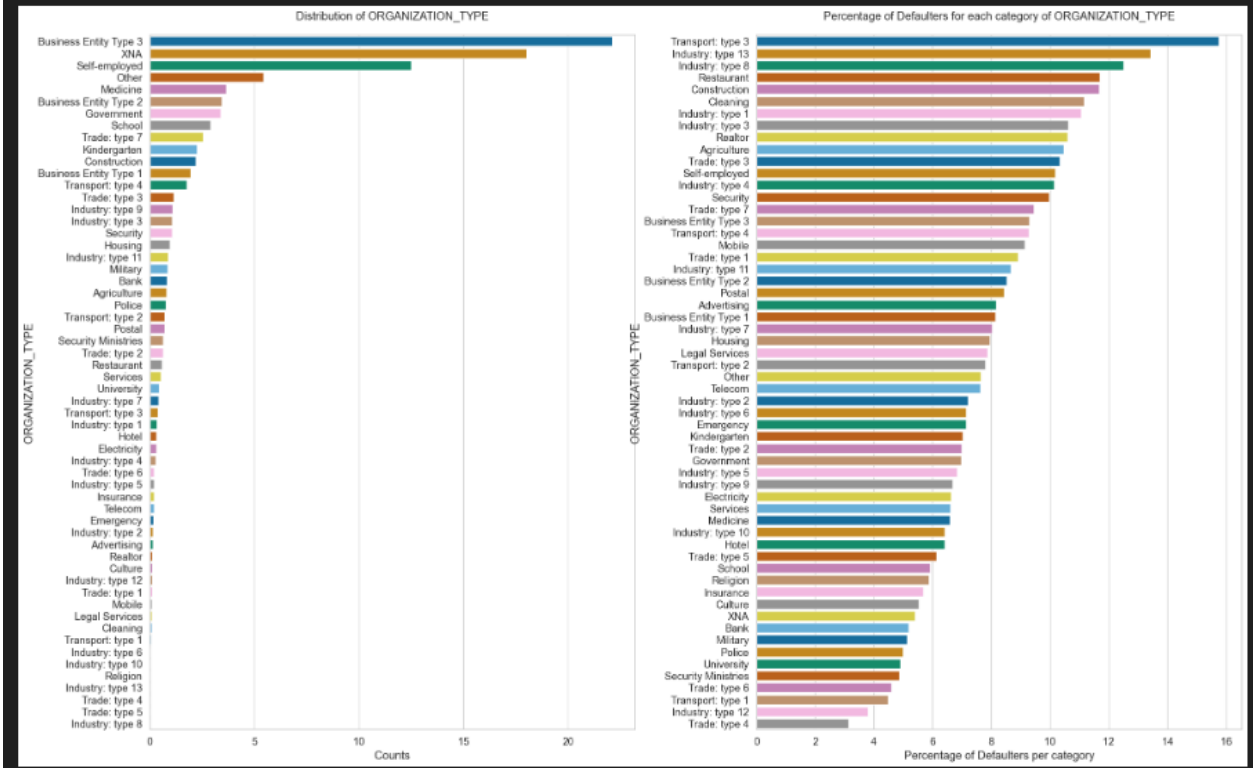
Συμπεράσματα

Οι πελάτες από επαγγέλματα όπως οι εργάτες, οι υπάλληλοι πωλήσεων και το βασικό προσωπικό αποτελούν την πλειοψηφία αυτών που λαμβάνουν δάνεια. Αυτή η κατανομή αντικατοπτρίζει την έντονη ζήτηση για δάνεια σε επαγγελματικές κατηγορίες που βρίσκονται στη μεσαία και εργατική τάξη, όπου τα δάνεια συχνά αποτελούν αναγκαίο εργαλείο. Οι επαγγελματικές κατηγορίες χαμηλότερων δεξιοτήτων, όπως οι εργάτες, οι οδηγοί και οι σερβιτόροι, έχουν υψηλότερα ποσοστά χρεοκοπίας. Αυτοί οι πελάτες, αν και συνήθως χρειάζονται δάνεια για τις καθημερινές ανάγκες τους, αντιμετωπίζουν μεγαλύτερες προκλήσεις στην αποπληρωμή τους, γεγονός που οδηγεί σε υψηλότερο οικονομικό ρίσκο. Αντίθετα, επαγγέλματα όπως οι λογιστές παρουσιάζουν μεγαλύτερη οικονομική σταθερότητα και χαμηλότερο κίνδυνο χρεοκοπίας.

5.1.12 Ανάλυση για τη Μεταβλητή ORGANIZATION_TYPE

Η μεταβλητή ORGANIZATION_TYPE αναφέρεται στον τύπο του οργανισμού στον οποίο εργάζεται ο πελάτης.

	dtype	count	unique	missing_count	missing_percentage
ORGANIZATION_TYPE	object	307511	58	0	0.0



Εικόνα 5-13: Κατανομή της ORGANIZATION_TYPE

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
- **Αριθμός κενών τιμών:** Δεν υπάρχουν απουσίες σε αυτή τη στήλη.
- **Μοναδικές τιμές:** Υπάρχουν πολλοί τύποι οργανισμών, όπως "Business Entity Type 3", "Self-employed", "Government", "Medicine", κλπ.

2. Κατανομή Οργανισμών:

- Το 22.11% των πελατών εργάζεται σε οργανισμούς τύπου "Business Entity Type 3" (67,992 περιπτώσεις).
- Το 18.01% αφορά περιπτώσεις με τιμή "XNA" (55,374 περιπτώσεις).
- Το 12.49% είναι αυτοαπασχολούμενοι (38,412 περιπτώσεις).
- Το 5.43% εργάζεται σε άλλους οργανισμούς (16,683 περιπτώσεις).
- Το 3.64% εργάζεται στον τομέα της ιατρικής (11,193 περιπτώσεις).

3. Χρεοκοπίες ανά Τύπο Οργανισμού:

- Οι πελάτες που εργάζονται σε "Transport: type 3" οργανισμούς έχουν το υψηλότερο ποσοστό χρεοκοπίας 15.75%.
- Οι πελάτες που εργάζονται σε "Industry: type 13" οργανισμούς έχουν ποσοστό χρεοκοπίας 13.43%.
- Οι πελάτες που εργάζονται σε "Restaurant" οργανισμούς έχουν ποσοστό χρεοκοπίας 11.71%.
- Οι πελάτες που εργάζονται στον κατασκευαστικό τομέα έχουν ποσοστό χρεοκοπίας 11.68%.
- Οι αυτοαπασχολούμενοι έχουν ποσοστό χρεοκοπίας 10.17%.

Συμπεράσματα

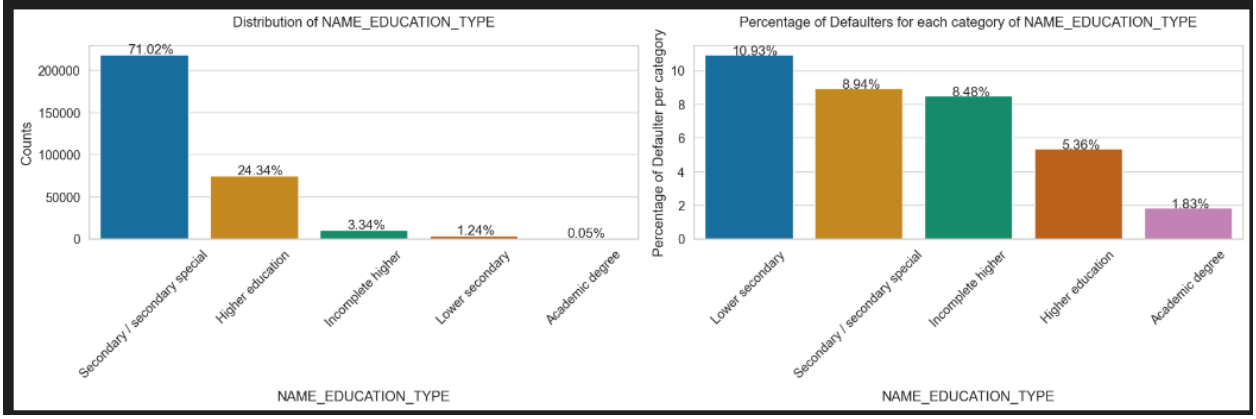
Οι πελάτες που εργάζονται σε επιχειρήσεις ή είναι αυτοαπασχολούμενοι αντιπροσωπεύουν το μεγαλύτερο ποσοστό των αιτούντων. Αυτό αντικατοπτρίζει την ανάγκη αυτών των επαγγελματικών κατηγοριών για πρόσβαση σε κεφάλαια, καθώς μπορεί να χρειάζονται δάνεια για την ανάπτυξη των δραστηριοτήτων τους. Παρατηρούμε ότι επαγγελματικές κατηγορίες όπως οι μεταφορές, η βιομηχανία και η εστίαση έχουν υψηλότερα ποσοστά χρεοκοπίας, γεγονός που πιθανώς οφείλεται στη φύση των εργασιών τους, οι οποίες μπορεί να υπόκεινται σε μεγαλύτερες οικονομικές αβεβαιότητες. Οι αυτοαπασχολούμενοι, επίσης, έχουν σχετικά υψηλό ρίσκο χρεοκοπίας, πιθανώς λόγω της αβεβαιότητας που συνδέεται με το εισόδημά τους. Αντίθετα, επαγγελματίες σε πιο σταθερές βιομηχανίες, όπως οι τράπεζες και το δημόσιο, εμφανίζουν χαμηλότερα ποσοστά χρεοκοπίας.

5.1.13 Ανάλυση για τη Μεταβλητή NAME_EDUCATION_TYPE

Η μεταβλητή NAME_EDUCATION_TYPE αναφέρεται στο επίπεδο εκπαίδευσης του πελάτη.

	dtype	count	unique	missing_count	missing_percentage
NAME_EDUCATION_TYPE	object	307511	5	0	0.0

```
The unique categories of 'NAME_EDUCATION_TYPE' are:
['Secondary / secondary special' 'Higher education' 'Incomplete higher'
 'Lower secondary' 'Academic degree']
```



Εικόνα 5-14: Κατανομή της NAME_EDUCATION_TYPE

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
- **Αριθμός κενών τιμών:** Δεν υπάρχουν.
- **Μοναδικές τιμές:** Υπάρχουν πέντε κατηγορίες εκπαίδευσης, όπως "Secondary / secondary special", "Higher education", "Incomplete higher", "Lower secondary", "Academic degree".

2. Κατανομή Εκπαιδευτικού Επιπέδου:

- Το 71.02% των πελατών έχει ολοκληρώσει δευτεροβάθμια εκπαίδευση (218,391 περιπτώσεις).
- Το 24.24% έχει ανώτατη εκπαίδευση (74,863 περιπτώσεις).
- Το 3.34% έχει ημιτελή ανώτατη εκπαίδευση (10,277 περιπτώσεις).
- Το 1.24% έχει κατώτερη δευτεροβάθμια εκπαίδευση (3,816 περιπτώσεις).
- Το 0.05% έχει ακαδημαϊκό πτυχίο (164 περιπτώσεις).

3. Πιθανότητα Αθέτησης ανά Επίπεδο Εκπαίδευσης:

- Οι πελάτες με κατώτερη δευτεροβάθμια εκπαίδευση έχουν ποσοστό χρεοκοπίας 10.93%.
- Οι πελάτες με δευτεροβάθμια/δευτεροβάθμια ειδική εκπαίδευση έχουν ποσοστό χρεοκοπίας 8.94%.
- Οι πελάτες με ημιτελή ανώτατη εκπαίδευση έχουν ποσοστό χρεοκοπίας 8.48%.
- Οι πελάτες με ανώτατη εκπαίδευση έχουν ποσοστό χρεοκοπίας 5.36%.
- Οι πελάτες με ακαδημαϊκό πτυχίο έχουν το χαμηλότερο ποσοστό χρεοκοπίας, 1.83%.

Συμπεράσματα

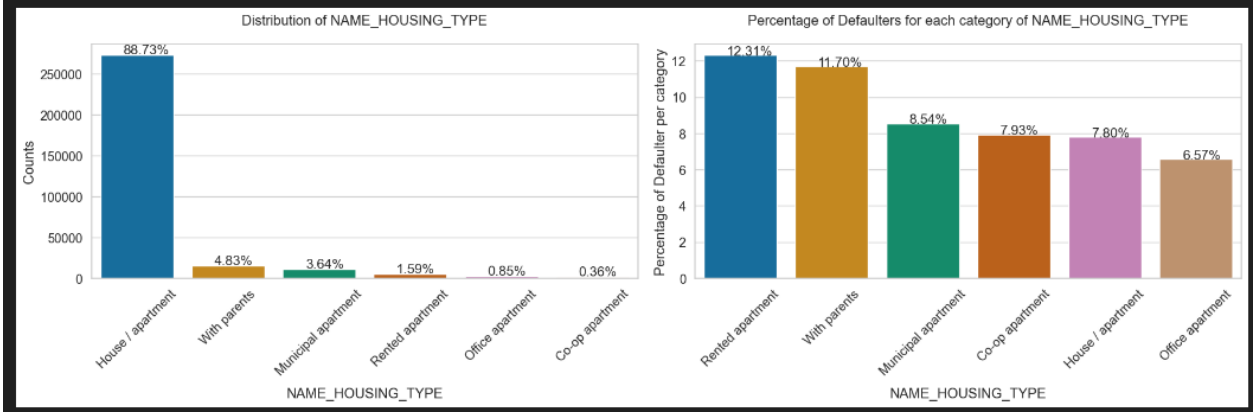
Η πλειοψηφία των πελατών έχει δευτεροβάθμια εκπαίδευση, κάτι που αντικατοπτρίζει το εκπαιδευτικό επίπεδο του γενικού πληθυσμού. Παρατηρούμε ότι οι πελάτες με υψηλότερο επίπεδο εκπαίδευσης παρουσιάζουν χαμηλότερο κίνδυνο χρεοκοπίας. Η ανώτατη και ακαδημαϊκή εκπαίδευση φαίνεται να παρέχει οικονομική σταθερότητα, ενώ οι πελάτες με κατώτερη δευτεροβάθμια εκπαίδευση έχουν το υψηλότερο ποσοστό χρεοκοπίας, πιθανώς λόγω της περιορισμένης πρόσβασης σε υψηλότερα εισοδήματα.

5.1.14 Ανάλυση για τη Μεταβλητή NAME_HOUSING_TYPE

Η μεταβλητή NAME_HOUSING_TYPE αναφέρεται στον τύπο κατοικίας του πελάτη.

	dtype	count	unique	missing_count	missing_percentage
NAME_HOUSING_TYPE	object	307511	6	0	0.0

```
----------------------------------------------------------------------------------------------------
The unique categories of 'NAME_HOUSING_TYPE' are:
['House / apartment' 'Rented apartment' 'With parents'
 'Municipal apartment' 'Office apartment' 'Co-op apartment']
----------------------------------------------------------------------------------------------------
```



Εικόνα 5-15: Κατανομή της NAME_HOUSING_TYPE

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα.
- **Αριθμός κενών τιμών:** Δεν υπάρχουν.
- **Μοναδικές τιμές:** Υπάρχουν έξι τύποι κατοικίας "House / apartment", "With parents", "Municipal apartment", "Rented apartment","Office apartment","Co-op apartment".

2. Κατανομή Τύπων Κατοικίας:

- Το 88.73% των πελατών μένει σε σπίτι ή διαμέρισμα (272,868 περιπτώσεις).
- Το 4.83% μένει με τους γονείς τους (14,840 περιπτώσεις).
- Το 3.64% μένει σε δημόσιο ακίνητο(11,183 περιπτώσεις).
- Το 1.59% μένει σε ενοικιαζόμενο διαμέρισμα (4,881 περιπτώσεις).
- Υπάρχουν μικρότερα ποσοστά για γραφεία (0.85%, 2,617 περιπτώσεις) και συνιδιοκτησίες (0.36%, 1,122 περιπτώσεις).

3. Χρεοκοπίες ανά Τύπο Κατοικίας:

- Οι πελάτες που μένουν σε ενοικιαζόμενο διαμέρισμα έχουν το υψηλότερο ποσοστό χρεοκοπίας, 12.31%.
- Οι πελάτες που μένουν με τους γονείς τους έχουν ποσοστό χρεοκοπίας 11.70%.
- Οι πελάτες που μένουν σε δημόσιο ακίνητο έχουν ποσοστό χρεοκοπίας 8.54%.
- Οι πελάτες που μένουν σε σπίτι ή διαμέρισμα έχουν ποσοστό χρεοκοπίας 7.80%.
- Οι πελάτες που μένουν σε διαμέρισμα με συνιδιοκτησία έχουν ποσοστό χρεοκοπίας 7.93%.
- Οι πελάτες που έχουν γραφείο έχουν το χαμηλότερο ποσοστό χρεοκοπίας, 6.56%.

Συμπεράσματα

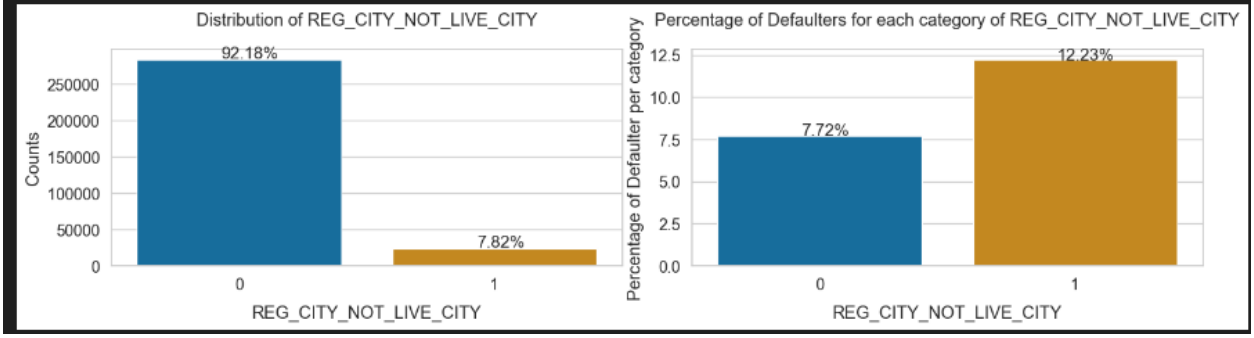
Η πλειοψηφία των πελατών κατοικεί σε ιδιόκτητα σπίτια ή διαμερίσματα, κάτι που αντικατοπτρίζει την προτίμηση για μόνιμη κατοικία. Οι μικρότερες κατηγορίες, όπως τα ενοικιαζόμενα σχετίζονται περισσότερο με οικονομικά πιο ασταθείς καταστάσεις. Η ανάλυση δείχνει ότι όσοι μένουν σε ενοικιαζόμενες κατοικίες ή με τους γονείς τους παρουσιάζουν αυξημένο ρίσκο χρεοκοπίας, πιθανώς λόγω της οικονομικής αστάθειας που συνδέεται με αυτούς τους τύπους κατοικιών. Αντίθετα, όσοι μένουν σε ιδιόκτητα σπίτια φαίνεται να έχουν καλύτερη οικονομική σταθερότητα.

5.1.15 Ανάλυση για τη Μεταβλητή REG_CITY_NOT_LIVE_CITY

Η μεταβλητή REG_CITY_NOT_LIVE_CITY δείχνει αν η πόλη της μόνιμης διαμονής του πελάτη διαφέρει από την πόλη στην οποία έχει δηλωθεί ως κάτοικος.

	dtype	count	unique	missing_count	missing_percentage
REG_CITY_NOT_LIVE_CITY	int64	307511	2	0	0.0

```
The unique categories of 'REG_CITY_NOT_LIVE_CITY' are:
[0 1]
```



Εικόνα 5-16: Κατανομή της REG_CITY_NOT_LIVE_CITY

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** int, δηλαδή δυαδικά αριθμητικά δεδομένα (0 ή 1).
- **Αριθμός κενών τιμών:** Δεν υπάρχουν απουσίες σε αυτή τη στήλη.
- **Μοναδικές τιμές:** Οι τιμές 0 (ίδια πόλη) και 1 (διαφορετική πόλη).

2. Κατανομή Πόλης Διαμονής και Κατοικίας:

- Το 92.18% των πελατών ζει στην ίδια πόλη στην οποία έχει δηλωθεί μόνιμα.
- Το 7.82% των πελατών ζει σε διαφορετική πόλη από εκείνη στην οποία είναι δηλωμένοι.

3. Χρεοκοπίες Ανάλογα με το Αν Ζει στην Ίδια ή Διαφορετική Πόλη:

- Οι πελάτες που ζουν στην ίδια πόλη έχουν ποσοστό χρεοκοπίας 7.72%.
- Οι πελάτες που ζουν σε διαφορετική πόλη έχουν ποσοστό χρεοκοπίας 12.23%.

Συμπεράσματα

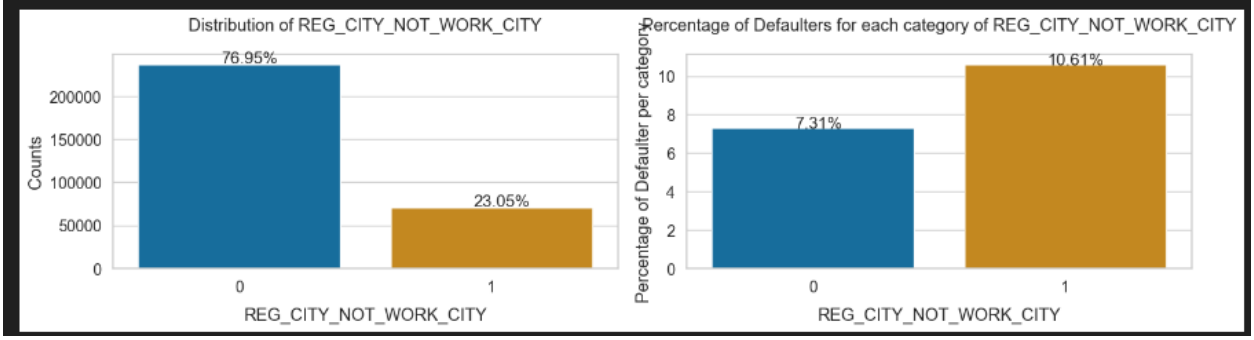
Αυτή η ανάλυση δείχνει ότι η συντριπτική πλειοψηφία των πελατών ζει στην ίδια πόλη με τη δηλωμένη τους διεύθυνση, γεγονός που υποδηλώνει σταθερότητα στη μόνιμη διαμονή. Οι πελάτες που ζουν σε διαφορετική πόλη από εκείνη στην οποία είναι δηλωμένοι φαίνεται να έχουν αυξημένο ρίσκο χρεοκοπίας, πιθανώς λόγω της οικονομικής αστάθειας ή των δυσκολιών που σχετίζονται με τη μετακίνηση και τη διαχείριση δύο διαφορετικών τόπων κατοικίας. Αντίθετα, όσοι ζουν στην ίδια πόλη έχουν χαμηλότερο ρίσκο χρεοκοπίας.

5.1.16 Ανάλυση για τη Μεταβλητή REG_CITY_NOT_WORK_CITY

Η μεταβλητή REG_CITY_NOT_WORK_CITY δείχνει αν η πόλη της μόνιμης κατοικίας του πελάτη διαφέρει από την πόλη εργασίας του.

	dtype	count	unique	missing_count	missing_percentage
REG_CITY_NOT_WORK_CITY	int64	307511	2	0	0.0

```
The unique categories of 'REG_CITY_NOT_WORK_CITY' are:
[0 1]
```



Εικόνα 5-17: Κατανομή της REG_CITY_NOT_WORK_CITY

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** int, αριθμητικές τιμές.
- **Αριθμός κενών τιμών:** Δεν υπάρχουν.
- **Μοναδικές τιμές:** Οι τιμές 0 (ίδια πόλη κατοικίας και εργασίας) και 1 (διαφορετική πόλη).

2. Κατανομή Πόλης Κατοικίας και Εργασίας:

- Το 76.95% των πελατών ζει και εργάζεται στην ίδια πόλη.
- Το 23.05**%** των πελατών ζει σε διαφορετική πόλη από την πόλη εργασίας τους.

3. Πιθανότητα Αθέτησης Ανάλογα με το Αν Ζει και Εργάζεται στην Ίδια Πόλη:

- Οι πελάτες που ζουν και εργάζονται στην ίδια πόλη έχουν ποσοστό αθέτησης 7.31%.
- Οι πελάτες που ζουν και εργάζονται σε διαφορετικές πόλεις έχουν ποσοστό αθέτησης 10.61%.

Συμπεράσματα

Αυτό δείχνει ότι η πλειονότητα των πελατών ζει στην ίδια πόλη με την πόλη εργασίας τους, γεγονός που μπορεί να υποδηλώνει σταθερότητα στη μετακίνηση και τη διαβίωση. Οι πελάτες που εργάζονται σε διαφορετική πόλη από εκείνη στην οποία ζουν φαίνεται να έχουν μεγαλύτερο

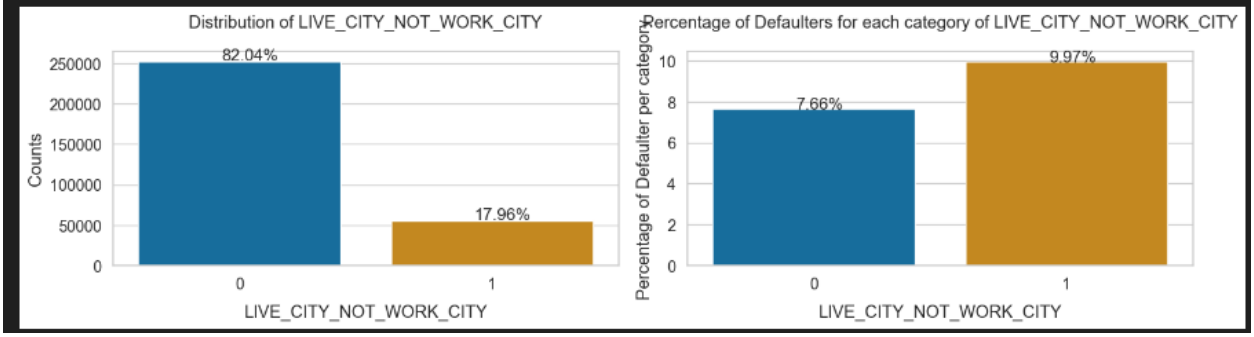
ρίσκο, πιθανώς λόγω των επιπλέον οικονομικών υποχρεώσεων ή των δυσκολιών που προκύπτουν από την καθημερινή μετακίνηση. Αντίθετα, όσοι ζουν και εργάζονται στην ίδια πόλη φαίνεται να αντιμετωπίζουν λιγότερες οικονομικές δυσκολίες, γεγονός που αντανακλάται στο χαμηλότερο ποσοστό αθέτησης.

5.1.17 Ανάλυση για τη Μεταβλητή LIVE_CITY_NOT_WORK_CITY

Η μεταβλητή LIVE_CITY_NOT_WORK_CITY δείχνει αν η πόλη της διαμονής του πελάτη διαφέρει από την πόλη εργασίας του.

	dtype	count	unique	missing_count	missing_percentage
LIVE_CITY_NOT_WORK_CITY	int64	307511	2	0	0.0

```
The unique categories of 'LIVE_CITY_NOT_WORK_CITY' are:
[0 1]
```



Εικόνα 5-18: Κατανομή της LIVE_CITY_NOT_WORK_CITY

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** int, αριθμητικά δεδομένα.
- **Αριθμός κενών τιμών:** Δεν υπάρχουν.
- **Μοναδικές τιμές:** Οι τιμές 0 (ίδια πόλη διαμονής και εργασίας) και 1 (διαφορετική πόλη).

2. Κατανομή Πόλης Διαμονής και Εργασίας:

- Το 82.04% των πελατών ζει και εργάζεται στην ίδια πόλη.
- Το 17.96% ζει σε διαφορετική πόλη από την πόλη εργασίας τους.

3. Πιθανότητα Αθέτησης Ανάλογα με το Αν Ζει και Εργάζεται στην Ίδια ή Διαφορετική Πόλη:

- Οι πελάτες που ζουν και εργάζονται στην ίδια πόλη έχουν ποσοστό αθέτησης 7.66%.
- Οι πελάτες που ζουν και εργάζονται σε διαφορετικές πόλεις έχουν ποσοστό αθέτησης 9.97%.

Συμπεράσματα

Αυτό δείχνει ότι η πλειονότητα των πελατών ζει και εργάζεται στην ίδια πόλη. Οι πελάτες που εργάζονται σε διαφορετική πόλη από εκείνη στην οποία ζουν αντιμετωπίζουν ελαφρώς αυξημένο

ρίσκο, πιθανώς λόγω των επιπλέον εξόδων ή των δυσκολιών μετακίνησης. Αντίθετα, όσοι ζουν και εργάζονται στην ίδια πόλη φαίνεται να έχουν χαμηλότερο οικονομικό ρίσκο.

5.1.18 Ανάλυση των FLAG_DOCUMENTS

Η ανάλυση των μεταβλητών FLAG_DOCUMENTS αφορά τα έγγραφα που προσκομίζονται ή λείπουν από τους πελάτες κατά τη διαδικασία αίτησης για δάνειο. Παρατηρούμε ότι οι περισσότερες μεταβλητές των εγγράφων έχουν πολύ χαμηλή συχνότητα εμφάνισης της τιμής 1 (που σημαίνει ότι το έγγραφο είναι διαθέσιμο), εκτός από την FLAG_DOCUMENT_3.

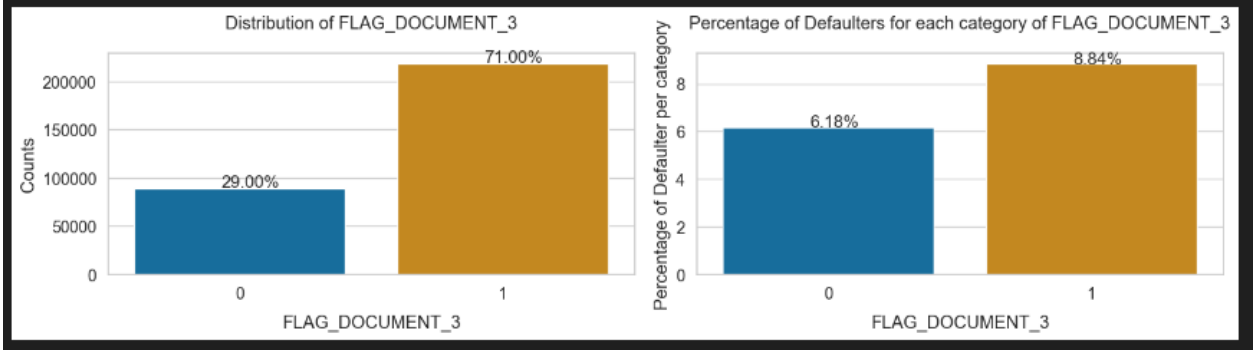
```
for column in flag_document_df:

    count_0 = flag_document_df[column].value_counts()[0]
    count_1 = flag_document_df[column].value_counts()[1]
    total_rows = flag_document_df.shape[0]

    percent_0 = np.round((count_0*100/total_rows),2)
    percent_1 = np.round(100 - percent_0,2)

    print(column, "contains percentage of 1's = ",percent_1,\
    "and percentage of 0's =", percent_0)
```

```
FLAG_DOCUMENT_2 contains percentage of 1's =  0.0 and percentage of 0's = 100.0
FLAG_DOCUMENT_3 contains percentage of 1's =  71.0 and percentage of 0's = 29.0
FLAG_DOCUMENT_4 contains percentage of 1's =  0.01 and percentage of 0's = 99.99
FLAG_DOCUMENT_5 contains percentage of 1's =  1.51 and percentage of 0's = 98.49
FLAG_DOCUMENT_6 contains percentage of 1's =  8.81 and percentage of 0's = 91.19
FLAG_DOCUMENT_7 contains percentage of 1's =  0.02 and percentage of 0's = 99.98
FLAG_DOCUMENT_8 contains percentage of 1's =  8.14 and percentage of 0's = 91.86
FLAG_DOCUMENT_9 contains percentage of 1's =  0.39 and percentage of 0's = 99.61
FLAG_DOCUMENT_10 contains percentage of 1's =  0.0 and percentage of 0's = 100.0
FLAG_DOCUMENT_11 contains percentage of 1's =  0.39 and percentage of 0's = 99.61
FLAG_DOCUMENT_12 contains percentage of 1's =  0.0 and percentage of 0's = 100.0
FLAG_DOCUMENT_13 contains percentage of 1's =  0.35 and percentage of 0's = 99.65
FLAG_DOCUMENT_14 contains percentage of 1's =  0.29 and percentage of 0's = 99.71
FLAG_DOCUMENT_15 contains percentage of 1's =  0.12 and percentage of 0's = 99.88
FLAG_DOCUMENT_16 contains percentage of 1's =  0.99 and percentage of 0's = 99.01
FLAG_DOCUMENT_17 contains percentage of 1's =  0.03 and percentage of 0's = 99.97
FLAG_DOCUMENT_18 contains percentage of 1's =  0.81 and percentage of 0's = 99.19
FLAG_DOCUMENT_19 contains percentage of 1's =  0.06 and percentage of 0's = 99.94
FLAG_DOCUMENT_20 contains percentage of 1's =  0.05 and percentage of 0's = 99.95
FLAG_DOCUMENT_21 contains percentage of 1's =  0.03 and percentage of 0's = 99.97
```

Εικόνα 5-19: Βαθμός Εμφάνισης FLAG_DOCUMENTS

	dtype	count	unique	missing_count	missing_percentage
FLAG_DOCUMENT_3	int64	307511	2	0	0.0

```
The unique categories of 'FLAG_DOCUMENT_3' are:
[1 0]
```



Εικόνα 5-20: Κατανομή της FLAG_DOCUMENT_3

1. Κατανομή του FLAG_DOCUMENT_3:

- Το 71.00% των πελατών έχει προσκομίσει το έγγραφο FLAG_DOCUMENT_3.
- Το 29.00% των πελατών δεν έχει προσκομίσει το έγγραφο αυτό.

2. Πιθανότητα Αθέτησης Ανάλογα με την Κατάσταση του FLAG_DOCUMENT_3:

- Οι πελάτες που δεν προσκόμισαν το έγγραφο έχουν ποσοστό αθέτησης 6.18%.
- Οι πελάτες που προσκόμισαν το έγγραφο έχουν ποσοστό αθέτησης 8.84%.

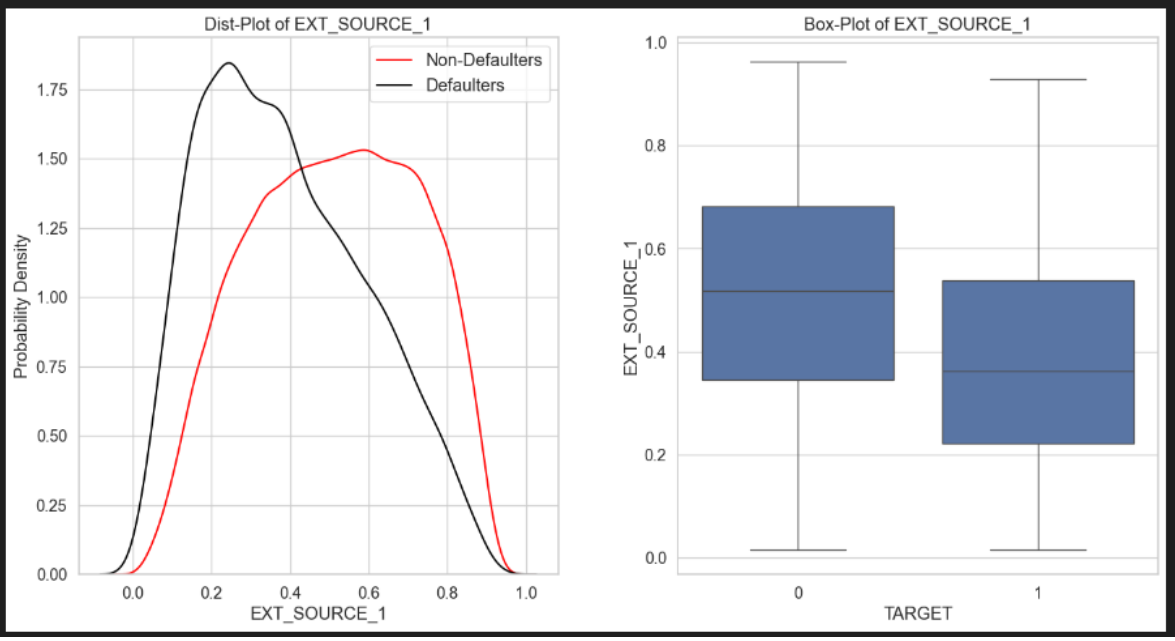
Συμπεράσματα

Αυτή η μεταβλητή φαίνεται να είναι η μόνη από τα FLAG_DOCUMENTS που έχει μια σχετικά ισορροπημένη κατανομή και μπορεί να είναι χρήσιμη στην ανάλυση, καθώς οι περισσότερες από τις άλλες μεταβλητές έχουν σχεδόν μηδενική παρουσία της τιμής 1. Το FLAG_DOCUMENT_3 εμφανίζει υψηλότερο ρίσκο χρεοκοπίας για όσους έχουν προσκομίσει το έγγραφο, αν και το ποσοστό της τιμής 1 είναι πολύ υψηλότερο από τις υπόλοιπες μεταβλητές εγγράφων. Οι υπόλοιπες μεταβλητές των εγγράφων (π.χ. FLAG_DOCUMENT_2, FLAG_DOCUMENT_4) δεν παρουσιάζουν παρόμοια ισορροπία, με το 99% ή και περισσότερο των πελατών να μην έχουν προσκομίσει τα αντίστοιχα έγγραφα.

