ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ ΚΑΙ ΥΠΗΡΕΣΙΕΣ»
ΚΑΤΕΥΘΥΝΣΗ: ΠΡΟΗΓΜΕΝΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Χρήση μεθόδων *Μηχανικής Μάθησης στο τομέα της Εφοδιαστικής Αλυσίδας: Πρόβλεψη Backorder*

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΥΘΥΜΙΟΣ ΟΡΦΑΝΙΔΗΣ

Υπεύθυνος Καθηγητής: Μιχαήλ Φιλιππάκης

Πρόλογος

Καθώς ο όγκος των δεδομένων που έχουν να διαχειριστούν οι επιχειρήσεις τα τελευταία χρόνια έχει αυξηθεί ραγδαία, γεγονός στο οποίο έχει συμβάλει αναμφίβολα η διάδοση του διαδικτύου και των ηλεκτρονικών πλατφορμών που το απαρτίζουν, ο τομέας της μηχανικής μάθησης και των μεγάλων δεδομένων γίνεται όλο και πιο χρήσιμος στην σωστή διαχείριση αυτών των δεδομένων, εξοικονομώντας χρόνο και πολύτιμους πόρους στις επιχειρήσεις.

Στη συγκεκριμένη διπλωματική εργασία θα εστιάσουμε στο κομμάτι του Supply Chain και πιο συγκεκριμένα στο κομμάτι που αφορά το Backorder Prediction, το οποίο αποτελεί πλέον ένα αντικείμενο ιδιαίτερης σημασίας σε μεσαίες και μεγάλες επιχειρήσεις λόγω της υπερκατανάλωσης αλλά και της απρόβλεπτης/εναλλασσόμενης φύσης της ζήτησης της αγοράς. Επειδή ο αριθμός των backorder είναι πολύ μικρότερος από τον αριθμό των παραγγελιών που αποστέλλονται εγκαίρως, η εφαρμογή ενός αποτελεσματικού μοντέλου πρόβλεψης για αυτόν τον τομέα είναι μια πρόκληση.

Θα γίνει χρήση και σύγκριση μεταξύ διαφόρων μεθόδων και μοντέλων μηχανικής μάθησης και deep learning για να υπολογίσουμε την περίπτωση που ένα προϊόν θα βρεθεί σε backorder, δηλαδή μια παραγγελία που δεν μπορεί να ολοκληρωθεί άμεσα λόγω έλλειψης του προϊόντος.

Η δομή της εργασίας αρχίζει με μία αυτοματοποιημένη διαδικασία γραμμένη σε Python, όπου λαμβάνονται όλα τα αρχεία μορφής csv που περιέχουν τις στήλες με τα δεδομένα μας και γίνεται μια βασική μετατροπή και καθαρισμός σε αυτά ώστε να έρθουν σε μορφή αναγνωρίσιμη από SQL database. Έπειτα γίνεται η επιθυμητή προ-επεξεργασία των δεδομένων και τέλος ανεβαίνουν στην cloud υπηρεσία της Microsoft, Azure SQL, όπου είναι δυνατή και η περεταίρω διαχείριση και ανάγνωσή τους. Στη συνέχεια, για να τρέξουμε τα μοντέλα μηχανικής μάθησης και να κάνουμε οπτικοποίηση, σύγκριση και ερμηνεία των αποτελεσμάτων, μετατρέπουμε τους πίνακες από την βάση δεδομένων σε dataframes ώστε να συνεχιστούν οι προαναφερθείς διαδικασίες στο Jupyter Notebook.

Abstract

As the amount of data that businesses have to manage has increased rapidly in recent years, a fact to which the spread of the internet and the electronic platforms that make it up have undoubtedly contributed, the field of machine learning and big data is becoming increasingly useful in the proper management of this data, saving time and valuable resources in businesses.

In this particular thesis we will focus on the part of the Supply Chain and more specifically on the part concerning Backorder Prediction, which is now a task of particular importance in medium and large companies due to the overconsumption and also the unpredictable/changing nature of market demand. Because the number of backorders is much smaller than the number of orders shipped on time, implementing an effective forecasting model for this sector is a challenge.

Various machine learning and deep learning methods and models will be used and compared, to calculate if a product will be backordered, i.e., an order that cannot be completed immediately due to a lack of availability of the product.

The structure of the thesis begins with an automated process written in Python, where all the csv files containing the columns with our data are received and a basic conversion and cleaning is done to them, so that they transform into a format recognizable by SQL database. Then, the desired pre-processing of the data is done and finally they are uploaded to Microsoft's cloud service, Azure SQL, where they can be further managed and read. Afterwards, in order to run the machine learning models and visualize, compare and interpret the results, we convert the tables from the database into dataframes to continue the aforementioned processes in Jupyter Notebook.

Table of Contents

Ενότητα **Σελίδα**

Εφοδιαστική Αλυσίδα (Supply Chain)

Στο εμπόριο, μια αλυσίδα εφοδιασμού αναφέρεται στο δίκτυο οργανισμών, ανθρώπων, δραστηριοτήτων, πληροφοριών και πόρων που εμπλέκονται στην παράδοση ενός προϊόντος ή μιας υπηρεσίας σε έναν καταναλωτή. Οι δραστηριότητες της εφοδιαστικής αλυσίδας περιλαμβάνουν τη μετατροπή φυσικών πόρων, πρώτων υλών και εξαρτημάτων σε τελικό προϊόν και την παράδοση του στον τελικό πελάτη. Σε εξελιγμένα συστήματα εφοδιαστικής αλυσίδας, τα χρησιμοποιημένα προϊόντα μπορούν να εισέλθουν ξανά στην αλυσίδα εφοδιασμού σε οποιοδήποτε σημείο όπου η υπολειμματική αξία είναι ανακυκλώσιμη. Οι αλυσίδες εφοδιασμού συνδέουν αλυσίδες αξίας. Μια αλυσίδα αξίας είναι ένα σύνολο δραστηριοτήτων που εκτελεί μια επιχείρηση που δραστηριοποιείται σε έναν συγκεκριμένο κλάδο προκειμένου να παραδώσει ένα πολύτιμο προϊόν (δηλαδή αγαθό ή/και υπηρεσία) στον τελικό πελάτη.

Μπορούμε να χωρίσουμε την διαχείριση των παραδοσιακών συστημάτων της εφοδιαστικής αλυσίδας σε πέντε στοιχεία:

- Σχεδιασμός (Planning): Περιλαμβάνει τον σχεδιασμό και την διαχείριση όλων των πόρων που απαιτούνται για την κάλυψη της ζήτησης των πελατών για το προϊόν ή την υπηρεσία μιας εταιρείας. Όταν δημιουργηθεί η εφοδιαστική αλυσίδα, καθορίζονται μετρήσεις για να φανεί εάν η αλυσίδα εφοδιασμού είναι αποτελεσματική, αποδοτική, προσφέρει αξία στους πελάτες και πληροί τους στόχους της εταιρείας.
- Προμήθειες (Sourcing): Επιλογή προμηθευτών για να παρέχουν τα αγαθά και τις υπηρεσίες που απαιτούνται για τη δημιουργία του προϊόντος. Στη συνέχεια, καθορισμός διαδικασιών για την παρακολούθηση και τη διαχείριση των σχέσεων με τους προμηθευτές. Οι βασικές διαδικασίες περιλαμβάνουν: παραγγελία, λήψη, διαχείριση αποθέματος και εξουσιοδότηση πληρωμών προμηθευτή.
- Βιομηχανοποίηση (Manufacturing): Οργάνωση των δραστηριοτήτων που απαιτούνται για την αποδοχή πρώτων υλών, την κατασκευή του προϊόντος, τον ποιοτικό έλεγχο, τη συσκευασία για την αποστολή και το χρονοδιάγραμμα παράδοσης.
- Παράδοση και Logistics (Delivery and Logistics): Συντονισμός παραγγελιών πελατών, προγραμματισμός παραδόσεων, αποστολή φορτίων, τιμολόγηση πελατών και λήψη πληρωμών.
- Επιστροφές (Returning): Δημιουργία ενός δικτύου ή μια διαδικασία για να λαμβάνονται πίσω τα ελαττωματικά ή ανεπιθύμητα προϊόντα.

Σύμφωνα με την ιστοσελίδα CIO.com, προσδιορίζονται τρία σενάρια όπου η αποτελεσματική διαχείριση της εφοδιαστικής αλυσίδας αυξάνει την αξία στον κύκλο της εφοδιαστικής αλυσίδας:

1. Εντοπισμός πιθανών προβλημάτων. Όταν ένας πελάτης παραγγέλνει περισσότερα προϊόντα από αυτά που μπορεί να παραδώσει ο κατασκευαστής, ο αγοραστής μπορεί να παραπονεθεί για κακή εξυπηρέτηση. Μέσω της ανάλυσης δεδομένων, οι κατασκευαστές μπορούν να είναι σε θέση να προβλέψουν την έλλειψη προτού απογοητευτεί ο αγοραστής. Εδώ κατατάσσεται και το πρόβλημα του Backorder prediction.

2. Δυναμική βελτιστοποίηση της τιμής. Τα εποχιακά προϊόντα έχουν περιορισμένη διάρκεια ζωής. Στο τέλος της σεζόν, αυτά τα προϊόντα συνήθως απορρίπτονται ή πωλούνται με μεγάλες εκπτώσεις. Π.χ. αεροπορικές εταιρείες, ξενοδοχεία και άλλα με ευπαθή «προϊόντα» προσαρμόζουν συνήθως τις τιμές δυναμικά για να καλύψουν τη ζήτηση. Με τη χρήση αναλυτικού λογισμικού, παρόμοιες τεχνικές πρόβλεψης μπορούν να βελτιώσουν τα περιθώρια, ακόμη και για τα σκληρά αγαθά (αγαθά που δεν φθείρουν εύκολα με τον καιρό, π.χ. αμάξια).

3. Βελτίωση της κατανομής του «διαθέσιμου προς υπόσχεση» αποθέματος. Τα εργαλεία αναλυτικού λογισμικού βοηθούν στη δυναμική κατανομή πόρων και στον προγραμματισμό της εργασίας με βάση την πρόβλεψη πωλήσεων, τις πραγματικές παραγγελίες και την υποσχόμενη παράδοση των πρώτων υλών. Οι κατασκευαστές μπορούν να επιβεβαιώσουν μια ημερομηνία παράδοσης προϊόντος κατά την υποβολή της παραγγελίας — μειώνοντας σημαντικά τις παραγγελίες που έχουν συμπληρωθεί λανθασμένα.

Οι σύγχρονες αλυσίδες εφοδιασμού εκμεταλλεύονται τεράστιες ποσότητες δεδομένων που παράγονται από τη διαδικασία της αλυσίδας (chain process) και επιμελούνται από ειδικούς αναλυτές και επιστήμονες δεδομένων (data scientists). Οι μελλοντικοί ηγέτες της εφοδιαστικής αλυσίδας και τα συστήματα Enterprise Resource Planning (ERP) που διαχειρίζονται, πιθανότατα θα επικεντρωθούν στη βελτιστοποίηση της χρησιμότητας αυτών των δεδομένων — στην ανάλυσή τους σε πραγματικό χρόνο με ελάχιστο λανθάνοντα χρόνο (latency).

Δεδομένα Αλυσίδας Εφοδιασμού

Τα δεδομένα στο πλαίσιο των αλυσίδων εφοδιασμού μπορούν να κατηγοριοποιηθούν σε δεδομένα πελατών, αποστολής, παράδοσης, παραγγελίας, πώλησης, καταστήματος και προϊόντος. Ως εκ τούτου, τα δεδομένα αλυσίδας εφοδιασμού προέρχονται από διαφορετικές (και τμηματοποιημένες) πηγές όπως οι πωλήσεις, το απόθεμα, η κατασκευή, η αποθήκευση και η μεταφορά.
Ο ανταγωνισμός, οι αστάθειες των τιμών, η τεχνολογική ανάπτυξη και οι ποικίλες δεσμεύσεις των πελατών θα μπορούσαν να οδηγήσουν σε υποεκτίμηση ή υπερεκτίμηση της ζήτησης σε καθιερωμένες προβλέψεις.

Επομένως, για να αυξηθεί η ακρίβεια της πρόβλεψης ζήτησης, τα δεδομένα της εφοδιαστικής αλυσίδας πρέπει να αναλυθούν προσεκτικά για να ενισχυθεί η γνώση σχετικά με τις τάσεις της αγοράς, τη συμπεριφορά των πελατών, τους προμηθευτές και τις τεχνολογίες. Η εξαγωγή τάσεων και προτύπων από τέτοια δεδομένα και η χρήση τους για τη βελτίωση της ακρίβειας των μελλοντικών προβλέψεων μπορεί να βοηθήσει στην ελαχιστοποίηση του κόστους της εφοδιαστικής αλυσίδας.

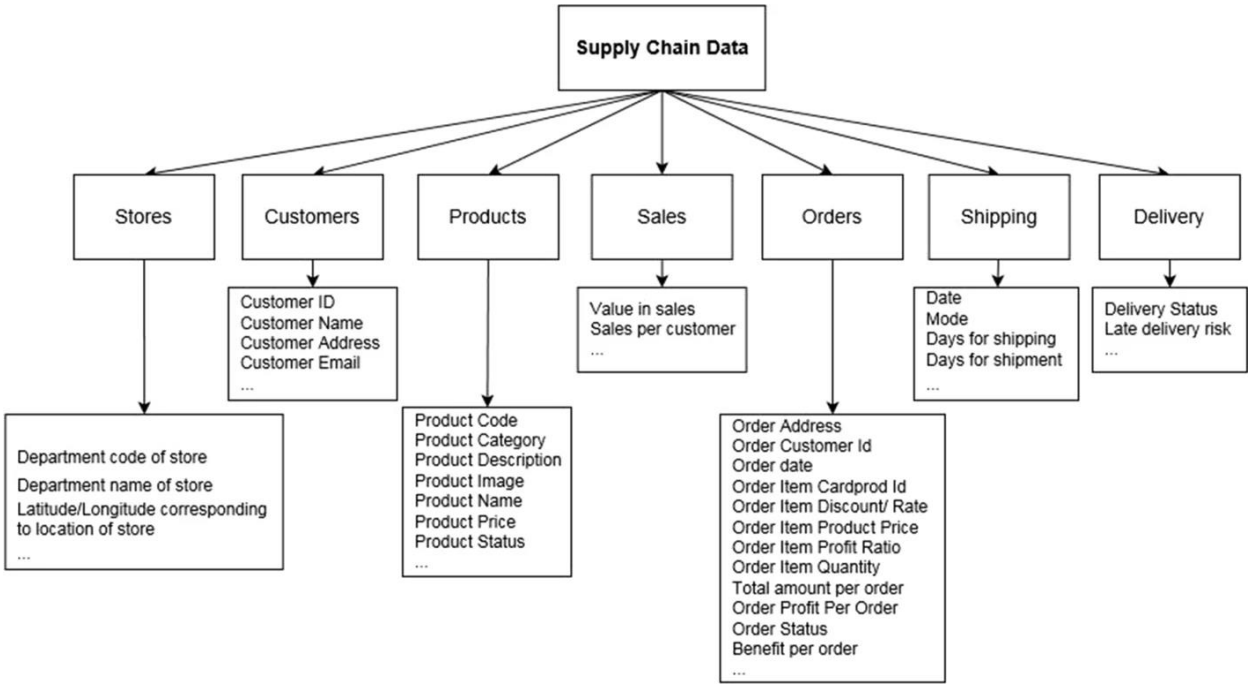


Figure 1: Ταξινόμηση δεδομένων εφοδιαστικής αλυσίδας

Πρόβλεψη Ζήτησης

Τα χαρακτηριστικά των δεδομένων ζήτησης στις σημερινές συνεχώς διευρυνόμενες και σποραδικές παγκόσμιες αλυσίδες εφοδιασμού καθιστούν την υιοθέτηση των προσεγγίσεων ανάλυσης μεγάλων δεδομένων (και μηχανικής μάθησης) απαραίτητη για την πρόβλεψη της ζήτησης.

Η ψηφιοποίηση των αλυσίδων εφοδιασμού και η ενσωμάτωση τεχνολογιών Blockchain για καλύτερη παρακολούθηση τους τονίζει περαιτέρω τον ρόλο της ανάλυσης μεγάλων δεδομένων. Τα δεδομένα μιας αλυσίδας εφοδιασμού είναι υψηλών διαστάσεων που παράγονται σε πολλά σημεία της αλυσίδας για διάφορους σκοπούς (προϊόντα, παραγγελίες, αποστολές, πελάτες, λιανοπωλητές κ.λπ.) σε μεγάλους όγκους λόγω της πληθώρας προμηθευτών, προϊόντων και πελατών και σε υψηλή ταχύτητα που αντικατοπτρίζεται από τις πολλές συναλλαγές που υποβάλλονται σε συνεχή επεξεργασία στα δίκτυα της εφοδιαστικής αλυσίδας.

Με την παρουσίαση τέτοιων πολυπλοκοτήτων, υπήρξε μια κλίση από τις συμβατικές (στατιστικές) προσεγγίσεις πρόβλεψης της ζήτησης που λειτουργούν με βάση τον εντοπισμό στατιστικών τάσεων (που χαρακτηρίζονται από χαρακτηριστικά μέσης τιμής και διακύμανσης) στα ιστορικά δεδομένα, προς έξυπνες προβλέψεις που μπορούν να μάθουν από τα ιστορικά δεδομένα και εξελίσσονται έξυπνα για να προσαρμόζονται και να μπορούν να προβλέψουν τη συνεχώς μεταβαλλόμενη ζήτηση στις αλυσίδες εφοδιασμού. Αυτή η ικανότητα καθιερώνεται χρησιμοποιώντας τεχνικές ανάλυσης μεγάλων δεδομένων που εξάγουν κανόνες πρόβλεψης μέσω της ανακάλυψης των υποκείμενων σχέσεων μεταξύ των δεδομένων ζήτησης στα δίκτυα της εφοδιαστικής αλυσίδας.

Διαχείριση ζήτησης στις αλυσίδες εφοδιασμού

Υπάρχουν δύο προσεγγίσεις για τη διαχείριση της ζήτησης. Η forward προσέγγιση που εξετάζει τη δυνητική ζήτηση για τα επόμενα αρκετά χρόνια και η backward προσέγγιση που βασίζεται σε προηγούμενες ή τωρινές ικανότητες για την ανταπόκριση στη ζήτηση.

Στη forward διαχείριση της ζήτησης, το επίκεντρο θα είναι η πρόβλεψη και ο σχεδιασμός της ζήτησης, η διαχείριση δεδομένων και οι στρατηγικές μάρκετινγκ. Η πρόβλεψη και ο προγραμματισμός της ζήτησης αναφέρονται στην πρόβλεψη των ποσοτήτων και των χρονισμών των αιτημάτων των πελατών. Τέτοιες προβλέψεις στοχεύουν στην επίτευξη της ικανοποίησης των πελατών καλύπτοντας τις ανάγκες τους έγκαιρα. Η ακριβής πρόβλεψη ζήτησης θα μπορούσε να βελτιώσει την αποτελεσματικότητα και την συνοχή των διαδικασιών παραγωγής (και των σχετικών αλυσίδων εφοδιασμού), καθώς οι πόροι θα αντιστοιχούν με τις απαιτήσεις που οδηγούν σε μείωση των αποθεμάτων.

Η διαχείριση της εφοδιαστικής αλυσίδας (Supply Chain Management ή SCM) εστιάζει στη ροή αγαθών, υπηρεσιών και πληροφοριών από τα σημεία προέλευσης στους πελάτες μέσω μιας αλυσίδας οντοτήτων και δραστηριοτήτων που συνδέονται μεταξύ τους. Σε τυπικά προβλήματα SCM, θεωρείται ότι η χωρητικότητα, η ζήτηση και το κόστος είναι γνωστές παράμετροι. Ωστόσο, αυτό δεν συμβαίνει στην πραγματικότητα, καθώς υπάρχουν αβεβαιότητες που προκύπτουν από διακυμάνσεις στη ζήτηση των πελατών, τη μεταφορά προμηθειών, τους οργανωτικούς κινδύνους και τους χρόνους παράδοσης. Οι αβεβαιότητες ζήτησης, ειδικότερα, έχουν τη μεγαλύτερη επιρροή στην απόδοση της εφοδιαστικής

αλυσίδας με εκτεταμένες επιπτώσεις στον προγραμματισμό της παραγωγής, τον προγραμματισμό αποθεμάτων και τη μεταφορά. Για αυτόν τον λόγο, η πρόβλεψη ζήτησης είναι μια βασική προσέγγιση για την αντιμετώπιση των αβεβαιοτήτων στις αλυσίδες εφοδιασμού. Πολλές φορές βέβαια δεν είναι αρκετή από μόνη της και για αυτό παρατηρούμε συχνά το φαινόμενο των αναδρομικών παραγγελιών.

Ποικιλία τεχνικών στατιστικής ανάλυσης έχει χρησιμοποιηθεί για την πρόβλεψη ζήτησης στο SCM, συμπεριλαμβανομένης της ανάλυσης χρονοσειρών και της ανάλυσης παλινδρόμησης. Με τις εξελίξεις στις τεχνολογίες της πληροφορίας και τη βελτιωμένη υπολογιστική απόδοση, η ανάλυση μεγάλων δεδομένων έχει αναδειχθεί ως μέσο για την επίτευξη ακριβέστερων προβλέψεων που αντικατοπτρίζουν καλύτερα τις ανάγκες των πελατών. Ακόμη, διευκολύνουν την αξιολόγηση της απόδοσης της εφοδιαστικής αλυσίδας, βελτιώνουν την αποτελεσματικότητα της, μειώνουν τον χρόνο αντίδρασης και υποστηρίζουν τρόπους αξιολόγησης κινδύνου σε αυτήν.

Το πρόβλημα της αναδρομικής παραγγελίας (Backorder)

Backorder είναι μια παραγγελία (ή μέρος μιας παραγγελίας) που περιμένει να εκπληρωθεί, συνήθως επειδή ο εν λόγω έμπορος δεν έχει αυτό το είδος αποθηκευμένο στην αποθήκη. Υποδηλώνει ότι η ζήτηση των πελατών για ένα προϊόν ή μια υπηρεσία υπερβαίνει την ικανότητα μιας εταιρείας να το παρέχει. Ανήκει στην Διαχείριση αποθεμάτων ενός supply chain. (Inventory Management)

Στο σύστημα της αλυσίδας εφοδιασμού, η παραγγελία υλικού είναι ένα κοινό πρόβλημα, που επηρεάζει το επίπεδο εξυπηρέτησης και την αποτελεσματικότητα του συστήματος απογραφής. Ο εντοπισμός αντικειμένων με τις μεγαλύτερες πιθανότητες έλλειψης πριν από την εμφάνισή της μπορεί να αποτελέσει μεγάλη ευκαιρία για τη βελτίωση της συνολικής απόδοσης μιας εταιρείας. Η αναδρομική παραγγελία προϊόντος μπορεί να είναι αποτέλεσμα ισχυρών επιδόσεων πωλήσεων (π.χ. το προϊόν έχει τόσο υψηλή ζήτηση που η παραγωγή δεν μπορεί να συμβαδίσει με τις πωλήσεις). Ωστόσο, τα backorders μπορεί να αναστατώσουν τους καταναλωτές, να οδηγήσουν σε ακυρώσεις παραγγελιών και μειωμένη αφοσίωση των πελατών. Οι εταιρείες θέλουν να τα αποφύγουν, αλλά και να αποφύγουν την υπερφόρτωση (overstocking) κάθε προϊόντος (που οδηγεί σε υψηλότερο κόστος αποθεμάτων). Σαφώς μια εταιρεία που έχει καταφέρει να «κτίσει» ένα πιστό αγοραστικό κοινό (brand loyalty), όπως για παράδειγμα η Apple, δεν διατρέχει ιδιαίτερο κίνδυνο να χάσει μεγάλο μέρος των πελατών της, ακόμα και σε μεγάλα επίπεδα ύπαρξης backorder. Αλλά κάτι τέτοιο δεν ισχύει για τις περισσότερες επιχειρήσεις.

Σύμφωνα με έρευνα της Oracle του 2021 στην Αμερική, Το 87 τοις εκατό των ερωτηθέντων ανθρώπων έχουν επηρεαστεί αρνητικά από ζητήματα εφοδιαστικής αλυσίδας κατά το περασμένο έτος, με πολλούς να μην μπορούν να αγοράσουν ορισμένα είδη λόγω ελλείψεων (60%), να αναγκάζονται να ακυρώσουν παραγγελίες λόγω καθυστερήσεων (51%) και ακόμη να οδηγούνται σε υπεραγορά βασικών προϊόντων από φόβο να μην ξεμείνουν (40%). Το 78% δήλωσε ότι είναι πιο πρόθυμοι να αγοράσουν από μια εταιρεία εάν ήξεραν ότι χρησιμοποιεί προηγμένες τεχνολογίες όπως η τεχνητή νοημοσύνη για τη διαχείριση της αλυσίδας εφοδιασμού της.

Η μηχανική εκμάθηση μπορεί να εντοπίσει μοτίβα που σχετίζονται με παραγγελίες προτού παραγγείλουν οι πελάτες. Στη συνέχεια, η παραγωγή μπορεί να προσαρμοστεί για να ελαχιστοποιήσει τις καθυστερήσεις, ενώ η εξυπηρέτηση πελατών μπορεί να παρέχει ακριβείς ημερομηνίες για να κρατά τους πελάτες

ενήμερους και χαρούμενους. Η προσέγγιση της προγνωστικής ανάλυσης επιτρέπει την μέγιστη δυνατή ποσότητα προϊόντος να φτάσει στα χέρια των πελατών με το χαμηλότερο κόστος για τον οργανισμό.

Machine Learning – Τεχνητή Νοημοσύνη (A.I.)

Τεχνητή Νοημοσύνη

Το πεδίο της τεχνητής νοημοσύνης (A.I.) αφορά τη μελέτη και κατασκευή συσκευών και μοντέλων που αντιλαμβάνονται το περιβάλλον τους και δρουν με τέτοιο τρόπο ώστε να μεγιστοποιήσουν τις πιθανότητες επίτευξης ενός συγκεκριμένου στόχου. Ενώ έχει εμφανιστεί αριθμός ορισμών της τεχνητής νοημοσύνης τις τελευταίες δεκαετίες, ο John McCarthy προσφέρει τον ακόλουθο ορισμό στο έγγραφο του «Τι είναι η τεχνητή νοημοσύνη», 2004: "Είναι η επιστήμη και η μηχανική κατασκευής έξυπνων μηχανών, ειδικά ευφυών προγραμμάτων υπολογιστών. Σχετίζεται με το παρόμοιο καθήκον της χρήσης υπολογιστών για την κατανόηση της ανθρώπινης νοημοσύνης, αλλά η τεχνητή νοημοσύνη δεν χρειάζεται να περιοριστεί σε μεθόδους που είναι βιολογικά παρατηρήσιμες».

Η έρευνα της τεχνητής νοημοσύνης προσπαθεί να επιτύχει έναν από τους τρεις ακόλουθους στόχους: Ισχυρή τεχνητή νοημοσύνη (Strong AI), εφαρμοσμένη τεχνητή νοημοσύνη (applied AI) ή γνωστική προσομοίωση (cognitive simulation).

Η ισχυρή τεχνητή νοημοσύνη στοχεύει να κατασκευάσει μηχανές που σκέφτονται. Η απόλυτη φιλοδοξία της είναι να παράγει μια μηχανή της οποίας η συνολική διανοητική ικανότητα δεν διακρίνεται από αυτή ενός ανθρώπου. Μερικοί κριτικοί αμφιβάλλουν εάν η έρευνα θα δημιουργήσει ακόμη και ένα σύστημα με τη συνολική ευφυΐα ενός μυρμηγκιού στο άμεσο μέλλον. Πράγματι, ορισμένοι ερευνητές που εργάζονται στους άλλους δύο κλάδους της τεχνητής νοημοσύνης θεωρούν ότι η ισχυρή τεχνητή νοημοσύνη δεν αξίζει να συνεχιστεί σε βάρος των άλλων δύο.

Η εφαρμοσμένη τεχνητή νοημοσύνη, γνωστή και ως προηγμένη επεξεργασία πληροφοριών, στοχεύει στην παραγωγή εμπορικά βιώσιμων «έξυπνων» συστημάτων – για παράδειγμα, «ειδικών» συστημάτων ιατρικής διάγνωσης και συστημάτων ανταλλαγής μετοχών. Μέχρι σήμερα έχει γνωρίσει σημαντική επιτυχία.

Στη γνωστική προσομοίωση, οι υπολογιστές χρησιμοποιούνται για τη δοκιμή θεωριών σχετικά με το πώς λειτουργεί το ανθρώπινο μυαλό - για παράδειγμα, θεωρίες σχετικά με το πώς οι άνθρωποι αναγνωρίζουν πρόσωπα ή ανακαλούν αναμνήσεις. Η γνωστική προσομοίωση είναι ήδη ένα ισχυρό εργαλείο τόσο στη νευροεπιστήμη όσο και στη γνωστική ψυχολογία.

Στις μέρες μας ο κλάδος της Τεχνητής Νοημοσύνης βρίσκεται σε μια φάση έντονης ανάπτυξης. Οι κύριοι λόγοι είναι ο τεράστιος όγκος δεδομένων που είναι πλέον διαθέσιμος σε κάθε τομέα και η εκθετική αύξηση της διαθέσιμης υπολογιστικής ισχύος. Χαρακτηριστικό παράδειγμα είναι η πρόοδος στον τομέα των νευρωνικών δικτυών.

Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning) είναι υποκατηγορία της Τεχνητής Νοημοσύνης. Με τον όρο Μηχανική Μάθηση εννοούμε το επιστημονικό πεδίο που ασχολείται με τους αλγόριθμους και τα στατιστικά μοντέλα που τα υπολογιστικά συστήματα χρησιμοποιούν για να εκτελέσουν μια εργασία ή και να εξαγάγουν συμπεράσματα εκμεταλλευόμενα μοτίβα μέσα στα δεδομένα. Ήδη από το έτος 1943 οι Warren McCulloch και Walter Pitts δημοσιεύουν το άρθρο "A logical calculus of the ideas immanent in nervous activity", στο οποίο δίνεται μια μαθηματική περιγραφή της λειτουργίας των νευρώνων.

Η Εκμάθηση Μηχανών (machine learning), η Τεχνητή Νοημοσύνη (artificial intelligence) και η Ρομποτική (robotics) είναι έννοιες στενά συνδεδεμένες. Τα τελευταία χρόνια έχει σημειωθεί σημαντική πρόοδος σε αυτούς τους τομείς οι οποίοι αναμένεται να επηρεάσουν τόσο την καθημερινή ζωή όσο και την εργασία. Η συνεισφορά των τεχνολογιών αυτών στην ανάδειξη νέων επιχειρηματικών μοντέλων και στην παραγωγική διαδικασία είναι αναμφίβολα μεγάλη.

Η χρήση Machine Learning και Artificial Intelligence έχει αποκτήσει μια πληθώρα χρήσεων στις μέρες μας. Μερικές από αυτές είναι για σκοπούς προσωποποιημένου marketing, για τον εντοπισμό απάτης (fraud detection), για φωνητική βοήθεια (voice assistant), αυτοοδηγούμενα οχήματα, βελτιστοποίηση μεταφορών, αυτοματοποίηση / βελτιστοποίηση διαδικασιών, φροντίδα υγείας και chatbots.

Μερικές σημαντικές διεργασίες που ανήκουν στον τομέα της βελτιστοποίησης διαδικασιών και του αυτοματισμού:

- Η πρόβλεψη για αυξημένες ροές χρηματοδότησης της επιχείρησης
- Η εκτίμηση ζήτησης ή και του φόρτου εργασίας με στόχο την βέλτιστη κάλυψή τους
- Ένας οργανισμός μπορεί να προβλέψει πότε θα χρειαστεί αυξημένους υπολογιστικούς πόρους και να τους μισθώσει από το cloud
- Η ανάπτυξη εφαρμογών για την αυτοματοποίηση επαναλαμβανόμενων υπηρεσιών.

