ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ & ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ

ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΟΙΚΗΤΙΚΗ ΚΙΝΔΥΝΟΥ

ΜΟΝΤΕΛΑ ΒΑΘΜΟΛΟΓΗΣΗΣ ΣΥΜΠΕΡΙΦΟΡΑΣ ΣΤΑ ΠΛΑΙΣΙΑ ΤΗΣ ΒΑΣΙΛΕΙΑΣ

Τούρτα Ν. Χρυσάνθη

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής

Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων

για την απόκτηση του Μεταπτυχιακού Διπλώματος στην

Αναλογιστική Επιστήμη και Διοικητική Κινδύνου

Πειραιάς

Σεπτέμβριος 2020

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστήμιου Πειραιώς στην υπ' αριθμόν ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Αναλογιστική Επιστήμη και Διοικητική Κινδύνου.

Τα μέλη της ήταν:

– Κούτρας Μάρκος (Επιβλέπων)

– Τήνιος Πλάτων

– Σεβρόγλου Βασίλης

Η έγκριση της Διπλωματικής Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS & INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN

ACTUARIAL SCIENCE AND RISK MANAGEMENT

BEHAVIOURAL SCORING MODELS IN THE BASEL FRAMEWORK

Tourta N. Chrysanthi

MSc Dissertation

Submitted to the Department of Statistics and

Insurance Science of the University of Piraeus in

partial fulfillment of the requirements for the degree of

Master of Science in Actuarial Science and Risk Management

Piraeus, Greece

This thesis was approved unanimously by the three-member Commission of Inquiry appointed by the Department of Statistics and Insurance Science of the University of Piraeus No. ..... meeting in accordance with the by-laws of the Postgraduate Program in Actuarial Science and Risk Management.

Committee members were:

– Koutras Markos (Supervisor)

– Tinios Platon

– Sevroglou Vassilios

The approval of the thesis by the Department of Statistics, University of Piraeus does not imply acceptance of opinions of the author.

Στην οικογένεια μου,

Γονείς και φίλους

Περίληψη

Ο πιστωτικός κίνδυνος για ένα χρηματοπιστωτικό ίδρυμα σχετίζεται με την πιθανότητα κάποιος αντισυμβαλλόμενος να προκαλέσει στο ίδρυμα οικονομική ζημία λόγω αθέτησης των οικονομικών του υποχρεώσεων. Η ακριβής αξιολόγηση και μέτρησή του αποτελεί την βάση για το ίδρυμα ώστε να τιμολογήσει και να διαχειριστεί τα ανοίγματά του ως προς αυτόν. Η μοντελοποίηση ενός δανειακού χαρτοφυλακίου είναι μία από τις δραστηριότητες που πραγματοποιεί κάθε χρηματοπιστωτικός οργανισμός.

Η παρούσα εργασία παρουσιάζει αρχικά το κανονιστικό πλαίσιο της Βασιλείας, σύμφωνο με το οποίο κάθε πιστωτικό ίδρυμα οφείλει να λειτουργεί για να είναι αξιόπιστο. Στην επόμενη ενότητα παρουσιάζουμε τις βασικές αρχές και τα χαρακτηριστικά γνωρίσματα της ανάλυσης επιβίωσης. Στο τρίτο κεφάλαιο παραθέτουμε τις τεχνικές πρόβλεψης αθέτησης του οφειλέτη. Η εργασία ολοκληρώνεται με την εφαρμογή των μοντέλων σε πραγματικά δεδομένα και με την παρουσίαση των αποτελεσμάτων του κάθε μοντέλου.

Για την εργασία αυτή, χρησιμοποιήθηκαν δεδομένα από μια πλατφόρμα “peer to peer” (δανεισμός από άτομο σε άτομο), το “Lending Club” με στόχο την πρόβλεψη αθέτησης ενός δανείου και αν ναι πότε με βάση τα χαρακτηριστικά – μεταβλητές που δίνονται στα δεδομένα.

Λέξεις-Κλειδιά: Ανάλυση επιβίωσης, Μέθοδοι μηχανικής μάθησης, Λογιστική Παλινδρόμηση, Πιστωτικός κίνδυνος, P2P δάνεια.

Abstract

Credit risk for a financial institution focuses on the possibility of a loss resulting from a borrower's failure to or meet contractual obligations. The accurate assessment and measurement of credit risk forms a basis for the institution to price and manage various credit risk exposures. Loan valuation modeling is one of the activities institutions should undertake for risk portfolio management.

In this thesis we give a short overview of the Basel framework, which credit institutions have to comply with. Following this section, in the next chapter we present the basic principles and features of survival analysis. In the third section, we list the default forecasting models used to evaluate and predict the default of a loan and the time of default. The task is completed by applying the models to real data and presenting the results of each model.

We investigate the usage of some of these methods and their performance on a data set from a Peer-to-peer (P2P) lending company, “Lending Club”. Before building a forecasting model we will provide the theoretical background of default forecasting models and survival analysis.

Keywords: Survival Analysis, Machine learning, Logistic Regression, Credit Risk, P2P loans.

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω τον Επιβλέποντα Καθηγητή κ. Κούτρα Μάρκο για την εμπιστοσύνη που μου έδειξε αναθέτοντας μου την εκπόνηση της συγκεκριμένης εργασία και την στήριξη κατά την διάρκειά της. Θα ήθελα ακόμη να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την συμπαράσταση και την υποστήριξή τους.

Πίνακας περιεχομένων

1. Πιστωτικός κίνδυνος

1.1 Εισαγωγή

Ο πιστωτικός κίνδυνος είναι άρρηκτα συνδεδεμένος με τη φύση και τη λειτουργία ενός πιστωτικού ιδρύματος. Η θεμελιώδης λειτουργία των τραπεζών είναι ο δανεισμός χρημάτων και ο μεγαλύτερος και σημαντικότερος κίνδυνος είναι η αθέτηση, δηλαδή η μη επιστροφή των χρημάτων από τους πελάτες. Παραδείγματα τραπεζών που απέτυχαν επειδή οι βασικοί οφειλέτες τους χρεοκόπησαν είναι εμφανή τόσο στην ιστορία, όπως η αθέτηση του βασιλιά Εδουάρδου Γ 'της Αγγλίας τον 14^{ο} αιώνα που οδήγησε στη διάλυση των κυριότερων τραπεζών της Φλωρεντίας, όσο και στην τωρινή εποχή με την κρίση των ενυπόθηκων στεγαστικών δανείων στις Ηνωμένες Πολιτείες που οδήγησε σε μια παγκόσμια χρηματοπιστωτική κρίση. Οι Nijskens και Wagner (2011) υποστήριξαν ότι μια από τις αιτίες της οικονομικής κρίσης ήταν ο τρόπος με τον οποίο οι τράπεζες μεταβίβαζαν τον πιστωτικό κίνδυνο στο οικονομικό σύστημα-αγορές. Συνεπώς είναι ζωτικής σημασίας οι τράπεζες να είναι σε θέση να εντοπίζουν, να μετρούν και να διαχειρίζονται τα ανοίγματα σε πιστωτικό κίνδυνο.

1.2 Βασικές Αρχές του πιστωτικού κινδύνου

Πιστωτικός κίνδυνος είναι ο κίνδυνος μη πληρωμής χρηματοοικονομικών υποχρεώσεων όταν καθίστανται απαιτητοί επειδή ο οφειλέτης ή ο αντισυμβαλλόμενος είτε δεν μπορεί είτε δεν επιθυμεί να πληρώσει.

Ο όρος που χρησιμοποιείται συνηθέστερα σε σχέση με τον πιστωτικό κίνδυνο είναι "αθέτηση" (default). Αν και δεν υπάρχει τυπικός ορισμός της αθέτησης, ο ορισμός που παρέχεται από την Επιτροπή της Βασιλείας για την Τραπεζική Εποπτεία [(Basel Committee on Banking Supervision June 2006), BCBS] καταγράφει τα πιο σημαντικά χαρακτηριστικά:

Αθέτηση θεωρείται ότι συνέβη σε σχέση με συγκεκριμένο οφειλέτη όταν πραγματοποιηθεί ένα από τα δύο ακόλουθα γεγονότα ή και τα δύο :

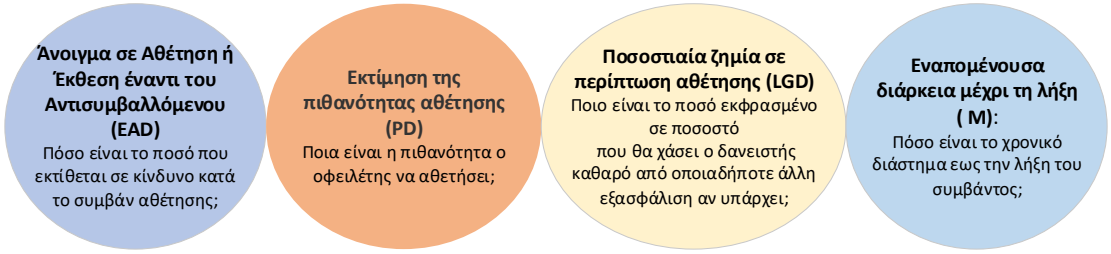
- Η τράπεζα εκτιμά ότι ο οφειλέτης δεν είναι πιθανό να εκπληρώσει πλήρως την πιστωτική του υποχρέωση έναντι του πιστωτικού ιδρύματος, εκτός εάν το πιστωτικό ίδρυμα προσφύγει σε μέτρα όπως η ρευστοποίηση της εξασφάλισης (εάν υπάρχει).
- Ο οφειλέτης έχει καθυστερήσει πέραν των 90 ημερών να εκπληρώσει την πιστωτική υποχρέωση στο πιστωτικό ίδρυμα. Με τις πιστωτικές διευκολύνσεις, η καθυστέρηση αρχίζει να τρέχει αμέσως μόλις ο οφειλέτης υπερβεί ένα εγκεκριμένο όριο ή ενημερωθεί ότι διαθέτει όριο χαμηλότερο από το τρέχον υπόλοιπο.

Ο βαθμός στον οποίο ένα πιστωτικό ίδρυμα εκτίθεται σε πιστωτικό κίνδυνο εξαρτάται από διάφορους παράγοντες, όπως:

1) *Όρια πιστωτικού κινδύνου:* Ο καθορισμός κατάλληλων ορίων πιστωτικού κινδύνου είναι καθοριστικής σημασίας για τη διαχείριση του κινδύνου. Τα όρια ορίζουν το επίπεδο πιστωτικής έκθεσης που μια τράπεζα είναι διατεθειμένη να αποδεχθεί. Όσο υψηλότερο είναι το όριο τόσο μεγαλύτερη είναι η έκθεση. Τα όρια διαφοροποιούνται ανάλογα με το θεσμικό πλαίσιο, το προϊόν και τα αντίστοιχα χαρακτηριστικά του. Για παράδειγμα, τίθενται διαφορετικά όρια για:
   - Ατομικά δάνεια ή άλλες συναλλαγές
   - Μεμονωμένοι δανειολήπτες / αντισυμβαλλόμενοι
   - Ομάδες συνδεδεμένων δανειοληπτών / αντισυμβαλλομένων
   - Ειδικές βιομηχανίες / τομείς
   - Ειδικά προϊόντα
   - Γεωγραφικές περιοχές
   - Το πιστωτικό χαρτοφυλάκιο στο σύνολό του
2) *Ο τύπος του προϊόντος :*
   - Τα δάνεια σταθερού επιτοκίου τα οποία είναι συνήθως πληρωτέα σε δόσεις και δεν μπορούν να αναδιαρθρωθούν. Εάν πραγματοποιούνται πληρωμές, η έκθεση μειώνεται. Ωστόσο, η έκθεση αυξάνεται εάν οι οφειλόμενοι τόκοι δεν πληρώνονται από τον οφειλέτη.
   - Τα προϊόντα ανακυκλούμενης πίστωσης, όπως οι πιστωτικές κάρτες. Αυτά μπορούν να αντληθούν, να πληρωθούν και να ανασυσταθούν. Επομένως, η έκθεση κυμαίνεται για τα προϊόντα αυτά.
   - Τα διαπραγματευόμενα προϊόντα, όπως οι συμβάσεις συναλλάγματος και τα παράγωγα. Ο πιστωτικός κίνδυνος για αυτά ποικίλλει κατά τη διάρκεια της σύμβασης, καθώς οι παράγοντες κινδύνου αγοράς όπως οι συναλλαγματικές ισοτιμίες και τα επιτόκια κυμαίνονται. Η πιστωτική έκθεση για τέτοια προϊόντα υπολογίζεται χρησιμοποιώντας μοντέλα τα οποία ενημερώνονται καθημερινά με τις τρέχουσες τιμές. Όταν επιτρέπεται ο συμψηφισμός των συμβάσεων, η έκθεση είναι το άθροισμα των συμβάσεων "στο χρήμα", μείον το άθροισμα των συμβάσεων "εκτός χρήματος".
3) *Η έκθεση σε συναλλαγματική ισοτιμία:* Τα δάνεια / ανοίγματα σε άλλα νομίσματα παρουσιάζουν διακυμάνσεις καθώς μεταβάλλονται οι συναλλαγματικές ισοτιμίες. Οι δυσμενείς μεταβολές στις συναλλαγματικές ισοτιμίες μπορούν να έχουν ως αποτέλεσμα την έκθεση σε κίνδυνο στο βασικό νόμισμα που υπερβαίνει το συμφωνημένο πιστωτικό όριο.
4) *Η έκθεση στον κίνδυνο διακανονισμού:* Την ημέρα κατά την οποία οι συμβάσεις καθίστανται απαιτητές, υφίσταται ο κίνδυνος διακανονισμού εάν αποδεσμευτεί η αξία

πριν ληφθεί η πληρωμή. Αυτός ο κίνδυνος αυξάνεται εάν ωριμάσουν την ίδια ημέρα πολλαπλά συμβόλαια, αλλά μπορεί να αποφευχθεί με την πληρωμή πριν από την αποδέσμευση αξίας ή με ταυτόχρονη ανταλλαγή.

Η **μέτρηση του πιστωτικού κινδύνου** δεν είναι πάντοτε μία εύκολη διαδικασία. Υπάρχουν τέσσερις βασικοί παράγοντες στη διαδικασία μέτρησης.



Η σημασία της ακριβούς αξιολόγησης και μέτρησης του πιστωτικού κινδύνου είναι ύψιστη. Αποτελεί τη βάση για να προβαίνουν οι τράπεζες στην τιμολόγηση και τη διαχείριση των διαφόρων πιστωτικών κινδύνων. Η χρηματοοικονομική κρίση της περασμένης δεκαετίας κατέδειξε επίσης τη σημασία του ύψους του κεφαλαίου που οφείλουν να κατέχουν τα χρηματοπιστωτικά ιδρύματα. Περισσότερες λεπτομέρειες δίνονται παρακάτω.

Α. Άνοιγμα σε Αθέτηση ή Έκθεση έναντι του Αντισυμβαλλόμενου (Exposure at Default-EAD): μετρά το δυνητικό επίπεδο έκθεσης σε μελλοντική ημερομηνία σε περίπτωση αθέτησης υποχρέωσης από τον οφειλέτη. Η έκθεση δεν είναι ίδια για όλα τα προϊόντα. Για τα απλά μακροπρόθεσμα δάνεια, η έκθεση είναι σε μεγάλο βαθμό γνωστή εκ των προτέρων - παρόλο που οι τράπεζες πρέπει να γνωρίζουν τη δυνητική επίδραση των δεδουλευμένων τόκων (που θα μπορούσαν να αυξήσουν την έκθεση) και τις προπληρωμές (που θα μπορούσαν να μειώσουν την έκθεση). Ωστόσο, μια τράπεζα μπορεί να έχει επεκτείνει μια διευκόλυνση ανακυκλούμενης πίστωσης που επιτρέπει στον οφειλέτη να αντλεί χρήματα στο μέλλον. Η τρέχουσα έκθεση σε αυτόν τον δανειολήπτη μπορεί να είναι μηδέν, αλλά αυτό θα μπορούσε να αλλάξει στο μέλλον. Μια μακροπρόθεσμη σύμβαση παραγώγων θα έχει συνήθως μικρή πιστωτική έκθεση κατά την έναρξη, εάν η σύμβαση εκτελείται στην τιμή της αγοράς. Όμως, με την πάροδο του χρόνου, η σύμβαση θα αναπτύξει θετική ή αρνητική έκθεση υπό το φως των αλλαγών της αγοράς. Η μελλοντική πιστωτική έκθεση δεν είναι η ίδια με την τρέχουσα έκθεση.

Β. Ποσοστιαία ζημία σε περίπτωση αθέτησης (Loss Given Default – LGD): είναι μια εκτίμηση του ποσοστού της EAD που θα χαθεί σε περίπτωση αθέτησης. Κατά την εκτίμηση των πιθανών πιστωτικών ζημιών, οι δανειστές πρέπει να εξετάσουν το ποσό που θα μπορούσε να ανακτηθεί σε περίπτωση αθέτησης. Για παράδειγμα, ένα δάνειο μπορεί να υποστηριχθεί από την ύπαρξη ασφάλειας. Αν ναι, το καθαρό άνοιγμα - γνωστό ως η ζημία

λόγω αθέτησης (LGD) - είναι η διαφορά μεταξύ του ανοίγματος σε αθέτηση και του εκτιμώμενου ποσού ανάκτησης. Σε ορισμένες περιπτώσεις, η καθαρή έκθεση ενδέχεται να είναι μηδενική.

Το LGD εκφράζεται κανονικά ως ποσοστό:

$$LGD = 1 - Ποσοστό\ ανάκτησης\ .$$

Το ποσοστό ανάκτησης αντιπροσωπεύει το ποσοστό του ποσού έκθεσης που ανακτάται σε περίπτωση αθέτησης, οπότε το LGD είναι το υπόλοιπο ποσοστό.

Γ. Πιθανότητα αθέτησης (Probability of Default-PD): είναι η πιθανότητα ένας πελάτης θα χρεοκοπήσει σε ένα συγκεκριμένο χρονικό διάστημα (συνήθως ένα έτος). Εκφράζεται ως ποσοστό. Οι πιθανότητες αθέτησης μπορούν να ληφθούν από δύο πηγές:

- Εξωτερικές πηγές: οι πιο γνωστές εξωτερικές πηγές είναι οι σημαντικότεροι οργανισμοί αξιολόγησης όπως το Standard & Poor's (S & P), το Moody's και το Fitch. Άλλοι, όπως η Dun & Bradstreet, παρέχουν αξιολογήσεις για μικρότερες εταιρείες.
- Εσωτερικά μοντέλα: οι μεγαλύτερες τράπεζες χρησιμοποιούν τα δικά τους εσωτερικά μοντέλα κινδύνου για τον υπολογισμό των PD για τους πελάτες από διάφορες εσωτερικές και εξωτερικές εισροές δεδομένων.

Οι PD είναι ενδεικτικές καθώς ενδέχεται να προκύψουν απροσδόκητα συμβάντα ή ενδέχεται τα χρησιμοποιούμενα μοντέλα να είναι ακατάλληλα, ανακριβή ή να έχουν κακή χρήση. Στην πράξη, ορισμένοι οφειλέτες δεν θα εκπληρώσουν τις υποχρεώσεις τους, ακόμη και αν η PD έχει χαμηλή τιμή - αλλά αυτές οι περιπτώσεις αθέτησης θα είναι πολύ λιγότερες από αυτές που βρίσκονται στο υψηλότερο σημείο κινδύνου της κλίμακας.

Δ. Εναπομένουσα διάρκεια μέχρι την λήξη (Maturity-M): είναι η συμφωνημένη περίοδος κατά την οποία λήγει ένα δάνειο και μπορεί να αλλάξει επανειλημμένα καθ' όλη τη διάρκεια ζωής ενός δανείου, σε περίπτωση που ένας δανειολήπτης ανανεώσει το δάνειο, σε περίπτωση αθέτησης υποχρεώσεων, σε περίπτωση που επιβάλει υψηλότερα τέλη τόκων ή εξοφλήσει τη συνολική υποχρέωση νωρίτερα. Σχετικά με την σύνδεση του με την πιθανότητα αθέτησης, για μεγαλύτερο χρονικό διάστημα είναι πιο δύσκολο να υπολογιστεί μια αξιόπιστη εκτίμηση της πιθανότητας αθέτησης, καθώς πολλοί παράγοντες ενδέχεται να αλλάξουν, είτε θετικά είτε αρνητικά.

1.3 Μέτρηση του πιστωτικού κινδύνου

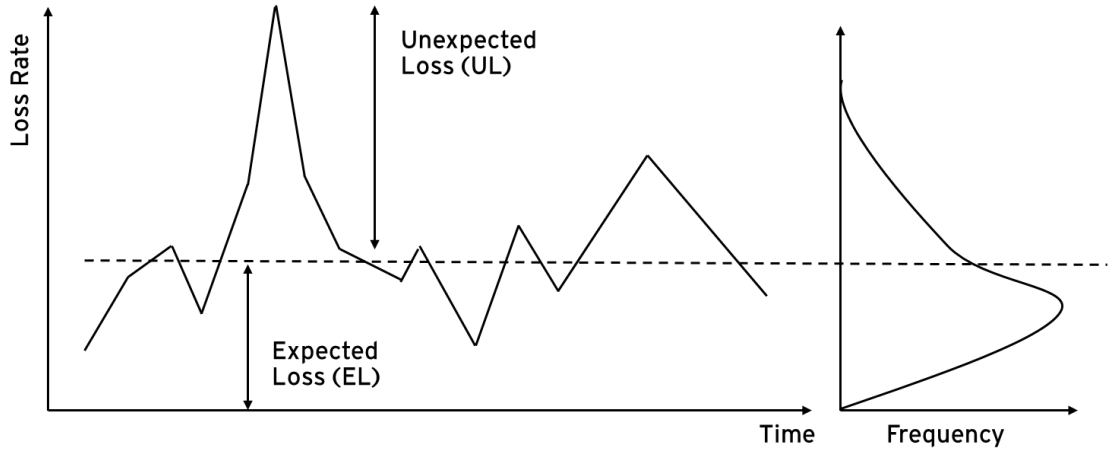
Με την εκτίμηση των παραπάνω ποσοτικών παραμέτρων προβλέπεται η απαιτούμενη κεφαλαιακή επάρκεια για κάθε άνοιγμα - η απαιτούμενη κάλυψη της **αναμενόμενης ζημιάς από αθέτηση** εκπλήρωσης των υποχρεώσεων του δανειολήπτη σε μια δεδομένη περίοδο **(Expected Loss-EL)** και δίνεται από τον παρακάτω τύπο:

$$EL = EAD\ x\ PD\ x\ LGD\ .$$

Αυτό παράγει ένα απόλυτο ποσό. Το EL ως ποσοστό, το οποίο είναι χρήσιμο ως σχετικό μέτρο κινδύνου, είναι απλά το γινόμενο PD x LGD. Υψηλότερες τιμές για PD, EAD ή LGD θα οδηγήσουν σε υψηλότερο EL. Ωστόσο, μια υψηλότερη τιμή PD μπορεί να είναι κάτι παραπάνω από αντιστάθμιση καθώς λαμβάνοντας οποιαδήποτε μορφή εξασφάλισης, το LGD θα είναι πολύ χαμηλότερο και θα έχει ως αποτέλεσμα χαμηλότερη κεφαλαιακή απαίτηση. Επιπρόσθετα, το συνολικό EL σε επίπεδο χαρτοφυλακίου είναι το άθροισμα των επιμέρους EL για κάθε ένα περιουσιακό στοιχείο, δηλαδή

$$EL = \sum_{k=0}^{n} EL_{\mathrm{k}} \ .$$

Η **μη αναμενόμενη απώλεια (UL)** είναι ένα μέτρο κατανομής των ζημιών που υπερβαίνει την αναμενόμενη ζημία (EL) για μια καθορισμένη χρονική περίοδο. Στην πράξη απρόβλεπτες απώλειες συμβαίνουν επειδή είναι απίθανο οι πραγματικές απώλειες να αντανακλούν τέλεια τις προσδοκίες - υπάρχει κάποια πιθανότητα οι απώλειες να είναι μεγαλύτερες από τις αναμενόμενες. Το UL είναι συνάρτηση του EL για ένα χαρτοφυλάκιο και υπολογίζεται λαμβάνοντας υπόψη τον αριθμό των πελατών και τους ενδεχόμενους κινδύνους συσχετισμού μεταξύ τους. Οι αθετήσεις δεν είναι γενικά ανεξάρτητα γεγονότα. Ο κίνδυνος συσχέτισης είναι ο κίνδυνος η αθέτηση υποχρέωσης ενός πελάτη να οδηγήσει σε αδυναμία πληρωμής και άλλων. Για παράδειγμα, η αποτυχία ενός δανειολήπτη μπορεί να είναι τυχαίο γεγονός μιας γενικής τάσης ή μπορεί να προκαλέσει την αδυναμία ενός άλλου. Εάν υπάρχει θετική συσχέτιση μεταξύ των δανειοληπτών, αυτό θα επηρεάσει τη συνολική κατανομή των αθετήσεων σε ένα χαρτοφυλάκιο. Όταν συσχετίζονται οι αθετήσεις, αυξάνεται η πιθανότητα πιο ακραίων αποτελεσμάτων. Στο ακόλουθο διάγραμμα στα αριστερά απεικονίζονται οι μεταβλητές EL,UL και στα δεξιά η κατανομή των πραγματικών απωλειών.

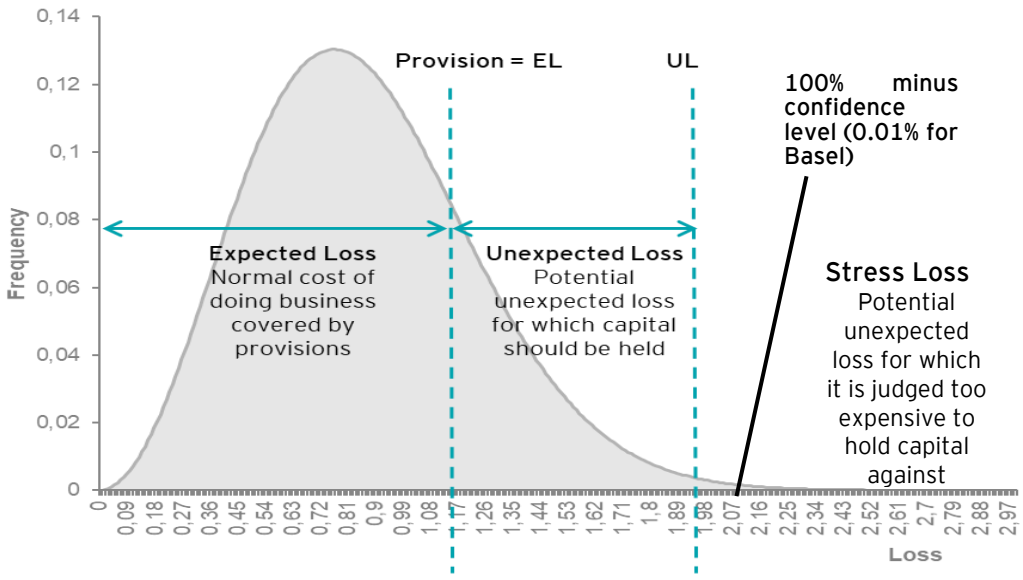


Γράφημα 1: Απεικόνιση της σχέσης μεταξύ EL και UL (BIS (2005)).

Η διακεκομμένη γραμμή είναι το EL. Η συνεχής γραμμή δείχνει τις πραγματικές απώλειες, οι οποίες μπορεί να είναι υψηλότερες ή χαμηλότερες από το EL. Αυτό συμβαίνει επειδή ο EL είναι ένας μέσος όρος, οπότε η πιθανότητα πραγματικών απωλειών να είναι ίση με EL σε οποιοδήποτε έτος είναι πολύ χαμηλή. Στο δεξί διάγραμμα φαίνεται ότι το πιο πιθανό αποτέλεσμα σε κάθε ένα έτος είναι οι πραγματικές απώλειες να είναι μικρότερες από το EL. Η σχέση αυτή, όμως, είναι μόνο ένας δείκτης των αποδόσεων, όχι συναρτήσει του χρόνου - ακόμα και αν το πιο ακραίο αποτέλεσμα εμφανίζεται μια φορά στα εκατό χρόνια, αυτό δεν σημαίνει ότι δεν μπορεί να συμβεί φέτος.

Δεδομένου ότι αναμένονται τέτοιες (απρόσμενες) απώλειες, πρόκειται για κόστος επιχειρηματικής δραστηριότητας που, όπως και όλα τα άλλα κόστη, πρέπει να καλύπτεται από τα κέρδη από πελάτες που πληρούν τις υποχρεώσεις τους για να είναι κερδοφόρα μια τράπεζα. Σε αυτές τις περιπτώσεις, το EL καλύπτεται από τα κέρδη και το UL από το κεφάλαιο της εταιρείας. Ωστόσο, δύναται οι ζημίες να υπερβούν τα κέρδη και να προκύψει καθαρή ζημία και μείωση του τραπεζικού κεφαλαίου.

Επιπρόσθετα, υπάρχει η πιθανότητα ένα χαρτοφυλάκιο να υποστεί απώλεια "ακραία" η οποία επισημαίνεται ως έκτακτη κατάσταση καθώς πλήττει σοβαρά την οικονομική υπόσταση της εταιρείας. Η πιθανότητα όμως είναι πολύ χαμηλή και είναι εξαιρετικά επώδυνο να διατηρείται επαρκές κεφάλαιο για την προστασία από μια τέτοια απώλεια. Αυτό σημαίνει ότι υπάρχει κάποιο ποσοστό ζημίας που δεν θα καλυφθεί από το κεφάλαιο και δεν θα πρέπει να ξεπερνάει το 1% του συνολικού κεφαλαίου σύμφωνα με τις εποπτικές αρχές. Τα παραπάνω απεικονίζονται στο διάγραμμα που ακολουθεί:



Γράφημα 2: Απεικόνιση της κατανομής ζημιών (Chatterjee (2015))**.**

Το άθροισμα της αναμενόμενης και μη αναμενόμενης απώλειας (EL + UL) υποδηλώνει το ποσό του κεφαλαίου που απαιτείται έτσι ώστε η τράπεζα να μπορέσει να προστατευτεί από την αποτυχία με ένα δεδομένο επίπεδο εμπιστοσύνης και συμφωνεί με την έννοια Αξία σε Κίνδυνο (Value at-Risk – VaR) η οποία εκτιμά τη μέγιστη δυνατή ζημιά για συγκεκριμένη χρονική περίοδο και με συγκεκριμένη πιθανότητα. Σύμφωνα με την Βασιλεία ΙΙ, το απαραίτητο κεφάλαιο καθορίζεται κατά τέτοιο τρόπο ούτως ώστε να διατηρείται ένα σταθερό επίπεδο εμπιστοσύνης σύμφωνα με τις απαιτήσεις των εποπτικών αρχών. Κατ' αυτή την έννοια, το απαραίτητο κεφάλαιο καθορίζεται σύμφωνα με το άνοιγμα μεταξύ EL και VaR. Εφόσον το EL καλύπτεται επαρκώς από προβλέψεις ή έσοδα, τότε η πιθανότητα η τράπεζα να γίνει αφερέγγυα (με χρονικό ορίζοντα ενός έτους) περιορίζεται σε πολύ μικρό ποσοστό. (π.χ. 0,01 %).

1.4 Κεφαλαιακή επάρκεια

Κεφαλαιακή Επάρκεια ορίζεται ως το μέτρο που δείχνει κατά πόσο το κεφάλαιο μιας τράπεζας (και γενικά ενός χρηματοπιστωτικού ιδρύματος) είναι αρκετό προκειμένου η τράπεζα να μπορεί να ανταπεξέλθει σε πιθανές ζημίες από δάνεια τα οποία έχει ήδη δώσει, και να μπορέσει μελλοντικά να είναι και εκείνη συνεπής απέναντι στις δικές της υποχρεώσεις και τα δικά της χρέη[1]. Αποτελεί επίσης το επίκεντρο του πυλώνα 1 του πλαισίου της Βασιλείας.

Ο **δείκτης κεφαλαιακής επάρκειας (capital adequacy ratio, CAR**) αντιπροσωπεύει το λόγο του κεφαλαίου προς τα σταθμισμένα περιουσιακά στοιχεία (risk-weighted assets, RWAs), πιο συγκεκριμένα:

$$\text{Δείκτης Κεφαλαιακής Επάρκειας } = \frac{Κεφάλαιο\ Tier\ 1 + Κεφάλαιο\ Tier\ 2}{Σταθμισμένα\ στον\ Κίνδυνο\ Περουσιακά\ Στοιχεία}$$

$$CAR\ = \frac{Qualified\ Capital}{Risk - weighted\ assets}$$

- Κεφάλαιο Tier 1 είναι το βασικότερο κομμάτι των ιδίων κεφαλαίων το οποίο είναι σχετικά διαφανές και ασφαλές, όπως το μετοχικό κεφάλαιο και τα καταγεγραμμένα αποθεματικά. Η Βασιλεία ΙΙΙ πρόσθεσε ένα επιπλέον μερίδιο στο Tier 1, κεφάλαια μετατρέψιμα σε αυτά που χαρακτηρίζονται ως Tier 1 όταν συμβεί κάποιο γεγονός που το επιβάλει. Το συνολικό κεφάλαιο Tier 1 έχει όριο 6%, με 4.5% να είναι το όριο για το επιπρόσθετο κεφάλαιο που επιβλήθηκε από την Βασιλεία ΙΙΙ.

[1] Πηγή: Οικονομικά της Καθημερινότητας, Κεφαλαιακή Επάρκεια (Διαθέσιμο στο: https://www.dailyeconomics.gr/oikonomikoi-oroi/kefalaiakh-eparkeia)

- Κεφάλαιο Tier 2 είναι ένα συμπληρωματικό κεφάλαιο το οποίο είναι πιο σύνθετο και ευμετάβλητο, όπως προβλέψεις για επισφαλή δάνεια και το χρέος μειωμένης εξασφάλισης

Ο κίνδυνος απώλειας κατά την παροχή πιστωτικών διευκολύνσεων, για παράδειγμα, ενός πελάτη με πιστοληπτική ικανότητα ΑΑΑ, είναι χαμηλότερος από έναν που βαθμολογείται με BBB. Ως αποτέλεσμα, θα πρέπει να διατηρηθεί περισσότερο το κεφάλαιο για έναν πελάτη που έχει βαθμολογηθεί με BBB. Το ύψος του κεφαλαίου που πρέπει να διατηρηθεί καθορίζεται με τη μετατροπή της ονομαστικής αξίας του ενεργητικού σε ένα σταθμισμένο με βάση το κίνδυνο στοιχείο, το οποίο είναι γνωστό ως σταθμισμένο περιουσιακό στοιχείο (RWA). Σε ένα περιουσιακό στοιχείο ισχύει:

$$Σταθμισμένο\ (RWA) = Ονομαστική\ αξία\ (Asset\ Value)\ x\ Ύψος\ κινδύνου\ (Risk\ weight)$$

Ενώ σε ένα χαρτοφυλάκιο με N περιουσιακά στοιχεία:

$$Σταθμισμένο\ (RWA) = \sum_{k=1}^{N} RWA\ Asset_k$$

Τα ύψη κινδύνου διαφέρουν ανάλογα με τον κίνδυνο απώλειας του περιουσιακού στοιχείου. Τα υψηλότερα ύψη συνδέονται με υψηλότερους κινδύνους. Όπου επιτρέπονται από τις ρυθμιστικές αρχές, οι τράπεζες μπορούν να χρησιμοποιούν τα δικά τους μοντέλα για τον υπολογισμό των RWA. Όπου αυτό δεν συμβαίνει, οι τράπεζες πρέπει να χρησιμοποιούν σταθμισμένα βάρη κινδύνου διαφοροποιημένα ανά τύπο αντισυμβαλλομένου και εξωτερική αξιολόγηση κινδύνου.

Το εποπτικό πλαίσιο συνιστά ένα ελάχιστο όριο απαιτούμενων κεφαλαίων ίσο με το 8% του σταθμισμένου ενεργητικού (δείκτης φερεγγυότητας). Η απαίτηση αυτή καθορίστηκε από την Επιτροπή Βασιλείας για την Τραπεζική Εποπτεία το 1988 όταν δημοσιεύθηκαν οι αρχικοί κανόνες για την κεφαλαιακή επάρκεια (Basel I). Η ελάχιστη αναλογία παρέμεινε αμετάβλητη όταν αναθεωρήθηκαν οι κεφαλαιακές απαιτήσεις ως μέρος των Βασιλεών ΙΙ και Βασιλεία ΙΙΙ, ωστόσο έχουν προστεθεί στον τύπο υπολογισμού νέες κατηγορίες κεφαλαιακών αποθεμάτων.

Η κεντρική ιδέα για τη μέτρηση των πιστωτικών ανοιγμάτων με σκοπό τον προσδιορισμό της κεφαλαιακής επάρκειας είναι τα σταθμισμένα περιουσιακά στοιχεία (RWA).Το κανονιστικό πλαίσιο προτείνει δύο εναλλακτικές προσεγγίσεις υπολογισμού των σταθμισμένων περιουσιακών στοιχείων για τον υπολογισμό της κεφαλαιακής επάρκειας με αυξημένο βαθμό πολυπλοκότητας και διαφοροποιούμενη ευαισθησία ως προς τον κίνδυνο (Χαραλαμπίδης 2004).