5.1.19 Ανάλυση των EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3

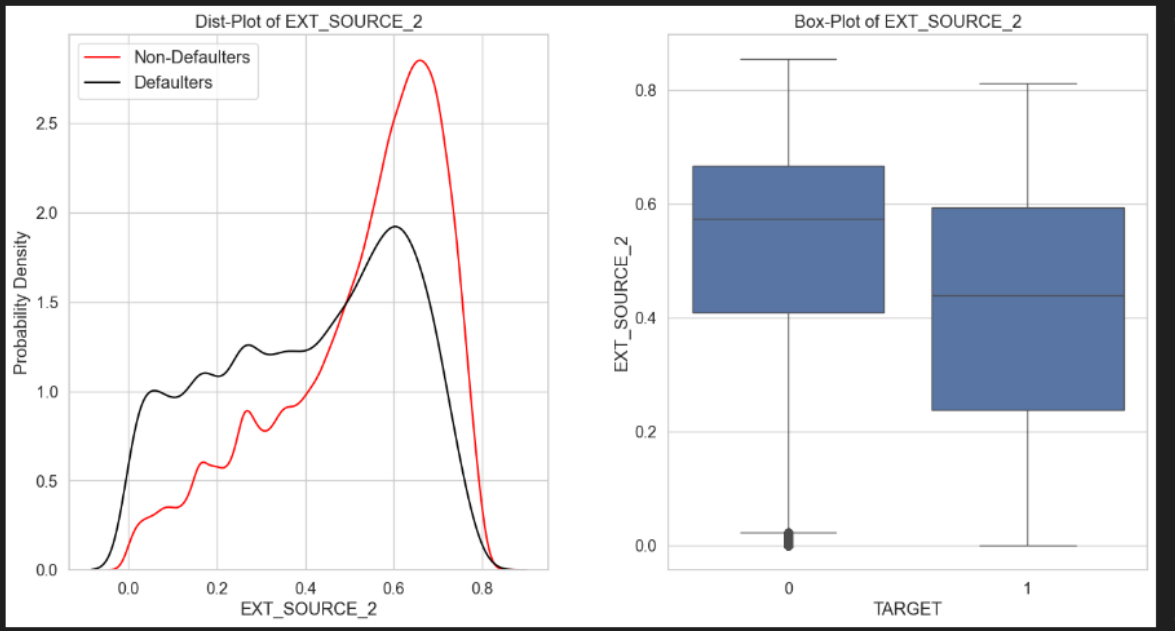
Οι μεταβλητές EXT_SOURCE_1, EXT_SOURCE_2, και EXT_SOURCE_3 είναι εξωτερικές πηγές αξιολόγησης του πιστωτικού κινδύνου και χρησιμοποιούνται από το σύστημα για την εκτίμηση της πιθανότητας αθέτησης του δανείου. Η ανάλυση αυτών των μεταβλητών είναι ιδιαίτερα σημαντική καθώς συσχετίζονται αρνητικά με την πιθανότητα αθέτησης .

que	top_value_counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
584	{0.62: 5, 0.44: 5, 0.55: 5}	173378	56.38	0.5	0.21	0.01	0.51	0.96	-0.16



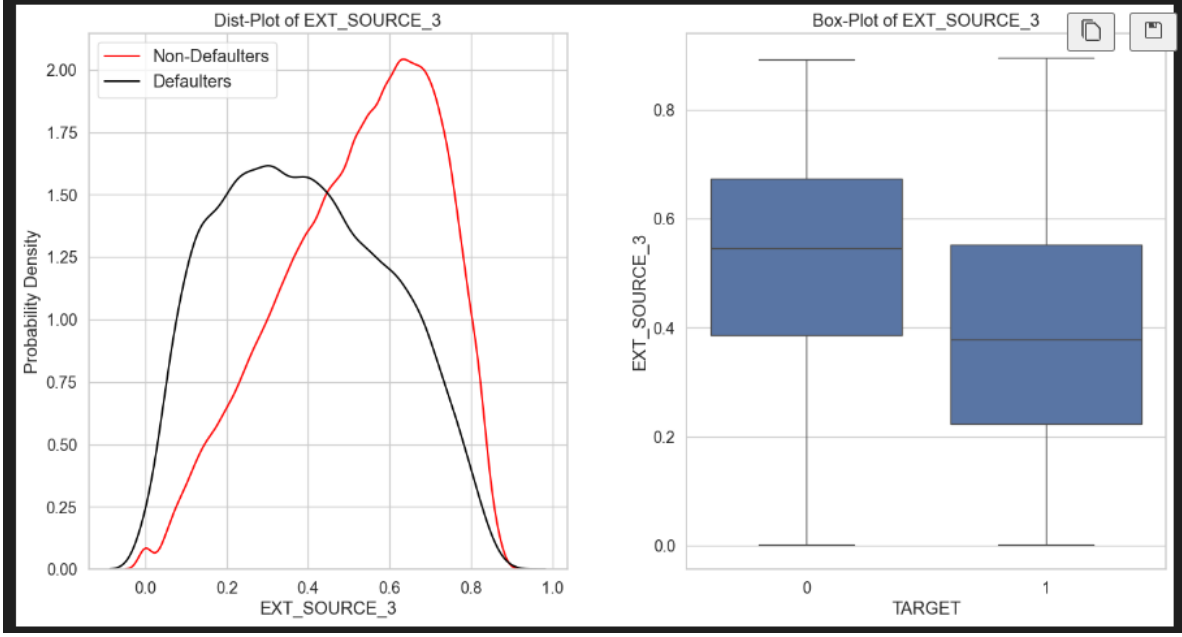
Εικόνα 5-21: Διαγράμματα κατανομής της EXT_SOURCE_1

que	top_value_counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
9831	{0.29: 721, 0.26: 417, 0.27: 343}	660	0.21	0.51	0.19	0.0	0.57	0.85	-0.16



Εικόνα 5-22: Διαγράμματα κατανομής της EXT_SOURCE_2

que	top_value_counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
314	{0.75: 1460, 0.71: 1315, 0.69: 1276}	60965	19.83	0.51	0.19	0.0	0.54	0.9	-0.18



Εικόνα 5-23: Διαγράμματα κατανομής της EXT_SOURCE_3

Από την ανάλυση προκύπτουν τα εξής:

1.Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** float, δηλαδή αριθμητικά δεδομένα.
- **EXT_SOURCE_1**: Αριθμητική μεταβλητή με 43.6% ελλείψεις.
- **EXT_SOURCE_2**: Αριθμητική μεταβλητή με σχεδόν πλήρη δεδομένα 0.21%.
- **EXT_SOURCE_3**: Αριθμητική μεταβλητή με 19.83% ελλείψεις.
- Οι τιμές κυμαίνονται από 0 έως 1, με υψηλότερες τιμές να υποδηλώνουν χαμηλότερο κίνδυνο.

2.Συσχέτιση με τη Πιθανότητα Αθέτησης

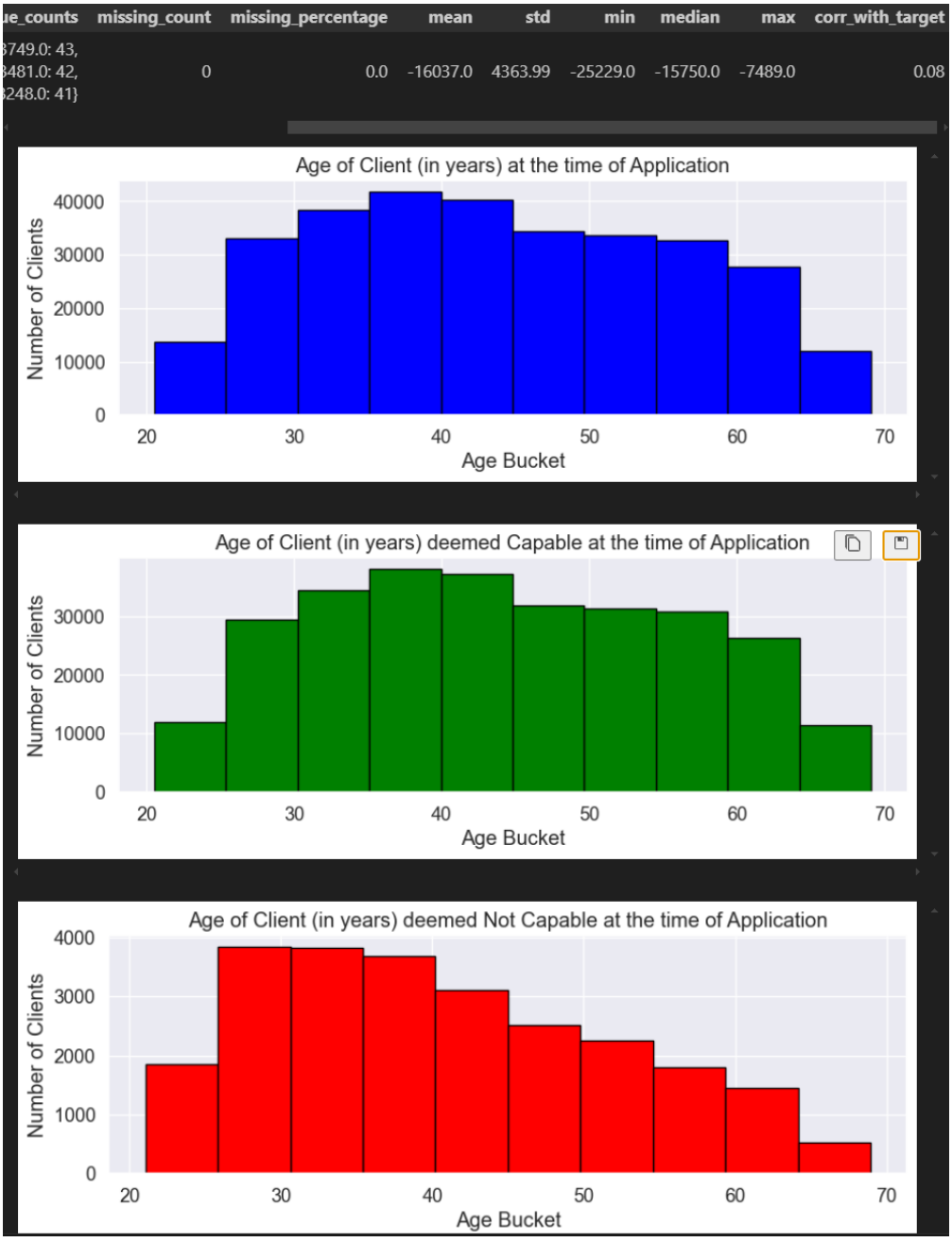
- Και οι τρεις μεταβλητές παρουσιάζουν αρνητική συσχέτιση(-0.16,-0.18) με την πιθανότητα αθέτησης. Όσο υψηλότερη η τιμή, τόσο χαμηλότερη η πιθανότητα αθέτησης δανείου. Η EXT_SOURCE_3 εμφανίζει την ισχυρότερη διαχωριστική ικανότητα.

Συμπεράσματα

Οι EXT_SOURCE μεταβλητές φαίνεται παίζουν σημαντικό ρόλο στην εκτίμηση της πιθανότητας αθέτησης δανείου. Αυτοί που προβλήματα αποπληρωμής τείνουν να έχουν χαμηλότερες τιμές στις μεταβλητές EXT_SOURCE ενώ αυτοί που δεν έχουν δείχνουν μεγαλύτερη πυκνότητα σε υψηλές τιμές αυτών των μεταβλητών, υποδεικνύοντας μικρότερο κίνδυνο. Οι μεταβλητές EXT_SOURCE_1 και EXT_SOURCE_3 διαχωρίζουν καλύτερα τους κακοπληρωτές από τους καλοπληρωτές συγκριτικά με την EXT_SOURCE_2 με την EXT_SOURCE_3 να είναι ο καλύτερος διαχωριστικός παράγοντας μεταξύ αυτών. Αν και η συσχέτιση τους με την πιθανότητα αθέτησης δεν είναι ιδιαίτερα ισχυρή, αυτές οι μεταβλητές μπορούν να βελτιώσουν την ακρίβεια μοντέλων πρόβλεψης.

5.1.20 Ανάλυση της Μεταβλητής DAYS_BIRTH

Η μεταβλητή DAYS_BIRTH αναφέρεται στην ηλικία του πελάτη κατά την υποβολή της αίτησης δανείου, εκφρασμένη σε ημέρες από την ημερομηνία γέννησης. Για την καλύτερη κατανόηση και ανάλυση των δεδομένων, οι ημέρες μετατράπηκαν σε έτη διαιρώντας τις τιμές με το 365.



Εικόνα 5-24: Ιστογράμματα κατανομής της DAYS_BIRTH

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- Τύπος δεδομένων: int64 , Αριθμητικές τιμές.
- Αριθμός κενών τιμών: Δεν υπάρχουν ελλείψεις

2. Στατιστικά Περιγραφής

- **Μέσος όρος ηλικίας**: 44 έτη. Οι πελάτες έχουν κατά μέσο όρο ηλικία 44 ετών κατά την αίτηση.
- **Τυπική απόκλιση**: 11.95 έτη, υποδηλώνοντας ποικιλία ηλικιών μεταξύ των πελατών.
- **Ελάχιστη ηλικία**: 21 έτη.
- **Διάμεσος ηλικία**: 43 έτη. Η διάμεσος τιμή είναι χαμηλότερη από τον μέσο όρο, γεγονός που δείχνει ότι αρκετοί πελάτες βρίσκονται σε μικρότερες ηλικίες.
- **Μέγιστη ηλικία**: 69 έτη.

3. Κατανομή της Μεταβλητής

- Η κατανομή δείχνει ότι οι περισσότεροι πελάτες είναι μεταξύ 30 και 60 ετών, με την πλειονότητα να συγκεντρώνεται γύρω στα 35-45 έτη.
- Η κατανομή είναι σχετικά συμμετρική, με μια μικρή προτίμηση σε μεγαλύτερες ηλικιακές ομάδες.

4. Συσχέτιση με την Πιθανότητα Αθέτησης

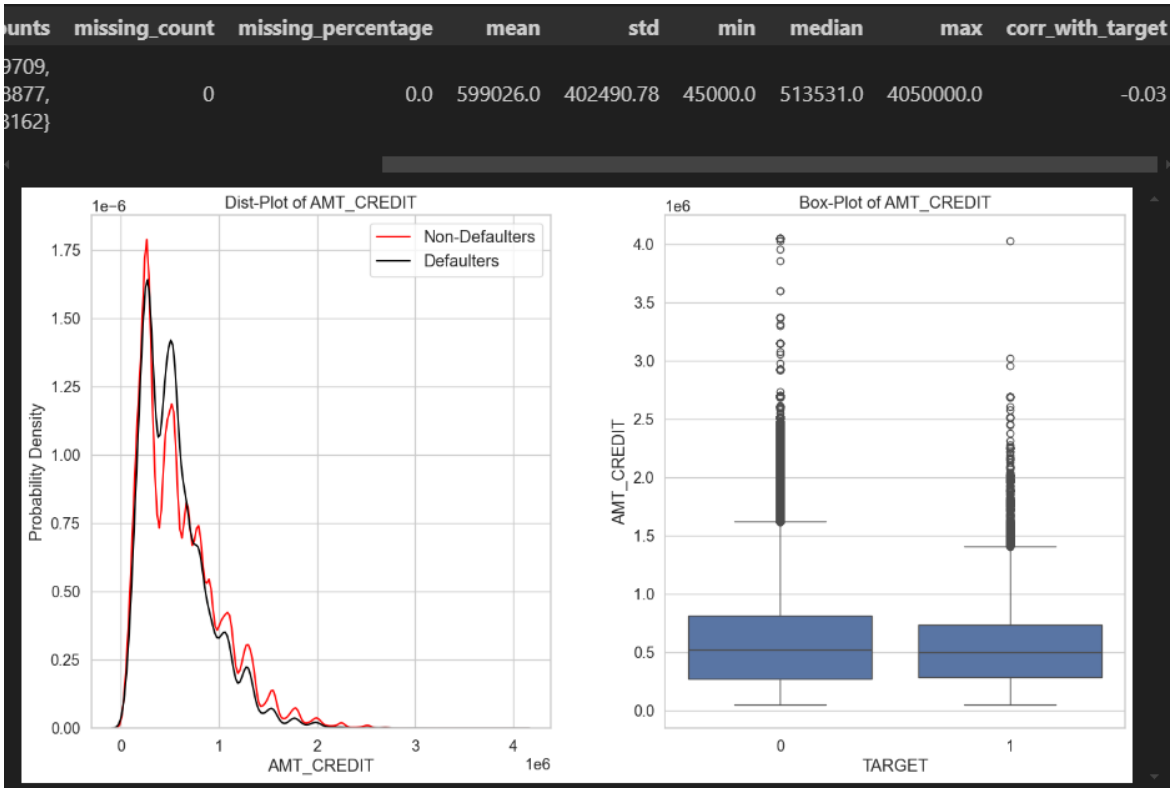
- Η μεταβλητή DAYS_BIRTH παρουσιάζει μια ελαφρά θετική συσχέτιση (0.08) με την πιθανότητα αθέτησης, υποδηλώνοντας ότι οι μεγαλύτερης ηλικίας πελάτες είναι ελαφρώς λιγότερο πιθανό να αθετήσουν το δάνειο.

Συμπεράσματα

Η μεταβλητή **DAYS_BIRTH**, αποκαλύπτει μια σαφή τάση όπου οι μεγαλύτερης ηλικίας πελάτες είναι πιο πιθανό να αποπληρώσουν τα δάνειά τους ενώ οι πελάτες ηλικίας 30-40 ετών φαίνεται να παρουσιάζουν μεγαλύτερη πιθανότητα χρεοκοπίας. Παρά τη θετική συσχέτιση της ηλικίας με την ικανότητα αποπληρωμής, η επίδραση είναι σχετικά μικρή, και άλλοι παράγοντες πιθανότατα επηρεάζουν περισσότερο την πιθανότητα χρεοκοπίας.

5.1.21 Ανάλυση της Μεταβλητής AMT_CREDIT

Η μεταβλητή AMT_CREDIT αναφέρεται στο ποσό δανείου που αιτείται ένας πελάτης κατά την υποβολή της αίτησης.



Εικόνα 5-25: Διαγράμματα κατανομής της AMT_CREDIT

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Τύπος δεδομένων**: float64, Αριθμητικές τιμές .
- **Αριθμός κενών τιμών**: Δεν υπάρχουν ελλείψεις.

2. Συμπεράσματα από τα Διαγράμματα

- Από το διάγραμμα κατανομής, βλέπουμε ότι η κατανομή της μεταβλητής AMT_CREDIT είναι έντονα δεξιά ασύμμετρη. Αυτό σημαίνει ότι το μεγαλύτερο μέρος των αιτήσεων δανείου αφορά μικρότερα ποσά, ενώ υπάρχουν λίγες αιτήσεις για πολύ μεγάλα δάνεια.

- Οι πελάτες που έχουν αθετήσει δάνειο έχουν γενικά χαμηλότερα ποσά δανείου σε σύγκριση με αυτούς που δεν έχουν. Παρατηρούμε ότι η κατανομή των καλοπληρωτών εμφανίζει μεγαλύτερη διασπορά, με αρκετές περιπτώσεις υψηλών δανείων. Οι κακοπληρωτές τείνουν να συγκεντρώνονται σε μικρότερα ποσά δανείου, κάτι που μπορεί να σχετίζεται με την οικονομική τους κατάσταση ή την έλλειψη εγγυήσεων για μεγαλύτερα ποσά.

3. Συσχέτιση με την Πιθανότητα Αθέτησης

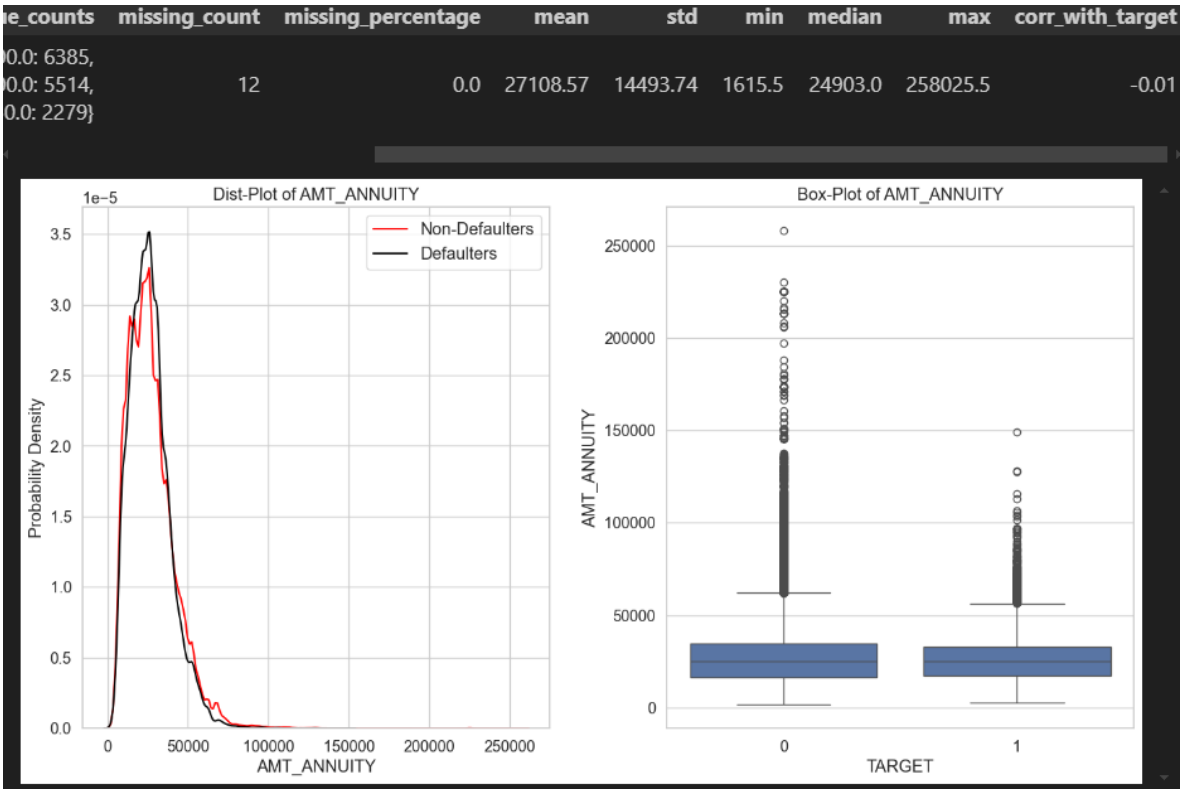
- Η συσχέτιση με την αθέτηση δανείου είναι ασθενής (-0.03), κάτι που δείχνει ότι το ποσό δανείου από μόνο του δεν είναι αποφασιστικός παράγοντας για την πρόβλεψη αθέτησης. Ωστόσο, τα παραπάνω ευρήματα από τα διαγράμματα δείχνουν ότι τα χαμηλότερα ποσά δανείου μπορεί να σχετίζονται με υψηλότερο κίνδυνο αθέτησης.

Συμπεράσματα

Η AMT_CREDIT παρουσιάζει σημαντική ασυμμετρία, με την πλειοψηφία των αιτήσεων να επικεντρώνεται σε μικρότερα ποσά δανείου, ενώ οι μεγάλοι δανειολήπτες αποτελούν μικρό ποσοστό. Οι κακοπληρωτές τείνουν να έχουν μικρότερα δάνεια, αλλά η γενική συσχέτιση με την πιθανότητα αθέτησης είναι αδύναμη. Η μεταβλητή αυτή μπορεί να συνεισφέρει στην ανάλυση κινδύνου, κυρίως όταν συνδυαστεί με άλλα χαρακτηριστικά.

5.1.22 Ανάλυση της Μεταβλητής AMT_ANNUITY

Η μεταβλητή **AMT_ANNUITY** αναφέρεται στο ποσό της ετήσιας δόσης που πρέπει να πληρώσει ο πελάτης για το δάνειο.



Εικόνα 5-26: Διαγράμματα κατανομής της AMT_ANNUITY

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Τύπος δεδομένων**: float64 (Αριθμητικές τιμές).
- **Αριθμός κενών τιμών**: 12 εγγραφές έχουν κενές τιμές.

2. Συμπεράσματα από τα Διαγράμματα

- Η κατανομή της AMT_ANNUITY δείχνει δεξιά ασυμμετρία, με την πλειοψηφία των πελατών να πληρώνει μικρότερες ετήσιες δόσεις και λίγους πελάτες να πληρώνουν πολύ υψηλότερες δόσεις. Οι κακοπληρωτές παρουσιάζουν γενικά ελαφρώς χαμηλότερες ετήσιες δόσεις σε σύγκριση με τους που δεν αθέτησαν. Οι καλοπληρωτές παρουσιάζουν μεγαλύτερη διασπορά στις τιμές των δόσεων, με αρκετές ακραίες τιμές που αντιπροσωπεύουν υψηλές ετήσιες δόσεις.

3. Συσχέτιση με την Πιθανότητα Αθέτησης

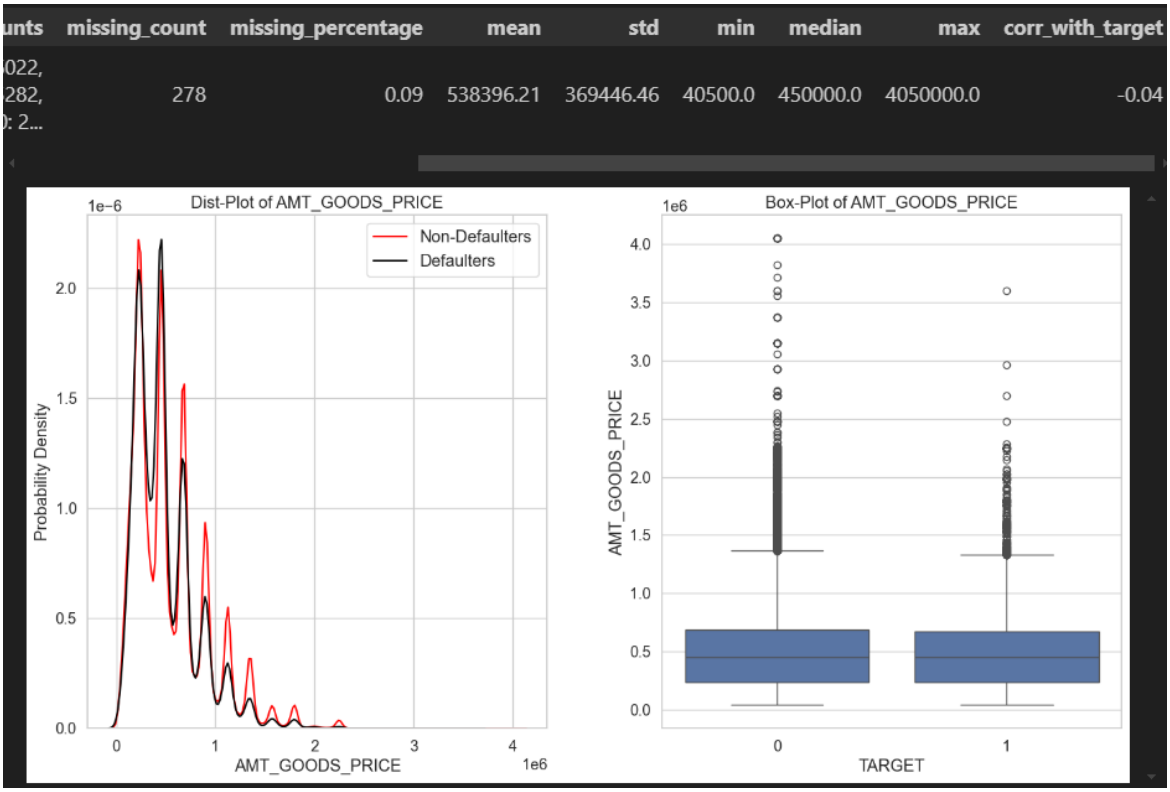
- Η συσχέτιση της AMT_ANNUITY με την πιθανότητα αθέτησης δανείου είναι πολύ χαμηλή (-0.01). Παρά το γεγονός ότι η συσχέτιση είναι χαμηλή, η διαφορά στην κατανομή ανάμεσα στους κακοπληρωτές και καλοπληρωτές δείχνει ότι οι πελάτες με υψηλότερες ετήσιες δόσεις έχουν καλύτερη δυνατότητα αποπληρωμής.

Συμπεράσματα

Η μεταβλητή AMT_ANNUITY παρουσιάζει μεγάλη διασπορά στις ετήσιες δόσεις των πελατών, με την πλειοψηφία να πληρώνει χαμηλότερες δόσεις. Οι κακοπληρωτές φαίνεται να πληρώνουν ελαφρώς χαμηλότερες δόσεις από τους καλοπληρωτές. Παρά τη χαμηλή συσχέτιση της ετήσιας δόσης με την πιθανότητα αθέτησης, η κατανομή δείχνει ότι οι πελάτες με υψηλότερες ετήσιες δόσεις έχουν λιγότερες πιθανότητες να αθετήσουν το δάνειο.

5.1.23 Ανάλυση της Μεταβλητής AMT_GOODS_PRICE

Η μεταβλητή AMT_GOODS_PRICE αναφέρεται στην τιμή των αγαθών που αγοράστηκαν μέσω του δανείου.



Εικόνα 5-27: Διαγράμματα κατανομής της AMT_GOODS_PRICE

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Τύπος δεδομένων**: float64 (Αριθμητικές τιμές
- **Αριθμός κενών τιμών**: Υπάρχουν 278 εγγραφές με κενές τιμές, οι οποίες αποτελούν πολύ μικρό ποσοστό του συνόλου (0.09%).

2. Ανάλυση Διαγραμμάτων

- Η κατανομή της AMT_GOODS_PRICE παρουσιάζει δεξιά ασυμμετρία. Η πλειοψηφία των τιμών συγκεντρώνεται σε χαμηλότερα επίπεδα, με λίγους πελάτες να έχουν αγοράσει αγαθά με πολύ υψηλές τιμές. Τα διαγράμματα αποκαλύπτουν ότι οι τιμές των αγαθών τείνουν να είναι σχετικά χαμηλές για τους περισσότερους πελάτες, με την κορυφή της κατανομής γύρω από την περιοχή των 450,000.
- Οι κακοπληρωτές εμφανίζουν τιμές για αγαθά ελαφρώς χαμηλότερες από τους καλοπληρωτές, κάτι που δείχνει ότι όσοι αγοράζουν πιο ακριβά αγαθά είναι πιθανότερο να αποπληρώσουν τα δάνεια τους. Οι καλοπληρωτές εμφανίζουν μεγαλύτερη διασπορά, με ακραίες τιμές σε ακριβότερα αγαθά.

3. Στατιστικά Περιγραφής

- **Μέσος όρος τιμής αγαθών**: 538,396.21.
- **Τυπική απόκλιση**: 369,446.46, που δείχνει μεγάλη διακύμανση στις τιμές αγαθών.
- **Ελάχιστη τιμή αγαθών**: 40,500
- **Διάμεσος τιμή**: 450,000 Η διάμεσος τιμή είναι πολύ κοντά στη συχνότερη τιμή, γεγονός που επιβεβαιώνει ότι μεγάλο ποσοστό των πελατών αγοράζει αγαθά σε αυτό το επίπεδο τιμών.
- **Μέγιστη τιμή αγαθών**: 4,050,000.0. Αυτή πιθανόν αποτελεί ακραία τιμή.

4. Συσχέτιση με την Πιθανότητα Αθέτησης

- Η συσχέτιση (-0.04) δείχνει ότι όσο αυξάνεται η τιμή των αγαθών, τόσο μικρότερη είναι η πιθανότητα χρεοκοπίας. Αυτό δείχνει ότι πελάτες που αγοράζουν ακριβότερα αγαθά έχουν μεγαλύτερη ικανότητα αποπληρωμής.

Συμπεράσματα

Η μεταβλητή AMT_GOODS_PRICE δείχνει ότι οι πελάτες που αγοράζουν ακριβότερα αγαθά είναι πιθανότερο να αποπληρώσουν το δάνειο τους παρά τη χαμηλή συσχέτιση της τιμής αγαθών με την πιθανότητα αθέτησης.

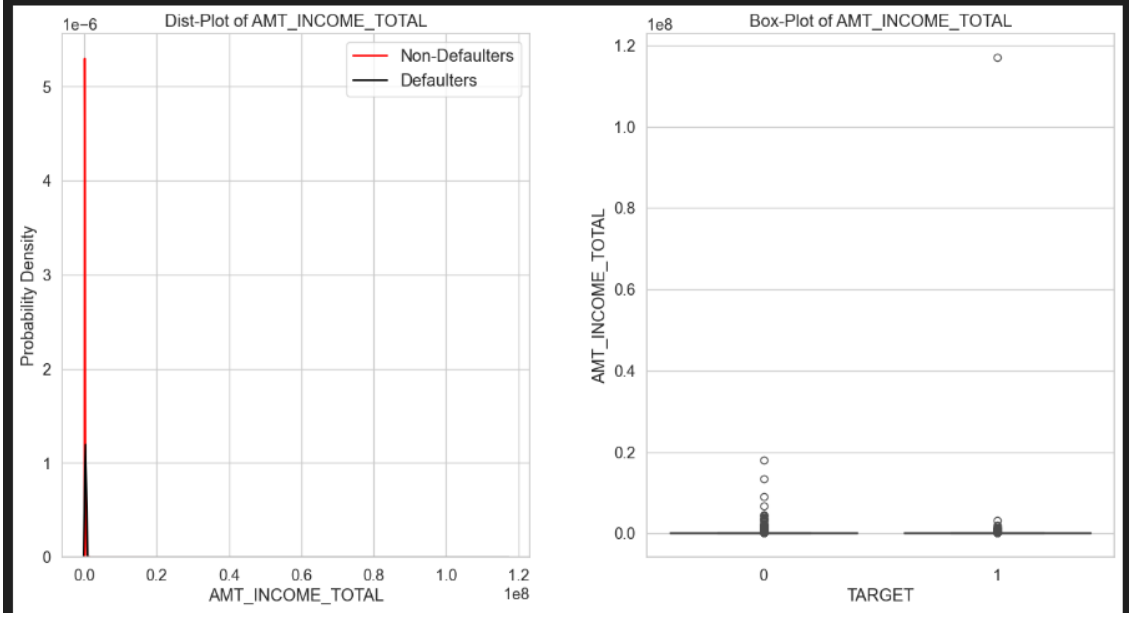
5.1.24 Ανάλυση της Μεταβλητής AMT_INCOME_TOTAL

Η μεταβλητή AMT_INCOME_TOTAL αναφέρεται στο συνολικό εισόδημα του πελάτη.

Από τα αρχικά διαγράμματα παρατηρούμε την ύπαρξη ακραίων τιμών ,οι οποίες δημιουργούν μια εικόνα που δυσκολεύει την εξαγωγή σαφών συμπερασμάτων σχετικά με την κατανομή των εισοδημάτων. Οι τιμές αυτές είναι τόσο μεγάλες που αλλοιώνουν την κατανόηση των κεντρικών τάσεων και της συνολικής διασποράς. Για να έχουμε μια πιο καθαρή εικόνα της κατανομής των εισοδημάτων, αποφασίσαμε να αφαιρέσουμε αυτές τις ακραίες τιμές, με threshold στο 1,000,000, επιτρέποντας καλύτερη οπτικοποίηση.

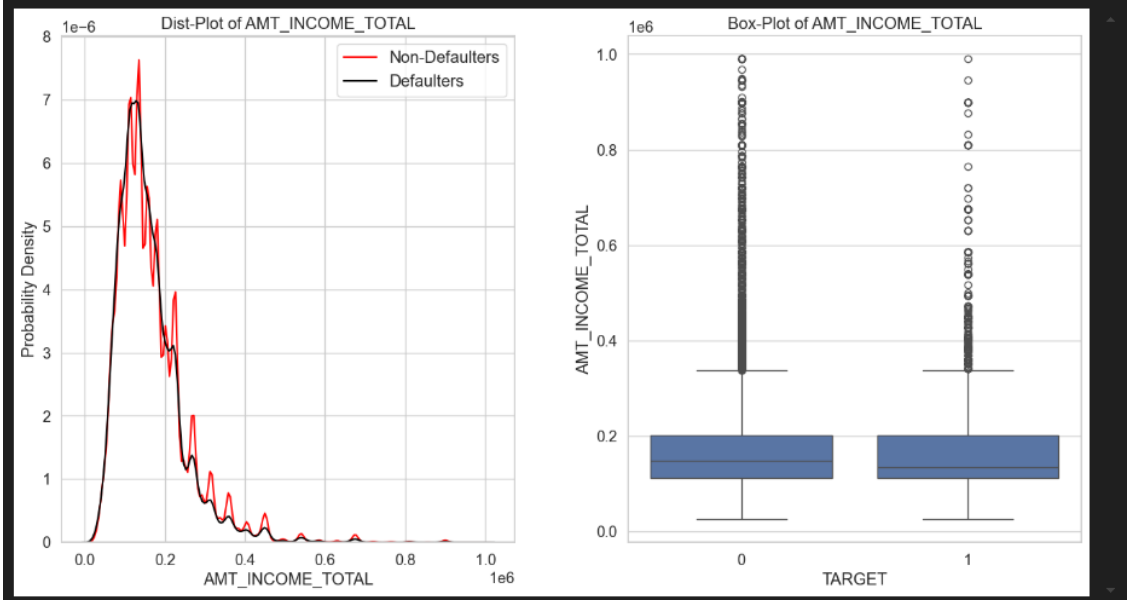
Μετά την αφαίρεση των ακραίων τιμών παρατηρούμε πιο καθαρή κατανομή των δεδομένων, δίνοντάς μας τη δυνατότητα να εξάγουμε πιο σαφή συμπεράσματα.

	dtype	count	unique	top_value_counts	missing_count	missing_percentage	
AMT_INCOME_TOTAL	float64	307511	2548	{135000.0: 35750, 112500.0: 31019, 157500.0: 2...	0	0.0	1687



Εικόνα 5-28: Διαγράμματα κατανομής της AMT_INCOME_TOTAL

counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
35750, 31019, 0.0: 2...	0	0.0	167180.9	89606.74	25650.0	144000.0	990000.0	-0.02



Εικόνα 5-29: Διαγράμματα κατανομής της AMT_INCOME_TOTAL2

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Τύπος δεδομένων**: float64 ,Αριθμητικές τιμές.
- **Πλήθος κενών τιμών**: Δεν υπάρχουν κενές τιμές.