Μέθοδοι Μηχανικής Μάθησης

Οι μέθοδοι χωρίζονται σε αυτούς της μάθησης με επίβλεψη (supervised learning), της μάθησης χωρίς επίβλεψη (unsupervised learning) και της ενισχυτικής μάθησης (reinforcement learning).
Στην πρώτη περίπτωση έχουμε ένα σύνολο δεδομένων όπου η τιμή της μεταβλητής που μας ενδιαφέρει είναι γνωστή. Η γνώση αυτή μπορεί να χρησιμοποιηθεί για την κατασκευή ενός μοντέλου.
Στη δεύτερη περίπτωση δεν έχουμε στα δεδομένα μας τις τιμές της μεταβλητής για την οποία ενδιαφερόμαστε.
Στην τρίτη περίπτωση ένα πρόγραμμα υπολογιστή αλληλεπιδρά με ένα δυναμικό περιβάλλον στο οποίο πρέπει να εκτελέσει έναν συγκεκριμένο στόχο (όπως η οδήγηση ενός οχήματος ή το παιχνίδι ενός παιχνιδιού εναντίον ενός αντιπάλου). Καθώς περιηγείται στον χώρο προβλημάτων του, το πρόγραμμα λαμβάνει ανατροφοδότηση (feedback) ανάλογη με τις ανταμοιβές, τις οποίες προσπαθεί να μεγιστοποιήσει.

Supervised Learning

Οι τύποι αλγορίθμων εποπτευόμενης μάθησης περιλαμβάνουν την ενεργητική μάθηση, την ταξινόμηση (classification) και την παλινδρόμηση (regression). Με τον όρο μοντέλα ταξινόμησης ή παλινδρόμησης εννοούμε μαθηματικά μοντέλα στόχος των οποίων είναι να ταξινομήσουν μια παρατήρηση σε μια από τις διαθέσιμες κατηγορίες. Η δημιουργία τέτοιων μοντέλων γίνεται με τη χρήση ενός αρχικού συνόλου παρατηρήσεων που είναι ήδη ταξινομημένες σε κατηγορίες. Το σύνολο αυτό λέγεται σύνολο εκπαίδευσης (training set).

Οι αλγόριθμοι ταξινόμησης χρησιμοποιούνται όταν οι έξοδοι περιορίζονται σε ένα περιορισμένο σύνολο τιμών. Το πρόβλημα που θα μελετήσουμε ανήκει στην κατηγορία αυτή, καθώς θα κατατάξουμε τα δεδομένά μας σε 2 τάξεις. Είναι δηλαδή ένα binary classification πρόβλημα. Κάποιες από τις βασικές μεθόδους αυτής της κατηγορίας είναι η λογιστική παλινδρόμηση (logistic regression), τα δένδρα αποφάσεων (decision trees), ο Naïve Bayes και τα Random Forests.

Οι αλγόριθμοι παλινδρόμησης χρησιμοποιούνται όταν οι έξοδοι μπορεί να έχουν οποιαδήποτε αριθμητική τιμή εντός μιας περιοχής. Κάποιοι από τους βασικούς αλγορίθμους είναι η γραμμική παλινδρόμηση (linear regression), η πολυωνυμική παλινδρόμηση (Polynomial Regression), τα δένδρα αποφάσεων και τα Random Forests.

Τα αρχικά διαθέσιμα δεδομένα για την ανάπτυξη ενός μοντέλου ταξινόμησης χωρίζονται σε δύο αντιπροσωπευτικά υποσύνολα. Το ένα χρησιμοποιείται για τη δημιουργία του μοντέλου ταξινόμησης, ενώ το άλλο για την αξιολόγησή του. Το πρώτο καλείται σύνολο εκπαίδευσης (training set), ενώ το δεύτερο σύνολο ελέγχου (test set). Το train set είναι το μεγαλύτερο, αποτελούμενο συνήθως από το 70% με 80% του αρχικού συνόλου.

Unsupervised Learning

Τα μοντέλα μάθησης χωρίς επίβλεψη χρησιμοποιούνται για τρεις κύριες εργασίες: ομαδοποίηση (clustering), συσχέτιση (association) και μείωση διαστάσεων (dimensionality reduction).

Η ομαδοποίηση/συσταδοποίηση είναι μια τεχνική εξόρυξης δεδομένων (data mining) που ομαδοποιεί δεδομένα χωρίς ετικέτα (label) με βάση τις ομοιότητες ή τις διαφορές τους. Οι αλγόριθμοι ομαδοποίησης χρησιμοποιούνται για την επεξεργασία ακατέργαστων, μη ταξινομημένων δεδομένων σε ομάδες που αντιπροσωπεύονται από δομές ή μοτίβα στις πληροφορίες. Οι αλγόριθμοι ομαδοποίησης μπορούν να κατηγοριοποιηθούν σε μερικούς τύπους, συγκεκριμένα σε exclusive, overlapping, hierarchical και probabilistic. Μερικοί από τους πιο γνωστούς είναι ο K-Means (centroid-based), ο DBSCAN (density-based) και ο Agglomerative (hierarchical).

Ένας κανόνας συσχέτισης είναι μια μέθοδος βασισμένη σε κανόνες για την εύρεση σχέσεων μεταξύ μεταβλητών σε ένα δεδομένο σύνολο δεδομένων. Αυτές οι μέθοδοι χρησιμοποιούνται συχνά για ανάλυση της αγοράς, επιτρέποντας στις εταιρείες να κατανοήσουν καλύτερα τις σχέσεις μεταξύ διαφορετικών προϊόντων.

Ενώ περισσότερα δεδομένα αποφέρουν γενικά πιο ακριβή αποτελέσματα, μπορούν επίσης να επηρεάσουν την απόδοση των αλγορίθμων μηχανικής εκμάθησης (π.χ. overfitting) και μπορούν επίσης να δυσκολέψουν την οπτικοποίηση των συνόλων δεδομένων. Η μείωση διαστάσεων είναι μια τεχνική που χρησιμοποιείται όταν ο αριθμός των χαρακτηριστικών ή διαστάσεων σε ένα σύνολο δεδομένων είναι πολύ υψηλός. Μειώνει τον αριθμό των εισροών δεδομένων σε ένα διαχειρίσιμο μέγεθος, ενώ παράλληλα διατηρεί την ακεραιότητα του συνόλου δεδομένων όσο το δυνατόν περισσότερο. Χρησιμοποιείται συνήθως στο στάδιο προεπεξεργασίας δεδομένων και υπάρχουν μερικοί μέθοδοι που μπορούν να χρησιμοποιηθούν, όπως: Principal Component Analysis (PCA), Autoencoders κ.α.

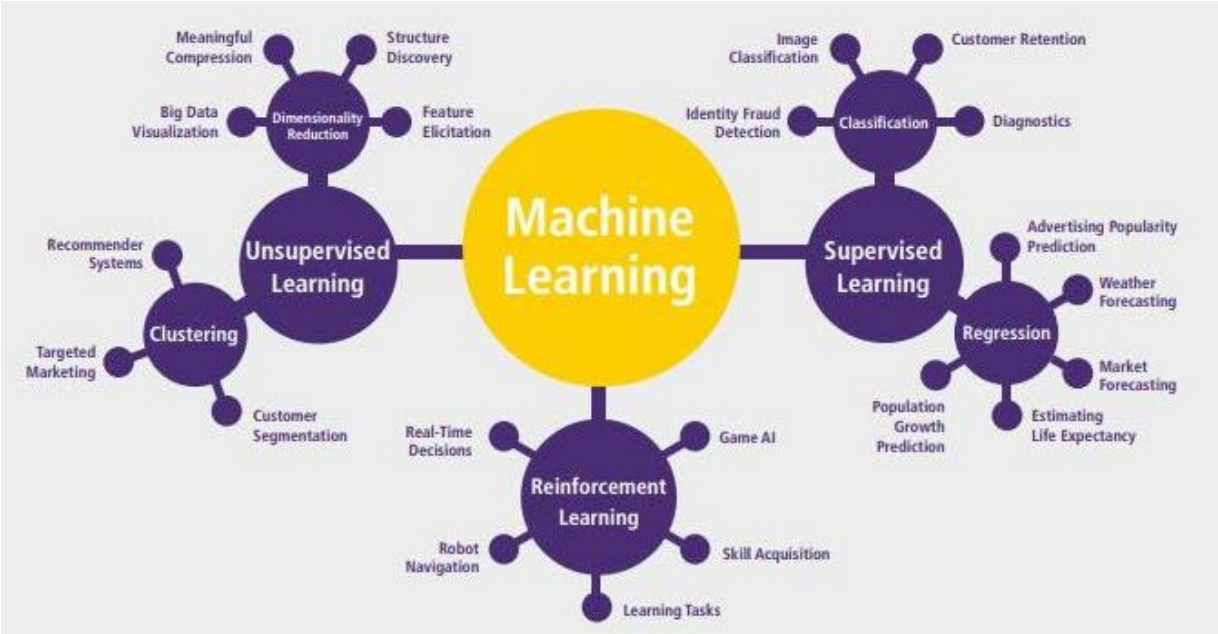


Figure 2: Ταξινόμηση μεθοδολογιών Μηχανικής Μάθησης

Βαθιά Μάθηση (Deep Learning)

Το deep learning είναι μέρος μιας ευρύτερης οικογένειας μεθόδων μηχανικής μάθησης που βασίζονται σε τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks) με μάθηση αναπαράστασης. Αποτελεί στην ουσία μία εξειδικευμένη μορφή μηχανικής μάθησης εκτελώντας "εκμάθηση από άκρο σε άκρο" - όπου δίνονται σε ένα δίκτυο ακατέργαστα δεδομένα και μια εργασία για εκτέλεση, όπως η ταξινόμηση, και μαθαίνει πώς να το κάνει αυτόματα.

Μια βασική διαφορά μεταξύ machine learning και deep learning είναι ότι η απόδοση αλγόριθμων βαθιάς μάθησης αυξάνεται ανάλογα με το πλήθος των δεδομένων. Το ίδιο δεν ισχύει απαραίτητα για την ρηχή μάθηση, η οποία αναφέρεται σε μεθόδους μηχανικής εκμάθησης που φτάνουν στο μέγιστο επίπεδο απόδοσής τους σε κάποιο σημείο, ανεξάρτητος εάν προθέτονται περισσότερα παραδείγματα και δεδομένα εκπαίδευσης στο νευρωνικό δίκτυο.

Επιπρόσθετα, το deep learning απαιτεί μεγάλη υπολογιστή ισχύ για να είναι δυνατή η διαχείριση μεγάλων όγκων δεδομένων, ειδικά σε περίπτωση εικόνων, και η εκπαίδευση πολύπλοκων μοντέλων. Ιδιαίτερο ρόλο παίζει η κάρτα γραφικών του υπολογιστή (GPU) καθώς το μεγαλύτερο βάρος της εργασίας αναλογεί σε αυτήν. Συγκεκριμένα, είναι επιθυμητό μεγάλο πλήθος GPU με αρκετή μνήμη (VRAM).

Data Mining

Η εξόρυξη δεδομένων είναι η διαδικασία ανάλυσης τεράστιων ποσοτήτων πληροφοριών και συνόλων δεδομένων, εξαγωγής (ή «εξόρυξης») χρήσιμης νοημοσύνης για να βοηθήσει τους οργανισμούς να λύσουν προβλήματα, να προβλέψουν τάσεις, να μειώσουν τους κινδύνους και να βρουν νέες ευκαιρίες.
Η εξόρυξη δεδομένων περιλαμβάνει επίσης τη δημιουργία σχέσεων και την εύρεση προτύπων, ανωμαλιών και συσχετισμών για την αντιμετώπιση προβλημάτων, δημιουργώντας πληροφορίες που μπορούν να χρησιμοποιηθούν κατά την διάρκεια της διαδικασίας. Εφαρμόζεται σε πάρα πολλούς κλάδους όπως το λιανεμπόριο, το marketing, η ιατρική περίθαλψη, ο τραπεζικός, κ.α.

Η Cross-Industry Standard διαδικασία για εξόρυξη δεδομένων (CRISP-DM) δημοσιεύτηκε το 1999 για την τυποποίηση των διαδικασιών εξόρυξης δεδομένων σε όλες τις βιομηχανίες και από τότε έχει γίνει η πιο κοινή μεθοδολογία για εξόρυξη και ανάλυση δεδομένων και έργα επιστήμης δεδομένων. Αποτελεί ένα αξιόπιστο μοντέλο εξόρυξης δεδομένων που αποτελείται από έξι φάσεις. Είναι μια κυκλική διαδικασία που παρέχει μια δομημένη προσέγγιση στη διαδικασία εξόρυξης. Οι έξι φάσεις μπορούν να εφαρμοστούν με οποιαδήποτε σειρά, αλλά αυτό μερικές φορές θα απαιτούσε την επιστροφή στα προηγούμενα βήματα και την επανάληψη των ενεργειών.

Οι έξι φάσεις του CRISP-DM περιλαμβάνουν:

1) Επιχειρηματική κατανόηση: Σε αυτό το βήμα, τίθενται οι στόχοι των επιχειρήσεων και ανακαλύπτονται οι σημαντικοί παράγοντες που θα βοηθήσουν στην επίτευξη του στόχου.

2) Κατανόηση δεδομένων: Εδώ συλλέγονται όλα τα δεδομένα και συμπληρώνονται στο εργαλείο (εάν χρησιμοποιείται οποιοδήποτε εργαλείο). Τα δεδομένα παρατίθενται με την πηγή τους, την τοποθεσία, τον τρόπο απόκτησής τους και εάν αντιμετωπίστηκε κάποιο πρόβλημα. Τα δεδομένα οπτικοποιούνται και εξετάζονται για να ελεγχθεί η πληρότητά τους.

3) Προετοιμασία δεδομένων: Αυτό το βήμα περιλαμβάνει την επιλογή των κατάλληλων δεδομένων, τον καθαρισμό τους, την κατασκευή χαρακτηριστικών από αυτά και την ενοποίηση τους από πολλές βάσεις δεδομένων.

4) Μοντελοποίηση: Η επιλογή της τεχνικής/μοντέλου εξόρυξης δεδομένων, όπως π.χ. το δέντρο αποφάσεων, η δημιουργία ελέγχου για την αξιολόγηση του επιλεγμένου μοντέλου και η κατασκευή μοντέλων από το σύνολο δεδομένων.

5) Αξιολόγηση: Αυτό το βήμα θα καθορίσει τον βαθμό στον οποίο το μοντέλο που προκύπτει πληροί τις επιχειρηματικές απαιτήσεις. Η αξιολόγηση μπορεί να γίνει δοκιμάζοντας το μοντέλο σε πραγματικές εφαρμογές. Το μοντέλο ελέγχεται για τυχόν λάθη ή βήματα που πρέπει να επαναληφθούν.

6) Ανάπτυξη: Σε αυτό το βήμα γίνεται ένα σχέδιο ανάπτυξης και διαμορφώνεται στρατηγική για την παρακολούθηση και τη διατήρηση των αποτελεσμάτων του μοντέλου, για τον έλεγχο της χρησιμότητάς του. Ακόμη, γίνονται τελικές αναφορές και επανεξέταση της όλης διαδικασίας για να ελεγχθεί οποιοδήποτε λάθος και εάν επαναλαμβάνεται κάποιο βήμα.

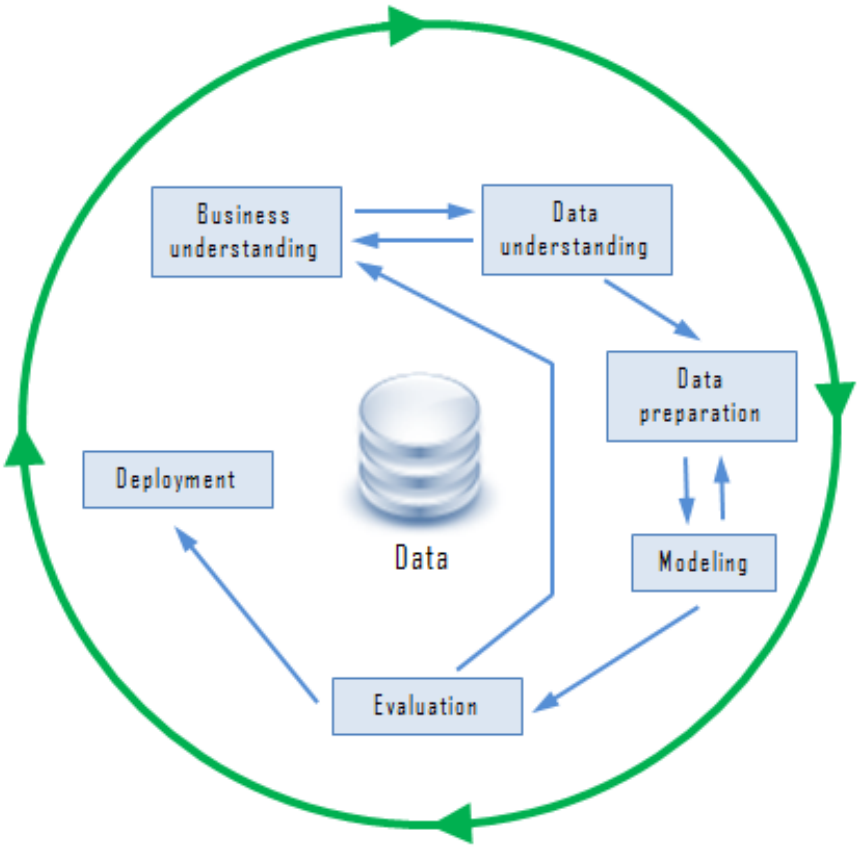


Figure 3: Διαδικασία Εξόρυξης Δεδομένων (CRISP-DM)

Χρονοσειρές

Οι χρονοσειρές είναι μεθοδολογίες για την εξόρυξη πολύπλοκων και διαδοχικών τύπων δεδομένων. Στα δεδομένα χρονοσειρών, ακολουθίας, δεδομένα που αποτελούνται από μεγάλες ακολουθίες αριθμητικών δεδομένων, που καταγράφονται σε ίσα χρονικά διαστήματα (π.χ. ανά λεπτό, ανά ώρα ή ανά ημέρα). Πολλές φυσικές και ανθρωπογενείς διεργασίες, όπως τα χρηματιστήρια, η ιατρική διάγνωση ή φυσικά φαινόμενα, μπορούν να δημιουργήσουν δεδομένα χρονοσειρών.

Χρήση Μηχανικής Μάθησης σε Αλυσίδες Εφοδιασμού

Οι περιπτώσεις χρήσης μηχανικής μάθησης στην αλυσίδα εφοδιασμού βοηθούν τους λιανοπωλητές, τους προμηθευτές και τους διανομείς να οδηγήσουν σε αλλαγές μετασχηματισμού που είναι απαραίτητες σήμερα, ιδιαίτερα μετά την πανδημία του covid. Η Μηχανική Μάθηση προσφέρει άνευ προηγουμένου αξία στις λειτουργίες της εφοδιαστικής αλυσίδας: από εξοικονόμηση κόστους μέσω μειωμένων λειτουργικών γενικών εξόδων και μετριασμού κινδύνου, έως βελτιωμένες προβλέψεις εφοδιαστικής αλυσίδας, γρήγορες παραδόσεις και βελτιωμένη εξυπηρέτηση πελατών. Σημαντικότερα οφέλη της θα είναι η παροχή στους επαγγελματίες της εφοδιαστικής αλυσίδας των πιο σημαντικών γνώσεων για το πώς μπορεί να βελτιωθεί η απόδοση της εφοδιαστικής αλυσίδας, προβλέποντας ανωμαλίες στο κόστος των logistics και στην απόδοση προτού εμφανιστούν. Η μηχανική μάθηση παρέχει επίσης πληροφορίες για το πού μπορεί ο αυτοματισμός να προσφέρει τα πιο σημαντικά πλεονεκτήματα κλίμακας.

Πιο συγκεκριμένα μπορεί να βελτιώσει σε πολύ μεγάλο βαθμό τις παρακάτω διεργασίες:

- Ασφάλεια: Οι αλγόριθμοι μηχανικής μάθησης μπορούν να αναλύσουν τεράστιες ποσότητες δεδομένων και να σχεδιάσουν μοτίβα για κάθε επιχείρηση για να την προστατεύσουν από απάτες όπως κατάχρηση διαπιστευτηρίων (credentials), επιτάχυνση των ερευνών απάτης και αυτοματοποίηση των διαδικασιών κατά της απάτης.
- Επιχειρηματικότητα: Από επιχειρηματική σκοπιά, η μηχανική μάθηση παρέχει πολύτιμες πληροφορίες που απλοποιούν και επιταχύνουν τη λήψη αποφάσεων. Δίνει τη δυνατότητα στα ανώτερα στελέχη να αξιολογούν γρήγορα τα καλύτερα και τα χειρότερα δυνατά σενάρια.
- Εξυπηρέτηση Πελατών: Μέσω της χρήσης chatbot τα οποία έχουν εκπαιδευτεί για να κατανοούν συγκεκριμένες λέξεις-κλειδιά και φράσεις που ενεργοποιούν την απάντησή τους. Χρησιμοποιούνται ευρέως στη διαχείριση σχέσεων προμηθευτών, στις πωλήσεις και στη διαχείριση προμηθειών. Ακόμη, αξιοποιούνται τεχνικές όπως: προσδιορισμός δεμάτων με κίνδυνο να παρουσιάσουν πρόβλημα και προτάσεις για μέτρα μετριασμού του κινδύνου αυτού, αυτοματοποίηση της ροής ειδοποιήσεων ανάλογα με τις προηγούμενες αλληλεπιδράσεις των καταναλωτών και καθορισμός της σωστής στιγμής επικοινωνίας με τους καταναλωτές για μέγιστη αφοσίωση.
- Παραγωγή: Είναι δυνατό να εντοπιστούν ζητήματα ποιότητας στη γραμμή παραγωγής στα αρχικά στάδια. Για παράδειγμα, με τη βοήθεια του computer vision, οι κατασκευαστές μπορούν να ελέγξουν εάν η τελική εμφάνιση των προϊόντων αντιστοιχεί στο απαιτούμενο επίπεδο ποιότητας. Εάν τα προϊόντα έχουν κάποια ελαττώματα, είναι εύκολο να εντοπιστούν πριν φτάσουν στους πελάτες. Επίσης, η μηχανική μάθηση συμβάλλει στη μείωση του αριθμού των περιπτώσεων που δεν βρέθηκαν σφάλματα (NFF). Ως NFF χαρακτηρίζεται μια μονάδα που αφαιρείται από τη λειτουργία μετά από παράπονο για αντιληπτή βλάβη του εξοπλισμού. Εάν δεν εντοπιστεί καμία ανωμαλία, η μονάδα επιστρέφεται σε λειτουργία χωρίς να πραγματοποιηθεί επισκευή.
- Logistics & μεταφορές: Βοηθά να κατανοήσουμε πού βρίσκεται ένα πακέτο σε ολόκληρο τον κύκλο των logistics. Επιτρέπει στους επαγγελματίες της εφοδιαστικής αλυσίδας να παρακολουθούν τη θέση των αγαθών κατά τη μεταφορά. Επίσης, παρέχει ορατότητα στις συνθήκες υπό τις οποίες μεταφέρεται το δέμα. Με τη βοήθεια αισθητήρων, οι λιανοπωλητές μπορούν να παρακολουθούν παραμέτρους όπως υγρασία, θερμοκρασία κ.λπ. Επιπλέον, βοηθά στη βελτιστοποίηση της διαδρομής μεταφοράς σε πραγματικό χρόνο. Παρακολουθεί τις καιρικές συνθήκες και τις συνθήκες του δρόμου και παρέχει συστάσεις για τη βελτιστοποίηση της διαδρομής και τη μείωση του χρόνου οδήγησης.

- Διαχείριση αποθήκης: Στις αποθήκες, η μηχανική εκμάθηση χρησιμοποιείται για την αυτοματοποίηση της χειρωνακτικής εργασίας, την πρόβλεψη πιθανών ζητημάτων και τη μείωση της γραφειοκρατίας για το προσωπικό της αποθήκης. Για παράδειγμα, στον αυτόματο εντοπισμό της άφιξη των δεμάτων και στην αλλαγή των καταστάσεων παράδοσης. Οι κάμερες σαρώνουν γραμμωτούς κώδικες και ετικέτες στη συσκευασία και όλες οι απαραίτητες πληροφορίες πηγαίνουν απευθείας στο σύστημα. Επίσης, βοηθά στον προγραμματισμό αυτόνομων οχημάτων και ρομπότ που χρησιμοποιούνται ευρέως σε αποθήκες (π.χ. Amazon). Με τη βοήθεια οδηγών που είναι ενσωματωμένοι στο σύστημα, αυτόνομα οχήματα και ρομπότ βοηθούν στη λήψη, τη συσκευασία/αποσυσκευασία και τη μεταφορά. Το computer vision σε αυτήν την περίπτωση βοηθά στην εύρεση μιας ελεύθερης θέσης για ένα κουτί, στον έλεγχο της σωστής τοποθέτησης και στην αποφυγή σύγκρουσης ρομπότ και οχημάτων στις αποθήκες.
- Διαχείριση αποθεμάτων: Η αποθήκευση και η διατήρηση του αποθέματος σε καλή κατάσταση είναι δαπανηρή. Επομένως, οι επαγγελματίες της εφοδιαστικής αλυσίδας θα πρέπει να προσεγγίσουν τον προγραμματισμό αποθεμάτων πολύ διεξοδικά, καθώς έχει άμεσο αντίκτυπο στις ταμειακές ροές και στα περιθώρια κέρδους μιας εταιρείας. Η διαχείριση αποθέματος είναι μια από τις πιο τυπικές περιπτώσεις χρήσης μηχανικής εκμάθησης στην αλυσίδα εφοδιασμού. Μπορεί να βοηθήσει στην επίλυση του προβλήματος της έλλειψης ή υπερβολικού εφοδιασμού (over-under stocking). Με βάση τα δεδομένα που μπορούν να ληφθούν από πολλούς τομείς όπως το περιβάλλον της αγοράς, οι εποχιακές τάσεις, οι προσφορές, οι πωλήσεις και η ιστορική ανάλυση, είναι δυνατή η πρόβλεψη της αύξηση της ζήτησης. Ένα άλλο παράδειγμα είναι η περίπτωση του computer vision στη διαχείριση αποθεμάτων. Πρώτον, εφαρμόζεται για την καταμέτρηση και την ταξινόμηση των αντικειμένων που φτάνουν. Επίσης, βοηθά στην ανίχνευση οπτικών βλαβών της συσκευασίας. Με τη βοήθεια του, το λογισμικό είναι επίσης σε θέση να ταξινομήσει αντικείμενα που «βλέπει». Για παράδειγμα, ρομπότ εξοπλισμένα με κάμερες θα επιθεωρήσουν τις αποθήκες και θα δημιουργήσουν αυτόματα μια εικόνα του αποθέματος σε πραγματικό χρόνο.

Χρήσιμα Μοντέλα

Λογιστική παλινδρόμηση

Η μέθοδος της λογιστικής παλινδρόμησης είναι μια μέθοδος ταξινόμησης/κατηγοριοποίησης. Στην απλή της μορφή χρησιμεύει για την ταξινόμηση σε δύο κατηγορίες. Η λογιστική παλινδρόμηση κατασκευάζει ένα μοντέλο το οποίο εκτιμά την πιθανότητα να ανήκει μια περίπτωση σε μια κατηγορία.

Η λογιστική παλινδρόμηση υπολογίζει τους συντελεστές $\beta_0, \beta_1, \ldots, \beta_n$ στην εξίσωση:

$$p = \frac{1}{1 + e^{[-(\beta]_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Όπου p είναι η πιθανότητα η περίπτωση με τιμές τις εξαρτημένες μεταβλητές $x_1, \ldots, x_n$ να ανήκει στη μια από τις δύο κατηγορίες.

Δένδρα αποφάσεων

Τα δέντρα απόφασης μπορούν να χρησιμοποιηθούν για προβλήματα ταξινόμησης καθώς και παλινδρόμησης. Το ίδιο το όνομα υποδηλώνει ότι χρησιμοποιούν ένα διάγραμμα ροής σαν δομή δέντρου για να δείξουν τις προβλέψεις που προκύπτουν από μια σειρά διαχωρισμών που βασίζονται σε χαρακτηριστικά. Ξεκινούν με έναν ριζικό κόμβο και τελειώνουν με μια απόφαση που λαμβάνεται από τα φύλλα.

Σε ένα δέντρο απόφασης, κάθε εσωτερικός κόμβος αντιπροσωπεύει μια «δοκιμή» σε ένα χαρακτηριστικό (π.χ. εάν ένα κέρμα φέρει κορώνα ή γράμματα), κάθε κλάδος αντιπροσωπεύει το αποτέλεσμα της δοκιμής και κάθε κόμβος φύλλου αντιπροσωπεύει μια ετικέτα κλάσης (η απόφαση που ελήφθη μετά τον υπολογισμό όλων των χαρακτηριστικών). Ένας κόμβος που δεν έχει «παιδιά» είναι ένα φύλλο (leaf).

Σε κάθε βήμα της διαδικασίας μια μεταβλητή του συνόλου δεδομένων επιλέγεται και το σύνολο χωρίζεται σε δύο (συνήθως) μέρη με βάση αυτήν τη μεταβλητή. Στη συνέχεια, για κάθε μέρος επαναλαμβάνεται η ίδια διαδικασία. Αυτό συνεχίζεται μέχρι να ικανοποιηθεί κάποια συνθήκη. Αρκετές φορές υπάρχει και ένα στάδιο «κλαδέματος» όπου μπορεί κάποιες διακλαδώσεις/σπασίματα να κοπούν από το μοντέλο. Χρησιμοποιείται μια συνάρτηση που εκτιμά πόσο καλά κάθε κόμβος του δένδρου περιέχει περιπτώσεις που ανήκουν σε μια μόνο κατηγορία. Για κάθε σπάσιμο υπολογίζεται ο σταθμισμένος μέσος όρος της καθαρότητας των κόμβων που προκύπτουν. Το σπάσιμο που έχει τον καλύτερο μέσο όρο καθαρότητας είναι αυτό που επιλέγεται.

Η τιμή της σε κάθε κόμβο υπολογίζεται από τον τύπο: μετρική Gini Impurity

$$\sum_{i=1}^{n} p_i(1 - p_i) = 1 - \sum_{i=1}^{n} p_i^2$$

όπου Pi είναι η αναλογία των στοιχείων του κόμβου που ανήκουν στην κλάση i. Μικρότερη τιμή του συντελεστή δείχνει μεγαλύτερη καθαρότητα.

Random Forests

Είναι ένας από τους πιο χρησιμοποιούμενους αλγόριθμους, λόγω της απλότητας και της ποικιλομορφίας του (μπορεί να χρησιμοποιηθεί τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης).

Η Random Forest μέθοδος δημιουργεί πολλαπλά δέντρα απόφασης από δείγματα του training set και τα συγχωνεύει για να αποκτήσει μια πιο ακριβή και σταθερή πρόβλεψη. Τα δένδρα αυτά συνήθως εκπαιδεύονται με τη μέθοδο «bagging». Η γενική ιδέα της μεθόδου bagging είναι ότι ένας συνδυασμός μοντέλων εκμάθησης αυξάνει το συνολικό αποτέλεσμα.

Η Random Forest προσθέτει επιπλέον τυχαιότητα στο μοντέλο, ενώ μεγαλώνει τα δέντρα. Αντί να αναζητά το πιο σημαντικό χαρακτηριστικό κατά τη διάσπαση ενός κόμβου, αναζητά το καλύτερο χαρακτηριστικό ανάμεσα σε ένα τυχαίο υποσύνολο χαρακτηριστικών. Αυτό έχει ως αποτέλεσμα μια μεγάλη ποικιλία που γενικά οδηγεί σε ένα καλύτερο μοντέλο.