Μέθοδοι υπολογισμού κεφαλαιακών απαιτήσεων

Βαθμός ευαισθησίας στον κίνδυνο	Πιστωτικός Κίνδυνος
χαμηλός	Τυποποιημένη (Standardized Approach, SA)
μέτριος	Θεμελιώδης Εσωτερικών Διαβαθμίσεων (IRB Foundation)
υψηλός	Προηγμένη Βάσει Εσωτερικών Διαβαθμίσεων (IRB Advanced)

Επιπρόσθετα, εκτός από την διαφορά ως προς τον βαθμό ευαισθησίας στον κίνδυνο, οι παραπάνω μέθοδοι διαφέρουν ως προς τον τρόπο υπολογισμού των βασικών παραγόντων που απαιτούνται για την μέτρηση του πιστωτικού κινδύνου. Ο πίνακας δείχνει τον τρόπο υπολογισμού της PD, LGD και EAD για την καθεμία μέθοδο.

	SA	IRB Foundation	IRB Advanced
PD	Παρέχεται από εξωτερικούς οργανισμούς	Εκτιμάται εσωτερικά	Εκτιμάται εσωτερικά
LGD	Παρέχεται από εξωτερικούς οργανισμούς	Παρέχεται από εξωτερικούς οργανισμούς	Εκτιμάται εσωτερικά
EAD	Παρέχεται από εξωτερικούς οργανισμούς	Παρέχεται από εξωτερικούς οργανισμούς	Εκτιμάται εσωτερικά

Για καθεμία από τις παραπάνω μεθόδους θα δοθούν περισσότερες πληροφορίες στις επόμενες παραγράφους.

1.4.1 Τυποποιημένη Μέθοδος

Η τυποποιημένη μέθοδος προσδιορίζει τους συντελεστές στάθμισης που αντιστοιχούν σε διαφορετικούς τύπους περιουσιακών στοιχείων με χρήση εξωτερικών πιστοληπτικών αξιολογήσεων, στους οποίους έχει δοθεί έγκριση από την Τράπεζα Ελλάδος ή από τη CEBS (Committee of European Banking Supervisors), για τον υπολογισμό των σταθμισμένων περιουσιακών στοιχείων και τον τελικό προσδιορισμό της κεφαλαιακής επάρκειας. Ο συντελεστής στάθμισης για τα κρατικά, τραπεζικά και εταιρικά ανοίγματα διαφοροποιείται ανάλογα με τις εξωτερικές αξιολογήσεις της πιστοληπτικής ικανότητας. Όσο χαμηλότερος είναι ο συντελεστής στάθμισης τόσο χαμηλότερο είναι το RWA. Τα περιουσιακά στοιχεία υψηλότερου κινδύνου έχουν συντελεστή στάθμισης 150%.

Ο παρακάτω πίνακας απεικονίζει το συντελεστή στάθμισης για κάθε τύπο πελάτη ανάλογα με την εξωτερική αξιολόγησή του.

Αξιολόγηση	Συντελεστές
ΑΑΑ έως ΑΑ	0%
Α+ έως Α-	20%
ΒΒΒ+ έως ΒΒΒ-	50%
ΒΒ+ έως Β-	100%
Κάτω του Β-	150%
Χωρίς βαθμολογία	100%

Η μέθοδος αυτή αποτελεί βελτιωμένη εκδοχή της Βασιλείας Ι αντικατοπτρίζοντας με μεγαλύτερη ακρίβεια τις αλλαγές του κινδύνου και περιορίζει τις κεφαλαιακές απαιτήσεις εφαρμόζοντας τεχνικές μείωσης έναντι αυτού. Αποδεκτές τεχνικές μείωσης σύμφωνα με την Τράπεζα Ελλάδος είναι η Χρηματοδοτούμενη Πιστωτική Προστασία, όπως οι χρηματοοικονομικές εξασφαλίσεις, και η Μη-χρηματοδοτούμενη Πιστωτική Προστασία, όπως εγγυήσεις και πιστωτικά παράγωγα (ΤτΕ (2007)). Επίσης, παρέχει ένα μικρό κίνητρο για τις τράπεζες να χρησιμοποιούν διάφορους οργανισμούς αξιολόγησης πιστοληπτικής ικανότητας για να σταθμίσουν τα ανοίγματα τους προσφέροντας μεγαλύτερη ακρίβεια σε αντίθεση με την χρήση ενός οργανισμού. Επιπρόσθετα, μέσω της τυποποιημένης μεθόδου το ενδεχόμενο του " αρμπιτράζ" στον πιστωτικό κίνδυνο είναι περιορισμένο (Roy (2005)).

1.4.2 Μέθοδος Εσωτερικών Διαβαθμίσεων

Η μέθοδος εσωτερικών διαβαθμίσεων προϋποθέτει την έγκριση της Εποπτικής Αρχής και επιβάλλεται η τήρηση αυστηρών κριτηρίων σχετικά με τη μεθοδολογία διαχείρισης και διαβάθμισης των ανοιγμάτων και την παροχή στοιχείων. Αποτελεί τον κορμό για την ανάπτυξη της πιστωτικής πολιτικής και κουλτούρας του χρηματοπιστωτικού ιδρύματος. Το τελευταίο βαθμολογεί τους πελάτες του ανάλογα με την πιστοληπτική τους ικανότητα επισυνάπτοντας ένα score ή rating στον καθένα με βάση με τα χαρακτηριστικά του και παρακολουθεί συστηματικά τόσο το χαρτοφυλάκιο όσο και τον ίδιο τον πιστωτικό κίνδυνο

Ο πιστωτικός κίνδυνος, όπως προαναφέρθηκε παραπάνω, είναι συνάρτηση ενός αριθμού παραγόντων - πιθανότητα αθέτησης, ποσοστιαία ζημιά λόγω αθέτησης, έκθεση σε αθέτηση και λήξη του ανοίγματος. Οι μεθοδολογίες IRB παράγουν εκτιμήσεις αυτών των βασικών πιστωτικών μεταβλητών από ένα κατάλληλο και εγκεκριμένο μοντέλο εσωτερικού κινδύνου. Αυτές οι τιμές παράγουν ένα RWA και μια κεφαλαιακή απαίτηση μέσω της εφαρμογής ενός κανονιστικού τύπου.

Σε αντίθεση με την τυποποιημένη μέθοδο, η μέθοδος εσωτερικών διαβαθμίσεων παρέχει στις τράπεζες τη δυνατότητα να εφαρμόζουν πολύ μικρότερη διαφοροποίηση μεταξύ των κινδύνων σε σχέση με τους τυποποιημένους.

Υπάρχουν δύο προσεγγίσεις IRB για τα ανοίγματα σε πελάτες. Και οι δύο προσεγγίσεις υπόκεινται σε αυστηρές μεθοδολογικές προδιαγραφές, καθώς υπόκεινται στην έγκριση των τραπεζικών εποπτικών αρχών. Οι προσεγγίσεις αυτές είναι οι εξής:

- Θεμελιώδης Μέθοδος Εσωτερικών Διαβαθμίσεων : οι τράπεζες επιτρέπεται να χρησιμοποιούν τις δικές τους εσωτερικές αξιολογήσεις κινδύνου αλλά πρέπει να χρησιμοποιούν παραμέτρους EAD, LGD και M που καθορίζονται από κανονιστικές ρυθμίσεις.
- Προηγμένη Μέθοδος Εσωτερικών Διαβαθμίσεων : οι τράπεζες μπορούν - με την έγκριση των εποπτικών αρχών - να χρησιμοποιούν μοντέλα πιστωτικού κινδύνου για να παράγουν τις δικές τους παραμέτρους για όλα τα στοιχεία (PD, EAD, LGD και M). Για να λάβουν έγκριση, πρέπει να ικανοποιούν τους εποπτικούς φορείς έτσι ώστε, τόσο αρχικά όσο και σε συνεχή βάση, να πληρούν λεπτομερείς απαιτήσεις που καλύπτουν δραστηριότητες όπως η δημιουργία μοντέλων, η επικύρωση, η εποπτεία και η έγκριση, η χρήση εσωτερικών αξιολογήσεων και η γνωστοποίηση.

Υπάρχει εν γένει κίνητρο για τις τράπεζες να μετακινηθούν από τις προσεγγίσεις της SA σε IRB επειδή η κεφαλαιακή μεταχείριση είναι συνήθως λιγότερο δαπανηρή. Ωστόσο, οι υπολογισμοί IRB απαιτούν μεγαλύτερη επένδυση σε συστήματα. Μικρότερες, λιγότερο εξελιγμένες τράπεζες έχουν την τάση να επιλέγουν την τυποποιημένη προσέγγιση.

1.4.3 Βασιλεία III και υπολογισμός σταθμισμένων περιουσιακών στοιχείων

Στο πλαίσιο της Βασιλείας II εισήχθησαν χαμηλότεροι συντελεστές στάθμισης για τους πελάτες με καλύτερες αξιολογήσεις πιστοληπτικής ικανότητας για την αντιμετώπιση των επικρίσεων της Βασιλείας I. Ο κύριος λόγος για αυτό ήταν ότι η έλλειψη διαφοροποίησης των κινδύνων ενθάρρυνε τις τράπεζες να δανείζουν στους πελάτες με χαμηλή βαθμολογία, καθώς οι κεφαλαιακές απαιτήσεις ήταν οι ίδιες για τους πελάτες AAA και BB-.

Με την επιφύλαξη κανονιστικής έγκρισης, η Βασιλεία II επέτρεψε επίσης στις τράπεζες να χρησιμοποιούν τα δικά τους μοντέλα με το κίνητρο να είναι χαμηλότερα τα RWA και συνεπώς χαμηλότερες κεφαλαιακές απαιτήσεις απ’ ό,τι εάν χρησιμοποιούσαν την τυποποιημένη προσέγγιση.

Όταν αυτές οι αλλαγές εισήχθησαν στα μέσα της δεκαετίας του 2000, η οικονομία στις περισσότερες χώρες είχε ρυθμούς ανόδου και υπήρχαν περιορισμένα δεδομένα

αθέτησης κατά την κατασκευή και επικύρωση PD και άλλων μοντέλων. Επιπλέον, οι επιχειρηματικές πιέσεις ήταν τέτοιες που οι τράπεζες επικεντρώνονταν περισσότερο στην αύξηση των κερδών από τη διαχείριση κινδύνων. Αυτό είχε ως αποτέλεσμα την πίεση για ανάπτυξη των επιχειρήσεων, αλλά χωρίς ταυτόχρονα αύξηση των RWA.

Η παγκόσμια χρηματοπιστωτική κρίση αποκάλυψε μια σειρά ελαττωμάτων στις ρυθμιστικές απαιτήσεις και πρακτικές, καθώς και στη διαχείριση του τραπεζικού κινδύνου. Επίσης, υπογράμμισε το γεγονός ότι, πολλές τράπεζες όχι μόνο είχαν ανεπαρκή επίπεδα κεφαλαίου, αλλά ορισμένες μορφές κεφαλαίου δεν είχαν επαρκή απορροφητικότητα ζημιών.

Η Βασιλεία ΙΙΙ εισήγαγε ένα σύνολο κριτηρίων καταλληλόλητας τόσο για τα Κεφάλαια Tier 1 και Tier 2 όσο και για RWA. Το Κεφάλαιο Tier 1 υποδιαιρείται στο Κεφάλαιο Κοινών Μετοχών Tier 1 και σε ένα συμπληρωματικό κεφάλαιο. Το Κεφάλαιο Κοινών Μετοχών Tier 1 θεωρείται ως η πιο απορροφητική μορφή κεφαλαίου. Η Βασιλεία ΙΙΙ αύξησε το ποσοστό του κεφαλαίου που πρέπει να διατηρήσουν οι τράπεζες από 2% σε 4.5% και επίσης επέβαλε αυστηρότερους κανόνες επιλεξιμότητας για να αποκλείσει διάφορες μορφές χρέους που είχαν γίνει αποδεκτές στο παρελθόν και δεν κατάφεραν να παράσχουν το αναμενόμενο επίπεδο προστασίας. Στην ίδια βάση λειτουργεί και το συμπληρωματικό κεφάλαιο, όπου επίσης παρέχει απορρόφηση ζημιών σε συνεχή βάση, έχει παρόμοια χαρακτηριστικά κινδύνου και αυξήθηκε το ποσοστό του από 4% σε 6%. Το Κεφάλαιο Tier 2 αποτελείται από στοιχεία που παρέχουν την απορρόφηση ζημιών στη βάση της συνεχιζόμενης δραστηριότητας, δηλαδή διαθέτει η τράπεζα πόρους που απαιτούνται για να λειτουργεί έως ότου παρέχει στοιχεία που το αποδεικνύουν. Αντίστοιχα για τους υπολογισμούς RWA, που είναι ένα μέτρο των ανοιγμάτων μιας τράπεζας, περιλαμβάνει την χρήση διαφορετικών μεθοδολογιών για κάθε τύπο κινδύνου (πιστωτικός, λειτουργικός, αγοράς).

Παρά τις αλλαγές μετά την κρίση και τη βελτίωση της κανονιστικής εποπτείας και πρόκλησης, υπήρχαν ακόμη ανεπάρκειες στον υπολογισμό του RWAs. Οι ανεπάρκειες αυτές σχετίζονται με:

- Κίνδυνοι που ελλοχεύουν στα μοντέλα: οι τιμές των PD, LGD και EAD προέρχονται από τα μοντέλα όπου υπάρχει κίνδυνος ως προς την εφαρμογή και οι τιμές αυτές χρησιμοποιούνται για την εκτίμηση των RWAs. Επιπλέον, η ερμηνεία των αποτελεσμάτων των μοντέλων μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα αν δεν ερμηνευτεί σωστά.
- Περιπλοκότητα: Οι κανονιστικές αρχές παραμένουν περίπλοκες και υπάρχει πιθανότητα να ερμηνευτούν ή να εφαρμοστούν λανθασμένα.

- Εξωτερικές αξιολογήσεις: Εξακολουθεί να υπάρχει κάποια εξάρτηση από εξωτερικές αξιολογήσεις, παρόλο που οι οργανισμοί πιστοληπτικής ικανότητας έχουν βελτιώσει την μεθοδολογία τους και την διαφάνεια τους.
- Πελάτες χωρίς αξιολόγηση: Υπάρχει μεγάλος αριθμός πελατών χωρίς πιστοληπτικό προϊστορικό.

Στην κατεύθυνση αυτή, ο **Δείκτης Μόχλευσης (Leverage Ratio)** έρχεται να συμπληρώσει την αναδιαμόρφωση των κανόνων κεφαλαιακής επάρκειας που επιφέρει η Βασιλεία ΙΙΙ και πρέπει να ικανοποιεί τον παρακάτω περιορισμό:

$$Δείκτης\ Μόχλευσης\ = \frac{Κεφάλαιο\ Tier\ 1}{Συνολική\ Έκθεση} \geq 3\%$$

Ο δείκτης Μόχλευσης εκφράζεται ως ποσοστό, μολονότι η Μόχλευση παρουσιάζεται συνήθως ως ακέραιος αριθμός. Θέτοντας ένα ανώτερο όριο στην Μόχλευση ενός χρηματοπιστωτικού ιδρύματος τίθεται και ανώτερο όριο στον κίνδυνο που αναλαμβάνει το ίδρυμα. Παράλληλα, ο δείκτης αυτός παρουσιάζει την ικανότητα των τραπεζών να ανταπεξέλθουν στις δύσκολες καταστάσεις.

2. Ανάλυση Επιβίωσης

2.1 Εισαγωγή

Η ανάλυση επιβίωσης αναφέρεται σε στατιστικές διαδικασίες που χρησιμοποιούνται για την ανάλυση δεδομένων όπου το επιθυμητό αποτέλεσμα είναι ο χρόνος εμφάνισης ενός γεγονότος, δηλαδή αποτελεί μια μέθοδο εκτίμησης του χρόνου μέχρι να συμβεί το συμβάν που μελετάται (Kleinbaum (1998)). Συχνά αναφέρεται ως **χρόνος επιβίωσης** για παράδειγμα σε ιατρικές μελέτες με πρόβλεψη τον θάνατο ενός ασθενούς, **χρόνος εκδήλωσης** σε μηχανικές εφαρμογές με πρόβλεψη την βλάβη μιας μηχανής ή **χρόνος αποτυχίας** σε οικονομικές επιστήμες με πρόβλεψη την αποτυχία εκπλήρωσης των υποχρεώσεων ενός δανειολήπτη. Στον τραπεζικό τομέα, στο πλαίσιο του πιστωτικού κινδύνου η ανάλυση επιβίωσης χρησιμοποιήθηκε πρώτη φορά από τον Narain (1992). Ο Narain εφάρμοσε το εκθετικό μοντέλο επιταχυνόμενων χρόνων ζωής σε δεδομένα προσωπικών δανείων και εκτίμησε τον αριθμό αποτυχιών σε κάθε χρονικό διάστημα, αποδεικνύοντας ότι οι αξιολογήσεις πιστοληπτικής ικανότητας μπορούν να βελτιωθούν με την χρήση της ανάλυσης επιβίωσης σε σχέση με την πολλαπλή παλινδρόμηση.

Επιπρόσθετα, ο Banasik (1999) ενίσχυσε αυτή τη διαπίστωση συγκρίνοντας την απόδοση του εκθετικού, Weibull και Cox μοντέλου με την λογιστική παλινδρόμηση και υπέδειξε ότι μπορεί να μοντελοποιηθεί ο χρόνος αθέτησης και όχι απλά να μελετηθεί η πιθανότητα αν ο αιτών δανείου αθετήσει ή όχι την δανειακή του υποχρέωση. Με την πάροδο του χρόνου πολλοί συγγραφείς υιοθέτησαν την ίδια διαδικασία και ανέπτυξαν πιο εξελιγμένα μοντέλα ανάλυσης επιβίωσης. Οι Stepanova and Thomas (2002) επέκτειναν τα μοντέλα Cox PH (μοντέλο αναλογικού κινδύνου του Cox) και AFT (μοντέλα επιταχυνόμενου χρόνου αποτυχίας) χρησιμοποιώντας, μεταξύ άλλων την τεχνική της ομαδοποίησης (coarse classification), και αργότερα εισήχθηκε στα μοντέλα και χρονικά μεταβαλλόμενη συν διακύμανση . Το επόμενο στάδιο έρευνας συμπεριέλαβε τα mixture cure models όπου το δείγμα αποτελείται ένα μείγμα ‘ευαίσθητων ατόμων‘ που μπορεί να βιώσουν το γεγονός που μελετάται και ‘μη ευαίσθητα άτομα’ που ποτέ δεν θα το βιώσουν. Οι Tong at el (2012) επισήμαναν ότι η ανάλυση επιβίωσης μπορεί να είναι χρήσιμη για ακριβή εκτίμηση της πιθανότητας αθέτησης για καθορισμένο 12μηνο, ορίζοντάς την για διάφορους τύπους δανείων, το οποίο είναι χρήσιμο για την εκτίμηση των πιθανοτήτων αθέτησης με βάση το Σύμφωνο της Βασιλείας II.

Στην εργασία των Zhang and Thomas (2012) για πρώτη φορά συγκρίθηκε η απόδοση μοντέλων ανάλυσης επιβίωσης σε ένα δείγμα. Στην πλειοψηφία των παραπάνω εργασιών τα κριτήρια αξιολόγησης των μοντέλων εστιάζουν στο ποσοστό ορθής ταξινόμησης και στο εμβαδόν κάτω από την καμπύλη ROC, γνωστό ως AUC (Area Under the Curve). Ο δείκτης AUC

αποτελεί έναν δείκτη διαχωρισμού μεταξύ των καλών και κακών οφειλετών. Όσο μεγαλύτερη είναι η τιμή του εμβαδού κάτω από την καμπύλη, τόσο μεγαλύτερη είναι η ακρίβεια του διαγνωστικού ελέγχου προς την ταξινόμηση των οφειλετών. Σημαντική ήταν επίσης η εργασία των Dirick, Claeskens & Baesens (2017) όπου συμπεριελάμβανε όλα τα παραπάνω μοντέλα και τα κριτήρια αξιολόγησης σε 5 διαφορετικά δάνεια.

Στον ακόλουθο πίνακα παρουσιάζονται με χρονική ακολουθία οι συγγραφείς, η επιλογή μοντέλων καθώς και το κριτήριο αξιολόγησης.

Δημοσίευση	Parametric/ AFT[2]	Cox PH[3]	AFT/ Cox PH + extensions	Non Parametric	Mixture cure	Multi-event mixture cure	Κριτήριο αξιολόγησης
Narain (1992)	X						Κανένα
Banasik *et al* (1999)	X	X					Ταξινόμηση
Stepanova and Thomas (2001)		X	X				Ταξινόμηση, AUC, μέτρο κέρδους
Stepanova and Thomas (2002)		X	X				Ταξινόμηση, AUC
Bellotti and Crook (2009)			X				Κόστος ανεπιθύμητου αποτελέσματος
Cao et al (2009)	X	X		X			AUC
Tong *et al* (2012)		X			X		AUC, *H*-measure, Kolmogorov–Smirnov

[2] Accelerated Failure Time model
[3] Cox proportional-hazards model

Zhang and Thomas (2012)	X	X	X				Λάθος στην πρόβλεψη αθέτησης χρόνου
Dirick *et al* (2015)					X	X	AUC
Lore Dirick, Gerda Claeskens & Bart Baesens (2017)	X	X	X	X	X	X	AUC, πρόβλεψη χρόνου αθέτησης, οικονομικά μέτρα

Γράφημα 3: Υπάρχουσα Βιβλιογραφία της ανάλυσης επιβίωσης στον πιστωτικό κίνδυνο (Journal of the Operational Research Society).

2.2 Αρχές της ανάλυση επιβίωσης

Στην ανάλυση επιβίωσης, στόχος είναι συμπεριφορά του ατόμου σε σχέση με το χρόνο και ειδικότερα ο χρόνος που μεσολαβεί μέχρι κάποιο συγκεκριμένο συμβάν. Για την ευκολότερη ερμηνεία συχνή είναι η στατιστική απεικόνιση της ανάλυσης επιβίωσης συναρτήσει του χρόνου, γνωστή ως καμπύλη επιβίωσης. Η κατανομή των χρόνων επιβίωσης μπορεί να περιγράφει από τρεις συναρτήσεις:

(α) τη συνάρτηση επιβίωσης (S),

(β) τη συνάρτηση πυκνότητας πιθανότητας (p.d.f.), και

(γ) τη συνάρτηση κινδύνου (h).

Η συνάρτηση επιβίωσης μπορεί να εκφραστεί ως η πιθανότητα επιβίωσης ενός ατόμου πέραν της χρονικής στιγμής t ή η πιθανότητα ένα άτομο να βιώσει το γεγονός που μελετάται μετά τη χρονική στιγμή t. Εδώ το άτομο είναι ο αιτών και γεγονός είναι η αθέτηση στον πιστωτικό κίνδυνο σε χρόνο t. Η συνάρτηση επιβίωσης είναι μη αρνητική, φθίνουσα συνάρτηση του t με $S(0) = 1$ και $S(\infty) = 0$ και η καμπύλης της είναι ιδιαίτερα σημαντική για την εξαγωγή συμπερασμάτων όπως θα δούμε και παρακάτω. Δίνεται από την σχέση :

$$S(t) = P(T > t)$$

Η αντίστοιχη συνάρτηση πυκνότητας δίνεται από το όριο της πιθανότητας ο πελάτης να αθετήσει την υποχρέωση του σε ένα πολύ μικρό χρονικό διάστημα $(t, t + \Delta t)$ προς το αντίστοιχο χρονικό διάστημα Δt:

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P(t \le T \le t + \Delta t)}}{\Delta t} = -\frac{dS(t)}{dt} = \frac{dF(t)}{dt}$$

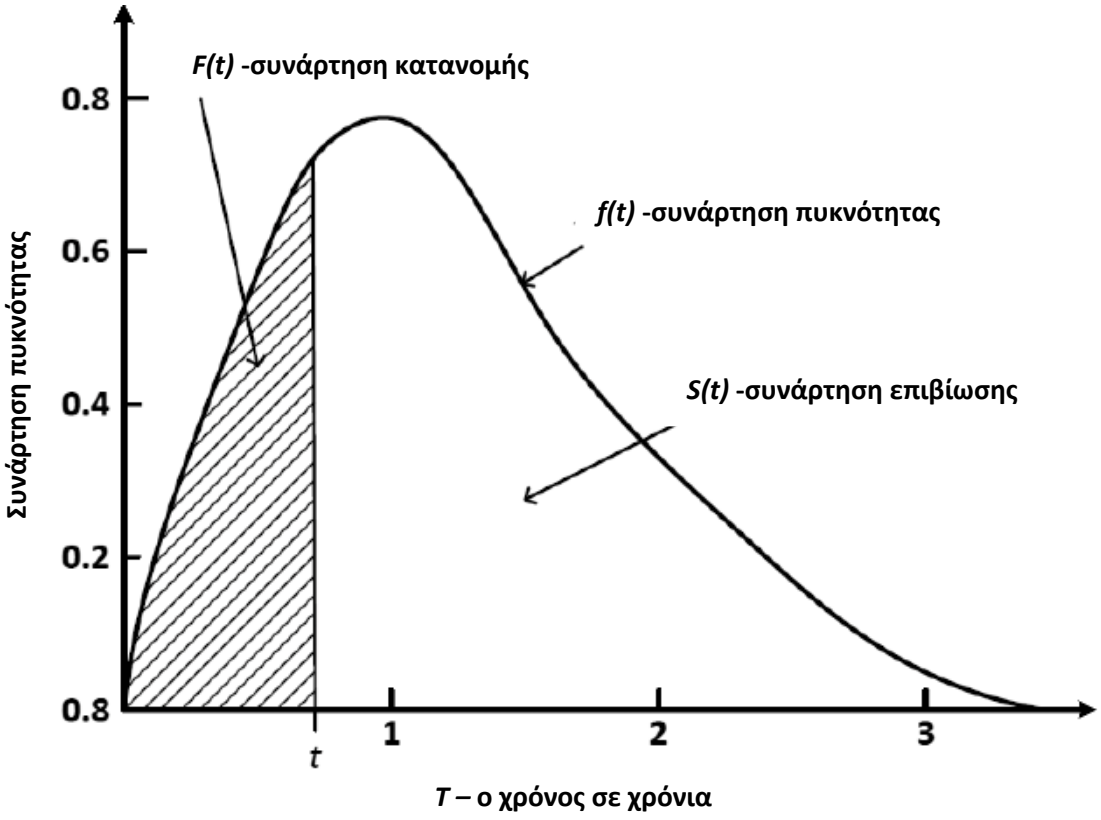
Η F(t) είναι η συνάρτηση κατανομής και η σύνδεσή της με την συνάρτηση επιβίωσης είναι η παρακάτω:

$$F(t) = P(T \le t) = 1 - S(t).$$

Η συνάρτηση κινδύνου ή ένταση θνησιμότητας $h(t)$ ορίζεται ως η πιθανότητα αποβίωσης (ή πραγμάτωσης του γεγονότος που εξετάζεται) τη χρονική στιγμή t, δεδομένου ότι το άτομο έχει επιβιώσει μέχρι τη χρονική στιγμή αυτή. Στον πιστωτικό κίνδυνο εκφράζει τον ρυθμό αποτυχίας υπό όρους δηλαδή ορίζεται ως η πιθανότητα αποτυχίας κατά τη διάρκεια ενός πολύ μικρού χρονικού διαστήματος, υποθέτοντας ότι ο πελάτης είναι φερέγγυος προς την τράπεζα στην αρχή του διαστήματος. Δίνεται από την εξής σχέση:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P(t \le T \le t + \Delta t | T > t)}}{\Delta t} = \frac{f(t)}{S(t)}.$$

Στο παρακάτω διάγραμμα απεικονίζονται οι παραπάνω συναρτήσεις ως προς το χρόνο καθώς και η σχέση μεταξύ τους:

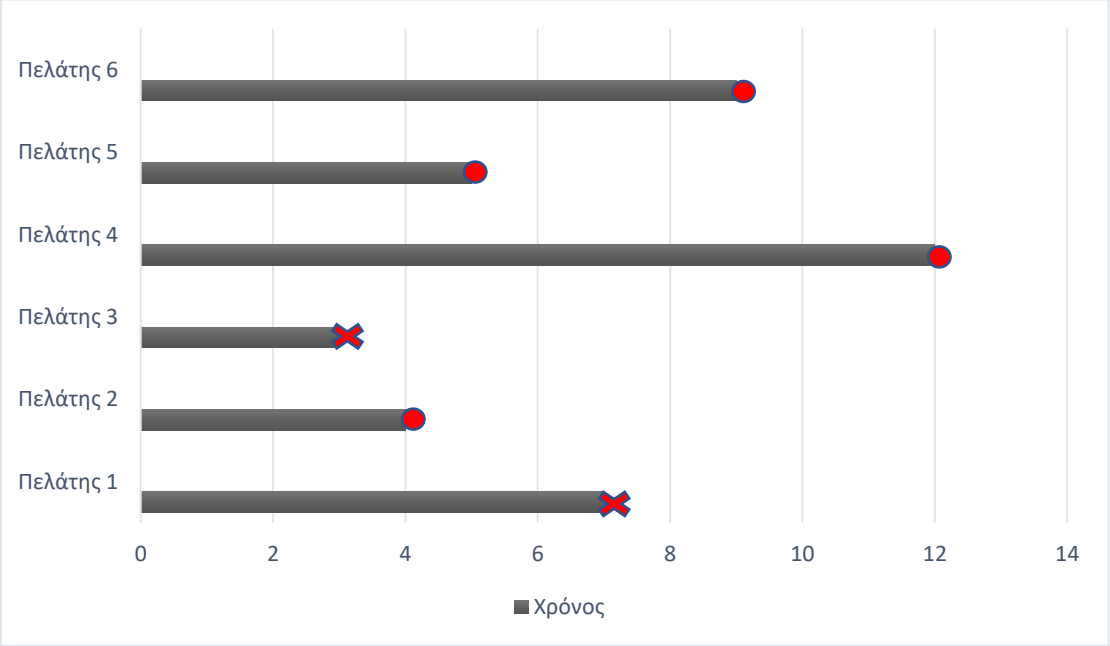


Γράφημα 4: Σχέσεις μεταξύ των S(t), F(t), f(t) και h(t) (CSUR (2017)).

2.3 Λογοκριμένα Δεδομένα (Censored data)

Ένα χαρακτηριστικό γνώρισμα της ανάλυσης επιβίωσης είναι ότι μπορεί να ενσωματώσει τα λογοκριμένα δεδομένα στην ανάλυση. Η λογοκρισία αναφέρεται στην ύπαρξη πληροφοριών σχετικά με το χρονικό διάστημα επιβίωσης των ατόμων, δηλαδή δεν γνωρίζουμε επακριβώς το χρόνο επιβίωσης. Στο δείγμα πιστοληπτικής ικανότητας παρατηρούνται λογοκριμένα δεδομένα για έναν από τους εξής λόγους:

- ένα άτομο δεν βιώνει το γεγονός πριν τελειώσει η μελέτη
- ένα άτομο χάθηκε από την καταγραφή-παρακολούθηση κατά τη διάρκεια της περιόδου μελέτης
- ένα άτομο αποσύρεται από τη μελέτη λόγω θανάτου (αν ο θάνατος δεν είναι το γεγονός ενδιαφέροντος) ή για κάποιο άλλο λόγο



Γράφημα 5: Ανάλυση επιβίωσης με ύπαρξη λογοκριμένων δεδομένων

Υπάρχουν διαφορετικά είδη λογοκριμένων δεδομένων, στην αξιολόγηση όμως της πιστοληπτικής ικανότητας εντοπίζονται μόνο δεξιά λογοκριμένα δεδομένα. Περιπτώσεις λογοκρισίας θεωρούνται ως δάνεια που δεν έτυχαν αθέτησης μέχρι τη στιγμή της συλλογής δεδομένων, η πρόωρη εξόφληση και οι ώριμες περιπτώσεις (ή πλήρεις, εκείνοι που φτάνουν στην προκαθορισμένη ημερομηνία λήξης πριν από τη στιγμή της συλλογής δεδομένων). Στην Εικόνα 2 καταγράφονται οι πληροφορίες για την εμφάνιση του συμβάντος και παρουσιάζονται τα δεδομένα που έχουν συλλεχθεί κατά τη διάρκεια 12 μηνών. Μπορούμε να διαπιστώσουμε ότι μόνο οι πελάτες 1 και 3 έχουν βιώσει το γεγονός (σημειωμένο με «Χ»)

κατά τη διάρκεια του χρόνου παρακολούθησης και ο χρόνος που παρατηρείται γι 'αυτούς είναι ο χρόνος της εκδήλωσης σε αντίθεση με τους πελάτες 2,4,5 και 6, οι οποίοι θεωρούνται ότι είναι λογοκριμένα δεδομένα και σημειώνονται με κόκκινους κύκλους στο σχήμα. Ειδικότερα, ο πελάτης 4 είναι λογοκριμένο δεδομένου ότι δεν συνέβη κανένα γεγονός κατά τη διάρκεια της περιόδου μελέτης, ενώ οι πελάτες 2,5 και 6 λογοκρίνονται λόγω της απόσυρσης ή απώλειας για παρακολούθηση εντός της χρονικής περιόδου μελέτης.

Είναι φανερό ότι η χρήση των λογοκριμένων δεδομένων στην μελέτη είναι σημαντική καθώς αν δεν συμπεριλαμβάνονταν, ο αριθμός του μεγέθους του δείγματος των πελατών προς έρευνα θα μειωνόταν σημαντικά. Για τον λόγο αυτό, το μοντέλο βαθμολόγησης της πιστοληπτικής ικανότητας πρέπει να λαμβάνει υπόψη όλες τις παρατηρήσεις, ακόμη και τις λογοκριμένες που δεν παραμένουν στην μελέτη όλο το χρονικό διάστημα. Είναι αναγκαίο να τονιστεί ότι οι παρατηρήσεις στην μελέτη είναι ανεξάρτητες, δηλαδή ένας πελάτης που είναι λογοκριμένος και φερέγγυος στο χρόνο *t* πρέπει να έχει τον ίδιο κίνδυνο αποτυχίας με έναν πελάτη που είναι μη λογοκριμένος.

2.4 Περικομμένα Δεδομένα (Truncated data)

΄Ένα άλλο είδος δεδομένων με ελλιπείς πληροφορίες όπως τα λογοκριμένα είναι τα περικομμένα δεδομένα για τα οποία δεν έχουμε καθόλου πληροφορίες. Στον αξιολόγηση πιστοληπτικής ικανότητας αφορά τα δάνεια που κινδύνευαν πριν από την είσοδο στη μελέτη, όπως για παράδειγμα τα δάνεια που εισέρχονται στο χαρτοφυλάκιο σε ένα συγκεκριμένο σημείο επειδή αγοράζονται.

Η διαφορά μεταξύ της λογοκρισίας και των περικομμένων δεδομένων συχνά δεν γίνεται διακριτή. Η λογοκρισία είναι η περίπτωση κατά την οποία τα δάνεια είναι γνωστά ότι θα αθετήσουν εντός συγκεκριμένου χρόνου, αλλά ο ακριβής χρόνος για την αθέτηση δεν είναι γνωστός. Η περικοπή είναι η περίπτωση που τα δάνεια δεν περιλαμβάνονται στο σύνολο δεδομένων επειδή δεν έχουν μελετηθεί. Η ενσωμάτωση των περικομμένων δεδομένων είναι πέρα από το πεδίο αυτής της εργασίας.

3. Μέθοδοι Ανάλυσης Επιβίωσης

3.1 Εισαγωγή

Τα πιστωτικά μοντέλα χρησιμοποιούνται για να εκτιμήσουν αν και κατά πόσο η πίστωση «κινδυνεύει» σε περίπτωση αθέτησης ή λόγω αλλαγών σε διάφορους παράγοντες πιστωτικού κινδύνου. Η μοντελοποίηση του πιστωτικού κινδύνου επιτρέπει στις τράπεζες να αναλαμβάνουν και να εκτιμούν τα πιστωτικά ανοίγματα που αντιμετωπίζουν πιο αποτελεσματικά, εκτός από το να τους βοηθούν να υπολογίσουν πόσα κεφάλαια χρειάζονται για την προστασία από τέτοιους κινδύνους.

Η παγκόσμια χρηματοπιστωτική κρίση ανέδειξε σημαντικές ελλείψεις στη μέτρηση των κινδύνων και έθεσε εκτεταμένες ανησυχίες σχετικά με την ικανότητα των τραπεζών και άλλων ιδρυμάτων να χρησιμοποιούν μοντέλα για την αποτελεσματική μέτρηση της πιστωτικής έκθεσης. Είναι ζωτικής σημασίας οι τράπεζες να αναγνωρίζουν, να κατανοούν και να διαχειρίζονται τον κίνδυνο από τα μοντέλα προκειμένου να αποτρέψουν τη λήψη ακατάλληλων επιχειρηματικών αποφάσεων σε επίπεδο πελάτη, χαρτοφυλακίου ή στρατηγικού επιπέδου.