2. Ανάλυση Διαγραμμάτων

- Μετά την αφαίρεση των ακραίων τιμών, η κατανομή των εισοδημάτων είναι περισσότερο ισορροπημένη και παρουσιάζει δεξιά ασυμμετρία με τα περισσότερα δεδομένα να συγκεντρώνονται σε χαμηλότερα επίπεδα εισοδήματος ενώ υπάρχει μικρότερος αριθμός πελατών με πολύ υψηλά εισοδήματα. Οι πελάτες που αποπληρώνουν τα δάνειά τους τείνουν να έχουν μεγαλύτερες τιμές εισοδήματος.

3. Συσχέτιση με την Χρεοκοπία

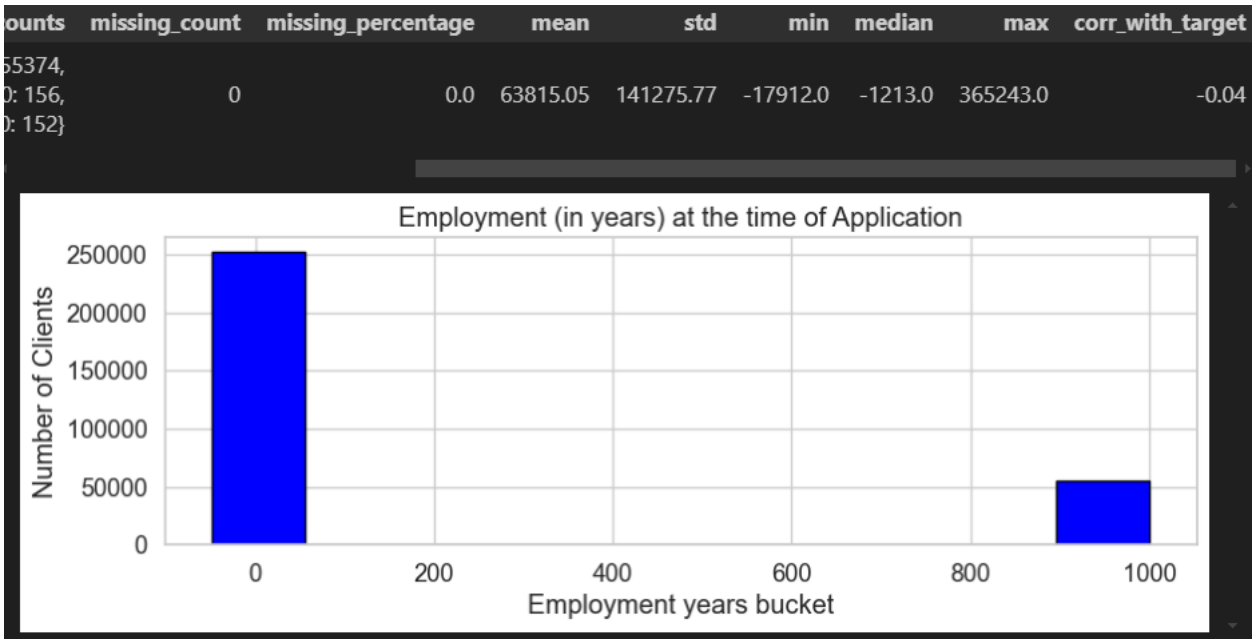
- Η μεταβλητή δεν εμφανίζει ισχυρή συσχέτιση με την πιθανότητα αθέτησης δανείου. Το ποσοστό συσχέτισης είναι κοντά στο -0.02, που δείχνει ότι το εισόδημα δεν είναι ιδιαίτερα προγνωστικός παράγοντας από μόνο του

Συμπεράσματα

Γενικά παρατηρούμε ο όγκος των πελατών έχει μικρότερα εισοδήματα και δεν μπορεί να γίνει κάποια διάκριση για τη δυνατότητα αποπληρωμής τους. Ωστόσο το μικρό ποσοστό πελατών με μεγάλα εισοδήματα φαίνεται να είναι πιο ικανό. Ο βαθμός συσχέτισης είναι χαμηλός ωστόσο, μπορεί να χρειάζεται συνδυασμός με άλλες μεταβλητές για να κατανοηθεί καλύτερα η σχέση του με τον στόχο.

5.1.25 Ανάλυση της Μεταβλητής DAYS_EMPLOYED

Η μεταβλητή DAYS_EMPLOYED αναφέρεται στις ημέρες απασχόλησης του πελάτη τη στιγμή της αίτησης δανείου. Τα δεδομένα για αυτή τη μεταβλητή μπορούν να μας προσφέρουν σημαντικές πληροφορίες σχετικά με την εργασιακή σταθερότητα και το επαγγελματικό ιστορικό των πελατών.



Εικόνα 5-30: Κατανομή της DAYS_EMPLOYED

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Τύπος δεδομένων**: int64
- **Πλήθος κενών τιμών**: Δεν υπάρχουν ελλείψεις.

2. Ανάλυση Διαγράμματος

- Από το διάγραμμα παρατηρούμε την ύπαρξη της τιμής των 365,243 ημερών αντιπροσωπεύει ένα εμφανές outlier που πιθανότατα υποδεικνύει είτε λάθος καταχώρηση είτε ειδική κωδικοποίηση για πελάτες που δεν έχουν εργαστεί. Αυτή η ακραία τιμή πρέπει να ληφθεί υπόψη κατά την ανάλυση, καθώς επηρεάζει σημαντικά τα στατιστικά δεδομένα.

3. Συσχέτιση με την Πιθανότητα Αθέτησης

- Ο βαθμός συσχέτισης είναι -0.04, που σημαίνει ότι υπάρχει αρνητική αλλά πολύ ασθενής συσχέτιση μεταξύ της διάρκειας απασχόλησης και της πιθανότητας χρεοκοπίας. Αυτό σημαίνει ότι όσο περισσότερο διάστημα απασχολείται ένας πελάτης, τόσο ελαφρώς μειώνεται η πιθανότητα να αθετήσει το δάνειο, αλλά αυτή η συσχέτιση δεν είναι ιδιαίτερα ισχυρή.

Συμπεράσματα

Η μεταβλητή DAYS_EMPLOYED παρέχει πολύτιμες πληροφορίες για την απασχόληση των πελατών, αλλά τα outliers όπως οι 365,243 ημέρες πρέπει να αντιμετωπιστούν προσεκτικά για την εξαγωγή ακριβέστερων συμπερασμάτων. Η συσχέτιση της διάρκειας απασχόλησης με την πιθανότητα χρεοκοπίας είναι πολύ ασθενής, αλλά ενδέχεται να έχει σημασία σε συνδυασμό με άλλες μεταβλητές.

5.2 BUREAU EDA

Ο πίνακας bureau περιέχει πληροφορίες για δάνεια των πελατών σε άλλα χρηματοπιστωτικά ιδρύματα. Συγχωνεύουμε το bureau με το application_train για να εξετάσουμε πώς οι εξωτερικές δανειακές υποχρεώσεις επηρεάζουν τον κίνδυνο χρεοκοπίας των πελατών. Αυτή η διαδικασία μας επιτρέπει να ενσωματώσουμε δεδομένα από εξωτερικές πηγές για μια πιο ολοκληρωμένη ανάλυση.

5.2.1 Ανάλυση της Μεταβλητής CREDIT_ACTIVE

Η μεταβλητή CREDIT_ACTIVE στον πίνακα bureau.csv περιγράφει την κατάσταση των δανείων των πελατών σε άλλα χρηματοπιστωτικά ιδρύματα.

Από την ανάλυση προκύπτουν τα εξής:

1. Τύπος Δεδομένων και Πλήθος Τιμών

- **Είδος δεδομένων:** object, δηλαδή κατηγορικά δεδομένα
- **Αριθμός κενών τιμών:** Δεν υπάρχουν απουσίες σε αυτή τη στήλη.
- **Μοναδικές τιμές:** Υπάρχουν τέσσερις κατηγορίες “Closed”,”Active”, “Sold”,”Bad debt”

2.Κατανομή της CREDIT_ACTIVE

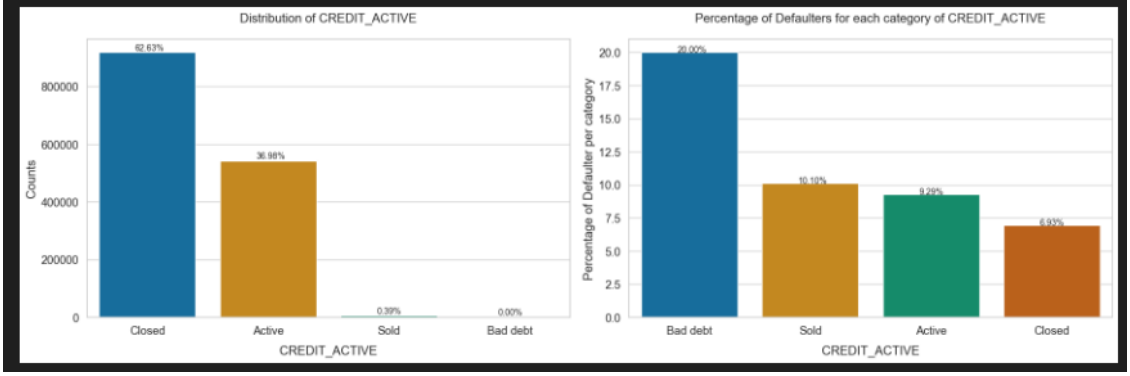
- Closed: Το 62.61% των δανείων είναι κλειστά.
- Active: Το 26.98% των δανείων είναι ενεργά.
- Sold: Το 0.39% των δανείων έχουν πωληθεί.
- Bad debt: Μόνο 20 δάνεια (0.00%) καταγράφονται ως "κακή οφειλή".

3.Πιθανότητα Αθέτησης ανά Κατηγορία

- Η κατηγορία με το μεγαλύτερο ποσοστό χρεοκοπίας είναι η Bad debt με ποσοστό 20.00%.
- Η κατηγορία Sold έχει ποσοστό χρεοκοπίας 10.10%.
- Η κατηγορία Active έχει ποσοστό χρεοκοπίας 9.29%.
- Η κατηγορία Closed έχει ποσοστό χρεοκοπίας 6.93%.

	dtype	count	unique	missing_count	missing_percentage
CREDIT_ACTIVE	object	1465325	4	44020	2.92

```
The unique categories of 'CREDIT_ACTIVE' are:
['Closed' 'Active' nan 'Sold' 'Bad debt']
```



Εικόνα 5-31: Κατανομή της CREDIT_ACTIVE

Συμπεράσματα

Η πλειοψηφία των εγγραφών έχει καθεστώς "Closed" (κλειστό), το οποίο σημαίνει ότι οι πελάτες έχουν αποπληρώσει τις πιστώσεις τους και δεν έχουν ενεργές υποχρεώσεις. Αυτό μπορεί να δείχνει ότι μεγάλο μέρος των πελατών έχει διατηρήσει υπεύθυνη συμπεριφορά δανεισμού στο παρελθόν. Ένα σημαντικό ποσοστό των πελατών (περίπου 27%) έχει ενεργά δάνεια, γεγονός που σημαίνει ότι εξακολουθούν να έχουν ανοιχτές χρηματοοικονομικές υποχρεώσεις προς το δανειστικό ίδρυμα. Οι πελάτες αυτοί είναι πιθανό να συνεχίσουν να αποτελούν αντικείμενο αξιολόγησης για την πιθανότητα αθέτησης.

5.2.2 Ανάλυση της Μεταβλητής CREDIT_TYPE

Η μεταβλητή CREDIT_TYPE περιγράφει τον τύπο πίστωσης ή δανείου που έχει λάβει ο πελάτης στον πίνακα bureau.

1.Τύπος Δεδομένων και Πλήθος Τιμών

- **Τύπος δεδομένων:** object, δηλαδή κατηγορικές τιμές.
- **Πλήθος κενών τιμών:** Δεν υπάρχουν σε αυτή τη στήλη.
- **Μοναδικές τιμές:** Υπάρχουν 15 μοναδικές κατηγορίες τύπων πίστωσης.

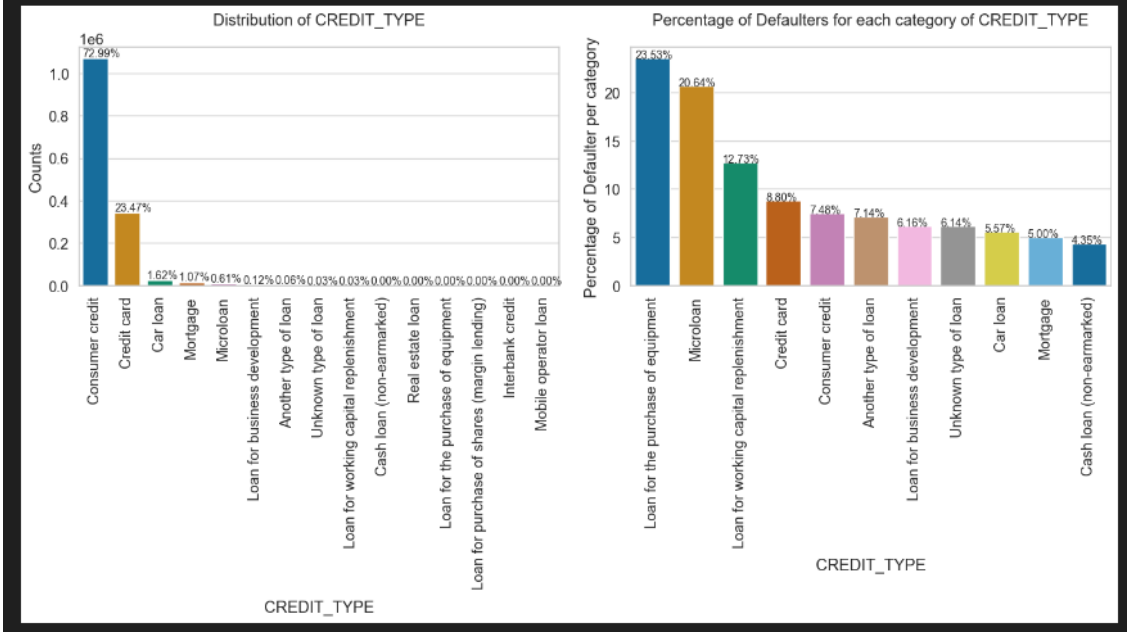
2. Κατανομή των Κατηγοριών

- Consumer credit: Το 72.99% των δανείων είναι τύπου καταναλωτικής πίστωσης, υποδεικνύοντας ότι η πλειονότητα των πελατών λαμβάνει δάνεια αυτού του τύπου.
- Credit card: Το 23.47% των δανείων είναι πιστωτικές κάρτες, η δεύτερη πιο συνηθισμένη κατηγορία.
- Car loan: Το 1.62% των δανείων αφορούν δάνεια για αγορά αυτοκινήτου.

- Mortgage: Το 1.07% είναι δάνεια για αγορά κατοικίας.
- Microloan: Το 0.61% είναι μικροδάνεια, τα οποία ίσως προορίζονται για άμεση και μικρή χρηματοδότηση.
- Οι υπόλοιπες κατηγορίες, όπως Loan for business development, Another type of loan, και Loan for working capital replenishment, αποτελούν μικρό ποσοστό της συνολικής κατανομής, με ποσοστά κάτω του 1%.

	dtype	count	unique	missing_count	missing_percentage
CREDIT_TYPE	object	1465325	15	44020	2.92

```
The unique categories of 'CREDIT_TYPE' are:
['Credit card' 'Consumer credit' nan 'Loan for business development'
 'Microloan' 'Mortgage' 'Car loan'
 'Loan for working capital replenishment' 'Another type of loan'
 'Unknown type of loan' 'Loan for the purchase of equipment'
 'Cash loan (non-earmarked)' 'Real estate loan' 'Interbank credit'
 'Loan for purchase of shares (margin lending)' 'Mobile operator loan']
```



Εικόνα 5-32: Κατανομή της CREDIT_TYPE

Συμπεράσματα

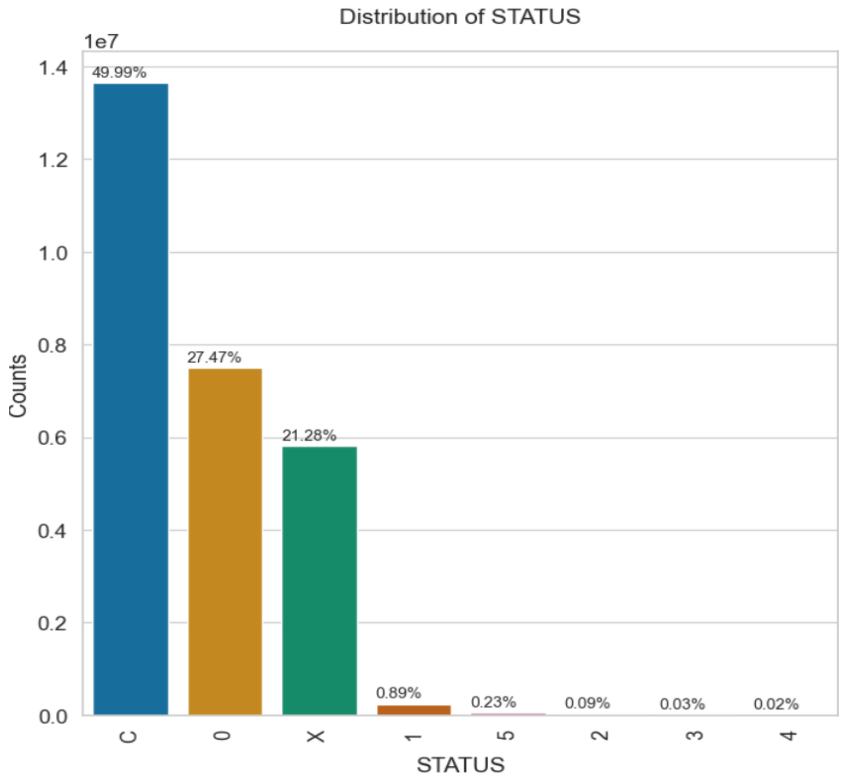
Η συντριπτική πλειονότητα των δανείων αφορά καταναλωτικά δάνεια και πιστωτικές κάρτες, γεγονός που αντανακλά την ανάγκη των πελατών για καταναλωτική πίστωση και διαχείριση του χρέους μέσω πιστωτικών καρτών ενώ ειδικοί τύποι δανείων, όπως δάνεια για επιχειρηματική ανάπτυξη ή αγορά κατοικίας, είναι πολύ λιγότερο συχνοί, γεγονός που ενδεχομένως δείχνει ότι αυτά τα δάνεια λαμβάνονται για συγκεκριμένους σκοπούς και από μικρότερες ομάδες πελατών.

5.3 BUREAU BALANCE EDA

Το bureau_balance είναι ένας πίνακας που περιέχει μηνιαίες πληροφορίες για το υπόλοιπο των πιστωτικών λογαριασμών που έχουν αναφερθεί σε εξωτερικά γραφεία πιστοληπτικής ικανότητας. Κάθε γραμμή αυτού του πίνακα αντιστοιχεί σε μια συγκεκριμένη χρονική περίοδο για έναν πελάτη.

5.3.1 Ανάλυση της Μεταβλητής STATUS

Η μεταβλητή STATUS από τον πίνακα bureau_balance αναφέρεται στην κατάσταση του πιστωτικού λογαριασμού ενός πελάτη για ένα συγκεκριμένο μήνα. Οι καταστάσεις αυτές παρέχουν μια ένδειξη για την κανονικότητα των πληρωμών και τον κίνδυνο που αντιπροσωπεύει ο κάθε πελάτης.



Εικόνα 5-33: Κατανομή της STATUS

1. **Μοναδικές Τιμές**

Η στήλη STATUS περιέχει τις εξής κατηγορίες:

- C: Ο λογαριασμός είναι κλειστός.
- 0: Καμία καθυστέρηση στις πληρωμές.
- X: Δεν υπάρχουν διαθέσιμες πληροφορίες για το συγκεκριμένο μήνα.
- 1, 2, 3, 4, 5: Καθυστέρηση πληρωμών κατά 1, 2, 3, 4 ή 5 μήνες αντίστοιχα.

2. **Κατανομή Τιμών**

- C (κλειστός λογαριασμός): 49.99% των περιπτώσεων, δηλαδή σχεδόν το ήμισυ των εγγραφών, αντιστοιχούν σε κλειστούς λογαριασμούς.
- 0 (καμία καθυστέρηση): 27.47% των περιπτώσεων αντιστοιχούν σε μήνες χωρίς καθυστέρηση πληρωμών.
- X (μη διαθέσιμες πληροφορίες): 21.28% των εγγραφών δεν περιέχουν πληροφορίες για τον λογαριασμό τον συγκεκριμένο μήνα.

- Καθυστέρηση πληρωμών (1 έως 5 μήνες): Μόνο ένα μικρό ποσοστό εγγραφών (0.89% για καθυστέρηση 1 μήνα, 0.23% για 2 μήνες, και ακόμα μικρότερα ποσοστά για 3, 4, και 5 μήνες) αντιστοιχεί σε καθυστερήσεις στις πληρωμές.

Αυτή η κατανομή δείχνει ότι οι περισσότερες εγγραφές αφορούν λογαριασμούς που είτε είναι κλειστοί είτε δεν έχουν καμία καθυστέρηση πληρωμών ενώ τα ποσοστά καθυστέρησης είναι σχετικά χαμηλά.

Συμπεράσματα

Το γεγονός ότι ένα μεγάλο ποσοστό των λογαριασμών είναι κλειστοί ή δεν έχουν καθυστέρηση δείχνει γενικά καλή συμπεριφορά πληρωμών των πελατών και η χαμηλή παρουσία καθυστερήσεων δείχνει ότι οι περιπτώσεις πελατών με προβλήματα στις πληρωμές είναι περιορισμένες, αλλά παραμένουν σημαντικές για την ανάλυση του κινδύνου.

5.4 PREVIOUS APPLICATION EDA

Ο πίνακας previous_application περιέχει στοιχεία για παλαιότερες αιτήσεις δανείων των πελατών, όπως την κατάσταση έγκρισης ή απόρριψης του δανείου, τον τύπο του δανείου, και άλλες συναφείς πληροφορίες.

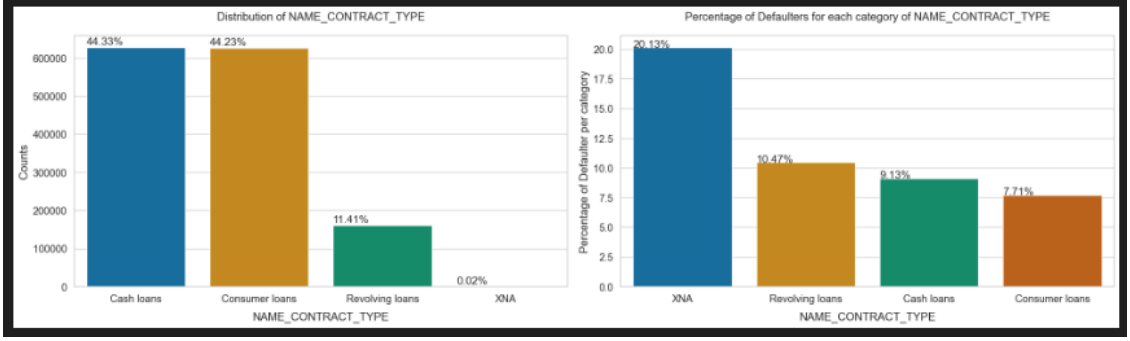
Αρχικά, προχωράμε στη συγχώνευση του πίνακα application_train με τον πίνακα previous_application ώστε να συνδέσουμε τη συμπεριφορά των πελατών στο παρελθόν με την τρέχουσα κατάστασή τους και την πιθανότητα αθέτησης πληρωμών

5.4.1Ανάλυση της Μεταβλητής NAME_CONTRACT_TYPE

Η μεταβλητή NAME_CONTRACT_TYPE περιγράφει τον τύπο της σύμβασης δανείου για προηγούμενες αιτήσεις.

	dtype	count	unique	missing_count	missing_percentage
NAME_CONTRACT_TYPE	object	1413701	4	16454	1.15

```
The unique categories of 'NAME_CONTRACT_TYPE' are:
['Consumer loans' 'Cash loans' 'Revolving loans' nan 'XNA']
```



Εικόνα 5-34: Κατανομή της NAME_CONTRACT_TYPE

1.Τύπος και Πλήθος Τιμών

- **Τύπος δεδομένων:** object (Κατηγορικές τιμές).
- **Αριθμός κενών τιμων:** 16.454 (1.15%).

- **Μοναδικές τιμές:** Υπάρχουν 4 κατηγορίες δανείων που έχουν καταγραφεί στις προηγούμενες αιτήσεις Consumer loans (Καταναλωτικά δάνεια) Cash loans (Δάνεια σε μετρητά) Revolving loans (Ανακυκλούμενα δάνεια) XNA (Άγνωστο/Μη καθορισμένο).

2. Κατανομή της Μεταβλητής

- Cash loans: 44.33% του συνόλου.
- Consumer loans: 44.23% του συνόλου.
- Revolving loans: 11.41% του συνόλου.
- XNA: 0.02% του συνόλου (αγνώστου τύπου δάνειο).

Αυτές οι κατηγορίες αντικατοπτρίζουν τις κύριες μορφές δανείων που προσφέρονται στους πελάτες.

3. Ποσοστά Αθέτησης Ανά Κατηγορία

- Cash loans: Το 9.13% των πελατών που έλαβαν αυτόν τον τύπο δανείου παρουσίασαν αθέτηση πληρωμών.
- Consumer loans: Το 7.71% των πελατών παρουσίασαν αθέτηση.
- Revolving loans: Το 10.47% των πελατών αθέτησαν τις πληρωμές τους.
- XNA: Παρά το χαμηλό ποσοστό εμφάνισης, το 20.13% των πελατών σε αυτή την κατηγορία αθέτησαν τις πληρωμές τους.

Συμπεράσματα

Η κατηγορία Revolving loans, παρόλο που είναι μικρότερη σε ποσοστό συγκριτικά με τα άλλα είδη, έχει σχετικά υψηλό ποσοστό αθέτησης. Τα Consumer loans παρουσιάζουν το χαμηλότερο ποσοστό αθέτησης, κάτι που ίσως σχετίζεται με τη φύση αυτών των δανείων, που μπορεί να είναι πιο διαχειρίσιμα.Το XNA, παρόλο που αποτελεί πολύ μικρό ποσοστό του συνόλου των δανείων, παρουσιάζει ένα εξαιρετικά υψηλό ποσοστό αθέτησης (20.13%), κάτι που μπορεί να υποδεικνύει κάποια ιδιαίτερη κατηγορία δανείων ή πελατών που δυσκολεύονται να αποπληρώσουν τα δάνεια αυτά.

5.4.2 Ανάλυση της Μεταβλητής NAME_CONTRACT_STATUS

Η μεταβλητή NAME_CONTRACT_STATUS περιγράφει την κατάσταση των προηγούμενων αιτήσεων δανείων.

1.Τύπος και Πλήθος Τιμών

- **Τύπος δεδομένων:** object (Κατηγορικές τιμές).
- **Αριθμός κενών τιμών**: 16.454 (1.15%).
- **Μοναδικές Τιμές:** Υπάρχουν 4 κύριες κατηγορίες για την κατάσταση των αιτήσεων: *Approved* (Εγκριμένες) ,*Canceled* (Ακυρωμένες) *Refused* (Απορριφθείσες) *Unused offer* (Μη χρησιμοποιηθείσα προσφορά)

2. Κατανομή της μεταβλητής

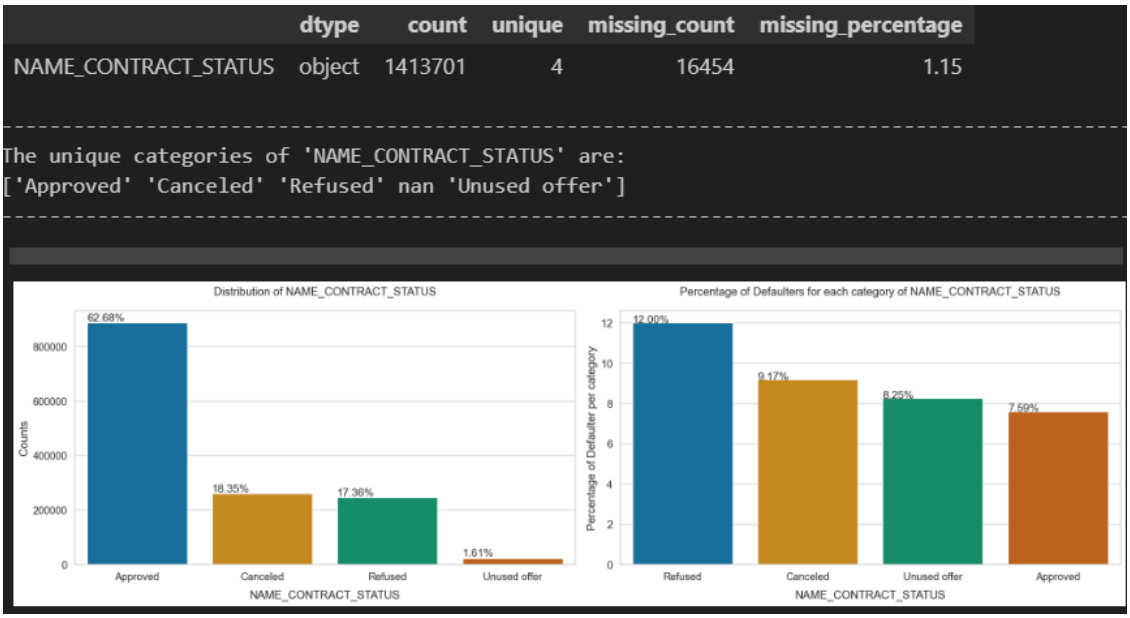
Οι διαθέσιμες κατηγορίες για αυτή τη μεταβλητή είναι οι εξής:

- Approved (Εγκρίθηκε): Αντιπροσωπεύουν το 62.68% των προηγούμενων αιτήσεων, γεγονός που υποδηλώνει ότι η πλειονότητα των αιτήσεων εγκρίθηκε.
- Canceled (Ακυρώθηκε):): Το 18.35% των αιτήσεων ακυρώθηκε από τον αιτούντα ή την τράπεζα.
- Refused (Απορρίφθηκε): Το 17.36% των αιτήσεων απορρίφθηκε, υποδεικνύοντας πιθανά προβλήματα πιστοληπτικής αξιοπιστίας.
- Unused offer (Αχρησιμοποίητη προσφορά): Μόνο το 1.61% των αιτήσεων δεν αξιοποιήθηκε τελικά από τον πελάτη.

3. Ποσοστά Αθέτησης Ανά Κατηγορία

Κάθε κατηγορία σύμβασης έχει διαφορετικό ποσοστό αθέτησης (default rate):

- Approved: Το 7.59% των εγκεκριμένων δανείων κατέληξαν σε αθέτηση πληρωμών.
- Canceled: Το 9.17% των ακυρωμένων συμβάσεων παρουσίασε αθέτηση, υποδεικνύοντας ότι οι πελάτες που ακυρώνουν δάνεια.
- Refused: Το 12.00% των απορριφθέντων δανείων κατέληξαν σε αθέτηση, δείχνοντας ότι η τράπεζα ήταν πιθανώς σωστή να απορρίψει την αίτηση.
- Unused offer: Το 8.25% των προσφορών που δεν χρησιμοποιήθηκαν οδήγησαν σε αθέτηση.



Εικόνα 5-35: Κατανομή της NAME_CONTRACT_STATUS

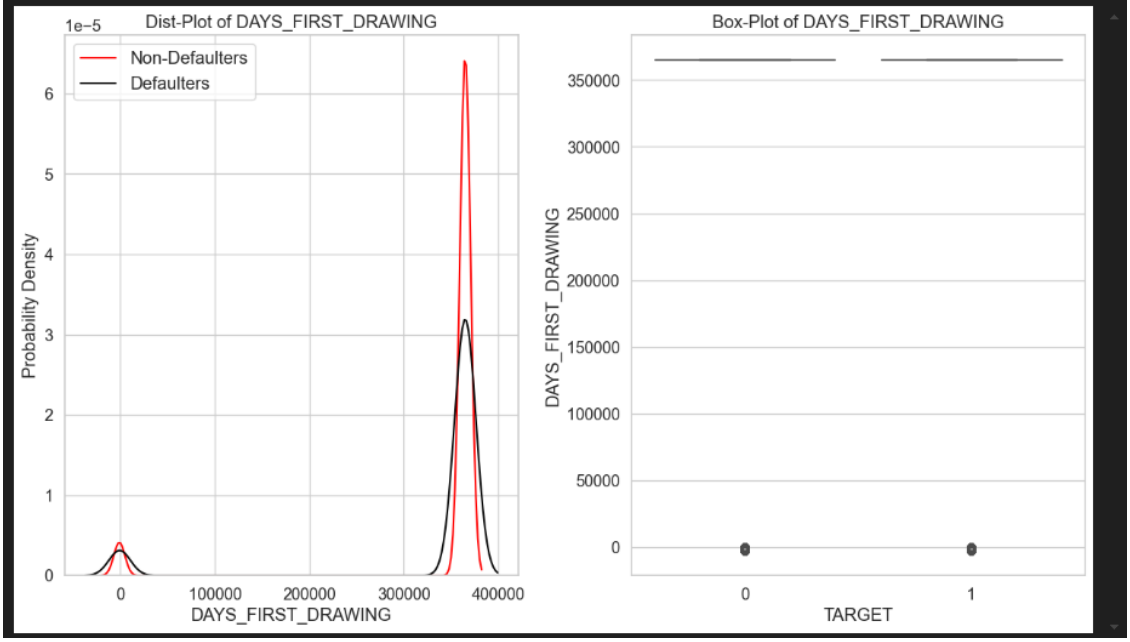
Συμπεράσματα

Οι εγκεκριμένες αιτήσεις έχουν ένα σχετικά χαμηλό ποσοστό αθέτησης (7.59%), το οποίο μπορεί να θεωρηθεί φυσιολογικό ενώ τα ακυρωμένα δάνεια και οι απορριφθείσες αιτήσεις παρουσιάζουν υψηλότερα ποσοστά αθέτησης, με τις απορριφθείσες αιτήσεις να έχουν το μεγαλύτερο ποσοστό αθέτησης (12.00%), υποδεικνύοντας ότι οι αιτούντες αυτοί είναι και οι πιο επικίνδυνοι. Οι αχρησιμοποίητες προσφορές έχουν επίσης ένα σημαντικό ποσοστό αθέτησης, υποδεικνύοντας πιθανούς πελάτες που αρχικά δεν χρειάστηκαν το δάνειο αλλά αργότερα βρέθηκαν σε οικονομική δυσκολία.

5.4.3 Ανάλυση της Μεταβλητής DAYS_FIRST_DRAWING

Η μεταβλητή DAYS_FIRST_DRAWING περιγράφει τον αριθμό ημερών πριν από την αίτηση που ο πελάτης έκανε την πρώτη χρήση της πίστωσης.

ounts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
243.0: 228.0: : 113}	577560	40.38	342257.66	88832.27	-2922.0	365243.0	365243.0	-0.03



Εικόνα 5-36: Διαγράμματα κατανομής της DAYS_FIRST_DRAWING

1. Τύπος και Πλήθος Τιμών

- **Τύπος δεδομένων:** float64 (Αριθμητικές τιμές).
- **Αριθμός κενών τιμών:** 557.560 (40.38%).

2. Παρατηρήσεις για τις Ακραίες Τιμές

Παρατηρούμε αρκετά ακραίες τιμές, ειδικά η μέγιστη τιμή των 365,243 ημερών (1,000 χρόνια πριν την αίτηση), που είναι σαφώς μη ρεαλιστική. Αυτές οι τιμές μπορούν να επηρεάσουν την ανάλυση και ενδέχεται να χρειάζονται προσεκτικό χειρισμό, όπως η αφαίρεση ή αντικατάστασή τους.

5.4.4 Ανάλυση της Μεταβλητής DAYS_FIRST_DUE

Η μεταβλητή DAYS_FIRST_DUE περιγράφει τον αριθμό ημερών πριν την αίτηση που ο πελάτης είχε την πρώτη προθεσμία πληρωμής για τη δανειακή του υποχρέωση.

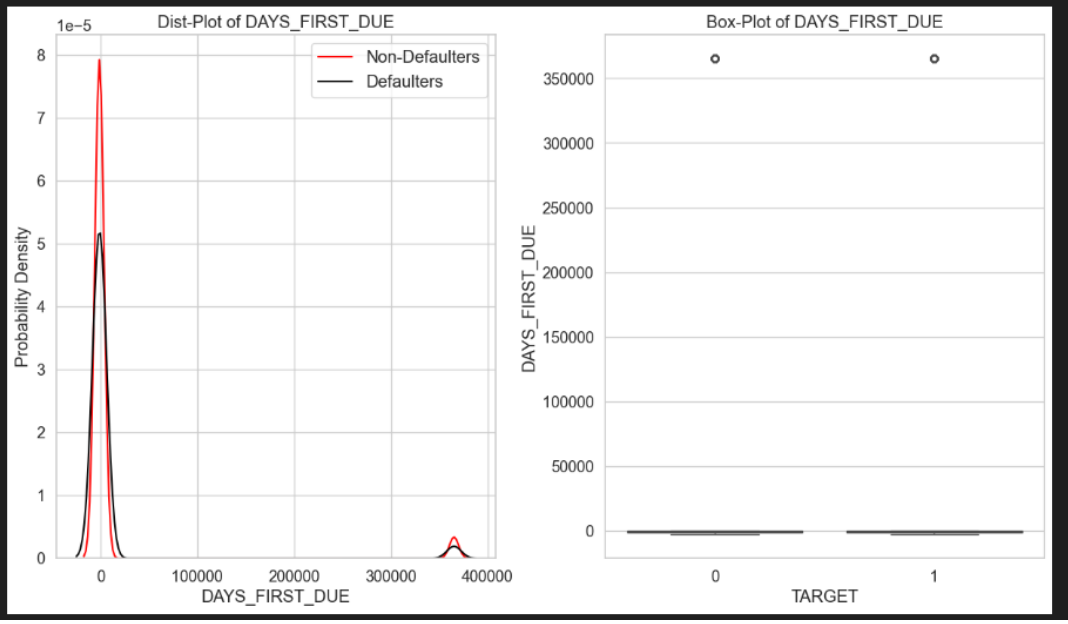
1.Τύπος και Πλήθος Τιμών

- **Τύπος δεδομένων:** float64 (Αριθμητικές τιμές).
- **Αριθμός κενών τιμών:** 577.560 (40.38%).