Ένα από τα βασικά πλεονεκτήματα της χρήσης των Random Forests είναι πως είναι λιγότερο πιθανό να εμφανίσουν υπερβολικά μεγάλη προσαρμογή στο training set που θα επηρέαζε αρνητικά τη γενικότερη απόδοσή τους (overfitting). Ένα δεύτερο πλεονέκτημά τους είναι ότι μπορούν να χρησιμοποιηθούν για να εκτιμηθεί πόσο σημαντική είναι κάθε μεταβλητή για την εκτίμηση που μας ενδιαφέρει να κάνει το μοντέλο και επίσης διαχειρίζονται καλά τα outliers.

Gradient Boosting

Η διαδικασία κατά την οποία ένα ensemble χτίζεται σταδιακά προσθέτοντας μοντέλα ώστε κάθε νέο μοντέλο να προσπαθεί να διορθώσει το σφάλμα του προηγούμενου μέχρι τότε συνδυασμού και να γίνεται πιο αποτελεσματικό, λέγεται boosting.

Ο Gradient Boosting έχει τρία κύρια συστατικά:

- Loss Function - Ο ρόλος της είναι να εκτιμήσει πόσο καλό είναι το μοντέλο στο να κάνει προβλέψεις με τα δεδομένα. Αυτό μπορεί να διαφέρει ανάλογα με το πρόβλημα.
- Weak Learner - Είναι αυτός που ταξινομεί τα δεδομένα αλλά το κάνει με κακό τρόπο, ίσως όχι καλύτερα από μία τυχαία εικασία. Με άλλα λόγια, έχει υψηλό ποσοστό σφαλμάτων. Αυτά είναι συνήθως δέντρα αποφάσεων (ονομάζονται επίσης decision stumps, επειδή είναι λιγότερο περίπλοκα από τα τυπικά δέντρα αποφάσεων).
- Additive Model - Είναι η επαναληπτική και διαδοχική προσέγγιση της προσθήκης των δέντρων (Weak Learners) ένα βήμα τη φορά. Μετά από κάθε επανάληψη, πρέπει να είμαστε πιο κοντά στο τελικό μοντέλο. Με άλλα λόγια, κάθε επανάληψη θα πρέπει να μειώνει την αξία της συνάρτησης απώλειας.

Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι μοντέλα υπολογισμού που βασίζονται σε φυσικά, βιολογικά δίκτυα νευρώνων. Το μοντέλο αυτό παρουσιάστηκε το 1943 από τους Warren McCulloch και Walter Pitts στο άρθρο τους "A logical calculus of the ideas immanent in nervous activity". Μετά από μια περίοδο το ενδιαφέρον για αυτά μειώθηκε. Κύριος λόγος ήταν ότι οι υπολογιστές εκείνης της εποχής δεν είχαν αρκετή υπολογιστική ισχύ.

Η ανακάλυψη της backpropagation, μιας μεθόδου για την εκπαίδευση νευρωνικών δικτύων, από τον Paul J Werbos το 1974 έδωσε ένα εργαλείο για την αποτελεσματική κατασκευή και εκπαίδευση νευρωνικών δικτύων. Σε συνδυασμό με τον εκθετικό ρυθμό αύξησης της υπολογιστικής ισχύος τα τελευταία χρόνια, η μέθοδος αυτή έχει κάνει τα νευρωνικά δίκτυα την κυρίαρχη μέθοδο σε εφαρμογές όπως computer vision, επεξεργασία φυσικής γλώσσας κ.α.

Ο συγγραφέας Carbonneau et al. (2007) στην ερευνητική του εργασία συνέκρινε διάφορες παραδοσιακές χρονοσειρές προβλέψεων όπως ο κινητός μέσος όρος, η γραμμική παλινδρόμηση με επαναλαμβανόμενα νευρωνικά δίκτυα και μηχανές διανυσμάτων υποστήριξης και κατέληξε στο συμπέρασμα ότι τα επαναλαμβανόμενα νευρωνικά δίκτυα είχαν καλύτερη απόδοση.

Τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν για τη δημιουργία μοντέλων ταξινόμησης (classification) αλλά και για τη δημιουργία μοντέλων πρόβλεψης (prediction).

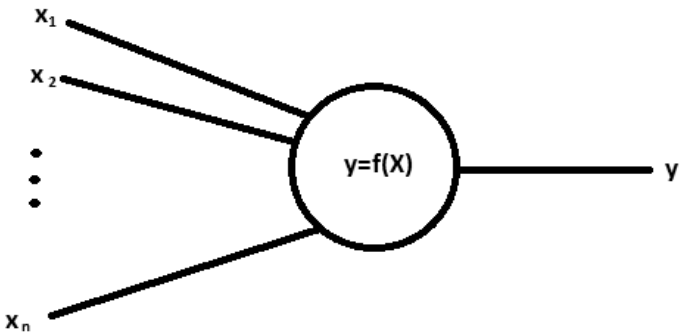


Figure 4: Νευρώνας (τεχνητού) νευρωνικού δικτύου

Ένας νευρώνας σε ένα τέτοιο δίκτυο δέχεται αριθμητικές τιμές ως είσοδο/ερεθίσματα. Στην εικόνα παριστάνονται ως x1, x2, …, xn. Ο νευρώνας αναθέτει ένα συγκεκριμένο βάρος αι σε κάθε είσοδο και στη συνέχεια υπολογίζει το άθροισμά τους:

$$X = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

Η έξοδος είναι η τιμή μιας συνάρτησης f υπολογισμένη πάνω στο X, δηλ. f(X). Η συνάρτηση αυτή λέγεται συνάρτηση ενεργοποίησης (activation function). Στη συνέχεια, η τιμή f(X) χρησιμοποιείται ως είσοδος σε άλλους νευρώνες ή αποτελεί μέρος της εξόδου του νευρωνικού δικτύου αν ο συγκεκριμένος νευρώνας είναι μέλος του στρώματος εξόδου.

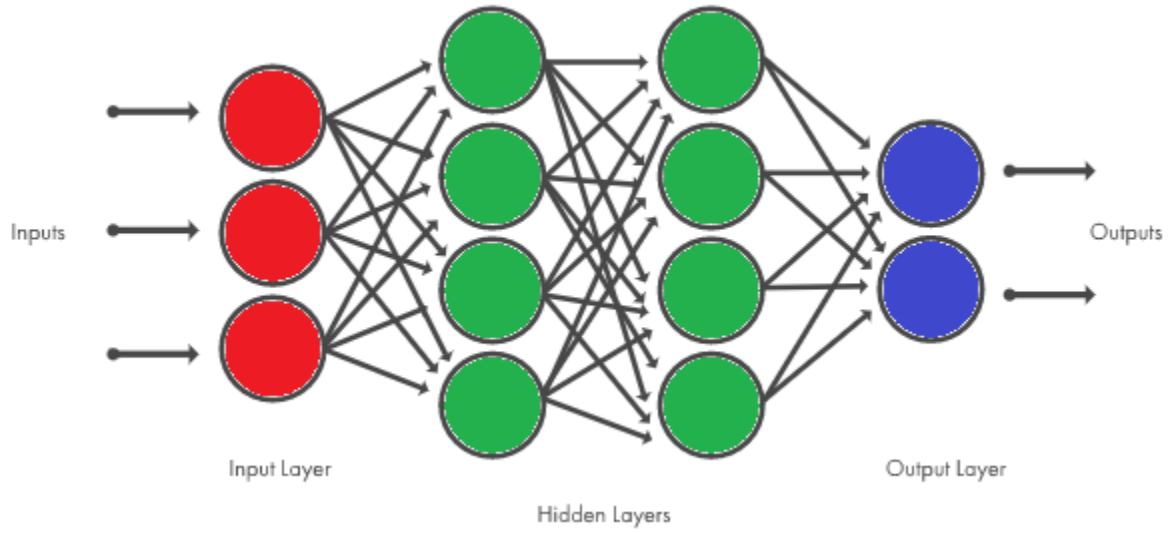


Figure 5: Παράδειγμα ενός Νευρωνικού Δικτύου

Ένα νευρωνικό δίκτυο αποτελείται από ομάδες νευρώνων που ονομάζονται στρώματα (layers). Αυτά είναι:

- Το στρώμα εισόδου (input layer) που απαρτίζεται από τους νευρώνες που δέχονται ως είσοδο τις τιμές/δεδομένα με τις οποίες τροφοδοτείται το νευρωνικό δίκτυο
- Το στρώμα εξόδου (output layer) που απαρτίζεται από τους νευρώνες των οποίων η έξοδος αποτελεί την έξοδο του νευρωνικού δικτύου
- Τα κρυφά στρώματα (hidden layers) τα οποία βρίσκονται ανάμεσα

Οι νευρώνες ενός στρώματος δεν επικοινωνούν μεταξύ τους. Δέχονται τις τιμές εξόδου νευρώνων από το προηγούμενο στρώμα και στέλνουν τις δικές τους τιμές εξόδου στους νευρώνες του επόμενου στρώματος. Υπάρχουν και νευρωνικά δίκτυα που αποτελούνται από ένα μόνο στρώμα ή ακόμη και από ένα μόνο νευρώνα.

- βαθύ νευρωνικό δίκτυο (deep neural net) λέγεται ένα νευρωνικό δίκτυο με πολλά κρυφά στρώματα
- πυκνό νευρωνικό δίκτυο (dense neural network) λέγεται ένα νευρωνικό δίκτυο του οποίου οι νευρώνες κάθε στρώματος συνδέονται με όλους τους νευρώνες του επόμενου

Συνάρτηση ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης είναι ένα από τα σημαντικότερα στοιχεία των νευρωνικών δικτύων. Η επιλογή της συνάρτησης ενεργοποίησης στο κρυφό επίπεδο θα ελέγχει πόσο καλά το μοντέλο δικτύου μαθαίνει από το σύνολο δεδομένων εκπαίδευσης, ενώ στο επίπεδο εξόδου θα καθορίσει τον τύπο των προβλέψεων που μπορεί να κάνει το μοντέλο. Μια συνάρτηση ενεργοποίησης σε ένα νευρωνικό δίκτυο ορίζει πώς το σταθμισμένο άθροισμα της εισόδου μετατρέπεται σε έξοδο από έναν κόμβο ή κόμβους σε ένα επίπεδο του δικτύου.

Η ενεργοποίηση για τα κρυφά στρώματα περιλαμβάνει 3 βασικές μεθόδους ενεργοποίησης:

- Rectified Linear Activation (**ReLU**): Είναι ίσως η πιο συνηθισμένη επειδή είναι τόσο απλή στην εφαρμογή της όσο και αποτελεσματική στην υπέρβαση των περιορισμών άλλων παλαιότερα δημοφιλών λειτουργιών ενεργοποίησης, όπως η Sigmoid και η Tanh. Συγκεκριμένα, είναι λιγότερο επιρρεπής σε εξαφανιζόμενες κλίσεις (vanishing gradients) που εμποδίζουν την εκπαίδευση των deep models, αν και μπορεί να υποφέρει από άλλα προβλήματα όπως κορεσμένες ή «νεκρές» μονάδες.
  Η συνάρτηση ReLU υπολογίζεται ως εξής: max(0.0, x). Αυτό σημαίνει ότι εάν η τιμή εισόδου (x) είναι αρνητική, τότε επιστρέφεται μια τιμή 0.0, διαφορετικά επιστρέφεται η τιμή.
- Logistic (**Sigmoid**): Είναι η ίδια συνάρτηση που χρησιμοποιείται στον αλγόριθμο ταξινόμησης λογιστικής παλινδρόμησης. Παίρνει οποιαδήποτε πραγματική τιμή ως είσοδο και εξάγει τιμές στην περιοχή από 0 έως 1. Όσο μεγαλύτερη είναι η είσοδος (πιο θετική), τόσο πιο κοντά η τιμή εξόδου θα είναι στο 1.0, ενώ όσο μικρότερη είναι η είσοδος (πιο αρνητική), τόσο πιο κοντά η έξοδος θα είναι στο 0.0.
- Hyperbolic Tangent (**Tanh**): Μοιάζει πολύ με τη λειτουργία ενεργοποίησης sigmoid και μάλιστα έχει το ίδιο σχήμα S. Η συνάρτηση παίρνει οποιαδήποτε πραγματική τιμή ως είσοδο και εξάγει τιμές στην περιοχή -1 έως 1. Όσο μεγαλύτερη είναι η είσοδος (πιο θετική), τόσο πιο κοντά η τιμή εξόδου θα είναι στο 1.0, ενώ όσο μικρότερη είναι η είσοδος (πιο αρνητική), τόσο πιο κοντά η έξοδος θα είναι στο -1.0.

Ένα νευρωνικό δίκτυο θα έχει σχεδόν πάντα την ίδια λειτουργία ενεργοποίησης σε όλα τα κρυφά του στρώματα.

Η ενεργοποίηση για τα στρώματα εξόδου περιλαμβάνει 3 βασικές μεθόδους ενεργοποίησης:

- Linear: Η συνάρτηση γραμμικής ενεργοποίησης δεν αλλάζει με κανέναν τρόπο το σταθμισμένο άθροισμα της εισόδου και αντ' αυτού επιστρέφει απευθείας την τιμή.
- Logistic (Sigmoid)
- Softmax: Η συνάρτηση softmax εξάγει ένα διάνυσμα τιμών που αθροίζονται σε 1.0 και μπορούν να ερμηνευθούν ως πιθανότητες συμμετοχής στην κλάση. Η Softmax είναι μια «πιο μαλακή» έκδοση της argmax που επιτρέπει μια έξοδο παρόμοια με πιθανότητες μιας συνάρτησης winner-take-all. Ως εκ τούτου, η είσοδος στη συνάρτηση είναι ένα διάνυσμα πραγματικών τιμών και η έξοδος είναι ένα διάνυσμα του ίδιου μήκους με τιμές που αθροίζονται σε 1.0 πιθανότητες.

Νευρωνικά Μοντέλα

Μερικά από τα σημαντικά μοντέλα νευρωνικών δικτύων είναι τα παρακάτω.

1. **CNN (Convolutional Neural Networks)**: Τα CNN διακρίνονται από άλλα νευρωνικά δίκτυα για την ανώτερη απόδοσή τους με εισόδους σημάτων εικόνας, ομιλίας ή ήχου. Έχουν τρεις κύριους τύπους layers, τα οποία είναι: Convolutional layer, Pooling layer, Fully-connected (FC) layer.

   Το convolution layer είναι το βασικό δομικό στοιχείο ενός CNN, και εκεί λαμβάνει χώρα η πλειοψηφία των υπολογισμών. Απαιτεί μερικά στοιχεία, τα οποία είναι δεδομένα εισόδου, ένα φίλτρο και ένας χάρτης χαρακτηριστικών. Μετά από κάθε λειτουργία convolution, ένα CNN εφαρμόζει έναν μετασχηματισμό ReLU στον χάρτη χαρακτηριστικών, εισάγοντας μη γραμμικότητα στο μοντέλο.
   Τα Pooling layers, γνωστά και ως downsampling, πραγματοποιούν μείωση διαστάσεων, μειώνοντας τον αριθμό των παραμέτρων στην είσοδο. Παρόμοια με το convolutional layer, η λειτουργία συγκέντρωσης σαρώνει ένα φίλτρο σε ολόκληρη την είσοδο, αλλά η διαφορά είναι ότι αυτό το φίλτρο δεν έχει βάρη. Αντίθετα, ο πυρήνας εφαρμόζει μια συνάρτηση συνάθροισης (aggregation) στις τιμές εντός του πεδίου υποδοχής, συμπληρώνοντας τον πίνακα εξόδου.
   Στο fully-connected layer, κάθε κόμβος στο επίπεδο εξόδου συνδέεται απευθείας με έναν κόμβο στο προηγούμενο επίπεδο. Αυτό το layer εκτελεί το καθήκον της ταξινόμησης με βάση τα χαρακτηριστικά που εξάγονται μέσω των προηγούμενων layers και των διαφορετικών φίλτρων τους. Ενώ τα convolutional και τα pooling layers τείνουν να χρησιμοποιούν συναρτήσεις ReLu, τα FC layers χρησιμοποιούν συνήθως μια συνάρτηση ενεργοποίησης softmax για την κατάλληλη ταξινόμηση των εισόδων, παράγοντας μια πιθανότητα από 0 έως 1.

2. **RNN (Recurrent Neural Networks)**: Το επαναλαμβανόμενο νευρωνικό δίκτυο (RNN) είναι ένας τύπος τεχνητού νευρωνικού δικτύου που χρησιμοποιεί διαδοχικά δεδομένα ή δεδομένα χρονοσειρών. Αυτοί οι αλγόριθμοι deep learning χρησιμοποιούνται συνήθως για τακτικά ή χρονικά προβλήματα, όπως η μετάφραση γλώσσας, η επεξεργασία φυσικής γλώσσας (nlp), η αναγνώριση ομιλίας κ.α. Ενσωματώνονται σε δημοφιλείς εφαρμογές όπως η Siri, η φωνητική αναζήτηση και η μετάφραση της Google. Όπως τα feedforward και convolutional νευρωνικά δίκτυα (CNN), τα RNN χρησιμοποιούν δεδομένα εκπαίδευσης για μάθηση. Διακρίνονται από τη «μνήμη» τους καθώς λαμβάνουν πληροφορίες από προηγούμενες εισόδους για να επηρεάσουν την τρέχουσα είσοδο και έξοδο. Ενώ τα παραδοσιακά βαθιά νευρωνικά υποθέτουν ότι οι είσοδοι και οι έξοδοι είναι ανεξάρτητες μεταξύ τους, η έξοδος των επαναλαμβανόμενων νευρωνικών δικτύων εξαρτάται από τα προηγούμενα στοιχεία της ακολουθίας. Ενώ μελλοντικά γεγονότα θα ήταν επίσης χρήσιμα για τον προσδιορισμό της εξόδου μιας δεδομένης ακολουθίας, τα επαναλαμβανόμενα νευρωνικά δίκτυα μονής κατεύθυνσης δεν μπορούν να λάβουν υπόψη αυτά τα συμβάντα στις προβλέψεις τους.

   Ένα άλλο διακριτικό χαρακτηριστικό των RNN είναι ότι μοιράζονται παραμέτρους σε κάθε layer του δικτύου. Ενώ τα feedforward δίκτυα έχουν διαφορετικά βάρη σε κάθε κόμβο, τα επαναλαμβανόμενα νευρωνικά δίκτυα μοιράζονται την ίδια παράμετρο βάρους σε κάθε layer του δικτύου. Παρόλα αυτά, τα βάρη εξακολουθούν να προσαρμόζονται κατά τη διάρκεια των διαδικασιών backpropagation και gradient descent για να διευκολυνθεί η ενισχυτική μάθηση.

Τα RNN τείνουν να αντιμετωπίζουν δύο προβλήματα, γνωστά ως exploding gradients και vanishing gradients. Όταν η κλίση είναι πολύ μικρή, συνεχίζει να γίνεται μικρότερη, ενημερώνοντας τις παραμέτρους βάρους μέχρι να γίνουν ασήμαντες, δηλαδή 0. Όταν συμβεί αυτό, ο αλγόριθμος δεν μαθαίνει πλέον. Οι Exploding gradients συμβαίνουν όταν η κλίση είναι πολύ μεγάλη, δημιουργώντας ένα ασταθές μοντέλο. Σε αυτήν την περίπτωση, τα βάρη του μοντέλου θα μεγαλώσουν πολύ και τελικά θα αναπαρασταθούν ως NaN. Μια λύση σε αυτά τα ζητήματα είναι η μείωση του αριθμού των hidden layers μέσα στο νευρωνικό δίκτυο, εξαλείφοντας μέρος της πολυπλοκότητας στο RNN μοντέλο.

3. **LSTM (Long Short-Term Memory)**: Οι αρχιτεκτονικές νευρωνικών δικτύων όπως τα MLP και τα δίκτυα που βασίζονται σε LSTM είναι κατάλληλα για προβλήματα παλινδρόμησης και χρονοσειρών πολλών σταδίων, καθώς μπορούν να εκπαιδευτούν ώστε να προβλέπουν ολόκληρη την ακολουθία εξόδου με ένα μοντέλο. Τα LSTM βασίζονται στην αρχιτεκτονική των RNN και περιέχουν εσωτερικές μεταβλητές που καταγράφουν ολόκληρη την ιστορία της χρονοσειράς, επιτρέποντας στο μοντέλο να μάθει από γεγονότα που συνέβησαν πριν το input περιθώριο της περιόδου πρόβλεψης. Τα layers των LSTM μπορούν να συνδυαστούν για να δημιουργήσουν μεγαλύτερες, βαθύτερες και πιο σύνθετες αρχιτεκτονικές μοντέλων που μπορούν να δημιουργήσουν προβλέψεις από μεγάλης κλίμακας και υψηλών διαστάσεων δεδομένα.

   Το (LSTM) μοντέλο σχεδιάστηκε για να ξεπερνά τα προβλήματα του απλού επαναλαμβανόμενου νευρωνικού δικτύου (RNN) επιτρέποντας στο δίκτυο να αποθηκεύει δεδομένα σε ένα είδος μνήμης στην οποία μπορεί να έχει πρόσβαση αργότερα. Εισάγει εκφράσεις, ειδικότερα, πύλες (gates). Συγκεκριμένα, υπάρχουν τρεις τύποι πυλών:

- πύλη που ξεχνάει – ελέγχει πόσες πληροφορίες θα λάβει το κελί μνήμης από το κελί μνήμης του προηγούμενου βήματος
- πύλη ενημέρωσης (εισόδου) – αποφασίζει εάν το κελί μνήμης θα ενημερωθεί. Επίσης, ελέγχει πόσες πληροφορίες θα λάβει το τρέχον κελί μνήμης από ένα δυνητικά νέο κελί μνήμης
- πύλη εξόδου – ελέγχει την τιμή της επόμενης κρυφής κατάστασης

   Το κλειδί του μοντέλου LSTM είναι η κατάσταση των κελιών (cell state). Η κατάσταση του κελιού ενημερώνεται δύο φορές με λίγους υπολογισμούς. Έχει επίσης μια κρυφή κατάσταση που λειτουργεί σαν βραχυπρόθεσμη μνήμη. Η παράμετρος return_sequences στη συνάρτηση LSTM καθορίζει εάν θα επιστρέψει την τελευταία έξοδο στην ακολουθία εξόδου ή την πλήρη ακολουθία. Εάν είναι True, το layer LSTM εξάγει έναν τρισδιάστατο tensor (batch_size, timesteps, units), ο οποίος μπορεί να τροφοδοτηθεί απευθείας στο επόμενο LSTM layer.

   Ένα LSTM με χρονικό βήμα (timestep) 1 είναι παρόμοιο με ένα MLP στο ότι δεν έχει επαναλαμβανόμενη σύνδεση με ένα προηγούμενο χρονικό βήμα. Ωστόσο, ένα LSTM με χρονικό βήμα 1 εξακολουθεί να έχει το κελί μνήμης και τους μηχανισμούς πύλης που του επιτρέπουν επιλεκτικά να διαβάζει, να γράφει και να ξεχνά πληροφορίες από προηγούμενα χρονικά βήματα. Επομένως, μπορεί ακόμα να χρησιμοποιήσει το κελί μνήμης του για να αποθηκεύσει και να ανακτήσει πληροφορίες από προηγούμενα χρονικά βήματα.

Η αρχιτεκτονική του μοντέλου κωδικοποιητή-αποκωδικοποιητή (encoder-decoder) αποτελείται από δύο στοιχεία. Τον κωδικοποιητή, ο οποίος κωδικοποιεί τα χαρακτηριστικά και τον αποκωδικοποιητή, ο οποίος επιχειρεί να ανακατασκευάσει χαρακτηριστικά με βάση μια συμπιεσμένη κατάσταση των δεδομένων εισόδου. Το δίκτυο κωδικοποιητή μαθαίνει μια αναπαράσταση της ακολουθίας εισόδου που καταγράφει τα χαρακτηριστικά του ή το περιεχόμενο και παράγει ένα context vector. Το δίκτυο αποκωδικοποιητή λαμβάνει το context vector και μαθαίνει να διαβάζει και να εξάγει την ακολουθία εξόδου από αυτό. Τα δύο components περιέχουν ένα LSTM layer που λαμβάνει και ερμηνεύει τη διαδοχική εισαγωγή δεδομένων.

4. **MLP (Multilayer Perceptron)**: Τα MLP ανήκουν στην κατηγορία των νευρωνικών δικτύων τροφοδοσίας με πολλαπλά layers perceptrons που έχουν λειτουργίες ενεργοποίησης. Αποτελούνται από ένα layer εισόδου και ένα layer εξόδου που είναι πλήρως συνδεδεμένα. Έχουν τον ίδιο αριθμό layers εισόδου και εξόδου, αλλά μπορεί να έχουν πολλαπλά hidden layers και μπορούν να χρησιμοποιηθούν για τη δημιουργία λογισμικού αναγνώρισης ομιλίας, αναγνώρισης εικόνας και μηχανικής μετάφρασης.

   Τα MLP τροφοδοτούν τα δεδομένα στο layer εισόδου του δικτύου. Τα layers των νευρώνων συνδέονται σε ένα γράφημα έτσι ώστε το σήμα να περνά προς μία κατεύθυνση. Υπολογίζουν την είσοδο με τα βάρη που υπάρχουν μεταξύ του layer εισόδου και των hidden layers. Έπειτα χρησιμοποιούν λειτουργίες ενεργοποίησης για να καθορίσουν ποιους κόμβους θα ενεργοποιήσουν. Οι συναρτήσεις ενεργοποίησης περιλαμβάνουν ReLU, sigmoid συναρτήσεις και tanh. Τέλος, εκπαιδεύουν το μοντέλο ώστε να κατανοεί τη συσχέτιση και να μαθαίνει τις εξαρτήσεις μεταξύ των ανεξάρτητων και των μεταβλητών-στόχων από ένα σύνολο δεδομένων εκπαίδευσης.

Cloud Computing

Με τον όρο cloud computing εννοούμε τη χρήση υπολογιστικών συστημάτων και υπηρεσιών τα οποία ως υλικό δεν συντηρούνται από τους χρήστες τους, αλλά προσφέρονται μέσω διαδικτύου. Είναι η διαδικασία παροχής υπολογιστικών πόρων όπως αποθηκευτικός χώρος, εφαρμογές, βάσεις δεδομένων, λογισμικό και υπηρεσίες. Το βασικό πλεονέκτημα της χρήσης cloud είναι ότι ο χρήστης αυτών των υπηρεσιών δεν επιβαρύνεται από το κόστος της αγοράς και συντήρησης εξοπλισμού. Αυτό λύνει τα χέρια σε πολλές επιχειρήσεις, καθώς πολλές από τις διαδικασίες που απαιτούνται για την λειτουργία τους διαφέρουν σε κόστος, περιπλοκότητα και χρησιμότητα, οπότε το να επενδύσουν οι ίδιες τους πόρους τους για όλες τις υπηρεσίες που μπορεί να χρειαστούν στο μέλλον ή ανά πάσα στιγμή δεν θα ήταν επικερδές για αυτές.

Παράλληλα, πολλοί πάροχοι δίνουν τη δυνατότητα αυξομείωσης των πόρων που μισθώνει κάποιος ώστε να υπάρχει η βέλτιστη χρήση τους και το κόστος τους να ανταποκρίνεται στις πραγματικές ανάγκες του κάθε χρήστη. Στα αρνητικά θα πρέπει να αναφέρουμε ότι το κόστος τέτοιων υπηρεσιών μπορεί, δυνητικά, να γίνει μεγάλο και χρειάζεται προσεκτική μελέτη πριν την υιοθέτηση μιας τέτοιας λύσης αλλά και συνεχής παρακολούθηση του κόστους χρήσης της.

Τα τελευταία χρόνια στην αγορά παροχής cloud έχουν μπει μεγάλοι τεχνολογικοί κολοσσοί όπως η Amazon με το Amazon Web Services (AWS), η Google με το Google Cloud και η Microsoft με το Azure. Το γεγονός αυτό έχει αυξήσει τον ανταγωνισμό. Το αποτέλεσμα είναι η διάθεση και άλλων υπηρεσιών cloud πέρα από την κλασική αποθήκευση δεδομένων και την παροχή virtual machines. Ένα ακόμα πολύ σημαντικό αρνητικό, ειδικά στην περίπτωση των μεγάλων παρόχων, είναι ότι κάθε πάροχος χρησιμοποιεί την δικιά του τεχνολογία και στήνει-οργανώνει διαφορετικά τις υπηρεσίες του. Συνεπώς αν μια επιχείρηση που έχει επενδύσει πολλές από τις διεργασίες της στον cloud τομέα αποφασίσει να αλλάξει πάροχο (για λόγους οικονομικούς ή και άλλους), τότε θα πρέπει να ακολουθήσει μια χρονοβόρα μεταφορά αρχείων και δεδομένων από την μια πλατφόρμα στην άλλη. Αυτό απαιτεί βέβαια μεγάλο βαθμό τεχνογνωσίας καθώς και εκπαίδευσης του προσωπικού ώστε να μπορεί να διαχειριστεί το νέο περιβάλλον του καινούργιου cloud παρόχου.

Για την παρούσα εργασία έχει γίνει χρήση της Cloud υπηρεσίας της Microsoft, το Azure. Η βάση δεδομένων που έχει χρησιμοποιηθεί, καθώς και τα δεδομένα, είναι αποθηκευμένα και τρέχουν μέσω cloud.

Η 4η βιομηχανική επανάσταση

Ο όρος «4η βιομηχανική επανάσταση» χρησιμοποιήθηκε το 2015 από τον Klaus Schwab, ιδρυτή και εκτελεστικό διευθυντή του World Economic Forum, στο άρθρο του "Mastering the Fourth Industrial Revolution". Σε αυτό, εκτιμούσε ότι η ανθρωπότητα είναι στα πρόθυρα μιας τεχνολογικής επανάστασης που θα αλλάξει τον τρόπο ζωής μας. Σύμφωνα με τον Schwab, η επανάσταση αυτή θα χτίσει πάνω στην Τρίτη Βιομηχανική Επανάσταση, τη λεγόμενη ψηφιακή επανάσταση, που αφορούσε την εκτεταμένη χρήση υπολογιστών και αυτοματισμών.

Ως βασικοί πυλώνες της, αναφέρονται η Τεχνητή Νοημοσύνη, η ρομποτική, το Internet of Things, τα αυτόνομα οχήματα, το 3-D printing, η νανοτεχνολογία, η βιοτεχνολογία, η επιστήμη υλικών, η αποθήκευση ενέργειας και το quantum computing. Οι Erik Brynjolfsson και Andrew McAfee στο βιβλίο τους "The Second Machine Age" υποστηρίζουν ότι η 4η βιομηχανική επανάσταση θα στηρίζεται στην αυτοματοποίηση των νοητικών εργασιών που θα επιτευχθεί μέσω της Τεχνητής Νοημοσύνης. Η 4η βιομηχανική επανάσταση προβλέπεται ότι θα φέρει μεγάλες αλλαγές στην οικονομία και την αγορά εργασίας.

Με τον όρο βιομηχανική επανάσταση εννοούμε μια περίοδο κατά την οποία μια ή και περισσότερες τεχνολογίες και τεχνικές αντικαθίστανται από άλλες σε ένα σχετικά σύντομο χρονικό διάστημα. Η περίοδος αυτή χαρακτηρίζεται από επιταχυνόμενη τεχνολογική πρόοδο που η εφαρμογή της εξαπλώνεται με γρήγορους ρυθμούς.

Οι τέσσερις βιομηχανικές επαναστάσεις στις οποίες αναφερόμαστε σήμερα είναι:

- Η πρώτη βιομηχανική επανάσταση που εξελίχθηκε κυρίως στην Ευρώπη και τις ΗΠΑ από τα μέσα του 18ου έως τα μέσα του 19ου αιώνα και χαρακτηρίστηκε από την αντικατάσταση της χειρονακτικής εργασίας από ατμοκίνητες μηχανές.
- Η δεύτερη βιομηχανική επανάσταση από τα τέλη του 19ου έως τις αρχές του 20ου αιώνα. Χαρακτηρίζεται κυρίως από την αντικατάσταση του ατμού με τον ηλεκτρισμό.
- Η τρίτη βιομηχανική επανάσταση ή ψηφιακή επανάσταση που έλαβε χώρα κυρίως μεταξύ 1950 και 1980. Συνίσταται στην εισαγωγή των ηλεκτρονικών υπολογιστών και την ψηφιοποίηση.
- Η τέταρτη βιομηχανική επανάσταση που σύμφωνα με αρκετούς σύγχρονους στοχαστές είναι προ των θυρών και θα χαρακτηρίζεται από την ταχύτατη υιοθέτηση καινοτομιών στους τομείς που αναφέρθηκαν πιο πάνω.