Οι μέθοδοι ανάλυσης επιβίωσης διακρίνονται σε 2 βασικές κατηγορίες, στις **στατιστικές μεθόδους** και τις **μεθόδους μηχανικής μάθησης.** Και οι δύο έχουν στόχο την πρόβλεψη του χρόνου επιβίωσης καθώς και την εκτίμηση της πιθανότητας επιβίωσης στον αντίστοιχο εκτιμώμενο χρόνο. Ωστόσο παρουσιάζουν διαφορές ως προς τον χειρισμό των δεδομένων όπως είναι κατανοητό και από τον ορισμό τους. Οι στατιστικές μέθοδοι εστιάζουν περισσότερο τόσο στην εύρεση των κατανομών του χρόνου επιβίωσης όσο και στην εκτίμηση των ιδιοτήτων των παραμέτρων με την χρήση επίσης των καμπυλών επιβίωσης. Οι μέθοδοι μηχανικής μάθησης συνδυάζουν τις σύγχρονες τεχνικές με την δύναμη των παραδοσιακών μεθόδων μέσω αλγορίθμων για την πρόβλεψη του συμβάντος και χρησιμοποιούνται για δεδομένα μεγάλης διάστασης.

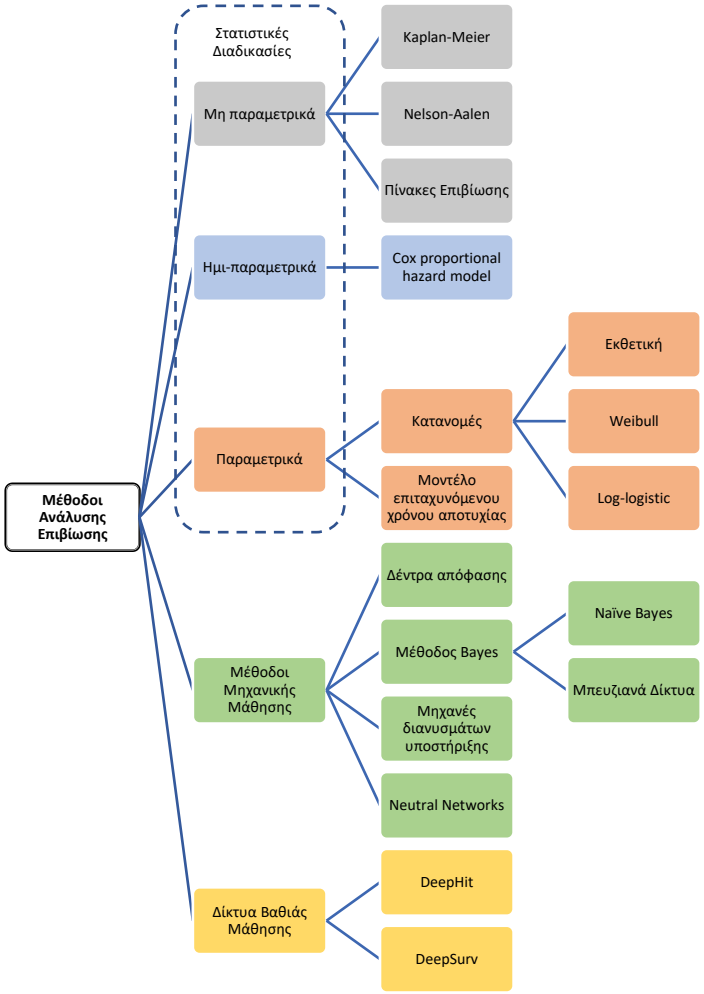
Κάθε μία από αυτές τις κατηγορίες υποδιαιρείται σε υποκατηγορίες. Οι 3 βασικές υποκατηγορίες των στατιστικών διαδικασιών είναι τα μη- παραμετρικά, τα ημι-παραμετρικά και τα παραμετρικά μοντέλα. Αντίστοιχα, οι μέθοδοι τεχνικής μάθησης αποτελούνται από δέντρα απόφασης, μάθηση κατά Bayes, μηχανές διανυσμάτων υποστήριξης,νευρωνικά δίκτυα και προηγμένες μεθόδους τεχνικής μάθησης.

Στην κατηγορία των μη παραμετρικών μοντέλων που χρησιμοποιούνται για τον υπολογισμό της εκτιμήτριας συνάρτησης επιβίωσης είναι η εκτιμήτρια Kaplan-Meier (KM), η Nelson-Aalen (NA) και οι πίνακες επιβίωσης (Life-Table ,LT). Η KM για τον υπολογισμό της πιθανότητας επιβίωσης λαμβάνει υπόψη ότι έχει επιβιώσει το άτομο σε όλο το

προηγούμενο χρονικό διάστημα, δηλαδή είναι προϊόν της ίδιας εκτίμησης μέχρι τον προηγούμενο χρόνο και αποτελεί το παρατηρούμενο ποσοστό επιβίωσης για αυτόν τον δεδομένο χρόνο. Για τον λόγο αυτό αναφέρεται και ως εκτιμήτρια Product limit όπως αναφέρουν το 1958 οι ίδιοι Kaplan-Meier αλλά και το 2003 οι Lee and Wang. Η NA εκτιμά την αθροιστική συνάρτηση κατανομής κινδύνου. Οι πίνακες επιβίωσης αναπτύχθηκαν νωρίτερα από την KM το 1958 από τους Cutler and Ederer αλλά η KM αποδείχτηκε ισχυρότερη. Παράγουν παρόμοιες καμπύλες επιβίωσης με την KM με τη διαφορά ότι οι πίνακες επιβίωσης παρουσιάζουν ιδιαίτερη ευαισθησία ως προς τον τρόπο που θα διαιρεθεί το χρονικό διάστημα μελέτης σε υπό-διαστήματα, με καλύτερη απόδοση όταν είναι άγνωστος ο χρόνος πραγμάτωσης τους συμβάντος. Για την εκτιμήτρια KM θα δοθούν περισσότερες πληροφορίες στην Ενότητα 3.2.1.

Στις παραμετρικές μέθοδοι χρησιμοποιούνται συγκεκριμένες κατανομές όπως η Εκθετική, η Weibull και η Log-logistic, μέθοδος γραμμικής παλινδρόμησης καθώς και το μοντέλο επιταχυνόμενου χρόνου αποτυχίας.

Δίνεται παρακάτω μία πλήρης απεικόνιση των μεθόδων αυτών και των υποδιαιρέσεων τους.



Γράφημα 6: Απεικόνιση των μεθόδων ανάλυσης επιβίωση

3.2 Στατιστικές μέθοδοι

Στην ενότητα αυτή θα αναφερθούν τα πλεονεκτήματα και τα μειονεκτήματα κάθε υποκατηγορίας καθώς και τα κυριότερα μοντέλα που χρησιμοποιούνται για την ανάλυση επιβίωσης κάθε υποκατηγορίας. Τα πλεονεκτήματα και τα μειονεκτήματα των 3 αυτών υποκατηγοριών μπορούν να βρεθούν στους Kalbfleisch and Prentice (1980), Lawless (1982), Cox and Oakes (1984) και πιο πρόσφατα με σημαντική συνεισφορά στην σύγκριση μεθόδων στους Keiding and Andersen (2006).

Τύπος	Πλεονεκτήματα	Μειονεκτήματα	Είδη μεθόδων
Μη-παραμετρικά	• Εύκολα στην εφαρμογή και στην ερμηνεία τους • Μεγαλύτερη αποδοτικότητα όταν υπάρχει έλλειψη γνωστών θεωρητικών κατανομών • Χρήση σε μικρά δείγματα πληθυσμών	• Έλλειψη ομαλότητας (smooth) των καμπυλών επιβίωσης • Δεν μπορεί να υπολογιστεί η συνάρτηση κινδύνου (κυρίως στο KM) • Ύπαρξη κατηγορικών μεταβλητών (λίγων) και όχι ποσοτικών (KM)	• Kaplan-Meier (KM) / Product Limit • Πίνακες επιβίωσης • Nelson-Aalen
Παραμετρικά	• Συνεχείς συναρτήσεις οδηγούν σε πιο ομαλή καμπύλη της συνάρτησης επιβίωσης • Εύκολη ενσωμάτωση μεταβλητών ελέγχου στο μοντέλο και στην διαδικασία συμπερασμάτων	• Αν θεωρηθεί λανθασμένα ότι ακολουθούν τα δεδομένα μια κατανομή , τα αποτελέσματα θα είναι λάθος	• Εκθετική • Weibull • Log-logistic • Μοντέλο επιταχυνόμενου χρόνου αποτυχίας • Μέθοδος Buckley -James (Μοντέλο γραμμικής παλινδρόμησης) • Penalized Παλινδρόμηση (Ποινικοποιημένη)
Ημι-παραμετρικά	• Εύκολη ενσωμάτωση μεταβλητών ελέγχου στο μοντέλο και στην διαδικασία συμπερασμάτων, πιο εύκολη από παραμετρικά • Εξομαλύνει τα “θορυβώδη” δεδομένα	• Δεν είναι δυνατός ο υπολογισμός των ποσοστών (απόλυτες διαφορές) • Μπορεί να ερμηνεύσει μόνο από την άποψη των σχετικών διαφορών	• Μοντέλο του Cox • Cox Boost

		• Δεν υπολογίζει την τιμή συνάρτησης κινδύνου	
Υπόθεση αναλογικού κινδύνου	• Η τιμή του κινδύνου είναι ίδια καθ' όλη την διάρκεια της μελέτης		• Μοντέλο του Cox αν πληροί αυτή την προϋπόθεση • Κάποια παραμετρικά μοντέλα

3.2.1 Η εκτιμήτρια Kaplan-Meier

Η ύπαρξη λογοκριμένων δεδομένων σε ένα δείγμα καθιστά μη αποδοτική την χρήση της εμπειρικής συνάρτησης επιβίωσης ($\hat{S}(t)$). Σε αυτή την περίπτωση και κυρίως όταν γνωρίζουμε τους παρατηρούμενους χρόνους επιβίωσης αποτελεσματική εκτιμήτρια συνάρτηση είναι η Kaplan-Meier (KM). Ο KM εκτιμά τον διάμεσο χρόνο επιβίωσης και όχι την μέση τιμή, ενώ έχει επίσης τη δυνατότητα να εκτιμά την συνάρτηση επιβίωσης ακόμη και αν μόνο το 50% του δείγματος φτάσει στο τελικό σημείο της μελέτης.

Για τον υπολογισμό της KM χωρίζουμε το χρονικό διάστημα που πραγματοποιήθηκε η μελέτη σε τόσα υπό-διαστήματα όσα και τα γεγονότα που εκδηλώθηκαν, έστω ρ και δεδομένου ότι ισχύει $0 \leq t_1 \leq t_2 \leq \ldots\ldots \leq t_\rho \leq \infty$. Κάθε υπό-διάστημα κατασκευάζεται με τέτοιο τρόπο ώστε να περιλαμβάνει μια αποτυχία στην αρχή του διαστήματος (ή περισσότερες αν πραγματοποιούνται ταυτόχρονα), εκτός από το πρώτο υπό-διάστημα $(0, t_1)$ που είναι εξ ορισμού η αρχή ως να συμβεί η πρώτη αποτυχία. Συνεπώς, σε κάθε υπό-διάστημα της μορφής (t_i, t_{i+1}) εκδηλώνεται μια αποτυχία. Έστω n_i ο αριθμός των ατόμων που συνεχίζουν να παραμένουν στην μελέτη λίγο πριν το t_i, γνωστός και όχι λογοκριμένος και d_i ο αριθμός των παρατηρούμενων ατόμων που αποτυγχάνουν (αθετούν/πεθαίνουν) στο t_i. Η εκτιμήτρια KM της συνάρτησης επιβίωσης $S(t)$ δίνεται από την σχέση:

$$KM(t) = \hat{S}(t) = \prod_{i=1}^{\rho} \frac{(n_i - d_i)}{n_i}$$

Η KM θεωρείται ότι ακολουθεί προσεγγιστικά την κανονική κατανομή για μεγάλο δείγμα και ένας τύπος για το 100(1-α) % διάστημα εμπιστοσύνης, με Z_α το α-στο ποσοστημόριο της κανονικής κατανομής, σύμφωνα με τον εκθετικό τύπο του Greenwood είναι ο εξής:

$$\hat{S}(t) \pm Z_{1-\alpha/2} \sqrt{V\hat{a}r[\hat{S}(t)]}$$

όπου και $V\hat{a}r[\hat{S}(t)]$ η εκτίμηση της διασποράς της KM η οποία δίνεται από τον τύπο:

$$V\hat{a}r[\hat{S}(t)] = \hat{S}(t)^2 \sum_{i=1}^{\rho} \frac{d_i}{n_i(n_i - d_i)}$$

Μπορεί να εκτιμηθεί επίσης η αθροιστική συνάρτηση κινδύνου από τον παρακάτω τύπο μέσω του $KM(t)$] :

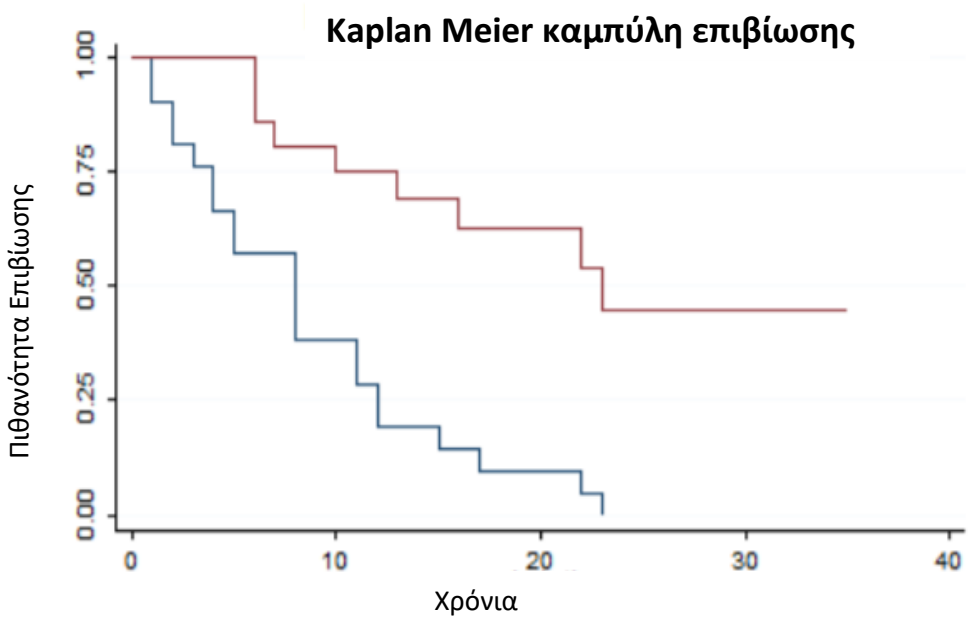
$$\widehat{H}(t) = -\log\left[KM(t)\right].$$

Ωστόσο για την αθροιστική συνάρτηση κινδύνου χρησιμοποιείται συνήθως ο τύπος των NA που είναι ο εξής:

$$\widehat{H}(t) = \sum_{i:t_i<t} \frac{d_i}{n_i}$$

Σε όλες τις παραπάνω σχέσεις το di μπορεί να πάρει τις τιμές 0 ή 1. Αξίζει να σημειωθεί ότι ο τύπος του Greenwood για το διάστημα εμπιστοσύνης παίρνει τιμές μεταξύ 0, 1 και όπως επισήμαναν οι Hosmer and Lemeshow (1999) συμπεριφέρεται καλά και σε μικρά δείγματα.

Η καμπύλη της ΚΜ έχει χαρακτηριστική σκαλωτή μορφή και αλλάζει επίπεδο κάθε φορά που πραγματοποιείται το γεγονός που μελετάται. Το παρακάτω σχήμα δείχνει δύο εκτιμήτριες ΚΜ για την πιθανότητα επιβίωσης σε σχέση με το χρόνο.



Γράφημα 7: Καμπύλη ΚΜ επιβίωσης

Για την σύγκριση των καμπυλών ΚΜ, τα πιο γνωστά κριτήρια είναι τα Mantel-Haenszel (LMH) και Gehan-Breslow-Wilcoxon. Το LMH αποτελεί ένα γενικευμένο log-rank τεστ και είναι κατάλληλο για χρήση όταν τα δεδομένα είναι στρωματοποιημένα και λογοκριμένα. Προσαρμόζει τις διακυμάνσεις και δίνει μια ολική σύγκριση των καμπυλών στηριζόμενο στη μηδενική υπόθεση ότι δεν υπάρχει καμία διαφορά μεταξύ των καμπυλών

$H_0 : S_1 = S_2$.

Το ΜΗ ακολουθεί μια κατανομή χ^2 με 1 βαθμό ελευθερίας οπότε αν $MH > X_1^2(\alpha)$ υπάρχουν επαρκή αποδεικτικά στοιχεία για να απορριφθεί η μηδενική υπόθεση, διαφορετικά οι δύο καμπύλες δίνουν ίδια αποτελέσματα.

Στο log-rank τεστ όλα τα σημεία t_i το ίδιο βάρος, έχουν την ίδια πιθανότητα να συμβεί το συμβάν που παρατηρείται. Επίσης, είναι πιο ισχυρό τεστ σε σύγκριση με το Gehan-Breslow-Wilcoxon τεστ όταν ισχύει η υπόθεση του αναλογικού κινδύνου στο δείγμα, δηλαδή ο λόγος των συναρτήσεων κινδύνου είναι σταθερός. Το Gehan-Breslow-Wilcoxon τεστ δίνει περισσότερο βάρος στους θανάτους-πραγματοποίηση συμβάντος σε πρώιμα χρονικά σημεία (Collett (2003)). Το τελευταίο αποτελεί τροχοπέδη όταν μεγάλο μέρος του δείγματος λογοκρίνεται σε πρώιμα χρονικά σημεία. Όσο αφορά τον λόγο συναρτήσεων κινδύνου το Gehan-Breslow-Wilcoxon τεστ δεν απαιτεί να είναι σταθερός, αλλά απαιτεί από τη μία ομάδα να έχει σταθερά υψηλότερο κίνδυνο από την άλλη.

Εάν οι δύο καμπύλες επιβίωσης διασταυρώνονται, τότε η μία ομάδα έχει υψηλότερο κίνδυνο σε πρώιμα χρονικά σημεία και η άλλη ομάδα έχει υψηλότερο κίνδυνο σε τελευταία. Αυτό θα μπορούσε απλώς να είναι σύμπτωση τυχαίας δειγματοληψίας, και η υπόθεση των αναλογικού κινδύνου θα μπορούσε να εξακολουθεί να ισχύει. Αλλά αν το μέγεθος του δείγματος είναι μεγάλο και οι καμπύλες επιβίωσης είναι κοντά η μία στην άλλη στη μέση του διαστήματος του χρόνου, ούτε το log-rank ούτε το Gehan-Breslow-Wilcoxon τεστ βοηθά στην σύγκριση.

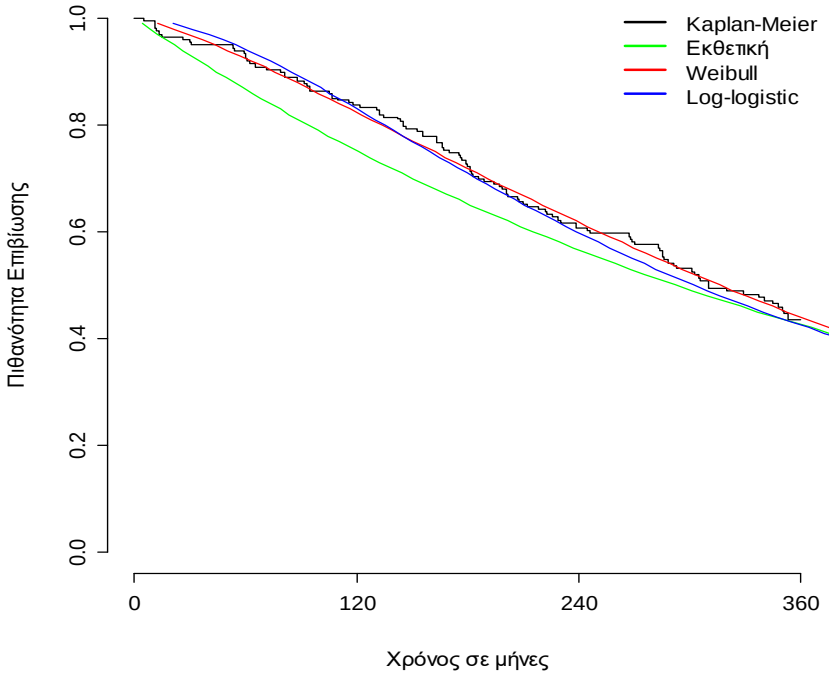
3.2.2 Μερικές παραμετρικές συναρτήσεις επιβίωσης

Η επιλογή παραμετρικών κατανομών σε ένα δείγμα με λογοκριμένα δεδομένα έχει αρκετά πλεονεκτήματα. Ο υπολογισμός των συναρτήσεων επιβίωσης και πυκνότητας πιθανότητας είναι ακριβής και οι στατιστικοί έλεγχοι είναι πιο αποτελεσματικοί και ισχυροί. Παρακάτω δίνεται ένας πίνακας με τα βασικά χαρακτηριστικά των κατανομών, Εκθετική, Weibull και Log-logistic.

Κατανομή	Παράμετρος	*f(t)*	*F(t)*	*S(t)*	*h(t)*	*H(t)*
Εκθετική	λ	$\lambda e^{-\lambda t}$	$1\text{-}e^{-\lambda t}$	$e^{-\lambda t}$	λ	λt
Weibull	Μορφή $\lambda>0$ Κλίμακα $k>0$	$k\lambda^k t^{k-1} e^{-(\lambda t)^k}$	$1\text{-}e^{-(\lambda t)^k}$	$e^{-(\lambda t)^k}$	$k\lambda^k t^{k-1}$	$(\lambda t)^k$
Log-logistic	Μορφή $\lambda>0$ Κλίμακα $k>0$	$\frac{(\frac{k}{\lambda})(\frac{t}{\lambda})^{k-1}}{(1+(\frac{t}{\lambda})^k)^2}$	$\frac{1}{1+(\frac{t}{\lambda})^{-k}}$	$\frac{1}{1+(\frac{t}{\lambda})^k}$	$\frac{k}{t(1+\left(\frac{t}{\lambda}\right)^{-k})}$	$\log(1+\left(\frac{t}{\lambda}\right)^k)$

Επιπρόσθετα σε σύγκριση με την εκτιμήτρια ΚΜ είναι πιο ομαλή η καμπύλη της συνάρτησης επιβίωσης σε δεδομένα που ακολουθούν κάποια από τις προαναφερθείσες

κατανομές, όπως φαίνεται και στο διάγραμμα παρακάτω, όπου απεικονίζονται οι τέσσερις καμπύλες. Χρησιμοποιήθηκε το πακέτο R με δεδομένα που υπάρχουν στην βάση του προγράμματος (creditR) και αναφέρονται σε 1000 οφειλέτες που έχουν ταξινομηθεί με βάση την πιθανότητα αθέτησης τους ή όχι.



Γράφημα 8*:* Καμπύλη επιβίωσης KM, Εκθετική, Weibull και Log-logistic.

3.2.3 Μοντέλο επιταχυνόμενου χρόνου αποτυχίας- ζωής

Το μοντέλο επιταχυνόμενου χρόνου αποτυχίας (AFT) ανήκει στην οικογένεια των παραμετρικών μοντέλων. Το AFT υποστηρίζει ότι η σχέση μεταξύ δύο συναρτήσεων επιβίωσης προσδιορίζεται με την χρήση των συμμεταβλητών και η επίδραση μιας συμμεταβλητής είναι να επιταχύνει ή να επιβραδύνει το χρόνο του συμβάντος ως προς μια σταθερά c όπως αναφέρουν οι Kalbfleisch and Prentice (1980). Αυτός είναι ο βασικός λόγος που το μοντέλο ονομάζεται επιταχυνόμενου χρόνου. Αν για κάθε χρονικό σημείο t ισχύει:

$$S_1(t) = S_2(ct), \;\; c > 0$$

δηλαδή ο χρόνος επιβίωσης του δείγματος 1 είναι c φορές του χρόνου επιβίωσης του δείγματος 2, το AFT μοντέλο μπορεί να μοντελοποιήσει την σχέση αυτή, θέτοντας μία συμμεταβλητή.

Η πιο απλή μορφή του AFT υποθέτει χρονικά ανεξάρτητες συμμεταβλητές και η συνάρτηση επιβίωσης και κινδύνου είναι οι παρακάτω:

$$S(t) = S_0[t\psi(z)] = S_0[texp(\beta x)]$$

$$\mathrm{h}(t) = \psi(z)h_0[t\psi(z)] = exp(\beta x)h_0[texp(\beta x)]$$

Το $\psi(z)$ αποτελεί τον παράγοντα επιτάχυνσης και είναι συνάρτηση των συμμεταβλητών z οι οποίες όπως αναφέρθηκε παραπάνω μειώνουν ή αυξάνουν τον χρόνο επιβίωσης. Η σχέση αυτή εδώ συνδέει την $S(t)$ με την συνάρτηση αναφοράς S_0, με την μια να είναι c φορές την άλλη. Κάτι παρόμοιο δεν ισχύει για την συνάρτηση κινδύνου $\mathrm{h}(t)$ με την αντίστοιχη h_0.

3.2.4 Πλήρως παραμετρικό μοντέλο αναλογικού κινδύνου

Ο Cox το 1972 περιέγραψε τα μοντέλα αναλογικού κινδύνου και αποτελεί ένα από τα σημαντικότερα στατιστικά έργα. Τα μοντέλα αυτά εκτιμούν την τιμή του κινδύνου σε αντίθεση με όλα τα προαναφερθείσα μοντέλα. Ο τύπος της συνάρτησης κινδύνου είναι ο εξής:

$$h(t/x) = h_0(t)exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_\rho x_\rho\right) = h_0(t)exp(\beta x)$$

Δηλαδή η $h(t/x)$ εξαρτάται από 2 παραμέτρους, την αναφορική συνάρτηση κινδύνου $h_0(t)$ και την επίδραση από την παράμετρο $exp(\beta x)$ που λειτουργεί με τον τρόπο που περιγράφτηκε παραπάνω. Συνεπώς, στο μοντέλο αναλογικού κινδύνου ισχύει ο παράγοντας επιτάχυνσης και στην συνάρτηση κινδύνου.

Η σύγκριση παραμετρικών μοντέλων γίνεται συχνά με το κριτήριο πληροφορίας του Akaike (Akaike information criterion, AIC) το οποίο είναι ένα μέτρο της καλής εφαρμογής των δεδομένων από το μοντέλο. Όσο πιο μικρή η τιμή του AIC τόσο καλύτερη η εφαρμογή του μοντέλου στα δεδομένα.

3.2.5 Το μοντέλο αναλογικού κινδύνου του Cox

Το μοντέλο αναλογικού κινδύνου του Cox αποτελεί ένα ημί-παραμετρικό μοντέλο όπου η χρήση του είναι αρκετά συχνή στην ανάλυση επιβίωσης και ειδικότερα σε περιπτώσεις ύπαρξης λογοκριμένων δεδομένων. Το μοντέλο χρησιμοποιεί λογοκριμένα δεδομένα αντί για τις πραγματικά και συνεπώς υπολογίζει τον σχετικό κίνδυνο και όχι τον πραγματικό. Θεωρείται ημί-παραμετρικό καθώς δεν χρειάζεται η γνώση της αναφορικής συνάρτησης κινδύνου $h_0(t)$. Η έλλειψη αυτή αποτελεί το μη παραμετρικό κομμάτι της

συνάρτησης και δίνει στο μοντέλο ευελιξία. Το υπόλοιπο μέρος συνεχίζει να εκτιμάται με τον ίδιο τρόπο που περιγράφτηκε στην ενότητα 3.2.4.

Το μοντέλο του Cox χρησιμοποιείται συχνά με την προϋπόθεση όμως ότι ισχύει η σχέση του αναλογικού κινδύνου καθ' όλη την διάρκεια της μελέτης. Δηλαδή ο λόγος των συναρτήσεων κινδύνου είναι πάντα ένας σταθερός αριθμός και ονομάζεται κινδυνότητα (Hazard Ratio-HR). Όταν δεν τηρείται η υπόθεση της αναλογικότητας, εφαρμόζουμε στρωματοποίηση των μεταβλητών αυτών. Αν πάρουμε δύο στρώματα της μεταβλητής x θα πρέπει να ισχύει η παρακάτω σχέση:

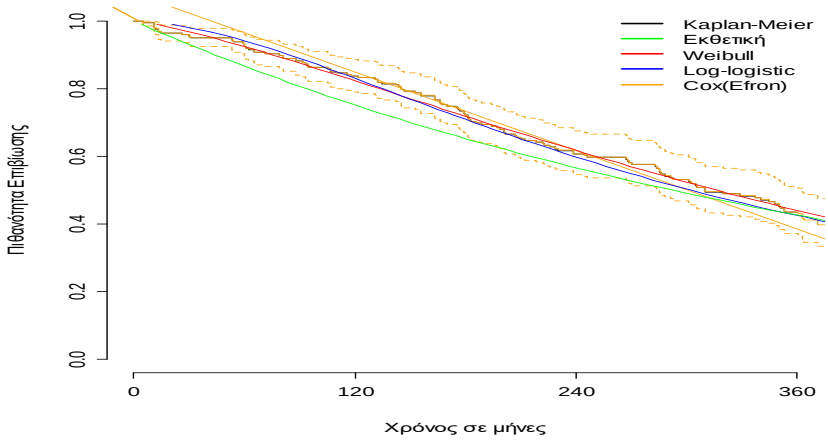
$$HR(t) = \frac{h(t/x_1)}{h(t/x_2)} = \frac{h_0(t)exp(\beta x_1)}{h_0(t)exp(\beta x_2)} = exp[\beta(x_1 - x_2)]$$

Ο Cox προτείνει την μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας για την εκτίμηση των συντελεστών β. Συνεπώς, σύμφωνα με τα παραπάνω και δεδομένου ότι ισχύει $0 \leq T_1 \leq T_2 \leq \ldots\ldots \leq T_\rho \leq \infty$ οι χρόνοι και $0 \leq R_1 \leq R_2 \leq \ldots\ldots \leq R_\rho \leq \infty$ τα άτομα που βρίσκονται σε κίνδυνο την αντίστοιχη χρονική στιγμή, ο εκτιμητής $\widehat{\beta}$ του β δίνεται από τον εξής τύπο:

$$\mathrm{L}(\beta) = \prod_{i=1}^{\rho} L_i(\beta) = \prod_{i=1}^{\rho} [\frac{\exp(x\mathrm{i}\beta)}{\sum_{j\varepsilon Ri} \exp(x\mathrm{i}\beta)}]$$

Ο εκτιμητής που προκύπτει είναι αμερόληπτος και ασυμπτωτικά κανονικός. Επίσης, η μεταβλητή στο μοντέλο αναλογικού κινδύνου μπορεί να είναι διακριτή αλλά και συνεχής.

Όσο αφορά την σύγκριση με τις υπόλοιπες μεθόδους προστίθεται στο γράφημα 9 και το μοντέλο αναλογικού κινδύνου του Cox παίρνουμε την παρακάτω απεικόνιση.



Γράφημα 9: Καμπύλη επιβίωσης ΚΜ, Εκθετική, Weibull, Log-logistic και Cox

Με το πέρασμα του χρόνου και τη συνεχή ανάπτυξη των τεχνικών συλλογής, ανίχνευσης δεδομένων και συσσώρευσης μεγάλης διάστασης δεδομένων το μοντέλο του Cox έχει παραλλαχθεί και εξελιχθεί. Ωστόσο στην εργασία αυτή γίνεται αναφορά μόνο στο αρχικό μοντέλο.

3.3 Λογιστική Παλινδρόμηση

Για την πρόβλεψη ενός δίτιμου αποτελέσματος, όπως η ταξινόμηση ενός αιτούντα δανείου σε καλό ή κακό, είναι γνωστή η επιλογή της λογιστικής παλινδρόμησης τα προηγούμενα χρόνια. Η μέθοδος ορίζει ως καλό ή κακό δανειολήπτη τον αιτούντα βάσει της απόδοσης της αποπληρωμής στην διάρκεια του χρόνου, και στο μοντέλο αποτελεί την εξαρτημένη μεταβλητή και στηρίζεται σε ανεξάρτητες μεταβλητές του δείγματος για την ταξινόμηση. Αυτές οι ανεξάρτητες μεταβλητές που έχουμε στη διάθεσή μας σχετικά με το δάνειο είναι είτε κατηγορικές είτε συνεχείς και μας βοηθούν να διαμορφώσουμε την πιθανότητα του συμβάντος (στην περίπτωσή μας την πιθανότητα αθέτησης). Αυτές οι μεταβλητές καλούνται επίσης μεταβλητές πρόβλεψης. Μερικά παραδείγματα μεταβλητών πρόβλεψης παρέχονται παρακάτω:

- Προσωπικά στοιχεία του οφειλέτη: ηλικία, κατάσταση απασχόλησης, επάγγελμα, εισόδημα, κατάσταση κατοικίας και αριθμός εξαρτώμενων ατόμων.
- Πιστωτικό ιστορικό: Διάρκεια πιστωτικού ιστορικού, αριθμός και αξία προηγούμενων δανείων, αριθμός και αξία προηγούμενων καθυστερούμενων δανείων.
- Δεδομένα συμπεριφοράς: Μοτίβο δαπανών, μοτίβα αποπληρωμής.

Η λογιστική παλινδρόμηση είναι ένα μη γραμμικό μοντέλο με σφάλματα ε_i που δεν ακολουθούν την κανονική κατανομή. Βασικός περιορισμός του μοντέλου είναι ο τύπος της εξαρτημένης μεταβλητής Y_i η οποία απαιτείται να είναι διακριτή. Το μοντέλο ορίζεται από την παρακάτω σχέση:

$$Y_i = E(Y_i) + \varepsilon_i$$

Ένα μέτρο μέτρησης της πιθανότητας να συμβεί ένα γεγονός σε σχέση με την πιθανότητα να μην συμβεί είναι η χρήση του λόγου

$$\frac{p}{1-p} = \frac{\pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ \nu\alpha\ \alpha\theta\varepsilon\tau\acute{\eta}\sigma\varepsilon\iota\ o\ o\varphi\varepsilon\iota\lambda\acute{\varepsilon}\tau\eta\varsigma}{\pi\iota\theta\alpha\nu\acute{o}\tau\eta\tau\alpha\ \nu\alpha\ \mu\eta\nu\ \alpha\theta\varepsilon\tau\acute{\eta}\sigma\varepsilon\iota\ o\ o\varphi\varepsilon\iota\lambda\acute{\varepsilon}\tau\eta\varsigma}$$

ο οποίος ονομάζεται odds. Η πιθανότητα αθέτησης με χρήση του μοντέλου λογιστικής παλινδρόμησης και ανεξάρτητες μεταβλητές x_i μπορεί να υπολογιστεί ως εξής:

$$P(\alpha\theta\acute{\epsilon}\tau\eta\sigma\eta\varsigma) = E(Y_i) = \frac{exp(a + \sum \beta_i x_i)}{1 + exp(a + \sum \beta_i x_i)} = \frac{1}{1 - exp[-(a + \sum \beta_i x_i)]}$$

Για τον προαναφερθέντα λόγο με βάση τον παραπάνω τύπο θα έχουμε:

$$\frac{p}{1-p} = \frac{\frac{exp(a + \sum \beta_i x_i)}{1 + exp(a + \sum \beta_i x_i)}}{1 - \frac{exp(a + \sum \beta_i x_i)}{1 + exp(a + \sum \beta_i x_i)}} = exp\left(a + \sum \beta_i x_i\right)$$

$$\ln\left(\frac{p}{1-p}\right) = \ln(exp(a + \sum \beta_i x_i)) = a + \sum \beta_i x_i$$

Οδηγούμαστε έτσι σε μια γραμμική συνάρτηση. Στόχος είναι η ερμηνεία των συντελεστών παλινδρόμησης. Το α είναι ίσο με τον λογάριθμο του odds όταν όλες οι ανεξάρτητες μεταβλητές πάρουν την τιμή 0 (κακός οφειλέτης εδώ). Τα β_i αποτελούν τους συντελεστές παλινδρόμησης και εκφράζουν το μέγεθος συνεισφοράς της αντίστοιχης μεταβλητής όταν αυτή πάρει την τιμή 1. Όσο μεγαλύτερη και θετική είναι η τιμή του συντελεστή, τόσο η πιθανότητα να συμβεί το γεγονός αυξάνεται ενώ η αρνητική τιμή σημαίνει ότι η επεξηγηματική μεταβλητή μειώνει την πιθανότητα αυτής της έκβασης. Υψηλή τιμή του συντελεστή σημαίνει ότι η ανεξάρτητη μεταβλητή επιδρά πολύ ισχυρά στην πιθανότητα να συμβεί το γεγονός ή μη, ενώ χαμηλή τιμή δηλώνει μικρή επιρροή της ανεξάρτητης μεταβλητής στην πιθανότητα εμφάνισης της ανάλογης έκβασης.