2.Παρατηρήσεις για τις Ακραίες Τιμές:

- Όπως και στη μεταβλητή DAYS_FIRST_DRAWING, εδώ παρατηρούμε ακραίες τιμές, με τη μέγιστη τιμή των 365,243 ημερών να αποτελεί ακραία τιμή. Αυτές οι τιμές επηρεάζουν σημαντικά την κατανομή και την ακρίβεια των στατιστικών δεδομένων.

e_counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
.0: 33962, 34.0: 661, 99.0: 653}	577560	40.38	13488.74	71650.15	-2892.0	-825.0	365243.0	-0.01



Εικόνα 5-37: Διαγράμματα κατανομής της DAYS_FIRST_DUE

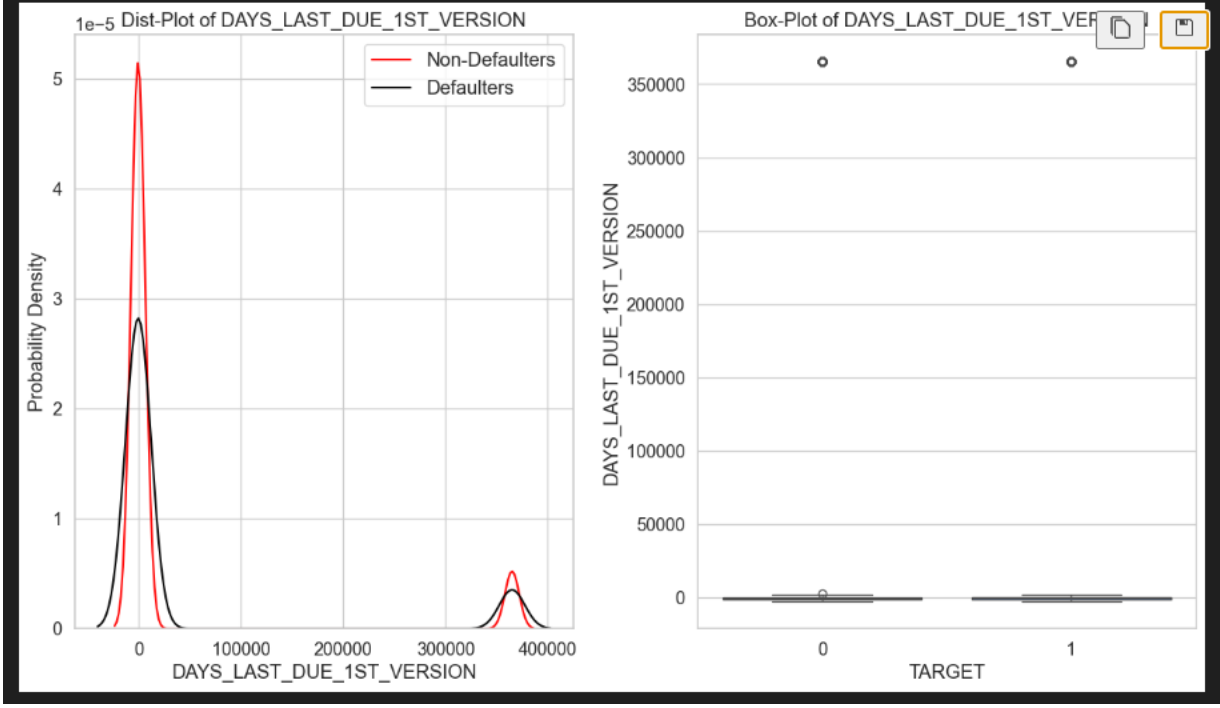
Συμπεράσματα

Συνοψίζοντας, η μεταβλητή DAYS_FIRST_DUE παρουσιάζει παρόμοια χαρακτηριστικά με άλλες μεταβλητές ημερών, με την ύπαρξη ακραίων τιμών να επηρεάζει την ακρίβεια των στατιστικών. Αυτές οι ακραίες τιμές απαιτούν διαχείριση για μια πιο σαφή εικόνα της κατανομής.

5.4.5Ανάλυση της Μεταβλητής DAYS_LAST_DUE_1ST_VERSION

Η μεταβλητή DAYS_LAST_DUE_1ST_VERSION αναφέρεται στον αριθμό ημερών πριν ή μετά την αίτηση, όταν έπρεπε να γίνει η τελευταία πληρωμή δανείου σύμφωνα με την πρώτη έκδοση των όρων του δανείου.

counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
79099, 1, 10.0: 601}	577560	40.38	33274.83	106161.78	-2801.0	-358.0	365243.0	0.02



Εικόνα 5-38: Διαγράμματα κατανομής της DAYS_LAST_DUE_1ST_VERSION

1.Τύπος και Πλήθος Τιμών

- **Τύπος δεδομένων:** float64 (Αριθμητικές τιμές).
- **Αριθμός κενών τιμών:** 577.56040.38%.

2. Παρατηρήσεις για τις Ακραίες Τιμές

•Η μέγιστη τιμή των 365,243 ημερών είναι προφανώς μια ανωμαλία. Αυτές οι τιμές επηρεάζουν σημαντικά την κατανομή και την ακρίβεια της ανάλυσης.

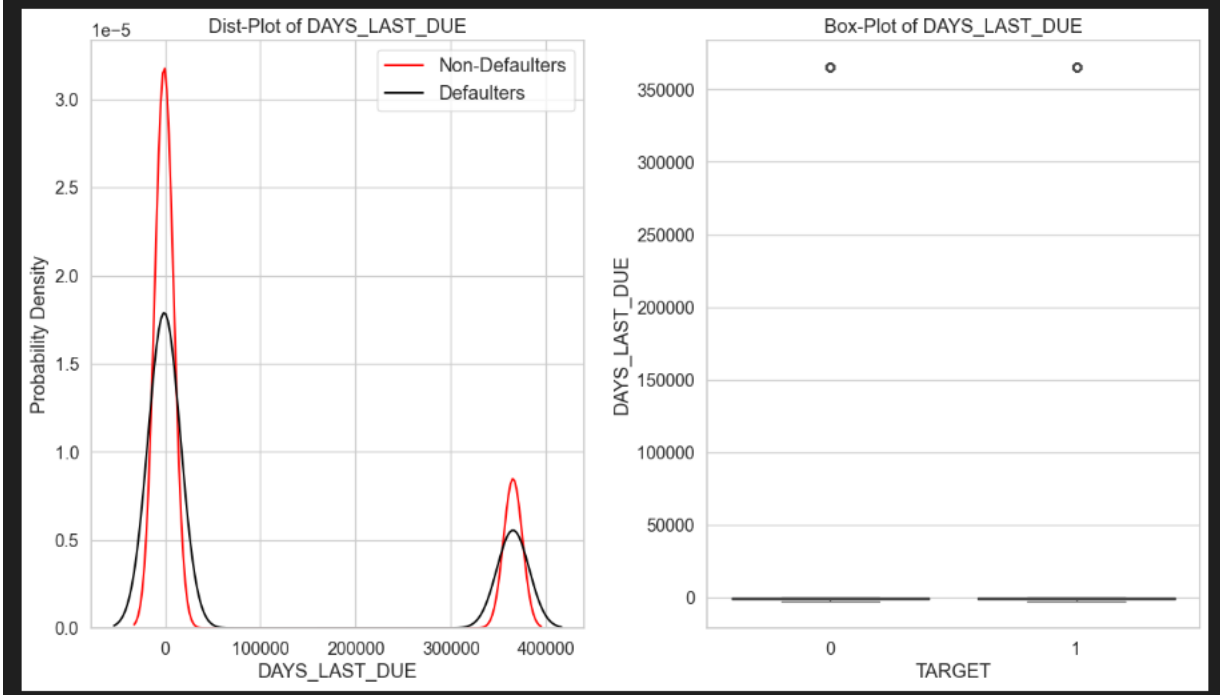
Συμπεράσματα

Το διάγραμμα κατανομής παρουσιάζει μια μεγάλη συγκέντρωση γύρω από χαμηλές τιμές, ενώ το box plot επιβεβαιώνει την παρουσία ακραίων τιμών. Η κατανομή των τιμών δείχνει ότι οι περισσότερες πληρωμές έχουν λογικά χρονικά πλαίσια, αλλά οι ακραίες τιμές επηρεάζουν την συνολική εικόνα και θα πρέπει κάπως να τις διαχειριστούμε.

5.4.6 Ανάλυση της Μεταβλητής DAYS_LAST_DUE

Η μεταβλητή DAYS_LAST_DUE αναφέρεται στον αριθμό ημερών πριν ή μετά την αίτηση, όταν έπρεπε να γίνει η τελευταία πληρωμή του προηγούμενου δανείου.

counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
55243.0: -239.0: 3.0: 564}	577560	40.38	76665.63	149704.72	-2889.0	-534.0	365243.0	0.02



Εικόνα 5-39: Διαγράμματα κατανομής της DAYS_LAST_DUE

1. **Τύπος και Πλήθος Τιμών**
   - **Τύπος δεδομένων:** float64 (Αριθμητικές τιμές).
   - **Αριθμός κενών τιμών:** 577.560 (40.38%).

2. **Παρατηρήσεις για Outliers**
   - Οι μέγιστες τιμές γύρω από 365,243 ημέρες είναι εμφανώς outliers. Αυτές οι ακραίες τιμές πρέπει να ληφθούν υπόψη κατά την ανάλυση, καθώς επηρεάζουν σημαντικά την κατανομή και την ακρίβεια των αποτελεσμάτων.

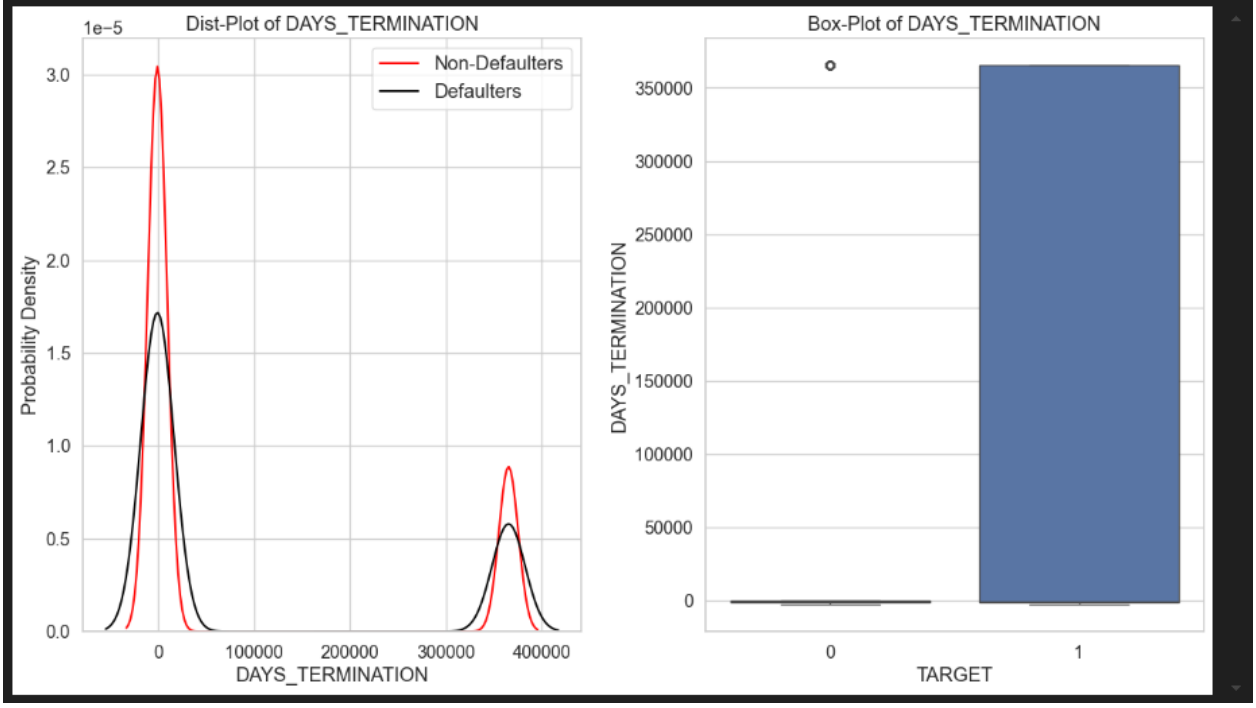
Συμπεράσματα

Το διάγραμμα κατανομής δείχνει σημαντική συγκέντρωση γύρω από χαμηλές τιμές, με μεγάλο αριθμό τιμών κοντά στις μηδενικές ημέρες. Το box plot επιβεβαιώνει την ύπαρξη ακραίων τιμών που επηρεάζουν τη συνολική εικόνα. Η μεταβλητή DAYS_LAST_DUE εμφανίζει παρόμοια μοτίβα με άλλες μεταβλητές ημερών, και οι ακρές τιμές (όπως οι τιμές 365,243 ημερών) επηρεάζουν σημαντικά τα στατιστικά αποτελέσματα και θα πρέπει να εξεταστούν για πιθανή απομάκρυνση.

5.4.7 Ανάλυση της Μεταβλητής DAYS_TERMINATION

Η μεταβλητή DAYS_TERMINATION αναφέρεται στον αριθμό ημερών πριν ή μετά την αίτηση, από όταν τερματίστηκε το προηγούμενο δάνειο.

counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
5243.0: -156.0: 0: 652}	577560	40.38	82353.17	153537.06	-2874.0	-494.0	365243.0	0.02



Εικόνα 5-40: Διαγράμματα κατανομής της DAYS_TERMINATION

1.Τύπος και Πλήθος Τιμών

- **Τύπος δεδομένων:** float64 (Αριθμητικές τιμές).
- **Αριθμός κενών τιμών:** 577.56040.38%.

2.Παρατηρήσεις για Outliers

- Οι μέγιστες τιμές των 365,243 ημερών, είναι προφανή outliers και επηρεάζουν την κατανομή της μεταβλητής, δημιουργώντας μια παραμορφωμένη εικόνα.

Συμπεράσματα

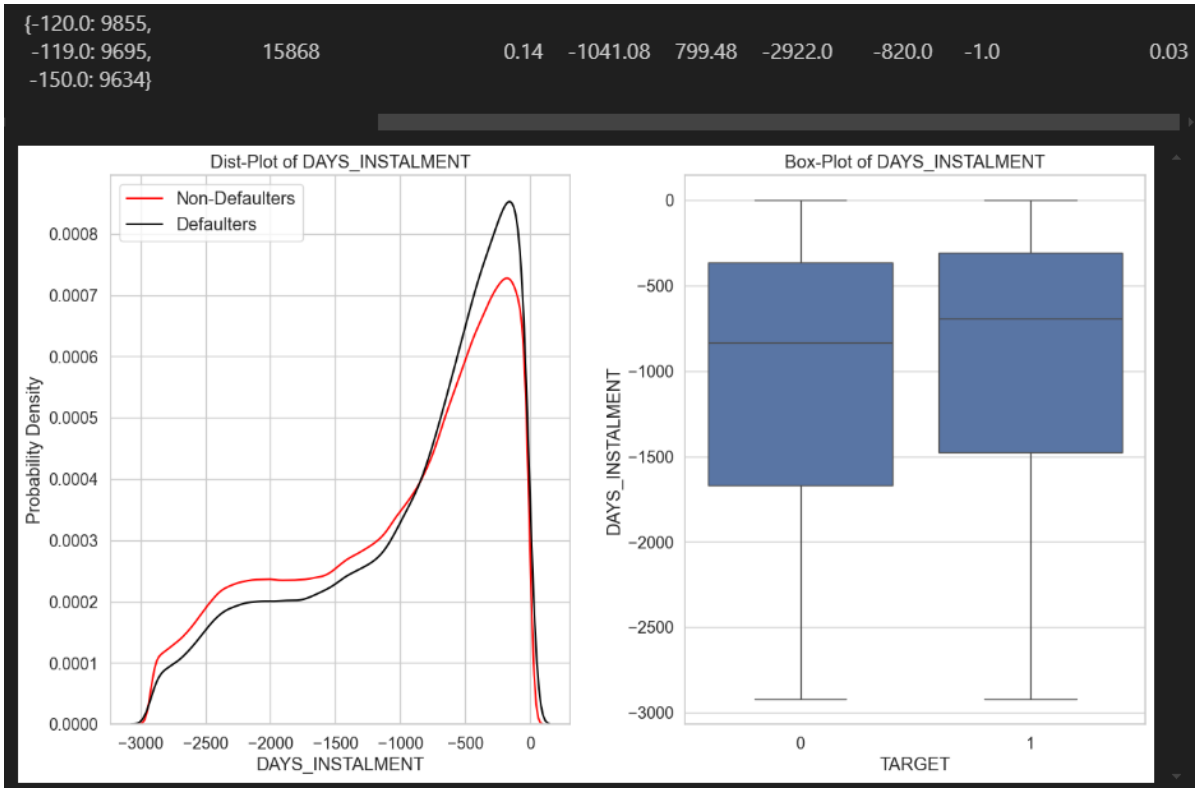
Η μεταβλητή DAYS_TERMINATION επηρεάζεται από outliers, όπως οι τιμές των 365,243 ημερών, τα οποία δημιουργούν μια παραμορφωμένη κατανομή και διασπορά. Αυτές οι ακραίες τιμές μπορούν θα πρέπει να δούμε πως θα τις διαχειριστούμε ώστε να εκπαιδεύσουμε καλύτερα τα μοντέλα μας.

5.5 INSTALLMENTS PAYMENTS EDA

Ο πίνακας installments_payments, περιλαμβάνει πληροφορίες σχετικά με τις πληρωμές που έγιναν για τα δάνεια. Η συνένωση του με τον πίνακα application_train έχει ως στόχο τη διερεύνηση της σχέσης των πληρωμών δόσεων με την πιθανότητα χρεοκοπίας. Με αυτή τη συνένωση, θα μπορέσουμε να εξετάσουμε την επιρροή του ιστορικού πληρωμών στις πιθανότητες αποπληρωμής του δανείου.

5.5.1 Ανάλυση της Μεταβλητής DAYS_INSTALMENT

Η μεταβλητή DAYS_INSTALMENT αναφέρεται στις ημέρες που απομένουν μέχρι την προθεσμία πληρωμής μιας δόσης από τον πελάτη.



Εικόνα 5-41: Διαγράμματα κατανομής της DAYS_INSTALMENT

1.Τύπος Δεδομένων και Πλήθος Τιμών

- **Τύπος δεδομένων:** float64 (Αριθμητικές τιμές).
- **Πλήθος κενώμ τιμών:** Δεν υπάρχουν κενές τιμές.

2.Ανάλυση Διαγράμματος

Παρατηρούμε ότι πολλοί πελάτες τείνουν να κάνουν πληρωμές σε συγκεκριμένες περιόδους, με την πλειοψηφία να καταβάλλει τις πληρωμές πολύ πριν από την προθεσμία. Μέσος όρος ημερών: -1041.08, δείχνοντας ότι οι πελάτες πληρώνουν κατά μέσο όρο πάνω από 1000 ημέρες πριν από την προθεσμία.

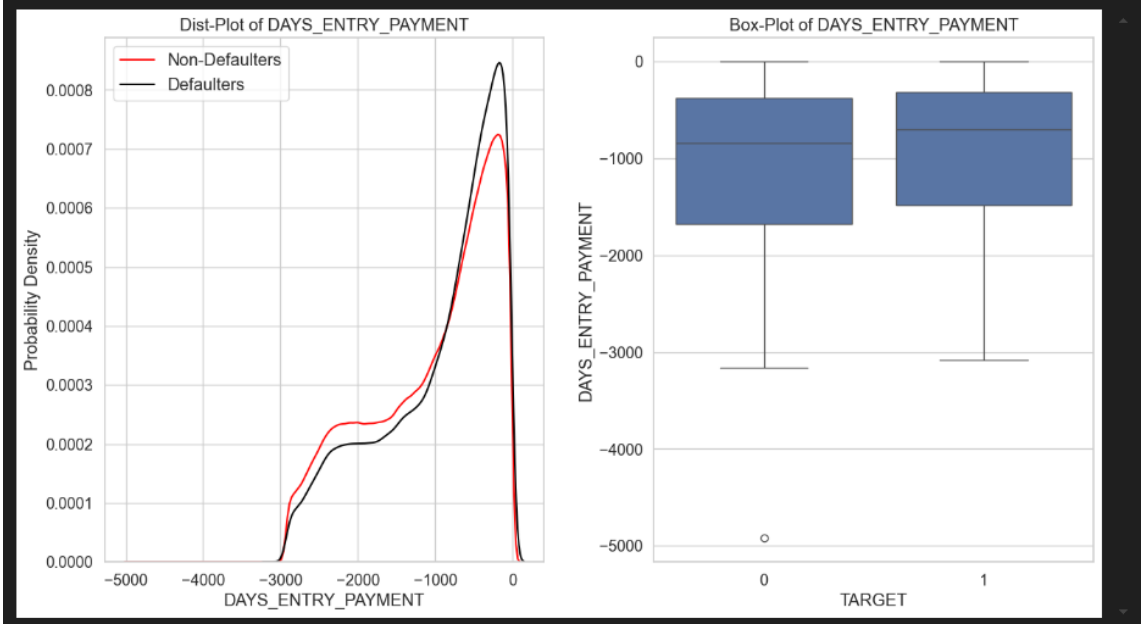
Συμπεράσματα

Η μεταβλητή DAYS_INSTALMENT προσφέρει πολύτιμες πληροφορίες σχετικά με την οικονομική συμπεριφορά των πελατών. Οι πελάτες που πληρώνουν πολύ πριν την προθεσμία παρουσιάζουν μικρότερο κίνδυνο αθέτησης, καθώς αυτό δείχνει σταθερότητα και πειθαρχία στις πληρωμές. Αντίθετα, οι πελάτες που πληρώνουν πολύ κοντά στην προθεσμία μπορεί να είναι πιο επιρρεπή σε οικονομικά προβλήματα ή αθέτηση των δανειακών τους υποχρεώσεων.

5.5.2 Ανάλυση της Μεταβλητής DAYS_ENTRY_PAYMENT

Η μεταβλητή DAYS_ENTRY_PAYMENT αναφέρεται στον αριθμό των ημερών που μεσολαβούν από την ημερομηνία καταβολής της προηγούμενης δόσης σε σχέση με την ημερομηνία αίτησης του τρέχοντος δανείου.

pp_value_counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
{-154.0: 11340, -91.0: 11197, -182.0: 11016}	18451	0.16	-1049.9	799.08	-4921.0	-829.0	-1.0	0.04



Εικόνα 5-42: Διαγράμματα κατανομής της DAYS_ENTRY_PAYMENT

1.Τύπος Δεδομένων και Πλήθος Τιμών

- **Τύπος δεδομένων:** float64 (Αριθμητικές τιμές).
- **Πλήθος κενών τιμών**: 18451 (0.16%).

2.Ανάλυση Διαγράμματος

- Στο διάγραμμα βλέπουμε ότι οι περισσότερες πληρωμές καταβάλλονται νωρίτερα από την προγραμματισμένη ημερομηνία. Το μεγαλύτερο μέρος του πληθυσμού εμφανίζει προπληρωμές, και υπάρχει μια μικρή ομάδα που πληρώνει πολύ νωρίτερα. Ο μέσος όρος ημερών είναι -1049.9, υποδεικνύοντας ότι, κατά μέσο όρο, οι πελάτες καταβάλλουν πληρωμές περισσότερες από 1,000 ημέρες πριν την ημερομηνία λήξης, κάτι που δείχνει αυξημένο ποσοστό έγκαιρων ή ακόμα και πρόωρων πληρωμών.
- Στο boxplot, παρατηρούμε κάποιες ακραίες τιμες , ιδιαίτερα στις πολύ μεγάλες αρνητικές τιμές, που αντιπροσωπεύουν πληρωμές που έγιναν εξαιρετικά νωρίς.

Συμπεράσματα

Παρατηρούμε ότι οι περισσότεροι πελάτες πληρώνουν τις δόσεις τους αρκετά πριν από την προθεσμία, κάτι που είναι θετικό σημάδι οικονομικής υγείας. Η ύπαρξη ακραίων τιμών μπορεί να σημαίνει πρόωρες εξοφλήσεις χρεών, ενώ η παρουσία πληρωμών κοντά στην προθεσμία μπορεί να υποδηλώνει πελάτες που έχουν κάποια οικονομική δυσκολία.

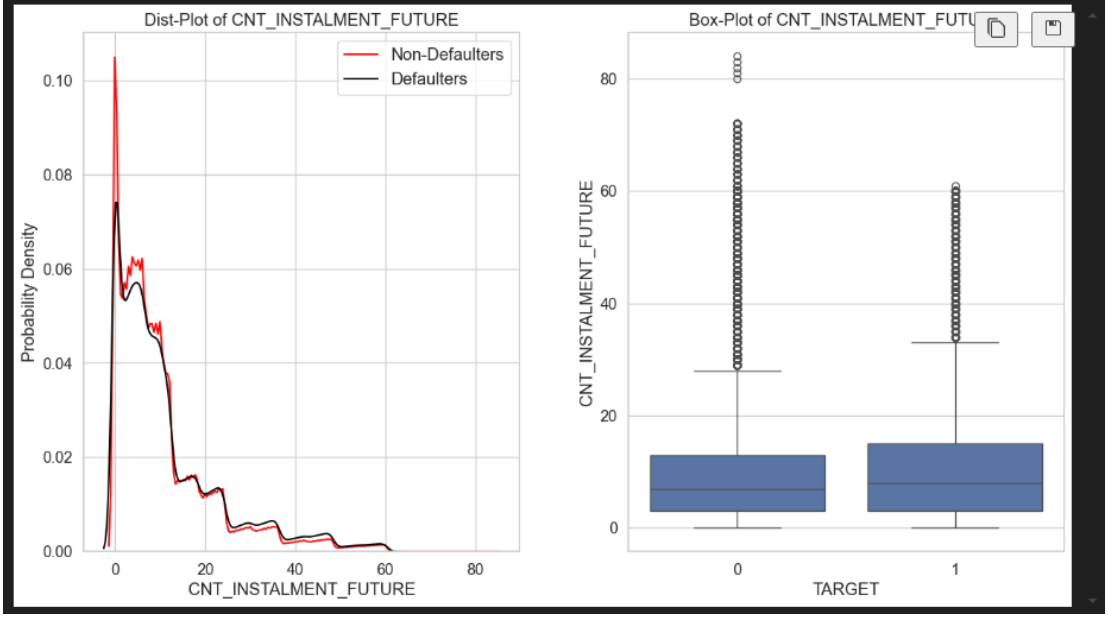
5.6 POS_CASH_BALANCE

Ο πίνακας POS_CASH_balance περιέχει λεπτομέρειες για τις συναλλαγές μέσω POS ή Cash Loans των πελατών. Συγχωνεύοντας τον με τον application_train, μπορούμε να αναλύσουμε τη συμπεριφορά των πελατών που σχετίζεται με POS/Cash Loans και να διερευνήσουμε την επίδραση αυτών των παραμέτρων στην πιθανότητα αθέτησης δανείου.

5.6.1 Ανάλυση της Μεταβλητής CNT_INSTALMENT_FUTURE

Η μεταβλητή CNT_INSTALMENT_FUTURE αναφέρεται στον αριθμό των δόσεων που απομένουν στους πελάτες από προηγούμενες πιστώσεις.

	dtype	count	unique	top_value_counts	missing_count	missing_percentage	n
CNT_INSTALMENT_FUTURE	float64	8521497	78	{0.0: 1021934, 4.0: 525573, 6.0: 524893}	39945	0.47	1



Εικόνα 5-43: Διαγράμματα κατανομής της CNT_INSTALMENT_FUTURE

1. **Τύπος Δεδομένων και Πλήθος Τιμών**
   - **Τύπος δεδομένων:** float64, αριθμητικές τιμές.
   - **Πλήθος κενών τιμών:** 39,945 κενές τιμές (0.47%).

2. **Ανάλυση των διαγραμμάτων**
   - Το διάγραμμα κατανομής δείχνει ότι υπάρχει μια συγκέντρωση γύρω από τις χαμηλές τιμές. Η πλειονότητα των δανείων φαίνεται να έχει λίγες απομένουσες δόσεις (0-7), ενώ υπάρχει μια σταδιακή μείωση στον αριθμό των δανείων καθώς αυξάνεται ο αριθμός των δόσεων που απομένουν. Ο μέσος όρος, υποδεικνύει ότι οι πελάτες έχουν περίπου 10 μελλοντικές δόσεις.

3. **Συσχέτιση με τον Στόχο**
   - Η μικρή θετική συσχέτιση 0.02 υποδεικνύει ότι οι πελάτες με περισσότερες μελλοντικές δόσεις έχουν ελαφρώς υψηλότερες πιθανότητες αθέτησης δανείων.

5.7 CREDIT_CARD_BALANCE

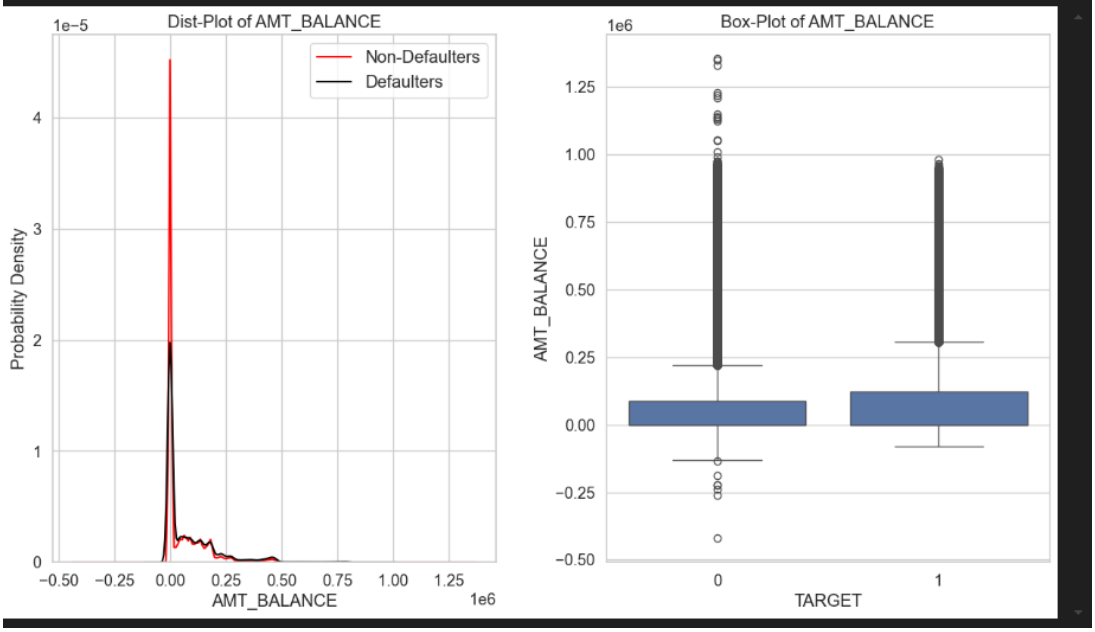
Συγχωνεύουμε τον πίνακα application_train με τον πίνακα credit_card_balance κάτι το οποίο θα μας επιτρέψει να αναλύσουμε τα υπόλοιπα των πιστωτικών καρτών των πελατών σε σχέση με τις βασικές πληροφορίες τους, καθώς και την πιθανότητα αθέτησης των δανείων τους.

Με τη συγχώνευση αυτή, θα έχουμε πρόσβαση στα δεδομένα των πελατών τόσο από τον αρχικό πίνακα των αιτήσεων, όσο και από τον πίνακα των υπολοίπων πιστωτικών καρτών, παρέχοντας έτσι πιο πλήρη πληροφόρηση για κάθε πελάτη.

5.7.1 Ανάλυση της Μεταβλητής AMT_BALANCE

Η μεταβλητή AMT_BALANCE αφορά το υπόλοιπο της πιστωτικής κάρτας των πελατών κατά τη διάρκεια του μήνα της προηγούμενης πίστωσης.

missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
220606	6.4	59073.32	106613.65	-420250.18	0.0	1354829.26	0.05



Εικόνα 5-44: Διαγράμματα κατανομής της AMT_BALANCE

1. **Τύπος Δεδομένων και Πλήθος Τιμών**
   - **Τύπος δεδομένων**: float64 (Αριθμητικές τιμές με δεκαδικά ψηφία).
   - **Πλήθος κενών τιμών**: 220,606 εγγραφές να λείπουν (6.4%)

2. **Ανάλυση απο τα Διαγράμματα**
   - Παρατηρούμε ότι η κατανομή των τιμών υπολοίπου συγκεντρώνεται κυρίως γύρω από το 0, υποδεικνύοντας ότι η πλειοψηφία των πελατών έχει μικρό ή μηδενικό υπόλοιπο στην πιστωτική τους κάρτα. Ωστόσο, υπάρχουν πελάτες με πολύ υψηλά υπόλοιπα.
   - Το διάγραμμα (boxplot) αποκαλύπτει την παρουσία αρκετών outliers, ειδικά στην περιοχή των υψηλών υπολοίπων, που δείχνουν σημαντικές διακυμάνσεις στο χρέος των πελατών.

3. Συσχέτιση με το Στοχο

•Ο συντελεστής συσχέτισης με την πιθανότητα χρεοκοπίας (TARGET) είναι περίπου 0.05, υποδεικνύοντας ότι υπάρχει μια πολύ αδύναμη θετική συσχέτιση ανάμεσα στο υπόλοιπο της πιστωτικής κάρτας και την πιθανότητα αθέτησης δανείου.

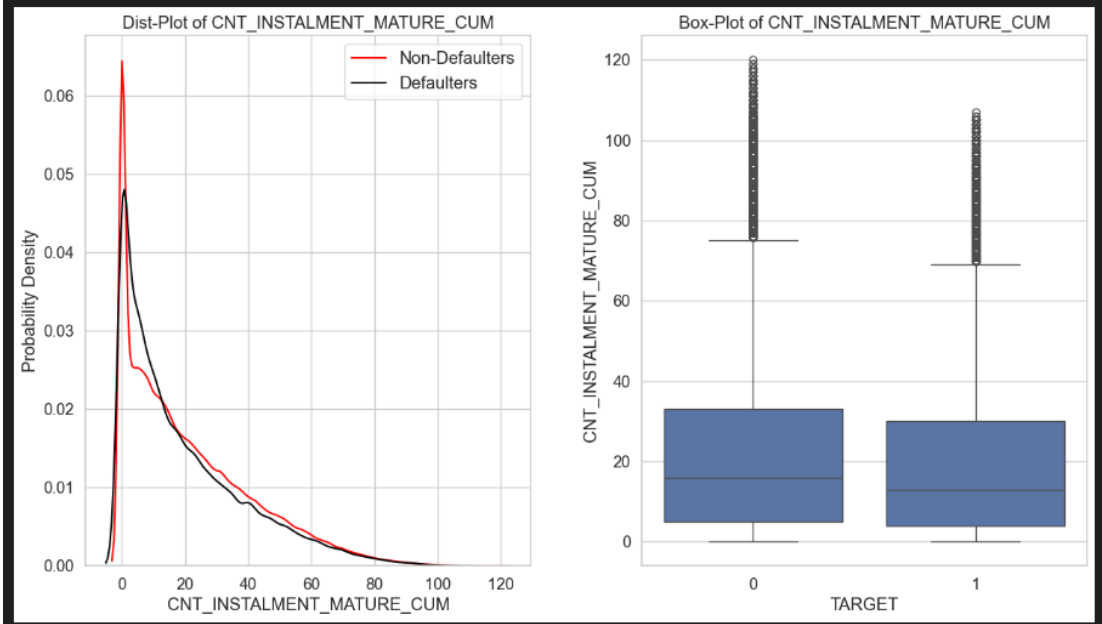
Συμπεράσματα

Η μεταβλητή AMT_BALANCE δείχνει ότι ένα μεγάλο ποσοστό πελατών έχει μηδενικά ή χαμηλά υπόλοιπα πιστωτικών καρτών, ενώ ορισμένοι πελάτες έχουν σημαντικά χρέη, κάτι που μπορεί να σχετίζεται με την αθέτηση πληρωμών. Παρόλο που η συσχέτιση με την πιθανότητα αθέτησης δανείου είναι αδύναμη, η μεταβλητή παραμένει χρήσιμη για τη γενικότερη ανάλυση της πιστοληπτικής συμπεριφοράς των πελατών.

5.7.2 Ανάλυση της Μεταβλητής CNT_INSTALMENT_MATURE_CUM

Η μεταβλητή CNT_INSTALMENT_MATURE_CUM αναφέρεται στον συνολικό αριθμό των δόσεων που έχουν πληρωθεί για την πιστωτική κάρτα του πελάτη μέχρι το συγκεκριμένο χρονικό σημείο.

e	top_value_counts	missing_count	missing_percentage	mean	std	min	median	max	corr_with_target
1	{0.0: 431057, 5.0: 77844, 2.0: 77182}	484990	14.06	21.09	19.99	0.0	16.0	120.0	-0.02



Εικόνα 5-45: Διαγράμματα κατανομής της CNT_INSTALMENT_MATURE_CUM

1. Τύπος Δεδομένων και Πλήθος Τιμών

- Τύπος δεδομένων: float64 (Αριθμητικές τιμές)
- Πλήθος κενών τιμών: 484,990 κενές τιμές(14.06%).

2. Ανάλυση από τα Διαγράμματα

•Η κατανομή είναι ασύμμετρη με σημαντική συγκέντρωση στις χαμηλές τιμές, υποδεικνύοντας ότι πολλοί πελάτες δεν έχουν πληρώσει πολλές δόσεις ή βρίσκονται στα αρχικά στάδια της αποπληρωμής.

- Το διάγραμμα boxplot αποκαλύπτει την ύπαρξη outliers για τις υψηλές τιμές, γεγονός που ενδέχεται να επηρεάζει την ερμηνεία της κατανομής και να δείχνει λίγους πελάτες με μεγάλο αριθμό δόσεων.