Οι βασικές της συνιστώσες

Έρχεται για να προσφέρει ένα καλύτερο βιοτικό επίπεδο και υπόσχεται υψηλούς ρυθμούς ανάπτυξης για όσες χώρες δεν φοβηθούν και επενδύσουν σε νέες τεχνολογίες. Η ταχύτητα εξέλιξής της, το εύρος των διαταραχών που επιφέρει και η περίοδος στην οποία συμβαίνει είναι μερικοί από τους λόγους που θα πρέπει να διαχωριστεί από οποιαδήποτε επανάσταση του παρελθόντος.

Η μεταφορά πλούτου από τη Δύση στην Ανατολή και η διαφαινόμενη κυριαρχία της Κίνας στον τομέα των τεχνολογιών μεταβάλλουν και τη γεωγραφική προέλευση της 4ης Βιομηχανικής Επανάστασης η οποία φαίνεται να ξεκινάει από τις αναδυόμενες χώρες. Εκμάθηση μηχανών (machine learning), Μεγάλα Δεδομένα (Big Data), Ρομποτική (Robotics), Τεχνητή Νοημοσύνη (Artificial Intelligence), Ίντερνετ των Πραγμάτων (Internet of Things-IOT) είναι μερικές από τις ανακαλύψεις που βρίσκονται στον πυρήνα της επανάστασης αυτής.

Η ανθρωπότητα βρίσκεται μπροστά σε μια νέα εποχή και σε μια νέα Βιομηχανική Επανάσταση με αντίκτυπο σε όλα τα κοινωνικά στρώματα, γεγονός που την διαφοροποιεί ουσιαστικά από τις προηγούμενες που είχαν αντίκτυπο στις χαμηλές βαθμίδες της κοινωνικής πυραμίδας. Οι ανακαλύψεις που συνοδεύουν την επανάσταση αυτή δεν έχουν προηγούμενο ιστορικά και αναμένεται να συντελέσουν τα μέγιστα στην πρόοδο διαφόρων επιστημών, όπως η μηχανική, η αρχιτεκτονική, η ηλεκτρονική, η βιολογία, αστροφυσική κ.α. Οι εργαζόμενοι με χαμηλές δεξιότητες θα έρθουν αντιμέτωποι με δυσκολίες και θα δουν τα ευρήματα της τεχνολογίας να τους αντικαθιστούν.

Η 4η Βιομηχανική Επανάσταση έρχεται για να φέρει μεγάλες ανακατατάξεις στον τομέα των υπηρεσιών. Ο χρόνος που απαιτείται για την έρευνα αγοράς έχει μειωθεί δραστικά, με τους καταναλωτές να στηρίζουν όλο και περισσότερο αυτές τις πρωτοβουλίες καθώς έχουν πληθώρα επιλογών. Υποστηρικτικά συστήματα όπως η ενημέρωση για τον χρόνο παράδοσης (delivery time process) και η ενημέρωση για την διαδρομή που ακολουθεί το προϊόν (tracking system) βοηθούν τον πελάτη να αισθάνεται ασφαλής για την αγορά που πραγματοποίησε και αυξάνουν την ποιότητα των παρεχόμενων υπηρεσιών. Η ευρεία χρήση των πιστωτικών καρτών είναι ένας ακόμη παράγοντας που ενισχύει το ηλεκτρονικό εμπόριο.

Ωστόσο θα πρέπει να σημειωθεί πως οι πιέσεις για εξεύρεση λύσης σε διάφορα θέματα οικονομικού, κοινωνικού και περιβαλλοντικού ενδιαφέροντος είναι αυτές που επιταχύνουν την έλευση της 4ης Βιομηχανικής Επανάστασης. Πλέον με την 4^{η} Βιομηχανική Επανάσταση η τεχνολογία εισβάλει σε κάθε κλάδο της παραγωγής με αποτέλεσμα να δημιουργείται η ανάγκη ισόρροπης ανάπτυξης όλων των κλάδων προκειμένου να επιβιώσουν σε ένα διεθνές ανταγωνιστικό περιβάλλον.

Τα Μεγάλα Δεδομένα (Big Data)

Μία ακόμη ιδιαίτερα σημαντική πτυχή της 4ης Βιομηχανικής Επανάστασης είναι τα Μεγάλα Δεδομένα (Big Data) και η ανάλυση τους. Τα Μεγάλα Δεδομένα ορίζονται ως δεδομένα που είναι υπερβολικά μεγάλα για να διαχειριστούν από τις παραδοσιακές βάσεις δεδομένων. Οι μεγάλες αυτές δομές δεδομένων αποτελούνται από ένα μείγμα δομημένων και μη δομημένων δεδομένων.

Οι επιχειρήσεις καλούνται να αναλύσουν όλα τα διαφορετικά δεδομένα και να τα αξιοποιήσουν έτσι ώστε να προσφέρουν στα προϊόντα τους μεγαλύτερη προστιθέμενη αξία, ενώ παράλληλα να τα μετασχηματίσουν με σκοπό να ανταποκριθούν στις μεταβαλλόμενες καταναλωτικές ανάγκες.

Τα Μεγάλα Δεδομένα απαιτούν για να αποθηκευτούν και να αναλυθούν, τις επενδύσεις από πλευράς επιχειρήσεων σε τεχνολογικό κεφάλαιο και σε εξειδικευμένο εργατικό δυναμικό που θα είναι ικανό να χειριστεί αυτή τη μεγάλη μάζα πληροφοριών. Η ανάλυση τους οδηγεί στη κατανόηση των τάσεων και στην εξατομίκευση προϊόντων και υπηρεσιών έτσι ώστε να υπάρχει πελατοκεντρική προσέγγιση στη παραγωγή. Πλέον δεν δίνεται βάση μόνο στο προϊόν αυτό καθαυτό αλλά και στην αίσθηση και εμπειρία του πελάτη πριν και μετά την αγορά του. Τα Μεγάλα Δεδομένα έρχονται για να μετασχηματίσουν τις στρατηγικές των επιχειρήσεων και να διευκολύνουν στη προσέλκυση των κατάλληλων πελατών με βάση τις προδιαγραφές του προϊόντος. Αυτό επιτυγχάνετε σε μεγάλο βαθμό με την τηλεμετρία (telemetry) που χρησιμοποιείται από πολλές υπηρεσίες και προγράμματα και έχει ως σκοπό την συλλογή πληροφοριών σχετικά με τις συνήθειες του χρήστη και με τα στοιχεία του. Αυτό γίνεται πολλές φορές εις βάρος της ιδιωτικότητας του ίδιου του χρήστη.

Η επιστήμη των μεγάλων δεδομένων θα αυξήσει την προβλεπτική ικανότητα των επιστημόνων προσθέτοντας μεγαλύτερη ακρίβεια και περισσότερη πληροφόρηση. Ωστόσο θα πρέπει να γίνει σαφές πως η διαδικασία της επεξεργασίας μεγάλης μάζας δεδομένων δεν αποτελεί μια απλή διαδικασία. Τα Μεγάλα Δεδομένα προκειμένου να διακρατήσουν τα χρήσιμα (δομημένα) δεδομένα και να απομακρύνουν τα μη δομημένα δεδομένα απαιτούν κάποιου είδους φιλτράρισμα.

Η δυνατότητα και η ανάγκη ανάλυσης τεραστίων όγκων δεδομένων είναι και η γενεσιουργός αιτία των Μεγάλων Δεδομένων. Όπως ακριβώς συμβαίνει και στους στατιστικούς ελέγχους, όσο μεγαλύτερη είναι η ποσότητα των δεδομένων που παράγεται και αναλύεται τόσο πιο αξιόπιστη είναι και η πρόβλεψη, με την προϋπόθεση ότι τα δεδομένα αυτά είναι έγκυρα. Η εγκυρότητα των δεδομένων είναι ένα ζήτημα που απασχολεί την ερευνητική κοινότητα. Όταν κανείς καταλήγει στην ανάλυση τους, αντιμετωπίζει πολλές φορές το πρόβλημα όχι της ποσότητας αλλά της ποιότητας των δεδομένων. Τα δεδομένα ενδέχεται να είναι λανθασμένα ή ακατάλληλα για τον σκοπό που έχουν συλλεχθεί. Το σωστό φιλτράρισμα τους είναι αναγκαίο προκειμένου να ελεγχθεί η αξιοπιστία και η συμβατότητα των ενδείξεων αυτών.

Big Data σε Εφοδιαστική Αλυσίδα

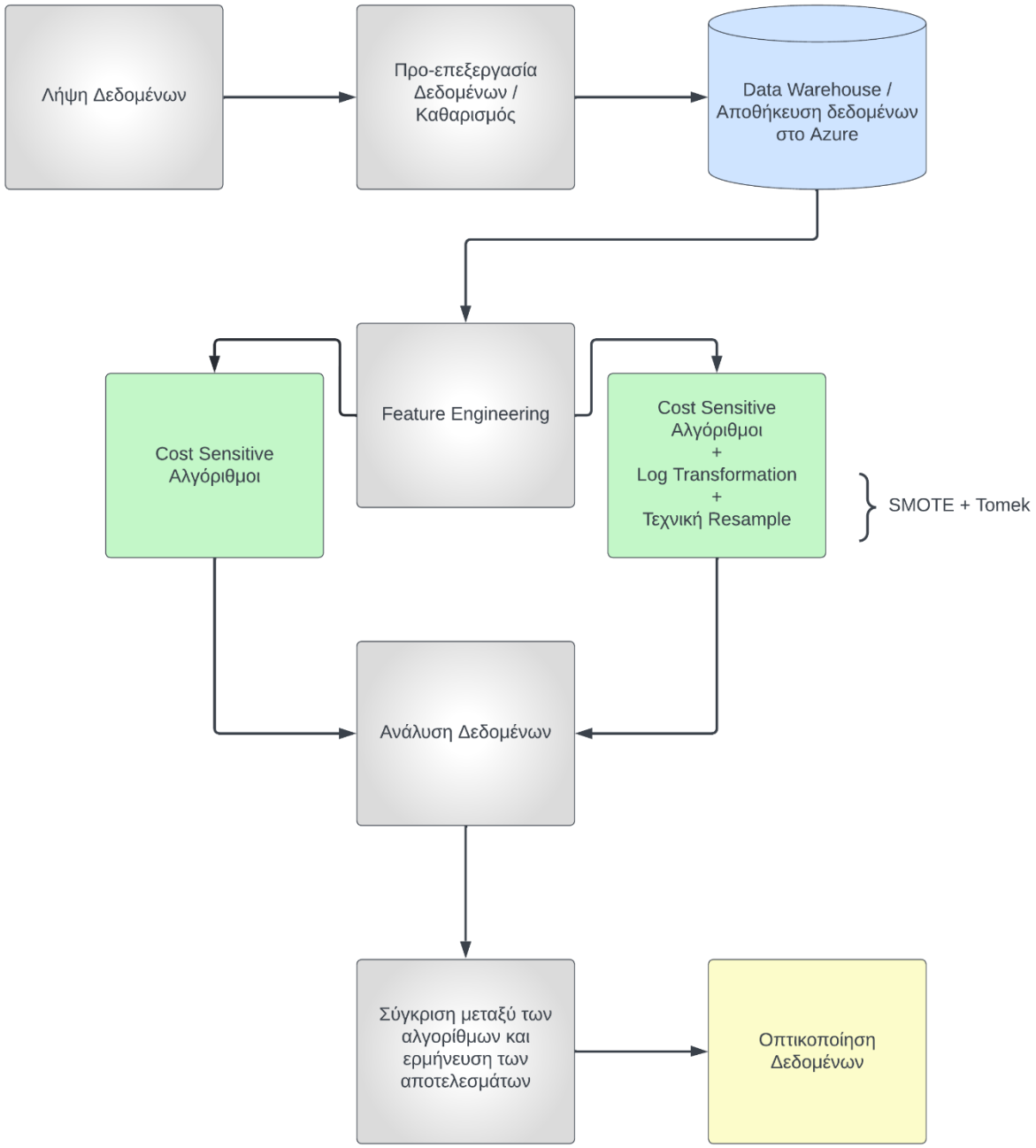
Η ανάλυση μεγάλων δεδομένων στη διαχείριση της αλυσίδας εφοδιασμού λαμβάνει ολοένα και μεγαλύτερη προσοχή. Αυτό οφείλεται στο γεγονός ότι έχει ένα ευρύ φάσμα εφαρμογών στο SCM, συμπεριλαμβανομένης της ανάλυσης συμπεριφοράς πελατών, της ανάλυσης τάσεων και της πρόβλεψης ζήτησης. Η αυξανόμενη ανάγκη για ανάλυση συμπεριφοράς πελατών και πρόβλεψη ζήτησης καθοδηγείται από την παγκοσμιοποίηση και τους αυξανόμενους ανταγωνισμούς της αγοράς, καθώς και από την αύξηση των πρακτικών ψηφιοποίησης της αλυσίδας εφοδιασμού.

Οι εφαρμογές ανάλυσης μεγάλων δεδομένων στην πρόβλεψη ζήτησης της εφοδιαστικής αλυσίδας έχουν χρησιμοποιηθεί και στις δύο κατηγορίες εποπτευόμενης και μη εποπτευόμενης μάθησης.

Στην εποπτευόμενη μάθηση, τα δεδομένα θα συσχετίζονται με ετικέτες, που σημαίνει ότι οι είσοδοι και οι έξοδοι είναι γνωστές. Οι εποπτευόμενοι αλγόριθμοι μάθησης προσδιορίζουν τις υποκείμενες σχέσεις μεταξύ των εισροών και των εξόδων σε μια προσπάθεια να χαρτογραφήσουν τις εισόδους σε αντίστοιχες εξόδους με βάση ένα νέο μη επισημασμένο σύνολο δεδομένων. Για παράδειγμα, στην περίπτωση ενός εποπτευόμενου μοντέλου μάθησης για την πρόβλεψη ζήτησης, η μελλοντική ζήτηση μπορεί να προβλεφθεί με βάση τα ιστορικά δεδομένα για τη ζήτηση προϊόντων.

Στην μάθηση χωρίς επίβλεψη, τα δεδομένα είναι χωρίς ετικέτα (δηλαδή άγνωστη έξοδο) και οι αλγόριθμοι προσπαθούν να βρουν τα υποκείμενα μοτίβα μεταξύ μη επισημασμένων δεδομένων αναλύοντας τις εισροές και τις αλληλεπιδράσεις τους. Η τμηματοποίηση πελατών είναι ένα παράδειγμα μάθησης χωρίς επίβλεψη στις αλυσίδες εφοδιασμού που συγκεντρώνει διαφορετικές ομάδες πελατών με βάση την ομοιότητά τους. Πολλοί αλγόριθμοι μηχανικής μάθησης/αναλυτικών δεδομένων μπορούν να διευκολύνουν τόσο την εποπτευόμενη μάθηση (εξαγωγή των σχέσεων εισόδου-εξόδου) όσο και την μάθηση χωρίς επίβλεψη (εξαγωγή εισροών, εκροών και των σχέσεών τους).

Project Workflow



Dataset

Όνομα Μεταβλητής	Επεξήγηση Μεταβλητής	Είδος Μεταβλητής
Sku (Stock Keeping Unit)	Αναγνωριστικός τυχαίος κωδικός για το προϊόν	Αριθμητική
national_inv	Τρέχον επίπεδο αποθέματος για το προϊόν	Αριθμητική
lead_time	Χρόνος μεταφοράς για το προϊόν (εάν υπάρχει)	Αριθμητική
in_transit_qty	Ποσότητα προϊόντων σε μεταφορά (από μία περιοχή σε άλλη)	Αριθμητική
forecast_3_month	Πρόβλεψη πωλήσεων για τους επόμενους 3 μήνες	Αριθμητική
forecast_6_month	Πρόβλεψη πωλήσεων για τους επόμενους 6 μήνες	Αριθμητική
forecast_9_month	Πρόβλεψη πωλήσεων για τους επόμενους 9 μήνες	Αριθμητική
sales_1_month	Ποσότητα πωλήσεων για τον προηγούμενο μήνα	Αριθμητική
sales_3_month	Ποσότητα πωλήσεων για τους προηγούμενους 3 μήνες	Αριθμητική
sales_6_month	Ποσότητα πωλήσεων για τους προηγούμενους 6 μήνες	Αριθμητική
sales_9_month	Ποσότητα πωλήσεων για τους προηγούμενους 9 μήνες	Αριθμητική
min_bank	Ελάχιστη συνιστώμενη ποσότητα για το απόθεμα	Αριθμητική
potential_issue	Εντοπισμένο πρόβλημα με το προϊόν/κομμάτι	Κατηγορική
pieces_past_due	Ανταλλακτικά ληξιπρόθεσμα/Ποσότητα εξαρτημάτων του προϊόντος σε καθυστέρηση, εάν υπάρχει	Αριθμητική
perf_6_month_avg	Απόδοση προϊόντος τους τελευταίους 6 μήνες	Αριθμητική
perf_12_month_avg	Απόδοση προϊόντος τους τελευταίους 12 μήνες	Αριθμητική
local_bo_qty	Ποσό καθυστερημένων παραγγελιών αποθεμάτων	Αριθμητική
deck_risk	Ένδειξη κινδύνου προϊόντος (Τα προϊόντα που ενδέχεται να παραμείνουν στο κατάστρωμα/κατάστημα)	Κατηγορική
oe_constraint	Ένδειξη κινδύνου προϊόντος (Προϊόντα που παρουσιάζουν λειτουργικούς περιοριστικούς παράγοντες, όπως bottleneck)	Κατηγορική
ppap_risk	Ένδειξη κινδύνου προϊόντος (Κίνδυνοι που συνδέονται με τη συσκευασία και την παραγωγή)	Κατηγορική
stop_auto_buy	Ένδειξη κινδύνου προϊόντος (η αυτόματη διαδικασία πώλησης έχει σταματήσει)	Κατηγορική
rev_stop	Ένδειξη κινδύνου προϊόντος (Κατάσταση εσόδων για το προϊόν)	Κατηγορική
went_on_backorder	Το προϊόν κατέληξε σε backorder. Αυτή είναι η τιμή στόχος	Κατηγορική

Επεξήγηση Δεδομένων και Παρατηρήσεις

Το σετ δεδομένων απαρτίζεται από ιστορικά δεδομένα για τα προϊόντα μιας επιχείρησης για 8 εβδομάδες πριν από την εβδομάδα που πρόκειται να προβλέψουμε. Τα δεδομένα λήφθηκαν ως εβδομαδιαία στιγμιότυπα στην αρχή κάθε εβδομάδας. Περιέχονται 23 στήλες, 15 αριθμητικές και 8 κατηγορηματικές. Στόχος μας είναι να κάνουμε πρόβλεψη εάν ένα προϊόν θα καταλήξει σε backorder ή όχι. Όπως παρατηρούμε υπάρχει μία πολύ μεγάλη διαφορά μεταξύ του πλήθους των παρατηρήσεων αυτών των δύο. Αυτό είναι πολύ λογικό αν σκεφτούμε ότι σχεδόν πάντα το ποσοστό των προϊόντων μιας επιχείρησης που καταλήγουν σε backorder είναι αρκετά μικρό. Στην περίπτωσή μας, μόνο το 0.7% έχει καταλήξει σε backorder.

Ιδανικά θα θέλαμε να έχουμε το ίδιο πλήθος παρατηρήσεων σε κάθε κλάση για να μπορέσουμε να εκπαιδεύσουμε τα μοντέλα μας σωστά (να μην έχουν bias), αλλά κάτι τέτοιο δεν συμβαίνει συχνά σε real world data. Επομένως πρέπει να χρησιμοποιήσουμε διάφορες τεχνικές για να εξισορροπήσουμε τα δεδομένα που ανήκουν στο training set.

Όλοι οι δείκτες κινδύνου που αφορούν τα προϊόντα δημιουργήθηκαν βάση υπολογισμών και μεθόδων, οι οποίοι δεν μας είναι γνωστοί. Μπορούμε να εικάσουμε ότι τέτοιοι δείκτες κινδύνου θα διαφέρουν σε μεγάλο βαθμό ανάλογα με την επιχείρηση και τις διαδικασίες που ακολουθεί. Τα δεδομένα είναι ήδη χωρισμένα σε train και test data.

Correlation μεταξύ διαφόρων χαρακτηριστικών

Το correlation matrix είναι ένα πολύ χρήσιμο εργαλείο που μπορεί να μας βοηθήσει να εντοπίσουμε τις συσχετίσεις που παρατηρούνται μεταξύ των χαρακτηριστικών των δεδομένων. Όσο ο αριθμός πλησιάζει το 1, τόσο μεγαλύτερη είναι η συσχέτιση μεταξύ των μεταβλητών.

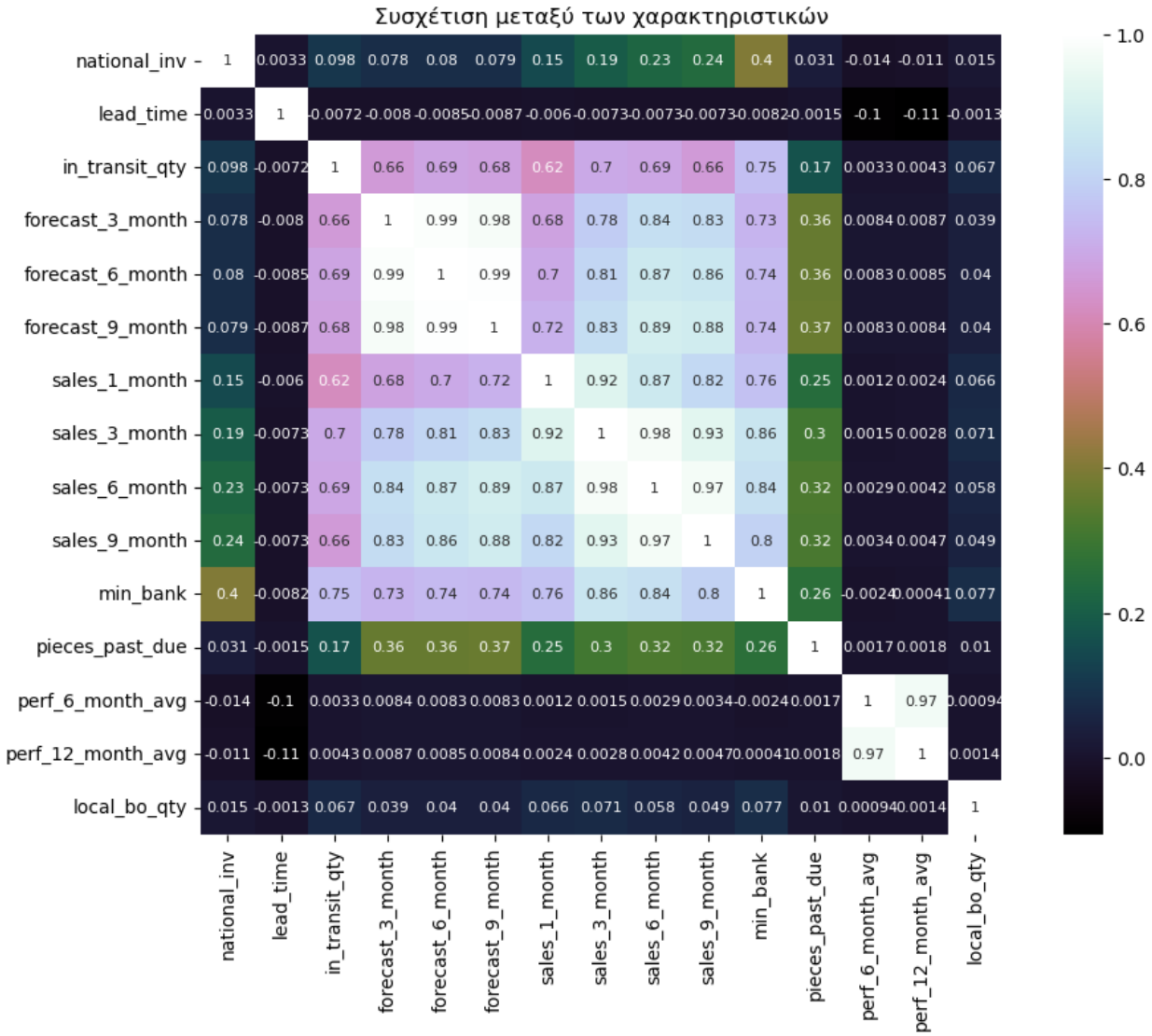


Figure 6: Correlation Matrix

Από τον πίνακα μπορούμε να κάνουμε αρκετές παρατηρήσεις σχετικά με τα χαρακτηριστικά. Μερικές από τις σημαντικές είναι:

1. Όλα τα χαρακτηριστικά των προβλέψεων έχουν μια πολύ μεγάλη συσχέτιση μεταξύ τους (0.98, 0.99). Το ίδιο ακριβώς ισχύει και για τις πωλήσεις. (> 0.82). Παρατηρούμε αρκετά μεγάλο συσχετισμό μεταξύ των προβλέψεων και των πωλήσεων, καθώς οι πωλήσεις στο παρελθόν, ανάλογα με το εάν είναι μεγάλες η μικρές, θα επηρεάσουν τις μελλοντικές προβλέψεις.
2. Η απόδοση του προϊόντος στους 6 και 12 μήνες παρουσιάζει υπερβολικά μεγάλη συσχέτιση (0.97).
3. Όλες οι σημαντικές συσχετίσεις που παρατηρήθηκαν είναι θετικές.
4. Η στήλη min_bank συσχετίζεται σε μεγάλο βαθμό με τις στήλες πωλήσεων και προβλέψεων καθώς το απόθεμα στις αποθήκες είναι ευθέως ανάλογο με τις πωλήσεις.
5. Η στήλη in_transit_qty συσχετίζεται σε μεγάλο βαθμό με τις στήλες των πωλήσεων, προβλέψεων και min_bank. Αυτό είναι λογικό επειδή υψηλές πωλήσεις ενός προϊόντος σημαίνουν περισσότερη ποσότητα από αυτό το προϊόν σε μεταφορά για αναπλήρωση του αποθέματος.
6. Η στήλη pieces_past_due παρουσιάζει μία ήπια συσχέτιση με τις στήλες πωλήσεων και προβλέψεων.
7. Η στήλη national_inv συσχετίζεται ήπια με την min_bank και σε μικρό βαθμό με τις πωλήσεις.

Καθώς πολλά χαρακτηριστικά συσχετίζονται, τα γραμμικά μοντέλα όπως η λογιστική παλινδρόμηση, το Linear SVM, κ.α. ενδέχεται να μην αποδίδουν καλά καθώς οι συντελεστές διαχωρισμού αλλάζουν. Ελέγχοντας την τιμή VIF (Variance Inflation Factor) μεταξύ των συσχετισμένων χαρακτηριστικών μπορούμε να αφαιρέσουμε περιττά χαρακτηριστικά εάν χρειάζεται ή χρησιμοποιώντας PCA μπορούμε να μειώσουμε τις διαστάσεις.

Ο παράγοντας πληθωρισμού διακύμανσης (VIF) ποσοτικοποιεί την έκταση της συσχέτισης μεταξύ ενός προγνωστικού παράγοντα και των άλλων προγνωστικών παραγόντων σε ένα μοντέλο. Χρησιμοποιείται για τη διάγνωση συγγραμμικότητας/πολυγραμμικότητας. Οι υψηλότερες τιμές υποδηλώνουν ότι είναι δύσκολο έως αδύνατο να εκτιμηθεί με ακρίβεια η συμβολή των προγνωστικών παραγόντων σε ένα μοντέλο.

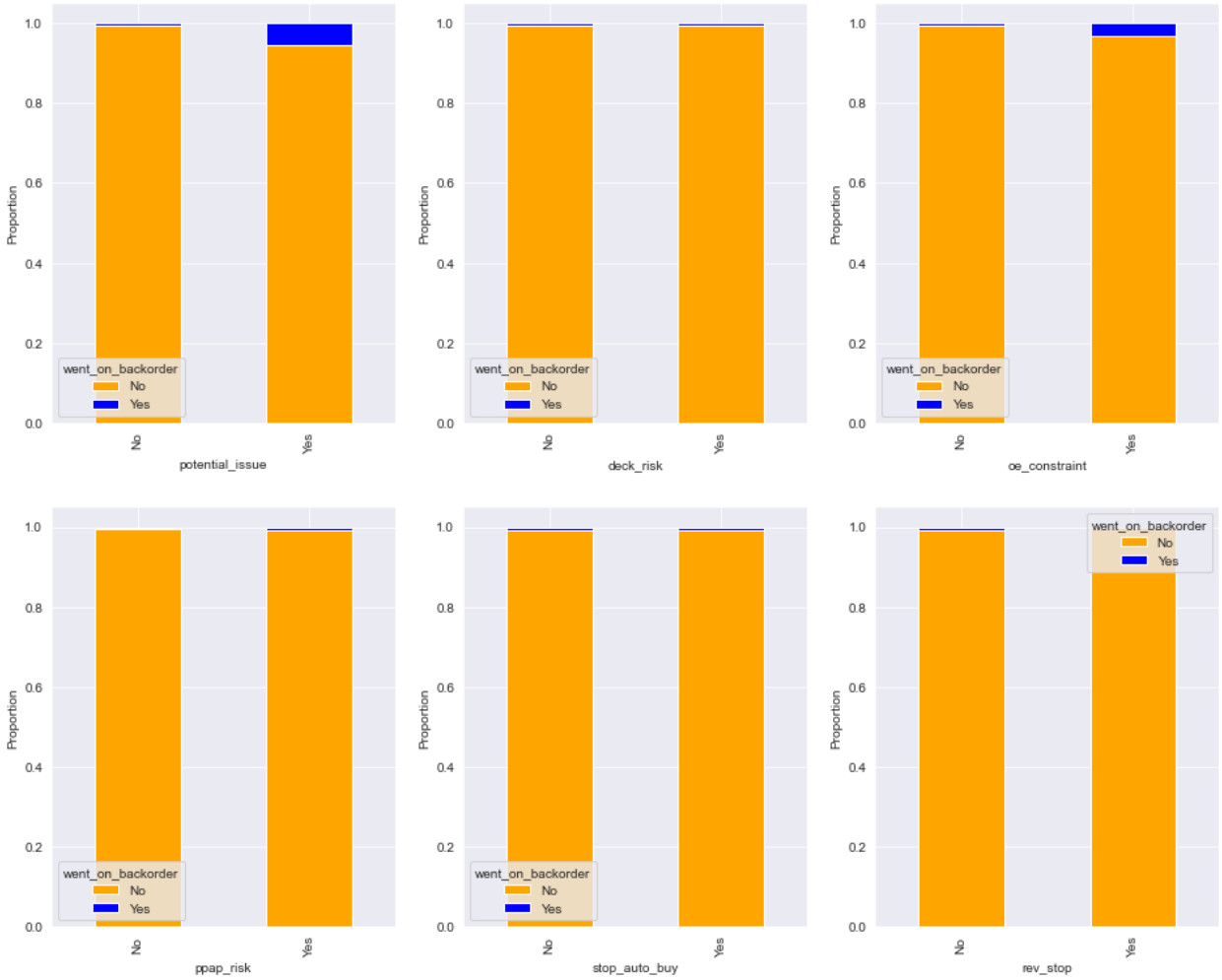
Τρέχοντας το Chi Square Test για τα κατηγορικά χαρακτηριστικά, διαπιστώνουμε ότι όλες οι κατηγορικές μεταβλητές παρουσιάζουν συσχέτιση με την μεταβλητή στόχο μας, went_on_backorder. Αυτό το τεστ ελέγχει στην ουσία την ανεξαρτησία μεταξύ των δύο κατηγορικών μεταβλητών, κοιτάζοντας για την τιμή p-value και την αληθότητα του null hypothesis. Ελέγχουμε την υπόθεση null χρησιμοποιώντας την p-value. Παράγοντας σημασίας ονομάζεται η τιμή Alpha. Συνήθως, επιλέγεται η τιμή alpha = 0,05. Εάν η τιμή p είναι μεγαλύτερη από την alpha, τότε ισχύει η null hypothesis. Στην περίπτωσή μας έχουμε απορρίψει όλες τις null hypothesis.