Ένα παράδειγμα εφαρμογής της λογιστικής παλινδρόμησης είναι αν θέσουμε ως ανεξάρτητη μεταβλητή το ετήσιο εισόδημα ενός οφειλέτη για την πρόβλεψη αθέτησης. Το ποσοστό των οφειλετών που αθέτησαν είναι μεγαλύτερο για αυτούς με χαμηλότερο εισόδημα συγκριτικά με αυτούς που έχουν υψηλό εισόδημα. Ως εκ τούτου όσο υψηλότερο το ετήσιο εισόδημα τόσο μικρότερη η πιθανότητα αθέτησης. Η εφαρμογή της λογιστικής παλινδρόμησης εμπίπτει στα δεδομένα της διπλωματικής παρακάτω και ένα από τα κριτήρια επιλογής του μοντέλου αποτελεί η εύκολη ερμηνεία του.

3.4 Μέθοδοι μηχανικής μάθησης

Η ανάγκη για την διαχείριση δεδομένων με μη γραμμικές σχέσεις, η πολυπλοκότητα και συγχρόνως η ακρίβεια της πρόβλεψης κινδύνου αποτελούν ισχυρό παράγοντα για την εισαγωγή και ευρεία χρήση μοντέλων μηχανικής μάθησης. Η ανάπτυξη προηγμένων αλγορίθμων προγνωστικής μοντελοποίησης εξόρυξης δεδομένων έχει στόχο την ενίσχυση της προγνωστικής ακρίβειας η οποία επιτυγχάνεται με την δημιουργία ενός μαθηματικού μοντέλου που βασίζεται στα δεδομένα εκπαίδευσης (training data). Σύμφωνα με τον Andriy Burkov το 2019 διακρίνονται σε τέσσερις κατηγορίες, εποπτευόμενα (supervised), ήμι-εποπτευόμενα (semi-supervised), μη εποπτευόμενα (unsupervised) και ενίσχυσης

(reinforcement). Τις τελευταίες δεκαετίες, αναπτύχθηκε πληθώρα προσεγγίσεων, κυρίως στον τομέα της μηχανικής μάθησης και βαθιάς μάθησης για την αντιμετώπιση του προβλήματος της πιστοληπτικής ποιότητας μιας εταιρείας, χρησιμοποιώντας τόσο ποσοτικές όσο και ποιοτικές πληροφορίες.

3.4.1 Δέντρα απόφασης

Οι δενδρικές μέθοδοι όπως αλλιώς ονομάζονται, έχουν πάρει το όνομά τους από το γεγονός ότι οι κανόνες διαχωρισμού, δηλαδή ο διαχωρισμός του συνόλου τιμών των επεξηγηματικών μεταβλητών σε υπό-περιοχές, μπορούν να απεικονιστούν υπό μορφή δέντρου. Ο διαχωρισμός αυτός γίνεται βάση ενός συγκεκριμένου κριτηρίου και στην ίδια υπό-περιοχή τοποθετούνται όμοια σύμφωνα με το κριτήριο αντικείμενα/άτομα. Αποτελούν μη παραμετρικά μοντέλα και η εφαρμογή τους επιτυγχάνεται τόσο με μοντέλα παλινδρόμησης όσο και με μοντέλα ταξινόμησης.

Η πρώτη αναφορά των δέντρων απόφασης στην ανάλυση επιβίωσης έγινε από τον Ciampi (1981), αλλά τέθηκαν υπό συζήτηση στο βιβλίο των Gordon and Olshen (1985). Τα κριτήρια διαχωρισμού στα δέντρα απόφασης επιβίωσης είναι δύο: (α) η ελαχιστοποίηση της ομοιογένειας εντός του κόμβου με στόχο την ελαχιστοποίηση του κόστους με χρήση πολλών διαφορετικών μεθόδων για τον υπολογισμό όπως ο εκτιμητής μέγιστης πιθανοφάνειας της εκθετικής συνάρτησης και (β) η μεγιστοποίηση της ετερογένειας μεταξύ διαφορετικών κόμβων με χρήση διάφορων μεθόδων όπως το τεστ log-rank. Τα δέντρα αποφάσεων είναι απλά μοντέλα παλινδρόμησης ή ταξινόμησης με εύκολη ερμηνεία, δηλαδή τα δέντρα απόφασης διακρίνονται σε δέντρα παλινδρόμησης και δέντρα ταξινόμησης.

Υπάρχουν διάφορες τεχνικές κατασκευής ενός δέντρου. Η πιο απλή και γνωστή είναι η CART (Classification and Regression Trees). Ο αλγόριθμος κατασκευής για τα **δένδρα παλινδρόμησης** ξεκινά εφαρμόζοντας την προσέγγιση recursive binary splitting στα δεδομένα εκπαίδευσης, δηλαδή το σημείο διαχωρισμού γίνεται με βάση τη μικρότερη τιμή του αθροίσματος των τετραγώνων των καταλοίπων (αποκλίσεων) των παρατηρήσεων

$$RSS = \sum_{j=1}^{J} \sum_{i \varepsilon Rj} (y_j - \widehat{y_{R_j}})^2$$

όπου $\widehat{y_{R_j}}$ να είναι η μέση τιμή της εξαρτημένης μεταβλητής για τις παρατηρήσεις εκπαίδευσης στην R_j περιοχή. Πιο συγκεκριμένα και με βάση όσα αναφέρθηκαν παραπάνω, αναζητούνται κατάλληλα j, s έτσι ώστε χρησιμοποιώντας τις περιοχές

$$R_1(j,s) = \{{}^{X}/_{X_j} < s\} \qquad \text{και} \qquad R_2(j,s) = \{{}^{X}/_{X_j} \geq s\}$$

να ελαχιστοποιείται η ποσότητα:

$$\sum_{i:X_i \varepsilon R_1(j,s)} (y_j - \widehat{y_{R_1}})^2 + \sum_{i:X_i \varepsilon R_2(j,s)} (y_j - \widehat{y_{R_2}})^2$$

Η διαδικασία αυτή σταματά να επαναλαμβάνεται όταν κάθε τερματικός κόμβος έχει λιγότερες παρατηρήσεις από έναν προκαθορισμένο μέγιστο αριθμό. Στην συνέχεια, επειδή είναι πιθανό να οδηγηθούμε σε υπερπροσαρμογή (overfitting) και άρα σε μη ικανοποιητική απόδοση στα δεδομένα ελέγχου λόγω δημιουργίας μεγάλων δένδρων, εφαρμόζεται τεχνική κλαδέματος του δένδρου (cost complexity pruning), έστω Το το μεγάλο δέντρο, με στόχο την κατασκευή μικρότερων δέντρων Τ (subtrees). Αυτή η τεχνική βασίζεται στην τιμή μιας παραμέτρου α (tuning) που αντιστοιχεί ένα μικρότερο δέντρο $T \subset T_0$ τέτοιο ώστε να ελαχιστοποιείται η ποσότητα:

$$RSS = \sum_{K=1}^{|T|} \sum_{i \varepsilon R_K} (y_j - \widehat{y_{R_K}})^2 + \alpha |T|$$

όπου με |Τ| συμβολίζουμε τον αριθμό των τερματικών κόμβων. Η βέλτιστη τιμή του α υπολογίζεται μέσω της K-fold cross validation, δηλαδή προσδιορίζεται η τιμή που ελαχιστοποιεί το μέσο σφάλμα των Κ υπό-ομάδων. Έτσι οδηγούμαστε στο τελικό μικρότερο δέντρο που αναζητούμε.

Τα **δέντρα ταξινόμησης** ακολουθούν την ίδια λογική με τα δέντρα παλινδρόμησης, παρουσιάζοντας όμως κάποιες διαφορές. Αναλυτικότερα, επιθυμούμε την πρόβλεψη μιας ποιοτικής αντί ποσοτικής μμεταβλητής, κριτήριο διαχωρισμού αποτελεί η συχνότητα των παρατηρήσεων για την ένταξη της σε έναν κλάδο και αντί για το κριτήριο RSS που δεν μπορεί να εφαρμοστεί εδώ χρησιμοποιείται το classification error rate ή το Gini index ή το Cross-entropy που δίνονται από τους παρακάτω τύπους:

- Classification error rate: $E = 1 - \max(\widehat{Pmk})|$

όπου $\widehat{Pmk}$ το ποσοστό των παρατηρήσεων εκπαίδευσης στην m περιοχή που προέρχονται από την k κατηγορία.

- Gini index: $G = \sum_{k=1}^{K} \widehat{Pmk}(1 - \widehat{Pmk})$

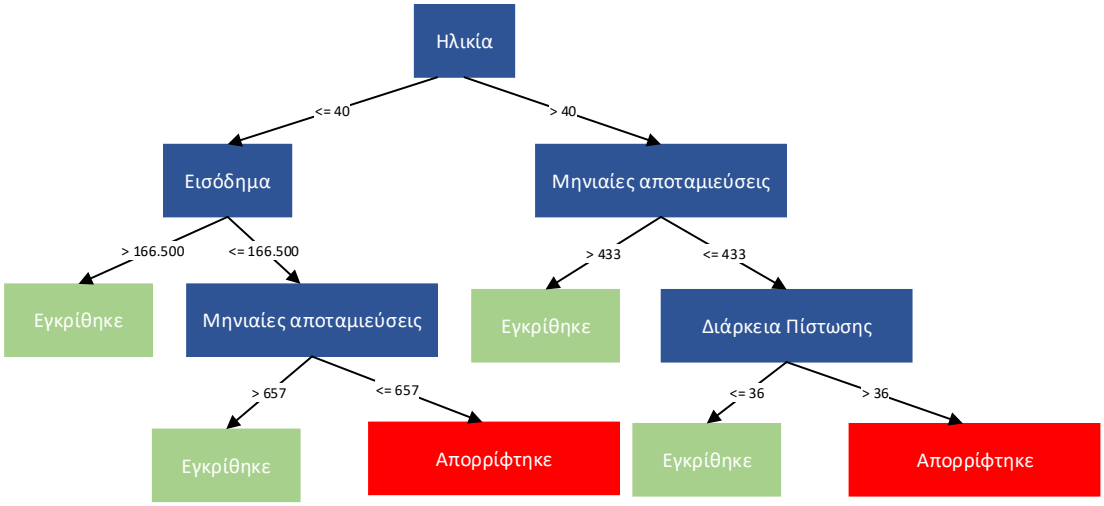
με το G να παίρνει μικρές τιμές όταν ο κόμβος αποτελείται κυρίως από παρατηρήσεις οι οποίες προέρχονται από την ίδια κατηγορία.

- Cross-entropy: $D = -\sum_{k=1}^{K} \widehat{Pmk}(log\widehat{Pmk})$

και δίνει παρόμοια αριθμητικά αποτελέσματα με το Gini Index.

Τα δύο τελευταία κριτήρια δίνουν καλύτερα αποτελέσματα καθώς το classification error δεν είναι αρκετά ευαίσθητο στο μέγεθος του δέντρου. Επίσης ο δείκτης Gini αναφέρεται συχνά ως μέτρο της "καθαρότητας" των κόμβων.

Οι Galindo και Tamayo (2010) εφάρμοσαν CART σε ένα χαρτοφυλάκιο ενυπόθηκων δανείων για την εκτίμηση της αθέτησης υποχρεώσεων και στην σύγκριση με άλλα μοντέλα που χρησιμοποίησαν (Νευρωνικά Δίκτυα, του πλησιέστερου k γείτονα και probit μοντέλα) κατέληξαν ότι τα CART μοντέλα παρέχουν την καλύτερη εκτίμηση. Ένα δέντρο ταξινόμησης πιστοληπτικής ικανότητας θα μπορούσε να είναι το παρακάτω.



Γράφημα 10: Δέντρο ταξινόμησης για την έγκριση δανείου.

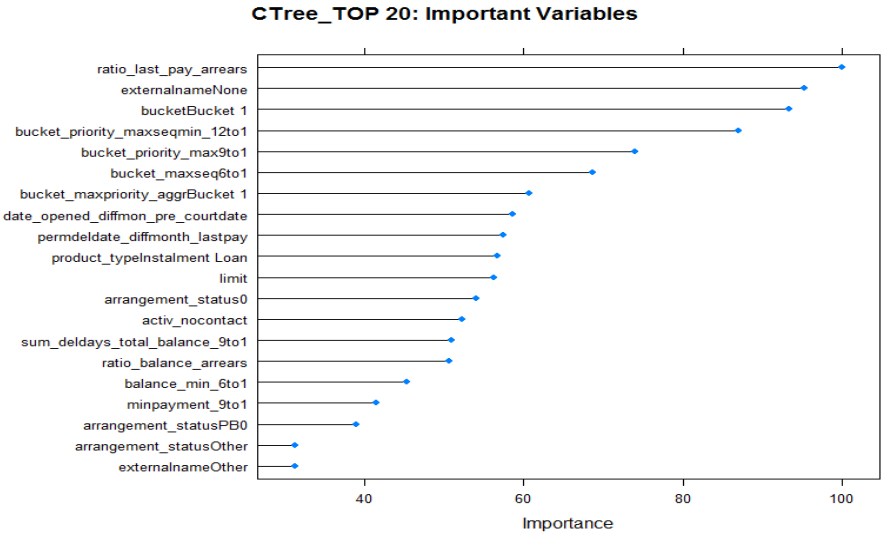
Τα πλεονεκτήματα των δένδρων απόφασης είναι στην εύκολη ερμηνεία και απεικόνιση τους καθώς και στον χειρισμό όλων των ειδών μεταβλητών, ποιοτικών και ποσοτικών. Στον αντίποδα αυτών των πλεονεκτημάτων βρίσκεται η πιο αδύναμη προγνωστική τους ικανότητα σε σχέση με άλλα μοντέλα καθώς και το γεγονός ότι μια αλλαγή στα δεδομένα μπορεί να αλλάξει την σύσταση όλου του δένδρου. Βελτίωση της προγνωστικής ικανότητας των CART αποτελούν οι προσεγγίσεις bagging, random forests και boosting.

Το bagging ή bootstrap aggregation όπως αλλιώς λέγεται έχει ως βάση την μείωση της διασποράς. Η μείωση αυτή επιτυγχάνεται μέσω της διαδικασίας bootstrap παίρνοντας επαναληπτικά δείγματα από τα δεδομένα εκπαίδευσης, στην συνέχεια κατασκευάζουμε ένα μοντέλο πρόβλεψης για κάθε επαναληπτικό δείγμα $\widehat{F^{*1}(x)}$, $\widehat{F^{*2}(x)}\ldots$ $\widehat{F^{*k}(x)}$ και τέλος υπολογίσουμε το μέσο όρο όλων των προβλέψεων

$$\widehat{Fbag(x)} = \frac{1}{K}\sum_{k=1}^{K}\widehat{F^{*k}(x)}$$

Στην μέθοδο αυτή η εκτίμηση του λάθους στα δεδομένα ελέγχου (test error) μπορεί να γίνει και χωρίς την διαδικασία της επικύρωσης (cross-validation), αλλά με την μέθοδο out-of-bag (OOB), δηλαδή η πρόβλεψη της εξαρτημένης μεταβλητής για κάθε παρατήρηση γίνεται χρησιμοποιώντας τα δέντρα για τα οποία η συγκεκριμένη παρατήρηση δεν ανήκε. Η τελική OOB πρόβλεψη για την κάθε παρατήρηση προκύπτει ως ο μέσος όρος των

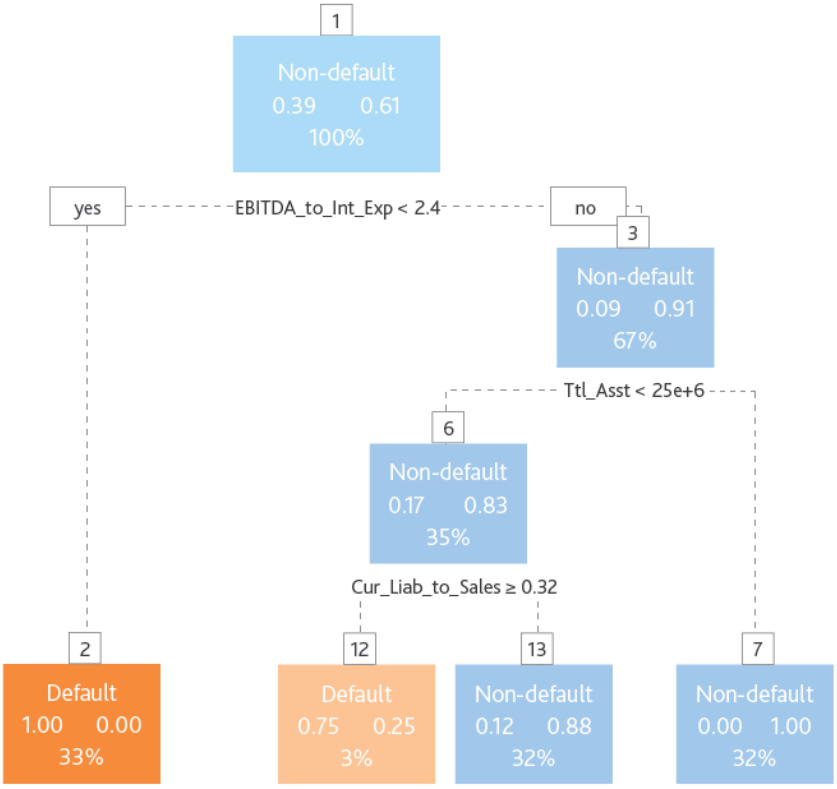
προβλέψεων στην περίπτωση παλινδρόμησης και στην περίπτωση ταξινόμησης αντίστοιχα ως η πιο συχνά εμφανιζόμενη κατηγορία (majority vote). Εκτός από το OOB σφάλμα μπορεί να χρησιμοποιηθεί και το RSS για παλινδρόμηση και το Gini για ταξινόμηση αντίστοιχα ως μέτρα, ειδικά όταν είναι δύσκολο να βρεθεί ποια μεταβλητή είναι πιο σημαντική για την πρόβλεψη της εξαρτημένης μεταβλητής.



Γράφημα 11: Εκτίμηση σημαντικότερης μεταβλητής με χρήση της μεθόδου CART (Angeliki Chatzimoschou, (2018))

Η προσέγγιση Random Forest ακολουθεί την ίδια βάση, δηλαδή bootstrap δείγματα. Ωστόσο, τα random forests λαμβάνουν υπόψη κάθε φορά που πραγματοποιείται ένας διαχωρισμός από όλο το σύνολο π των μεταβλητών ένα υποσύνολο των επεξηγηματικών μεταβλητών, έστω μ, ως υποψήφιες για διαχωρισμό. Οι διαχωρισμοί αυτοί δεν λαμβάνουν όλοι υπόψη την πολύ σημαντική επεξηγηματική μεταβλητή, κατά μέσο όρο (π − μ)/π των διαχωρισμών δεν λαμβάνει, δίνοντας την ευκαιρία σε άλλες επεξηγηματικές μεταβλητές να χρησιμοποιηθούν (ασυσχέτιστες) που αυτό οδηγεί σε μεγαλύτερη μείωση της διακύμανσης και συνεπώς πιο αξιόπιστες προβλέψεις.

Ο Breiman περιγράφει όλη την προσέγγιση των Random Forest (2001). Εξίσου σημαντική και συχνή είναι η χρήση τους στον πιστωτικό κίνδυνο. Πολλές μελέτες εστιάζουν στην υπεροχή των Random Forests έναντι των υπόλοιπων μεθόδων για την πρόβλεψη της αθέτησης σε ένα χαρτοφυλάκιο. Ένα παράδειγμα εφαρμογής τους παραθέτουν οι Moody's Analytics όπου το δέντρο καθορίζει την πιθανότητα αθέτησης με βάση τρεις μεταβλητές και απεικονίζεται παρακάτω:



Γράφημα 12: Random Forest διάγραμμα σε δεδομένα δανείων με στόχο την εκτίμηση αθέτησης (Bacham, *Zhao* (2017)).

Μια επιπλέον προσέγγιση είναι η boosting. Κινείται στην ίδια γραμμή με bootstrap δείγματα όπως και οι παραπάνω προσεγγίσεις με την διαφορά ότι κάθε δέντρο που δημιουργείται χρησιμοποιεί πληροφορία από τα προηγούμενα δέντρα. Η διαδικασία αυτή «μαθαίνει» με αργό τρόπο να εκτιμά την εξαρτημένη μεταβλητή εφαρμόζοντας το δέντρο στα κατάλοιπα και βελτιώνοντας με τον τρόπο αυτό αργά την εκτίμηση του δέντρου σε περιοχές που δεν εφαρμόζει καλά η πρόβλεψη. Στο τέλος, όπως περιγράψαμε παραπάνω υπολογίζουμε το σύνολο των προβλέψεων και όχι τον μέσο όρο και επιπλέον ο τύπος τώρα περιλαμβάνει μια παράμετρο λ που ονομάζεται shrinkage παράμετρος, και ελέγχει το ρυθμό με τον οποίο «μαθαίνει» η προσέγγιση boosting. Ο τύπος δηλαδή μετατρέπεται σε:

$$\widehat{F(x)} = \sum_{k=1}^{K} \lambda \widehat{F^{*k}}(x)$$

Εξαιτίας της διαδικασίας αυτής υπερτερεί στην αποφυγή της υπερπροσαρμογής. Η προσέγγιση boosting μπορεί να οδηγήσει σε overfitting αν ο αριθμός των δέντρων είναι πολύ μεγάλος με αυτό να συμβαίνει αργά όμως, αν τελικά συμβεί. Για την επιλογή του αριθμού των δένδρων χρησιμοποιείται η μέθοδος επικύρωσης (cross-validation).

Η προσέγγιση boosting χρησιμοποιείται συχνά στην εκτίμηση της πιθανότητας αθέτησης ενός δανειολήπτη στον πιστωτικό κίνδυνο. Ο αλγόριθμος ξεκινά δίνοντας σε όλους

τους αιτούντες πίστωση το ίδιο βάρος w (μηδενικό). Μετά την κατασκευή ενός ταξινομητή, το βάρος κάθε αιτούντος αλλάζει σύμφωνα με την ταξινόμηση που δίνεται από τον εν λόγω ταξινομητή. Στη συνέχεια, κατασκευάζεται ένας δεύτερος ταξινομητής χρησιμοποιώντας το επαναστάθμιστο δείγμα εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται συνήθως αρκετές εκατοντάδες φορές. Η τελική κατάταξη του αιτούντος πίστωση ή το «score» είναι ένας σταθμισμένος μέσος όρος των επιμέρους ταξινομήσεων με τους ταξινομητές με μεγάλες θετικές τιμές να δηλώνουν τους καλούς πιστωτές και το αντίστροφο τους κακούς.

3.4.2 Μέθοδος Bayes

Το θεώρημα Bayes είναι μία από τις πιο θεμελιώδεις αρχές στη θεωρία πιθανοτήτων και στη Στατιστική. Η Μπευζιανη προσέγγιση μας επιτρέπει να αλλάξουμε την πρόβλεψή μας με βάση την εκ των προτέρων πιθανότητα και το εκ των υστέρων αποτέλεσμα και δίνεται από τον εξής τύπο:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B/A)}{P(B)}$$

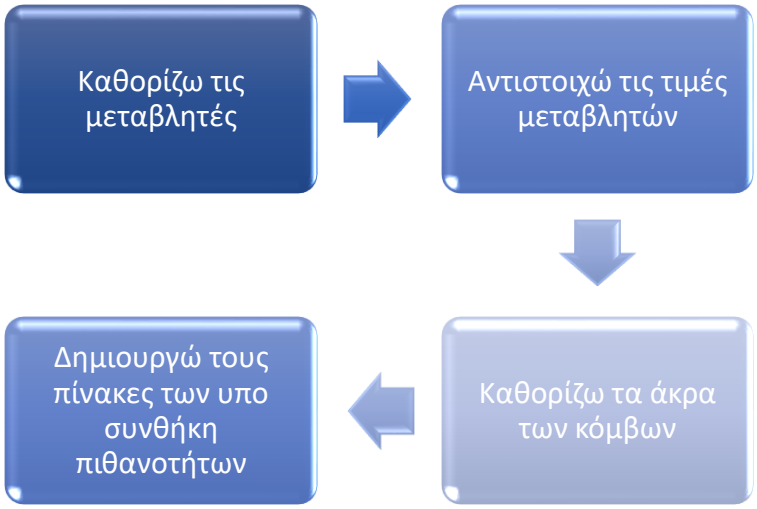
με Α, Β δύο ενδεχόμενα του δειγματοχώρου Ω και $P(B) > 0$.

Υπάρχουν δύο μοντέλα που βασίζονται στην θεωρία αυτή, το Naïve Bayes (NB) και τα μπευζιανά δίκτυα. Αρχικά το NB είναι ένας από τους απλούστερους αλγόριθμους πρόβλεψης αλλά απαιτεί την γνώση ή εκτίμηση των εκ των προτέρων πιθανοτήτων για κάθε ένα συμβάν που μελετάται. Προϋποθέτει να είναι ανεξάρτητες οι μεταβλητές μεταξύ τους. Η απόδοση αυτής της τεχνικής είναι καλή σε εποπτευόμενα μοντέλα και όταν τα δεδομένα είναι συνεχή και διακριτά. Αποδίδει ακόμη και με μικρό δείγμα εκπαίδευσης αν έχουμε καλή τιμή για την εκ των προτέρων πιθανότητα. Ωστόσο όταν έχουμε ελλιπή δεδομένα, είτε οι παρατηρήσεις αυτές απομακρύνονται από τα δεδομένα, είτε αντικαθίστανται οι τιμές από την εκ των προτέρων κατανομή είτε εξομαλύνονται με την χρήση της εκτιμήτριας Laplace αν δεν επηρεάζει τα δεδομένα σε μεγάλο βαθμό.

Στην ανάλυση επιβίωσης συναντάται συχνά η μέθοδος NB σε ιατρικά δεδομένα. Ωστόσο, η ανεξαρτησία μεταξύ όλων των χαρακτηριστικών, η οποία μπορεί να μην ισχύει για πολλά προβλήματα στην ανάλυση επιβίωσης αποτελεί τροχοπέδη.

Τα Μπευζιανά δίκτυα είναι ένας άκυκλος γράφος. Κάθε μεταβλητή αναπαρίσταται με έναν κόμβο και κάθε κόμβος διαθέτει ένα σύνολο από πιθανές τιμές που αντιστοιχούν σε κάθε μεταβλητή και έναν πίνακα υπό συνθήκη πιθανοτήτων (conditional probability table). Οι κόμβοι συνδέονται μεταξύ τους με κατευθυνόμενα βέλη τα οποία δείχνουν την αλληλεξάρτηση των μεταβλητών καθώς και την κατεύθυνση της επιρροής. Τα αρχικά βήματα των μπευζιανών δικτύων παρουσιάζω παρακάτω ως ένα δίκτυο:

:



Τα μπευζιανά δίκτυα μπορούν να αντιπροσωπεύουν οπτικά όλες τις σχέσεις μεταξύ των μεταβλητών, καθιστώντας εύκολα ερμηνεύσιμη την ερμηνεία για τον χρήστη. Συχνά γίνεται χρήση των δικτύων σε συνδυασμό με άλλες τεχνικές με σκοπό την ευκολότερη κατανόηση. Οι Fard et al (2016) συνδύασαν μπευζιανά δίκτυα με τα AFT μοντέλα χρησιμοποιώντας την εκ των προτέρων πιθανότητα εμφάνισης ενός συμβάντος για την πρόβλεψη της ανάλυσης επιβίωσης σε μελλοντικά χρονικά σημεία.

3.4.3 Μηχανές διανυσμάτων υποστήριξης (SVM)

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines -SVM) είναι μια κατηγορία μεθόδων ταξινόμησης. Για την περιγραφή της διαδικασίας αρχικά θα δοθεί ο ορισμός του υπερεπιπέδου καθώς αποτελεί βασικό στοιχείο της μεθόδου. Σε έναν p-διάστατο χώρο, ένα υπερεπίπεδο είναι ένας υποχώρος διάστασης p−1, δηλαδή για παράδειγμα στις δύο διαστάσεις ένα υπερεπίπεδο είναι μια γραμμή. Ένα υπερεπίπεδο περιγράφεται από την παρακάτω εξίσωση:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_P X_p = 0$$

με $X = X_{i,i=1,,p}$ να αποτελούν τις παρατηρήσεις εκπαίδευσης του p-διάστατο χώρου. Για την κατασκευή του ταξινομητή εργαζόμαστε ως εξής: $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_P X_p > 0$ τότε το X βρίσκεται από τη μία πλευρά του υπερεπιπέδου, ενώ $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_P X_p < 0$ τότε το X βρίσκεται από την άλλη πλευρά του υπερεπιπέδου. Συνεπώς ένα υπερεπίπεδο χωρίζει τον p-διάστατο χώρο σε δύο μισά. που αντιστοιχούν σε δύο ομάδες $y_1, \ldots, y_n \in \{-1, 1\}$ όπου με -1 συμβολίζεται η μία ομάδα και με 1 η άλλη. Στόχος είναι η σωστή ταξινόμηση

της παρατήρησης ελέγχου βάσει του ταξινομητή που κατασκευάζουμε για τα δεδομένα εκπαίδευσης.

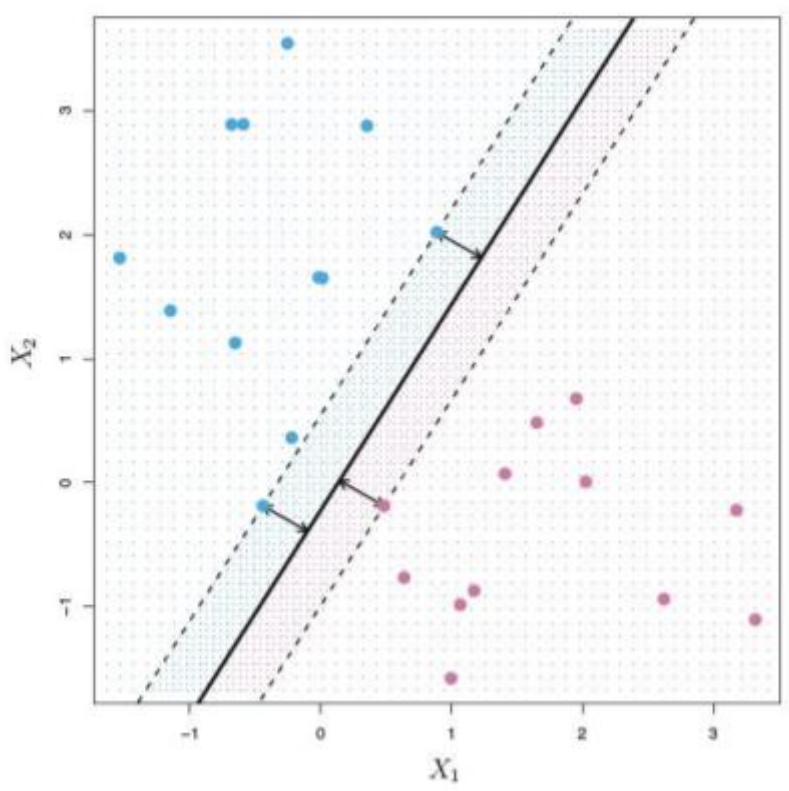
Ισοδύναμα:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_P X_p > 0 \quad \text{αν } y_i = 1$$

και

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_P X_p < 0 \quad \text{αν } y_i = -1$$

Συνεπώς η ταξινόμηση καθορίζεται από το πρόσημο της παραπάνω σχέσης. ΄Οσο πιο μακρυά από το 0 είναι τόσο πιο σίγουροι είμαστε για τη σωστή ταξινόμηση της. Με βάση αυτό οδηγούμαστε στον υπολογισμό της απόστασης ανάμεσα σε κάθε παρατήρηση εκπαίδευσης και το υπερεπίπεδο για την κατασκευή του καλύτερου ταξινομητή. Η μικρότερη τέτοια απόσταση είναι η ελάχιστη απόσταση των παρατηρήσεων από το υπερεπίπεδο, γνωστή ως περιθώριο (margin).



Γράφημα 13: Κανόνας ταξινόμησης

Στο παραπάνω γράφημα φαίνονται δύο ομάδες παρατηρήσεων που απεικονίζονται με μπλε και μωβ χρώμα αντίστοιχα. Το περιθώριο είναι η απόσταση ανάμεσα στη μαύρη γραμμή και τις διακεκομμένες γραμμές. Τα δύο μπλε σημεία και το μωβ σημείο που βρίσκονται πάνω στις διακεκομμένες γραμμές ονομάζονται support vectors και βάσει αυτών καθορίζεται η απόσταση μεταξύ των δύο γραμμών.

Η δημιουργία ένας κανόνα ταξινόμησης βάσει ενός υπερεπιπέδου μπορεί να ταξινομεί τέλεια τις παρατηρήσεις αλλά ο ταξινομητής που προκύπτει μπορεί να έχει μεγάλη

ευαισθησία σε ορισμένες. Μεγάλη ευαισθησία στις επιμέρους παρατηρήσεις αποτελεί ένδειξη overfitting των παρατηρήσεων εκπαίδευσης. Συνεπώς, στην προσέγγιση του SVM ο ταξινομητής επιτρέπει κάποιες παρατηρήσεις να βρίσκονται στη λάθος πλευρά του περιθωρίου ή και στη λάθος πλευρά του υπερεπιπέδου ώστε να έχει και καλύτερη ταξινόμηση για τις περισσότερες από αυτές. Η παραπάνω διαδικασία αποτελεί λύση ενός προβλήματος βελτιστοποίησης και περιγράφεται από τους εξής μαθηματικούς τύπους:

$$maximize_{\beta i,ei}\, M$$

$$y_i(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_P X_p) \geq M(1 - e_i)\ \mu\varepsilon\ y_i\ i = 1, \ldots . n$$

Με $e_i \geq 0\ \kappa\alpha\iota\ \sum_{i=1}^{n} e_i \leq C\ \kappa\alpha\iota\ \sum_{J=1}^{p} {\beta_j}^2 = 1$

Η ποσότητα M είναι το εύρος του περιθωρίου, C μια μη αρνητική tuning παράμετρος και τα e_i μεταβλητές που δηλώνουν που βρίσκεται η i-παρατήρηση σε σχέση με το υπερεπίπεδο και σε σχέση με το περιθώριο. Οι πιθανές τιμές του e_i είναι:

- $e_i = 0$, η i-παρατήρηση βρίσκεται στη σωστή πλευρά του περιθωρίου.
- $0 < e_i < 1$, η i-παρατήρηση βρίσκεται στη λάθος πλευρά του περιθωρίου.
- $e_i > 1$, η i-παρατήρηση βρίσκεται στη λάθος πλευρά του υπερεπιπέδου.

Αντίστοιχα, η μεταβλητή C αποτελεί το όριο του αθροίσματος των ei και καθορίζει τον αριθμό και τη σοβαρότητα των παραβιάσεων και επιλέγεται η τιμή μέσω διαδικασίας επικύρωσης (cross-validation). Όσο μεγαλύτερη είναι η τιμή του C τόσο μεγαλύτερη η ανοχή μας σε λάθος ταξινομήσεις. Είναι ύψιστης σημασίας η επιλογή αυτή για τον ταξινομητή καθώς καθορίζεται κυρίως από τις λάθος τιμές ή από αυτές που βρίσκονται πάνω στο περιθώριο, δηλαδή στην διακεκομμένη γραμμή στην παραπάνω εικόνα.

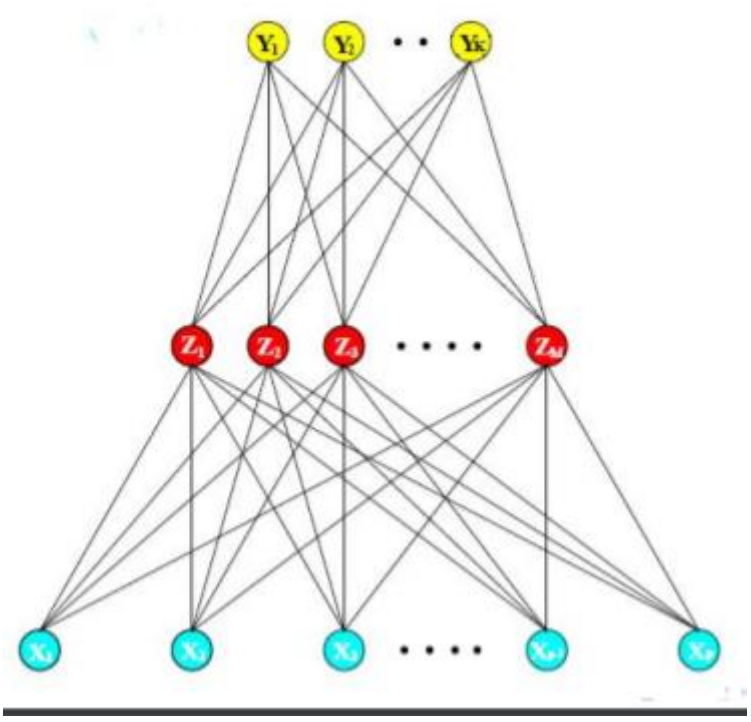
Ο SVM έχει μικρή ευαισθησία σε παρατηρήσεις που βρίσκονται μακριά από το υπερεπίπεδο. Αντίστοιχη είναι η συμπεριφορά της λογιστικής παλινδρόμησης και παρουσιάζουν αρκετά κοινά. Ο SVM λειτουργεί τόσο με γραμμικές σχέσεις μεταξύ επεξηγηματικών μεταβλητών και εξαρτημένης μεταβλητής όσο και μη γραμμικές χρησιμοποιώντας συναρτήσεις των επεξηγηματικών μεταβλητών, όπως τετραγωνικούς ή κυβικούς όρους, πολυωνυμικούς όρους των επεξηγηματικών μεταβλητών ή άλλες συναρτήσεις (όχι πολυώνυμά) των επεξηγηματικών μεταβλητών. Αυτό συμβαίνει με τη δημιουργία πυρήνων (kernel) συναρτήσεων στους παραπάνω τύπους, δηλαδή συναρτήσεις που ποσοτικοποιούν την ομοιότητα δύο παρατηρήσεων. Γνωστή είναι η εφαρμογή των radial kernel που έχουν την ιδιότητα να επηρεάζονται μόνο οι παρατηρήσεις εκπαίδευσης που είναι γειτονικές στην παρατήρηση ελέγχου επιδρούν στην ταξινόμησή της.