3. Συσχέτιση με το Στόχο

- Ο συντελεστής συσχέτισης με την πιθανότητα αθέτησης δανείου είναι περίπου -0.02, υποδεικνύοντας πολύ αδύναμη αρνητική συσχέτιση. Αυτό σημαίνει ότι οι πελάτες με περισσότερες πληρωμένες δόσεις είναι ελαφρώς λιγότερο πιθανό να αθετήσουν την αποπληρωμή τους.

5.8 Συμπεράσματα από την Διερευνητική Ανάλυση Δεδομένων (EDA)

Κατά την διερευνητική ανάλυση δεδομένων (EDA), προέκυψαν ορισμένα σημαντικά ευρήματα τα οποία θα αποτελέσουν τη βάση για τα επόμενα βήματα της ανάλυσης:

- **Κενές τιμές (Missing Values):** Εντοπίστηκε μεγάλος αριθμός μεταβλητών με κενές τιμές. Ένα σημαντικό ποσοστό από αυτές τις κενές τιμές αφορούν χαρακτηριστικά των κατοικιών και άλλων στατικών δεδομένων που πιθανώς δεν καταγράφηκαν πλήρως για όλους τους πελάτες. Η διαχείριση αυτών των κενών τιμών θα είναι κρίσιμη για την ακεραιότητα των μοντέλων, καθώς θα πρέπει να αποφασιστεί ποιες μεταβλητές θα παραμείνουν και πώς θα συμπληρωθούν.

- **Ακραίες τιμές (Outliers):** Παρατηρήθηκαν πολλά ακραία δεδομένα σε μεταβλητές που σχετίζονται με το χρόνο, όπως η μεταβλητή DAYS_EMPLOYED, όπου εμφανίστηκαν τιμές των 1000 ετών, κάτι που είναι εξωπραγματικό και μάλλον αποτελεί σφάλμα καταγραφής. Αντίστοιχα, στη μεταβλητή DAYS_FIRST_DRAWING, υπήρχαν περιπτώσεις με 1000 χρόνια, οι οποίες δεν είναι ρεαλιστικές. Αυτά τα δεδομένα θα χρειαστεί να απομονωθούν ή να διορθωθούν, ώστε να μην επηρεάσουν την απόδοση των μοντέλων. Παρόμοια ακραία τιμή παρατηρήθηκε σε πολλές χρονικές μεταβλητές, υποδεικνύοντας είτε λάθη στα δεδομένα είτε μη καταχωρημένες πληροφορίες.

- **Κατηγορικές και αριθμητικές μεταβλητές:** Κατά την ανάλυση των κατηγορικών μεταβλητών, εντοπίσαμε σημαντικές αποκλίσεις στις κατανομές μεταξύ των διαφόρων κατηγοριών. Αυτές οι μεταβλητές παρουσιάζουν ενδιαφέρον για τη διαμόρφωση χαρακτηριστικών, ενώ οι αριθμητικές μεταβλητές παρουσίασαν ακραίες τιμές, ασυμμετρίες και διαφορετικές κατανομές. Αυτά τα χαρακτηριστικά θα αξιοποιηθούν με κατάλληλες μεθόδους στο *Feature Selection* και *Feature Engineering*.

Συνολικά, το επόμενο στάδιο θα επικεντρωθεί στη διαχείριση των κενών τιμών και των ακραίων τιμών, ειδικά αυτών που σχετίζονται με μη ρεαλιστικές χρονικές τιμές. Η προσεκτική προετοιμασία των δεδομένων θα είναι καθοριστική για τη βελτίωση της ακρίβειας των μοντέλων και τη βέλτιστη επιλογή χαρακτηριστικών για τη μηχανική μάθηση.

6 Data Preparation and Feature Engineering

Από την διερευνητική ανάλυση των δεδομένων (EDA) που κάναμε παρατηρούμε ότι υπάρχουν κάποιες ακραίες τιμές και πολλές τιμές που λείπουν τις οποίες θα πρέπει να διαχειριστούμε.

6.1 Application_train.csv

Στο application_train.csv (307511,122) έχουμε τέσσερις(4) εγγραφές με τιμή 'XNA' στο 'CODE_GENDER' τις οποίες και θα αφαιρέσουμε από το δείγμα μας καθώς είναι μια ασυνήθιστη τιμή που πιθανότατα δεν προσφέρει χρήσιμες πληροφορίες για την εκπαίδευση των μοντέλων. Επίσης στη στήλη 'DAYS_EMPLOYED' έχουμε τιμές 365.243, η οποία πιθανότατα αντιπροσωπεύει εσφαλμένα δεδομένα, δηλαδή ότι κάποιος εργάζεται για 1000 χρόνια, τις οποίες θα αντικαταστήσουμε με την κενή τιμή NaN (Not-a-Number).

```
df.dtypes.value_counts()
✓ 0.0s

float64    66
int64      40
object     16
Name: count, dtype: int64
```

Εικόνα 6-1: Τύποι δεδομένων του application_train

Στη συνέχεια θα μετατρέψουμε τα (16) κατηγορικά χαρακτηριστικά σε αριθμητικά ώστε να μπορούν να τα χειριστούν τα μοντέλα μας. Για τα χαρακτηριστικά 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY' τα οποία έχουν μόνο δύο κατηγορίες (M/F, Y/N) θα χρησιμοποιήσουμε την μέθοδο **Label-encoding** και για τις υπόλοιπες κατηγορικές μεταβλητές που έχουν περισσότερες από δύο κατηγορίες, θα εφαρμόσουμε **One-Hot Encoding**.

Τέλος θα δημιουργήσουμε μερικά νέα χαρακτηριστικά υπολογίζοντας ποσοστιαίες σχέσεις μεταξύ των ήδη υπαρχόντων.

```
# Some simple new features (percentages)
df['DAYS_EMPLOYED_PERC'] = df['DAYS_EMPLOYED'] / df['DAYS_BIRTH']
df['INCOME_CREDIT_PERC'] = df['AMT_INCOME_TOTAL'] / df['AMT_CREDIT']
df['INCOME_PER_PERSON'] = df['AMT_INCOME_TOTAL'] / df['CNT_FAM_MEMBERS']
df['ANNUITY_INCOME_PERC'] = df['AMT_ANNUITY'] / df['AMT_INCOME_TOTAL']
df['PAYMENT_RATE'] = df['AMT_ANNUITY'] / df['AMT_CREDIT']
```

Εικόνα 6-2: Δημιουργία χαρακτηριστικών application_train

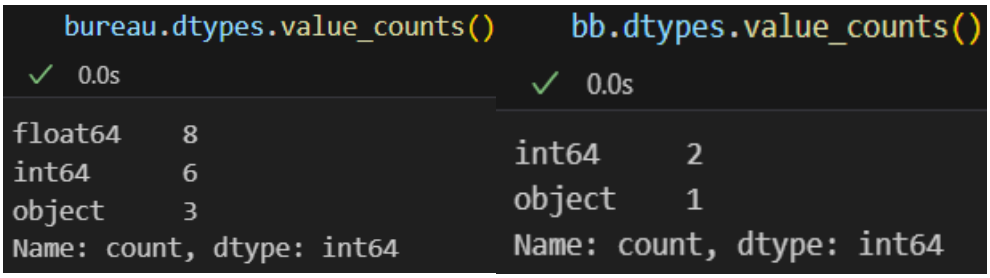
Αυτά το νέα χαρακτηριστικά μπορούν να προσφέρουν πρόσθετες πληροφορίες για τη συμπεριφορά των αιτούντων και να βελτιώσουν τις προβλέψεις των μοντέλων.

- **DAYS_EMPLOYED_PERC**: Ποσοστό ημερών εργασίας σε σχέση με την ηλικία.
- **INCOME_CREDIT_PERC**: Ποσοστό συνολικού εισοδήματος προς το ποσό του πιστωτικού δανείου.
- **INCOME_PER_PERSON**: Συνολικό εισόδημα ανά μέλος οικογένειας.
- **ANNUITY_INCOME_PERC**: Ποσοστό της ετήσιας δόσης προς το συνολικό εισόδημα.
- **PAYMENT_RATE**: Ποσοστό ετήσιας δόσης προς το ποσό του δανείου.

Μετά από την επεξεργασία των δεδομένων και την δημιουργία των νέων χαρακτηριστικών το σύνολο των δεδομένων application_train.csv έχει διαστάσεις (307508,247).

6.2 bureau.csv και bureau_balance.csv

Στα bureau.csv και bureau_balance.csv ξεκινάμε με το one-hot encoding των κατηγορικών μεταβλητών.



Εικόνα 6-3: Τύποι δεδομένων των bureau και bureau_balance

Το αρχείο bureau_balance.csv περιέχει μηνιαίες πληροφορίες για την κατάσταση (status) των δανείων του κάθε αιτούντα. Για αυτά τα δεδομένα δημιουργήσουμε νέα χαρακτηριστικά χρησιμοποιώντας συγκεντρωτικά στατιστικά που συμπυκνώνουν τα μηνιαία δεδομένα σε χρήσιμες μετρήσεις (όπως ελάχιστες, μέγιστες ή μέσες τιμές).

```
# Bureau balance: Perform aggregations and merge with bureau.csv
bb_aggregations = {'MONTHS_BALANCE': ['min', 'max', 'size']}
for col in bb_cat:
    bb_aggregations[col] = ['mean']
bb_agg = bb.groupby('SK_ID_BUREAU').agg(bb_aggregations)
bb_agg.columns = pd.Index([e[0] + "_" + e[1].upper() for e in bb_ag
bureau = bureau.join(bb_agg, how='left', on='SK_ID_BUREAU')
bureau.drop(['SK_ID_BUREAU'], axis=1, inplace= True)
```

Εικόνα 6-4: Δημιουργία συγκεντρωτικών χαρακτηριστικών bureau_balance

Για **αριθμητικές μεταβλητές** όπως για παράδειγμα το MONTHS_BALANCE, θα υπολογίσουμε μετρήσεις όπως:

- **Ελάχιστο**: Ο πρώτος μήνας για τον οποίο υπάρχουν δεδομένα (MONTHS_BALANCE_min).
- **Μέγιστο**: Ο πιο πρόσφατος μήνας στον οποίο υπάρχει καταγραφή (MONTHS_BALANCE_max).
- **Μέγεθος**: Το πλήθος των μηνών με καταγεγραμμένα δεδομένα (MONTHS_BALANCE_size).

Για **κατηγορικές μεταβλητές** μετά το one-hot encoding, θα υπολογίσουμε:

- **Μέσος όρος**: Το μέσο ποσοστό εμφάνισης κάθε κατηγορίας κατά τη διάρκεια των μηνών.

Το bureau_balance το οποίο σχετίζεται με το bureau όπως βλέπουμε από το σχεσιακό μοντέλο των δεδομένων με το 'SK_BUREAU_ID' αφού ομαδοποιήσουμε τα δεδομένα του ανά 'SK_ID_BUREAU' (ο μοναδικός κωδικός για κάθε δάνειο) και υπολογίσουμε τις συγκεντρωτικές μετρήσεις, τα συγχωνεύουμε με το κύριο σύνολο δεδομένων bureau.csv χρησιμοποιώντας το SK_ID_BUREAU. Αυτό μας επιτρέπει να έχουμε όλα τα σχετικά δεδομένα για κάθε δάνειο σε ένα ενιαίο σύνολο δεδομένων.

Τώρα που το σύνολο δεδομένων bureau.csv έχει εμπλουτιστεί με τα δεδομένα του bureau_balance.csv , προχωράμε στη δημιουργία νέων χαρακτηριστικών ομαδοποιώντας τα δεδομένα για κάθε αιτούντα και υπολογίζοντας συγκεντρωτικές μετρήσεις για τα δάνεια.

Για τις **Αριθμητικές μεταβλητές** θα υπολογίσουμε μετρήσεις όπως:

- **Ελάχιστο/Μέγιστο/Μέσος όρος/Άθροισμα/Διακύμανση**

```
# Bureau and bureau_balance numeric features
num_aggregations = {
    'DAYS_CREDIT': ['min', 'max', 'mean', 'var'],
    'DAYS_CREDIT_ENDDATE': ['min', 'max', 'mean'],
    'DAYS_CREDIT_UPDATE': ['mean'],
    'CREDIT_DAY_OVERDUE': ['max', 'mean'],
    'AMT_CREDIT_MAX_OVERDUE': ['mean'],
    'AMT_CREDIT_SUM': ['max', 'mean', 'sum'],
    'AMT_CREDIT_SUM_DEBT': ['max', 'mean', 'sum'],
    'AMT_CREDIT_SUM_OVERDUE': ['mean'],
    'AMT_CREDIT_SUM_LIMIT': ['mean', 'sum'],
    'AMT_ANNUITY': ['max', 'mean'],
    'CNT_CREDIT_PROLONG': ['sum'],
    'MONTHS_BALANCE_MIN': ['min'],
    'MONTHS_BALANCE_MAX': ['max'],
    'MONTHS_BALANCE_SIZE': ['mean', 'sum']
}
```

Εικόνα 6-5: Συγκεντρωτικά χαρακτηριστικά bureau με bureau_balance

Για τις **κατηγορικές μεταβλητές** ,θα υπολογίσουμε τον **μέσο όρο** για κάθε δάνειο. Αυτό δίνει μια ιδέα για το πόσο συχνά εμφανίζονται ορισμένες κατηγορίες στην πιστωτική ιστορία του κάθε αιτούντα.

Ομαδοποιούμε τα δεδομένα ανά SK_ID_CURR (ο μοναδικός κωδικός για κάθε αιτούντα) και εφαρμόζουμε τις συναρτήσεις συγκέντρωσης για τα δεδομένα κάθε αιτούντα:

```
bureau_agg = bureau.groupby('SK_ID_CURR').agg({**num_aggregations, **cat_aggregations})
bureau_agg.columns = pd.Index(['BURO_' + e[0] + "_" + e[1].upper() for e in bureau_agg.columns.tolist()])
```

Εικόνα 6-6: Ομαδοποίηση συναθροίσεων bureau_agg

Επίσης ξεχωρίζουμε τα δεδομένα σε δύο μέρη:

- **Ενεργά δάνεια**: Για δάνεια που είναι ενεργά (CREDIT_ACTIVE_Active), υπολογίζουμε τις ίδιες συγκεντρωτικές μετρήσεις. Αυτό παρέχει μια εικόνα για τις ενεργές πιστωτικές υποχρεώσεις ενός αιτούντα.

```
# Bureau: Active credits - using only numerical aggregations
active = bureau[bureau['CREDIT_ACTIVE_Active'] == 1]
active_agg = active.groupby('SK_ID_CURR').agg(num_aggregations)
active_agg.columns = pd.Index(['ACTIVE_' + e[0] + "_" + e[1].upper() for e in
bureau_agg = bureau_agg.join(active_agg, how='left', on='SK_ID_CURR')
```

Εικόνα 6-7: Δημιουργία συγκεντρωτικών χαρακτηριστικών για τα ενεργά δάνεια του bureau

- **Κλειστά δάνεια**: Ομοίως, υπολογίζουμε τις συγκεντρωτικές μετρήσεις για τα κλειστά δάνεια, δίνοντας πληροφορίες για το ιστορικό των κλειστών πιστωτικών λογαριασμών.

```
# Bureau: Closed credits - using only numerical aggregations
closed = bureau[bureau['CREDIT_ACTIVE_Closed'] == 1]
closed_agg = closed.groupby('SK_ID_CURR').agg(num_aggregations)
closed_agg.columns = pd.Index(['CLOSED_' + e[0] + "_" + e[1].upper() for
bureau_agg = bureau_agg.join(closed_agg, how='left', on='SK_ID_CURR')
```

Εικόνα 6-8: Δημιουργία συγκεντρωτικών χαρακτηριστικών για τα κλειστά δάνεια του bureau

Τέλος, συγχωνεύουμε τα συγκεντρωτικά δεδομένα (active και closed) στο κύριο σύνολο δεδομένων bureau_agg, δίνοντας μια ολοκληρωμένη σύνοψη της πιστωτικής ιστορίας κάθε αιτούντα, συμπεριλαμβανομένων των ενεργών και κλειστών δανείων. Το τελικό αποτέλεσμα είναι ένα σύνολο δεδομένων διαστάσεων (305811,116) όπου κάθε γραμμή αντιπροσωπεύει έναν αιτούντα (SK_ID_CURR), και κάθε στήλη αντιπροσωπεύει μια σημαντική μέτρηση για την πιστωτική ιστορία του αιτούντα.

6.3 previous_application.csv

Στο previous_application (1670214,37) αρχικά θα εφαρμόσουμε One-Hot Encoding για τις κατηγορικές μεταβλητές ώστε να μπορούν να χρησιμοποιηθούν στα μοντέλα μηχανικής μάθησης.

```
prev.dtypes.value_counts()
✓ 0.0s

object     16
float64    15
int64       6
Name: count, dtype: int64
```

Εικόνα 6-9: Τύποι δεδομένων του previous_application

Μερικές στήλες, όπως είδαμε από την ανάλυση, που αφορούν ημερομηνίες περιέχουν την τιμή (365.243), η οποία πιθανώς υποδηλώνει "άγνωστη" ή "λανθασμένη" τιμή. Αυτές οι τιμές θα τις αντικαταστήσουμε με NaN (κενές τιμές).

```
prev['DAYS_FIRST_DRAWING'].replace(365243, np.nan, inplace= True)
prev['DAYS_FIRST_DUE'].replace(365243, np.nan, inplace= True)
prev['DAYS_LAST_DUE_1ST_VERSION'].replace(365243, np.nan, inplace= True)
prev['DAYS_LAST_DUE'].replace(365243, np.nan, inplace= True)
prev['DAYS_TERMINATION'].replace(365243, np.nan, inplace= True)
```

Εικόνα 6-10: Αντικατάσταση ακραίων τιμών του previous_application

Στην συνέχεια θα δημιουργήσουμε το χαρακτηριστικό APP_CREDIT_PERC που είναι το ποσοστό του ποσού που ζητήθηκε προς το ποσό που εγκρίθηκε και συνέχεια θα δημιουργήσουμε νέα χαρακτηριστικά από τις συγκεντρωτικές μετρήσεις.

```
# Add feature: value ask / value received percentage
prev['APP_CREDIT_PERC'] = prev['AMT_APPLICATION'] / prev['AMT_CREDIT']
# Previous applications numeric features
num_aggregations = {
    'AMT_ANNUITY': ['min', 'max', 'mean'],
    'AMT_APPLICATION': ['min', 'max', 'mean'],
    'AMT_CREDIT': ['min', 'max', 'mean'],
    'APP_CREDIT_PERC': ['min', 'max', 'mean', 'var'],
    'AMT_DOWN_PAYMENT': ['min', 'max', 'mean'],
    'AMT_GOODS_PRICE': ['min', 'max', 'mean'],
    'HOUR_APPR_PROCESS_START': ['min', 'max', 'mean'],
    'RATE_DOWN_PAYMENT': ['min', 'max', 'mean'],
    'DAYS_DECISION': ['min', 'max', 'mean'],
    'CNT_PAYMENT': ['mean', 'sum'],
}
# Previous applications categorical features
cat_aggregations = {}
for cat in cat_cols:
    cat_aggregations[cat] = ['mean']
```

Εικόνα 6-11: Συγκεντρωτικά χαρακτηριστικά previous_application

Αφού διαχωρίζουμε τις εγκεκριμένες από τις απορριφθείσες αιτήσεις και ομαδοποιήσουμε τα δεδομένα με βάση το SK_ID_CURR υπολογίζουμε ξεχωριστά τις συγκεντρωτικές μετρήσεις για τα αριθμητικά χαρακτηριστικά των προηγούμενων αιτήσεων (όπως ελάχιστες, μέγιστες ,μέσες τιμές η διακύμανση), ώστε να έχουμε μια ξεχωριστή εικόνα για τα ποσοστά επιτυχίας ή αποτυχίας των αιτούντων καθώς και τον μέσο ορο για κάθε κατηγορία που προέκυψε από το one-hot encoding. Το αποτέλεσμα είναι ένα σύνολο δεδομένων όπου κάθε γραμμή αντιπροσωπεύει έναν αιτούντα, και κάθε στήλη περιέχει μια συνοπτική πληροφορία για τις προηγούμενες αιτήσεις του.

```
# Previous Applications: Approved Applications - only numerical features
approved = prev[prev['NAME_CONTRACT_STATUS_Approved'] == 1]
approved_agg = approved.groupby('SK_ID_CURR').agg(num_aggregations)
approved_agg.columns = pd.Index(['APPROVED_' + e[0] + "_" + e[1].upper() for e i
prev_agg = prev_agg.join(approved_agg, how='left', on='SK_ID_CURR')
# Previous Applications: Refused Applications - only numerical features
refused = prev[prev['NAME_CONTRACT_STATUS_Refused'] == 1]
refused_agg = refused.groupby('SK_ID_CURR').agg(num_aggregations)
refused_agg.columns = pd.Index(['REFUSED_' + e[0] + "_" + e[1].upper() for e in
prev_agg = prev_agg.join(refused_agg, how='left', on='SK_ID_CURR')
```

Εικόνα 6-12: Υπολογισμός συγκεντρωτικών μετρήσεων για εγκεκριμένα και απορριφθέντα δάνεια

Το τελικό μέγεθος του συνόλου δεδομένων του previous_applications φτάνει τις (338857,249).

6.4 POS_CASH_BALANCE

Στο POS_CASH_BALANCE (10001358, 8) αρχικά θα εφαρμόσουμε **One-Hot Encoding** στη κατηγορική μεταβλητή.

```
pos.dtypes.value_counts(
✓ 0.0s

int64      5
float64    2
object     1
Name: count, dtype: int64
```

Εικόνα 6-13: Τύποι δεδομένων του POS_CASH_BALANCE

Στη συνέχεια, θα δημιουργήσουμε συγκεντρωτικά χαρακτηριστικά για κάθε αίτηση και θα ομαδοποιήσουμε τα δεδομένα με βάση το SK_ID_CURR . Επίσης υπολογίζουμε τον μέσο όρο της παρουσίας της κάθε κατηγορίας στις εγγραφές POS για κάθε αιτούντα.

```
aggregations = {
    'MONTHS_BALANCE': ['max', 'mean', 'size'],
    'SK_DPD': ['max', 'mean'],
    'SK_DPD_DEF': ['max', 'mean']
}
for cat in cat_cols:
    aggregations[cat] = ['mean']
```

Εικόνα 6-14: Συγκεντρωτικά χαρακτηριστικά POS_CASH_BALANCE

Τέλος προσθέτουμε και ένα νέο χαρακτηριστικό POS_COUNT που υπολογίζει το πλήθος των λογαριασμών POS που έχει ο κάθε αιτών. Αυτή η πληροφορία μπορεί να είναι χρήσιμη για να κατανοήσουμε τον όγκο των συναλλαγών του κάθε αιτούντα.

Μετά από την επεξεργασία το POS_CASH_balance έχει μέγεθος (337252 , 18).

6.5 Installments_Payments.csv

Το INSTALLMENTS_PAYMENTS (13605401, 08) περιέχει πληροφορίες για τις πληρωμές δόσεων των αιτούντων. Αρχικά θα εφαρμόσουμε One-Hot Encoding στις κατηγορικές μεταβλητές, ώστε να μπορούν να χρησιμοποιηθούν σε μοντέλα μηχανικής μάθησης

```
ins.dtypes.value_counts()
✓ 0.0s

float64    5
int64      3
Name: count, dtype: int64
```

Εικόνα 6-15: Τύποι δεδομένων του Installments_Payments

Στην συνέχεια θα δημιουργήσουμε μερικά νέα χαρακτηριστικά συνδυάζοντας τα ήδη υπάρχοντα.

```
# Percentage and difference paid in each installment (amount paid and installment value)
ins['PAYMENT_PERC'] = ins['AMT_PAYMENT'] / ins['AMT_INSTALMENT']
ins['PAYMENT_DIFF'] = ins['AMT_INSTALMENT'] - ins['AMT_PAYMENT']
# Days past due and days before due (no negative values)
ins['DPD'] = ins['DAYS_ENTRY_PAYMENT'] - ins['DAYS_INSTALMENT']
ins['DBD'] = ins['DAYS_INSTALMENT'] - ins['DAYS_ENTRY_PAYMENT']
ins['DPD'] = ins['DPD'].apply(lambda x: x if x > 0 else 0)
ins['DBD'] = ins['DBD'].apply(lambda x: x if x > 0 else 0)
```

Εικόνα 6-16: Δημιουργία νέων χαρακτηριστικών του Installments_payments

- **PAYMENT_PERC**: Υπολογίζουμε το ποσοστό της πληρωμής σε κάθε δόση, δηλαδή το ποσό που πληρώθηκε προς την οφειλόμενη δόση.
- **PAYMENT_DIFF**: Υπολογίζουμε τη διαφορά μεταξύ της οφειλόμενης δόσης και του ποσού που τελικά πληρώθηκε. Αυτό μπορεί να δείξει αν κάποιος πληρώνει περισσότερο ή λιγότερο από ό,τι οφείλει.
- **DPD (Days Past Due)**: Υπολογίζουμε τις ημέρες καθυστέρησης (αν πληρώθηκε μετά την προθεσμία).
- **DBD (Days Before Due)**: Υπολογίζουμε τις ημέρες πριν την προθεσμία (αν πληρώθηκε πριν την προθεσμία).

Για να αποφύγουμε αρνητικές τιμές, θα εφαρμόσουμε μια συνάρτηση που κρατά μόνο τις θετικές τιμές.

Τέλος ομαδοποιούμε τα δεδομένα με βάση το SK_ID_CURR και υπολογίζουμε διάφορες συγκεντρωτικές μετρήσεις, όπως:

- **DPD και DBD**: Μέγιστες, μέσες τιμές και άθροισμα για τις καθυστερήσεις και τις προθεσμίες πληρωμών.
- **PAYMENT_PERC και PAYMENT_DIFF**: Υπολογίζετε τον μέγιστο, τον μέσο όρο, το άθροισμα και τη διακύμανση.

```
# Features: Perform aggregations
aggregations = {
    'NUM_INSTALMENT_VERSION': ['nunique'],
    'DPD': ['max', 'mean', 'sum'],
    'DBD': ['max', 'mean', 'sum'],
    'PAYMENT_PERC': ['max', 'mean', 'sum', 'var'],
    'PAYMENT_DIFF': ['max', 'mean', 'sum', 'var'],
    'AMT_INSTALMENT': ['max', 'mean', 'sum'],
    'AMT_PAYMENT': ['min', 'max', 'mean', 'sum'],
    'DAYS_ENTRY_PAYMENT': ['max', 'mean', 'sum']
}
for cat in cat_cols:
    aggregations[cat] = ['mean']
ins_agg = ins.groupby('SK_ID_CURR').agg(aggregations)
```

Εικόνα 6-17: Συγκεντρωτικά χαρακτηριστικά Installment_Payments

Ακόμα θα προσθέσουμε και μια νέα στήλη INSTAL_COUNT, η οποία υπολογίζει το πλήθος των πληρωμών δόσεων για κάθε αιτούντα.

```
# Count installments accounts
ins_agg['INSTAL_COUNT'] = ins.groupby('SK_ID_CURR').size()
```

Εικόνα 6-18: Δημιουργία του χαρακτηριστικού INSTAL_COUNT

Από την επεξεργασία και τη δημιουργία των νέων αυτών χαρακτηριστικών το μέγεθος του συνόλου των δεδομένων Installment_payments είναι (339587, 26).

6.6 credit_card_balance.csv

Το σύνολο credit_card_balance (3840312, 23) περιέχει πληροφορίες για τις συναλλαγές και τα υπόλοιπα πιστωτικών καρτών. Εφαρμόζουμε One-Hot Encoding για τις κατηγορικές μεταβλητές.

```
cc.dtypes.value_counts()
✓ 0.0s

float64    15
int64       7
object      1
Name: count, dtype: int64
```

Εικόνα 6-19: Τύποι δεδομένων του credit_card_balance

Στην συνέχεια ομαδοποιούμε τα δεδομένα ανά SK_ID_CURR και υπολογίζουμε διάφορες συγκεντρωτικές μετρήσεις για όλες τις αριθμητικές στήλες, όπως:

- **Ελάχιστο , Μέγιστο , Μέσος όρος , Άθροισμα και Διακύμανση.**

```
# General aggregations
cc.drop(['SK_ID_PREV'], axis= 1, inplace = True)
cc_agg = cc.groupby('SK_ID_CURR').agg(['min', 'max', 'mean', 'sum', 'var'])
cc_agg.columns = pd.Index(['CC_' + e[0] + "_" + e[1].upper() for e in cc_ag
cc_agg.columns = ['CC_' + '_'.join(col).upper() for col in cc_agg.columns]
# Count credit card lines
cc_agg['CC_COUNT'] = cc.groupby('SK_ID_CURR').size()
```

Εικόνα 6-20: Δημιουργία συγκεντρωτικών χαρακτηριστικών του credit_card_balance

Προσθέτουμε και μια νέα στήλη CC_COUNT που υπολογίζει το πλήθος των γραμμών πιστωτικής κάρτας για κάθε αιτούντα. Αυτή η πληροφορία μπορεί να είναι χρήσιμη για να κατανοήσουμε τον αριθμό των πιστωτικών καρτών ή συναλλαγών που έχει κάθε ο κάθε αιτών.

Το τελικό μέγεθος του συνόλου credit_card_balance που θα χρησιμοποιήσουμε είναι (103558, 141).

6.7 Συνένωση όλων των Συνόλων Δεδομένων.

Τέλος όλα τα σύνολα δεδομένων που περιέχουν τι συγκεντρωτικές πληροφορίες που φτιάξαμε (όπως το bureau_agg, prev_agg, pos_agg, ins_agg, και cc_agg) συγχωνεύεται με το κύριο σύνολο δεδομένων, βάση του σχεσιακού μοντέλου που μας έχει δοθεί, χρησιμοποιώντας το SK_ID_CURR.

```
print("Bureau df shape:",bureau_agg.shape)
df = df.join(bureau_agg, how='left', on='SK_ID_CURR')
del bureau_agg
gc.collect()
print("Previous applications df shape:", prev_agg.shape)
df = df.join(prev_agg, how='left', on='SK_ID_CURR')
del prev_agg
gc.collect()
print("Pos-cash balance df shape:", pos_agg.shape)
df = df.join(pos_agg, how='left', on='SK_ID_CURR')
del pos_agg
gc.collect()
print("Installments payments df shape:", ins_agg.shape)
df = df.join(ins_agg, how='left', on='SK_ID_CURR')
del ins_agg
gc.collect()
print("Credit card balance df shape:", cc_agg.shape)
df = df.join(cc_agg, how='left', on='SK_ID_CURR')
del cc_agg
gc.collect()
print("Merged df shape:", df.shape)
```

Εικόνα 6-21: Συνένωση όλων των συνόλων δεδομένων

Το τελικό σύνολο δεδομένων, μετά από την προεπεξεργασία και δημιουργία των νέων χαρακτηριστικών που φτιάξαμε έχει διαστάσεις (307507,797), δηλαδή 797 χαρακτηριστικά.

```
Bureau df shape: (305811, 116)
Previous applications df shape: (338857, 249)
Pos-cash balance df shape: (337252, 18)
Installments payments df shape: (339587, 26)
Credit card balance df shape: (103558, 141)
Merged df shape: (356251, 797)
```

Εικόνα 6-22: Διαστάσεις των συνόλων δεδομένων μας μετά την δημιουργίαχαρακτηριστικών και τη συνένωση τους

6.8 Διαχείριση ειδικών χαρακτήρων στα ονόματα των στηλών.

Από την δημιουργία των συγκεντρωτικών χαρακτηριστικών δημιουργήθηκαν στήλες με ονόματα που περιέχουν ειδικούς χαρακτήρες. Για να διασφαλιστεί η ομαλή λειτουργία των μοντέλων, πρέπει να αντιμετωπιστούν τυχόν στήλες του συνόλου δεδομένων που περιέχουν τέτοιους χαρακτήρες, όπως αυτοί που χρησιμοποιούνται σε δομές JSON (π.χ. {}, [], :, ,, \"). Αυτοί οι χαρακτήρες θα μπορούσαν να προκαλέσουν προβλήματα στην ανάλυση και την εκπαίδευση των μοντέλων, καθώς δεν είναι πάντα αποδεκτοί στα ονόματα των στηλών. Για τον λόγο αυτό, αναπτύχθηκε μια συνάρτηση που εντοπίζει και αντικαθιστά αυτούς τους χαρακτήρες με πιο κατάλληλους.

```python
def find_and_replace_json_special_columns(df):
    special_chars = set('{}[]:,\"')
    problematic_columns = [col for col in df.columns
                           if any(char in col for char in special_chars)]

    # Create a dictionary to map original column names to new column
    col_replacements = {}
    for col in problematic_columns:
        new_col = col
        for char in special_chars:
            new_col = new_col.replace(char, '_')
        col_replacements[col] = new_col

    # Rename columns in the DataFrame
    df.rename(columns=col_replacements, inplace=True)

    return col_replacements

# Find and replace columns with special JSON characters
col_replacements = find_and_replace_json_special_columns(df)

print("Columns with special JSON characters replaced:")
print(col_replacements)
print("\nUpdated DataFrame columns:")
print(df.columns.tolist())
```

Εικόνα 6-23: Συνάρτηση αντικατάστασης ειδικών χαρακτήρων

Η συνάρτηση αυτή εξασφαλίζει ότι όλες οι στήλες που περιέχουν χαρακτήρες που μπορεί να προκαλέσουν προβλήματα, μετατρέπονται σε κατάλληλες για χρήση στη συνέχεια της διαδικασίας προετοιμασίας των δεδομένων και της εκπαίδευσης των μοντέλων.

Έχοντας κατά νου ότι θα χρησιμοποιήσουμε τα μοντέλα **Logistic Regression**, **Random Forest**, **LightGBM** και **XGBoost**, η προετοιμασία των δεδομένων (συμπεριλαμβανομένης της διόρθωσης των ονομάτων των στηλών) θα διασφαλίσει ότι τα δεδομένα είναι συμβατά και κατάλληλα για κάθε μοντέλο. Αυτή η διαδικασία προετοιμασίας είναι κρίσιμη, ειδικά για τα μοντέλα **Logistic Regression** και **Random Forest**, τα οποία είναι πιο ευαίσθητα σε ακατάλληλα ονόματα στηλών και σε μη αριθμητικές ή μη κανονικοποιημένες τιμές.

Στην παρούσα ενότητα προετοιμασίας δεδομένων, υλοποιήθηκαν διαδικασίες για την αναγνώριση και διαχείριση ειδικών περιπτώσεων προβληματικών τιμών, όπως οι άπειρες τιμές (inf) και οι υπερβολικά μεγάλες αριθμητικές τιμές. Για τον λόγο αυτό, αναπτύχθηκαν και εφαρμόστηκαν συγκεκριμένες συναρτήσεις για την αντικατάσταση αυτών των τιμών και τη διαχείρισή τους.

6.9 Αντικατάσταση των άπειρων τιμών με NaN

Οι άπειρες τιμές (inf και -inf) μπορεί να εμφανιστούν λόγω υπολογιστικών σφαλμάτων ή μη ρεαλιστικών τιμών στα δεδομένα. Αυτές οι τιμές πρέπει να αντικατασταθούν, καθώς οι περισσότεροι αλγόριθμοι μηχανικής μάθησης δεν μπορούν να τις διαχειριστούν. Για τον σκοπό αυτό, χρησιμοποιήθηκε η συνάρτηση replace_inf_with_nan(), η οποία:

- Εντοπίζει και μετρά τον αριθμό των άπειρων τιμών στο σύνολο των δεδομένων.
- Αντικαθιστά τις άπειρες τιμές με NaN, ώστε να μπορέσουμε να τις διαχειριστούμε κατάλληλα στη συνέχεια (π.χ. μέσω της συμπλήρωσής τους με το μέσο όρο της στήλης).

```
# Function to identify and replace inf values, and count replacements
def replace_inf_with_nan(df):
    inf_count = np.isinf(df).sum().sum()
    df_replaced = df.replace([np.inf, -np.inf], np.nan)
    return df_replaced, inf_count

# Function to check and clip large values, and count clipped values
def clip_large_values(df, max_value=1e6):
    clipped_count = (df > max_value).sum().sum()
    df_clipped = df.clip(upper=max_value)
    return df_clipped, clipped_count
```

Εικόνα 6-24: Συναρτήσεις αντικατάστασης κενών και απεριόριστων τιμών

6.10 Περιορισμός υπερβολικά μεγάλων αριθμητικών τιμών (Clipping)

Υπερβολικά μεγάλες τιμές σε συγκεκριμένες στήλες μπορούν να αλλοιώσουν τις προβλέψεις των μοντέλων, ειδικά σε αλγορίθμους που είναι ευαίσθητοι σε μεγάλες διακυμάνσεις τιμών, όπως το Logistic Regression. Για την αντιμετώπιση αυτών των περιπτώσεων, χρησιμοποιήθηκε η συνάρτηση clip_large_values(), η οποία:

- Εντοπίζει τις τιμές που υπερβαίνουν ένα προκαθορισμένο ανώτατο όριο (εδώ ορίστηκε στο 1.000.000).
- Αντικαθιστά τις τιμές αυτές με το ανώτατο όριο που έχει οριστεί.

6.11 Διαχείριση κενών τιμών

Σε αυτό το σημείο το σύνολο δεδομένων μας είναι σχεδόν έτοιμο για να χρησιμοποιηθεί από τα μοντέλα μας. Αυτό που μένει είναι να δούμε πως θα διαχειριστούμε τις κενές τιμές NaN.