Όταν έχουμε Null Hypothesis δεν υπάρχει συσχέτιση μεταξύ του κατηγορικού χαρακτηριστικού και της μεταβλητής στόχου και όταν έχουμε Alternate Hypothesis υπάρχει συσχέτιση μεταξύ του χαρακτηριστικού και της μεταβλητής στόχου.

Barplots

Τα Barplots παρουσιάζουν κατηγορικά δεδομένα και δεν μπορούν να χρησιμοποιηθούν για την εύρεση ακραίων τιμών ή τον έλεγχο της λοξότητας (skewness). Μερικές ενδιαφέρουσες πληροφορίες που μπούμε να αποκομίσουμε κάνοντας barplot για τις κατηγορικές μεταβλητές είναι:

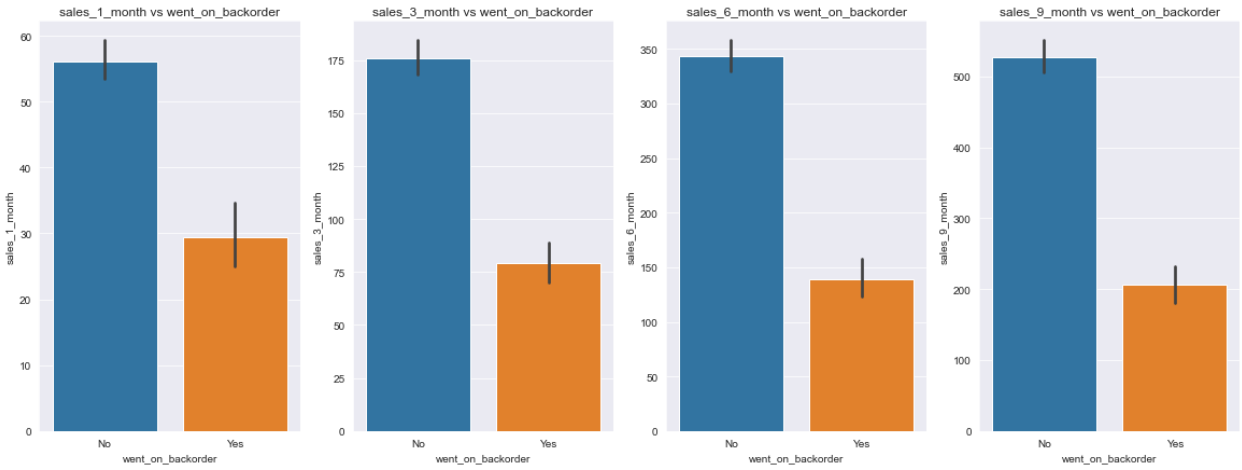
Figure 7: Barplots κατηγορικών χαρακτηριστικών



Εάν τα χαρακτηριστικά potential_issue και oe_cosntraint είναι «ναι», τότε υπάρχει μεγαλύτερη πιθανότητα το προϊόν να καταλήξει σε backorder. Σε αντίθεση, εάν η μεταβλητή rev_stop είναι «ναι», τότε δεν υπάρχει πιθανότητα η παραγγελία να μπει σε backorder. Οι υπόλοιπες κατηγορίες δεν έδειξαν μεγάλη επιρροή.

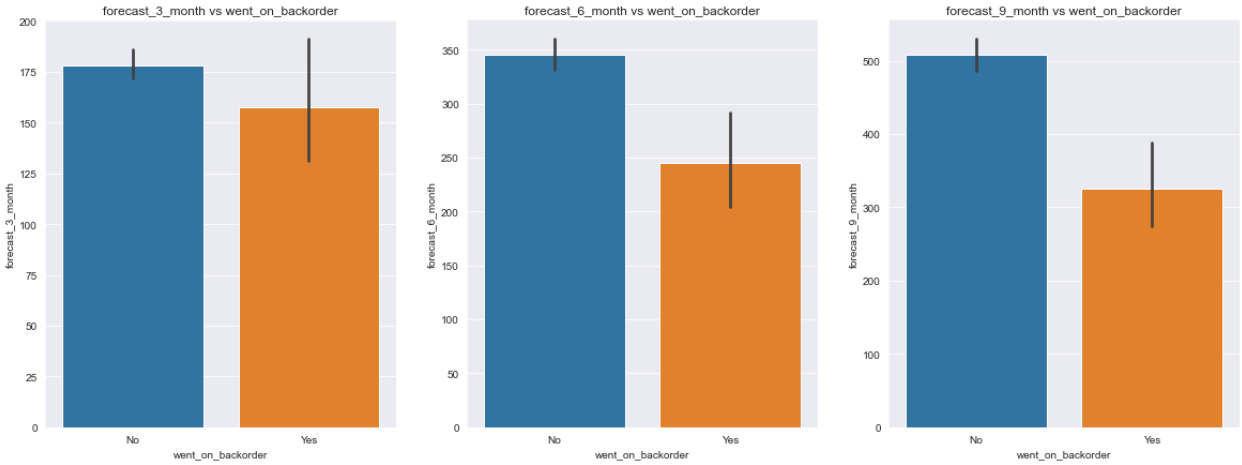
Τα ποσοστό των προϊόντων που κατέληξε σε backorder απεικονίζεται με μπλε.

Figure 8: Barplots Πωλήσεων



Από το σύνολο των barplots των πωλήσεων, καταλαβαίνουμε ότι ο μέσος αριθμός παραγγελιών που εισήχθησαν σε backorder στο πέρασμα των μηνών μειώνεται, όσο αυξάνεται ο αριθμός των παραγγελιών.

Figure 9: Barplots Προβλέψεων



Από το σύνολο των barplots των προβλέψεων, μπορούμε να πούμε ότι σε διάστημα 3, 6 και 9 μηνών, οι μέσες προβλεπόμενες πωλήσεις μειώνονται συνολικά για τη θετική κλάση ενώ φαίνεται να είναι σταθερές για την αρνητική κλάση.

Countplots

Τα countplots μοιάζουν με τα barplots, αλλά αντί για τον μέσο όρο μιας ποσοτικής μεταβλητής μεταξύ των παρατηρήσεων σε κάθε κατηγορία, δείχνουν τον αριθμό των παρατηρήσεων σε κάθε κατηγορία.

Figure 10: Countplots κατηγορικών χαρακτηριστικών

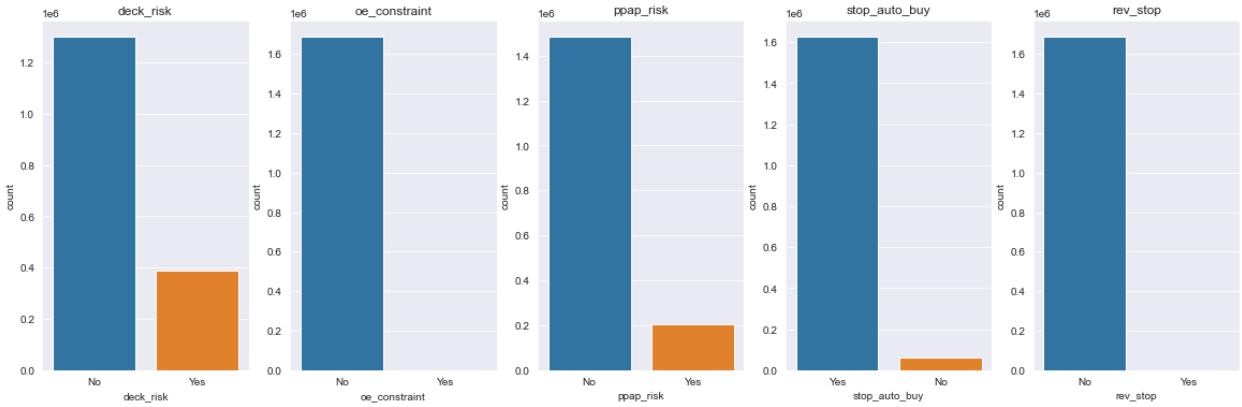
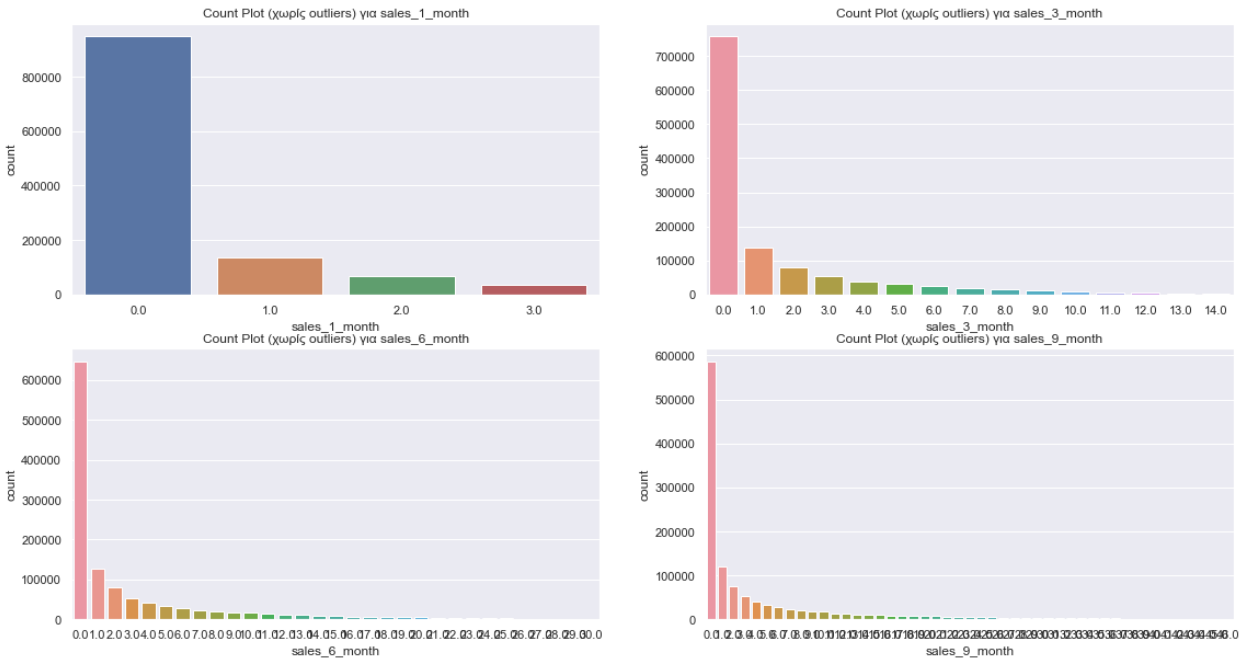


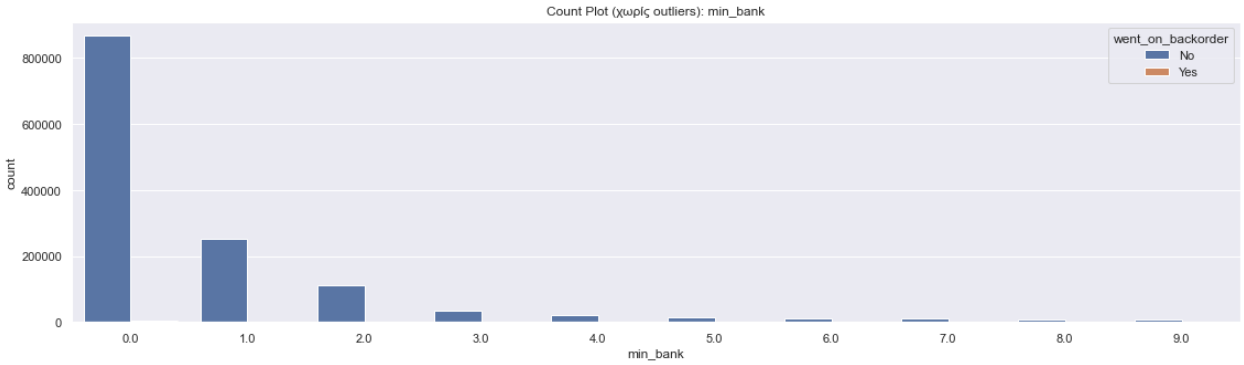
Figure 11: Countplots Πωλήσεων χωρίς outliers



Βλέπουμε ότι υπάρχουν πολλά προϊόντα με μηδενικό αριθμό μονάδων που πωλήθηκαν όλους τους προηγούμενους μήνες. Τα σημεία δεδομένων με πωλήσεις τουλάχιστον μίας μονάδας προϊόντος είναι περισσότερα σε σύγκριση με τα σημεία δεδομένων με τουλάχιστον 3 πωλήσεις για όλα τα χαρακτηριστικά.

Καθώς εξετάζουμε την ποσότητα πωλήσεων τους προηγούμενους 9 μήνες, βλέπουμε ότι ο αριθμός των μονάδων που πωλήθηκαν είναι μεγαλύτερος από την ποσότητα πωλήσεων για τους προηγούμενους 3 ή 6 μήνες, που είναι ιδανικό.

Figure 12: Countplot min_bank χωρίς outliers

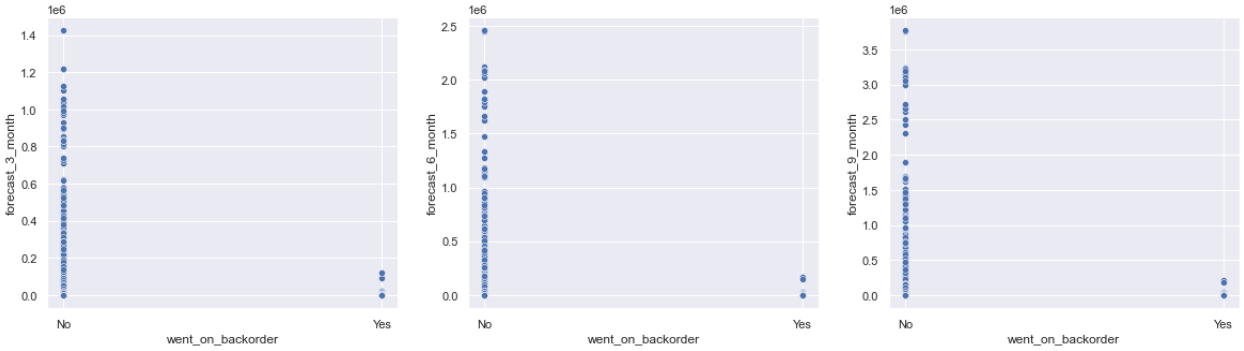


Από την γραφική παράσταση countplot του min_bank, μπορούμε να επιβεβαιώσουμε ότι οι περισσότερες από τις τιμές τείνουν να είναι μηδέν και υπάρχουν πολύ λιγότερα σημεία δεδομένων με τιμή μεγαλύτερη ή ίση του 3.

Scatterplots

Μπορούν να δείξουν συσχέτιση ή σχέση μεταξύ 2 μεταβλητών και να βοηθήσουν στον εντοπισμό των ακραίων τιμών.

Figure 13: Scatterplots Προβλέψεων



Από ότι βλέπουμε έχουμε backorder μόνο όταν η πρόβλεψη είναι κάτω από 0,5 και αυτό ισχύει και για τις τρεις μεταβλητές προβλέψεων.

Boxplots

Μια γραφική παράσταση box and whiskers ή αλλιώς boxplot εμφανίζει τη σύνοψη πέντε αριθμών ενός συνόλου δεδομένων. Η σύνοψη πέντε αριθμών είναι το ελάχιστο, πρώτο τεταρτημόριο, διάμεσος, τρίτο τεταρτημόριο και μέγιστο. Σε ένα boxplot, σχεδιάζουμε ένα πλαίσιο από το πρώτο τεταρτημόριο στο τρίτο τεταρτημόριο. Μια κατακόρυφη γραμμή περνά μέσα από το πλαίσιο στη μέση.

Figure 14: Boxplots χαρακτηριστικών

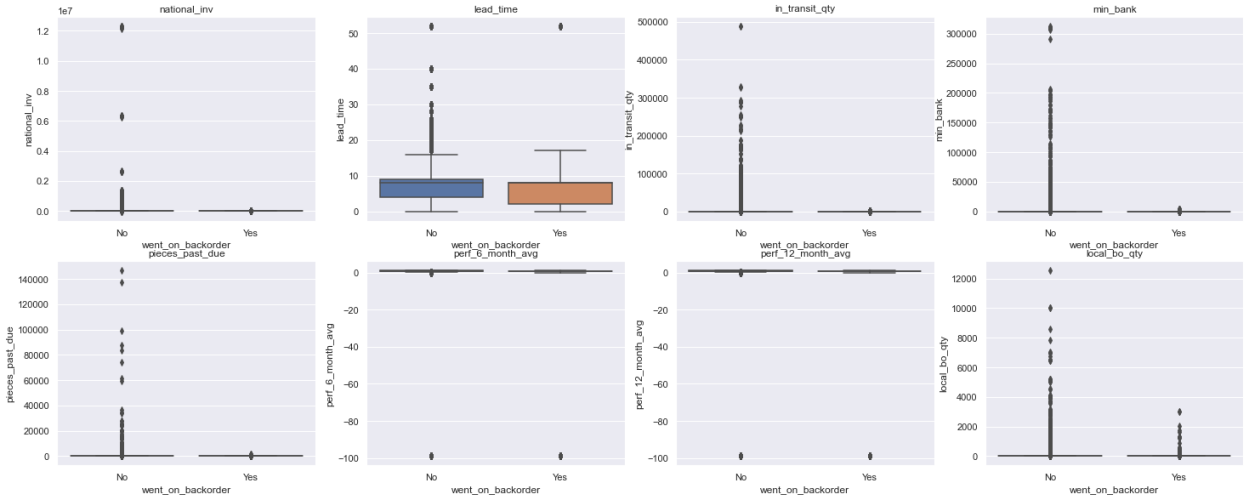
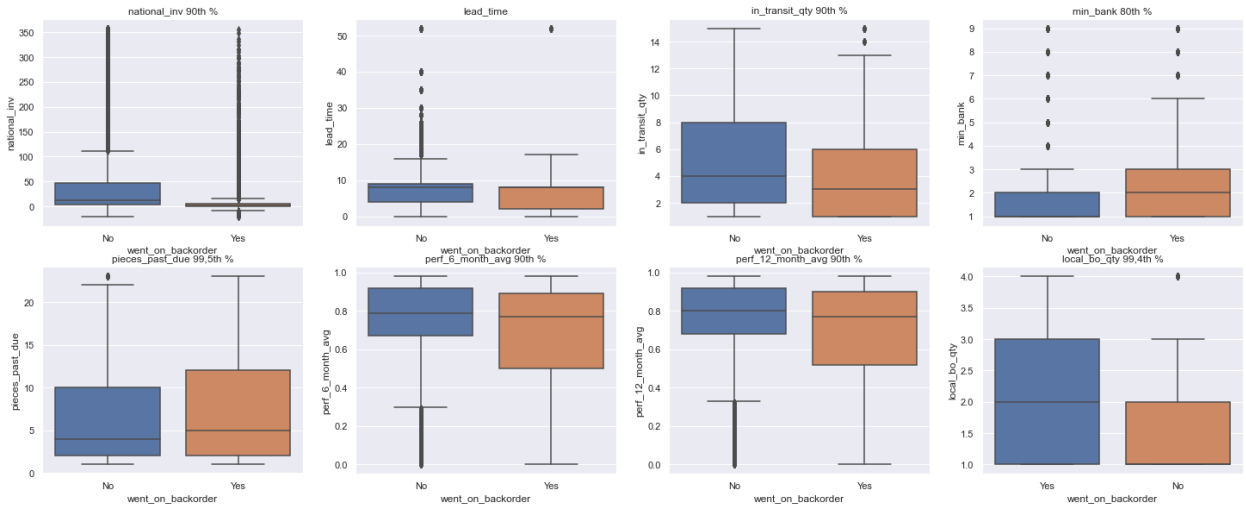


Figure 15: Boxplots χαρακτηριστικών από 0 έως 90% εκατοστιαίας τιμής



- national_inv: Η εξάπλωση (spread) είναι πολύ μεγάλη για το χαρακτηριστικό και το εύρος IQR είναι πολύ μικρό. Τα εύρη IQR και για τις δύο κλάσεις επικαλύπτονται (overlap), επομένως η χρήση αυτού του χαρακτηριστικού μεμονωμένα δεν θα είναι πολύ χρήσιμη για τον διαχωρισμό κλάσεων. Όταν κάνουμε boxplots λαμβάνοντας υπόψη μόνο τις τιμές μεταξύ του 0,1ου και του 90ου εκατοστημόριου του national_inv, βλέπουμε ότι η πλειονότητα των ακραίων τιμών στη θετική πλευρά αφαιρείται και η

μέγιστη τιμή μειώθηκε από 10^7 σε εύρος 350. Επίσης, τα IQR και των δύο κατηγοριών δεν κάνουν overlap όπως πριν. Με λογαριθμική μετατροπή θα μπορούσαμε να δούμε το χαρακτηριστικό σε μεγαλύτερη λεπτομέρεια.

- lead_time: Υπάρχουν μερικές ακραίες τιμές στο χαρακτηριστικό, ειδικά για την αρνητική κλάση. Το IQR και των δύο κλάσεων κάνει overlap σε κάποιο βαθμό. Το χαρακτηριστικό είναι εξαιρετικά λοξό (skewed) προς τη θετική πλευρά. Επίσης, βλέπουμε ότι δεν υπάρχει διακριτή διάμεσος για τη θετική κλάση. (Επομένως αυτό το χαρακτηριστικό μπορεί να είναι χρήσιμο για τον διαχωρισμό και των δύο κλάσεων τουλάχιστον σε εκείνες τις περιπτώσεις όπου ο χρόνος παράδοσης είναι μικρότερος από την τιμή του 25ου εκατοστημόριου της αρνητικής κλάσης.) Η διάμεσος είναι πιο κοντά στην τιμή του 75ου εκατοστημόριου.

- in_transit_qty: Η κατανομή είναι παρόμοια με του national_inv. Το IQR είναι πολύ μικρό και υπάρχουν πολλά outliers στις παραγγελίες που δεν πήγαν σε backorder. Μετά την αφαίρεση των τιμών που είναι μεγαλύτερες από την τιμή του 90ου τεταρτημόριου και μικρότερες από το 0ο τεταρτημόριο, βλέπουμε ότι οι τιμές μειώθηκαν από 5*10^5 σε 16. Το IQR και των δύο κλάσεων κάνει overlap σε κάποιο βαθμό. Μετά την αφαίρεση των ακραίων τιμών, για πολλά προϊόντα των οποίων η τιμή είναι μικρότερη από 2, πήγαν σε backorder και αυτών που είναι μεγαλύτερη από 6, δεν πήγαν σε backorder.

- min_bank: Το χαρακτηριστικό έχει δεξί skewness, δηλαδή τα δεδομένα βρίσκονται με υψηλή διαφορά μετά την τιμή του 75ου εκατοστημόριου. Η θετική κλάση (went_on_backorder="Yes") δεν εμφανίζει κανένα skewness μετά το 75ο εκατοστημόριο. Εάν είναι υψηλή η τιμή, υπάρχει μεγάλη πιθανότητα αυτό το προϊόν να μην πάει σε backorder. Εάν λάβουμε υπόψη μόνο τιμές μικρότερες από την τιμή του 80ου τεταρτημόριου, τότε εάν min_bank > 2, υπάρχει πολύ μεγάλη πιθανότητα το προϊόν να μπει σε backorder.

- pieces_past_due: Τα εύρη IQR και των δύο κατηγοριών κάνουν overlap για αυτό το χαρακτηριστικό. Η διαφορά είναι πολύ υψηλή για την αρνητική κλάση μετά το 75ο εκατοστημόριο και μόνο το 1% των τιμών έχει μη μηδενική τιμή.

- perf_6_month_avg, perf_12_month_avg: Η υπόθεση που κάνουμε στην συνέχεια ότι το -99 αντιπροσωπεύει μια τιμή που λείπει για τις στήλες απόδοσης φαίνεται να είναι αληθινή καθώς δεν υπάρχουν σημαντικές αρνητικές τιμές στο διάγραμμα. Κατά την επανασχεδίαση του boxplot για τιμές μεταξύ 0ου και 90ου τεταρτημόριου, τα στοιχεία που είχαν χαμηλή απόδοση τους τελευταίους 6 και 12 μήνες πήγαν σε backorder.

- local_bo_qty: Η πλειονότητα των σημείων δεδομένων είναι στο μηδέν. Για να βρούμε τις ακριβείς τιμές, έχουμε υπολογίσει τα εκατοστημόρια. Βλέπουμε ότι το 98% των σημείων δεδομένων είναι ίσο με μηδέν και το 99% του σημείου δεδομένων είναι μικρότερο ή ίσο με 1. Άρα, το χαρακτηριστικό μπορεί να θεωρηθεί ως outlier από μόνο του.

Figure 16: Boxplots για Πωλήσεις

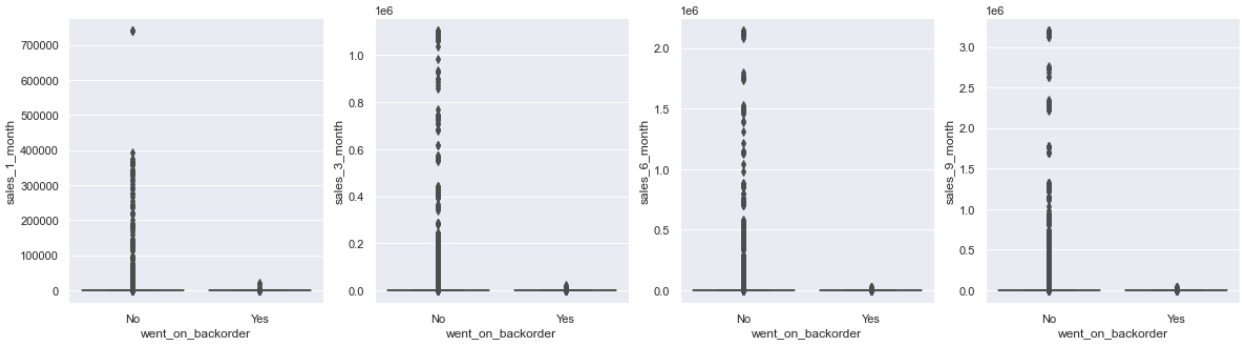


Figure 17: Boxplots για Πωλήσεις από 0 έως 90% εκατοστιαίας τιμής

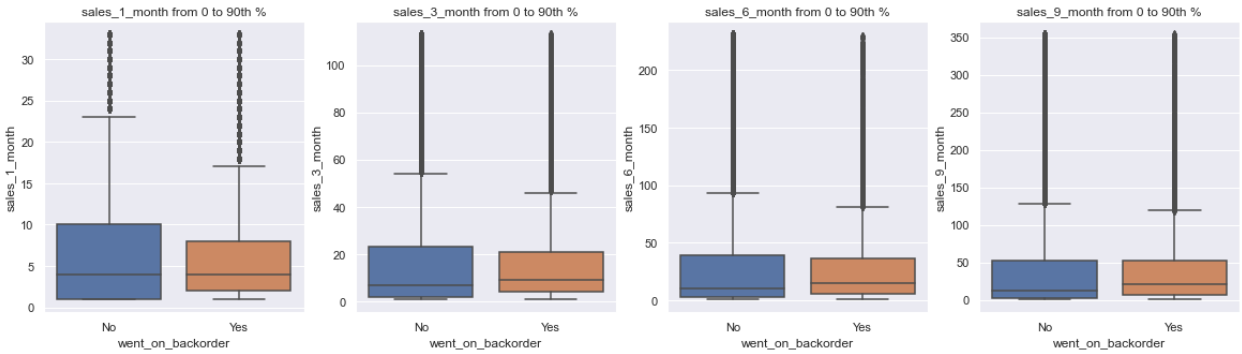
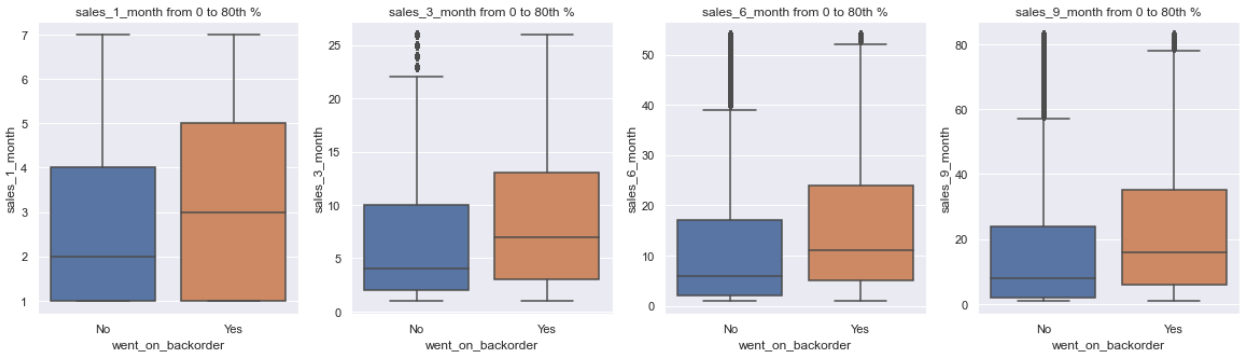


Figure 18: Boxplots για Πωλήσεις από 0 έως 80% εκατοστιαίας τιμής



Τα Boxplots μας δείχνουν ξεκάθαρα ότι τα backorder συνέβησαν περισσότερο όταν οι πωλήσεις ήταν υψηλές. Όπως είναι λογικό, οι πωλήσεις είναι ένα από τα πιο σημαντικά χαρακτηριστικά που επηρεάζουν την πρόβλεψη των backorders.

Τα boxplots πωλήσεων είναι παρόμοια με αυτά των προβλέψεων. Οι τιμές τους είναι skewed από δεξιά. Ακόμη και μετά την αφαίρεση τιμών μεγαλύτερων από του 90ου τεταρτημόριου, τα IQR κάνουν overlap πλήρως με πολλές τιμές, μετά το whisker που αντιπροσωπεύει τη μέγιστη τιμή. Έτσι, λαμβάνουμε υπόψη μόνο τις τιμές μέχρι το 80ο τεταρτημόριο σε νέα boxplots.