Η προσέγγιση SVM έχει χρησιμοποιηθεί σε προβλήματα ανάλυσης επιβίωσης και πιστωτικού κινδύνου με στόχο την καλύτερη επεξηγηματική ισχύ και προβλεπτική

ικανότητα. Η μέθοδος αυτή ξεκίνησε να χρησιμοποιείται στην βαθμολόγηση του πιστωτικού κινδύνου το 2004 από τον Huang, ασχολήθηκε με την ταξινόμηση των ομολόγων με δεδομένα από Ταιβάν και ΗΠΑ και στην συνέχεια καθιερώθηκε από πολλούς άλλους όπως οι Chen & Shih (2006) σε δεδομένα από την Ταιβάν και τον Harris (2013) ο οποίος σύγκρινε τα μοντέλα πιστοληπτικής ικανότητας που έχουν κατασκευαστεί με χρήση διάρκειας Broad (λιγότερο από 90 ημέρες ληξιπρόθεσμων δανείων) και Narrow (άνω των 90 ημερών ληξιπρόθεσμων δανείων) με SVM.

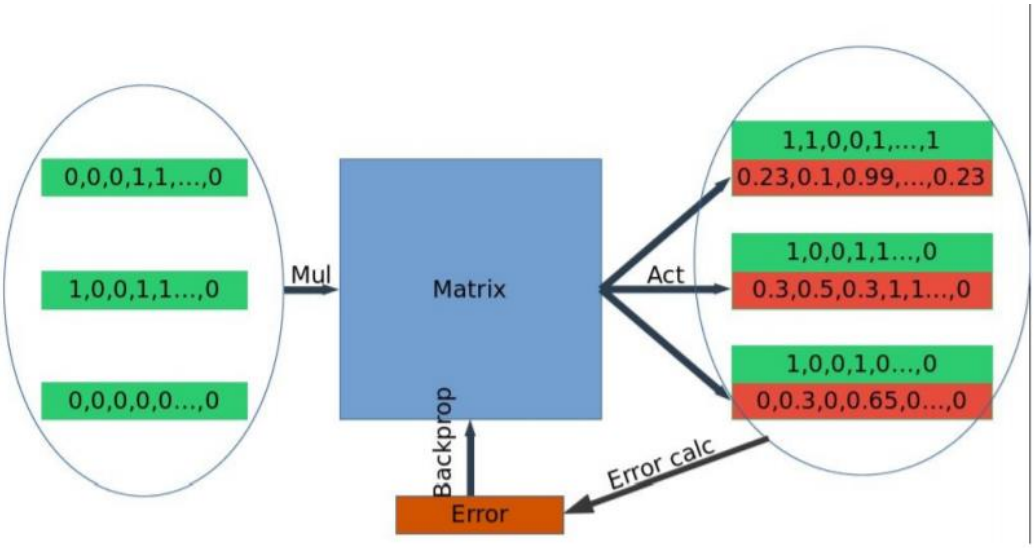
3.4.4 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα αποτελούν μια πολύ ισχυρή μέθοδο με κεντρική ιδέα την δημιουργία γραμμικών συνδυασμών των επεξηγηματικών μεταβλητών ως παράγωγα χαρακτηριστικά και μοντελοποίηση της εξαρτημένης μεταβλητής ως μια μη-γραμμική συνάρτηση αυτών των χαρακτηριστικών. Θα αναφερθούμε στα πιο γνωστά νευρωνικά δίκτυα "vanilla" ή single hidden layer back-propagation network ή single layer perceptron. Εφαρμόζονται τόσο για σκοπούς παλινδρόμησης κατασκευάζοντας μοντέλα πρόβλεψης όσο και ταξινόμησης κατασκευάζοντας μοντέλα κατηγοριοποίησης. Η κατασκευή του δικτύου γίνεται με την κατάλληλη αρχιτεκτονική όπως το παρακάτω γράφημα :



Γράφημα 14: Νευρωνικό Δίκτυο (Hastie, Tibshirani, & Friedman, J. (2009))

Επιλέγεται ένα σύνολο δεδομένων και εκπαιδεύεται το δίκτυο. Επεξεργάζεται τα στοιχεία ένα κάθε φορά και μαθαίνει συγκρίνοντας τη κατηγοριοποίηση τους για μια εγγραφή με τη γνωστή κύρια κατηγοριοποίηση της εγγραφής. Στην συνέχεια επανατροφοδοτεί το δίκτυο με τα σφάλματα από την αρχική κατηγοριοποίηση της πρώτης εγγραφής και η διαδικασία αυτή επαναλαμβάνεται κάθε φορά που ο αλγόριθμος τρέχει. Ο αλγόριθμος οπισθοδιάδοσης (Back propagation) σύμφωνα με τον LeCun (1998) ονομάζεται η διαδικασία όπου το σφάλμα που προκύπτει από τις εκτιμήσεις γυρνά πίσω στα βάρη του δικτύου ώστε να ενημερωθούν. Η παραπάνω διαδικασία απεικονίζεται στην εξής εικόνα:



Γράφημα 15: Εκπαίδευση Νευρωνικού δικτύου (Μαλινάκης, (2016))

Συγκεκριμένα, όπως φαίνεται στο διάγραμμα υπάρχουν στην είσοδο του διαγράμματος Κ ομάδες με καθεμία να μοντελοποιεί την πιθανότητα της αντίστοιχης ομάδας. Στόχος είναι οι μετρήσεις $Y_{i,\mathrm{i=1,..K}}$ να ταξινομηθούν στις αντίστοιχες ομάδες με 1 αν η παρατήρηση ανήκει στην αντίστοιχη ομάδα και 0 διαφορετικά. Τα στοιχεία $Z_{j,\mathrm{j=1,..M}}$ τα οποία ονομάζονται κρυφές μονάδες καθώς δεν φαίνονται άμεσα και μπορεί να υπάρχουν περισσότερα από ένα επίπεδα-στρώματα, προκύπτουν ως γραμμικοί μετασχηματισμοί των επεξηγηματικών μεταβλητών $X_{r,\mathrm{r=1,..p}}$.Το κάθε στοιχείο Y_i μοντελοποιείται ως συνάρτηση γραμμικών συνδυασμών των Z_i όπως φαίνεται παρακάτω:

$$Z_j = \sigma\left(\beta_{0_j} + \beta_j^T X\right) \mu\varepsilon\ j = 1,..M$$

$$T_k = \alpha_{0_k} + \alpha_k^T Z\ \mu\varepsilon\ k=1,...K$$

$$f_\kappa(X) = g_\kappa(T)\ \mu\varepsilon\ \kappa = 1,...K$$

Στην πρώτη σχέση, το σ αποτελεί συνήθως μια σιγμοειδής συνάρτηση $\sigma(t) = \frac{1}{1+e^{-t}}$ (αν όπου σ πάρουμε την ταυτοτική συμπεραίνουμε ότι προκύπτει γραμμικό μοντέλο). Το $g_\kappa(T)$ αποτελεί έναν τελικό μετασχηματισμό του διανύσματος Τ όπου σε

προβλήματα παλινδρόμησης είναι η ταυτοτική συνάρτηση, δηλαδή $g_\kappa(T) = T_K$ ενώ σε προβλήματα ταξινόμησης $g_\kappa(T) = \frac{e^{T_\kappa}}{\sum_{\kappa=1}^{K} e^{T_\kappa}}$ που ονομάζεται SoftMax συνάρτηση, αποτελεί μια γενίκευση της σιγμοειδούς και χρησιμοποιείται για την μετατροπή τιμών σε πιθανότητες όταν υπάρχουν περισσότερες από δύο ομάδες. Τα $\alpha_{0_k}, \alpha_k, με\ k = 1, .. K\ και\ \beta_{0_j}, \beta_j$ *j=1,…M* αποτελούν άγνωστες παραμέτρους, ονομάζονται βάρη και σκοπός είναι η εύρεση τιμών για τα βάρη που οδηγούν σε καλή εφαρμογή του Νευρωνικού δικτύου στα δεδομένα εκπαίδευσης. Για τα βάρη με τιμές κοντά στο μηδέν, το νευρωνικό δίκτυο αποτελεί ένα προσεγγιστικά γραμμικό μοντέλο.

Η ύπαρξη πολλών βαρών μπορεί να οδηγήσει σε overfitting του μοντέλου. Για την αποφυγή αυτού μία από τις προτεινόμενες λύσεις είναι ο κανόνας πρώιμης διακοπής, δηλαδή το μοντέλο 'εκπαιδεύεται' για λίγο και σταματάει πριν ο αλγόριθμος συγκλίνει σε ένα συνολικό ελάχιστο με αποτέλεσμα τη συρρίκνωση του τελικού μοντέλου. Για τον καθορισμό του χρόνου διακοπής απαιτείται ένα σύνολο δεδομένων επικύρωσης.

Για την μέτρηση της καλής εφαρμογής του μοντέλου χρησιμοποιούνται διάφορες συναρτήσεις ανάλογα με το είδος προβλημάτων. Σε προβλήματα παλινδρόμησης χρησιμοποιείται το άθροισμα τετραγώνων των σφαλμάτων που δίνεται από τον παρακάτω τύπο:

$$R(\theta) = \sum_{\kappa=1}^{K} \sum_{i=1}^{Z} (y_{ik} - f_k(x_i))^2$$

Σε προβλήματα ταξινόμησης ο αντίστοιχος τύπος που χρησιμοποιείται είναι το τετραγωνικό σφάλμα είτε η Cross-entropy (deviance) που είναι ο εξής:

$$R(\theta) = -\sum_{i=1}^{Z} \sum_{\kappa=1}^{K} y_{ik} log\ (f_k(x_i))$$

Στο τελικό αποτέλεσμα σημαντικό ρόλο παίζει η τυποποίηση των μεταβλητών εισόδου, ώστε να έχουν μέση τιμή ίση με το 0 και τυπική απόκλιση ίση με 1. Η τυποποίηση εξασφαλίζει ότι όλες οι μεταβλητές εισόδου αντιμετωπίζονται ισοδύναμα κατά τη διαδικασία της συστηματοποίησης (regularization). Ένα εύρος τιμών για τα τυχαία αρχικά βάρη είναι συνήθως το [-0.7,0.7]. Όσο αφορά τις κρυφές μονάδες, όσο μεγαλύτερος είναι ο αριθμός τόσο καλύτερα για το δίκτυο. Μεγαλύτερος αριθμός μεταβλητών εισόδου οδηγεί σε περισσότερα στρώματα εντός του δικτύου. Συνήθης αριθμός στρωμάτων είναι από 5 εως 100 και επιλέγεται βάσει πειραματισμού και παρελθοντικής γνώσης.

Τα νευρωνικά δίκτυα χρησιμοποιούνται ευρέως στην ανάλυση επιβίωσης. Έχουν χρησιμοποιηθεί τόσο για την πρόβλεψη του χρόνου επιβίωσης ενός ατόμου απευθείας από τις μεταβλητές που εισάγονται όσο και για την συνάρτηση κινδύνου ως επέκταση του

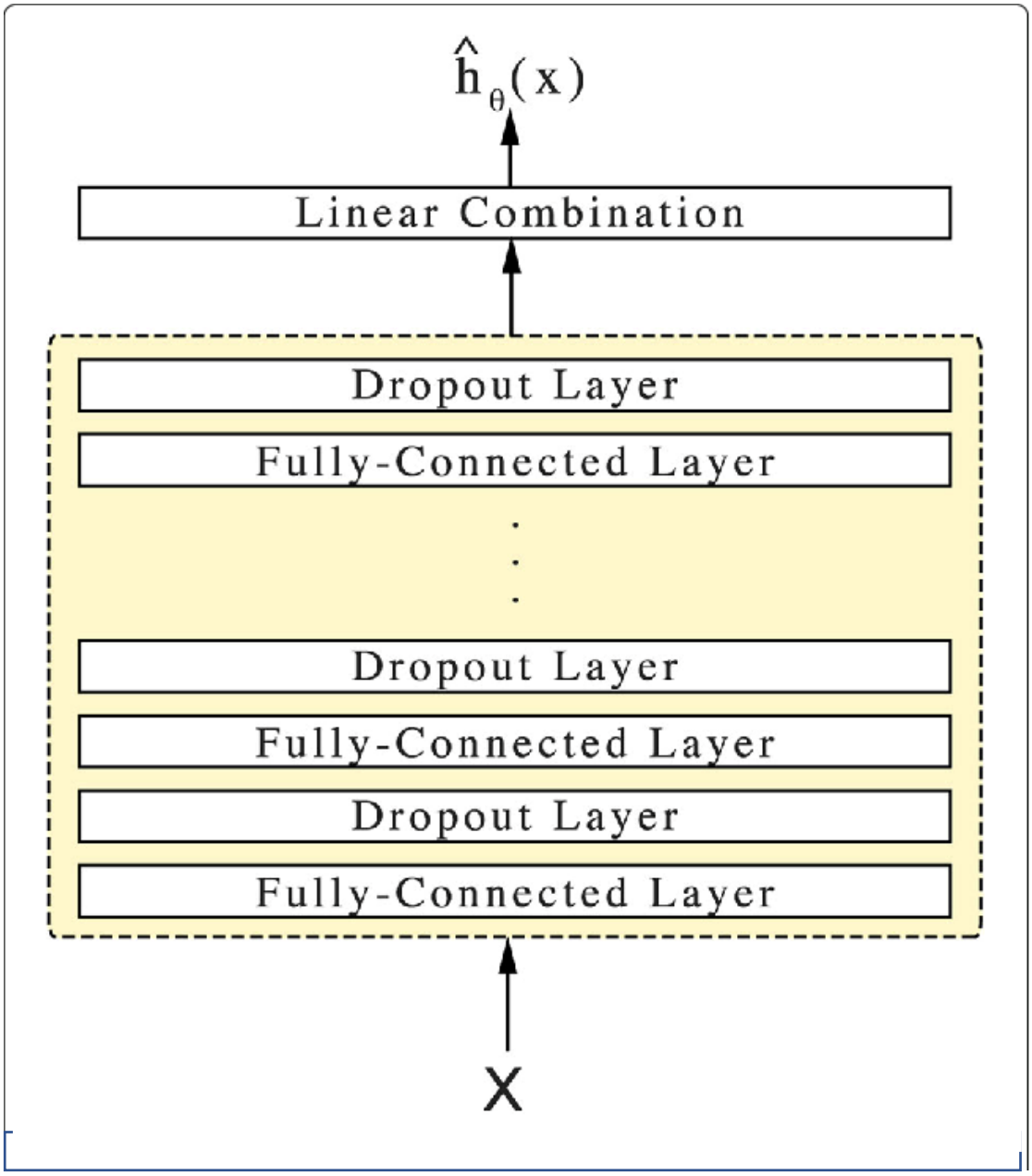
μοντέλου αναλογικού κινδύνου του Cox με πρώτη χρήση το 1995 από τους Faraggi και Simon. Έχουν μια εξαιρετική ικανότητα να χειρίζονται τα λογοκριμένα δεδομένα στην ανάλυση επιβίωσης. Επιπρόσθετα, τα αναδρομικά νευρωνικά δίκτυα έχουν συνδυαστεί με επιτυχία στην ανάλυση επιβίωσης για τη μελέτη επαναλαμβανόμενων γεγονότων, ιδιαίτερα στο πλαίσιο των εφαρμογών μοντελοποίησης συμπεριφοράς των χρηστών, όπως στην βαθμολόγηση συμπεριφοράς δανειοληπτών στον πιστωτικό κίνδυνο.

3.4.5 Δίκτυα Βαθιάς Μάθησης -DeepSurv- DeepHit

Το DeepSurv αποτελεί ένα δίκτυο βαθιάς μάθησης που έχει ως βάση τον υπολογισμό της συνάρτησης κινδύνου και αποτελεί εξέλιξη του μοντέλου αναλογικού κινδύνου του Cox. Η ανάπτυξη των νευρωνικών δικτύων σε συνδυασμό με το μοντέλο του Cox τις μεγάλες βάσεις δεδομένων με μη γραμμικές σχέσεις μεταξύ των συμμεταβλητών οδήγησαν στην δημιουργία του αλγορίθμου αυτού. Η δομή του είναι όπως περιγράψαμε παραπάνω στα νευρωνικά δίκτυα με στόχο τον υπολογισμό της επίδρασης του κάθε ατόμου στην συνάρτηση κινδύνου του σε σχέση με τα αντίστοιχα βάρη του δικτύου και τα βήματα που ακολουθούμε είναι τα εξής:

- Τα δεδομένα εισόδου X αποτελούνται από τις συμμεταβλητές που έχουν παρατηρηθεί
- Ο κορμός αποτελείται από τα κρυφά στρώματα, πλήρως συνδεδεμένα στρώματα κόμβων, ακολουθούμενα από ένα επίπεδο εγκατάλειψης (dropout) με στόχο την αποφυγή της υπερπροσαρμογής
- Το επίπεδο εξόδου έχει μόνο έναν κόμβο με μια γραμμική λειτουργία ενεργοποίησης που δίνει την έξοδο $\widehat{h_\theta(x)}$ (Λογαριθμικές εκτιμήσεις κινδύνου)

Η παραπάνω διαδικασία απεικονίζεται στην παρακάτω εικόνα.



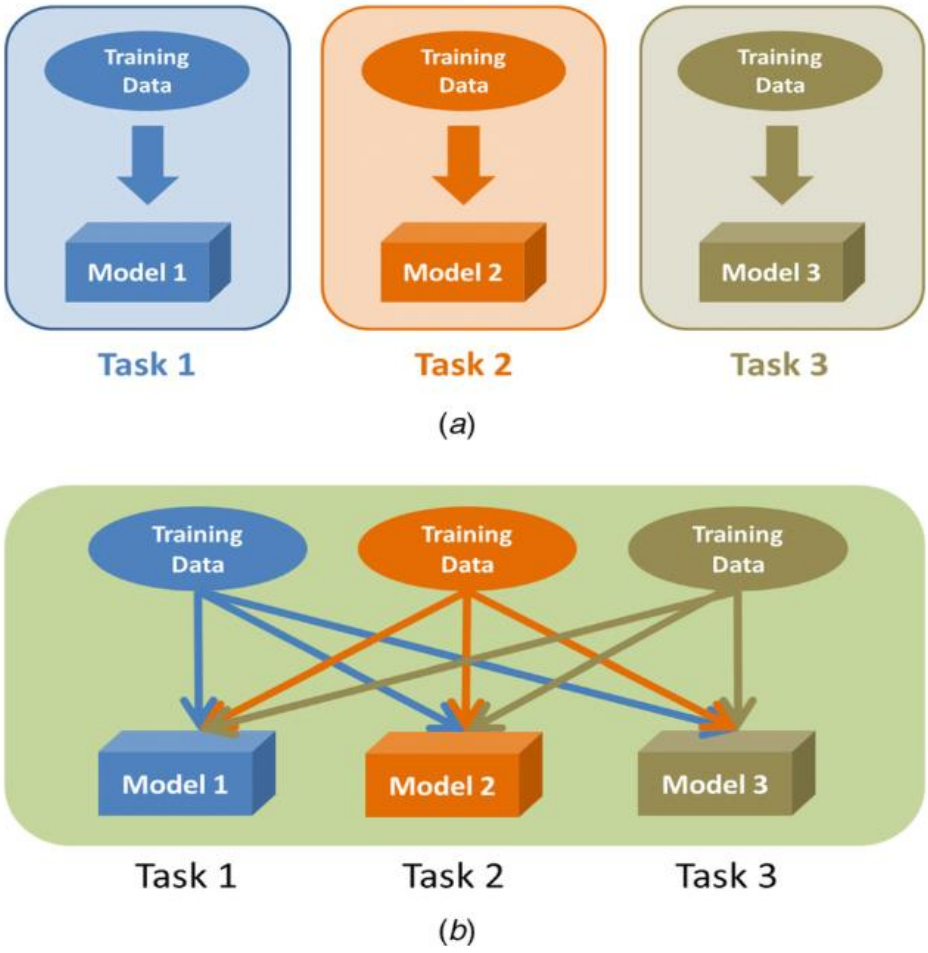
Γράφημα 16: Διάγραμμα DeepSurv (Katzman, Shaham, Cloninger, Bates, Jiang, Kluger (2018))

Το DeepHit αποτελεί μια εξελιγμένη μέθοδο μηχανικής μάθησης που προτάθηκε από τους Lee et. al. (2018) και χρησιμοποιεί ένα βαθύ νευρωνικό δίκτυο για την «εκμάθηση» της κατανομής των χρόνων επιβίωσης. Δεν κάνει υποθέσεις σχετικά με την υποκείμενη στοχαστική διαδικασία μεταξύ των συμμεταβλητών και των χρόνων επιβίωσης σε σύγκριση με τα προηγούμενα μοντέλα αλλά επιτρέπει την πιθανότητα η σχέση μεταξύ των συμμεταβλητών και των κινδύνων να αλλάζει με την πάροδο του χρόνου.

Ένα από τα σημαντικότερα πλεονεκτήματα της προσέγγισης αυτής είναι η δυνατότητα χρήσης του δικτύου για πρόβλεψη ενός κινδύνου όπως αυτόν την αθέτησης σε ένα χρηματοπιστωτικό ίδρυμα αλλά επίσης και περισσότερων «ανταγωνιστικών» από έναν κινδύνους, όπως για παράδειγμα στο ίδιο περιβάλλον θα μπορούσε να θεωρηθεί η πρώιμη αποπληρωμή (payoff). Ανταγωνιστικός κίνδυνος είναι ο κίνδυνος του οποίου η εμφάνιση

αποκλείει την ύπαρξη του πρωτογενούς γεγονότος. Στην ως τώρα μελέτη, οι πελάτες με πρόωρη αποπληρωμή χαρακτηρίζονται λογοκριμένοι, επειδή το μόνο γεγονός ενδιαφέροντος ήταν το γεγονός της αθέτησης. Εάν προσθέσουμε δεύτερο κίνδυνο ως επίκεντρο προσοχής από την άποψη της ανάλυσης επιβίωσης, θα προσθέσουμε αντίστοιχα μια επιπλέον στήλη ενδιαφέροντος με ετικέτα αποπληρωμή (αθέτηση =1, αποπληρωμή = 2). Το DeepHit έχει την δυνατότητα να κάνει πρόβλεψη και για τους δύο κινδύνους.

Το δίκτυο αυτό ανήκει στην κατηγορία των μοντέλων πολλαπλής μάθησης. Αρχικά κάθε μοντέλο εκπαιδεύεται ώστε να μαθαίνει από ένα μόνο σύνολο δεδομένων όπως φαίνεται στο πρώτο μισό της παρακάτω εικόνας. Αν όμως τα δεδομένα αυτά σχετίζονται μεταξύ τους, μπορεί να κατασκευαστεί ένα μοντέλο εκμάθησης πολλαπλών εργασιών με στόχο τη βελτίωση της εκμάθησης ενός μοντέλου με τη χρήση των γνώσεων που επιτυγχάνονται σε όλη τη διάρκεια της εκμάθησης των εργασιών παράλληλα. Κατά την διαδικασία αυτή στόχος είναι η βελτίωση της απόδοσης όλων των παραμέτρων και δεν υπάρχει ιεραρχία ως προς τους κινδύνους που εκτιμώνται.



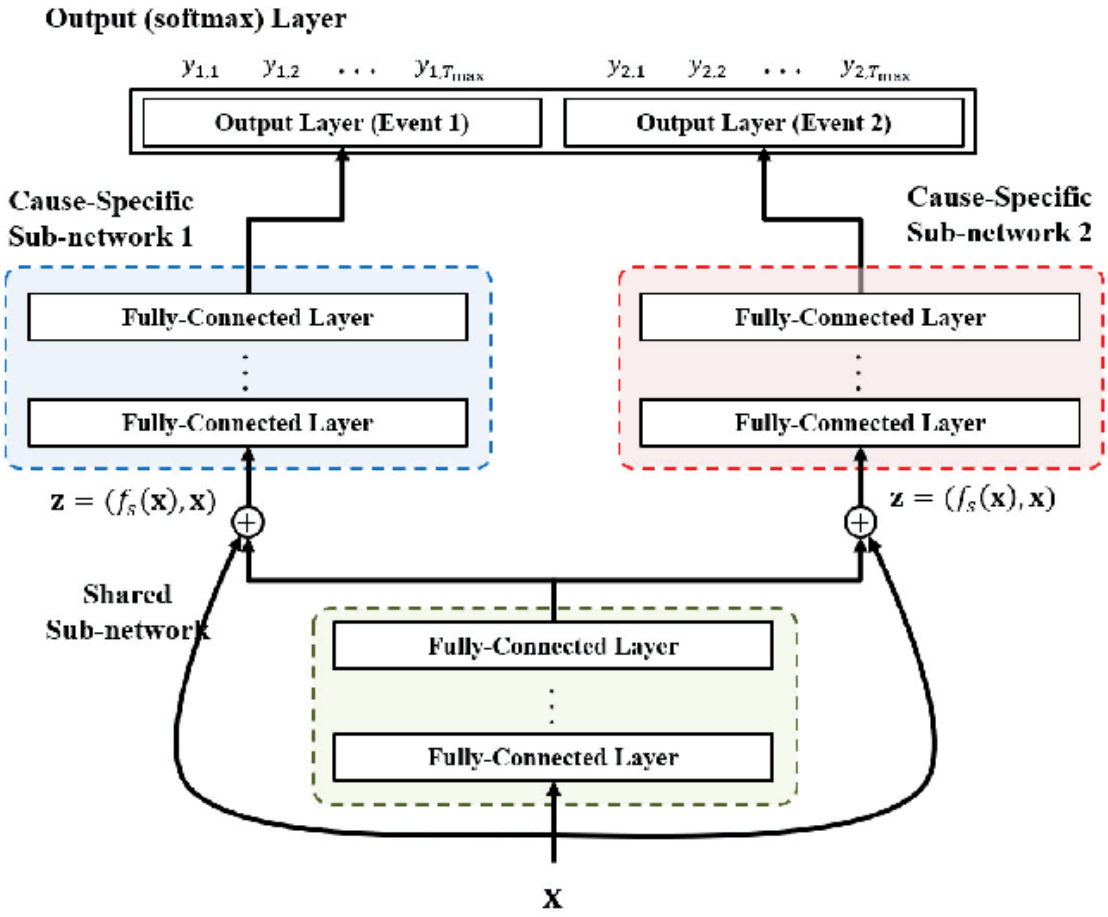
Γράφημα 17: Διαφορά μεταξύ της μάθησης μίας εργασίας και της εκμάθησης πολλαπλών εργασιών (Shao (2016))

Υπάρχουν δύο προσεγγίσεις στην εκμάθηση πολλαπλών εργασιών. Στην πρώτη προσέγγιση (soft parameter sharing) κάθε εργασία έχει το δικό της μοντέλο με τις δικές της

παραμέτρους ενώ στην δεύτερη (hard parameter sharing) το μοντέλο αποτελείται από δύο μέρη, ένα «κοινόχρηστο» υπό-δίκτυο, όπου εκεί το μοντέλο μαθαίνει τις κοινές λειτουργίες-εργασίες μεταξύ των διαφορετικών γεγονότων ενδιαφέροντος και μια οικογένεια, έστω K, υπό-δικτύων που αφορούν συγκεκριμένες εργασίες, μη κοινές. Αυτή η προσέγγιση παρουσιάζεται εδώ. Η διαδικασία που ακολουθεί είναι παρόμοια με τα νευρωνικά δίκτυα που αναφέραμε παραπάνω. Το μοντέλο εκπαιδεύεται μέσω της οπισθοδιάδοσης ανάμεσα στα κρυφά στρώματα και βάρη. Ωστόσο, υπάρχουν και διαφορές. Η πρώτη είναι ότι υπάρχει μια έξτρα σύνδεση μεταξύ των αρχικών συμμεταβλητών και της εισόδου των υπό-δικτύων για συγκεκριμένη εργασία. Το τελευταίο συνεπάγεται ότι τα αποτελέσματα από τα υπό-δίκτυα αυτά δεν προέρχονται μόνο από τα ίδια αλλά και από τις αρχικές συμμεταβλητές και με τον τρόπο αυτό το μοντέλο μαθαίνει καλύτερα. Η δεύτερη διαφορά συναντάται στο αποτέλεσμα, είναι ένα διάνυσμα Y ενιαίο για τα στρώματα εξόδου δηλαδή:

$$y = (y_{1.1}, y_{1.2} \cdots y_{1.T_{max}}, y_{2.1}, y_{2.2} \cdots y_{2.T_{max}})$$

όπου τα $y_{1,1}$ εκφράζουν την πιθανότητα η συμμεταβλητή x να βιώσει το συμβάν k για κάθε χρονική σήμανση t εντός του χρόνου παρατήρησης. Οι πιθανότητες κάθε κινδύνου αθροίζουν στο 1. Η διαδικασία αυτή του μοντέλου απεικονίζεται στην Εικόνα 15.



Εικόνα 15: Μοντέλο DeepHit με δύο ανταγωνιστικούς κινδύνους (Lee, Zame, Yoon, Schaar (2018))

Τα πλεονεκτήματα επιλογής μοντέλων εκμάθησης πολλαπλών εργασιών είναι πολλά. Έχοντας την δυνατότητα να χρησιμοποιήσει δεδομένα από τα υπόλοιπα στρώματα αυξάνεται το μέγεθος του δείγματος που χρησιμοποιείται για την εκπαίδευση του μοντέλου, και συνεπώς βελτιώνεται η απόδοση του. Επιπλέον, μαθαίνοντας ταυτόχρονα πολλές διαφορετικές εργασίες μπορεί το μοντέλο να αποκτήσει και μια γενικότερη αναπαράσταση όλων των εργασιών. Η προηγμένη επέκταση των μοντέλων ανάλυσης επιβίωσης που χρησιμοποιούν πρακτικές μηχανικής μάθησης δίνει περισσότερη μεθοδολογική ελευθερία. Με τη σωστή διαδικασία συντονισμού των παραμέτρων είναι δυνατό να επιτευχθούν ακριβέστερες προβλέψεις της μεταβλητής χρόνου-γεγονότος.

4. Εφαρμογή στο Χαρτοφυλάκιο

4.1 Πηγή και παρουσίαση δεδομένων

Τα δεδομένα που θα χρησιμοποιηθούν στην συνέχεια προέρχονται από μια εταιρεία δανεισμού στην Αμερική γνωστή ως «Lending Club», η οποία δημοσιεύει πληροφορίες σχετικά με όλα τα εκ ’δοθέντα δάνεια. Τα δεδομένα είναι διαθέσιμα στον ιστότοπο «Kaggle» όπου υπάρχει μια βάση δεδομένων, εργαλείων, πηγών και αλγορίθμων. Η ανάλυση των δεδομένων θα γίνει με χρήση της γλώσσας προγραμματισμού Python και του προγράμματος Jupiter.

Το χαρτοφυλάκιο αποτελείται από ένα σύνολο καταναλωτικών δανείων που εκδόθηκαν μεταξύ Ιανουαρίου 2007 και Δεκεμβρίου 2014. Τα δεδομένα αφορούν περισσότερα από 400.000 δάνεια αιτούντων που αθέτησαν και αιτούντων που εκπλήρωσαν τις υποχρεώσεις τους. Για το χρονικό διάστημα από το 2007 ως το 2014 έχουν 466285 εγγραφές τις οποίες θα χρησιμοποιήσουμε για την ανάπτυξη αλγορίθμων ανάλυσης επιβίωσης και μηχανικής εκμάθησης δυαδικής ταξινόμησης (καλός-κακός πελάτης). Η βάση αποτελείται από 75 μεταβλητές-χαρακτηριστικά τόσο των αιτούντων όσο και των αντίστοιχων δανείων, δηλαδή συνοψίζει ένα ετερογενές σύνολο χαρακτηριστικών σε σχέση με τα χαρακτηριστικά του δανείου (ημέρες παραβατικότητας, κατηγορία δανείου, συνολικό ποσό που έχει υποσχεθεί που καταβλήθηκε εντός του μήνα κ.λπ.) και ορισμένα προσωπικά χαρακτηριστικά του δανειολήπτη (επάγγελμα, υπόλοιπο λογαριασμού, μέρος διαμονής κ.λπ.). Παρακάτω δίνεται ο πίνακας με τις μεταβλητές και την αντίστοιχη περιγραφή.