Στο πλαίσιο της διερευνητικής ανάλυσης, διαπιστώθηκε ότι αρκετά χαρακτηριστικά περιείχαν υψηλό ποσοστό κενών τιμών .Επίσης κάποια από τα χαρακτηριστικά που δημιουργήσαμε από τις συναθροίσεις ,όπως και οι Inf τιμές που αντικαταστήσαμε NaN δημιούργησαν μεγάλο αριθμό κενών τιμών.

```
REFUSED_RATE_DOWN_PAYMENT_MAX                                 85.234287
REFUSED_RATE_DOWN_PAYMENT_MEAN                                85.234287
REFUSED_AMT_DOWN_PAYMENT_MEAN                                 85.234287
REFUSED_AMT_DOWN_PAYMENT_MAX                                  85.234287
REFUSED_AMT_DOWN_PAYMENT_MIN                                  85.234287
REFUSED_RATE_DOWN_PAYMENT_MIN                                 85.234287
REFUSED_APP_CREDIT_PERC_VAR                                   83.658432
CC_C_C___A_M_T___P_A_Y_M_E_N_T___C_U_R_R_E_N_T___V_A_R         79.901249
CC_C_C___A_M_T___D_R_A_W_I_N_G_S___P_O_S___C_U_R_R_E_N_T___V_A_R 79.875986
CC_C_C___C_N_T___D_R_A_W_I_N_G_S___P_O_S___C_U_R_R_E_N_T___V_A_R 79.875986
CC_C_C___C_N_T___D_R_A_W_I_N_G_S___A_T_M___C_U_R_R_E_N_T___V_A_R 79.875986
CC_C_C___A_M_T___D_R_A_W_I_N_G_S___A_T_M___C_U_R_R_E_N_T___V_A_R 79.875986
CC_C_C___C_N_T___D_R_A_W_I_N_G_S___O_T_H_E_R___C_U_R_R_E_N_T___V_A_R 79.875986
CC_C_C___A_M_T___D_R_A_W_I_N_G_S___O_T_H_E_R___C_U_R_R_E_N_T___V_A_R 79.875986
CC_C_C___A_M_T___P_A_Y_M_E_N_T___C_U_R_R_E_N_T___M_A_X         79.755846
CC_C_C___A_M_T___P_A_Y_M_E_N_T___C_U_R_R_E_N_T___M_E_A_N       79.755846
CC_C_C___A_M_T___P_A_Y_M_E_N_T___C_U_R_R_E_N_T___M_I_N         79.755846
CC_C_C___C_N_T___D_R_A_W_I_N_G_S___O_T_H_E_R___C_U_R_R_E_N_T___M_I_N 79.735074
CC_C_C___A_M_T___D_R_A_W_I_N_G_S___O_T_H_E_R___C_U_R_R_E_N_T___M_I_N 79.735074
CC_C_C___C_N_T___D_R_A_W_I_N_G_S___A_T_M___C_U_R_R_E_N_T___M_I_N 79.735074
CC_C_C___C_N_T___D_R_A_W_I_N_G_S___A_T_M___C_U_R_R_E_N_T___M_A_X 79.735074
CC_C_C___C_N_T___D_R_A_W_I_N_G_S___A_T_M___C_U_R_R_E_N_T___M_E_A_N 79.735074
CC_C_C___A_M_T___D_R_A_W_I_N_G_S___A_T_M___C_U_R_R_E_N_T___M_I_N 79.735074
CC_C_C___A_M_T___D_R_A_W_I_N_G_S___A_T_M___C_U_R_R_E_N_T___M_A_X 79.735074
CC_C_C___A_M_T___D_R_A_W_I_N_G_S___A_T_M___C_U_R_R_E_N_T___M_E_A_N 79.735074
...
YEARS_BUILD_MODE                                              66.330761
YEARS_BUILD_MEDI                                              66.330761
```

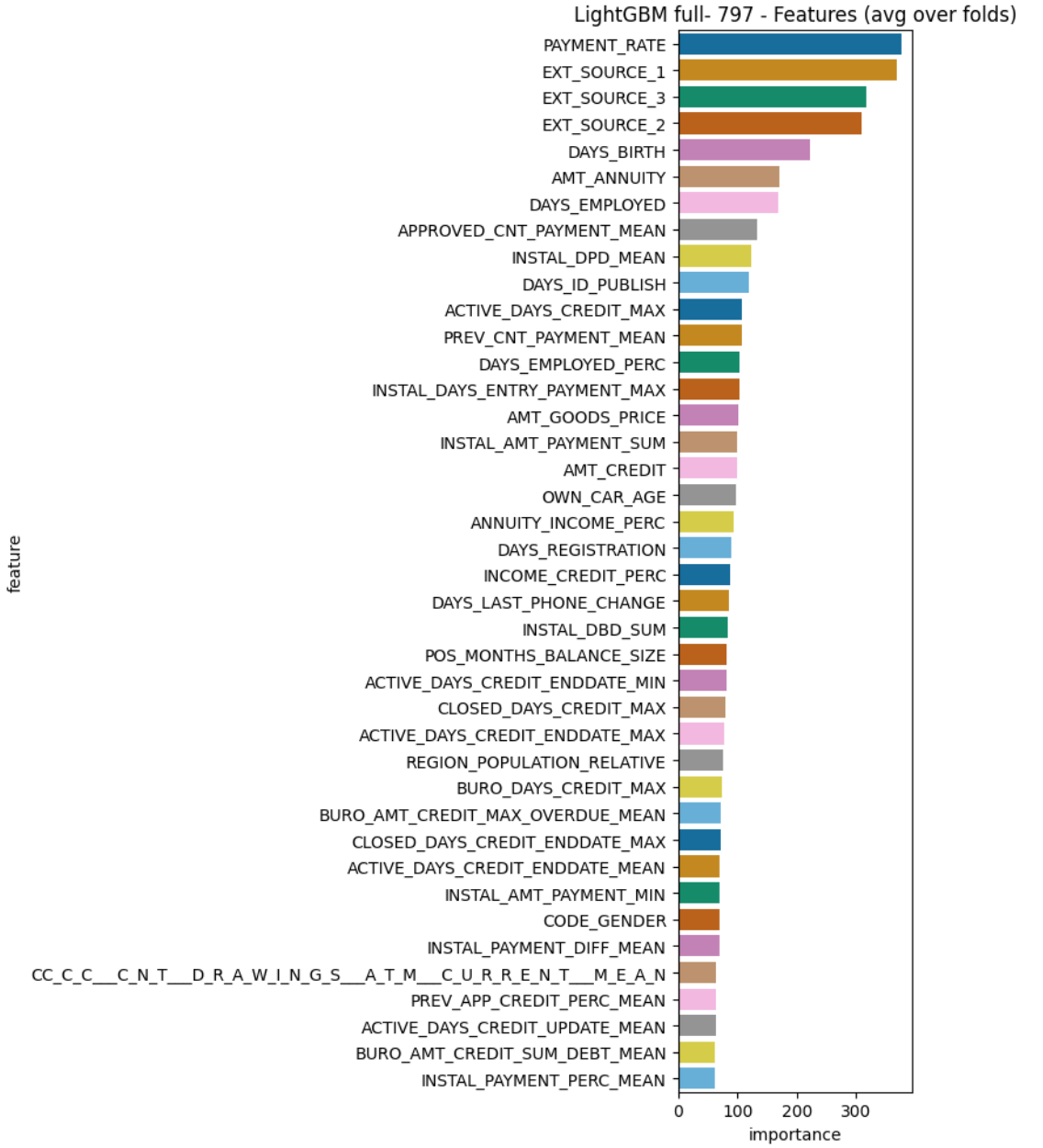
Εικόνα 6-25: Ποσοστά κενών τιμών στο τελικό σύνολο δεδομένων

Για να αντιμετωπιστεί αυτό το ζήτημα, αποφασίστηκε η αφαίρεση ενός μέρους αυτών των χαρακτηριστικών από το σύνολο δεδομένων. Συγκεκριμένα, τα χαρακτηριστικά που περιείχαν πάνω από το 65% των τιμών τους κενές αφαιρέθηκαν, καθώς θεωρήθηκαν μη αξιόπιστα για την εκπαίδευση των μοντέλων.

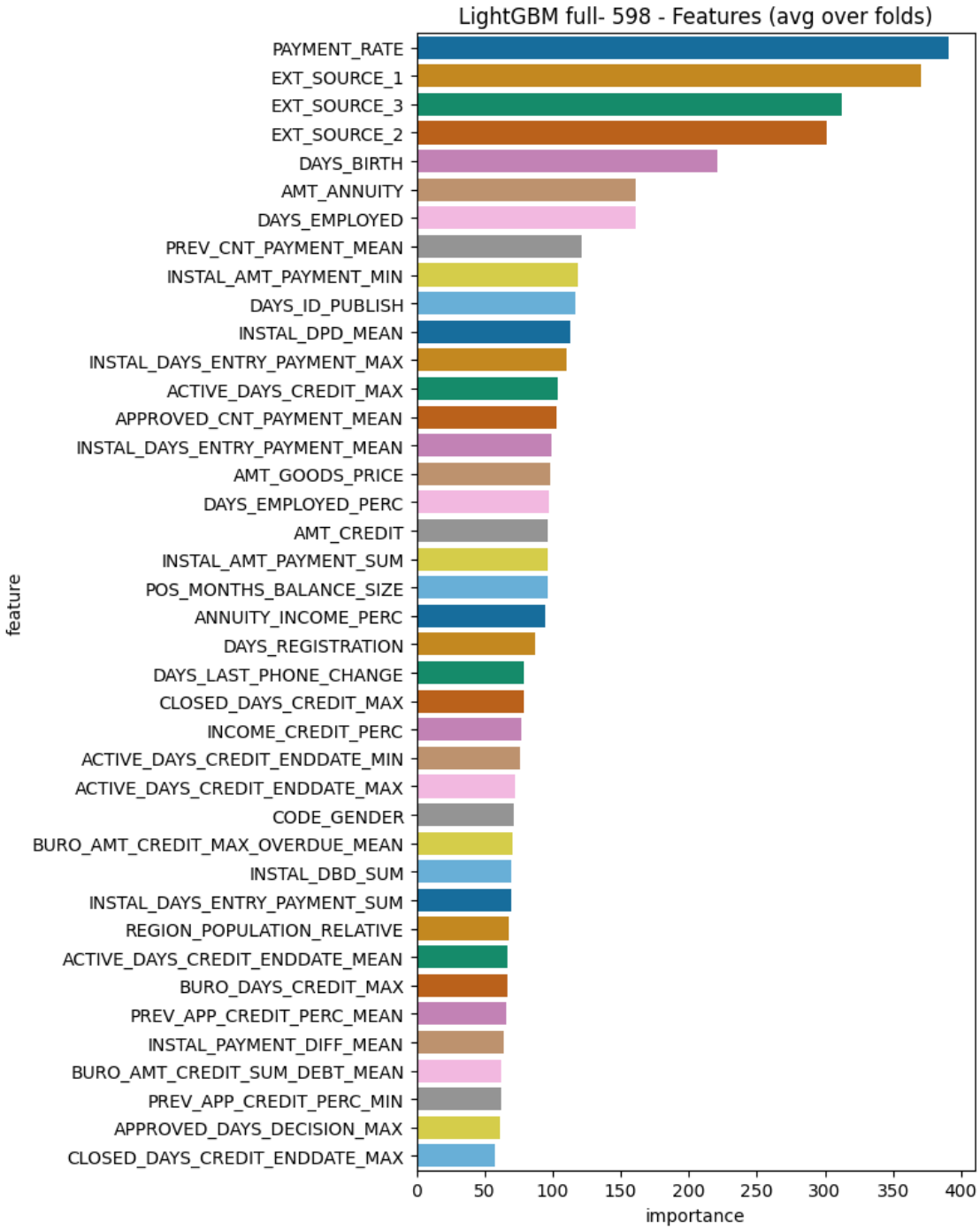
```python
# Find missing percentages
missing_percentage = df.isnull().mean() * 100
# Filter and sort the percentages from highest to lowest
missing_percentage = missing_percentage[missing_percentage > 65].sort_values(ascending=False)
# Display the percentage of missing
print(missing_percentage)
# Drop columns where the missing percentage is greater than the threshold
df_cleaned = df.drop(columns=missing_percentage.index)
```

Εικόνα 6-26: Αφαίρεση χαρακτηριστικών με περισσότερο από 65% κενές τιμές

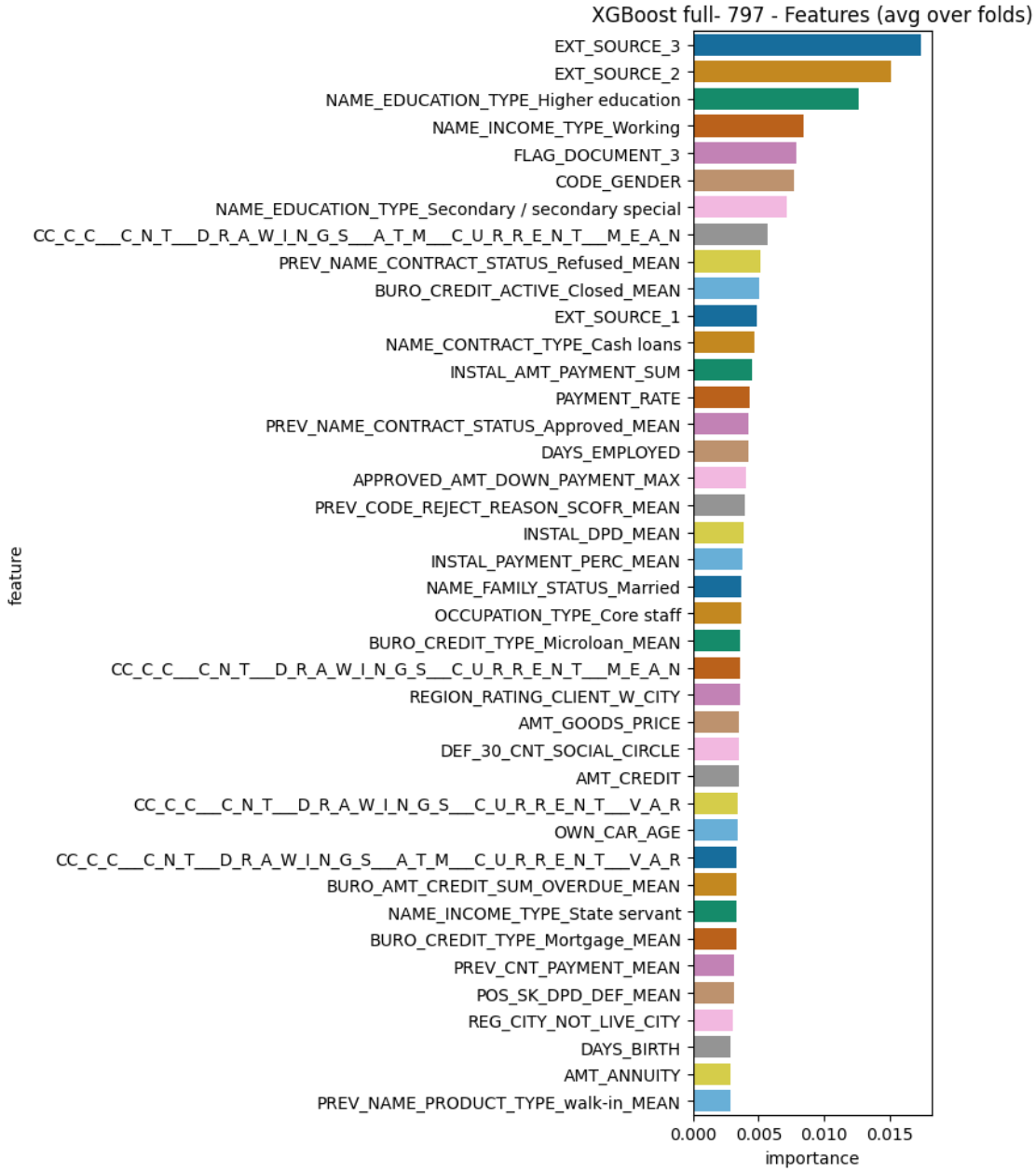
Για να δικαιολογήσουμε την απόφαση αυτή χρησιμοποιήσαμε τα μοντέλα **LightGBM, XGBoost** , τα οποία μπορούν να διαχειριστούν κενές τιμές κατά την εκπαίδευση τους και τα οποία έχουν την δυνατότητα να μας επιστρέψουν τον βαθμό σημαντικότητας (**feature importance**) κάθε χαρακτηριστικού που χρησιμοποιήθηκε. Από αυτή τη δοκιμή παρατηρήθηκε ότι τα χαρακτηριστικά αυτά δεν είχαν σημαντική συνεισφορά στη σημασία των χαρακτηριστικών κατά την εκπαίδευση του μοντέλου οπότε μπορούμε να τα αφαιρέσουμε.



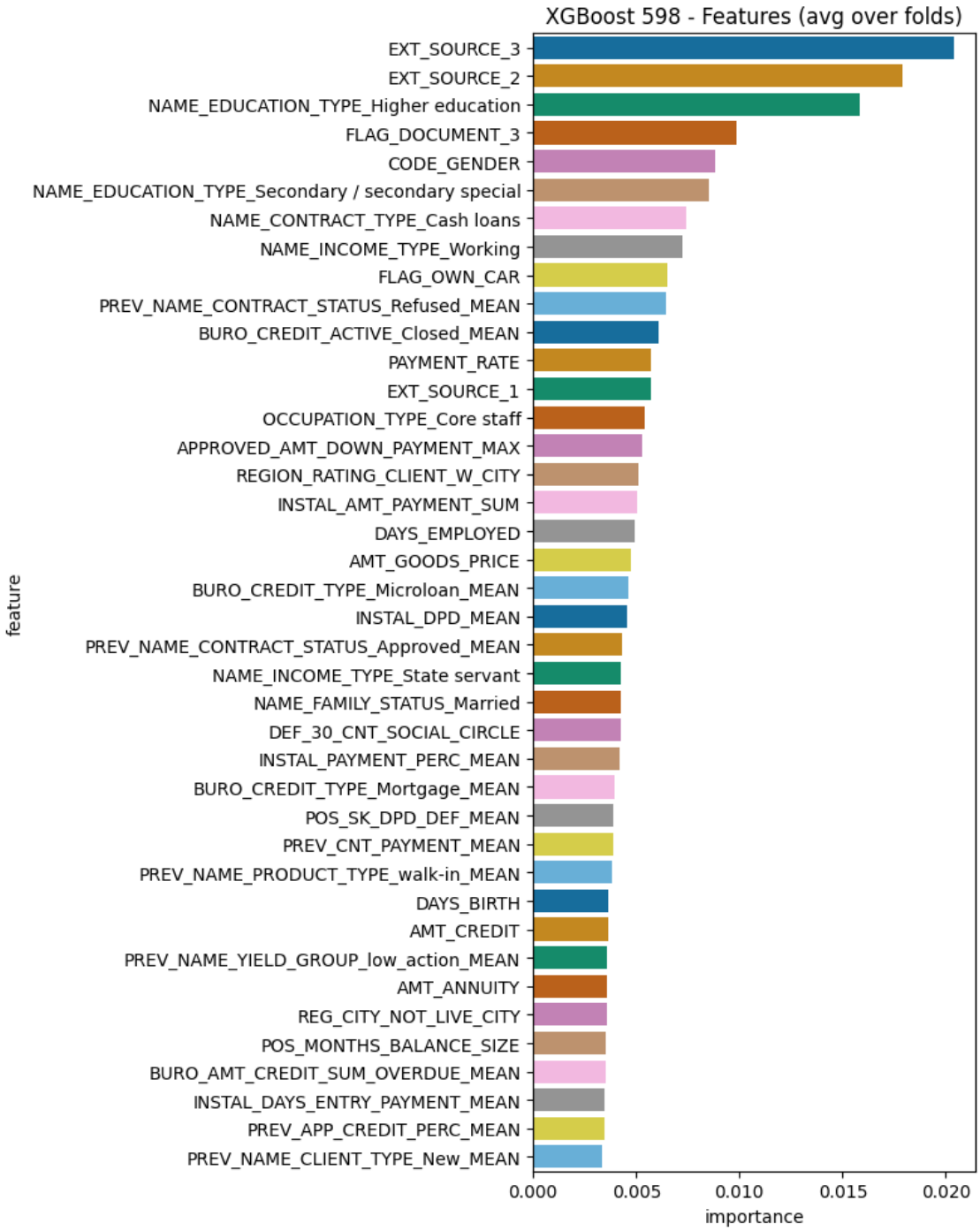
Εικόνα 6-27: Βαθμός σημαντικότητας των 797 χαρακτηριστικών με χρήση του Lightgbm



Εικόνα 6-28: Βαθμός σημαντικότητας των 598 χαρακτηριστικών με χρήση του Lightgbm



Εικόνα 6-29: Βαθμός σημαντικότητας των 797 χαρακτηριστικών με χρήση του XGBoost



Εικόνα 6-30: Βαθμός σημαντικότητας των 598 χαρακτηριστικών με χρήση του XGBoost

Με αυτόν τον τρόπο, διατηρούνται μόνο τα πιο αξιόπιστα χαρακτηριστικά για την εκπαίδευση, ενώ ταυτόχρονα απλοποιείται το σύνολο δεδομένων χωρίς να επηρεάζεται αρνητικά η ακρίβεια των μοντέλων.

Σε αυτό το σημείο έχουμε καταφέρει να μειώσουμε τον αριθμό των χαρακτηριστικών του συνόλου δεδομένων από 797 σε 598.

6.12 Συμπλήρωση Κενών Τιμών και Κανονικοποίηση (imputation and scaling)

Στην συνέχεια θα συμπληρώσουμε τις υπόλοιπες κενές τιμές(Imputation) με τη μέση τιμή της στήλης τους (median) και θα κανονικοποιήσουμε όλες τις τιμές (scaling) μεταξύ του εύρους τιμών [0-1] ώστε να φέρουμε όλα τα χαρακτηριστικά στην ίδια κλίμακα, διότι το μοντέλο της λογιστικής παλινδρόμησης είναι ευαίσθητο σε διαφορετικές κλίμακες χαρακτηριστικών και μπορεί να επηρεαστεί από τιμές με μεγάλη διακύμανση.

AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
0.181506	0.378636	0.090032	0.323606
0.250783	1.000000	0.132924	1.000000
0.042952	0.094241	0.020025	0.098489
0.112229	0.280296	0.109477	0.267327
0.098373	0.490052	0.078975	0.492444

Εικόνα 6-31: Παράδειγμα κανονικοποιημένων τιμών στο εύρος [0-1]

Με αυτόν τον τρόπο ολοκληρώνεται η προετοιμασία του συνόλου δεδομένων και είναι έτοιμο για την εκπαίδευση των μοντέλων.

7 Αποτελέσματα

7.1 Πρώτη δοκιμή των Μοντέλων

Μετά την ολοκλήρωση της διαδικασίας προετοιμασίας των δεδομένων, τα χαρακτηριστικά είναι πλέον έτοιμα για την πρώτη δοκιμή στα μοντέλα. Αρχικά, θα πραγματοποιηθεί εκπαίδευση των μοντέλων Logistic Regression, Random Forest, LightGBM και XGBoost χρησιμοποιώντας τα προεπεξεργασμένα δεδομένα, χωρίς βελτιστοποίηση υπερπαραμέτρων (unoptimized models). Για την αξιολόγηση των αποτελεσμάτων, θα χρησιμοποιηθεί η μεθοδολογία της Stratified K-Fold Cross-Validation, ώστε να διατηρηθεί η ισορροπία των κατηγοριών κατά την εκπαίδευση και να εκτιμηθεί η απόδοση των μοντέλων με ακρίβεια. Οι παράμετροι που θα εξεταστούν περιλαμβάνουν την ROC AUC, τον πίνακα σύγχυσης (Confusion Matrix), και την αναφορά ταξινόμησης (Classification Report) η οποία περιλαμβάνει τα Precision, Recall, F1-score και Accuracy παρέχοντας μια πλήρη εικόνα της απόδοσης κάθε μοντέλου πριν από την εφαρμογή πιο προηγμένων τεχνικών βελτιστοποίησης και αντιμετώπισης της ανισορροπίας στις κατηγορίες. Ακολουθούν οι συγκεντρωτικοί πίνακες αποτελεσμάτων της πρώτης δοκιμής με 598 χαρακτηριστικά.

AUC	Logistic Regression		Random Forest		XGBoost		LightGBM	
Fold	Train	Valid	Train	Valid	Train	Valid	Train	Valid
0	0.77260	0.76332	1.0	0.71137	0.86606	0.76781	0.86415	0.78467
1	0.77168	0.77094	1.0	0.71335	0.86646	0.77653	0.85724	0.78310
2	0.77154	0.77222	1.0	0.72451	0.85647	0.77558	0,85989	0.79614
3	0.77334	0.76652	1.0	0.71648	0.88533	0.76812	0.83757	0.79065
4	0.77158	0.77082	1.0	0.72105	0.86239	0.77622	0.84901	0.78444
5	0.77160	0.77385	1.0	0.71923	0.86970	0.78330	0.87488	0.78239
6	0.77274	0.76432	1.0	0.70835	0.87468	0.77489	0.87716	0.78809
7	0.77182	0.76368	1.0	0.72513	0.86650	0.77577	0.84043	0.78725
8	0.77206	0.77027	1.0	0.71838	0.87136	0.77868	0.88973	0.78546
9	0.77166	0.76483	1.0	0.71553	0.86073	0.77327	0.87873	0.78341
overall	0.77206	0.76805	1.0	0.71733	0.86797	0.77493	0.86288	0.78631

Πίνακας 7-1: AUC της δοκιμής με 598 για κάθε βήμα της επικύρωσης

Model	Category	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	0	0.92	1.00	0.96	0.9194	0.7681
	1	0.51	0.03	0.05		
Random Forest	0	0.92	1.00	0.96	0.9194	0.7173
	1	0.73	0.00	0.003		
XGBoost	0	0.92	0.99	0.96	0.9190	0.7749
	1	0.49	0.06	0.10		
LightGBM	0	0.96	0.75	0.85	0.7479	0.7863
	1	0.19	0.67	0.30		

Πίνακας 7-2: Αποτελέσματα με 598 χαρακτηριστικά

Model	TP	TN	FP	FN
Logistic Regression	282048	664	24161	634
Random Forest	282665	46	24779	17
XGBoost	281181	1429	23396	1501
LightGBM	213266	16720	8105	69416

Πίνακας 7-3: Πίνακας σύγχυσης με 598 χαρακτηριστικά

Παρακάτω ακολουθεί η ανάλυση των αποτελεσμάτων:

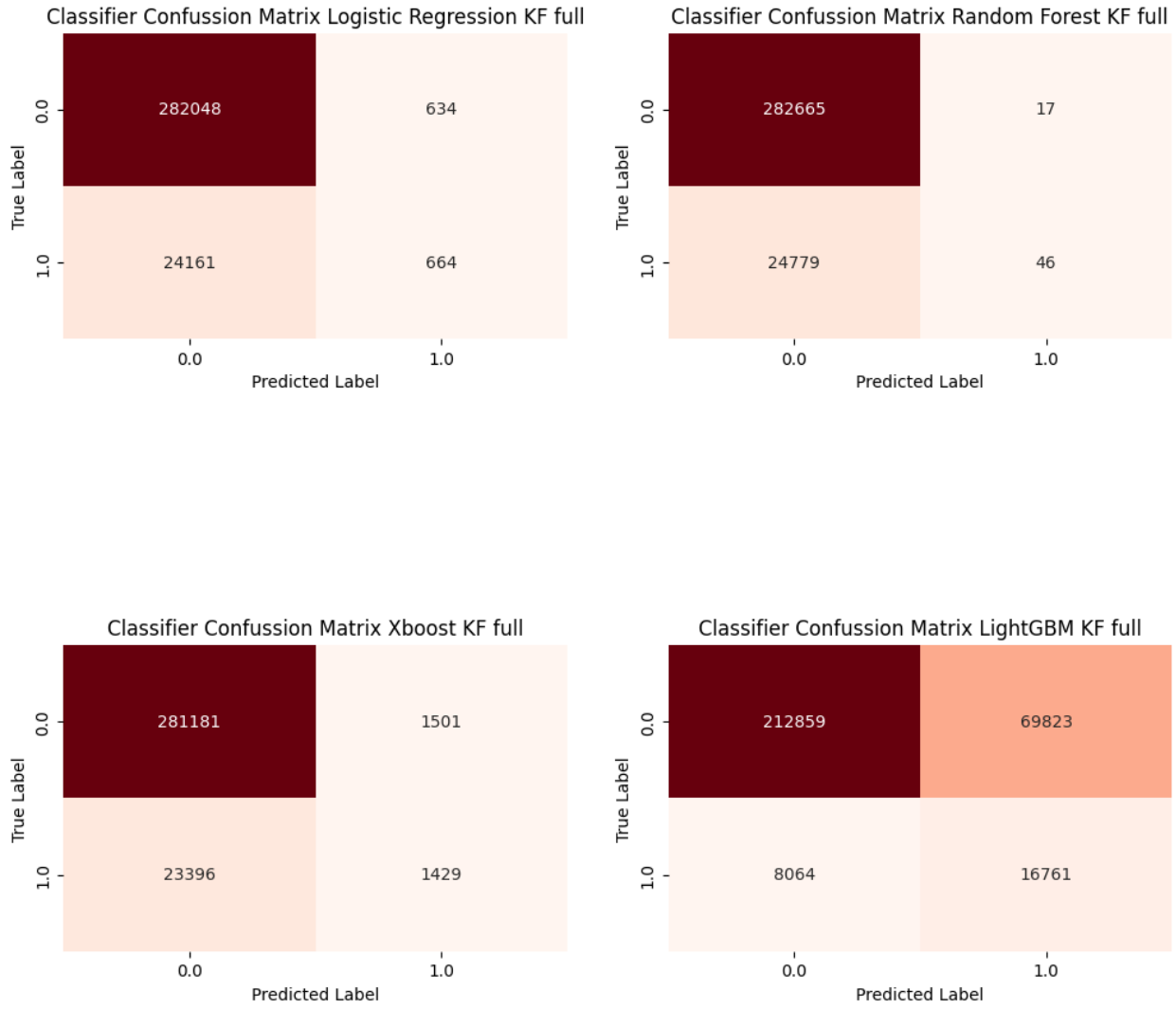
1. Classification Report

- **Logistic Regression**: Εμφανίζει υψηλό precision και recall για την κατηγορία 0 κάτι που μας δείχνει ότι αναγνώρισε σωστά 92% των προβλέψεων για αυτή την κατηγορία και ότι σχεδόν είχε 100% ποσοστό επιτυχίας σε αυτές. Ωστόσο, στην κατηγορία 1 μόνο το 51% των προβλέψεων ήταν σωστές με 664 TN και 634 FN ενώ από αυτές μόνο το 3% προβλέφθηκε σωστά κάτι που αντικατοπτρίζεται από τον μεγάλο αριθμό 24.161 FP. Τέλος το F1-Score (0.05), είναι πάρα πολύ χαμηλό πράγμα που δείχνει ότι το μοντέλο έχει πολύ κακή απόδοση στην ανίχνευση των περιπτώσεων της κατηγορίας 1.

- **Random Forest**: Εμφανίζει τα ίδια αποτελέσματα στην κατηγορία 0. Στην κατηγορία 1 το precision μας δείχνει ότι 71% των προβλέψεων ανήκαν πράγματι σε αυτή με 46 TN και 17 FN, όμως το μηδενικό recall μας δείχνει ότι δεν κατάφερε να αναγνωρίσει σχεδόν το 100% αυτής της κατηγορίας με 24.779 FP. Επίσης το σχεδόν μηδενικό F1-Score(0.003) επιβεβαιώνει ότι το μοντέλο αδυνατεί να προβλέψει την κατηγορία 1.

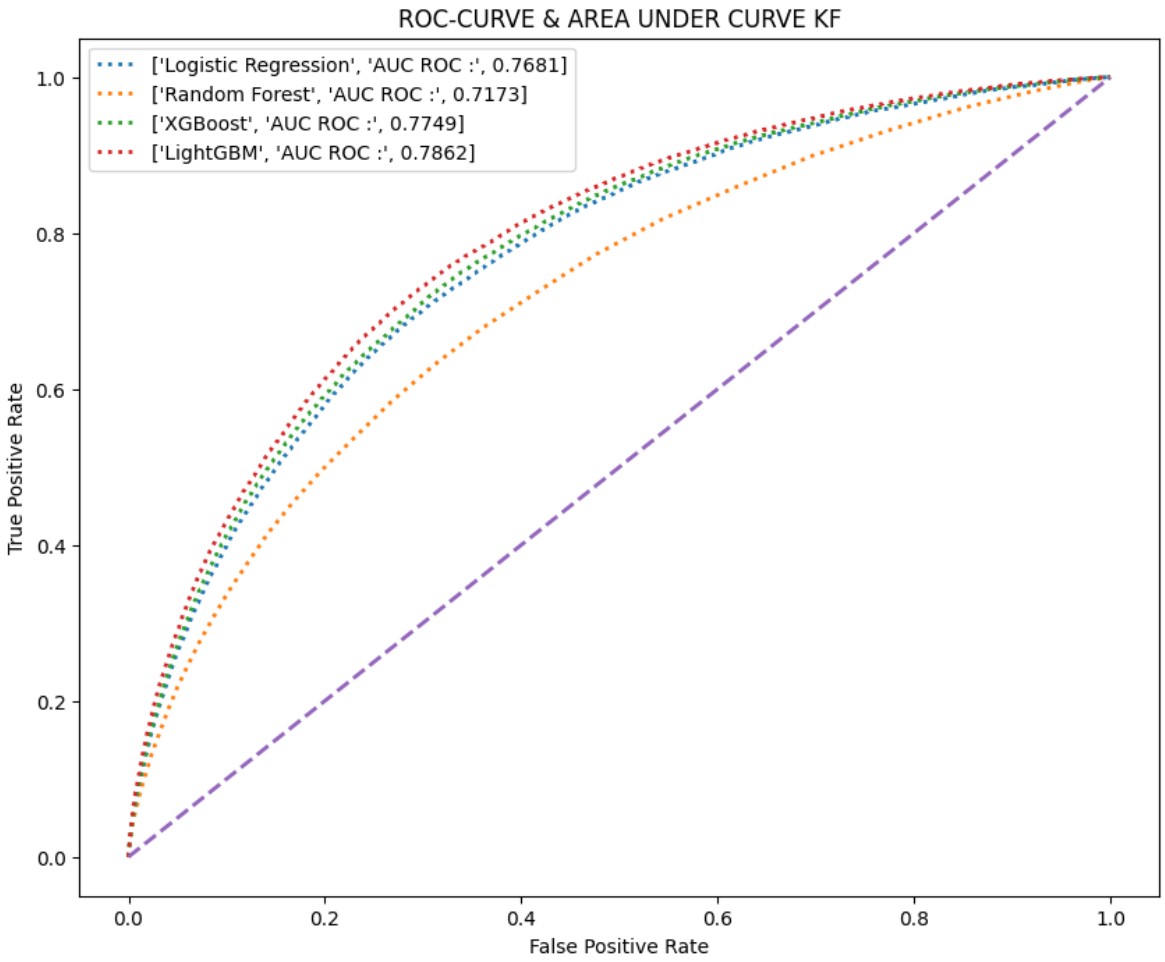
- **XGBoost**: Το μοντέλο τα πήγε καλύτερα από το Random Forest και το Logistic Regression με F1-Score 0.10 για την κατηγορία 1 παραμένοντας όμως ακόμα πολύ χαμηλό. Το precision μας δείχνει ότι 49% των προβλέψεων, 1429 αντί 1501, για την κατηγορία 1 ανήκαν πράγματι σε αυτή και το recall μας δείχνει ότι κατάφερε να αναγνωρίσει σωστά μόνο το 6% αυτής της κατηγορίας με 23.396 FP.

- **LightGBM**: Το LightGBM είχε το καλύτερο συνολικό αποτέλεσμα με F1-Score 0.30 για την κατηγορία 1. Το precision είναι χαμηλότερο από τα προηγούμενα μοντέλα με μόνο το 19% των προβλέψεων να ανήκουν πράγματι στην κατηγορία 1 με 16.720 TN κάτι που σημαίνει ότι δημιουργεί πολλά FN 69.416. Το recall όμως είναι αισθητά υψηλότερο δείχνοντας ότι έχει 67% ποσοστό επιτυχίας σε αυτά που αναγνώρισε κάτι που σημαίνει μείωση των FP στα 8.105.



Εικόνα 7-1: Πίνακες σύγχυσης με 598 χαρακτηριστικά

2. ROC-AUC



Εικόνα 7-2: Καμπύλες ROCAUC με 598 χαρακτηριστικά

- **Logistic Regression**: Το AUC για το Logistic Regression ήταν 0.7681 στο validation set, γεγονός που δείχνει μέτρια ικανότητα του μοντέλου να διαχωρίσει τις δύο κατηγορίες.
- **Random Forest**: Παρά το υψηλό AUC στην εκπαίδευση (1.0), το AUC στο validation set ήταν αρκετά χαμηλό (0.7173), το οποίο υποδηλώνει overfitting στα δεδομένα της κατηγορίας 0.
- **XGBoost**: Το AUC του XGBoost ήταν 0.7749 στο validation set, δείχνοντας καλύτερη ικανότητα διαχωρισμού από το Logistic Regression και το Random Forest.
- **LightGBM**: Έιχε το καλύτερο AUC με 0.7862, κάτι που υποδεικνύει ότι είναι το ισχυρότερο μοντέλο σε αυτό το στάδιο της διαδικασίας. Αυτό δείχνει την ικανότητά του να διαχωρίσει καλύτερα τις κατηγορίες σε σύγκριση με τα υπόλοιπα μοντέλα.

3. Accuracy

- Αν και η ακρίβεια accuracy ήταν υψηλή για όλα τα μοντέλα, αυτό οφείλεται κυρίως στην κυριαρχία της κατηγορίας 0 (καλοπληρωτές) στο dataset. Τα μοντέλα κατάφεραν να προβλέψουν σωστά την πλειοψηφία των περιπτώσεων, δηλαδή τους καλοπληρωτές, αλλά απέτυχαν να ανιχνεύσουν την κατηγορία 1 (κακοπληρωτές). Συνεπώς, η υψηλή ακρίβεια δεν αποτελεί καλή ένδειξη απόδοσης σε αυτό το πρόβλημα με μεγάλη ανισορροπία.