Figure 19: Boxplots για Προβλέψεις

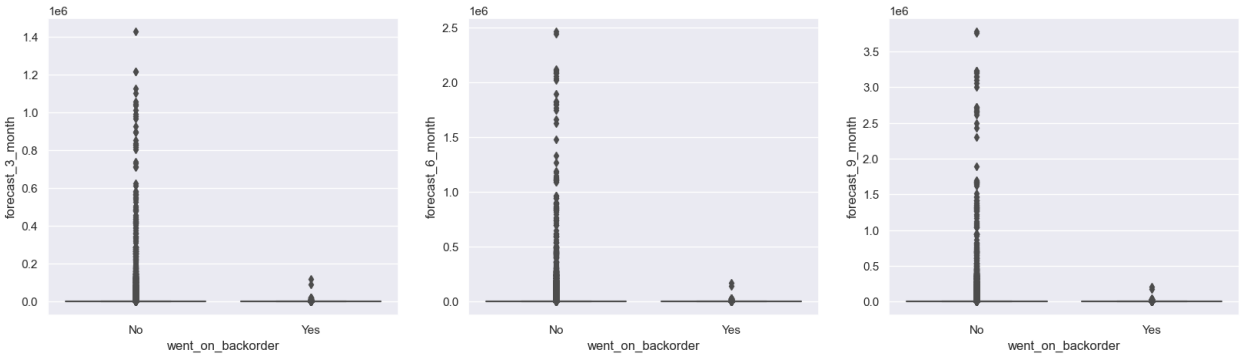


Figure 20: Boxplots για Προβλέψεις από 0 έως 90% εκατοστιαίας τιμής

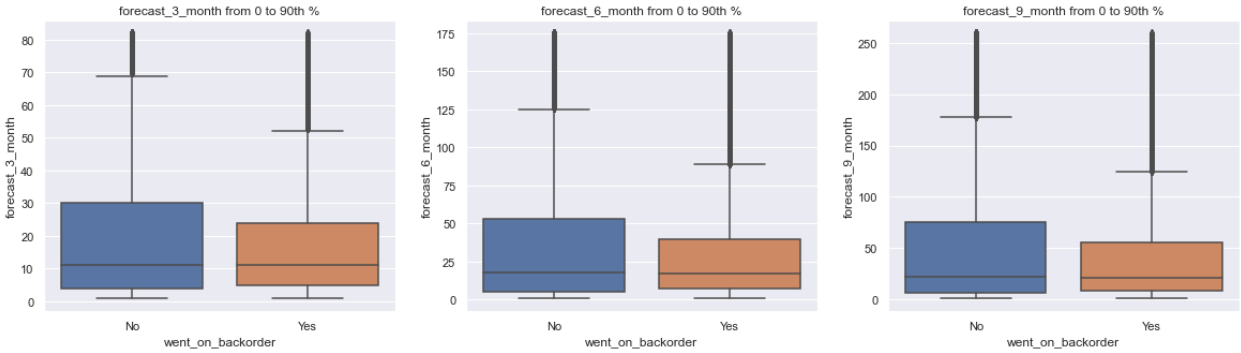
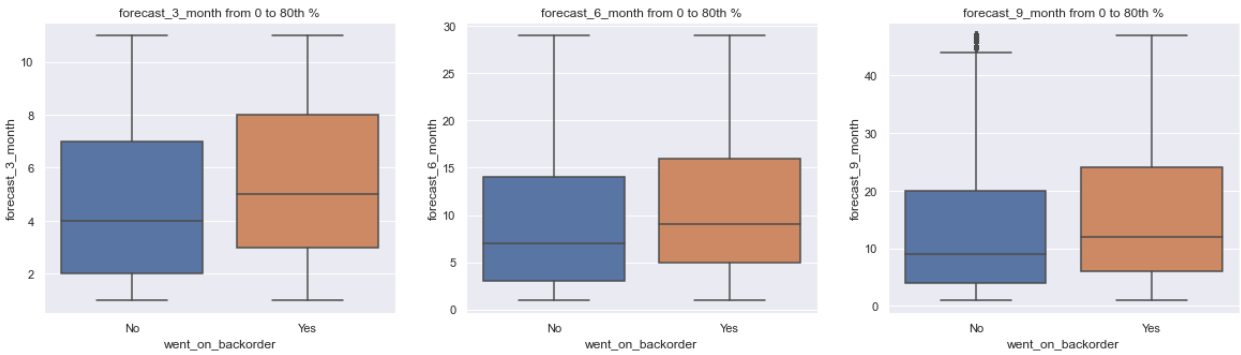


Figure 21: Boxplots για Προβλέψεις από 0 έως 80% εκατοστιαίας τιμής



Η κατανομή και των 3 χαρακτηριστικών πρόβλεψης είναι παρόμοια. Οι ακραίες τιμές υπάρχουν και στα 3 χαρακτηριστικά, ειδικά για την αρνητική κλάση, και το ποσό τους αυξάνεται με την περίοδο πρόβλεψης. Ακόμη και μετά την κατάργηση των τιμών που είναι πάνω από το 90ο τεταρτημόριο και κάτω από του 0ο τεταρτημόριο, τα IQR κάνουν overlap. Καθώς οι τιμές υπήρχαν μετά από τα whiskers στην γραφική παράσταση boxplot των τιμών 0-90ου εκατοστημόριου, αν λάβουμε υπόψη μόνο το 0-80ο εκατοστημόριο, τότε παρατηρούμε ότι εάν η πρόβλεψη είναι υψηλότερη τότε υπάρχει μεγαλύτερη πιθανότητα να υπάρξει backorder.

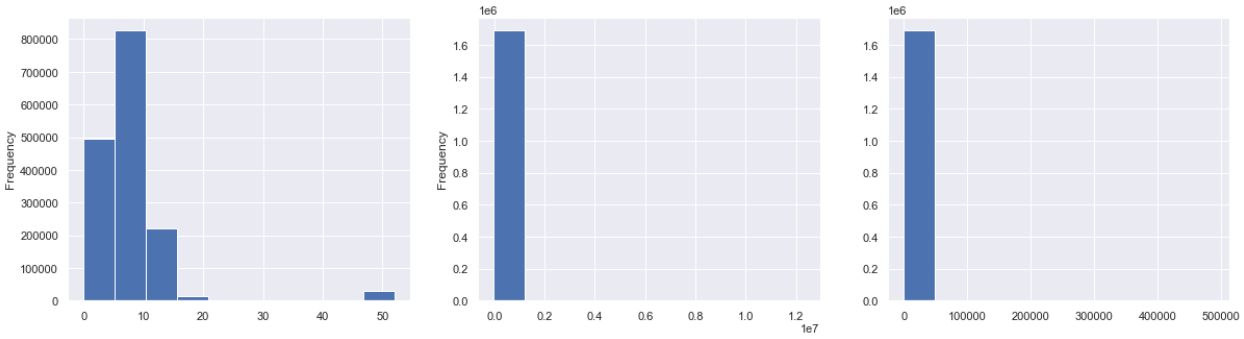
Γενικά συμπεράσματα των boxplots:

- Αρκετοί προγνωστικοί παράγοντες είναι skewed ή έχουν τεράστιες ακραίες τιμές.
- Οι ποσότητες προϊόντων (απόθεμα, πωλήσεις κ.λπ.) μπορεί να είναι σε πολύ διαφορετικές κλίμακες.
- Τα backorders κατά μέσο όρο σχετίζονται με: χαμηλά αποθέματα, υψηλές προβλέψεις πωλήσεων, υψηλό ιστορικό πωλήσεων, πιο συχνά potential risks.
- Αρκετοί προγνωστικοί παράγοντες συσχετίζονται σε μεγάλο βαθμό.

Histograms

Η βασική ιδέα του Ιστογράμματος είναι να δείξει την κατανομή συχνότητας του συνόλου των συνεχών σημείων που θα μας βοηθούσε να κατανοήσουμε την κατανομή των σημείων. Μπορούμε ακόμη να το βάλουμε ως plot για να ελέγξουμε για ακραίες τιμές, την λοξότητα και την κατανομή.

Figure 22: Histograms για lead time (1), national inv, in transit qty



Το μεγαλύτερο μέρος του χρόνου παράδοσης (lead_time) είναι μεταξύ 0 και 15, επομένως οι τιμές που λείπουν μπορούν να ληφθούν ως διάμεσος του χαρακτηριστικού, 0 αν θεωρήσουμε ότι δεν υπήρχε (ήταν ελάχιστος) ή να αφαιρεθούν εντελώς.

Προεπεξεργασία Δεδομένων (Pre-Processing)

Η προεπεξεργασία δεδομένων είναι απαραίτητη για τον μετασχηματισμό ακατέργαστων δεδομένων σε μια κατανοητή μορφή ώστε να μπορούν να χρησιμοποιηθούν για πρόβλεψη. Τα βήματα προεπεξεργασίας που χρησιμοποιούνται περιλαμβάνουν την διαχείριση των τιμών που λείπουν, την μετατροπή χαρακτηριστικών (feature conversion) και την διαχείριση των ακραίων τιμών.

Αφαίρεση Στηλών – Χαρακτηριστικών

Η στήλη 'sku' είναι το αναγνωριστικό και η 'went_on_backorder' είναι η ετικέτα κλάσης. Επομένως, θα τις αφαιρέσουμε και τις δύο.

Ελλιπείς Τιμές

Αρχικά, όλα τα χαρακτηριστικά της τελευταίας γραμμής των datasets αποτελούνται από nan τιμές. Επομένως αφαιρείται εξολοκλήρου.

Η στήλη με τον χρόνο παράδοσης αποτελείται από 100894 ελλιπείς τιμές. Για να αντιμετωπίσουμε το πρόβλημα των τιμών που λείπουν έχουμε δύο επιλογές. Πρώτον, μπορούμε να αφαιρέσουμε τελείως το χαρακτηριστικό εάν δεν μας είναι χρήσιμο ή να διαγράψουμε τις γραμμές στις οποίες εμφανίζεται η ελλιπής τιμή. Έτσι βέβαια χάνουμε πολύτιμα δεδομένα από τα υπόλοιπα χαρακτηριστικά. Δεύτερων, μας δίνεται η δυνατότητα να αντικαταστήσουμε την κενή τιμή με την μέση τιμή των υπολοίπων παρατηρήσεων του χρόνου παράδοσης, την διάμεσο, με 0 ή με κάποιον αριθμό που δεν θα εμφανίζεται στα δεδομένα (για να αναγνωρίζουν οι αλγόριθμοι τα δεδομένα ως διαφορετικά).

Στην δεδομένη περίπτωση επιλέγουμε να συμπληρώσουμε τις τιμές με 0. Για την ακρίβεια, για τον χρόνο παράδοσης (lead time), οι ελλιπείς τιμές είναι πιο πιθανό να αντιπροσωπεύουν την απουσία του, δηλαδή να μην υπάρχει χρόνος παράδοσης. Άρα η αντικατάσταση τιμών με 0 είναι μία καλή στρατηγική.

Outliers

Είναι ένα αντικείμενο ή μία παρατήρηση που αποκλίνει σημαντικά από τις υπόλοιπες. Μπορούν να προκληθούν από σφάλμα μέτρησης ή εκτέλεσης. Η ανάλυση των outlier αναφέρεται ως ανάλυση ακραίων τιμών ή εξόρυξη ακραίων τιμών. Οι περισσότερες μέθοδοι εξόρυξης δεδομένων απορρίπτουν τον ακραίο θόρυβο ή τις εξαιρέσεις. Ωστόσο, σε ορισμένες εφαρμογές όπως η ανίχνευση απάτης, τα σπάνια συμβάντα μπορεί να είναι πιο ενδιαφέροντα από τα πιο τακτικά συμβάντα και ως εκ τούτου, η ανάλυση ακραίων τιμών γίνεται σημαντική σε αυτήν την περίπτωση.

Πολλά από τα συνεχή χαρακτηριστικά (διαθέσιμο απόθεμα, κ.λπ.) έχουν πολύ λοξές κατανομές (skewed distributions), δηλαδή τα περισσότερα στοιχεία επικεντρώνονται γύρω από μια μικρή τιμή, αλλά ορισμένες τιμές σε ένα στοιχείο είναι πολύ μεγάλες. Θεωρείται ως καλή πρακτική η αντιμετώπιση αυτού του προβλήματος και ο διαχωρισμός των δεδομένων μέσω του log transformation των χαρακτηριστικών. Ο μετασχηματισμός log παίρνει τη μαθηματική τιμή εισόδου του x και την αντικαθιστά με την τιμή του log(x). Αυτός ο μετασχηματισμός είναι χρήσιμος σε συνεχή χαρακτηριστικά με δεδομένα που είναι

υπερβολικά right-skewed, καθώς μπορεί να αποδυναμώσει τα ακραία σημεία (την σημασία τους) και να ομαλοποιήσει την κατανομή των χαρακτηριστικών με μεγάλα εύρη. Είναι μια αποδεδειγμένα επιτυχής τεχνική έναντι των outliers και αυτήν θα επιλέξουμε για το 2^{ο} σετ δεδομένων μας.

Σε paper που έχει δημοσιεύσει η IBM για την "Defense Logistics Agency (DLA)" για την αντιμετώπιση του προβλήματος των backorder, χρησιμοποιείται η μέθοδος του log transformation.

Μερικές ακόμη τεχνικές που θα μπορούσαμε να χρησιμοποιήσουμε είναι:

Feature Clipping: Είναι η περικοπή των χαρακτηριστικών σε μία σταθερή αριθμητική τιμή (ή σταθερή τιμή +- 3 std (τυπική απόκλιση)). Αυτό θα οδηγούσε σε απώλεια πληροφοριών, αλλά θα καταπολεμούσε αποτελεσματικά την επίδραση των ακραίων τιμών.

Robust Scaler: Υπολογίζει τη διάμεσο (50ο εκατοστημόριο) και το 25ο και 75ο εκατοστημόριο. Στη συνέχεια, αφαιρεί την διάμεσο κάθε μεταβλητής από την τιμή της και διαιρεί με το διατεταρτημόριο (IQR) που είναι η διαφορά μεταξύ του 75ου και του 25ου εκατοστημόριου.

value = (value – median) / (p75 – p25)

Η προκύπτουσα μεταβλητή έχει μηδενικό μέσο και διάμεσο και τυπική απόκλιση 1, αν και δεν παραμορφώνεται από ακραίες τιμές, αυτές εξακολουθούν να υπάρχουν με τις ίδιες σχετικές σχέσεις με άλλες τιμές. Σε αντίθεση με άλλους scalers, τα στατιστικά στοιχεία κεντραρίσματος και κλιμάκωσης του Robust Scaler βασίζονται σε εκατοστημόρια και επομένως δεν επηρεάζονται από έναν μικρό αριθμό πολύ μεγάλων ακραίων τιμών. Αυτήν την τεχνική θα επιλέξουμε για το 1^{ο} σετ δεδομένων μας.

Τέλος, υπάρχει και η επιλογή αλγορίθμων για την ανάλυση που βασίζονται σε δέντρα ή νευρωνικά δίκτυα, τα οποία είναι περισσότερο ανθεκτικά σε ακραίες τιμές.

-99 Τιμές

Στις perf_6_month_avg και perf_12_month_avg στήλες παρατηρούμε αρνητική τιμή -99 και χαμηλότερη τιμή -0,99. Αυτές πρέπει να αντικατασταθούν. Οι τιμές -99 είναι πιο αντιμετωπίσιμες με ορισμένους αλγόριθμους, όπως τα δέντρα απόφασης και μοντέλα deep learning, επειδή μπορούν εύκολα να διαχωριστούν από τις άλλες αριθμητικές τιμές. Όμως με γραμμικά μοντέλα υπάρχει πρόβλημα. Παρόλα αυτά, εμείς θα τις θεωρήσουμε ως nan τιμές και θα αντικαταστήσουμε με την διάμεσο των χαρακτηριστικών από όπου προέρχονται. Μια άλλη λύση ήταν να τις αντικαταστήσουμε με 0.

Feature Engineering

Εδώ περιέχονται διαδικασίες όπως η μετατροπή χαρακτηριστικών (Feature Transformation), feature scaling, η κωδικοποίηση των labels κατηγορικών μεταβλητών, ο χωρισμός των δεδομένων σε train και test σετ δεδομένων, η μείωση διαστάσεων, κ.α.

Κατηγορικές Τιμές

Οι αλγόριθμοι μηχανικής μάθησης δεν είναι καλοί στον να διαχειρίζονται κατηγορικές τιμές. Επομένως, θα γίνει χρήση μιας κοινής τεχνικής για τη μετατροπή κατηγορηματικών τιμών σε δυαδική αριθμητική αναπαράσταση, η One-Hot encoding. Με αυτόν τον τρόπο θα είναι δυνατό για τον αλγόριθμο μηχανικής μάθησης να ερμηνεύει τις σχέσεις μεταξύ κώδικα και άλλων χαρακτηριστικών. Το 0 θα αντικαταστήσει το False και το 1 το True.

Χωρισμός δεδομένων σε train, test και validation σετ

Αρχικά θα ενώσουμε τα dataframes των δύο σετ δεδομένων (train και test), τα οποία έχουν υποστεί και τα 2 τις διαδικασίες προεπεξεργασίας. Στην συνέχεια, θα κάνουμε διαχωρισμό των δεδομένων σε 80% στο train σετ και 20% στο test σετ. Μετά, θα χωρίσουμε το train σετ σε 90% train σετ και σε 10% validation σετ. Θα χρειαστούμε ένα validation σετ για να βοηθήσουμε στη μοντελοποίηση και στην πιθανή βελτίωση των αποτελεσμάτων. Έτσι θα αφήσουμε το test σετ καθαρά για το evaluation των μοντέλων μας. Τέλος, θα αντιγράψουμε τα δεδομένα μας γιατί θα ακολουθήσουμε 2 διαφορετικές προσεγγίσεις στο κομμάτι του feature engineering για να δούμε πως θα συμπεριφερθούν τα μοντέλα που έχουμε επιλέξει και να συγκρίνουμε τα αποτελέσματα σε κάθε περίπτωση. Επιπλέον, θα δοκιμαστεί η PCA σε ξεχωριστό σετ.

Το PCA σετ δεδομένων θα ακολουθήσει την παρακάτω διαδικασία feature engineering:

Εφαρμογή Standard Scaler → Έλεγχος με PCA

Το πρώτο σετ δεδομένων θα ακολουθήσει την παρακάτω διαδικασία feature engineering:

Εφαρμογή Robust Scaler → Μοντέλα

Το δεύτερο σετ δεδομένων θα ακολουθήσει την παρακάτω διαδικασία feature engineering:

Εφαρμογή Log Transformation → Εφαρμογή Standard Scaler → Εφαρμογή Smote Tomek → Μοντέλα

Λογαριθμική Μετατροπή

Όπως αναφέραμε προηγουμένως, η λογαριθμική μετατροπή είναι ένας αποτελεσματικός τρόπος αντιμετώπισης των outliers, χωρίς να χρειάζεται να γίνει διαγραφή τιμών από το dataset. Θα την εφαρμόσουμε στο $2^{ο}$ σετ δεδομένων μας.

Feature Scaling

Ένα κανονικοποιημένο (normalized) σύνολο δεδομένων θα έχει πάντα τιμές που κυμαίνονται μεταξύ 0 και 1, ενώ ένα τυποποιημένο (standardized) σύνολο δεδομένων θα έχει μέσο όρο 0 και τυπική απόκλιση 1, αλλά δεν υπάρχει συγκεκριμένο άνω ή κάτω όριο για τις μέγιστες και ελάχιστες τιμές.

Συνήθως κανονικοποιούμε δεδομένα όταν εκτελούμε κάποιο είδος ανάλυσης στην οποία έχουμε πολλαπλές μεταβλητές που μετρούνται σε διαφορετικές κλίμακες και θέλουμε κάθε μία από τις μεταβλητές να έχει το ίδιο εύρος. Αυτό εμποδίζει μια μεταβλητή να έχει υπερβολική επιρροή, ειδικά αν μετράται σε διαφορετικές μονάδες. Στην περίπτωση των ακραίων τιμών, η τυποποίηση δεν βλάπτει τη θέση τους, ενώ η κανονικοποίηση συλλαμβάνει όλα τα σημεία δεδομένων στο εύρος τους.

Πολλές μέθοδοι μηχανικής μάθησης, όπως η PCA, είναι ευαίσθητοι σε ακραίες τιμές ή χαρακτηριστικά που βρίσκονται σε διαφορετική κλίμακα. Σε τέτοιες περιπτώσεις βοηθάει ιδιαίτερα η κανονικοποίηση. Συγκεκριμένα η PCA ενδείκνυται για χρήση έπειτα από εφαρμογή του standard scaler.

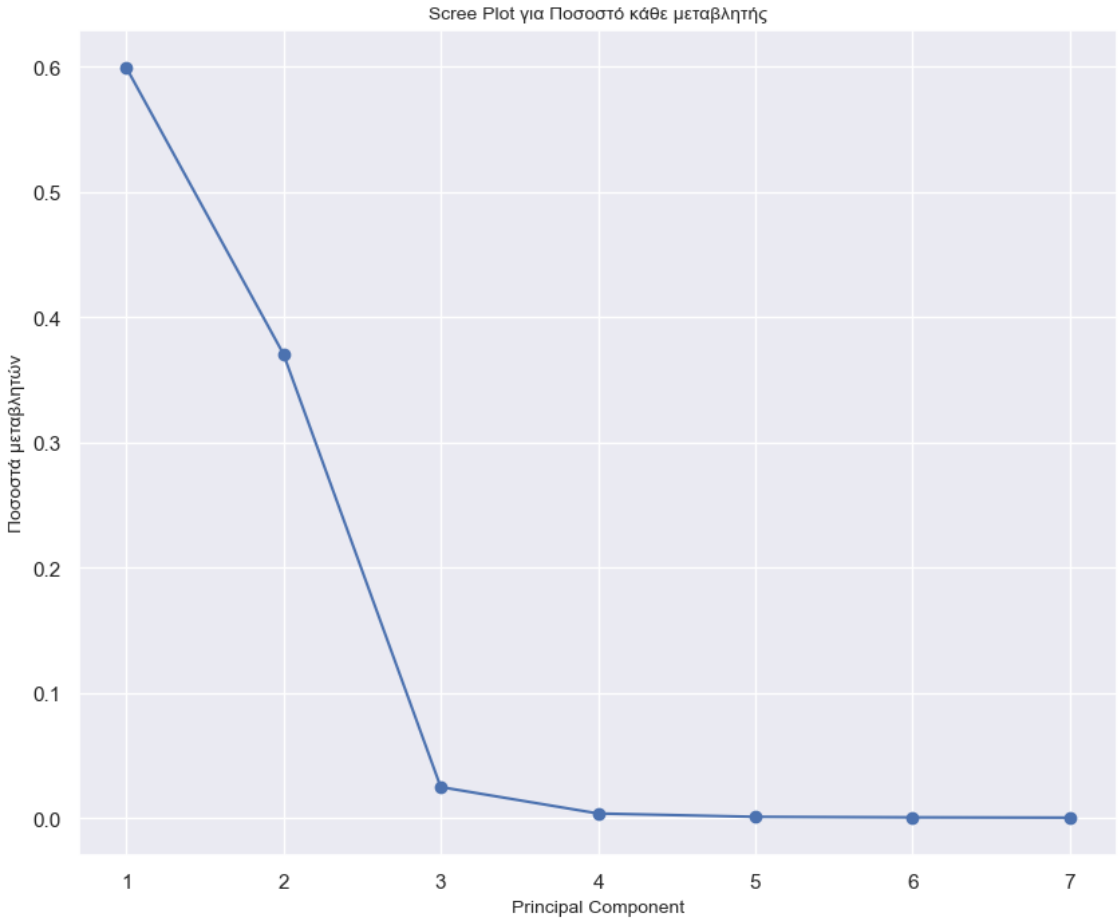
Θα εφαρμόσουμε τον Robust scaler στο πρώτο σετ δεδομένων για να αντιμετωπίσουμε τα outliers και τον Standard scaler στο δεύτερο σετ, καθώς τα outliers θα έχουν αντιμετωπισθεί από την λογαριθμική μετατροπή που θα έχει προηγηθεί.

Dimensionality Reduction

Μία από τις πιο αποτελεσματικές τεχνικές για να διαλέξουμε τα στοιχεία που επηρεάζουν πιο πολύ το μοντέλο μας και να μειώσουμε την πολυπλοκότητά του, κάνοντας ταυτόχρονα ευκολότερη και την οπτικοποίηση του, είναι η ανάλυση κυρίων συνιστωσών (Principal Component Analysis - PCA). Πρόκειται για μια στατιστική διαδικασία μείωσης διαστάσεων με την οποία αναπαριστούμε ένα πίνακα συνδιακύμανσης ενός συνόλου «αρχικών» μεταβλητών μέσα από ένα διαφορετικό (και συνήθως μικρότερο) σύνολο «νέων» μεταβλητών οι οποίες προκύπτουν από τον γραμμικό συνδυασμό των «αρχικών» μεταβλητών. Διατηρώντας τη βασική σημασία του συνόλου δεδομένων εξοικονομείται πολύς χρόνος κατά την εκπαίδευση των μοντέλων ταξινόμησης και κατά την πρόβλεψη. Συνήθως χρησιμοποιούμε την PCA όταν έχουμε μεταβλητές που παρουσιάζουν μεγάλη συσχέτιση μεταξύ τους. Λόγω της υψηλής διάστασης των δεδομένων και των συσχετίσεών τους στην περίπτωσή μας, μια μείωση διάστασης θα μπορούσε να φανεί χρήσιμη.

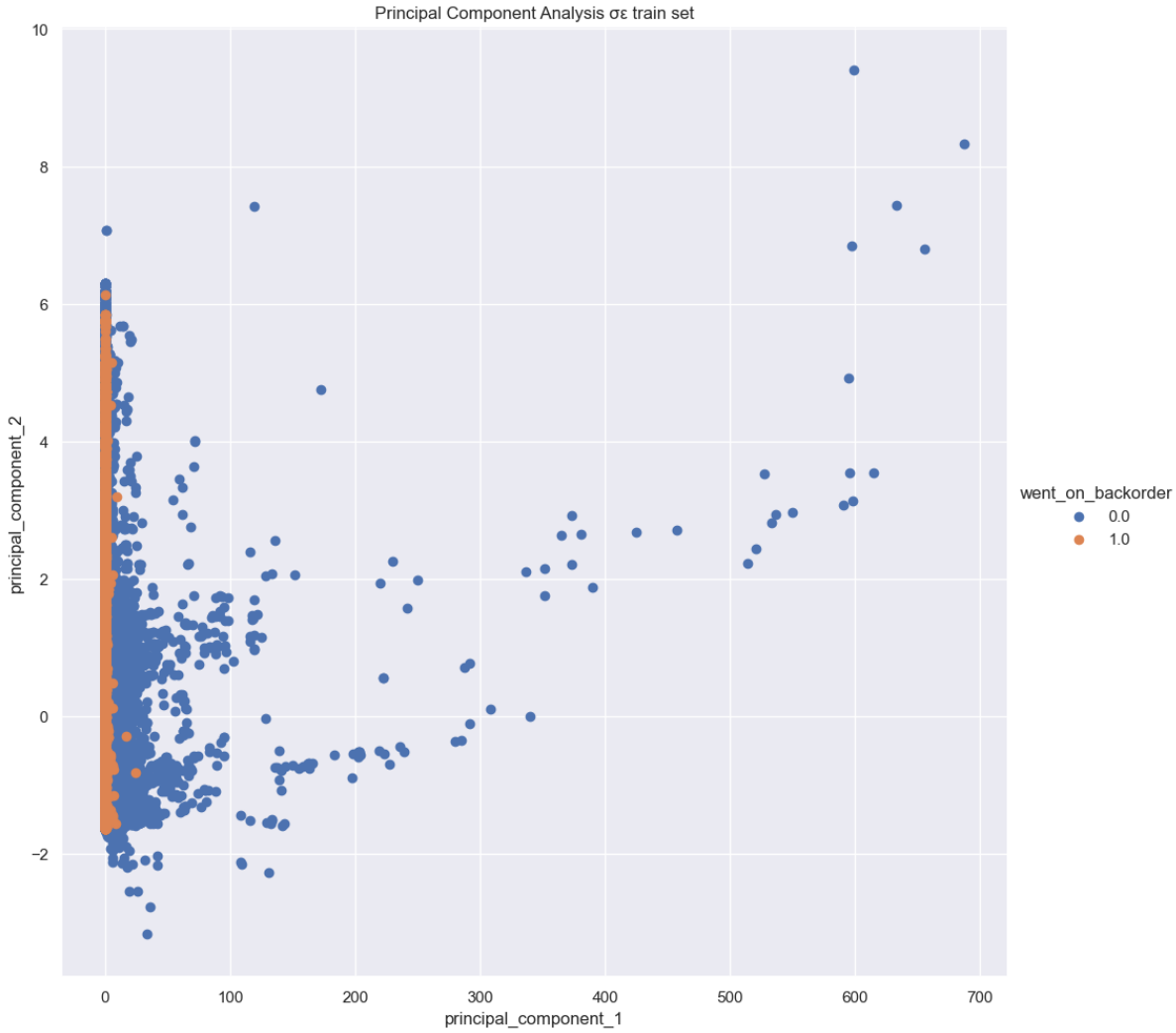
Εμείς θα δοκιμάσουμε την PCA στο πρώτο σετ δεδομένων, χωρίς να την εφαρμόσουμε τελικά, γιατί μας ενδιαφέρει το Feature Importance των δεδομένων.

Figure 23: Eigenvalues plot



Βλέπουμε από το σχήμα, το οποίο αποτυπώνει τα ποσοστά των μεταβλητών που αντιστοιχούν σε κάθε principal component (βάλαμε 7), ότι ο συνιστώμενος αριθμός χαρακτηριστικών για την PCA είναι 2, καθώς αιτιολογούν με διαφορά το μεγαλύτερο ποσοστό των μεταβλητών.
Τα eigenvalues μας πληροφορούν για τις κατευθύνσεις και το μέγεθος του spread των δεδομένων μας.

Figure 24: PCA σε data



Χρησιμοποιήσαμε τεχνική μείωσης διαστάσεων, σε αυτήν την περίπτωση την Principal Component Analysis, για να συλλάβουμε την ουσία των δεδομένων. Από την παραπάνω γραφική παράσταση βλέπουμε ότι τα περισσότερα από τα σημεία δεδομένων βρίσκονται κοντά στο 0. Αυτό το πόρισμα είναι αληθής επειδή έχουμε δει πολλά χαρακτηριστικά με ως επί το πλείστον τιμές 0 στην ανάλυση δεδομένων που κάναμε. Υπάρχουν outliers στα δεδομένα, αλλά αυτά τα σημεία δεδομένων δεν είναι αναγκαίο να είναι outliers. Επιπλέον, αυτές οι πιθανές ακραίες τιμές είναι περισσότερο της αρνητικής κλάσης σε σύγκριση με τη θετική κλάση. Για τη θετική κλάση, σχεδόν όλα τα σημεία δεδομένων βρίσκονται κοντά στο 0.

Βλέπουμε κάποιο διαχωρισμό και επίσης overlap μεταξύ της θετικής κλάσης και της αρνητικής κλάσης. Αυτό σημαίνει ότι το μοντέλο θα πρέπει να μπορεί να διακρίνει εύκολα μεταξύ ενός προϊόντος που πήγε σε backorder έναντι ενός προϊόντος που δεν πήγε.

Figure 25: Biplot σε data

Biplot είναι ένα βελτιωμένο διάγραμμα διασποράς που χρησιμοποιεί τόσο σημεία όσο και διανύσματα για να αναπαραστήσει τη δομή. Το συγκεκριμένο Biplot μας απεικονίζει την διασπορά των μεταβλητών στα PCA δεδομένα. Τα αρχικά δεδομένα αντιπροσωπεύονται από principal components που εξηγούν το μεγαλύτερο μέρος της διακύμανσης δεδομένων χρησιμοποιώντας τα loading vectors και PC scores.

Class Imbalance

Για να αντιμετωπίσουμε το πρόβλημα ανισορροπίας μεταξύ των δεδομένων έχουμε τις εξής επιλογές. Μπορούμε να χρησιμοποιήσουμε cost-sensitive learning, δηλαδή αλγορίθμους στους οποίους μπορούμε να τροποποιήσουμε τη συνάρτηση απώλειας (loss function) για να προστεθεί βάρος στην κλάση μειοψηφίας. Με τη μοντελοποίηση της συνάρτησης απώλειας για να ληφθεί υπόψη το μεταβλητό κόστος λανθασμένης ταξινόμησης, έχουμε πλέον έναν ταξινομητή με ευαισθησία στο κόστος (cost-sensitive).

Εκτός αυτού, μπορούμε να εφαρμόσουμε τεχνικές υπερδειγματοληψίας (oversampling), υποδειγματοληψίας (undersampling) ή και συνδυασμό των δύο. Οι μέθοδοι oversampling αντιγράφουν ή δημιουργούν νέα συνθετικά παραδείγματα στην κλάση μειοψηφίας (minority class), ενώ οι μέθοδοι undersampling διαγράφουν ή συγχωνεύουν παραδείγματα στην κλάση πλειοψηφίας (majority class). Και οι δύο τύποι επαναδειγματοληψίας μπορούν να είναι αποτελεσματικοί όταν χρησιμοποιούνται μεμονωμένα, αν και μπορεί να είναι πιο αποτελεσματικοί όταν χρησιμοποιούνται ταυτόχρονα.