LoanStatNew	Περιγραφή
addr_state	Η χώρα διαμονής που καταχωρείται από τον δανειολήπτη στην αίτηση δανείου
annual_inc	Το ετήσιο εισόδημα που δηλώνεται από τον δανειολήπτη κατά την εγγραφή του.
annual_inc_joint	Το συνδυασμένο ετήσιο εισόδημα που δηλώνεται από τους υπόλοιπους δανειστές (κοινό δάνειο) κατά την εγγραφή
application_type	Υποδεικνύει εάν το δάνειο είναι ατομική αίτηση ή κοινή αίτηση με άλλους
collection_recovery_fee	τέλος είσπραξης μετά την επιβάρυνση
collections_12_mths_ex_med	Αριθμός εισπράξεων σε 12 μήνες εκτός από ιατρικές εισπράξεις
delinq_2yrs	Ο αριθμός των 30+ ημερών ληξιπρόθεσμων περιπτώσεων εγκληματικότητας στο πιστωτικό αρχείο του δανειολήπτη για τα τελευταία 2 χρόνια
desc	Περιγραφή δανείου που παρέχεται από τον δανειολήπτη

dti	Ένας δείκτης που υπολογίζεται με βάση τις συνολικές μηνιαίες πληρωμές χρέους του δανειολήπτη για το σύνολο των δανειακών υποχρεώσεων, εξαιρουμένων των ενυπόθηκων δανείων και του αιτούμενου δανείου LC, διαιρούμενο με το μηνιαίο εισόδημα του δανειολήπτη.
dti_joint	Ένας λόγος που υπολογίζεται με βάση τις συνολικές μηνιαίες πληρωμές των συνειδικευθέντων επί του συνόλου των δανειακών υποχρεώσεων, εξαιρουμένων των ενυπόθηκων δανείων και του αιτούμενου δανείου LC, διαιρούμενος με το συνδυασμένο μηνιαίο εισόδημα των συνειδικευθέντων
earliest_cr_line	Η μέρα που άνοιξε το πρώτο πιστωτικό όριο για τον δανειολήπτη
emp_length	Διάρκεια απασχόλησης σε χρόνια. Οι πιθανές τιμές είναι μεταξύ 0 και 10 όπου 0 σημαίνει λιγότερο από ένα έτος και 10 σημαίνει δέκα ή περισσότερα έτη.
emp_title	Ο τίτλος εργασίας που παρέχεται από τον δανειολήπτη κατά την υποβολή αίτησης για το δάνειο.*
funded_amnt	Το συνολικό ποσό που είχε δεσμευθεί για το δάνειο αυτό εκείνη τη στιγμή.
funded_amnt_inv	Το συνολικό ποσό που δεσμεύτηκε από τους επενδυτές για το δάνειο αυτό εκείνη τη στιγμή.
grade	Βαθμός δανείου που χορηγείται στην LC
home_ownership	Το καθεστώς ιδιοκτησίας κατοικίας που παρέχεται από τον δανειολήπτη κατά τη στιγμή της εγγραφής: RENT, OWN, MORTGAGE, OTHER.
id	Ένα μοναδικό ID αντιστοίχισης LC για την καταχώρηση δανείου.
initial_list_status	Το καθεστώς αρχικής εγγραφής του δανείου. Πιθανές τιμές είναι – W, F
inq_last_6mths	Ο αριθμός των ερευνών κατά τους τελευταίους 6 μήνες (εκτός από τις έρευνες για αυτοκίνητα και ενυπόθηκα δάνεια)
installment	Η μηνιαία πληρωμή που οφείλει ο δανειολήπτης
int_rate	Επιτόκιο του δανείου
is_inc_v	Υποδεικνύει εάν το εισόδημα επαληθεύτηκε από την LC, δεν επαληθεύτηκε ή αν η πηγή εισοδήματος επαληθεύτηκε
issue_d	Ο μήνας που χρηματοδοτήθηκε το δάνειο
last_credit_pull_d	Ο πιο πρόσφατος μήνας LC τράβηξε πίστωση για αυτό το δάνειο
last_pymnt_amnt	Τελευταίο συνολικό ποσό πληρωμής που εισπράχθηκε.
last_pymnt_d	Τον τελευταίο μήνα που ελήφθη η πληρωμή
loan_amnt	Το απαριθμούμενο ποσό του δανείου που ζήτησε ο δανειολήπτης. Εάν κάποια στιγμή, το πιστωτικό τμήμα μειώνει το ποσό του δανείου, τότε θα αντικατοπτρίζεται σε αυτή την αξία.
loan_status	Τρέχουσα κατάσταση του δανείου
member_id	Ένα μοναδικό LC που έχει εκχωρηθεί id για το μέλος δανειολήπτη.
mths_since_last_delinq	Ο αριθμός των μηνών από την τελευταία παραβατικότητα του δανειολήπτη.

mths_since_last_major_derog	Μήνες από την πιο πρόσφατη βαθμολογία των τελευταίων 90 ημερών ή χειρότερης βαθμολόγισης
mths_since_last_record	Ο αριθμός των μηνών από την τελευταία δημόσια εγγραφή.
next_pymnt_d	Επόμενη προγραμματισμένη ημερομηνία πληρωμής
open_acc	Ο αριθμός των ανοικτών πιστωτικών γραμμών στο πιστωτικό αρχείο του δανειολήπτη.
policy_code	δημόσια διαθέσιμα προιόντα στο κοινό policy_code=1 νέα προϊόντα που δεν είναι διαθέσιμα στο κοινό policy_code=2
pub_rec	Αριθμός υποβαθμισμένων δημόσιων αρχείων
purpose	Μια κατηγορία που παρέχεται από τον δανειολήπτη για την αίτηση δανείου.
pymnt_plan	Υποδεικνύει εάν έχει τεθεί σε εφαρμογή σχέδιο πληρωμής για το δάνειο
recoveries	Ακαθάριστη ανάκτηση μετά τη χρέωση
revol_bal	Συνολικό πιστωτικό ανακυκλούμενο υπόλοιπο
revol_util	Ανακυκλούμενο ποσοστό χρησιμοποίησης γραμμών, ή το ποσό πίστωσης που ο δανειολήπτης χρησιμοποιεί σε σχέση με όλες τις διαθέσιμες ανακυκλούμενες πιστώσεις.
sub_grade	Δεύτερη βαθμολόγιση LC
term	Ο αριθμός των πληρωμών για το δάνειο. Οι τιμές είναι σε μήνες και μπορεί να είναι είτε 36 ή 60.
title	The loan title provided by the borrower
total_acc	Ο συνολικός αριθμός των πιστωτικών γραμμών που βρίσκονται επί του παρόντος στο πιστωτικό αρχείο του δανειολήπτη
total_pymnt	Πληρωμές που έχουν ληφθεί μέχρι σήμερα για το συνολικό χρηματοδοτούμενο ποσό
total_pymnt_inv	Πληρωμές που έχουν ληφθεί μέχρι σήμερα για μέρος του συνολικού ποσού που χρηματοδοτείται από επενδυτές
total_rec_int	Τόκοι που έχουν εισπραχθεί μέχρι σήμερα
total_rec_late_fee	Καθυστερημένες αμοιβές που έχουν ληφθεί μέχρι σήμερα
total_rec_prncp	Αρχές που ισχύουν μέχρι την ημερομηνία αυτή.
verified_status_joint	Υποδεικνύει εάν το κοινό εισόδημα των συνεπειδικών επαληθεύτηκε από την LC, δεν επαληθεύτηκε ή εάν η πηγή εισοδήματος επαληθεύτηκε
zip_code	Οι πρώτοι 3 αριθμοί του ταχυδρομικού κώδικα που παρέχονται από τον δανειολήπτη στην αίτηση δανείου.
open_acc_6m	Αριθμός ανοικτών συναλλαγών τους τελευταίους 6 μήνες
open_il_6m	Αριθμός συναλλαγών ανοιχτοί επί του παρόντος
open_il_12m	Αριθμός λογαριασμών δόσεων που άνοιξαν τους τελευταίους 12 μήνες
open_il_24m	Αριθμός λογαριασμών δόσεων που άνοιξαν τους τελευταίους 24 μήνες
mths_since_rcnt_il	Μήνες από το άνοιγμα των πιο πρόσφατων λογαριασμών δόσεων
total_bal_il	Συνολικό τρέχον υπόλοιπο όλων των λογαριασμών δόσεων
il_util	Λόγος του συνολικού τρέχοντος υπολοίπου προς το υψηλό πιστωτικό/πιστωτικό όριο σε όλες τις εγκαταστάσεις
open_rv_12m	Αριθμός ανακυκλούμενων συναλλαγών που άνοιξαν τους τελευταίους 12 μήνες

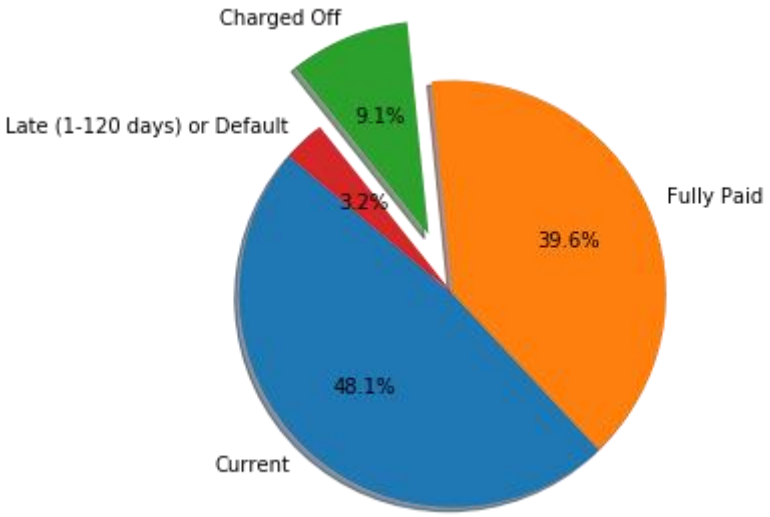
open_rv_24m	Αριθμός ανακυκλούμενων συναλλαγών που άνοιξαν τους τελευταίους 24 μήνες
max_bal_bc	Μέγιστο τρέχον οφειλόμενο υπόλοιπο σε όλους τους ανακυκλούμενους λογαριασμούς
all_util	Υπόλοιπο σε πιστωτικό όριο για όλες τις συναλλαγές
total_rev_hi_lim	Συνολικό ανακυκλούμενο υψηλό πιστωτικό όριο
inq_fi	Αριθμός ερευνών για τα προσωπικά οικονομικά
total_cu_tl	Αριθμός χρηματοοικονομικών συναλλαγών
inq_last_12m	Αριθμός πιστωτικών ερευνών τους τελευταίους 12 μήνες
acc_now_delinq	Ο αριθμός των λογαριασμών στους οποίους ο δανειολήπτης είναι πλέον παραβάτης.
tot_coll_amt	Σύνολο ποσών είσπραξης που οφείλονται
tot_cur_bal	Συνολικό τρέχον υπόλοιπο όλων των λογαριασμών

* Ο τίτλος εργοδότη αντικαθιστά το όνομα εργοδότη για όλα τα δάνεια που αναφέρονται μετά την 23/9/2013

Η βάση για την δημιουργία του αλγορίθμου και την προγνωστική μοντελοποίηση είναι η διατύπωση της εξαρτημένης μεταβλητής, δηλαδή ο ορισμός του καλού και του κακού πελάτη στην περίπτωση της βαθμολόγησης πίστωσης. Συνήθως αυτός ο ορισμός βασίζεται στον αριθμό των ημερών του πελάτη μετά την ημερομηνία λήξης (ημέρες ληξιπρόθεσμες, DPD) και το ποσό που είναι ληξιπρόθεσμο. Πρέπει να θέσουμε κάποιο επίπεδο ανοχής στην περίπτωση του ληξιπρόθεσμου ποσού. Αυτό σημαίνει ότι πρέπει να καθορίσουμε τι θεωρείται χρέος και τι όχι. Ένας κακός πελάτης είναι αυτός που δεν έχει πληρώσει καμία δόση δανείου για 3-μήνες ή 90 ημέρες. Η εξαρτημένη μεταβλητή αποτελεί στο δείγμα μας το loan_status. Οι τιμές της μεταβλητής καθώς και το πλήθος εμφανίσεων της καθεμίας παρουσιάζεται παρακάτω.

Current /Ενεργό	**224226**	**48.1%**
Fully Paid /Πληρωμένο πλήρως	184739	39.6%
Charged Off /Απενεργοποιημένο	42475	9.1%
Late (31-120 days) /Καθυστέρηση (31-120 μέρες)	6900	1.5%
In Grace Period / Περίοδος χάριτος	3146	0.7%
Does not meet the credit policy. Status:Fully Paid / Δεν πληροί την πιστωτική πολιτική κατάσταση: Πληρωμένος πλήρως	1988	0.4%
Late (16-30 days) /Καθυστέρηση (16-30 μέρες)	1218	0.3%
Default / Αθέτηση	832	0.2%
Does not meet the credit policy. Status:Charged Off / Δεν πληροί την πιστωτική πολιτική κατάσταση: Απενεργοποιημένο	761	0.2%

Παρατηρούμε ότι το 48.1% περίπου το μισό όλων των δανείων είναι ενεργά όπως φαίνεται στο γράφημα παρακάτω ενώ αμέσως μετά το 39.6% είναι πλήρως πληρωμένα. Στο μοντέλο μας θα εστιάσουμε μόνο στις 2 κατηγορίες, στα πλήρως πληρωμένα και στα απενεργοποιημένα δάνεια όπου η εταιρεία δεν αναμένει ο οφειλέτης να πληρώσει το χρέος του και με 81.3% ακρίβεια μπορούμε να προβλέψουμε ότι το δάνειο να πληρωθεί.



Γράφημα 18: Κατανομή δανείων.

Παραθέτουμε τις πρώτες 4 σειρές του συνόλου δεδομένων και μια προβολή των αρχικών χαρακτηριστικών, τα οποία διακρίνονται σε δύο τύπους, τις συνεχείς και τις κατηγορικές μεταβλητές .

Unnamed: 0	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	hor
466280	8598660	1440975	18400	18400	18400.0	60 months	14.47	432.64	C	C2	Financial Advisor	4 years	
466281	9684700	11536848	22000	22000	22000.0	60 months	19.97	582.50	D	D5	Chief of Interpretation (Park Ranger)	10+ years	
466282	9584776	11436914	20700	20700	20700.0	60 months	16.99	514.34	D	D1	patrol	7 years	
466283	9604874	11457002	2000	2000	2000.0	36 months	7.90	62.59	A	A4	Server Engineer Lead	3 years	

home_ownership	annual_inc	verification_status	issue_d	loan_status	pymnt_plan	url	desc	purpose	title
MORTGAGE	110000.0	Source Verified	Jan-14	Current	n	https://www.lendingclub.com/browse/loanDetail....	NaN	debt_consolidation	Debt consolidation
MORTGAGE	78000.0	Verified	Jan-14	Charged Off	n	https://www.lendingclub.com/browse/loanDetail....	NaN	debt_consolidation	Debt consolidation
MORTGAGE	46000.0	Verified	Jan-14	Current	n	https://www.lendingclub.com/browse/loanDetail....	Borrower added on 12/06/13 > I am going to c...	debt_consolidation	Debt consolidation
OWN	83000.0	Verified	Jan-14	Fully Paid	n	https://www.lendingclub.com/browse/loanDetail....	NaN	credit_card	Credit card refinancing

zip_code	addr_state	dti	delinq_2yrs	earliest_cr_line	inq_last_6mths	mths_since_last_delinq	mths_since_last_record	open_acc	pub_rec	revol_bal	revol_util	total_acc
773xx	TX	19.85	0.0	Apr-03	2.0	NaN	NaN	18.0	0.0	23208	77.6	36.0
377xx	TN	18.45	0.0	Jun-97	5.0	NaN	116.0	18.0	1.0	18238	46.3	30.0
458xx	OH	25.65	0.0	Dec-01	2.0	65.0	NaN	18.0	0.0	6688	51.1	43.0
913xx	CA	5.39	3.0	Feb-03	1.0	13.0	NaN	21.0	0.0	11404	21.5	27.0

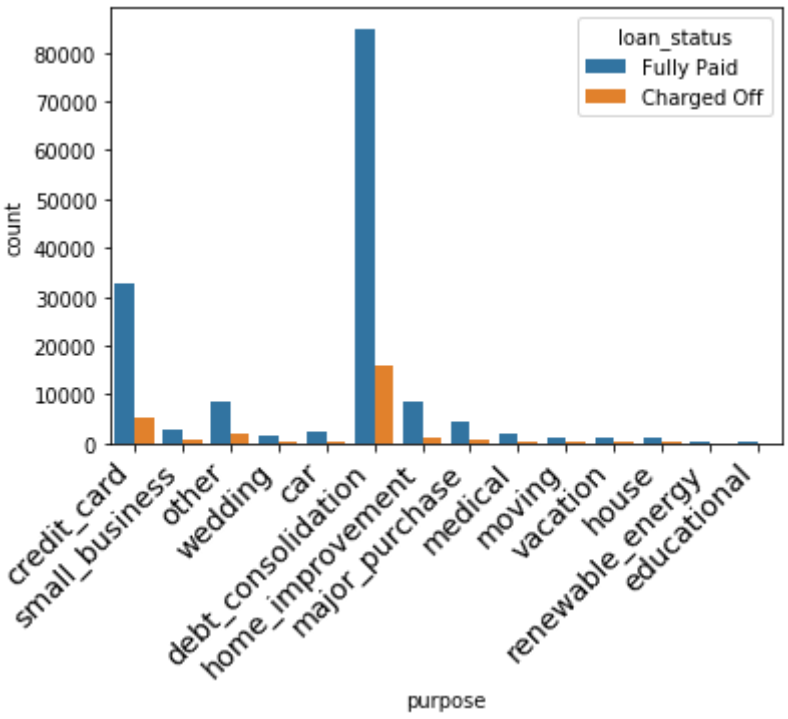
initial_list_status	out_prncp	out_prncp_inv	total_pymnt	total_pymnt_inv	total_rec_prncp	total_rec_int	total_rec_late_fee	recoveries	collection_recovery_fee	last_pymnt_d
w	12574.00	12574.00	10383.360000	10383.36	5826.00	4557.36	0.0	0.0	0.0	Jan-16
f	0.00	0.00	4677.920000	4677.92	1837.04	2840.88	0.0	0.0	0.0	Dec-14
f	14428.31	14428.31	12343.980000	12343.98	6271.69	6072.29	0.0	0.0	0.0	Jan-16
w	0.00	0.00	2126.579838	2126.58	2000.00	126.58	0.0	0.0	0.0	Dec-14

last_pymnt_amnt	next_pymnt_d	last_credit_pull_d	collections_12_mths_ex_med	mths_since_last_major_derog	policy_code	application_type	annual_inc_joint	dti_joint	ver
432.64	Feb-16	Jan-16	0.0	NaN	1	INDIVIDUAL	NaN	NaN	
17.50	NaN	Jan-16	0.0	NaN	1	INDIVIDUAL	NaN	NaN	
514.34	Feb-16	Dec-15	0.0	NaN	1	INDIVIDUAL	NaN	NaN	
1500.68	NaN	Apr-15	0.0	NaN	1	INDIVIDUAL	NaN	NaN	

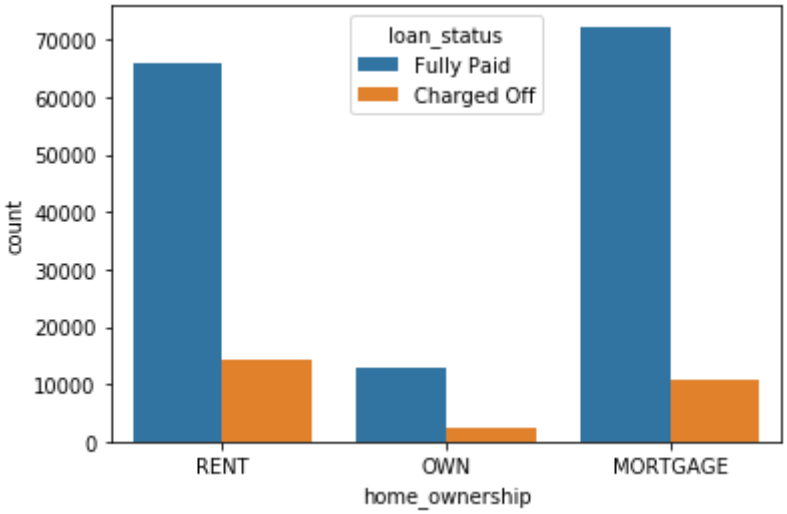
verification_status_joint	acc_now_delinq	tot_coll_amt	tot_cur_bal	open_acc_6m	open_il_6m	open_il_12m	open_il_24m	mths_since_rcnt_il	total_bal_il	il_util	open_rv_12m
NaN	0.0	0.0	294998.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	0.0	0.0	221830.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	0.0	0.0	73598.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	0.0	0.0	591610.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

open_rv_24m	max_bal_bc	all_util	total_rev_hi_lim	inq_fi	total_cu_tl	inq_last_12m
NaN	NaN	NaN	29900.0	NaN	NaN	NaN
NaN	NaN	NaN	39400.0	NaN	NaN	NaN
NaN	NaN	NaN	13100.0	NaN	NaN	NaN
NaN	NaN	NaN	53100.0	NaN	NaN	NaN

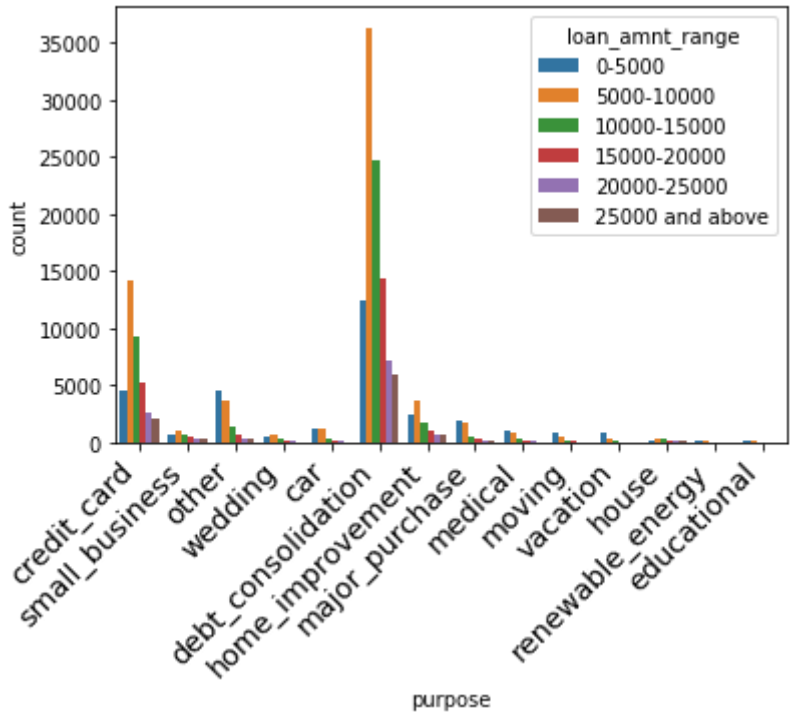
Παρακάτω δίνουμε μια πρώτη εικόνα για την σχέση των μεταβλητών μεταξύ τους αλλά και με την εξαρτημένη μεταβλητή με σκοπό να γίνει πιο κατανοητό το δείγμα δεδομένων.



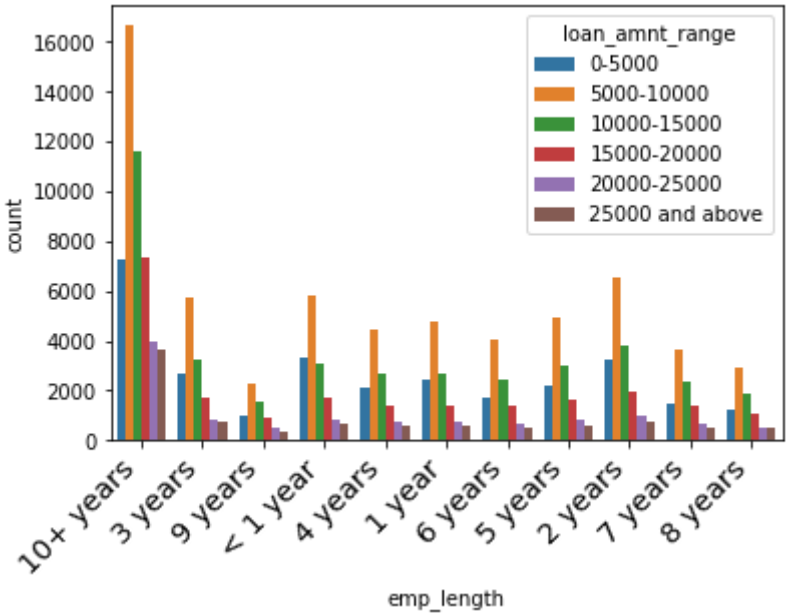
Γράφημα 19: Απεικόνιση της σχέσης της εξαρτημένης μεταβλητής αθέτησης δανείου με τον σκοπό αίτησης δανείου.



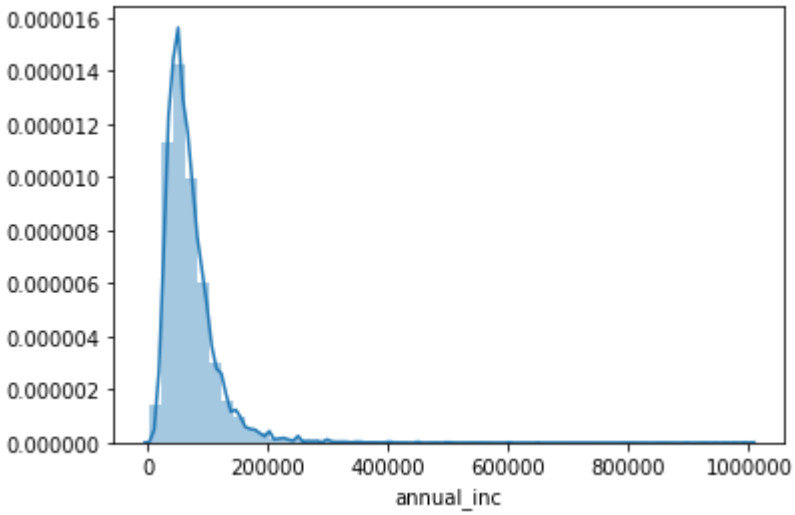
Γράφημα 20: Απεικόνιση της σχέσης της εξαρτημένης μεταβλητής αθέτησης δανείου με την κατάσταση του οφειλέτη ως προς την κατοικία.



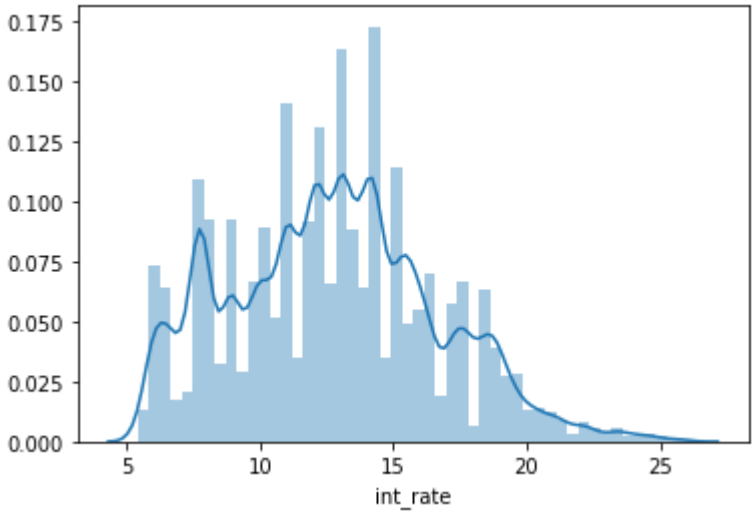
Γράφημα 21: Απεικόνιση της σχέσης του ποσού δανείου με τον σκοπό αίτησης δανείου.



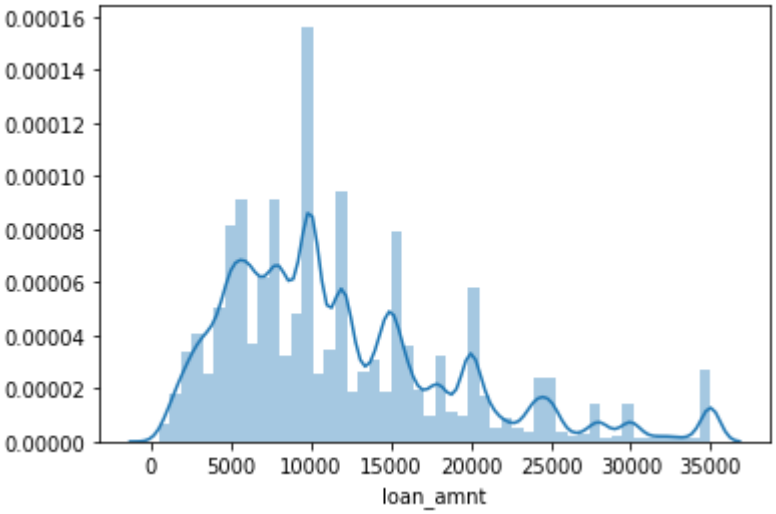
Γράφημα 22: Απεικόνιση της σχέσης των κλάσεων του ποσού δανείου με την εργασιακή διάρκεια του οφειλέτη.



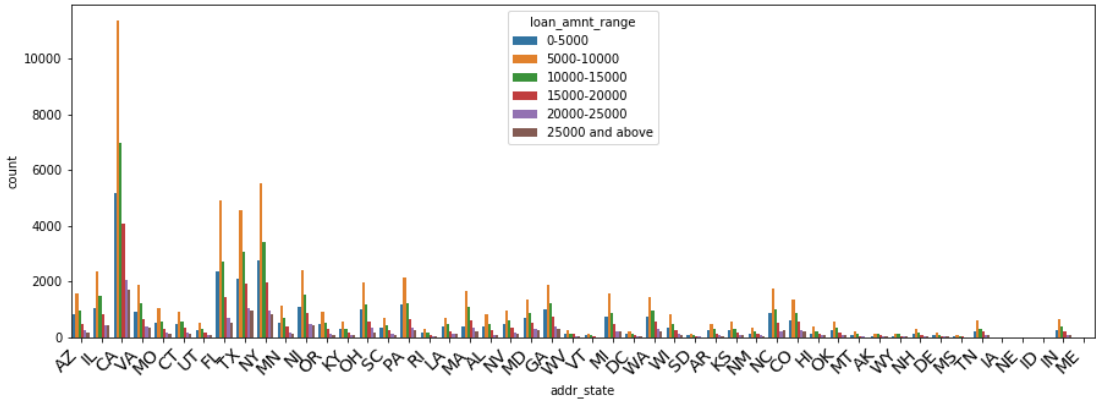
Γράφημα 23: Ιστόγραμμα ετήσιου εισοδήματος του οφειλέτη.



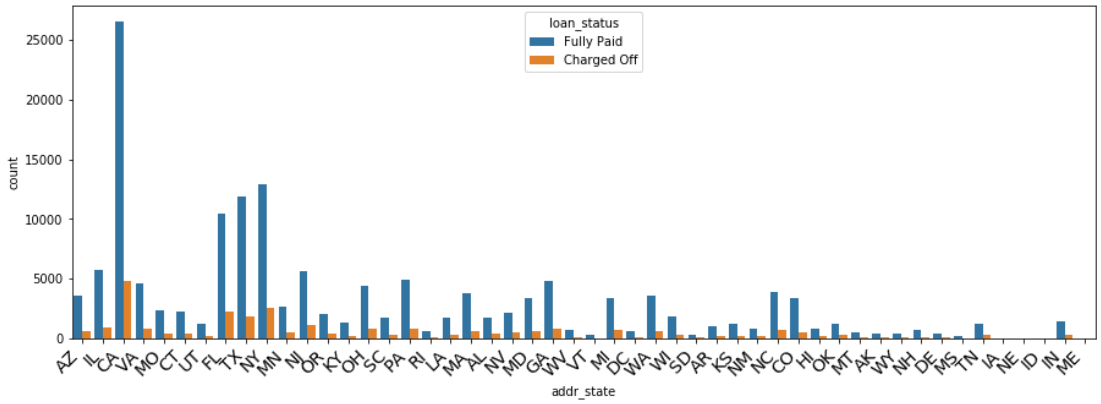
Γράφημα 24: Ιστόγραμμα του επιτοκίου.



Γράφημα 25: Ιστόγραμμα του ποσού δανείου.



Γράφημα 26: Απεικόνιση της σχέσης του ποσού του δανείου με την πολιτεία κατοικίας του οφειλέτη.



Γράφημα 27: Απεικόνιση της σχέσης της εξαρτημένης μεταβλητής αθέτησης δανείου με το μέρος διαμονής του οφειλέτη (ΗΠΑ πολιτείες)

Επιλέγουμε να μελετήσουμε τα δάνεια με διάρκεια 36 μήνες ως μέρος της έρευνας καθώς για τους 60 μήνες δεν έχει ολοκληρωθεί μελέτη ως το 2014 για να μπορούμε να χρησιμοποιήσουμε τα δεδομένα, με τελικό πλήθος εγγραφών 227214. Τελικό βήμα πριν την επεξεργασία των δεδομένων είναι να χωρίσουμε το δείγμα σε σε δύο μέρη, ένα δείγμα εκπαίδευσης (training sample) και ένα δείγμα ελέγχου (test sample), με το πρώτο να αποτελεί το 80% του συνολικού.

4.2 Προετοιμασία δεδομένων για την μοντελοποίηση

Ένα από τα σημαντικότερα στάδια μοντελοποίησης των δεδομένων είναι η επεξεργασία αυτών ώστε να έρθουν στην κατάλληλη μορφή που απαιτείται για να εισαχθούν στα μοντέλα. Η πρώτη κίνηση είναι να εντοπίσουμε το ποσοστό των ελλιπών δεδομένων σε κάθε μεταβλητή και να θέσουμε ένα όριο έτσι ώστε πάνω από αυτό να αφαιρεθούν οι μεταβλητές αυτές από το δείγμα καθώς δεν μπορούμε να βγάλουμε ασφαλή συμπεράσματα. Στον παρακάτω πίνακα εμφανίζουμε τα αποτελέσματα και στην συνέχεια αφαιρούμε τις μεταβλητές με ποσοστό ελλιπών τιμών πάνω από 70%.

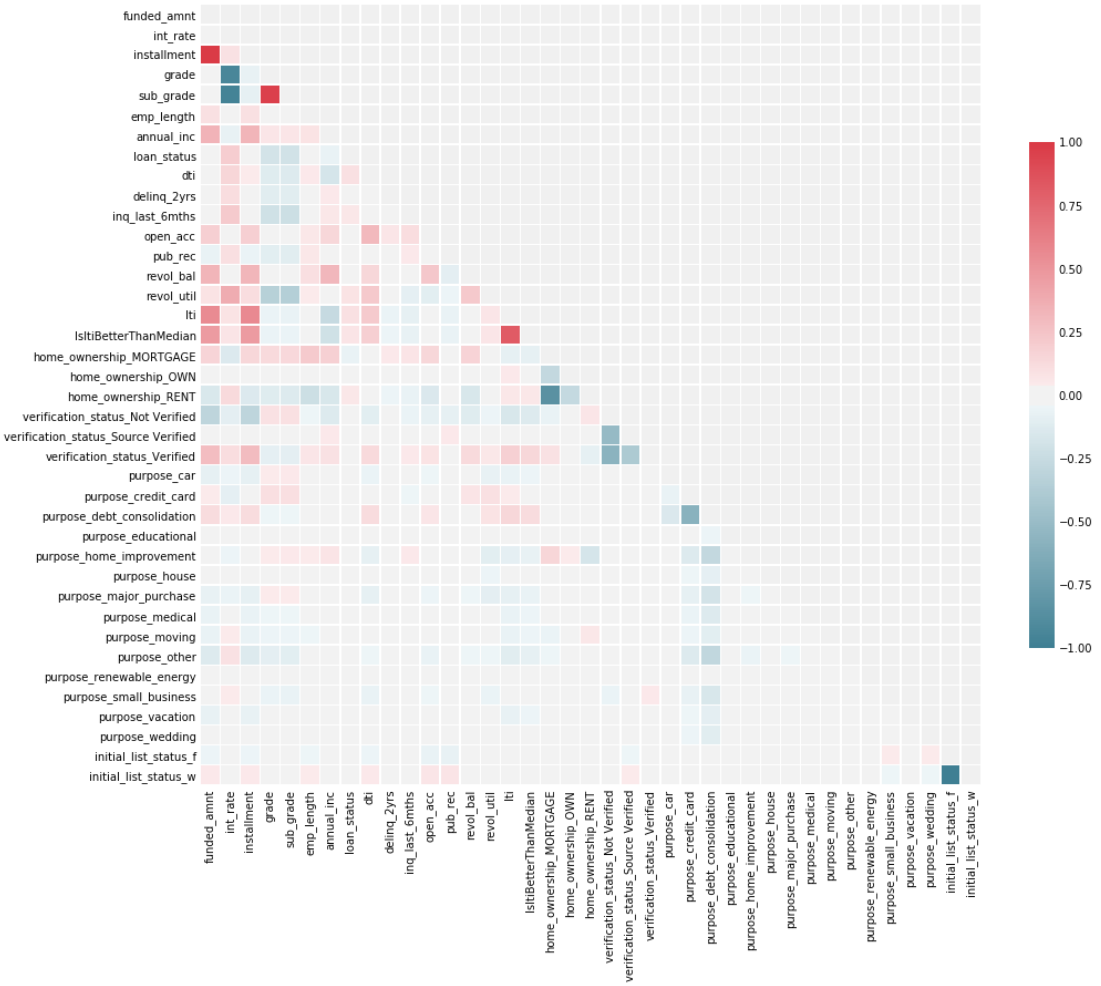
	Missing Values	% of Total Values
open_acc_6m	227214	100.0
open_il_6m	227214	100.0
inq_last_12m	227214	100.0
total_cu_tl	227214	100.0
inq_fi	227214	100.0

all_util	227214	100.0
max_bal_bc	227214	100.0
open_rv_24m	227214	100.0
open_rv_12m	227214	100.0
il_util	227214	100.0
total_bal_il	227214	100.0
mths_since_rcnt_il	227214	100.0
open_il_24m	227214	100.0
open_il_12m	227214	100.0
credit_age	227214	100.0
verification_status_joint	227214	100.0
dti_joint	227214	100.0
annual_inc_joint	227214	100.0
next_pymnt_d	227214	100.0
mths_since_last_record	200569	88.3
mths_since_last_major_derog	187196	82.4
mths_since_last_delinq	128059	56.4
tot_cur_bal	63708	28.0
tot_coll_amt	63708	28.0
total_rev_hi_lim	63708	28.0
emp_length	8673	3.8
last_pymnt_d	364	0.2
revol_util	186	0.1
collections_12_mths_ex_med	56	0.0
last_credit_pull_d	21	0.0
title	13	0.0

Για τις μεταβλητές που περιέχουν μικρότερο πλήθος ελλιπών δεδομένων και αποτελούν σημαντικό κομμάτι της ανάλυσης θα αντικατασταθούν οι τιμές αυτές είτε με την μέση τιμή της μεταβλητής, όπως η μεταβλητή revol_until, είτε με κάποια άλλη τιμή με βάση τον ορισμό της αρχικής. Παράδειγμα της τελευταίας αναφοράς είναι η total_rev_hi_lim όπου βάση ορισμού υποθέτουμε ότι ισούται με την funded_amnt και στην θέση των ελλιπών τιμών αντικαθιστούμε τις αντίστοιχες της μεταβλητής funded_amnt . Αφαιρούμε μεταβλητές όπως collections_12_mths_ex_md, poliy_code, application_type καθώς αποτελούνται από μια μόνο τιμή ως χαρακτηριστικό. Επαναλαμβάνουμε την διαδικασία για να δούμε με τι ποσοστό ακρίβειας προβλέπουμε ότι ο οφειλέτης δεν θα αθετήσει και παρατηρούμε ότι ο λόγος $\frac{\text{Fully Paid} \quad /\Pi\lambda\eta\rho\omega\mu\acute{\varepsilon}\nu\text{ο }\pi\lambda\acute{\eta}\rho\omega\varsigma}{\text{Charged Off} \quad /\text{Απενεργοποιημένο}}$ παραμένει σχεδόν ίδιος και μετά την επεξεργασία των δεδομένων εξασφαλίζοντας ότι η αφαίρεση των μεταβλητών με ελλιπείς τιμές δεν επηρέασε την κατανομή και δεν υπήρξε κάποια διάκριση.