Η ανάλυση των αποτελεσμάτων των τεσσάρων μοντέλων παρουσιάζει μια σημαντική πρόκληση, την ανισορροπία των κατηγοριών. Στην παρούσα μελέτη, η συντριπτική πλειονότητα των περιπτώσεων ανήκει στην κατηγορία 0 (μη προβληματικές περιπτώσεις), ενώ η κατηγορία 1 (προβληματικές περιπτώσεις) αποτελεί ένα μικρό ποσοστό των δεδομένων. Αυτό δημιουργεί προβλήματα στην ακρίβεια της πρόβλεψης των μοντέλων για την κατηγορία 1, κάτι που αποδεικνύεται από τις χαμηλές τιμές Precision, Recall, και F1-score στην κατηγορία αυτή.

Το LightGBM ξεχωρίζει ως το πιο ισορροπημένο μοντέλο, καθώς προσφέρει τον υψηλότερο AUC και καλύτερη απόδοση στην κατηγορία 1. Αντίθετα, το Logistic Regression, το Random Forest και XGBoost παρουσιάζουν προβλήματα στην κατηγορία 1 και αποδίδουν καλά μόνο στην κατηγορία 0.

7.2 Δεύτερη δοκιμή των Μοντέλων - FEATURE SELECTION

Αφού πραγματοποιήθηκε η αρχική δοκιμή με τα τέσσερα μοντέλα ,διαπιστώθηκε ότι το LightGBM παρουσίασε τα καλύτερα αποτελέσματα στις παραμέτρους αξιολόγησης. Λόγω αυτής της απόδοσης, επιλέχθηκε να χρησιμοποιηθεί για τον υπολογισμό της αθροιστικής σημαντικότητας των χαρακτηριστικών (cumulative feature importance), προκειμένου να μειωθεί περαιτέρω ο αριθμός των χαρακτηριστικών που θα χρησιμοποιηθούν στην τελική εκπαίδευση των μοντέλων.

	feature	importance	contribution_percentage
363	PAYMENT_RATE	428.8	3.465051
170	EXT_SOURCE_1	357.2	2.886465
171	EXT_SOURCE_2	307.4	2.484040
172	EXT_SOURCE_3	299.7	2.421818
154	DAYS_BIRTH	261.2	2.110707
22	AMT_ANNUITY	176.3	1.424646
157	DAYS_ID_PUBLISH	139.9	1.130505
155	DAYS_EMPLOYED	134.1	1.083636
219	INSTAL_AMT_PAYMENT_MIN	128.6	1.039192
410	PREV_CNT_PAYMENT_MEAN	124.6	1.006869
229	INSTAL_DPD_MEAN	119.1	0.962424
159	DAYS_REGISTRATION	114.2	0.922828
222	INSTAL_DAYS_ENTRY_PAYMENT_MAX	113.3	0.915556
32	ANNUITY_INCOME_PERC	109.0	0.880808
55	APPROVED_CNT_PAYMENT_MEAN	107.9	0.871919
23	AMT_CREDIT	107.2	0.866263
24	AMT_GOODS_PRICE	105.9	0.855758
158	DAYS_LAST_PHONE_CHANGE	104.9	0.847677
223	INSTAL_DAYS_ENTRY_PAYMENT_MEAN	104.0	0.840404
16	ACTIVE_DAYS_CREDIT_MAX	103.5	0.836364
367	POS_MONTHS_BALANCE_SIZE	99.3	0.802424
212	INCOME_CREDIT_PERC	99.0	0.800000
156	DAYS_EMPLOYED_PERC	97.5	0.787879
220	INSTAL_AMT_PAYMENT_SUM	90.7	0.732929
145	CLOSED_DAYS_CREDIT_MAX	85.6	0.691717
15	ACTIVE_DAYS_CREDIT_ENDDATE_MIN	82.6	0.667475
571	REGION_POPULATION_RELATIVE	82.4	0.665859
227	INSTAL_DBD_SUM	81.8	0.661010
153	CODE_GENDER	81.6	0.659394
57	APPROVED_DAYS_DECISION_MAX	76.1	0.614949

Εικόνα 7-3: Βαθμός αθροιστικής σημαντικότητας των χαρακτηριστικών

Αφού υπολογίσαμε τη αθροιστική σημαντικότητα των χαρακτηριστικών κρατήσαμε αυτά τα οποία συνεισφέρουν στο 95% της απόδοσης του μοντέλου με αποτέλεσμα τη μείωση τους στα 289 από τα 598.

```
# Calculate Cumulative Importance
feat_x['Cumulative_Importance'] = feat_x['contribution_percentage'].cumsum()

#  Determine the Threshold
threshold = 95  # 95% cumulative importance

#  Select Features Based on Threshold
features_to_drop = feat_x[feat_x['Cumulative_Importance'] > threshold]['feature']

#Strip Whitespace from Column Names
features_to_drop = features_to_drop.str.strip()

#Drop Features
df2 = df.drop(columns=features_to_drop)
```

Εικόνα 7-4: Αφαίρεση χαρακτηριστικών κάτω του 95% συνεισφοράς

Η διαδικασία αυτή επιτρέπει την επιλογή των πιο σημαντικών χαρακτηριστικών, διατηρώντας μόνο εκείνα που συνεισφέρουν σημαντικά στην απόδοση του μοντέλου, ενώ τα υπόλοιπα χαρακτηριστικά αφαιρούνται. Με αυτό τον τρόπο, μειώνεται η πολυπλοκότητα του μοντέλου, χωρίς να επηρεάζεται αρνητικά η ακρίβεια των προβλέψεων.

Αφού αφαιρέθηκαν τα χαρακτηριστικά που δεν συνέβαλαν σημαντικά στην απόδοση του μοντέλου, εκπαιδεύσαμε και πάλι τα τέσσερα μοντέλα με τα 289 επιλεγμένα χαρακτηριστικά.

Ακολουθούν οι συγκεντρωτικοί πίνακες αποτελεσμάτων για τα μοντέλα της δοκιμής με 289 χαρακτηριστικά.

AUC	Logistic Regression		Random Forest		XGBoost		LightGBM	
Folds	Train	Valid	Tr	Valid	Train	Valid	Train	Valid
0	0.76927	0.76082	1.0	0.71264	0.87742	0.77188	0.78562	0.87752
1	0.76890	0.76899	1.0	0.72140	0.84682	0.77710	0.78229	0.87636
2	0.76868	0.77192	1.0	0.72679	0.85875	0.78054	0.79594	0.87840
3	0.76838	0.76438	1.0	0.71620	0.85909	0.76804	0.79173	0.84349
4	0.76813	0.77040	1.0	0.72158	0.86082	0.77799	0.78442	0.85847
5	0.76821	0.77080	1.0	0.72352	0.84786	0.78205	0.78224	0.85247
6	0.76892	0.76172	1.0	0.71934	0.86173	0.77473	0.78838	0.83802
7	0.76853	0.76298	1.0	0.72494	0.87345	0.77861	0.78785	0.87435
8	0.76778	0.76834	1.0	0.72249	0.86911	0.77984	0.78566	0.89579
9	0.76933	0.76315	1.0	0.71338	0.85722	0.77199	0.78352	0.87654
Overall	0.76861	0.76635	1.0	0.72022	0.86123	0.77619	0.78676	0.78653

Πίνακας 7-4: AUC της δοκιμής με 289 χαρακτηριστικά για κάθε βήμα της επικύρωσης

Μοντέλο	Κατηγορία	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	0	0.92	1.00	0.96	0.9194	0.7664
	1	0.52	0.02	0.05		
Random Forest	0	0.92	1.00	0.96	0.9194	0.7202
	1	0.73	0.00	0.00		
XGBoost	0	0.92	0.99	0.96	0.9190	0.7762
	1	0.49	0.06	0.10		
LightGBM	0	0.96	0.76	0.85	0.7498	0.7865
	1	0.19	0.67	0.30		

Πίνακας 7-5: Αποτελέσματα με 289 χαρακτηριστικά

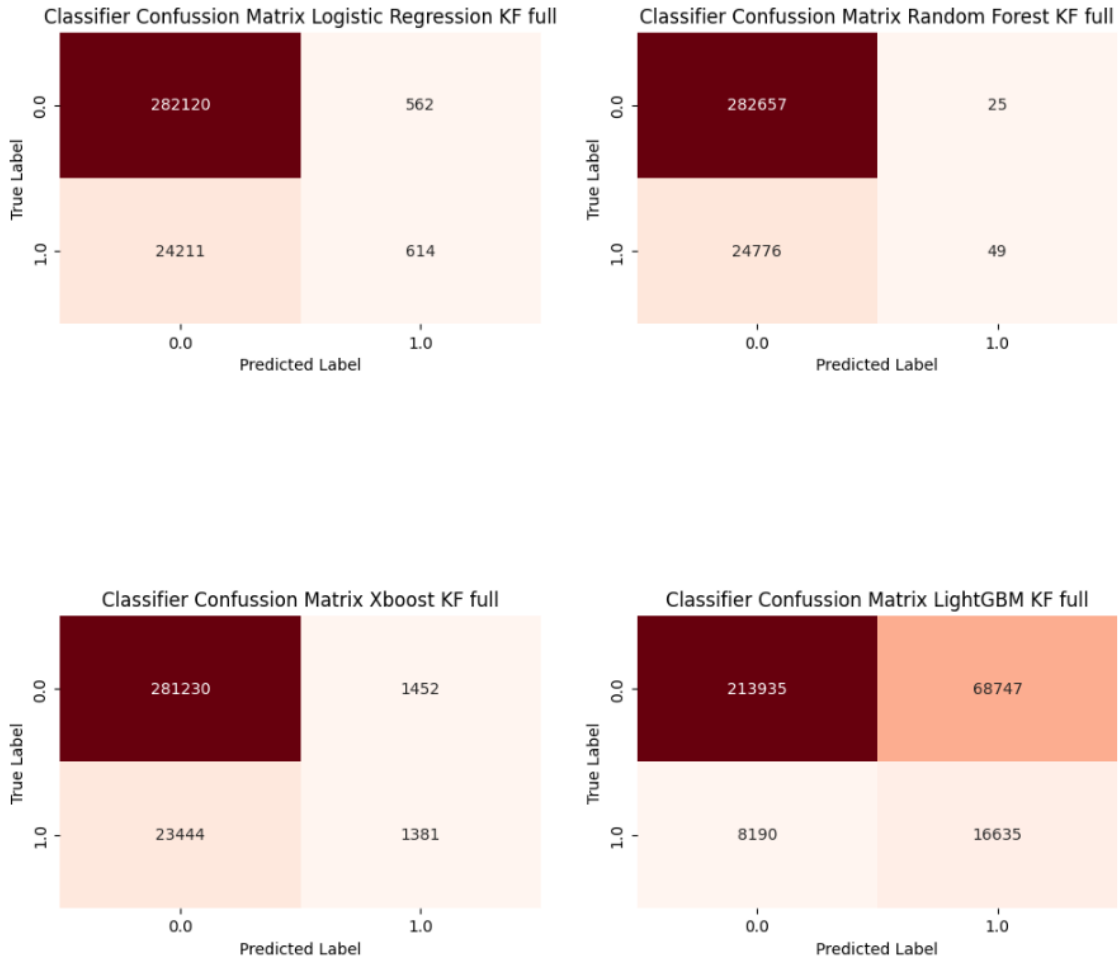
Μοντέλο	TP	TN	FP	FN
Logistic Regression	282120	614	24211	562
Random Forest	282657	49	24776	25
XGBoost	281230	1381	23444	1452
LightGBM	213935	16635	8190	68747

Πίνακας 7-6: Πίνακας σύγχυσης με 289 χαρακτηριστικά

Παρακάτω ακολουθεί η ανάλυση των αποτελεσμάτων:

1.Classification Report και Confusion Matrix

Από τη σύγκριση των αποτελεσμάτων μεταξύ της πρώτης και της δεύτερης δοκιμής, παρατηρούμε ότι τα αποτελέσματα στο classification report παραμένουν σταθερά, χωρίς ουσιαστικές διαφορές. Οι παράμετροι αξιολόγησης όπως το Precision, το Recall και το F1-score για όλες τις κατηγορίες παρουσιάζουν μικρές μόνο αποκλίσεις. Στο confusion matrix, παρατηρούμε επίσης ότι οι True Positives (TP) και True Negatives (TN) παραμένουν σχεδόν σταθερά, ενώ οι διαφορές στους False Positives (FP) και False Negatives (FN) είναι αμελητέες. Αυτό δείχνει ότι το μοντέλο, μετά την αφαίρεση των χαρακτηριστικών, συνεχίζει να κάνει τις ίδιες προβλέψεις και να αντιμετωπίζει παρόμοια προβλήματα στην ανίχνευση της κατηγορίας 1.

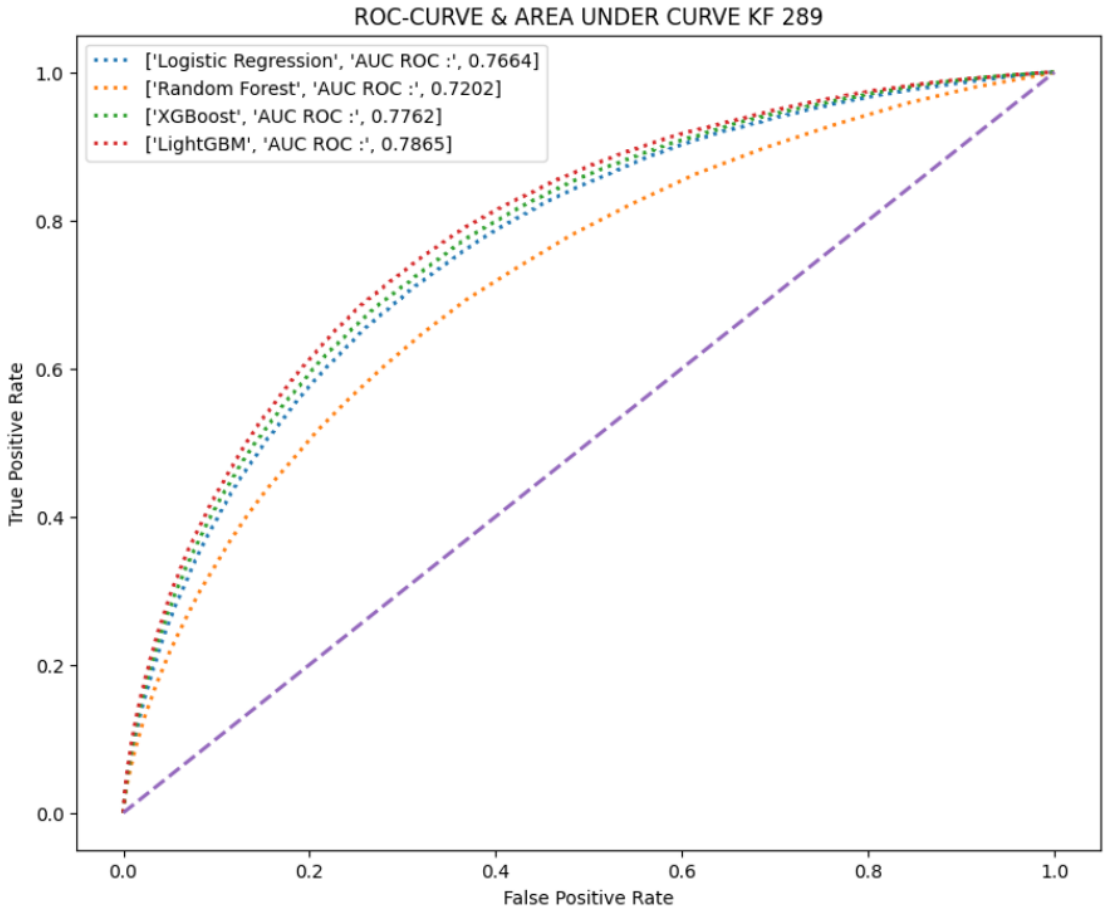


Εικόνα 7-5: Πίνακες σύγχυσης με 289 χαρακτηριστικά

2.ROC-AUC

Όσον αφορά τη ROC-AUC scores, και πάλι παρατηρούμε μικρές διαφορές που δεν επηρεάζουν ουσιαστικά την ικανότητα των μοντέλων να διαχωρίζουν τις δύο κατηγορίες. Συγκεκριμένα, τα AUC παραμένουν σχεδόν τα ίδια, γεγονός που επιβεβαιώνει ότι η αφαίρεση των χαρακτηριστικών δεν επιδρά σημαντικά στην απόδοση των μοντέλων.

- **Logistic Regression:** Ελάχιστη αλλαγή (0.7681 σε 0.7664).
- **Random Forest:** Μικρή βελτίωση (0.7173 σε 0.7202), αλλά εξακολουθεί να μην είναι ικανοποιητικό.
- **XGBoost:** Ελάχιστη βελτίωση (0.7749 σε 0.7762).
- **LightGBM:** Παρόμοιο AUC(0.7862 σε 0.7865.)



Εικόνα 7-6: Καμπύλες ROC-AUC με 289 χαρακτηριστικά

Η μείωση των χαρακτηριστικών από 598 σε 289 δεν βελτίωσε σημαντικά την απόδοση των μοντέλων. Αυτό είναι λογικό, δεδομένου ότι ο όγκος των χαρακτηριστικών που αφαιρέθηκαν ήταν είτε μηδενικής είτε πολύ χαμηλής σημαντικότητας, σύμφωνα με τη συνεισφορά τους στη αθροιστική σημαντικότητα. Η αφαίρεση αυτών των χαρακτηριστικών είχε σκοπό να μειώσει την πολυπλοκότητα των δεδομένων χωρίς να θυσιάσει την απόδοση των μοντέλων, όπως φαίνεται και από τα αποτελέσματα. Τα Precision, Recall και F1-Score παρέμειναν παρόμοια, ιδίως στην κατηγορία 1, όπου τα περισσότερα μοντέλα εξακολουθούν να δυσκολεύονται να προβλέψουν αυτούς που αθετούν.

7.3 Τρίτη δοκιμή των μοντέλων– PCA Feature Selection

Μετά την αφαίρεση των χαρακτηριστικών που αντιστοιχούν σε 95% της αθροιστικής σημαντικότητας χρησιμοποιήσαμε την τεχνική PCA (Principal Component Analysis) για να μειώσουμε περαιτέρω τις διαστάσεις του συνόλου δεδομένων και να διατηρήσουμε το 95% της συνολικής διακύμανσης των δεδομένων. Αυτή η διαδικασία μείωσε των αριθμό των χαρακτηριστικών ακόμη περισσότερο καταλήγοντας σε 95. Στη συνέχεια εκτελέσαμε εκ νέου δοκιμές με τα τέσσερα μοντέλα.

Ακολουθούν οι συγκεντρωτικοί πίνακες αποτελεσμάτων για τα μοντέλα της δοκιμής με εφαρμογή PCA.

AUC	Logistic Regression		Random Forest		XGBoost		LightGBM	
Folds	Train	Valid	Tr	Valid	Train	Valid	Train	Valid
0	0.75571	0.74926	1.0	0.67682	0.83979	0.73628	0.82431	0.74398
1	0.75462	0.75871	1.0	0.68448	0.83761	0.74484	0.82963	0.75099
2	0.75514	0.75471	1.0	0.68432	0.83861	0.74186	0.82240	0.75895
3	0.75554	0.75078	1.0	0.67835	0.85124	0.73909	0.82289	0.76039
4	0.75467	0.75787	1.0	0.68349	0.82181	0.74402	0.83056	0.75385
5	0.75474	0.75757	1.0	0.68521	0.83502	0.74268	0.83878	0.74423
6	0.75540	0.75172	1.0	0.67584	0.83914	0.73653	0.84256	0.75577
7	0.75528	0.75309	1.0	0.68171	0.82658	0.73726	0.82031	0.75612
8	0.75493	0.75618	1.0	0.67729	0.84384	0.74123	0.83436	0.74972
9	0.75545	0.75066	1.0	0.67727	0.82672	0.73418	0.81791	0.75199
Overall	0.75515	0.75404	1.0	0.68048	0.83604	0.73973	0.82837	0.75251

Πίνακας 7-7: AUC της δοκιμής με χρήση PCA για κάθε βήμα της επικύρωσης

Μοντέλο	Κατηγορία	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	0	0.92	1.00	0.96	0.9193	0.7540
	1	0.50	0.02	0.03		
Random Forest	0	0.92	1.00	0.96	0.9193	0.6805
	1	0.80	0.00	0.00		
XGBoost	0	0.92	1.00	0.96	0.9189	0.7397
	1	0.45	0.02	0.04		
LightGBM	0	0.96	0.73	0.83	0.7194	0.7525
	1	0.17	0.65	0.27		

Πίνακας 7-8: Αποτελέσματα με χρήση PCA

Μοντέλο	TP	TN	FP	FN
Logistic Regression	282289	395	24430	393
Random Forest	282680	9	24816	2
XGBoost	282027	545	24280	655
LightGBM	205189	16033	8792	77493

Πίνακας 7-9: Πίνακας σύγχυσης με χρήση PCA

Μοντέλο / Δοκιμές	Precision		Recall		F1-Score		AUC		TP		TN		FP		FN	
	2η	3η	2η	3η	2η	3η	2η	3η	2η	3η	2η	3η	2η	3η	2η	3η
Logistic Regression	0.92	0.92	1.00	1.00	0.96	0.96	0.7664	0.7540	282120	282289	614	395	24211	24430	562	393
	0.52	0.50	0.02	0.02	0.05	0.03										
Random Forest	0.92	0.92	1.00	1.00	0.96	0.96	0.7202	0.6805	282657	282680	49	9	24776	24816	25	2
	0.73	0.80	0.00	0.00	0.00	0.00										
XGBoost	0.92	0.92	0.99	1.00	0.96	0.96	0.7762	0.7397	281230	282027	1381	545	23444	24280	1452	655
	0.49	0.45	0.06	0.02	0.10	0.04										
LightGBM	0.96	0.96	0.76	0.73	0.85	0.83	0.7865	0.7397	213935	205189	16635	16033	8190	8792	68747	77493
	0.19	0.17	0.67	0.65	0.30	0.27										

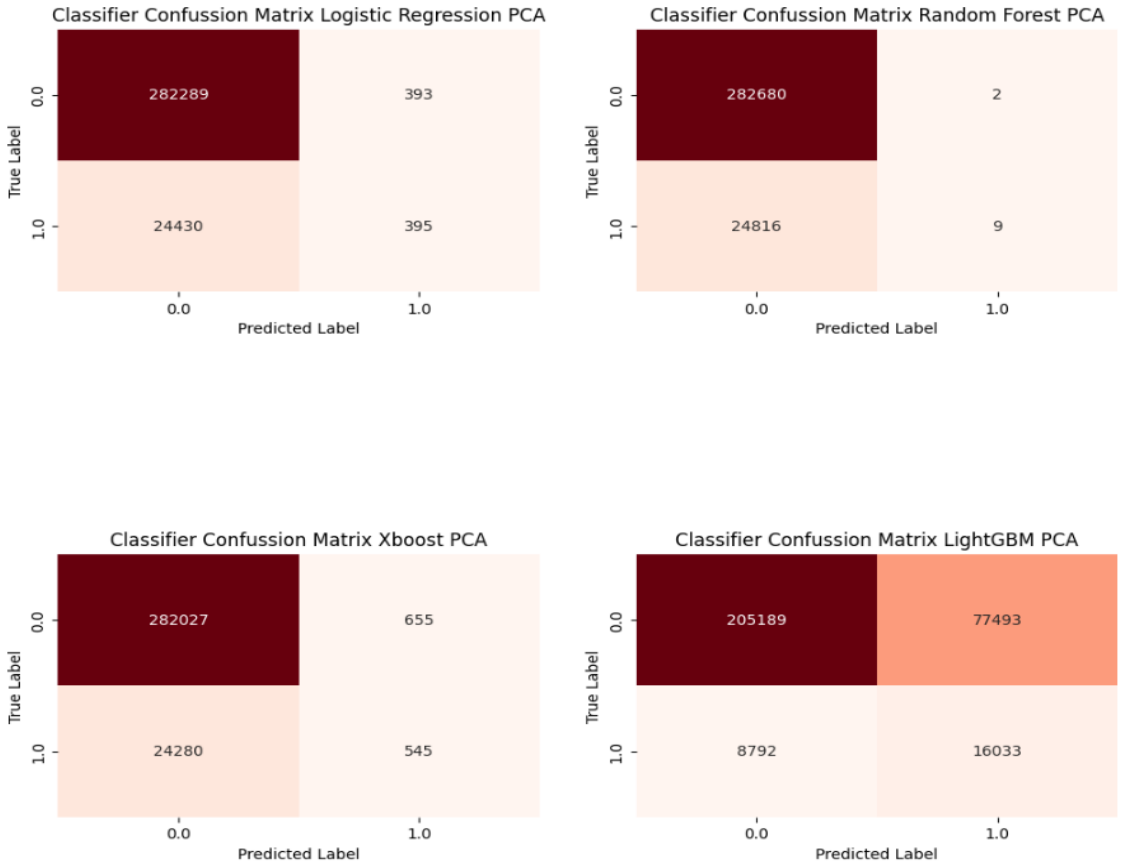
Πίνακας 7-10: Σύγκριση αποτελεσμάτων δεύτερης δοκιμής και δοκιμής με χρήση PCA

Η σύγκριση των αποτελεσμάτων μεταξύ της δεύτερης δοκιμής και της τρίτης, όπου εφαρμόσαμε PCA (Principal Component Analysis) για τη μείωση των χαρακτηριστικών, δείχνει σημαντικές αλλαγές στην απόδοση των μοντέλων, τόσο στο classification report και στο confusion matrix όσο και το ROC AUC.

1.Confusion Matrix και Classification report

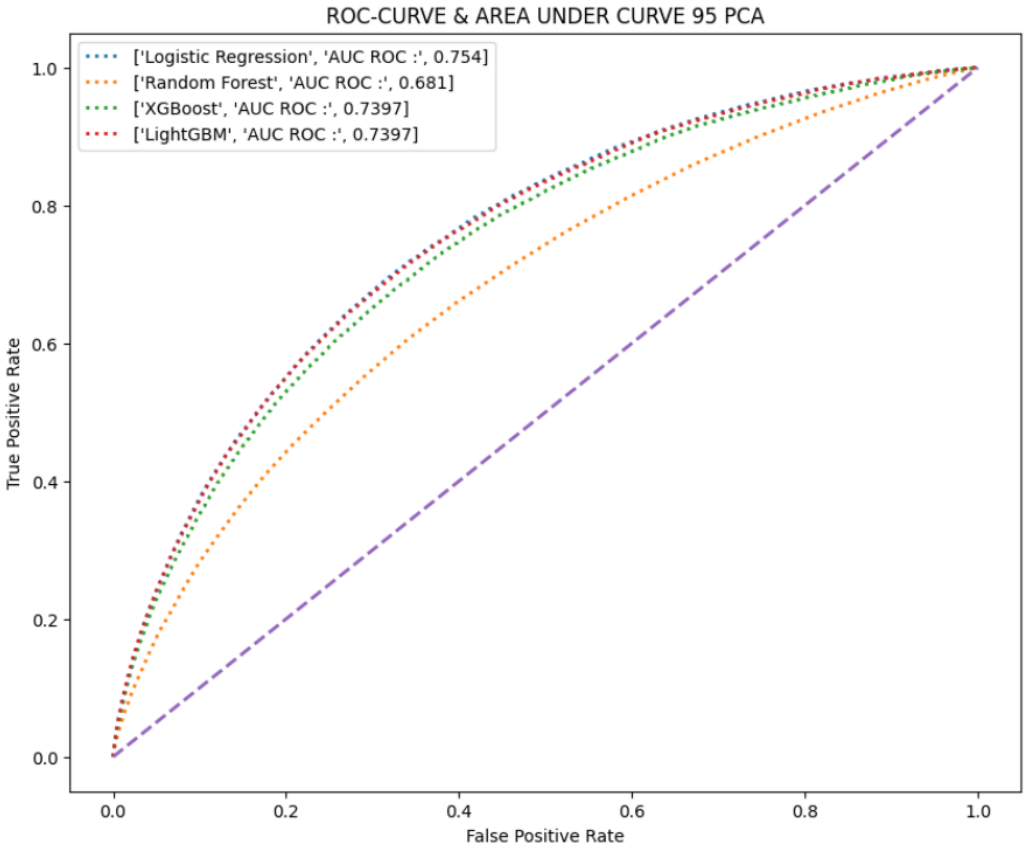
Με την εφαρμογή της PCA οι παράμετροι αξιολόγησης για την κατηγορία 1 μειώθηκαν σε όλα τα μοντέλα:

- **Logistic Regression:** Precision από 0.51 σε 0,5, Recall από 0.03 σε 0.02 και F1-score από 0.05 σε 0.03
- **Random Forest:** Το precision αυξήθηκε από 0.73 σε 0.8 όμως τα μηδενικά recall και f1-score δείχνουν ότι το μοντέλο αδυνατεί να αναγνωρίσει σχεδόν τελείως την κατηγορία 1 καθιστώντας αυτή την αύξηση ασήμαντη με μόλις 9 σωστές παρατηρήσεις.
- **XGBoost** και **Lightgbm:** Παρατηρούμε την ίδια μείωση σε όλες τις παραμέτρους.



Εικόνα 7-7: Πίνακες σύγχυσης με χρήση PCA και 95 χαρακτηριστικά

2.ROC-AUC



Εικόνα 7-8: Καμπύλες ROCAUC με χρήση PCA και 95 χαρακτηριστικά

- **Logistic Regression**: Η AUC μειώθηκε από 0.76634 σε 0.7540, κάτι που δείχνει ότι το PCA δεν βελτίωσε την ικανότητα του μοντέλου να διακρίνει μεταξύ των κατηγοριών.
- **Random Forest**: Η AUC μειώθηκε από 0.7202 σε 0.6805, γεγονός που δείχνει ότι η απόδοση του μοντέλου υπέστη μεγάλη υποβάθμιση λόγω της χρήσης του PCA.
- **XGBoost**: Επίσης μειώθηκε από 0.7762 σε 0.7397.
- **LightGBM**: Η AUC του LightGBM μειώθηκε και αυτή αρκετά 0.7397, σε σχέση με την προηγούμενη δοκιμή 0.7865, υποδεικνύοντας ότι η χρήση του PCA μείωσε την ικανότητα διαχωρισμού του μοντέλου.

Η εφαρμογή του PCA για τη μείωση της διάστασης των χαρακτηριστικών δεν απέδωσε τα επιθυμητά αποτελέσματα αντίθετα, φαίνεται να επηρέασε αρνητικά την ικανότητα των μοντέλων να ανιχνεύουν σωστά τις περιπτώσεις της κατηγορίας 1, με χειρότερα αποτελέσματα σε όλες τις παραμέτρους αξιολόγησης.

Ως αποτέλεσμα, αποφασίσαμε να απορρίψουμε τη χρήση της PCA και να συνεχίσουμε με τα 289 χαρακτηριστικά που επιλέχθηκαν στο προηγούμενο βήμα και να προχωρήσουμε στη βελτιστοποίηση των μοντέλων με αυτά τα χαρακτηριστικά.

Ένας επιπλέον λόγος για την απόρριψη της PCA ήταν η απώλεια της ερμηνευσιμότητας των χαρακτηριστικών. Η χρήση ττης PCA μετασχηματίζει τα αρχικά χαρακτηριστικά σε νέες συνιστώσες, οι οποίες δεν έχουν άμεση σχέση με τα αρχικά δεδομένα και τις μεταβλητές που είναι γνωστές από την ανάλυση του προβλήματος. Αυτό καθιστά δύσκολη την κατανόηση των επιμέρους παραγόντων που επηρεάζουν τις προβλέψεις του μοντέλου, καθώς δεν μπορούμε να γνωρίζουμε ποια χαρακτηριστικά συνέβαλαν περισσότερο στη λήψη αποφάσεων. Η απώλεια αυτής της ερμηνευσιμότητας είναι ιδιαίτερα κρίσιμη σε προβλήματα όπου η ανάλυση των

χαρακτηριστικών έχει σημασία, όπως στην περίπτωση της αθέτησης δανείων, όπου είναι σημαντικό να κατανοήσουμε ποιοι παράγοντες συμβάλλουν στην αθέτηση.

7.4 Τέταρτη δοκιμή των μοντέλων - Χρήση SMOTE

Από τις προηγούμενες δοκιμές, έγινε ξεκάθαρο ότι το μεγαλύτερο πρόβλημα που αντιμετωπίζουμε είναι η έντονη ανισορροπία των κατηγοριών στο σύνολο δεδομένων, όπου η πλειοψηφία των δειγμάτων (91%) ανήκει στην κατηγορία 0 (καλοπληρωτές), ενώ μόνο ένα μικρό ποσοστό (9%) αφορά την κατηγορία 1(κακοπληρωτές). Αυτό το γεγονός είχε ως αποτέλεσμα τα μοντέλα να επικεντρώνονται στην κατηγορία με τα περισσότερα δείγματα, αποδίδοντας υψηλές τιμές accuracy, αλλά αποτυγχάνοντας να αναγνωρίσουν τους πελάτες που είναι πιθανό να μην αποπληρώσουν το δάνειο.

Συγκεκριμένα, όπως είδαμε στα classification reports, οι παράμετροι αξιολόγησης για την κατηγορία 1 (κακοπληρωτές), όπως η precision, η recall και το F1-score, ήταν εξαιρετικά χαμηλές. Τα μοντέλα, ειδικά το Logistic Regression και το Random Forest, αδυνατούσαν να αναγνωρίσουν αποτελεσματικά τους πελάτες που ανήκουν στην κατηγορία 1.

Για να αντιμετωπίσουμε αυτό το πρόβλημα της ανισορροπίας, αποφασίσαμε να εφαρμόσουμε τη μέθοδο **SMOTE (Synthetic Minority Over-sampling Technique)**, η οποία έχει σχεδιαστεί για την ενίσχυση της μειονοτικής κατηγορίας μέσω της δημιουργίας συνθετικών δειγμάτων. Στόχος μας με το SMOTE είναι να ισορροπήσουμε τις δύο κατηγορίες, παρέχοντας στα μοντέλα περισσότερα δεδομένα από την κατηγορία 1, και έτσι να βελτιώσουμε την ικανότητά τους να αναγνωρίζουν τους πελάτες υψηλού κινδύνου.

Με τη χρήση του SMOTE, επιδιώκουμε να λύσουμε το πρόβλημα της ανισορροπίας, ελπίζοντας ότι οι παράμετροι αξιολόγησης Precision, Recall, και F1-score θα βελτιωθούν, χωρίς να επηρεαστεί αρνητικά η συνολική απόδοση των μοντέλων.

Παρακάτω ακολουθούν οι πίνακες που συγκεντρώνουν τα αποτελέσματα για τα μοντέλα της δοκιμής με εφαρμογή SMOTE με 289 χαρακτηριστικά.