Υπάρχουν αρκετοί μέθοδοι oversampling και undersampling. Ακολουθεί σύντομη παρουσίαση μερικών από τους πιο γνωστούς.

Μέθοδοι Oversampling

SMOTE: Δημιουργεί συνθετικές παρατηρήσεις της μικρής κλάσης κάνοντας εύρεση των k-πλησιέστερων γειτόνων (nearest neighbor) για παρατηρήσεις της μικρής κλάσης (εύρεση παρόμοιων παρατηρήσεων) και επιλέγει τυχαία έναν από τους k-πλησιέστερους γείτονες και τον χρησιμοποιεί για να δημιουργήσει μια παρόμοια, αλλά τυχαία τροποποιημένη νέα παρατήρηση.

ADASYN: Η προσέγγιση του αλγορίθμου είναι μια γενικευμένη μορφή του SMOTE, με στόχο την υπερδειγματοληψία της μειονοτικής κλάσης δημιουργώντας συνθετικά δεδομένα. Σε αντίθεση με τον SMOTE που δημιουργεί αυθαίρετο αριθμό συνθετικών μειονοτήτων δεδομένων για τη διόρθωση της ανισορροπίας στο σύνολο δεδομένων, ο ADASYN χρησιμοποιεί σταθμισμένη κατανομή για διαφορετικά παραδείγματα της μειονοτικής κλάσης σύμφωνα με το επίπεδο δυσκολίας τους στη μάθηση. Φτιάχνονται περισσότερα συνδετικά δεδομένα για τα παραδείγματα που μαθαίνουν πιο δύσκολα.

Μέθοδοι Undersampling

TomekLinks: Οι σύνδεσμοι Tomek είναι ζεύγη παραδειγμάτων αντίθετων κλάσεων σε κοντινή απόσταση. Αυτός ο αλγόριθμος, αφαιρεί το στοιχείο της πλειοψηφίας από τη σύνδεση Tomek, η οποία παρέχει ένα καλύτερο όριο απόφασης για τον ταξινομητή.

RUS: Είναι μια τεχνική υποδειγματοληψίας που περιλαμβάνει την τυχαία επιλογή ενός μικρού τμήματος της πλειοψηφική κλάσης χωρίς να αλλάζει τον αριθμό της μειονοτικής κλάσης. Ωστόσο, στη δειγματοληψία RUS, σημαντικές πληροφορίες για την πλειοψηφική κλάση μπορεί να χαθούν, γεγονός το οποίο με τη σειρά του θα επηρεάσει την απόδοση του προγνωστικού μοντέλο.

Χρήση μεθόδων Oversampling και Undersampling

Όπως αναφέραμε πριν, υπάρχει μια μεγάλη διαφορά μεταξύ του πλήθους των παρατηρήσεων στα δεδομένα. Μια τεχνική undersampling θα μείωνε σε τεράστιο βαθμό το μέγεθος του dataset (στην περίπτωσή μας) και δεν θα ήταν πλέον αρκετά αντιπροσωπευτικό γιατί θα χάναμε πληροφορίες. Από την άλλη μεριά, η εφαρμογή oversampling θα μπορούσε να οδηγήσει πάλι σε μη αντιπροσωπευτικά δεδομένα, λόγω του μεγάλου πλήθους των πλασματικών δεδομένων που θα πρέπει να δημιουργηθούν. Η υποδειγματοληψία δεν φέρει τον κίνδυνο του overfitting και δεν χρειάζεται παραπάνω υπολογιστική ισχύ γιατί δεν δημιουργεί συνθετικά δεδομένα, επομένως είναι λιγότερο προβληματική σαν τεχνική και αποτελεί την πιο ασφαλή επιλογή μεταξύ των δύο.

Μια τεχνική που συνδυάζει και τις δύο μεθόδους και μπορεί να φέρει το καλύτερο αποτέλεσμα σε περιπτώσεις μεγάλης ανισορροπίας σύμφωνα με έρευνες, είναι η εφαρμογή oversampling στην μικρή κλάση (λιγότερες παρατηρήσεις) και έπειτα η εφαρμογή undersampling στην μεγάλη κλάση (περισσότερες παρατηρήσεις).

Να τονίσουμε βέβαια ότι αυτό δεν αποτελεί κανόνα σε καμία περίπτωση, καθώς η αποτελεσματικότητα κάθε τεχνικής αλλάζει ανάλογα με τους αλγορίθμους πρόβλεψης που επρόκειτο να χρησιμοποιηθούν, αλλά και με το ίδιο το dataset. Ιδανικά, για να είμαστε απολύτως σίγουροι, θα πρέπει να γίνει σύγκριση όλων των πιθανών συνδυασμών, κάτι που είναι πολύ χρονοβόρο και μη διαχειρίσιμο σε πραγματικές συνθήκες (ειδικά σε περιπτώσεις τεραστίου όγκου δεδομένων). Οπότε αυτό που κάνουμε είναι να προσπαθήσουμε να εφαρμόσουμε τεχνικές που έχουν αποδειχθεί αποτελεσματικές σε αντίστοιχα προβλήματα.

Από τη στιγμή που η μεταβλητότητα στα δεδομένα είναι πολύ υψηλή, είναι βολικό να γίνει χρήση ενός αλγορίθμου τύπου Nearest Neighbor, όπως ο SMOTE για να υπερδειγματίσει τα δεδομένα. Με αυτόν τον τρόπο παρέχεται καλύτερη ποιότητα συνθετικών δειγμάτων που θα είναι πιο κοντά στα πραγματικά δεδομένα.

Σε paper που έχει δημοσιευθεί στο Science Direct, έγινε έρευνα σε πολλαπλά dataset ιατρικού περιεχομένου τα οποία παρουσιάζουν πολύ μεγάλη ανισορροπία, όπως το δικό μας. Όπως αναφέρεται αρχικά, τεχνικές oversampling όπως η SMOTE και συνδυασμοί τεχνικών oversampling-undersampling χρησιμοποιούνται συχνά και έχουν αποδειχτεί αποτελεσματικές στο παρελθόν. Η έρευνα που έγινε επικεντρώνεται σε αλγορίθμους οι οποίοι μπορούν να διαχειριστούν καλά τις ανισορροπίες μεταξύ των κλάσεων στα δεδομένα. Τα πειραματικά αποτελέσματα έδειξαν ότι οι cost sensitive εκδόσεις των random forest, του XGBoost και της λογιστικής παλινδρόμησης απέκτησαν εξαιρετική απόδοση, στα τέσσερα σύνολα δεδομένων που δοκιμάστηκαν, σε σύγκριση με άλλους αλγόριθμους.

SMOTE + Tomek Links

Η προσέγγιση που θα επιλέξουμε για να εφαρμόσουμε στο δεύτερο σετ δεδομένων μας είναι μία υβριδική (πηγή 29). Αφού γίνει η υπερδειγματοληψία με SMOTE, τα clusters κλάσεων μπορεί να εισβάλλουν το ένα στο χώρο του άλλου. Ως αποτέλεσμα, το μοντέλο ταξινόμησης θα παρουσιάσει overfitting. Οι σύνδεσμοι Tomek είναι τα αντίθετα ζευγαρωμένα δείγματα κλάσης που είναι οι πιο κοντινοί γείτονες μεταξύ τους. Επομένως, η πλειονότητα των παρατηρήσεων της κλάσης από αυτούς τους συνδέσμους αφαιρείται καθώς πιστεύεται ότι αυξάνει τον διαχωρισμό των κλάσεων κοντά στα όρια απόφασης. Για να αποκτήσουμε καλύτερα clusters κλάσης, οι σύνδεσμοι Tomek εφαρμόζονται στα δείγματα της μειοψηφικής κλάσης που έγιναν oversampled από την SMOTE. Έτσι, αντί να αφαιρούμε τις παρατηρήσεις μόνο από την πλειοψηφική κλάση, αφαιρούμε και τις δύο παρατηρήσεις κλάσης από τους συνδέσμους Tomek.

Hyperparameter Tuning

Υπερπαράμετροι είναι οι ρυθμίσεις που μπορούν να συντονιστούν πριν από την εκτέλεση μιας εργασίας εκπαίδευσης για τον έλεγχο της συμπεριφοράς ενός αλγορίθμου μηχανικής μάθησης. Μπορούν να έχουν μεγάλο αντίκτυπο στην εκπαίδευση μοντέλων, καθώς σχετίζονται με τον χρόνο εκπαίδευσης, τις απαιτήσεις πόρων υποδομής (και κατά συνέπεια το κόστος), τη σύγκλιση του μοντέλου και την ακρίβειά του. Οι παράμετροι του μοντέλου μαθαίνονται ως μέρος της διαδικασίας εκπαίδευσης, ενώ οι τιμές των υπερπαραμέτρων ορίζονται πριν από την εκτέλεση της εργασίας εκπαίδευσης και δεν αλλάζουν κατά τη διάρκεια της εκπαίδευσης.

Υπερπαράμετροι του μοντέλου — ορίζουν τη θεμελιώδη κατασκευή του ίδιου του μοντέλου, π.χ. χαρακτηριστικά νευρωνικών δικτύων όπως το filter size, pooling, κ.α.

Υπερπαράμετροι του Optimizer — σχετίζονται με τον τρόπο με τον οποίο το μοντέλο μαθαίνει τα μοτίβα με βάση τα δεδομένα. Αυτοί οι τύποι υπερπαραμέτρων περιλαμβάνουν βελτιστοποιητές όπως ο gradient descent και ο stochastic gradient descent (SGD), Adam, RMSprop, Adadelta και ούτω καθεξής.

Μερικές από τις κοινές τεχνικές που χρησιμοποιούνται για τον συντονισμό υπερπαραμέτρων περιλαμβάνουν την Αναζήτηση Πλέγματος (Grid Search), την Τυχαία Αναζήτηση (Random Search), τη Βελτιστοποίηση Bayes (Bayesian Optimization) και άλλες.

Grid Search — Ρυθμίζει ένα πλέγμα που αποτελείται από υπερπαραμέτρους και τις διαφορετικές τιμές τους. Για κάθε πιθανό συνδυασμό, εκπαιδεύεται ένα μοντέλο και παράγεται μία βαθμολογία στα δεδομένα επικύρωσης. Με αυτήν την προσέγγιση δοκιμάζεται κάθε συνδυασμός των πιθανών τιμών υπερπαραμέτρων. Ενώ η προσέγγιση εκτελεί εκτεταμένη σάρωση σε όλους τους πιθανούς συνδυασμούς, μπορεί να είναι πολύ αναποτελεσματική όσον αφορά τον χρόνο και το κόστος της εκπαίδευσης. Θα αποτελέσει και την μέθοδο που θα χρησιμοποιήσουμε στα μοντέλα μας.

Random Search — παρόμοια με την αναζήτηση πλέγματος, αλλά αντί για εκπαίδευση και βαθμολόγηση σε κάθε πιθανό συνδυασμό υπερπαραμέτρων, επιλέγονται τυχαίοι συνδυασμοί. Μπορεί να οριστεί ο αριθμός των επαναλήψεων αναζήτησης με βάση τους περιορισμούς χρόνου και πόρων.

Αυτές οι 2 προσεγγίσεις εξακολουθούν να είναι πεπαλαιωμένες και απαιτούν “trial and error” μέχρι να επιτευχθούν ικανοποιητικά αποτελέσματα. Τα τελευταία χρόνια χρησιμοποιούνται αυτοματοποιημένες λύσεις συντονισμού υπερπαραμέτρων που διατίθενται από παρόχους υπηρεσιών cloud, όπως η Amazon, η Google και η Microsoft. Αυτές οι λύσεις χρησιμοποιούν μεθόδους όπως gradient descent, bayesian βελτιστοποίηση, κ.α. για τη διεξαγωγή μιας καθοδηγούμενης αναζήτησης για τις καλύτερες ρυθμίσεις. Π.χ. η Amazon SageMaker ή η Google AutoML βρίσκουν την καλύτερη έκδοση ενός μοντέλου εκτελώντας πολλές εργασίες εκπαίδευσης στο σύνολο δεδομένων χρησιμοποιώντας τον αλγόριθμο και τις περιοχές υπερπαραμέτρων που καθορίζουμε. Στη συνέχεια, επιλέγουν τις τιμές που καταλήγουν σε ένα μοντέλο που έχει την καλύτερη απόδοση για την επιθυμητή μέτρηση.

Optuna

Ένα πολύ καλό open source εργαλείο για hyperparameter optimization είναι το Optuna. Το Optuna λειτουργεί ανεξαρτήτως framework. Μπορεί να χρησιμοποιηθεί με οποιοδήποτε αλγόριθμο μηχανικής μάθησης ή βαθιάς μάθησης. Ο αλγόριθμος του θα αποφασίσει εάν ο συνδυασμός υπερπαραμέτρων αξίζει για εκπαίδευση μετά από μερικές επαναλήψεις και θα σταματήσει τη διαδικασία εκμάθησης αυτού του συνδυασμού υπερπαραμέτρων εάν υπάρχει περιορισμένη βελτίωση.

Μέθοδοι Αξιολόγησης Αποτελεσμάτων

Μετρητές Απόδοσης σε Προβλήματα Ταξινόμησης

Εφόσον το πρόβλημα μας είναι ένα πρόβλημα ταξινόμησης, θα χρησιμοποιηθούν οι πιο γνωστοί μετρητές απόδοσης ταξινόμησης, οι οποίοι είναι οι παρακάτω.

- Accuracy: Είναι ο λόγος των σωστά ταξινομημένων στοιχείων προς το συνολικό πλήθος τους. Επίσης είναι ο πιο κοινός μετρητής αξιολόγησης όταν πρόκειται για μοντέλα προγνωστικής ανάλυσης. Παρόλα αυτά ένα μεγάλο ποσοστό Accuracy δεν σημαίνει απαραίτητα ότι έχουμε ένα αποτελεσματικό μοντέλο, ειδικά όταν δουλεύουμε με unbalanced δεδομένα. Στην περίπτωσή μας ειδικά που πρόκειται για binary classification πρόβλημα με μεγάλο class imbalance, ο μετρητής Accuracy μας είναι άχρηστος.
- Precision: Ο δείκτης precision υπολογίζεται για κάθε κλάση και ισούται με το πλήθος των στοιχείων που έχουν ταξινομηθεί σωστά στην κλάση προς το σύνολο των στοιχείων που έχουν ταξινομηθεί ότι ανήκουν στην κλάση αυτή.
- Recall: Ο δείκτης recall υπολογίζεται για κάθε κλάση και ισούται με το πλήθος των στοιχείων που έχουν ταξινομηθεί σωστά στην κλάση προς το σύνολο των στοιχείων που πραγματικά ανήκουν στην κλάση αυτή.
- F1-Score: Ο δείκτης F1-score για κάθε κλάση ισούται με το διπλάσιο του γινομένου των precision και recall προς το άθροισμά τους. Ο δείκτης F1-score για το σύνολο του μοντέλου ισούται με τον μέσο όρο των F1-scores κάθε κατηγορίας. Ο δείκτης F1-score προσπαθεί στην ουσία να παντρέψει τους precision και recall δείκτες σε μια μόνο τιμή.

- Area Under Curve (ROC-AUC): Το ROC AUC score μας δείχνει την ικανότητα του μοντέλου να διαφοροποιεί μεταξύ θετικής και αρνητικής κλάσης. Όταν έχουμε AUC=0,5, τότε ο ταξινομητής δεν μπορεί να διακρίνει μεταξύ θετικών και αρνητικών σημείων κλάσης. Αυτό σημαίνει ότι προβλέπει είτε μια τυχαία κλάση είτε μια σταθερή κλάση για όλα τα σημεία δεδομένων. Έτσι, όσο υψηλότερη είναι η τιμή AUC για έναν ταξινομητή, τόσο καλύτερη είναι η ικανότητά του να διακρίνει μεταξύ θετικών και αρνητικών κλάσεων. Είναι ιδανική για binary class classification. Η υψηλότερη τιμή AUC μπορεί να οδηγήσει σε ένα πιο ακριβές μοντέλο πρόβλεψης. Η καμπύλη ROC σχεδιάζεται λαμβάνοντας υπόψη τα True Positive Rates (TPR) στον άξονα y και τα False Positive Rates (FPR) στον άξονα x σε μια κλίμακα από 0 έως 1. Τα TPR και FPR υπολογίζονται για κάθε threshold point στη διαδικασία ταξινόμησης. Τα threshold points είναι οι τιμές της πιθανότητας που έχουν χρησιμοποιηθεί για τον προσδιορισμό της κλάσης. Στην περίπτωση που έχουμε μεγάλη ανισορροπία στις κλάσεις των δεδομένων, ο δείκτης ROC μπορεί να είναι παραπλανητικός, για αυτό προτιμάται ο δείκτης Precision-Recall για το evaluation του μοντέλου.
- Precision - Recall Curve (PR-AUC): Μας δείχνει το tradeoff μεταξύ precision και recall για διαφορετικό όριο. Υψηλό Area under curve αντιστοιχεί σε υψηλό recall και precision. Υψηλό precision αντιστοιχεί σε λίγα false positives και υψηλό recall αντιστοιχεί σε λίγα false negatives. Τα precision και recall εστιάζουν στην θετική κλάση (την μικρή) και αγνοούν τα true negatives (μεγάλη κλάση) και για αυτό το λόγο τα Precision-recall curves προτιμώνται για πολύ skewed και unbalanced δεδομένα, σαν τα δικά μας, όπου η ROC curve θα μας δίνει αισιόδοξα αποτελέσματα της απόδοσης των μοντέλων.
- Macro F1-Score: Σε μη ισορροπημένο σύνολο δεδομένων καλό είναι να χρησιμοποιείται το macro F1 score, καθώς αυτό θα εξακολουθεί να αντικατοπτρίζει την πραγματική απόδοση του μοντέλου ακόμα και όταν οι κλάσεις είναι skewed με μεγάλη ανισότητα. Το Macro F1 score είναι ο μέσος όρος των F1 scores των θετικών και των αρνητικών κλάσεων.

Confusion Matrix

Μας περιγράφει με εύκολο και κατανοητό τρόπο τα ακόλουθα:

1. True Positives – Το μοντέλο σωστά προέβλεψε ότι θα έχουμε backorder
2. True Negatives – Το μοντέλο σωστά προέβλεψε ότι δεν θα έχουμε backorder
3. False Positives – Το μοντέλο λανθασμένα προέβλεψε ότι θα έχουμε backorder
4. False Negatives – Το μοντέλο λανθασμένα προέβλεψε ότι δεν θα έχουμε backorder

Επεξήγηση Μετρητών Απόδοσης

- Precision: Μας δείχνει την αναλογία των προβλεπόμενων backorders που όντως καταλήγουν σε backorder, που είναι η αναλογία των true positive προβλεπόμενων προς των συνολικά προβλεπόμενων positive.
- Recall: Μας δείχνει την αναλογία των backorder αντικειμένων που προβλέπεται ότι θα καταλήξουν σε backorder, που σημαίνει αναλογία των true positive προβλεπόμενων προς τα συνολικά πραγματικά positive.

Αν ορίσουμε ένα χαμηλό όριο ταξινόμησης, προβλέπουμε ότι τα αντικείμενα πάνε σε backorder πιο συχνά. Αυτό οδηγεί σε υψηλότερη recall και χαμηλότερη precision. Εάν ορίσουμε ένα υψηλό όριο ταξινόμησης, δεν προβλέπουμε ότι τα αντικείμενα θα πάνε τόσο συχνά σε backorder. Αυτό οδηγεί σε χαμηλότερη recall και μεγαλύτερη precision.

Ανάλογα με τον στόχο μας, μπορεί να προτιμούμε το precision ή το recall έναντι του άλλου. Για παράδειγμα, εάν το κόστος της αποτυχίας πρόβλεψης ενός backorder (π.χ. ακυρωμένες παραγγελίες, απώλεια πελατών κ.λπ.) είναι μεγαλύτερο από το κόστος μιας εσφαλμένης πρόβλεψης ενός backorder (π.χ. υπερβολικό απόθεμα και υψηλότερο κόστος αποθέματος), τότε θα επιλέξουμε χαμηλότερο όριο (σε PR curve) και θα δώσουμε προτεραιότητα στο recall έναντι του precision.

Loss Function (Νευρωνικά)

Στη μηχανική μάθηση, η συνάρτηση απώλειας (loss function) χρησιμοποιείται για την εύρεση σφάλματος ή απόκλισης στη διαδικασία εκμάθησης. Το Keras απαιτεί συνάρτηση απώλειας κατά τη διαδικασία του compile του μοντέλου. Η βελτιστοποίηση είναι μια σημαντική διαδικασία που βελτιστοποιεί τα βάρη εισόδου (input weights) συγκρίνοντας την πρόβλεψη και την απώλεια. Εμείς θα χρησιμοποιήσουμε την συνάρτηση απώλειας στο validation σετ και όχι στο training για να αποφύγουμε το overfitting. Η συνάρτηση απώλειας που χρησιμοποιείται σε αυτή τη μελέτη είναι η δυαδική συνάρτηση διασταυρούμενης εντροπίας (binary cross entropy) για τον υπολογισμό της απώλειας μεταξύ της πραγματικής κλάσης και της προβλεπόμενης κλάσης.

SHAP (Νευρωνικά)

Το SHAP (SHapley Additive exPlanations) είναι μια ενοποιημένη προσέγγιση για την εξήγηση των αποτελεσμάτων οποιουδήποτε μοντέλου μηχανικής εκμάθησης. Συνδυάζει τη θεωρία παιγνίων με τα μοντέλα μηχανικής μάθησης. Έχει βελτιστοποιημένες συναρτήσεις για την ερμηνεία μοντέλων που βασίζονται σε δέντρα και μια συνάρτηση αγνωστικής επεξήγησης μοντέλων για την ερμηνεία οποιουδήποτε μοντέλου μαύρου κουτιού (όπως τα νευρωνικά δίκτυα) για το οποίο είναι γνωστές οι προβλέψεις. Συνοπτικά, οι τιμές του Shapley υπολογίζουν τη σημασία ενός χαρακτηριστικού συγκρίνοντας τι προβλέπει ένα μοντέλο με και χωρίς αυτό το χαρακτηριστικό. Ωστόσο, δεδομένου ότι η σειρά με την οποία ένα μοντέλο βλέπει τα χαρακτηριστικά μπορεί να επηρεάσει τις προβλέψεις του, αυτό γίνεται με όλους τους δυνατούς τρόπους, έτσι ώστε τα χαρακτηριστικά να συγκρίνονται δίκαια. Αυτή η προσέγγιση είναι εμπνευσμένη από τη θεωρία παιγνίων. Εκπαιδεύουμε, συντονίζουμε και δοκιμάζουμε το μοντέλο μας. Στη συνέχεια, μπορούμε να χρησιμοποιήσουμε τα δεδομένα μας και το μοντέλο για να δημιουργήσουμε ένα πρόσθετο μοντέλο SHAP που εξηγεί το μοντέλο ταξινόμησης μας. Οι shap values [0] είναι επεξηγήσεις σε σχέση με την αρνητική κλάση, ενώ οι shap values [1] είναι επεξηγήσεις σε σχέση με τη θετική κλάση.

Δεν έγινε ο υπολογισμός του Feature Importance μέσω shap values στα LSTM μοντέλα, γιατί υπάρχει ασυμβατότητα με τα LSTM layers και τον deep explainer για όλες τις εκδόσεις του tensorflow 2.

Προσέγγιση Εκπαίδευσης

Η τεχνική που θα ακολουθήσουμε για την εκπαίδευση των μοντέλων μας είναι η Coarse to Fine. Η "Coarse to Fine" συνήθως αναφέρεται στη βελτιστοποίηση υπερπαραμέτρων ενός αλγορίθμου κατά την οποία θα θέλαμε να δοκιμάσουμε διαφορετικούς συνδυασμούς των υπερπαραμέτρων και να αξιολογήσουμε την απόδοσή του. Ωστόσο, λόγω του μεγάλου αριθμού παραμέτρων και του μεγάλου εύρους των τιμών τους, είναι σχεδόν αδύνατο να ελεγχθούν όλοι οι διαθέσιμοι συνδυασμοί. Για αυτόν τον λόγο, συνήθως διαχωρίζουμε το διαθέσιμο εύρος τιμών κάθε παραμέτρου σε ένα πλέγμα τιμών (π.χ. val = 5,6,7,8,9) για να εκτιμήσουμε το αποτέλεσμα της αύξησης ή της μείωσης της τιμής αυτής της παραμέτρου. Αφού διαλέξουμε την τιμή που φαίνεται πιο ελπιδοφόρα, μπορούμε να κάνουμε μια μικρή αναζήτηση γύρω από αυτήν για να την βελτιώσουμε ακόμη περισσότερο.

Για το hyperparameter tuning των μοντέλων μηχανικής μάθησης έχει γίνει χρήση του Grid Search.

Ρύθμιση Παραμέτρων στα Νευρωνικά Δίκτυα

Η λειτουργία ενεργοποίησης που χρησιμοποιούμε για όλα τα layers των νευρωνικών, πέρα του output που είναι sigmoid γιατί έχουμε binary classification πρόβλημα, είναι η ReLU. Ο λόγος για τον οποίο η ReLU υιοθετήθηκε περισσότερο είναι ότι επιτρέπει καλύτερη βελτιστοποίηση με χρήση Stochastic Gradient Descent, πιο αποτελεσματικούς υπολογισμούς και είναι αμετάβλητη σε κλίμακα, που σημαίνει ότι τα χαρακτηριστικά της δεν επηρεάζονται από την κλίμακα της εισόδου.

Τα μικρότερα batch sizes έχουν ως αποτέλεσμα να έχουμε περισσότερα weight updates ανά εποχή. Για αυτό το λόγο στα μοντέλα που έχουν γίνει resampled με Smote Tomek θα μπορούμε να χρησιμοποιήσουμε μεγαλύτερο batch size για να επιταχύνουμε την εκπαίδευση του δικτύου, χωρίς να φοβόμαστε ότι τα batches δεν θα περιλαμβάνουν παρατηρήσεις από την σπάνια/θετική κλάση. Εμείς επιλέγουμε batch size 64.

Σε όλα τα μοντέλα έχει γίνει χρήση Early Stopping και dropout layers (0.4/0.2) για την αποφυγή του overfitting.

Ως optimizer χρησιμοποιήθηκε ο Adam με default learning rate (0,001). Στην περίπτωση των μοντέλων που δεν έχουν γίνει resampled ($1^{ο}$ σετ δεδομένων) έχει γίνει αύξηση του weight της μικρής κλάσης κατά την διάρκεια του training, ώστε να την θεωρούν σημαντικότερη.

Έχουμε ορίσει το όριο των epochs στα 100, το οποίο δεν πρόκειται να επιτευχθεί λόγω της χρήσης early stopping και checkpoint.

Για το hyperparameter tuning των νευρωνικών έχει γίνει χρήση του Grid Search με Keras Classifier.

Η εκπαίδευση των μοντέλων έγινε με την TensorFlow 2.10

MLP Summary

```
Layer (type)                 Output Shape              Param #
=================================================================
 dense (Dense)                (None, 256)               5632

 dense_1 (Dense)              (None, 256)               65792

 dropout (Dropout)            (None, 256)               0

 dense_2 (Dense)              (None, 256)               65792

 dropout_1 (Dropout)          (None, 256)               0

 dense_3 (Dense)              (None, 1)                 257
=================================================================
```

LSTM Summary

```
Layer (type)                 Output Shape              Param #
=================================================================
 lstm (LSTM)                  (None, 1, 128)            76800

 lstm_1 (LSTM)                (None, 64)                49408

 dropout (Dropout)            (None, 64)                0

 dense (Dense)                (None, 1)                 65
=================================================================
```

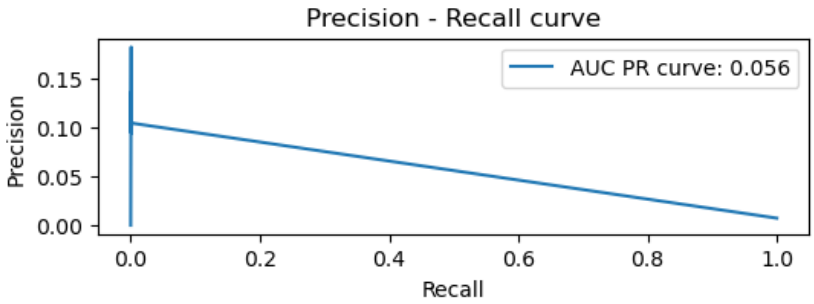
Machine Learning Models

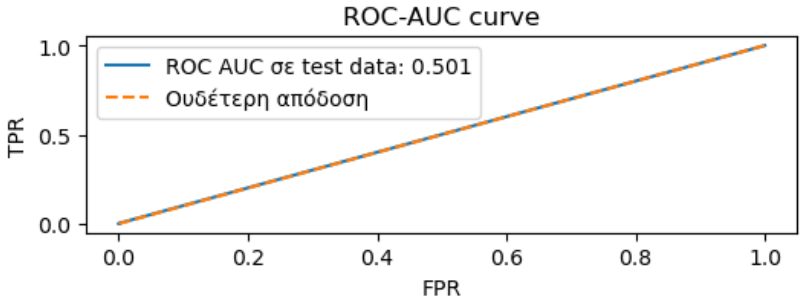
Λογιστική παλινδρόμηση

Επιλογή Παραμέτρων: penalty = l1 + l2 (δηλαδή elastic net) και διάφορες τιμές C για την εκπαίδευση και Grid Search για ρύθμιση των παραμέτρων.

Penalty: Αναφέρεται σε μια παράμετρο που χρησιμοποιείται σε τύπους κανονικοποιημένων αλγορίθμων λογιστικής παλινδρόμησης, όπως η κανονικοποίηση (regularization) L1 (Lasso) και L2 (Ridge). Αυτή η παράμετρος καθορίζει την ισχύ του όρου κανονικοποίησης στο μοντέλο παλινδρόμησης. Γίνεται επιβολή penalty στην αύξηση του μεγέθους των τιμών των παραμέτρων προκειμένου να μειωθεί το overfitting.
C: Αντίστροφη δύναμη κανονικοποίησης. Μια μεταβλητή ελέγχου που διατηρεί την τροποποίηση δύναμης της κανονικοποίησης, όντας αντίστροφη του ρυθμιστή λάμδα (C = 1/λ). Οι μικρότερες τιμές υποδηλώνουν ισχυρότερη κανονικοποίηση.