Καθώς αποτελεί σημαντικό κομμάτι στην έρευνα ο χρόνος αθέτησης δημιουργούμε μια νέα μεταβλητή που ισούται με τους μήνες διαφοράς από την μέρα που άνοιξε το πιστωτικό όριο για τον δανειολήπτη και την μέρα που εκδόθηκε. Ταυτόχρονα, μελετώντας όλες τις μεταβλητές παρατηρήσαμε ότι για την μεταβλητή που την κατοικία (home_ownership) υπάρχουν υποκατηγορίες με πολύ μικρό πλήθος σε σύγκριση με το υπόλοιπο δείγμα (ANY, NONE, OTHER) οπότε αφαιρέθηκαν οι αντίστοιχες γραμμές. Ελέγχουμε αν όλες οι μεταβλητές έχουν την κατάλληλη μορφή και αν δεν υπάρχει καμία τιμή που να είναι ελλιπής. Επιπρόσθετα, εισάγουμε άλλη μία μεταβλητή στο δείγμα μας που μετρά το ποσό του δανείου ως προς το αντίστοιχο ετήσιο εισόδημα και την ονομάζουμε “lti”. Παράλληλα, όπως γίνεται αντιληπτό από το γράφημα 28 λόγω των πολλών κρατιδίων κατοικίας η πληροφορία χάνεται ως προς την σχέση αθέτησης με το μέρος διαμονής. Ως εκ τούτου δημιουργούμε μια νέα μεταβλητή "IsltiBetterThanMedian" όπου ομαδοποιεί με βάση την διάμεσο του lti και του κρατιδίου κατοικίας και αφαιρούμε την μεταβλητή addr_state από το δείγμα. Έχοντας ολοκληρώσει την διαδικασία επεξεργασίας των δεδομένων ως προς την ποιότητα των μεταβλητών το συνολικό πλήθος εγγραφών στο δείγμα είναι 171059.

Κατασκευάζουμε το διάγραμμα συσχετίσεων το οποίο μας βοηθά να οπτικοποιήσουμε με πολύ παραστατικό τρόπο τους συντελεστές συσχέτισης πολλών μεταβλητών όπως φαίνεται παρακάτω και να διαπιστώσουμε ποιες συσχετίζονται ισχυρά μεταξύ τους.



Γράφημα 28: Διάγραμμα συσχετίσεων

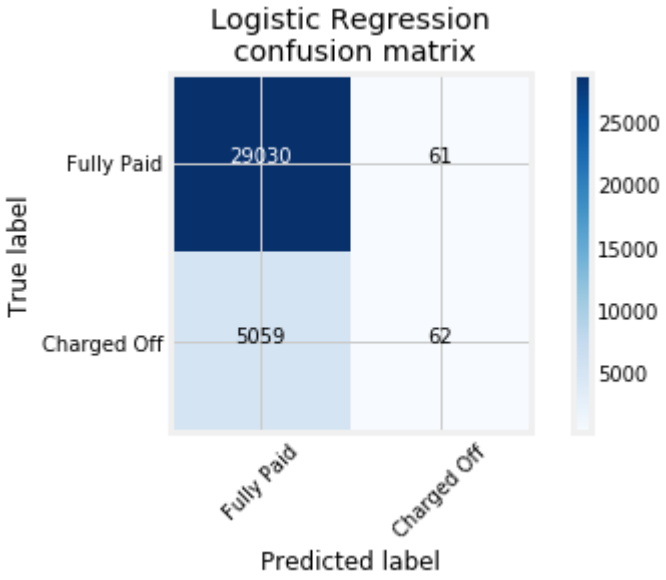
Στο τελικό βήμα της προετοιμασίας για την εφαρμογή των μοντέλων προχωρούμε στην επεξεργασία των ανεξάρτητων μεταβλητών ώστε να μετατρέψουμε όλες τις κατηγορικές μεταβλητές σε ψευδομεταβλητές με τιμές 0 και 1 και τις αντίστοιχες συνεχείς μεταβλητές σε μεταβλητές με μέση τιμή 0 και διακύμανση 1 (κανονικοποίηση). Επισημαίνεται ότι η εξαρτημένη μεταβλητή, η κατάσταση του δανείου, κωδικοποιήθηκε έτσι ώστε η τιμή του 1 να αντιπροσωπεύει την κατάσταση αθέτησης και 0 αντιπροσωπεύει ένα πλήρως καταβληθέν δάνειο Ολοκληρώνοντας την επεξεργασία των δεδομένων μπορούμε να χωρίσουμε με τυχαίο τρόπο το δείγμα μας σε δεδομένα εκπαίδευσης (training) και σε δεδομένα ελέγχου (testing).

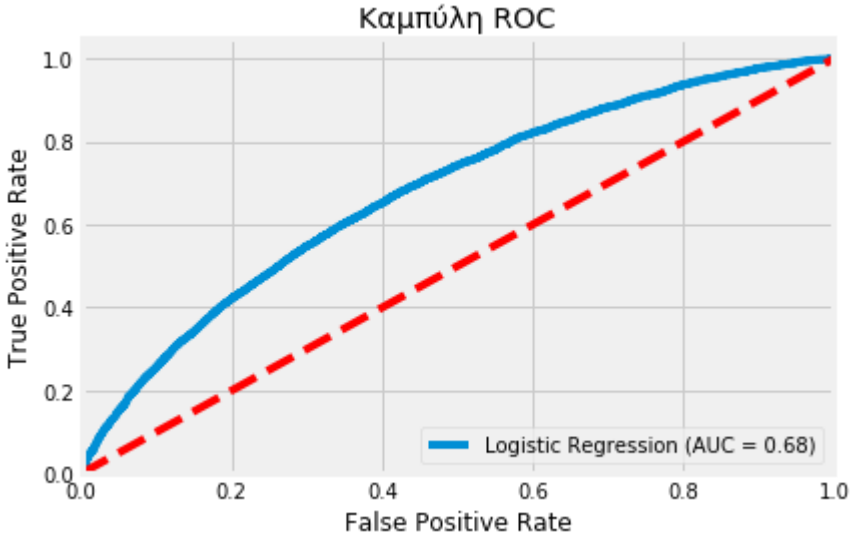
Ο δείκτης ακρίβεια ισούται με το μέγεθος των αποκλίσεων των αποτελεσμάτων από την πραγματική μέση τιμή που ισούται. Το ROC AUC Score που δείχνει το εμβαδόν κάτω από την καμπύλη ROC και με τιμή 1 είναι τέλειος ενώ με 0.5 όπως εδώ δεν δίνει κάποια πληροφορία διαγνωστικής ικανότητας. Προχωράμε στην παρουσίαση των υπόλοιπων μοντέλων στις παρακάτω ενότητες.

4.3 Εφαρμογή Λογιστικής Παλινδρόμησης

Αρχικά, εφαρμόζουμε το μοντέλο λογιστικής παλινδρόμησης, το οποίο αποτελούσε η επικρατέστερη επιλογή τα παλιότερα χρόνια. Τα αντίστοιχα αποτελέσματα των δεικτών απόδοσης παρουσιάζονται παρακάτω. Βλέπουμε ότι ο δείκτης καμπύλης ROC παρουσιάζει μια άνοδο από το βασικό μοντέλο, από 0.68 ενώ το αρχικό έχει 0.5 ενώ υπάρχει πτώση στην ακρίβεια.

ROC AUC Score: 0.68
Accuracy: 0.8503449082193383

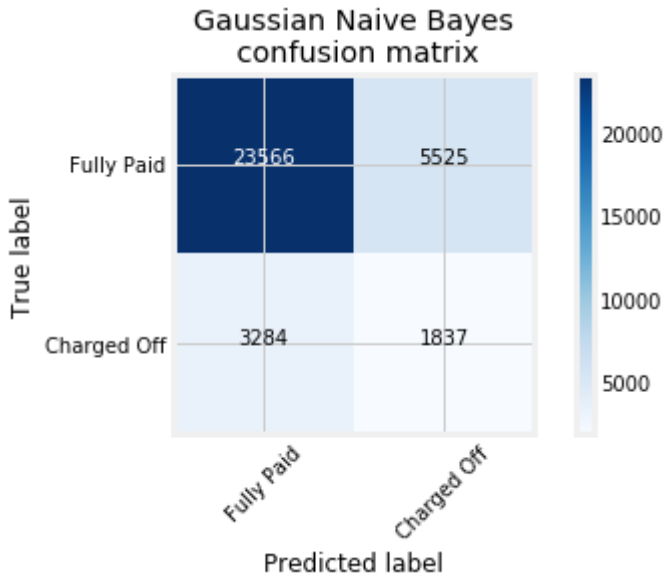


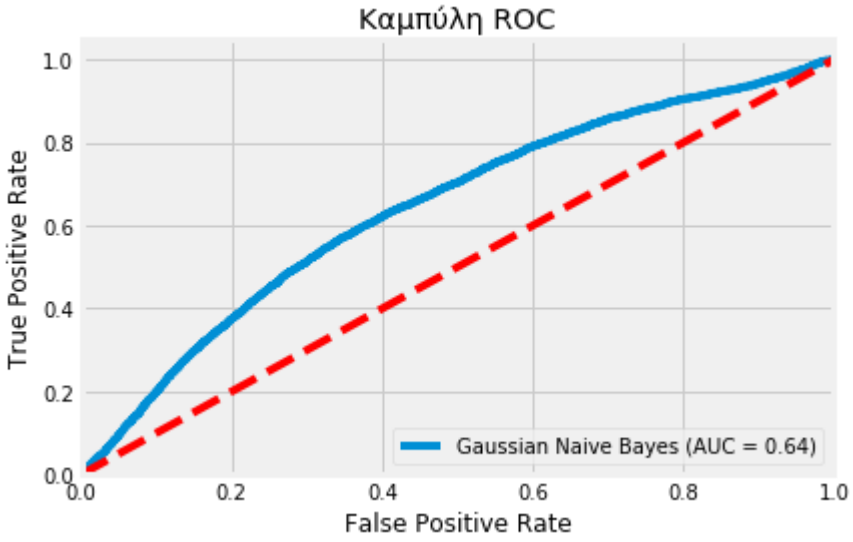


4.4 Μοντέλο Με την μέθοδο Bayes

Το δεύτερο μοντέλο μας έχει την παρακάτω συμπεριφορά:

ROC AUC Score: 0.64
Accuracy: 0.7425172454109669

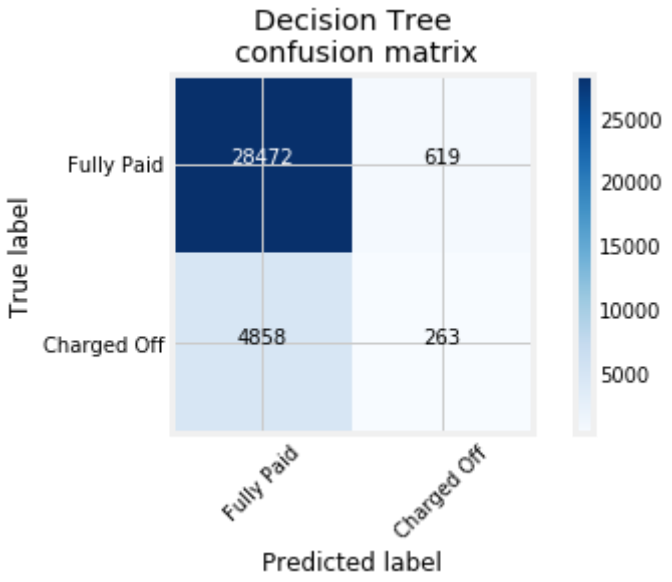


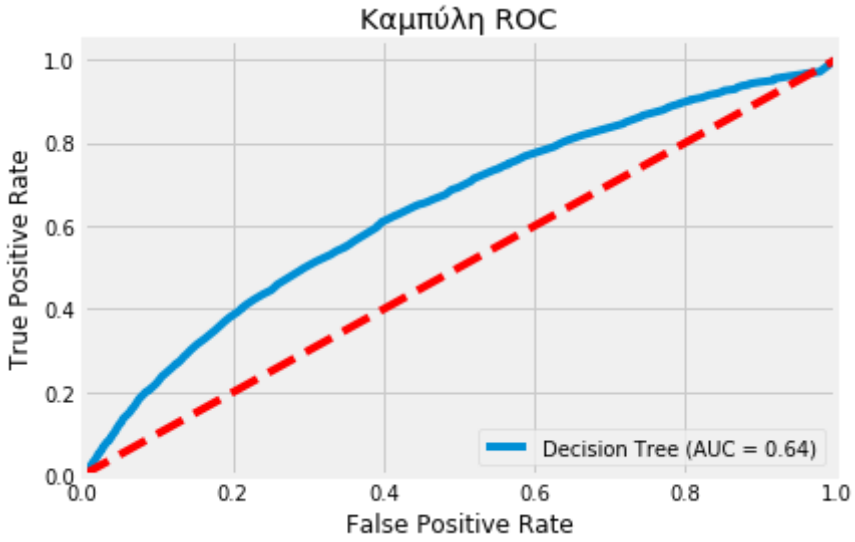


4.5 Μοντέλο Δέντρο απόφασης

Ακολουθεί το δέντρο απόφασης.

ROC AUC Score:	0.64
Accuracy:	0.8399099731088507

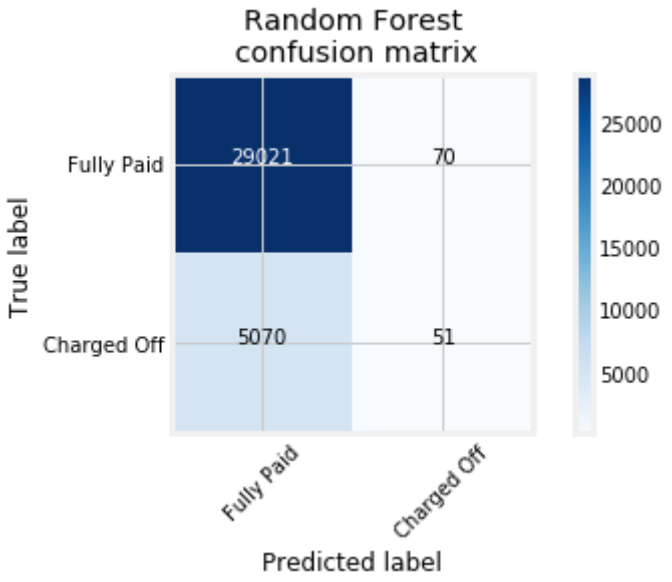


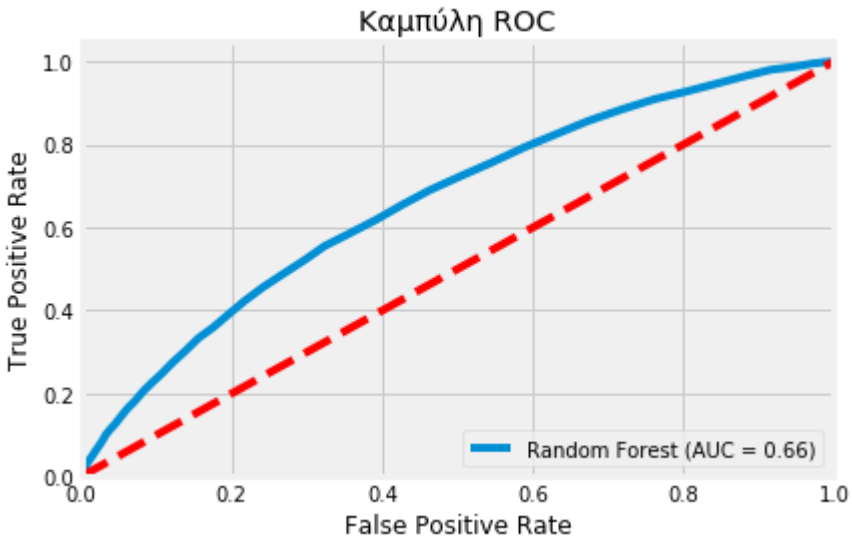


4.6 Μοντέλο Random forest

Το μοντέλο Random Forest δίνει τα εξής αποτελέσματα:

ROC AUC Score:	0.66
Accuracy:	0.8497603180170701

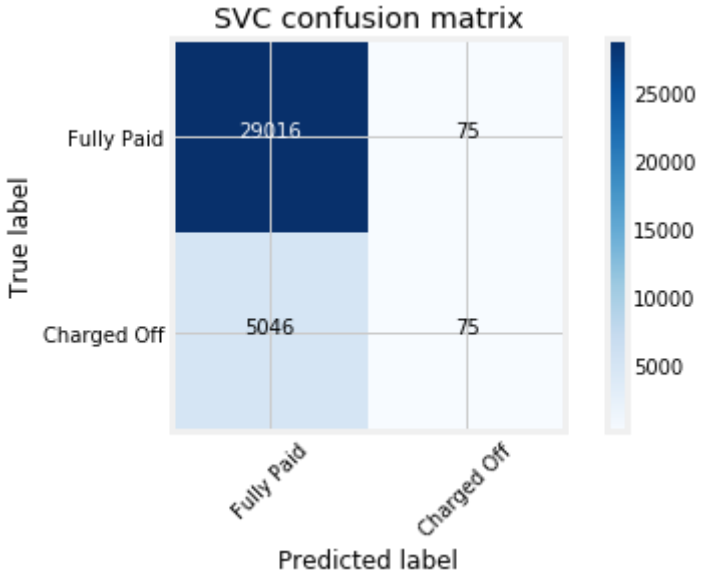


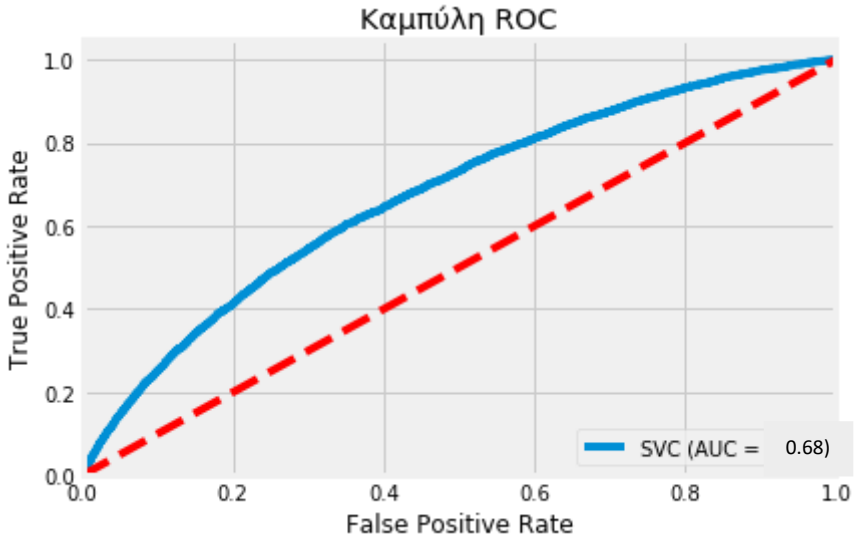


Εύκολα γίνεται φανερό ότι μεταξύ των δύο παραπάνω μοντέλων δεν παρατηρείται κάποια σημαντική διαφορά, αντιθέτως η απόδοση τους είναι σχεδόν ίδια.

4.7 Μοντέλο Μηχανής Διανυσμάτων υποστήριξης με (SVC)

ROC AUC Score: 0.684
Accuracy: 0.8503156787092249

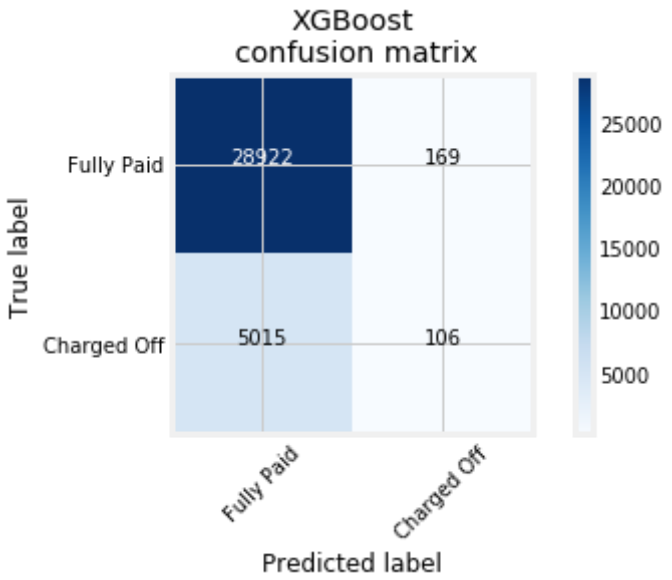


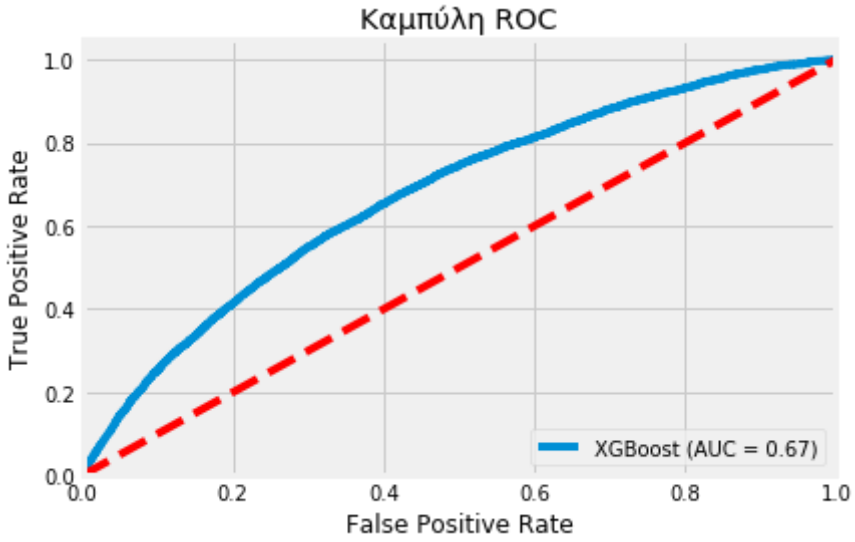


4.8 Μοντέλο XGBoost

Ακολουθώντας την ίδια διαδικασία με τα παραπάνω παίρνουμε:

ROC AUC Score:	0.67
Accuracy:	0.84847421957208

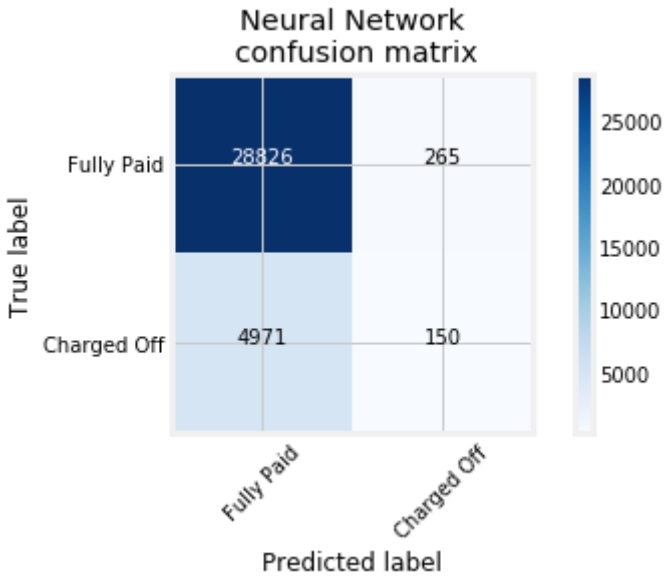


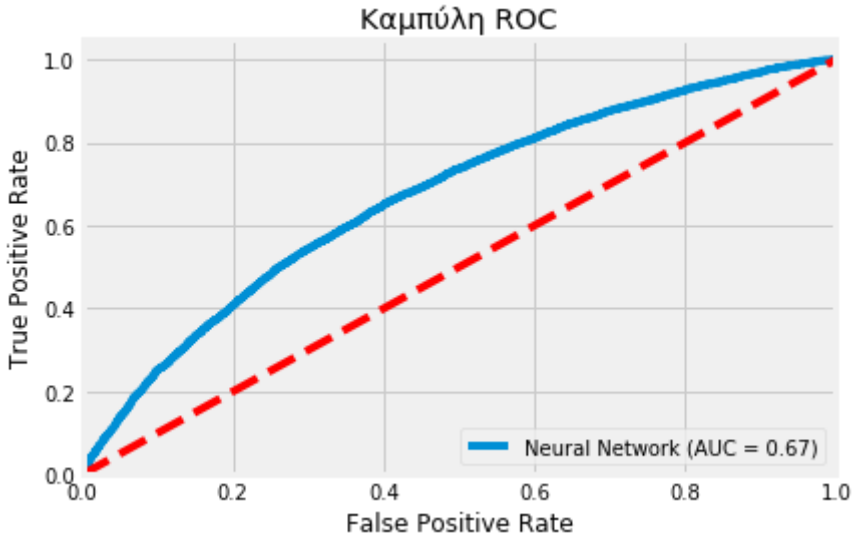


4.9 Μοντέλο Νευρωνικών Δικτύων

Το τελευταίο μοντέλο στην έρευνα μας δίνει:

ROC AUC Score:	0.67
Accuracy:	0.8469542850461826

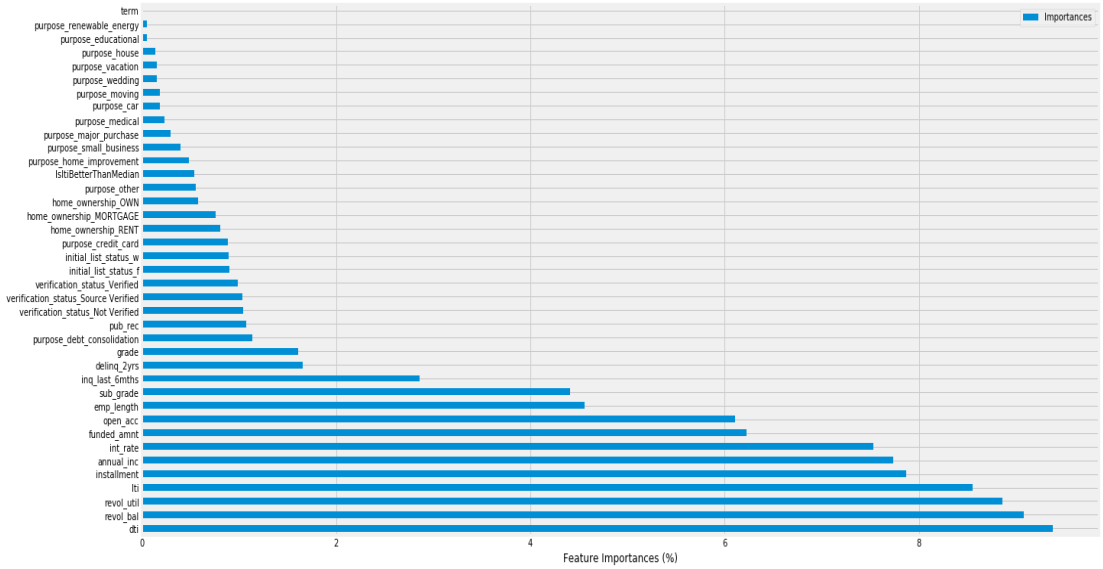




4.10 Σύνοψη μοντέλων

Παρουσιάζουμε την σημασία της κάθε μεταβλητής (%) στο μοντέλο με βάση τον ταξινομητή Random Forest στο δείγμα στον παρακάτω πίνακα και το αντίστοιχο γράφημα.

	Importances
dti	***9.381273***
revol_bal	***9.080829***
revol_util	***8.860815***
lti	***8.555005***
installment	***7.872315***
annual_inc	***7.734726***
int_rate	***7.530042***
funded_amnt	***6.223583***
open_acc	***6.109154***
emp_length	***4.554919***
sub_grade	***4.411575***
inq_last_6mths	***2.859022***
delinq_2yrs	***1.653680***
grade	***1.610087***
purpose_debt_consolidation	***1.138938***
pub_rec	***1.074892***
verification_status_Not Verified	***1.044121***
verification_status_Source Verified	***1.035767***
verification_status_Verified	***0.992174***
initial_list_status_f	***0.899876***
initial_list_status_w	***0.896780***
purpose_credit_card	***0.885377***
home_ownership_RENT	***0.807662***
home_ownership_MORTGAGE	***0.761755***
home_ownership_OWN	***0.581417***
purpose_other	***0.553206***
IsltiBetterThanMedian	***0.540187***
purpose_home_improvement	***0.483342***
purpose_small_business	***0.399646***
purpose_major_purchase	***0.296169***
purpose_medical	***0.232446***
purpose_car	***0.188656***
purpose_moving	***0.187530***
purpose_wedding	***0.157574***
purpose_vacation	***0.153560***
purpose_house	***0.143774***
purpose_educational	***0.055591***
purpose_renewable_energy	***0.052538***
term	***0.000000***



Γράφημα 29: Γράφημα σημαντικότητας μεταβλητών

Παρατηρούμε ότι η πιο σημαντική μεταβλητή είναι το dti δηλαδή ο δείκτης που αναφέρεται στις συνολικές μηνιαίες πληρωμές χρέους του δανειολήπτη για το σύνολο των δανειακών υποχρεώσεων, εξαιρουμένων των ενυπόθηκων δανείων και του αιτούμενου δανείου LC, διαιρούμενο με το μηνιαίο εισόδημα του δανειολήπτη και αφορά το πιστωτικό ιστορικό του. Πιο αδύναμες είναι η διάρκεια και αυτό λόγω δικής μας επιλογής στην έρευνα να μην συμπεριλάβουμε δάνεια διαφορετικής προθεσμίας και ο σκοπός δανείου.

Στην συνέχεια παρουσιάζουμε τον συγκεντρωτικό πίνακα με τις αποδόσεις των παραπάνω μοντέλων.

Model	Accuracy Scores	ROC | Auc
Logistic Regression	0.850	0.680
Random Forest	0.850	0.660
Support Vector	0.850	0.684
XGBoost	0.848	0.670
Neural Network	0.847	0.670
Decision Tree	0.839	0.640
Naive Bayes	0.743	0.640

Με βάση τις μετρήσεις για όλα τα μοντέλα που αναφέρονται στον παραπάνω πίνακα, η απόδοση των μοντέλων ως προς την ακρίβεια και την ROC καμπύλη δεν παρουσιάζουν μεγάλες διαφορές. Ωστόσο παρέχει την καλύτερη απόδοση στο σύνολο

δεδομένων επικύρωσης ο SVC, ο οποίος πέτυχε την υψηλότερη ακρίβεια διατηρώντας παράλληλα παρόμοια βαθμολογία AUC σε σύγκριση με άλλα εκπαιδευμένα μοντέλα

4.11 Εφαρμογή Ανάλυσης Επιβίωσης με Kaplan Meier

Στην προηγούμενη ενότητα εξετάσαμε την αθέτηση ενός οφειλέτη υπό την προοπτική της δυαδικής ταξινόμησης. Στην ενότητα αυτή θα μοντελοποιήσουμε το χρόνο αθέτησης μέσω της ανάλυσης επιβίωσης. Αυτό θα μας επιτρέψει να κατανοήσουμε πότε είναι πιθανό να χρεοκοπήσει το δάνειο. Θα γίνει χρήση του πακέτου **lifelines** όπου μας βοηθά να εφαρμόσουμε τις τεχνικές της ανάλυσης επιβίωσης.

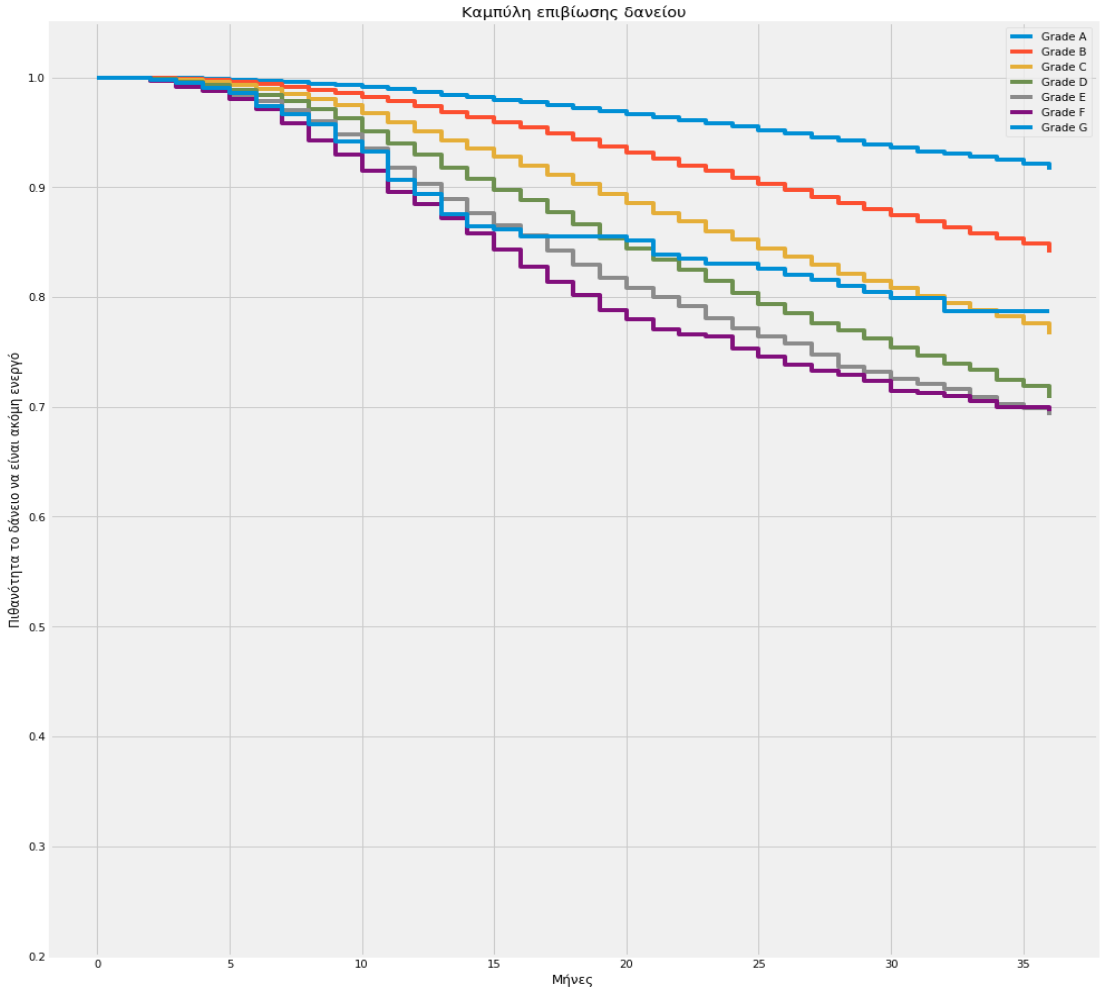
Καθώς μελετάμε τα δεδομένα από μια άλλη οπτική διαφέρει επίσης και το κομμάτι της προ-επεξεργασίας. Οι μεταβλητών με ελλιπείς τιμές (>35%) όπως περιγράψαμε στην ενότητα 2 του κεφαλαίου καθώς επίσης και μεταβλητές που δεν παρέχουν πληροφορία στην μελέτη μας αφαιρούνται. Η διαδικασία είναι ίδια επιπρόσθετα και ως προς την μορφή των μεταβλητών για να εισέλθουν στο μοντέλο μας. Τελευταία και πιο σημαντική προσθήκη μεταβλητής στο δείγμα μας είναι η διάρκεια δανείου, καθώς δεν μας δινόταν από το “Lending Club” και υπολογίσαμε με την βοήθεια των μεταβλητών του επιτοκίου μετατρέποντας το σε μηνιαίο, την μηνιαία δόση και το ποσό που δεσμεύεται εκείνη την στιγμή καθώς το δάνειο τρέχει.