AUC	Logistic Regression		Random Forest		XGBoost		LightGBM	
Folds	Train	Valid	Tr	Valid	Train	Valid	Train	Valid
0	0.79925	0.74827	1.0	0.70399	0.98660	0.75062	0.98972	0.77977
1	0.79805	0.76020	1.0	0.70977	0.98607	0.75719	0.99364	0.77768
2	0.79859	0.75912	1.0	0.70741	0.98669	0.75808	0.99335	0.78957
3	0.79773	0.75683	1.0	0.70527	0.98652	0.75432	0.99037	0.78544
4	0.79748	0.75968	1.0	0.70930	0.98618	0.76055	0.99230	0.77593
5	0.79751	0.75835	1.0	0.70653	0.98643	0.76539	0.99214	0.77661
6	0.79955	0.75254	1.0	0.70733	0.98670	0.75479	0.99507	0.78098
7	0.79802	0.75029	1.0	0.70737	0.98671	0.76105	0.99131	0.78211
8	0.79736	0.75628	1.0	0.71332	0.98645	0.76643	0.98996	0.77858
9	0.79842	0.75040	1.0	0.70174	0.98655	0.75634	0.99471	0.77925
Overall	0.79819	0.75517	1.0	0.70719	0.98649	0.75845	0.99226	0.78048

Πίνακας 7-11: AUC με 289 χαρακτηριστικά και χρήσh SMOTE για κάθε βήμα της επικύρωσης

Μοντέλο	Κατηγορία	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	0	0.96	0.75	0.82	0.7093	0.7552
	1	0.17	0.66	0.27		
Random Forest	0	0.92	0.99	0.95	0.9109	0.7072
	1	0.26	0.06	0.09		
XGBoost	0	0.92	0.99	0.96	0.9174	0.7585
	1	0.43	0.07	0.12		
LightGBM	0	0.92	1.00	0.96	0.9193	0.7805
	1	0.50	0.06	0.10		

Πίνακας 7-12: Αποτελέσματα με χρήση SMOTE

Μοντέλο	TP	TN	FP	FN
Logistic Regression	201637	16493	8332	81045
Random Forest	278701	1416	23409	3981
XGBoost	289446	1657	23168	2236
LightGBM	281333	1370	23455	1349

Πίνακας 7-13: Πίνακας σύγχυσης με χρήση SMOTE

Μοντέλο / Δοκιμές	Precision		Recall		F1-Score		AUC		TP		TN		FP		FN	
	2η	4η	2η	4η	2η	4η	2η	4η	2η	4η	2η	4η	2η	4η	2η	4η
Logistic Regression	0.92	0.96	1.00	0.75	0.96	0.82	0.7664	0.7552	282120	201637	614	16493	24211	8332	562	81045
	0.52	0.17	0.02	0.66	0.05	0.27										
Random Forest	0.92	0.92	1.00	0.99	0.96	0.95	0.7202	0.7072	282657	278701	49	1416	24776	23409	25	3981
	0.73	0.26	0.00	0.06	0.00	0.09										
XGBoost	0.92	0.92	0.99	0.99	0.96	0.96	0.7762	0.7585	281230	289446	1381	1657	23444	23168	1452	2236
	0.49	0.43	0.06	0.07	0.10	0.12										
LightGBM	0.96	0.92	0.76	1.00	0.85	0.96	0.7865	0.7805	213935	281333	16635	1370	8190	23455	68747	1349
	0.19	0.50	0.67	0.06	0.30	0.10										

Πίνακας 7-14: Σύγκριση αποτελεσμάτων δεύτερης δοκιμής και δοκιμής με χρήση SMOTE

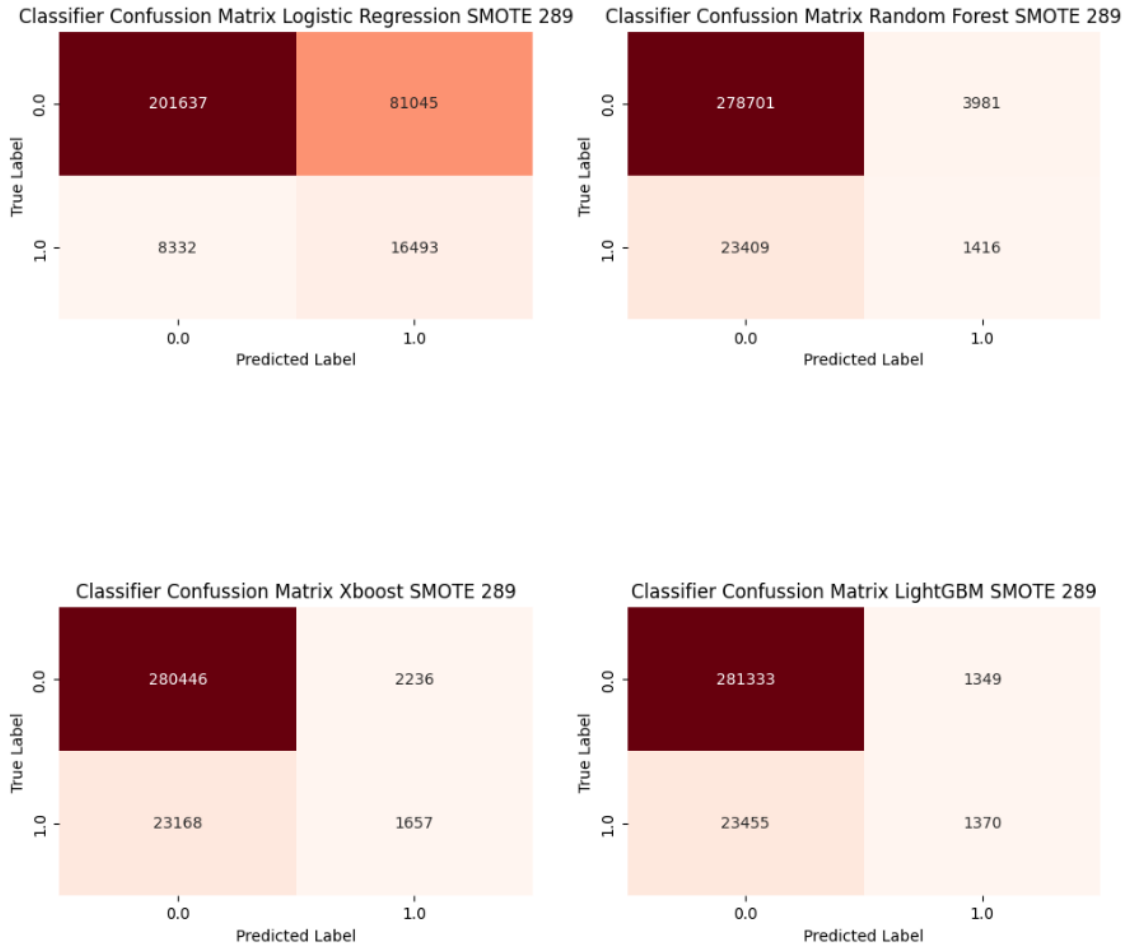
1.Classification Report και Confusion Matrix

- **Logistic Regression**: Από την εφαρμογή της SMOTE παρατηρούμε μια μεγάλη αύξηση 64% της αναγνώρισης της κατηγορίας 0 (recall) όμως έχουμε μείωση στο 75% των σωστών προβλέψεων της (precision) κάτι που αντικατοπτρίζεται και από την μεγάλη αύξηση των FN από 562 σε 81045 και μείωση των TP από 281120 σε 201637. Όσων αφορά την κατηγορία 1 έχουμε μείωση του Precision από 52% σε 17% όμως υπάρχει βελτίωση στην σωστή αναγνώριση αυτών από 2% σε 66% κάτι που φαίνεται από την μείωση των FP από 24211 σε 8332.

- **Random Forest**: Όσων αφορά την κατηγορία 0 συνεχίζει να έχει πάρα πολύ καλή απόδοση. Στην κατηγορία 1 παρατηρούμε μια μικρή αύξηση στην ικανότητα του να την αναγνωρίσει από 0% σε 6% αλλά και το ποσοστό των σωστών προβλέψεων μειώθηκε από 73% σε 26%.

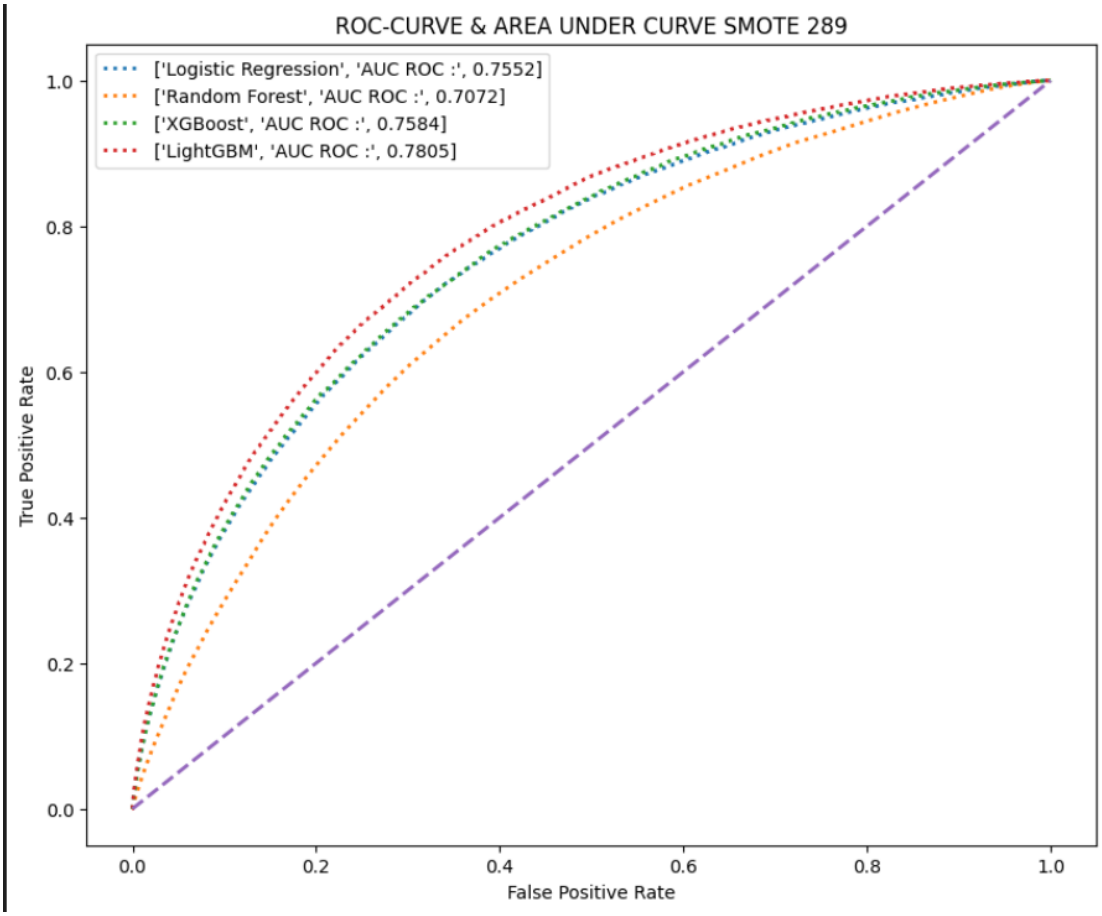
- **XGBoost**: Το precision στην κατηγορία 1 μειώθηκε από 0.49 σε 0.43, ενώ το recall βελτιώθηκε από 0.06 σε 0.07. Το F1-Score στην κατηγορία 1 αυξήθηκε ελαφρώς από 0.10 σε 0.12, υποδεικνύοντας ελαφρά βελτίωση της ισορροπίας μεταξύ precision και recall για την κατηγορία 1 στη δοκιμή με SMOTE αλλά παραμένουν πολύ χαμηλά.

- **LightGBM**: Το LightGBM εμφανίζει το καλύτερο αποτέλεσμα με precision 0.50 το οποίο είναι βελτιωμένο σε σχέση με το 0.19 της δεύτερης δοκιμής αλλά τα recall και f1-score έπεσαν αισθητά. Παρά την αύξηση του precision, το μοντέλο φαίνεται να έχασε την κανότητα του να τα αναγνωρίζει την κατηγορία 1 με recall 6% σε σχέση με το 67% που είχε.



Εικόνα 7-9: Πίνακες σύγχυσης με χρήση SMOTE και 289 χαρακτηριστικά

2.ROC AUC



Εικόνα 7-10: Καμπύλες ROCAUC με χρήση SMOTE και 289 χαρακτηριστικά

- Το **LightGBM** εξακολουθεί να έχει την καλύτερη απόδοση στο **AUC** (0.7805), παρά τη μείωση της ικανότητα του να διακρίνει μεταξύ των δύο κατηγοριών. Παρατηρήθηκε μια πολύ μικρή μείωση σε σχέση με τη Δοκιμή 2 (0.7865)

- Το **Random Forest** και το **XGBoost** παρουσίασαν ακρετή μείωση στο **AUC** λόγω της εφαρμογής του **SMOTE** από 0.7202 σε 0.7072 και 0.7762 σε 0.7585 και εξακολουθεί να μην αποδίδουν ικανοποιητικά στην αναγνώριση της κατηγορίας 1.

- Το **Logistic Regression** εμφανίζει και αυτό μια μικρή μείωση στο **AUC** 0.7552 από 0.7664 παρόλα αυτά βελτιώθηκε αρκετά στην αναγνώριση της κατηγορίας 1 με μείωση όμως της απόδοσης του στην κατηγορία 0.

Η χρήση του **SMOTE** βελτίωσε ελαφρώς στην αναγνώριση της κατηγορίας 1, ειδικά για το **Logistic Regression**, το οποίο παρουσίασε τη μικρότερη απώλεια κακοπληρωτών (FP). Παρά τη βελτίωση αυτή, τα μοντέλα εξακολουθούν να δυσκολεύονται να εντοπίσουν τους κακοπληρωτές με ικανοποιητική ακρίβεια. Το πρόβλημα της ανισορροπίας παραμένει σημαντικό εμπόδιο στην ικανότητα των μοντέλων να αναγνωρίζουν αποτελεσματικά τους πελάτες υψηλού κινδύνου.

7.5 Πέμπτη δοκιμή των μοντέλων - Χρήση SMOTEENN

Μετά τις προηγούμενες δοκιμές, ήταν ξεκάθαρο ότι η ανισορροπία των κατηγοριών είχε σημαντική επίδραση στα αποτελέσματα των μοντέλων μας, καθώς τα περισσότερα μοντέλα δυσκολεύονταν να εντοπίσουν αποτελεσματικά την κατηγορία των αθετημένων δανείων (κατηγορία 1). Η εφαρμογή του **SMOTE** στη δοκιμή της τέταρτης προσπάθειας βελτίωσε μερικώς τα αποτελέσματα, αλλά τα μοντέλα παρουσίασαν ακόμα υψηλά ποσοστά ψευδώς θετικών και ψευδώς αρνητικών προβλέψεων.

Για να προσπαθήσουμε να βελτιώσουμε περαιτέρω τα αποτελέσματα, επιλέξαμε να χρησιμοποιήσουμε την τεχνική **SMOTEENN**, η οποία συνδυάζει την υπερδειγματοληψία της κατηγορίας 1 μέσω της δημιουργίας συνθετικών δειγμάτων (**SMOTE**) και τη μείωση των δειγμάτων από την κατηγορία 0 μέσω της τεχνικής **Edited Nearest Neighbors (ENN)**. Αυτή η προσέγγιση μπορεί να μας βοηθήσει να διατηρήσουμε περισσότερα αντιπροσωπευτικά δείγματα για την κατηγορία 1, ενώ ταυτόχρονα αφαιρεί ψευδώς ταξινομημένα ή άσχετα δείγματα από την κατηγορία 0, προσφέροντας μια πιο ισορροπημένη εκπαίδευση στα μοντέλα μας.

Παρακάτω ακολουθούν οι πίνακες που συγκεντρώνουν τα αποτελέσματα για τα μοντέλα της δοκιμής με εφαρμογή SMOTEENN με 289 χαρακτηριστικά.

AUC	Logistic Regression		Random Forest		XGBoost		LightGBM	
Folds	Train	Valid	Tr	Valid	Train	Valid	Train	Valid
0	0.88056	0.75041	1.0	0.72163	0.99347	0.75144	0.98893	0.75339
1	0.87996	0.76183	1.0	0.72128	0.99339	0.75910	0.98896	0.75539
2	0.88080	0.76074	1.0	0.71361	0.99369	0.76221	0.98990	0.76989
3	0.88019	0.75675	1.0	0.71579	0.99348	0.75746	0.98897	0.76533
4	0.87953	0.75997	1.0	0.72263	0.99329	0.76560	0.98901	0.75750
5	0.88053	0.75896	1.0	0.71810	0.99332	0.76162	0.98902	0.75275
6	0.88044	0.75439	1.0	0.71613	0.99333	0.75359	0.98855	0.75731
7	0.88136	0.75679	1.0	0.71801	0.99344	0.76062	0.98901	0.76138
8	0.88080	0.75669	1.0	0.71826	0.99312	0.75700	0.98932	0.76026
9	0.88025	0.75420	1.0	0.71364	0.99330	0.75657	0.98878	0.75805
Overall	0.88044	0.75706	1.0	0.71788	0.99338	0.75851	0.98904	0.75911

Πίνακας 7-15: AUC με χρήση SMOTEEN και 289 χαρακτηριστικά για κάθε βήμα της επικύρωσης

Μοντέλο	Κατηγορία	Precision	Recall	F1-Score	Accuracy	AUC
Logistic Regression	0	0.98	0.40	0.57	0.4384	0.7571
	1	0.12	0.89	0.20		
Random Forest	0	0.95	0.74	0.84	0.7299	0.7179
	1	0.16	0.56	0.25		
XGBoost	0	0.95	0.83	0.89	0.8036	0.7585
	1	0.21	0.52	0.30		
LightGBM	0	0.94	0.94	0.94	0.8893	0.7591
	1	0.29	0.26	0.27		

Πίνακας 7-16: Αποτελέσματα με χρήση SMOTEENN

Μοντέλο	TP	TN	FP	FN
Logistic Regression	112609	22196	2629	170073
Random Forest	210,545	13901	10924	72137
XGBoost	234240	12872	11953	48442
LightGBM	267047	6414	18411	15635

Πίνακας 7-17: Πίνακας σύγχυσης με χρήση SMOTEENN

Μοντέλο / Δοκιμές	Precision		Recall		F1-Score		AUC		TP		TN		FP		FN	
	2η	5η	2η	5η	2η	5η	2η	5η	2η	5η	2η	5η	2η	5η	2η	5η
Logistic Regression	0.92	0.98	1.00	0.40	0.96	0.57	0.7664	0.7571	282120	112609	614	22196	24211	2629	562	170073
	0.52	0.12	0.02	0.89	0.05	0.20										
Random Forest	0.92	0.95	1.00	0.74	0.96	0.84	0.7202	0.7179	282657	210545	49	13901	24776	10924	25	72137
	0.73	0.16	0.00	0.56	0.00	0.25										
XGBoost	0.92	0.95	0.99	0.83	0.96	0.89	0.7762	0.7585	281230	234240	1381	12872	23444	11953	1452	48442
	0.49	0.21	0.06	0.52	0.10	0.30										
LightGBM	0.96	0.94	0.76	0.94	0.85	0.94	0.7865	0.7591	213935	267047	16635	6414	8190	18411	68747	15635
	0.19	0.29	0.67	0.26	0.30	0.27										

Πίνακας 7-18: Σύγκριση αποτελεσμάτων δεύτερης δοκιμής με τη δοκιμή με χρήση SMOTEENN

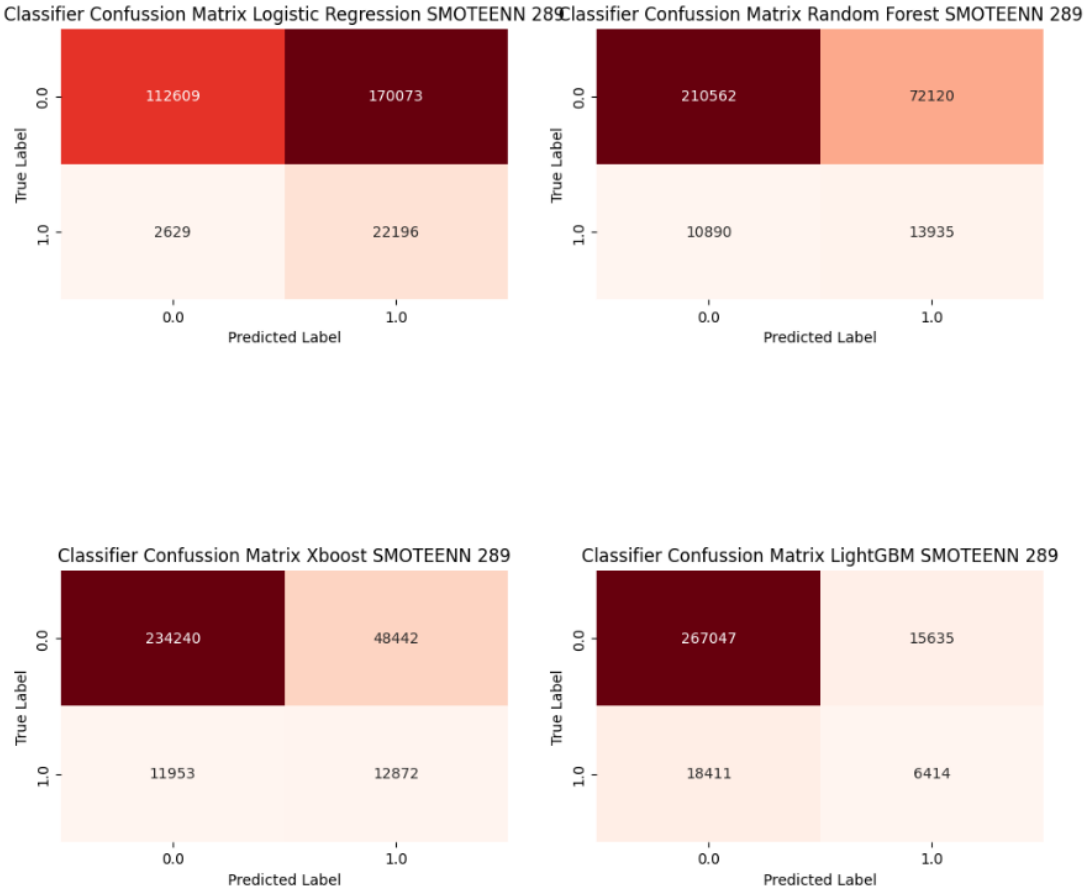
1.Classification Report και Confusion Matrix

- **Logistic Regression**: Από την εφαρμογή της SMOTEENN παρατηρούμε μια βελτίωση στην απόδοσης τη κατηγορίας 0 από 92% σε 98% (precision) όμως η ικανότητα του μοντέλου να την αναγνωρίσει σωστά μειώθηκε στο 40%(recall) με αύξηση των FN στα 170.073 από 562 και μείωση των TP σε 112.609 από 282.120. Όσων αφορά την κατηγορία 1 παρατηρήθηκε αύξηση στην ικανότητα του να την αναγνωρίσει σωστά με recall από 2% σε 89% και TN 22.196 από 614 και FP 2.629 από 24.211.

- **Random Forest**: Εδώ είχαμε πάλι αύξηση της πρόβλεψης της κατηγορίας 0 από 92% σε 95% όμως από αυτές αναγνωρίστηκε σωστά το 74% σε αντίθεση με τη δοκιμή 2 που είχε σχεδόν 100%. Αυτό φαίνεται και από την αύξηση των FN σε 72.137 από 25. Όσων αφορά την κατηγορία 1 έχουμε μείωση των σωστών προβλέψεων της από 73% σε 16% όμως από αυτές αναγνωρίζει σωστά το 56% ενώ πριν ήταν κοντά στο μηδέν. Έτσι παρουσιάζει μεγάλη αύξηση στα TN 13.901 από 49 και μείωση των FP από 24.776 σε 10.924.

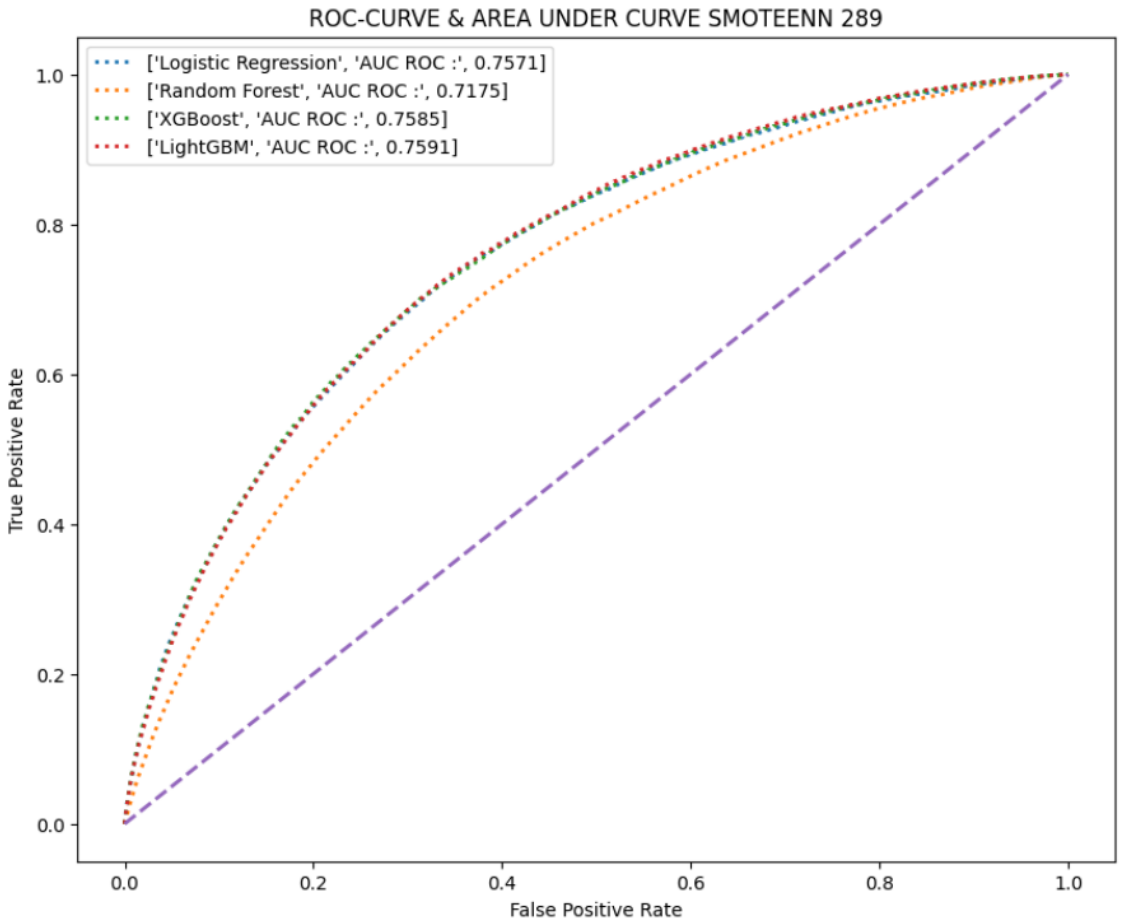
- **XGBoost**: Το XGBoost και αυτό μια μικρη αύξηση στην πρόβλεψη της κατηγορίας 0 και το 83% αυτών ήταν σωστές με αποτέλεσμα την αύξηση των FN. Στην κατηγορία 1 η ικανότητα πρόβλεψης της έπεσε από 49% σε 21% αλλά από αυτές αναγνώρισε σωστά το 52% σε αντίθεση με το 6% της δεύτερης δοκιμής. Αυτό είχε ως αποτέλεσμα την αύξηση των TN σε 12.872 από 1381 και μείωση των FP.

- **LightGBM**: Τέλος με την εφαρμογή της SMOTEENN στο LightGBM παρατηρούμε μία βελτίωση στην σωστή αναγνώριση της κατηγορίας 0 από 76% recall σε 94% με την αύξηση των TP από 213.935 σε 267.047 και μείωση των FN. Στην κατηγορία 1 παρατηρούμε βελτίωση στην πρόβλεψη της από 19% σε 29% όμως υπάρχει μείωση στη σωστή αναγνώριση της από 67% σε 26% οπότε και έχουμε μείωση των TN και αύξηση των FP.



Εικόνα 7-11: Πίνακες σύγχυσης με χρήση SMOTEENN και 289 χαρακτηριστικά

2.ROC AUC



Εικόνα 7-12: Καμπύλες ROCAUC με χρήση SMOTEENN και 289 χαρακτηριστικά

- Το **LightGBM** είχε την υψηλότερη **ROC AUC** (0.7591), επιβεβαιώνοντας ότι είναι το πιο αποδοτικό μοντέλο στη διάκριση των δύο κατηγοριών, ακολουθούμενο από το **XGBoost** με **AUC** 0.7585 και το **Random Forest** με 0.7571.

- Το **Logistic Regression** παρουσιάζει τη χαμηλότερη τιμή **AUC** γύρω από το 0.7179, υποδεικνύοντας τη δυσκολία του να διακρίνει με ακρίβεια τις δύο κατηγορίες.

Η χρήση του **SMOTEENN** βελτίωσε σε έναν βαθμό τις επιδόσεις των μοντέλων Logistic Regression, Random Forest και XGBoost όσων αφορά τη σωστή αναγνώριση τις μειονοτικής κατηγορίας 1 εις βάρος όμως της κατηγορίας 0 που παρατήρησε μεγάλη αύξηση στα False Negatives. Παρά τις βελτιώσεις, το πρόβλημα της ανισορροπίας παραμένει δύσκολο να επιλυθεί πλήρως, με τα μοντέλα να συνεχίζουν να εμφανίζουν αρκετά False Positives και False Negatives. Το F1-score σε όλα τα μοντέλα ήταν πολύ χαμηλό κοντά στο μηδέν δείχνοντας μας ότι τα μοντέλα δεν μπορούν να διαχωρίσουν καλά αν ο πελάτης θα αθετήσει το δάνειο. Για την παράμετρο ROC-AUC όλα τα μοντέλα είχαν μείωση στο πόσο καλά μπορούν να ξεχωρίσουν ανάμεσα στις δύο κατηγορίες.

8 Συμπεράσματα-Προτάσεις για μελλοντική έρευνα

Στην παρούσα εργασία, χρησιμοποιώντας δεδομένα δανείων από τη διαδικτυακή πλατφόρμα Home Credit, όπως αυτή είναι διαθέσιμη στο Kaggle, εξερευνήσαμε μεθοδολογικά τη προβλεπτική ικανότητα υποδειγμάτων μηχανικής μάθησης στην αξιολόγηση του πιστωτικού κινδύνου. Τα ευρήματά μας ακολουθώντας τη βιβλιογραφία, επιβεβαιώνουν ότι τα σύγχρονα μοντέλα μηχανικής μάθησης GBDT, όπως το XGBoost και το LightGBM, παρέχουν μια μικρή αλλά σημαντική βελτίωση, της προβλεπτικής ικανότητας σε σχέση με κλασσικά οικονομετρικά μοντέλα Logistic Regression και του Random Forest.

Παρά τα υψηλά επίπεδα ακρίβειας τα μοντέλα αδυνατούν να αναγνωρίσουν επαρκώς τους δανειολήπτες οι οποίοι δεν μπορούν να αποπληρώσουν τις οφειλές τους κάτι που επιβεβαιώθηκε από τα χαμηλά F1-Scores και το χαμηλό Recall για την κατηγορία. Ειδικά το Random Forest και το Logistic Regression είχαν χαμηλές επιδόσεις, ενώ η εφαρμογή του LightGBM έδειξε τα καλυτέρα αποτελέσματα φτάνοντας AUC 0.7865 ακολουθούμενο από το XGBoost όμως αυτά τα θετικά αποτελέσματα έχουν επηρεαστεί από το μεγάλο βαθμό ανισορροπίας των δεδομένων. Το LightGBM ήταν το καλύτερο στην αναγνώριση των κατηγοριών όμως παρουσίασε και μεγάλο αριθμό False Positives κάτι το οποίο στην περίπτωση των δανείων μπορεί να αποφέρει μεγάλες ζημιές για τους δανειοδότες.

Παρά την μείωση των πολυάριθμων χαρακτηριστικών του συνόλου δεδομένων δεν παρατηρήθηκε κάποια ουσιαστική βελτίωση στην απόδοση των μοντέλων, γεγονός που υποδεικνύει ότι τα περισσότερα χαρακτηριστικά που αφαιρέθηκαν ήταν χαμηλής σημαντικότητας. Αντίθετα μέθοδοι που χρησιμοποιήθηκαν για την επίλυση της ανισορροπίας όπως το SMOTE και το SMOTEENN, είχε πιο θετική επίδραση, βελτιώνοντας την ισορροπία μεταξύ των κατηγοριών με αισθητή αύξηση όμως των False Positives και μείωση της AUC.

Στη μελλοντική έρευνα, ένα από τα βασικά ζητήματα που πρέπει να εξεταστεί είναι η αντιμετώπιση της ανισορροπίας των κατηγοριών, καθώς αυτό παραμένει ένα σημαντικό ζήτημα που επηρεάζει την ικανότητα των μοντέλων να ανιχνεύουν αποτελεσματικά τις κατηγορίες. Οι τεχνικές που εφαρμόστηκαν, όπως το SMOTE και το SMOTEENN, προσέφεραν κάποιες βελτιώσεις, αλλά το πρόβλημα παραμένει σε μεγάλο βαθμό. Για το λόγο αυτό, προτείνεται η εξέταση πιο εξελιγμένων μεθόδων, καθώς και η εύρεση πιο ισορροπημένων συνόλων δεδομένων. Επιπλέον, η ενσωμάτωση μεθόδων βελτιστοποίησης υπερπαραμέτρων, όπως η GridSearchCV και η RandomizedSearchCV, θα μπορούσε να επιτρέψει τη βελτιστοποίηση των μοντέλων με τέτοιο τρόπο ώστε να αποφεύγονται τα υπερεκπαιδευμένα μοντέλα που έχουν χαμηλή απόδοση σε ανισόρροπα δεδομένα.

Τέλος ένα άλλος πολύ σημαντικός τομέας είναι η ενίσχυση των τεχνικών feature engineering. Από το βαθμό σημαντικότητας των χαρακτηριστικών που δημιουργήσαμε παρατηρήθηκε ότι ο ρυθμός αποπληρωμής του δανείου έπαιξε σημαντικό ρόλο κάτι που μας δείχνει ότι βάσει του υπάρχοντος γνωστικού υποβάθρου, είναι εφικτό να δημιουργηθούν πιο στοχευμένα χαρακτηριστικά τα οποία θα βελτιώσουν την απόδοση των μοντέλων. Επιπλέον, η χρήση μοντέλων που αυτόματα δημιουργούν και επιλέγουν χαρακτηριστικά, βασισμένα σε μεθόδους όπως η deep feature synthesis και recursive feature elimination θα μπορούσε να βελτιώσει τη διαδικασία της πρόβλεψης.

Βιβλιογραφία

Andreas Blöchlinger and Leippold M. (2006). “Economic benefit of powerful credit scoring”, Journal of Banking & Finance, 30-3, 2006, pp. 851-873.

André A. Montevechi, de Carvalho Miranda R., Medeiros A.L., Montevechi J.A.B. (2024). “Advancing credit risk modelling with Machine Learning: A comprehensive review of the state-of-the-art”, Engineering Applications of Artificial Intelligence, 137, 109082.

Anna Montoya, Odintsov K., Kotek M. (2018). “Home Credit Default Risk”, Kaggle. htpps://kaggle.com/competitions/home-credit-default-risk

Baesems, Bart, et al.,2003.Benchmarking state-of-the-art classification algorithms for credit scoring. J. Oper. Res. Soc. 54(6), 627-635

Barasch, Ron. 2017. “Leveraging Alternative Data to Energize Your Lending Portfolio.” Yodlee. com. https://www.yodlee.com/blog/leveragingalternative-data-energize-lending-portfolio.

Blazquez, Desamparados, and Josep Domenech. 2018. “Big Data Sources and Methods for Social and Economic Analyses.” Technological Forecasting and Social Change 130: 99–113.

CGFS (Committee on the Global Financial System) and FSB (Financial Stability Board). 2017. “FinTech Credit: Market Structure, Business Models and Financial Stability Implications.” Working Group Report. http://www.fsb.org/wp-content/uploads/ CGFS-FSB-Report-on-FinTech-Credit.pdf.

Dastile, Xolani, Celik, Turgay, Potsane, Moshe, 2020. Statistical and machine learning models in credit scoring: a systematic literature survey. Appl. Soft Comput. 91, 106263.

Dumitrescu, Elena-Ivona, et al., 2021. Machine Learning or Econometrics for Credit Scoring: Let’s Get the Best of Both Worlds.

FCA (Financial Conduct Authority). 2019. “FCA Innovate.” https://www.fca.org.uk/firms/fcainnovate

FSB (Financial Stability Board). 2017. “Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications.” FSB, Basel, Switzerland. http://www.fsb.org/wp-content/uploads/P011117. pdf.

Giri, Parimal Kumar, et al., 2021. Biogeography based optimization for mining rules to assess credit risk. Intell. Syst. Account. Finance Manag. 28 (1), 35–51.

GPFI (Global Partnership for Financial Inclusion). 2018. G-20 High-Level Principles for Digital Financial Inclusion. https://www.gpfi.org/sites/ gpfi/files/documents/G20%20High%20Level%20Principles%20for%20Digital%20Financial%20 Inclusion%20-%20Full%20version-.pdf.

Grab. 2018. “Grab and Credit Saison Form Financial Services Joint Venture to Expand Access to Credit for Southeast Asia’s Unbanked.” March. https://www. grab.com/sg/press/others/grab-and-credit-saisonform-financial-services-joint-venture-to-expandaccess-to-credit-for-southeast-asias-unbanked/.

Hurlin, Christophe, P´erignon, Christophe, 2019. Machine learning and data new sources for credit scoring. Rev. Econ. Financ. 135 (3), 21–50.

ICCR (International Committee on Credit Reporting). 2018. “Use of Alternative Data to Enhance Credit Reporting to Enable Access to Digital Financial Services by Individuals and SMEs Operating in the Informal Economy.” Guidance Note. https://www.gpfi.org/sites/gpfi/files/ documents/Use_of_Alternative_Data_to_Enhance_ Credit_Reporting_to_Enable_Access_to_Digital_ Financial_Services_ICCR.pdf.

Kaggle, X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, X. Niu (2018). “Study on A Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms according to Different High Dimensional Data Cleaning”, Electronic Commerce Research and Applications, 31, pp. 24-39.

Lessmann, Stefan, et al., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. Eur. J. Oper. Res. 247 (1), 124–136.

Lin, Wei-Yang, Hu, Ya-Han Hu Tsai, Tsai, Chih-Fong, 2011. Machine learning in financial crisis prediction: a survey. IEEE Trans. Syst. Man Cybern. C Appl. Rev. 42 (4), 421–436.

Louzada, Francisco, Ara, Anderson, Fernandes, Guilherme B., 2016. Classification methods applied to credit scoring: systematic review and overall comparison. Surveys in Operations Research and Management Science 21 (2), 117–134.

Ma Xiaojun, Jinglan S., Dehua W., Yuanbo Y., Qian Y. and Xueqi N. (2018). "Study on a Prediction Of P2P Network Loan Default Based on the Machine Learning Lightgbm and Xgboost Algorithms According to Different High Dimensional Data Cleaning", Electronic Commerce Research and Applications, 31, pp.24-39.

Marques, A.I., García, Vicente, S´ anchez, Jos´e Salvador, 2013. A literature review on the application of evolutionary computing to credit scoring. J. Oper. Res. Soc. 64, 1384–1399.

Márquez, Javier. 2008. "An Introduction to Credit Scoring for Small and Medium Size Enterprises." Unpublished paper. https://pdfs.semanticscholar. org/d07d/271a218920514f219890de5cce82448d7 efe.pdf?_ga=2.67253399.783922217.1570826234- 2114202053.1568730918

Rusli, Evelyn M. 2013. "Bad Credit? Start Tweeting: Startups Are Rethinking How to Measure Creditworthiness beyond FICO." Wall Street Journal, April 1

SAS. 2019. "Artificial Intelligence: What It Is and Why It Matters." https://www.sas.com/en_us/ insights/analytics/what-is-artificial-intelligence.

Siddiqi, Naeem. 2017. Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards. 2nd ed. Hoboken, NJ: Wiley and Sons.

Trujillo, George, Charles Kim, Steve Jones, Rommel Garcia, and Justin Murray. 2015. "Understanding the Big Data World." In Virtualizing Hadoop: How to Install, Deploy, and Optimize Hadoop in a Virtualized Architecture, 1–22. Hoboken, NJ: VMware Press. http://www.pearsonitcertification. com/store/virtualizing-hadoop-how-to-installdeploy-and-optimize-9780133811025?w_ptgrevart cl=Understanding+the+Big+Data+World_2427073.

Viani B. Djeundje, Crook J., Calabrese R., Hamid M. (2021). "Enhancing credit scoring with alternative data", Expert Systems with Applications, 163, 113766.

Vincenzo Moscato, Antonio Picariello, Giancarlo Sperlí (2021)."A benchmark of machine learning approaches for credit score prediction", Expert Systems with Applications, Volume 165,113986.

West, David, 2000. Neural network credit scoring models. Comput. Oper. Res. 27 (11-12), 1131-1152

World Bank Group (WBG) and International Committee on Credit Reporting (ICCR) (2019). "Credit Scoring Approaches Guidelines".

Wu, D. D., Chen, S.-H., & Olson, D. L. (2014). Business intelligence in risk management: Some recent progresses. Information Sciences, 256, 1–7, Business Intelligence in Risk Management.