Έχοντας ολοκληρώσει την παραπάνω διαδικασία το δείγμα μας είναι έτοιμο για την εφαρμογή του εκτιμητή Kaplan-Meier. Η λίστα με τις μεταβλητές που χρησιμοποιούνται παρουσιάζεται παρακάτω:

funded_amnt
term
int_rate
installment
grade
sub_grade
emp_length
home_ownership
annual_inc
verification_status
issue_d
purpose
dti
delinq_2yrs
inq_last_6mths
open_acc
pub_rec
revol_util
initial_list_status
total_pymnt

recoveries
time
status
title_word_count
title_polarity
title_subjectivity
title_length
lti

Υπολογίζουμε τον εκτιμητή Kaplan-Meier ο οποίος απαιτεί την γνώση δύο μεταβλητών, το χρόνο επιβίωσης, εδώ του δανείου και την εξαρτημένη μεταβλητή, αν συμβεί δηλαδή η αθέτηση. Η καμπύλη επιβίωσης του Kaplan Meier απεικονίζει την πιθανότητα επιβίωσης σε ένα δεδομένο χρονικό διάστημα. Αποτελεί σημαντικό κομμάτι ανάλυσης καθώς ερμηνεύεται εύκολα η πληροφορία που δίνει το διάγραμμα. Συνεπώς θα κατασκευάσουμε και θα μελετήσουμε τις παρακάτω καμπύλες:

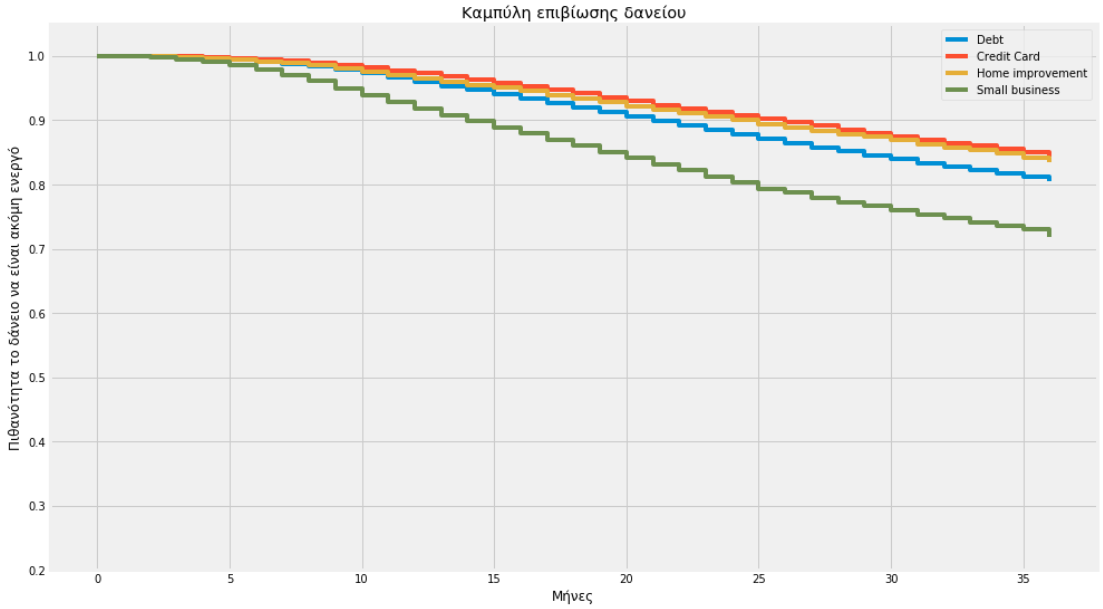


Γράφημα 30: Καμπύλη επιβίωσης δανείου σε σχέση με τον βαθμό αξιολόγησης.

Στο γράφημα παρατηρούμε τις καμπύλες επιβίωσης ως προς τους βαθμούς αξιολόγησης με την καμπύλη που βρίσκεται πιο ψηλά από όλες να αντιπροσωπεύει το

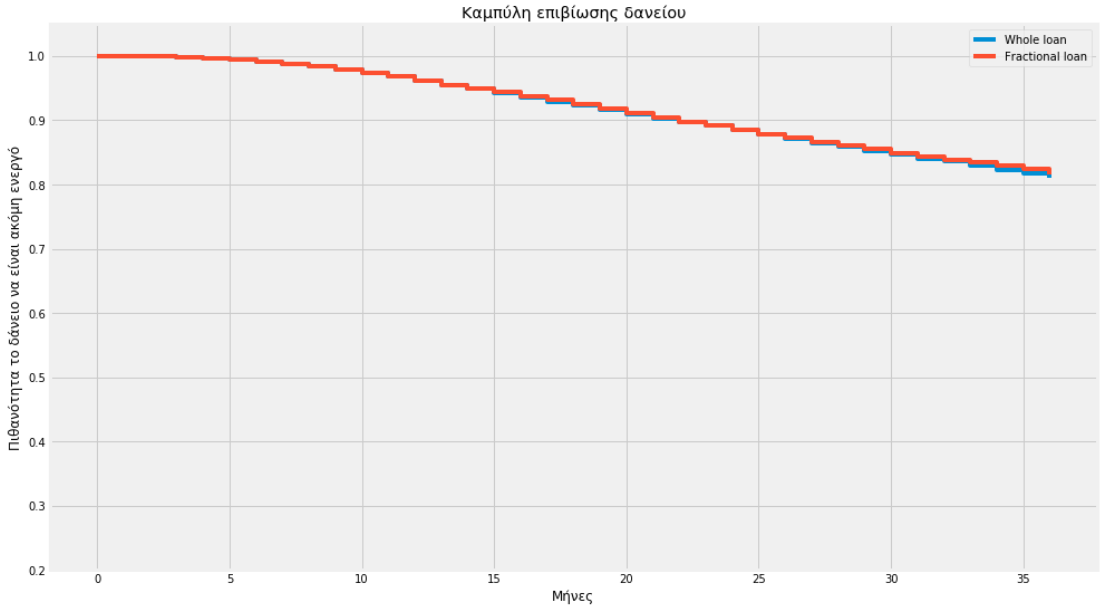
βαθμό «A» και με καθοδικό ρυθμό αντίστοιχα να ακολουθούν και οι υπόλοιπες, δηλαδή δεύτερη από πάνω να είναι η καμπύλη με βαθμό «B». Η καμπύλη επιβίωσης με βαθμό αξιολόγησης «A» έχει υψηλότερους χρόνους επιβίωσης από τις υπόλοιπες. Αυτό είναι μια ένδειξη ότι είναι πιο ασφαλής στην τήρηση των υποχρεώσεων τους οι οφειλέτες που βαθμολογήθηκαν με «A». Η κοινή άποψη δηλαδή ότι όσο καλύτερη η βαθμολόγηση τόσο πιο αξιόπιστος είναι ο οφειλέτης απεικονίζεται και στην καμπύλη επιβίωσης.

Παρατηρούμε επίσης ότι για τον οφειλέτη με βαθμό «G» στον οποίο αντιστοιχεί η τελευταία καμπύλη επιβίωσης, διασταυρώνεται με τις καμπύλες με βαθμό «F», «E», «D» και «C» και η σκαλωτή μορφή της δεν είναι ομαλή όπως οι υπόλοιπες. Η διασταύρωση αυτή αποτελεί ένδειξη ότι δεν ισχύει η υπόθεση του αναλογικού κινδύνου που είναι απαραίτητη για την εφαρμογή του μοντέλου Cox και επίσης, ότι για την καμπύλη αυτή σε σχέση με τις υπόλοιπες που διασταυρώνεται δεν υπάρχει σαφής διαφορά ως προς την ανάλυση επιβίωσης.

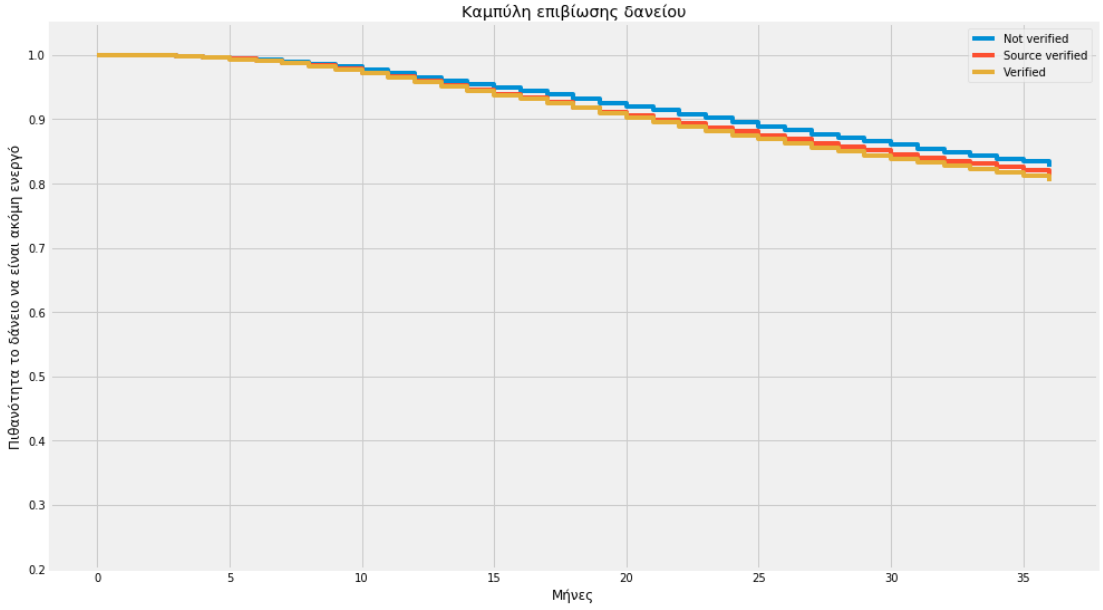


Γράφημα 31: Καμπύλη επιβίωσης δανείου σε σχέση με τον σκοπό αίτησης.

Όσο αφορά τον σκοπό δανειοδότησης, την καλύτερη συμπεριφορά ως προς το χρόνο εκπλήρωσης της υποχρέωσης τους, παρουσιάζουν οι οφειλέτες που σχετίζονται με την πιστωτικής κάρτα και την επισκευή του σπιτιού.



Γράφημα 32: Καμπύλη επιβίωσης δανείου σε σχέση με τον αρχική καταχώρηση του δανείου.



Γράφημα 33: Καμπύλη επιβίωσης δανείου σε σχέση με την επιβεβαίωση ή όχι της πηγής εισοδήματος του δανειολήπτη.

Η πιθανότητα αποπληρωμής σε σχέση με την διάρκεια του δανείου όπως φαίνεται στο γράφημα 35 δεν παρουσιάζει μεγάλη διαφορά ως προς την πιστοποίηση της πηγής εισοδήματος.

Στην συνέχεια, παραθέτουμε τον πίνακα γεγονότων που μας παρέχει πληροφορίες για κάθε χρονική στιγμή ως προς τα δεδομένα και την συμπεριφορά των οφειλετών.

event_at	removed	observed	censored	entrance	at_risk
0	55	0	55	105117	105117
1	980	19	961	0	105062
2	791	77	714	0	104082
3	996	134	862	0	103291
4	1107	192	915	0	102295
5	1171	235	936	0	101188
6	1410	307	1103	0	100017
7	1565	341	1224	0	98607
8	1916	424	1492	0	97042
9	2000	505	1495	0	95126
10	2122	573	1549	0	93126
11	2148	644	1504	0	91004
12	2258	634	1624	0	88856
13	2467	625	1842	0	86598
14	4343	593	3750	0	84131
15	5961	531	5430	0	79788
16	5163	496	4667	0	73827
17	3901	493	3408	0	68664
18	4317	509	3808	0	64763
19	3765	491	3274	0	60446
20	3617	455	3162	0	56681
21	3851	445	3406	0	53064
22	4066	370	3696	0	49213
23	3617	336	3281	0	45147
24	3938	317	3621	0	41530
25	3380	284	3096	0	37592
26	3322	235	3087	0	34212
27	3148	229	2919	0	30890
28	2767	221	2546	0	27742
29	2487	159	2328	0	24975
30	2683	152	2531	0	22488
31	2606	133	2473	0	19805
32	2212	100	2112	0	17199
33	1945	98	1847	0	14987
34	2284	69	2215	0	13042
35	4408	84	4324	0	10758
36	6350	65	6285	0	6350

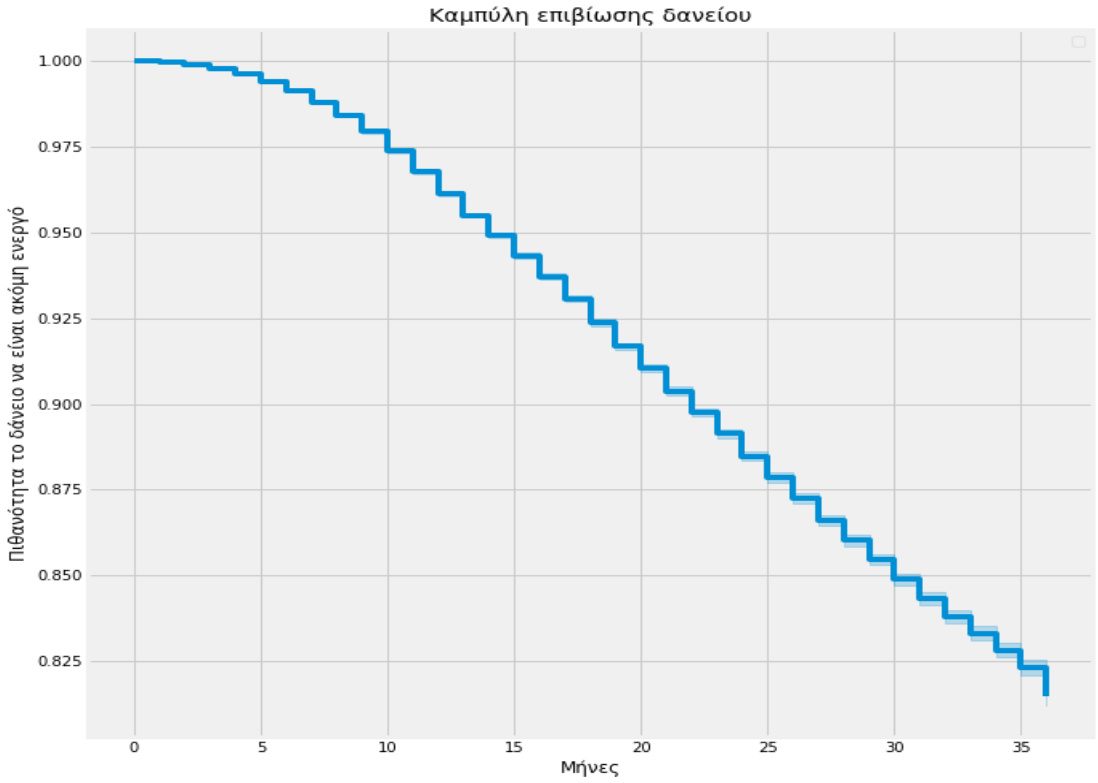
Η στήλη REMOVED αποτελείται από το άθροισμα των άλλων δύο στηλών, OBSERVED και CENSORED , δηλαδή είναι το άθροισμα των παρατηρήσεων που αφαιρέθηκαν είτε επειδή συνέβη το γεγονός, η αθέτηση είτε επειδή ήταν λογοκριμένα. Η στήλη ENTRANCE αποτελείται από νέους οφειλέτες που μπαίνουν στο δείγμα εκείνη την περίοδο.

Βάση του πίνακα μπορούμε να υπολογίσουμε την συνάρτηση ανάλυσης επιβίωσης με τον τύπο να μετατρέπεται στον εξής:

$$t = 0, \qquad \hat{S}(t) = \frac{AT\ RISK - OBSERVED}{AT\ RISK}$$

Οι εκτιμώμενες πιθανότητες επιβίωσης και η αντίστοιχη καμπύλη επιβίωσης δίνονται παρακάτω:

timeline	Verified	timeline	Verified	timeline	Verified	timeline	Verified
0.0	1.000000	10.0	0.971577	20.0	0.902899	30.0	0.838341
1.0	0.999819	11.0	0.964701	21.0	0.895327	31.0	0.832712
2.0	0.999079	12.0	0.957818	22.0	0.888596	32.0	0.827870
3.0	0.997783	13.0	0.950905	23.0	0.881982	33.0	0.822457
4.0	0.995911	14.0	0.944203	24.0	0.875250	34.0	0.818105
5.0	0.993598	15.0	0.937919	25.0	0.868638	35.0	0.811717
6.0	0.990548	16.0	0.931618	26.0	0.862671	36.0	0.803408
7.0	0.987122	17.0	0.924929	27.0	0.856276		
8.0	0.982809	18.0	0.917659	28.0	0.849454		
9.0	0.977592	19.0	0.910205	29.0	0.844047		

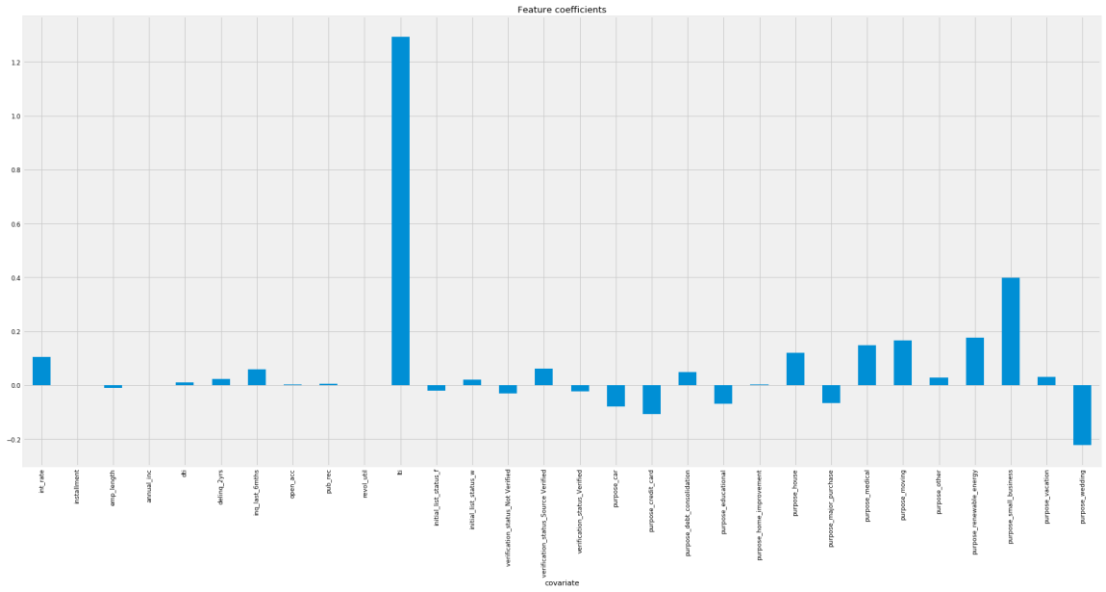


Γράφημα 34: Καμπύλη επιβίωσης δανείου.

Η διάρκεια των δανείων στα δεδομένα να μας είναι 36 μήνες. Παρατηρούμε ότι στην λήξη των δανείου η πιθανότητα να αποπληρώσει το δάνειο ο οφειλέτης είναι 80.34%.

4.12 Εφαρμογή Ανάλυσης Επιβίωσης με το μοντέλο του Cox

Έχοντας υπολογίσει τον εκτιμητή Kaplan Meier και αναπαραστήσει τις αντίστοιχες καμπύλες επιβίωσης για τις μεταβλητές, αφαιρούμε την μεταβλητή «grade» από το δείγμα μας καθώς δεν τηρεί την προϋπόθεση αναλογικού κινδύνου και εφαρμόζουμε το μοντέλο του Cox. Στην ανάλυση του μοντέλου Cox ενδιαφερόμαστε κυρίως για την επίδραση των μεταβλητών στον κίνδυνο, αθέτηση και όχι τόσο για το αν θα συμβεί και πότε το συμβάν. Αρχικά παρουσιάζουμε την σημασία της κάθε μεταβλητής στο μοντέλο με βάση τον Cox στο δείγμα.



Γράφημα 35: Σημαντικότητα μεταβλητών βάσει του μοντέλου του Cox

Παρατηρούμε ότι αποτελεί η μεταβλητή «lti» την σημαντικότερη μεταβλητή στο δείγμα. Στην συνέχεια, παρουσιάζουμε αναλυτικά τα αποτελέσματα που δίνει το μοντέλο ωw προς τους συντελεστές, με έμφαση να δίνουμε στον λόγο κινδύνου (hazard ratio=exp(coef)), στο διάστημα εμπιστοσύνης καθώς και στα p-values.

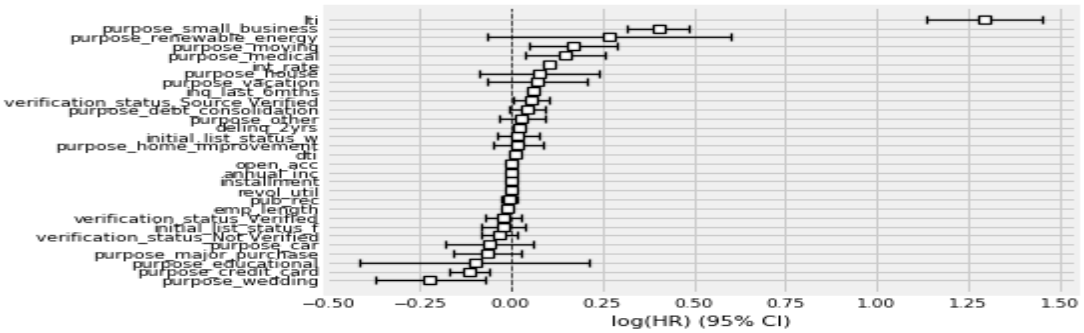
covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p
int_rate	0.105801	1.111601	1.480224e-03	0.102900	0.108703	1.108381	1.114831	71.476578	0.000000e+00
installment	-0.000195	0.999805	3.189974e-05	-0.000258	-0.000133	0.999742	0.999867	-6.122539	9.209582e-10
emp_length	-0.010070	0.989981	1.515346e-03	-0.013040	-0.007100	0.987045	0.992925	-6.645356	3.024861e-11
annual_inc	-0.000002	0.999998	1.828054e-07	-0.000002	-0.000001	0.999998	0.999999	-9.719980	2.478290e-22
dti	0.010202	1.010254	7.463152e-04	0.008739	0.011665	1.008777	1.011733	13.669523	1.544039e-42
delinq_2yrs	0.023113	1.023382	6.390538e-03	0.010588	0.035638	1.010644	1.036281	3.616783	2.982877e-04

covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p
inq_last_6mths	0.058959	1.060732	4.314399e-03	0.050503	0.067415	1.051800	1.069740	13.665694	1.627451e-42
open_acc	0.001658	1.001660	1.191037e-03	-0.000676	0.003993	0.999324	1.004001	1.392405	1.637997e-01
pub_rec	0.004260	1.004269	9.372944e-03	-0.014111	0.022631	0.985989	1.022889	0.454509	6.494628e-01
revol_util	-0.000186	0.999814	2.525726e-04	-0.000681	0.000309	0.999320	1.000309	-0.734789	4.624681e-01
lti	1.293131	3.644178	7.166061e-02	1.152679	1.433583	3.166664	4.193699	18.045211	8.603455e-73
initial_list_status_f	-0.020239	0.979965	2.677074e-02	-0.072708	0.032231	0.929872	1.032756	-0.756003	4.496474e-01
initial_list_status_w	0.020239	1.020445	2.677074e-02	-0.032231	0.072708	0.968283	1.075417	0.756003	4.496474e-01
verification_status_Not Verified	-0.030464	0.969996	2.243407e-02	-0.074434	0.013506	0.928269	1.013598	-1.357922	1.744883e-01
verification_status_Source Verified	0.061074	1.062978	2.274281e-02	0.016499	0.105649	1.016636	1.111432	2.685430	7.243651e-03
verification_status_Verified	-0.024239	0.976053	2.261897e-02	-0.068571	0.020094	0.933727	1.020297	-1.071614	2.838933e-01
purpose_car	-0.078080	0.924890	5.587375e-02	-0.187591	0.031430	0.828954	1.031929	-1.397442	1.622807e-01
purpose_credit_card	-0.107692	0.897904	2.400266e-02	-0.154736	-0.060647	0.856641	0.941155	-4.486659	7.234870e-06
purpose_debt_consolidation	0.048504	1.049699	2.288122e-02	0.003657	0.093350	1.003664	1.097846	2.119795	3.402333e-02
purpose_educational	-0.069654	0.932716	1.388907e-01	-0.341875	0.202567	0.710437	1.224542	-0.501502	6.160178e-01
purpose_home_improvement	0.001451	1.001452	3.139937e-02	-0.060090	0.062993	0.941679	1.065019	0.046219	9.631360e-01
purpose_house	0.120625	1.128202	7.258186e-02	-0.021633	0.262883	0.978600	1.300674	1.661918	9.652912e-02
purpose_major_purchase	-0.066302	0.935849	4.275045e-02	-0.150091	0.017488	0.860630	1.017642	-1.550898	1.209261e-01
purpose_medical	0.148152	1.159690	4.929437e-02	0.051537	0.244767	1.052888	1.277324	3.005461	2.651787e-03
purpose_moving	0.164890	1.179263	5.517884e-02	0.056741	0.273038	1.058382	1.313951	2.988280	2.805525e-03
purpose_other	0.026916	1.027282	2.951733e-02	-0.030936	0.084769	0.969537	1.088466	0.911886	3.618285e-01
purpose_renewable_energy	0.175254	1.191549	1.569539e-01	-0.132370	0.482878	0.876017	1.620733	1.116597	2.641667e-01
purpose_small_business	0.398194	1.489132	3.891353e-02	0.321925	0.474463	1.379781	1.607151	10.232784	1.413954e-24
purpose_vacation	0.030609	1.031083	6.382443e-02	-0.094484	0.155703	0.909842	1.168479	0.479585	6.315223e-01

Επιπλέον, αν αφαιρέσουμε από το δείγμα μας τις μεταβλητές με δείκτη πληθωρισμού διακύμανσης πάνω από 5 με στόχο την αντιμετώπιση της συγγραμικότητας στο μοντέλο λαμβάνουμε:

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p
int_rate	0.11	1.12	0.00	0.11	0.11	1.11	1.12	77.24	<0.005
installment	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-5.59	<0.005
emp_length	-0.01	0.99	0.00	-0.01	-0.01	0.99	0.99	-6.80	<0.005
annual_inc	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-9.06	<0.005
dti	0.01	1.01	0.00	0.01	0.01	1.01	1.01	13.88	<0.005
delinq_2yrs	0.02	1.03	0.01	0.01	0.04	1.01	1.04	3.89	<0.005
inq_last_6mths	0.06	1.06	0.00	0.05	0.07	1.05	1.07	13.51	<0.005
open_acc	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.69	0.49
pub_rec	0.01	1.01	0.01	-0.01	0.03	0.99	1.03	0.98	0.33
revol_util	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-2.46	0.01
lti	1.27	3.55	0.07	1.13	1.41	3.09	4.08	17.86	<0.005

Για τις μεταβλητές με τιμή κινδύνου μεγαλύτερη του 1, συμπεραίνουμε ότι όσο μεγαλύτερη η τιμή τόσο μεγαλύτερος ο κίνδυνος και τόση μικρότερη η διάρκεια επιβίωσης. Αντίστοιχα, όταν η τιμή του κινδύνου είναι μικρότερη του 1, ο κίνδυνος μειώνεται και αυξάνεται η διάρκεια επιβίωσης. Γνωρίσουμε ότι για $p<0.05$ η μεταβλητή είναι σημαντική για την απόδοση του μοντέλου. Στο δείγμα μας η μεταβλητή lti έχει p-value<0.005 και ο κίνδυνος είναι 3.55. Αυτό αποτελεί ισχυρή ένδειξη μεταξύ του κινδύνου και της μεταβλητής, 255% υψηλότερος ο κίνδυνος για τα άτομα με υψηλότερο lti. Η απεικόνιση των παραπάνω τιμών σε γράφημα δίνεται παρακάτω για την καλύτερη ερμηνεία.



Γράφημα 36: Αναλογία κινδύνου σε σχέση με τις μεταβλητές (μοντέλο του Cox).

Στο επόμενο βήμα, υπολογίζουμε την συνάρτηση επιβίωσης που παίρνουμε με το μοντέλο του Cox η οποία θα μπορούσε να συγκριθεί με την αντίστοιχη του Kaplan Meier.

timeline	Verified	timeline	Verified	timeline	Verified	timeline	Verified
0.0	1.000.000	10.0	0.977210	20.0	0.920986	30.0	0.864870
1.0	0.999835	11.0	0.971852	21.0	0.915077	31.0	0.859653
2.0	0.999250	12.0	0.966278	22.0	0.909410	32.0	0.854707
3.0	0.998218	13.0	0.960654	23.0	0.903813	33.0	0.850008
4.0	0.996881	14.0	0.955305	24.0	0.897916	34.0	0.845153
5.0	0.994901	15.0	0.950190	25.0	0.892220	35.0	0.840351
6.0	0.992519	16.0	0.944791	26.0	0.886818	36.0	0.832658
7.0	0.989602	17.0	0.939074	27.0	0.880664		
8.0	0.986140	18.0	0.932908	28.0	0.875413		
9.0	0.982085	19.0	0.926702	29.0	0.870242		

Παρατηρούμε ότι οι διαφορές είναι πολύ μικρές, με την πιθανότητα επιβίωσης στην λήξη του δανείου να είναι 83.26% σύμφωνα με το μοντέλο του Cox έναντι 80.34% με τον εκτιμητή Kaplan Meier.

5. Σύνοψη

Ο σκοπός της διπλωματικής αυτής εργασίας όπως ορίστηκε και στις παραπάνω ενότητες ήταν η εκτίμηση της πιθανότητας αθέτησης δανείων με μια ευρύτερη έννοια δηλαδή τόσο με την χρήση της λογιστικής παλινδρόμησης, μοντέλο που χρησιμοποιείται πολλά χρόνια για τον σκοπό αυτό όσο και από την πλευρά των μεθόδων μηχανικής μάθησης που εντείνεται η χρήση και η απόδοση τους συνεχόμενα τα τελευταία χρόνια με την χρήση μεγάλων βάσεων δεδομένων. Ξεχωριστό κομμάτι αποτελεί η ανάλυση επιβίωσης στην πρόβλεψη, ο χρόνος αθέτησης του δανείου προσθέτει μια πολύ σημαντική πληροφορία στο τελικό συμπέρασμα και επιπρόσθετα χειρίζεται λογοκριμένα δεδομένα όπως αναλύθηκε στην Ενότητα 2. Σε αυτόν τον σκοπό, προστέθηκε η μελέτη της ανάλυσης επιβίωσης με χρήση του εκτιμητή Kaplan Meier και του μοντέλου του Cox.

Για την επιλογή του καλύτερου μοντέλου πρόβλεψης αποτελεί σημαντικό κομμάτι η ποιότητα και η διαθεσιμότητα δεδομένων σε όλο το δείγμα. Αυτό βασίζεται αφενός στην επιδίωξη μεγαλύτερης ακρίβειας στα δεδομένα καθώς και δυνατότητα προσαρμογή τους στην κατάλληλη μορφή ώστε να εισέλθουν στο μοντέλο. Στο τέταρτο Κεφάλαιο δόθηκε η μορφή των δεδομένων ως προς την περιγραφή των μεταβλητών τόσο των δανειοληπτών όσο και της αγοράς την χρονική στιγμή που εκδόθηκαν.

Η απόδοση του μοντέλου της Μηχανής Διανυσμάτων υποστήριξης είναι η καλύτερη συγκριτικά με τα υπόλοιπα μοντέλα. Ακολουθούν η Λογιστική Παλινδρόμηση, τα Νευρωνικά Δίκτυα και το XGBoost όπου είναι πολύ κοντά ως προς την ακρίβεια και το εμβαδόν AUC. Η απλότητα των μεταβλητών και το μικρό σχετικά πλήθος του δείγματος, καθώς και η έλλειψη διαφορετικών δανείων σε διάρκεια αποτελεί έναν από τους λόγους που η επιλογή μοντέλου δεν είναι εμφανής και τα κριτήρια ακρίβειας και AUC δίνουν παραπλήσια αποτελέσματα.

Η καλύτερη επεξεργασία δεδομένων και η δημιουργία-προσθήκη περισσότερων μεταβλητών στην εργασία θα αποτελούσε σημαντικό κομμάτι για μελλοντική βελτίωση.

6. Βιβλιογραφία

Petropoulos, A., Siakoulis,V., Stavroulakis, E. and Klamargias. A. (2018). *A robust machine learning approach for credit risk analysis of large loan-level datasets using deep learning and extreme gradient boosting.* Bank of Greece.

Τράπεζα της Ελλάδος. (2007). *Υπολογισμός Κεφαλαιακών Απαιτήσεων έναντι του Πιστωτικού Κινδύνου σύμφωνα με την Τυποποιημένη Προσέγγιση.* Πράξη Διοικητή Αριθμ.2588/20.8.2007.

Χαραλαμπίδης. (2004). *Οι νέες προτάσεις για την αναθεώρηση του Πλαισίου Κεφαλαιακής Επάρκειας των Πιστωτικών Ιδρυμάτων: Παρουσίαση, ανάλυση και κριτική, Προβόπουλος Γ. και Γκόρτσος Χ. (επιμέλεια),* Ένωση Ελληνικών Τραπεζών-Εκδόσεις Αντ. Ν. Σάκκουλα, 151-196.

A. Ciampi, R. S. Bush, M. Gospodarowicz, and J. E. Till. (1981). An approach to classifying prognostic factors related to survival experience for non-Hodgkin's lymphoma patients: Based on a series of 982 patients, Cancer. 621–627.

Addo, P., Guegan, D. and Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models, University Ca' Foscari of Venice, Dept. of Economics Research Paper Series No. 08/WP/2018.

Agresti. (2007). *An Introdution to Categorical Data Analysis,* Wiley.

Andersen, P. K. and Keiding, N. (2006). Survival and event history analysis, Wiley.

Austin, P. C., Lee, D. S., Fine, J. P. (2016). Introduction to the Analysis of Survival Data in the Presence of Competing Risks .

Baesens, B., Egmont-Petersen, M., Castelo, R., and Vanthienen, J. (2001). Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search, Utrecht University, Institute of Computer and Information Sciences.

Banasik J, Crook J and Thomas L. (1999). Not if but when will borrowers default, The Journal of the Operational Research Society 50: 1185-1190.

Basel Committee on Banking Supervision, (June (2006a)). Results of the fifth quantitative impact study (QIS 5), Bank for International Settlements.

Basel Committee on Banking Supervision, (June(2006b)). International Convergence of Capital Measurement and Capital Standards: A Revised Framework.

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis, The Journal of the Operational Research Society.

Breiman, L., (2001). Random Forests, Machine Learning 45 (1) pp. 5–32.

Cao R, Vilar JM, Devia A (2009). Modelling consumer credit risk via survival analysis. SORT 33 (1) January-June 2009, 3-30.

Chen,W., Shih, J. (2006). A study of Taiwan's issuer credit rating systems using support vector machines, Expert Systems with Applications.

Committee, Basel., (2014). Basel III leverage ratio framework and disclosure requirements.

Cox, D. R. (1972). Regression models and life-tables, J. Royal Statist. Society, 187-220.

Cox, D. R. (1975). Partial likelihood, Biometrika, 269-276.

Cox, D. R. and Oakes. (1984): Analysis of Survival Data. London: Chapman and Hall.

Davis, R. H., Edelman, D. B., Gammerman, A. J. (1992). Machine-learning algorithms for credit-card applications, IMA Journal of Management Mathematics, 4.

DG., Altman. (1992). Analysis of Survival times.In:Practical statistics for Medical research, Chapman and Hall.

Dirick L, Claeskens G, Baesens B. (2015). An Akaike information criterion for multiple event mixture cure models, European Journal of Operational Research.

Dirick,L., Claeskens, G. and Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study, Journal of the Operational Research Society.

Ederer, Sidney J. Cutler and Fred. (1958). Maximum utilization of the life table method in analyzing survival, Journal of Chronic Diseases.

Faraggi, D. and Simon, R. (1995). A neural network model for survival data, Statistics in Medicine.

Fard, M. J., Wang, P., Chawla, S. and Reddy, C. K. (2016). A bayesian perspective on early stage event prediction in longitudinal data, Transactions on Knowledge and Data Engineering.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning.

Galindo, J. and Tamayo, P. (2000).Credit Risk Assessment using Statistical and Machine Learning, Basic Methodology and Risk Modeling Applications,Computational Economics, Vol. 15.

Hastie, T., Tibshirani, R. and Friedman, J. (2009).Elements of Statistical Learning, Springer.

Hosmer, D. and Lemeshow, L. (1999). Applied survival analysis: regression modeling of time to event data, John Wiley & Sons.

Hosmer and Lemeshow. (2000) Applied Logistic Regression. Wiley.

Kaplan, E and Meier, P. (1958). Nonparametric estimation from incomplete observations, Journal of American Statistical Association

Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T. and Kluger, K.(2018). "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Medical Research Methodology volume.

Kleinbaum, D.G. and Klein, M. (2005). Statistics for Biology and Health, Survival Analysis: A Self Learning Text. 2. New York: Springer-Verlag Publishers.

Lee, E., Wang, J. (2003). Statistical Methods for Survival Data Analysis, Fourth Edition, John Wiley & Sons.

Lee, C., Zame, W., Yoon, J., Schaar, M., (2018). DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks, Department of Electrical and Computer Engineering.

Narain, B. (1992) Survival analysis and the credit granting decision. In: Thomas L, Crook JN, Edelman DB (eds.) Credit Scoring and Credit Control. OUP: Oxford. pp: 109-121.

Olshen, Gordon, L. and Richard A. (1985).Tree-structured survival analysis. Cancer Treatment Reports.

Prentice, Kalbfleisch J. D. and Ross L. (2011). The Statistical Analysis of Failure Time Data. John Wiley & Sons.

Raftery, A. E. (1995). Bayesian model selection in social research, Sociological Methodology.

Roy, P, V. (2005). Credit ratings and the standardised approach to credit risk in Basel II.Social Science Research Network.

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data, Operations Research Quarterly.

Y. LeCun, L. Bottou, G. Orr, K. Muller. (1998). Neural Networks: Tricks of the Trade, Springer Berlin.

Nijskens, R. & Wagner, W. (2011). Credit risk transfer activities and systemic risk: How banks became less risky individually but posed greater risks to the financial system at the same time, Journal of Banking & Finance.