Figure 26: Λογιστική παλινδρόμηση σε πρώτο σετ δεδομένων





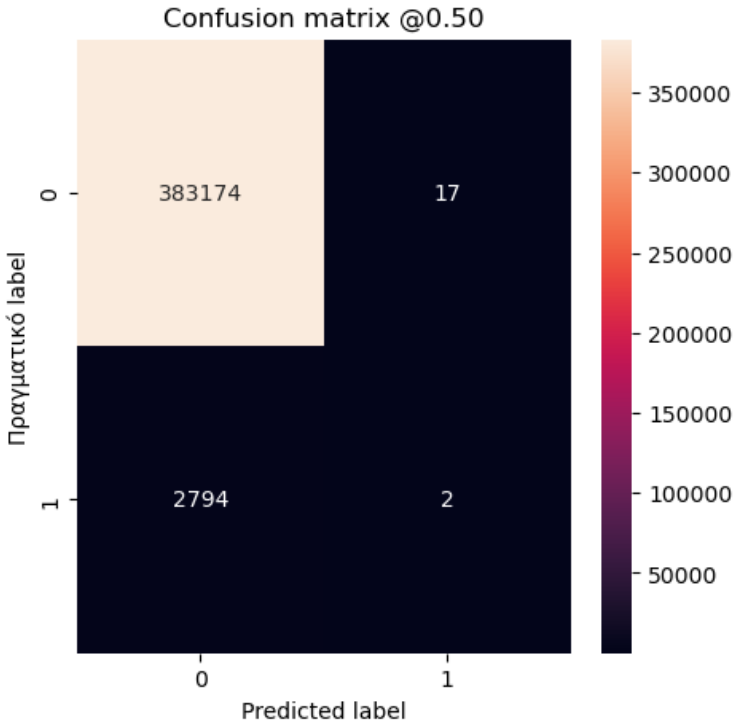
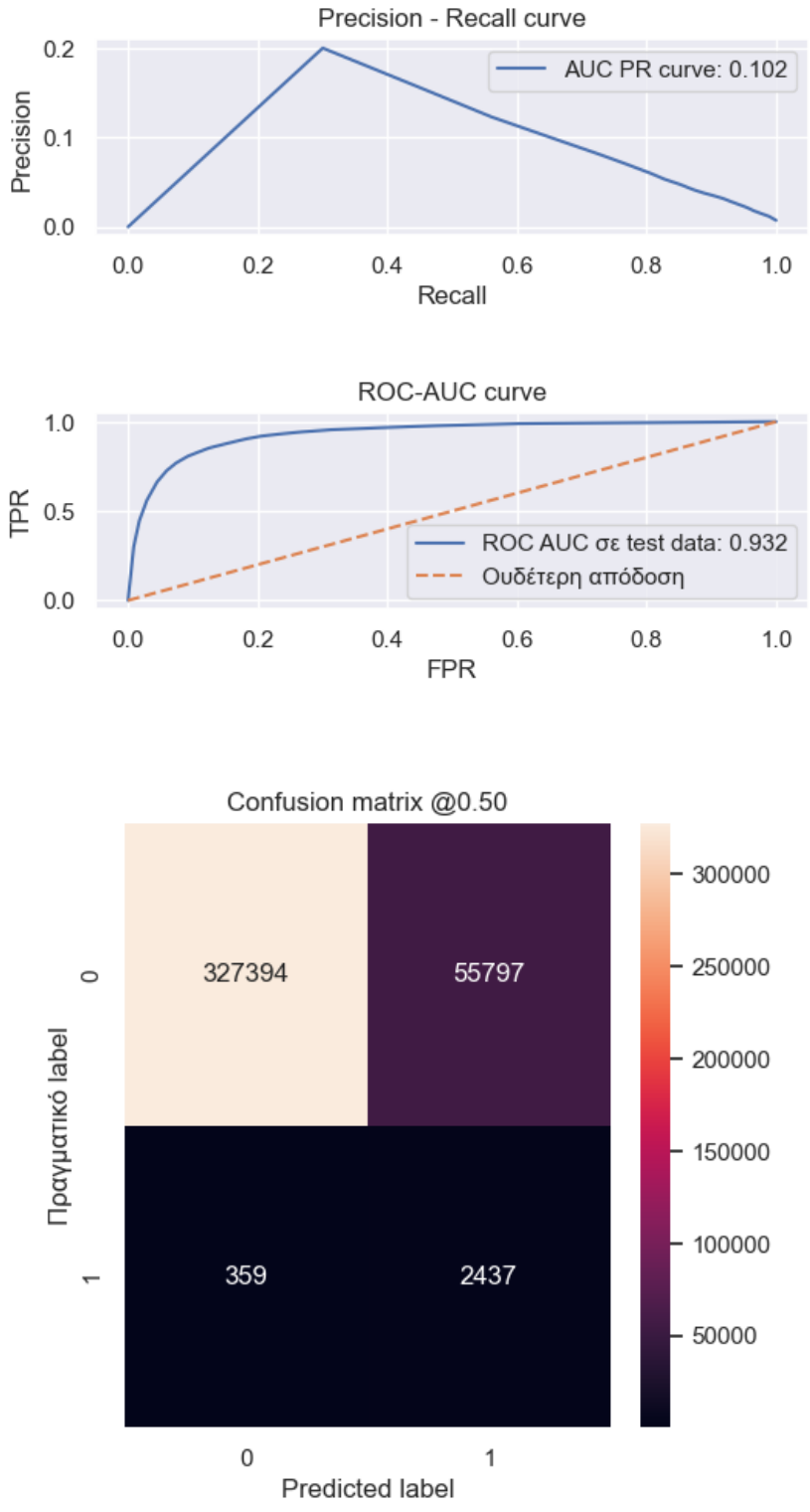


Figure 27: Λογιστική παλινδρόμηση σε δεύτερο σετ δεδομένων

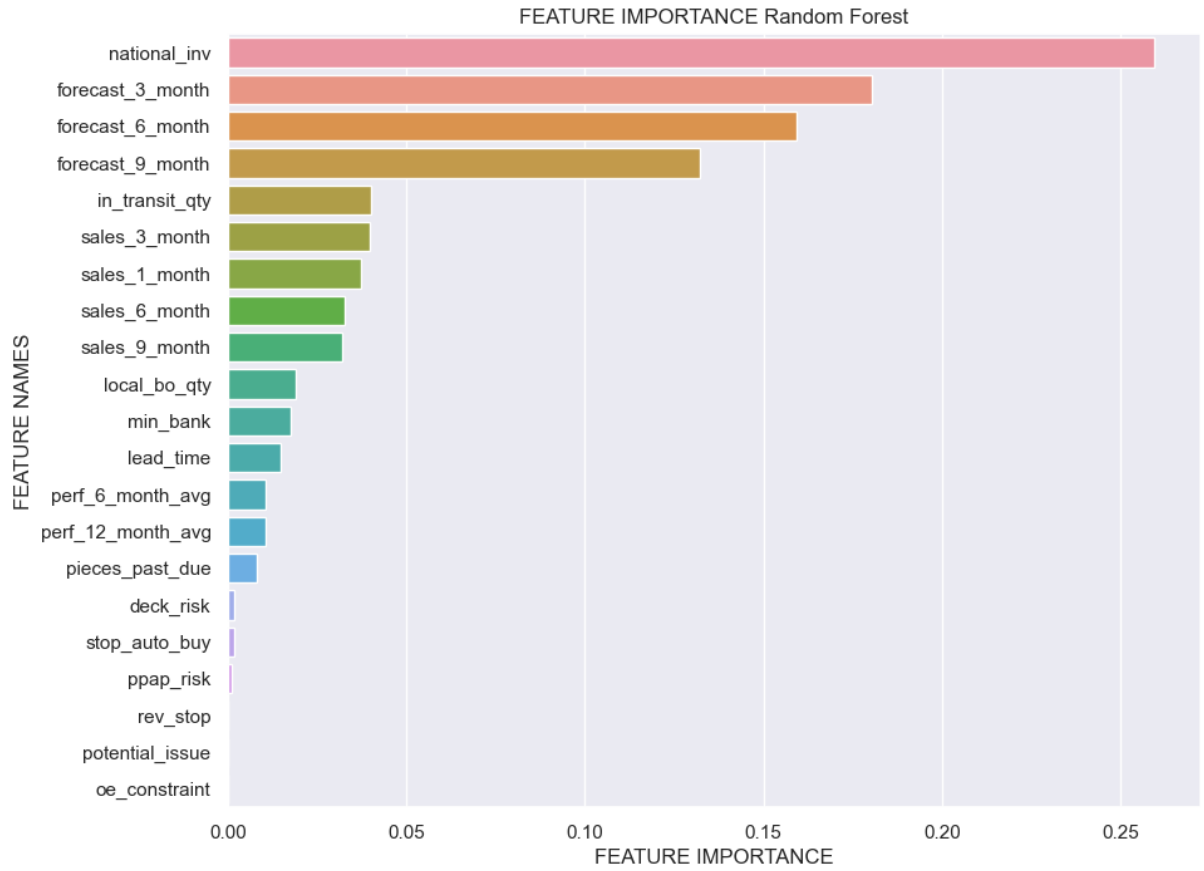


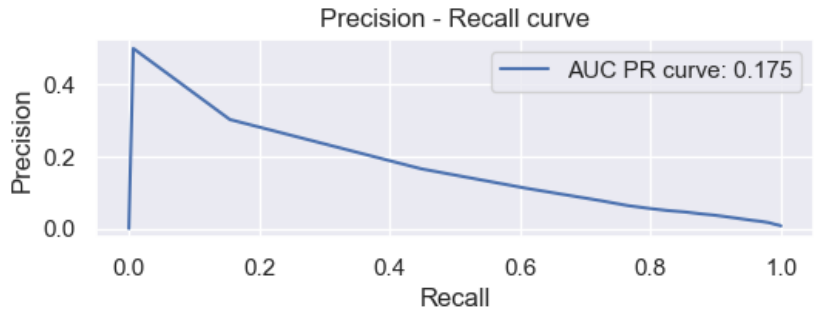
Random Forests

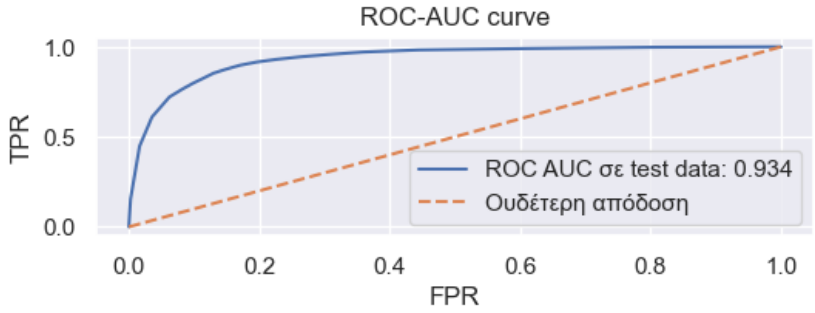
Επιλογή Παραμέτρων: class_weight = balanced_subsample, διάφορες τιμές n_estimators και max_depth για την εκπαίδευση και Grid Search για ρύθμιση των παραμέτρων.

Class weight: Αναφέρεται στην αντιστοίχιση βαρών κάθε κλάσης κατά την εκπαίδευση για τη διαχείριση της ανισορροπίας μεταξύ των κλάσεων. Η λειτουργία "balanced" χρησιμοποιεί τις τιμές του y για να προσαρμόζει αυτόματα τα βάρη αντιστρόφως ανάλογα με τις συχνότητες κλάσης στα δεδομένα εισόδου. Η λειτουργία "balanced_subsample" είναι η ίδια με την "balanced" εκτός από το ότι τα βάρη υπολογίζονται με βάση το δείγμα bootstrap (bootstrapping είναι ένας τύπος επαναδειγματοληψίας όπου μεγάλοι αριθμοί μικρότερων δειγμάτων του ίδιου μεγέθους λαμβάνονται επανειλημμένα, με αντικατάσταση από ένα μόνο αρχικό δείγμα) για κάθε δέντρο που αναπτύσσεται.
N estimators: Ο αριθμός των δέντρων στο δάσος.
Max depth: Το μέγιστο βάθος του δέντρου. Αν δεν δηλωθεί, τότε οι κόμβοι επεκτείνονται μέχρι να γίνουν όλα τα φύλλα καθαρά ή έως ότου όλα τα φύλλα περιέχουν λιγότερα από min_samples_split (ο ελάχιστος αριθμός δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου, default=2) δείγματα.

Figure 28: Random Forests σε πρώτο σετ δεδομένων




Precision - Recall curve
AUC PR curve: 0.175
Precision
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
Recall


ROC-AUC curve
TPR
0.0
0.5
1.0
ROC AUC σε test data: 0.934
Ουδέτερη απόδοση
0.0
0.2
0.4
0.6
0.8
1.0
FPR

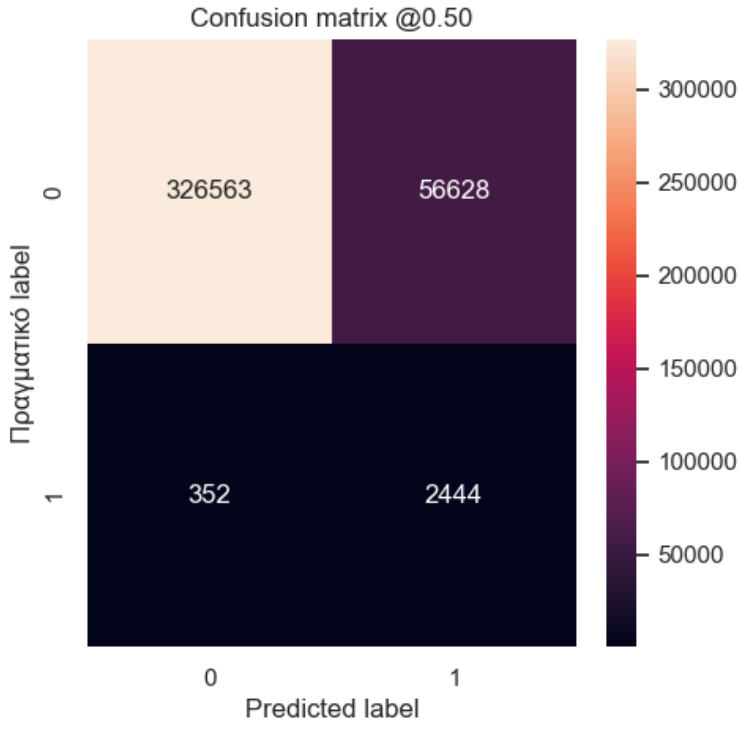
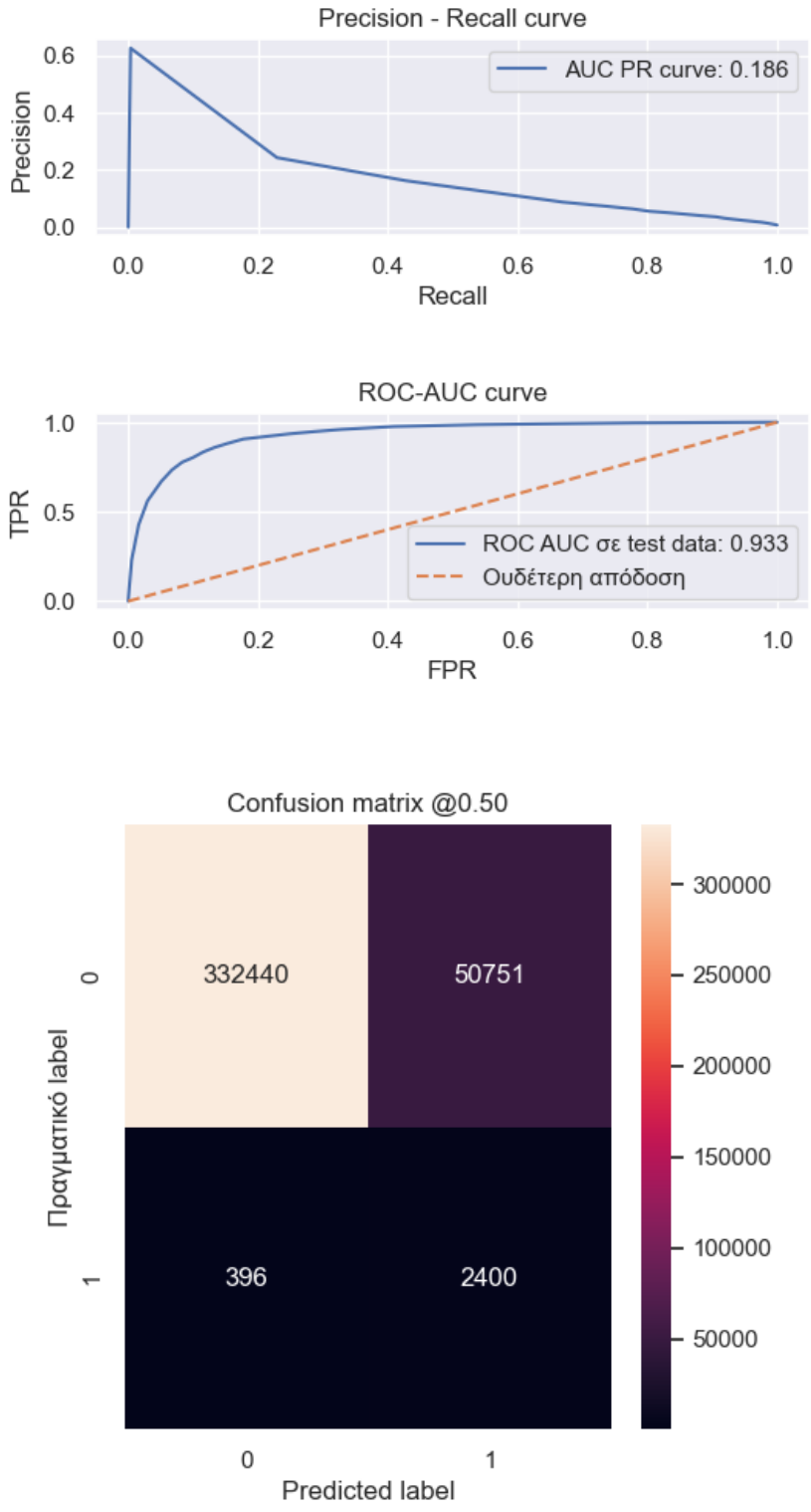
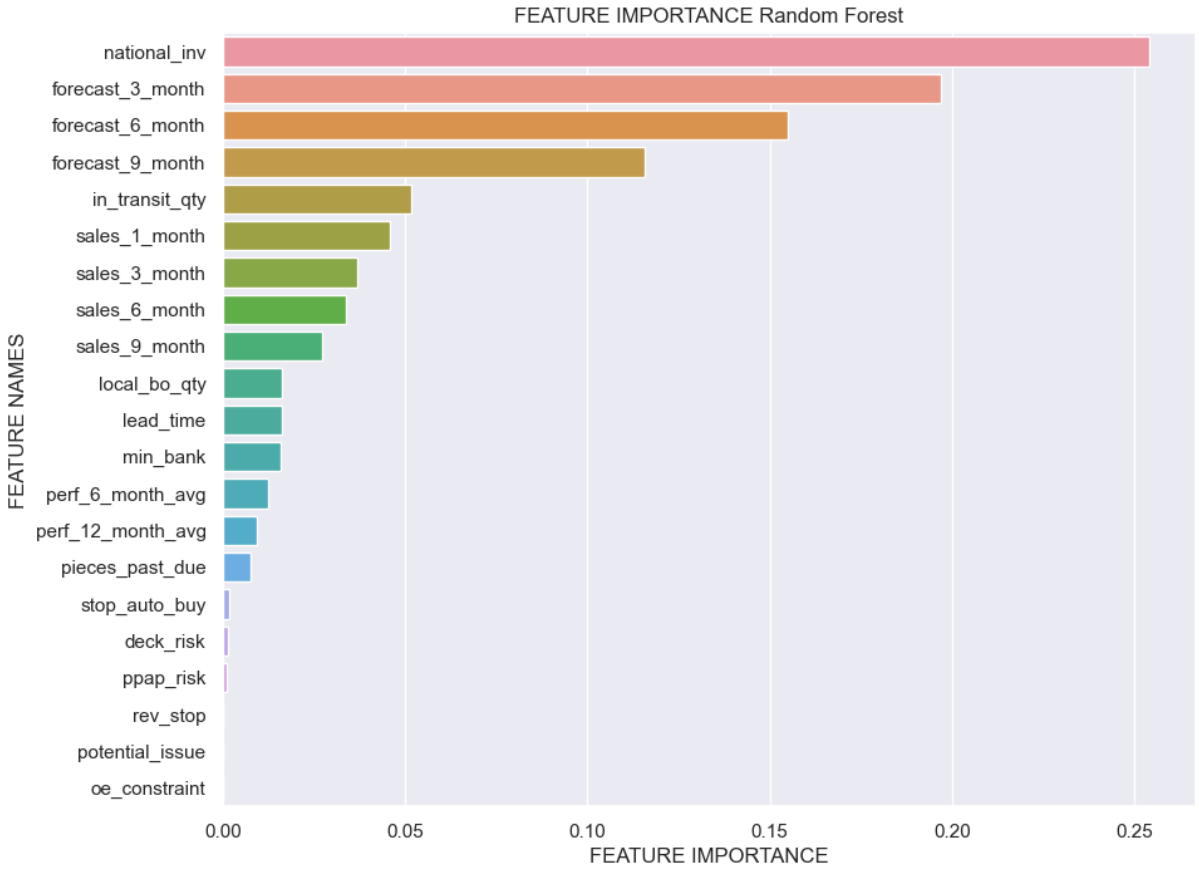

Confusion matrix @0.50
326563
56628
352
2444
Πραγματικό label
0
1
0
1
Predicted label
300000
250000
200000
150000
100000
50000

Figure 29: Random Forests σε δεύτερο σετ δεδομένων



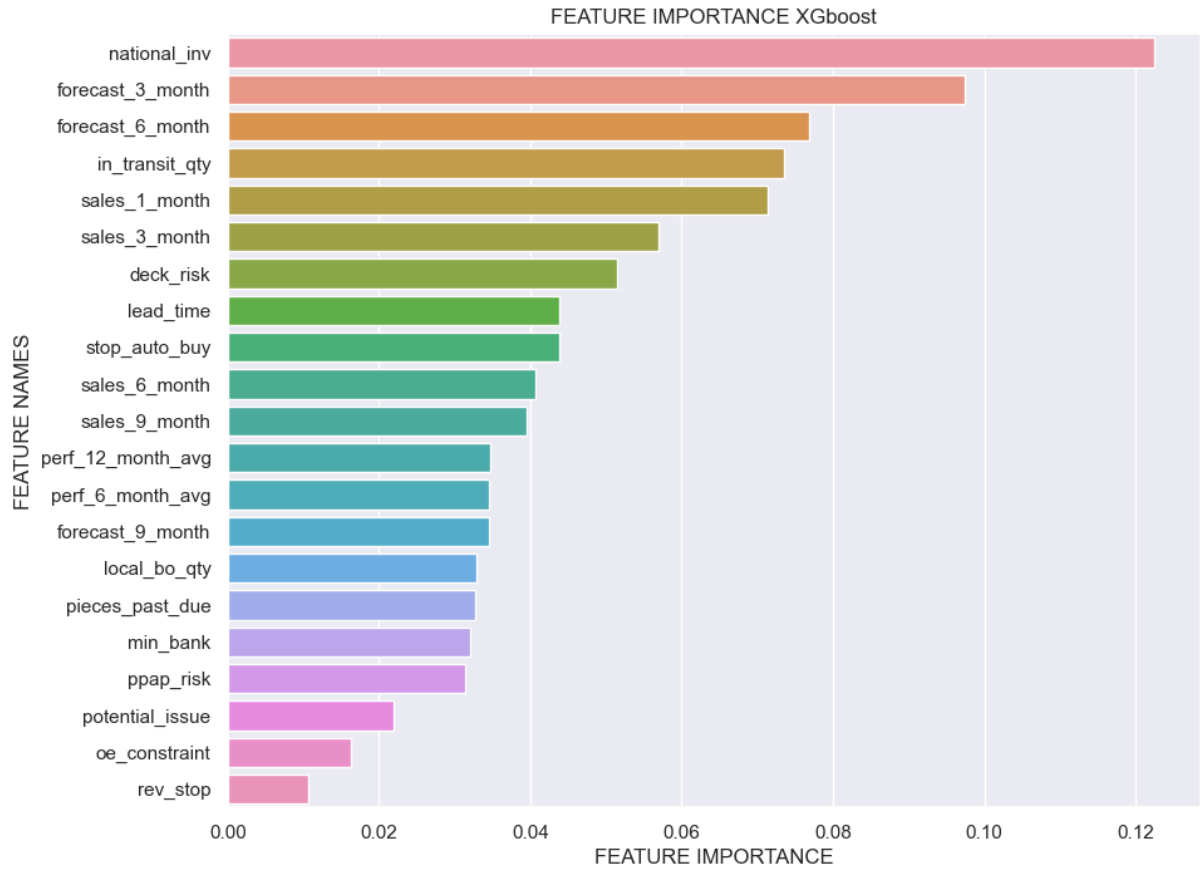

FEATURE IMPORTANCE Random Forest
national_inv
forecast_3_month
forecast_6_month
forecast_9_month
in_transit_qty
sales_1_month
sales_3_month
sales_6_month
sales_9_month
local_bo_qty
lead_time
min_bank
perf_6_month_avg
perf_12_month_avg
pieces_past_due
stop_auto_buy
deck_risk
ppap_risk
rev_stop
potential_issue
oe_constraint
FEATURE NAMES
0.00
0.05
0.10
0.15
0.20
0.25
FEATURE IMPORTANCE

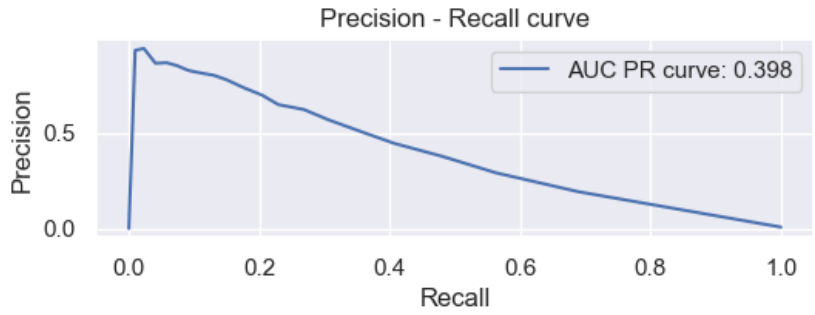
XG Boost

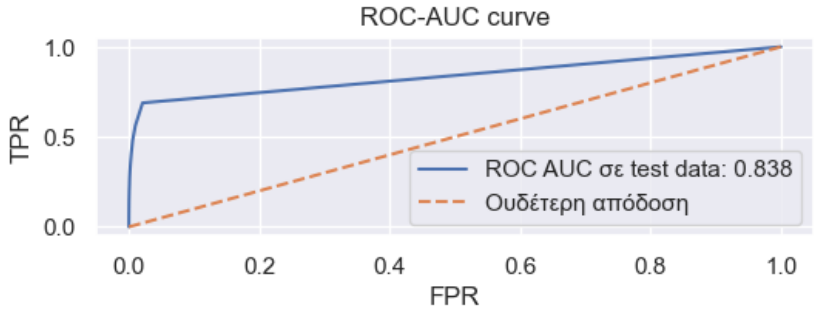
Επιλογή Παραμέτρων: Διάφορες τιμές n_estimators για την εκπαίδευση, eval_metric = logloss (negative log-likelihood) και Grid Search για ρύθμιση των παραμέτρων.

N estimators: Αριθμός δέντρων ενισχυμένων με κλίση (gradient boosted). Ισοδυναμεί με τον αριθμό των boosting rounds (ή αλλιώς δένδρων).

Figure 30: XG Boost σε πρώτο σετ δεδομένων




Precision - Recall curve
AUC PR curve: 0.398
Precision
0.0
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Recall


ROC-AUC curve
TPR
1.0
0.5
0.0
ROC AUC σε test data: 0.838
Ουδέτερη απόδοση
0.0
0.2
0.4
0.6
0.8
1.0
FPR

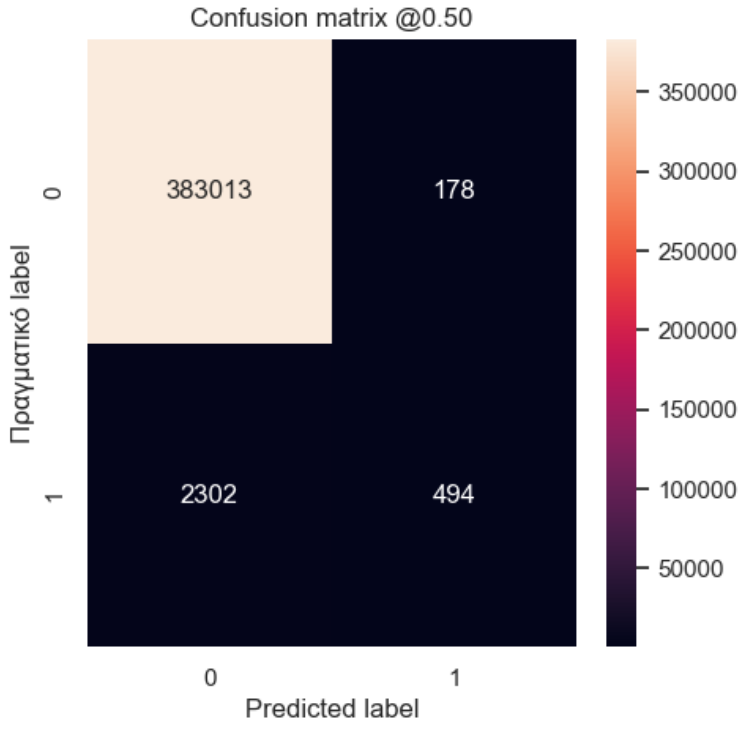
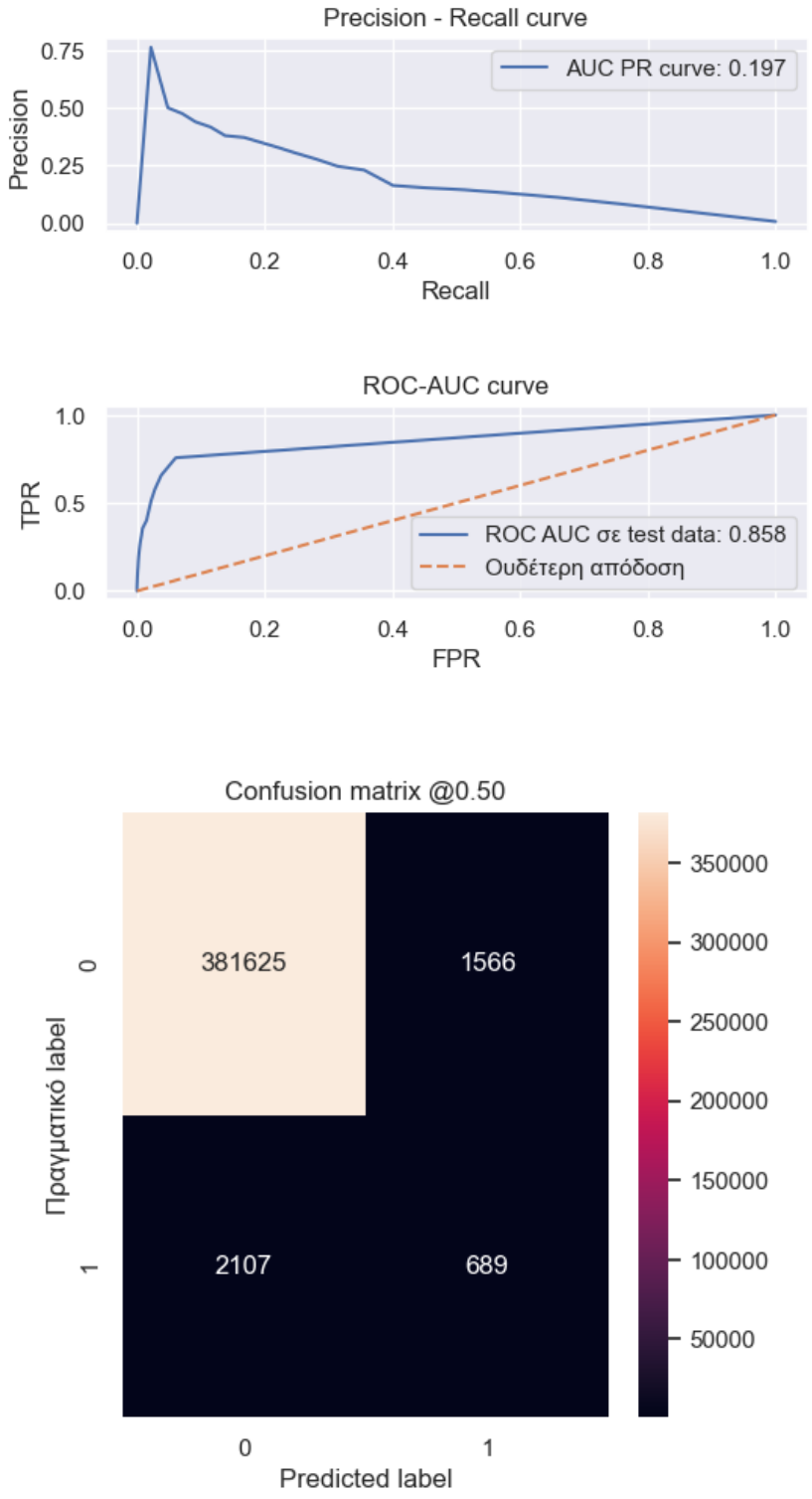
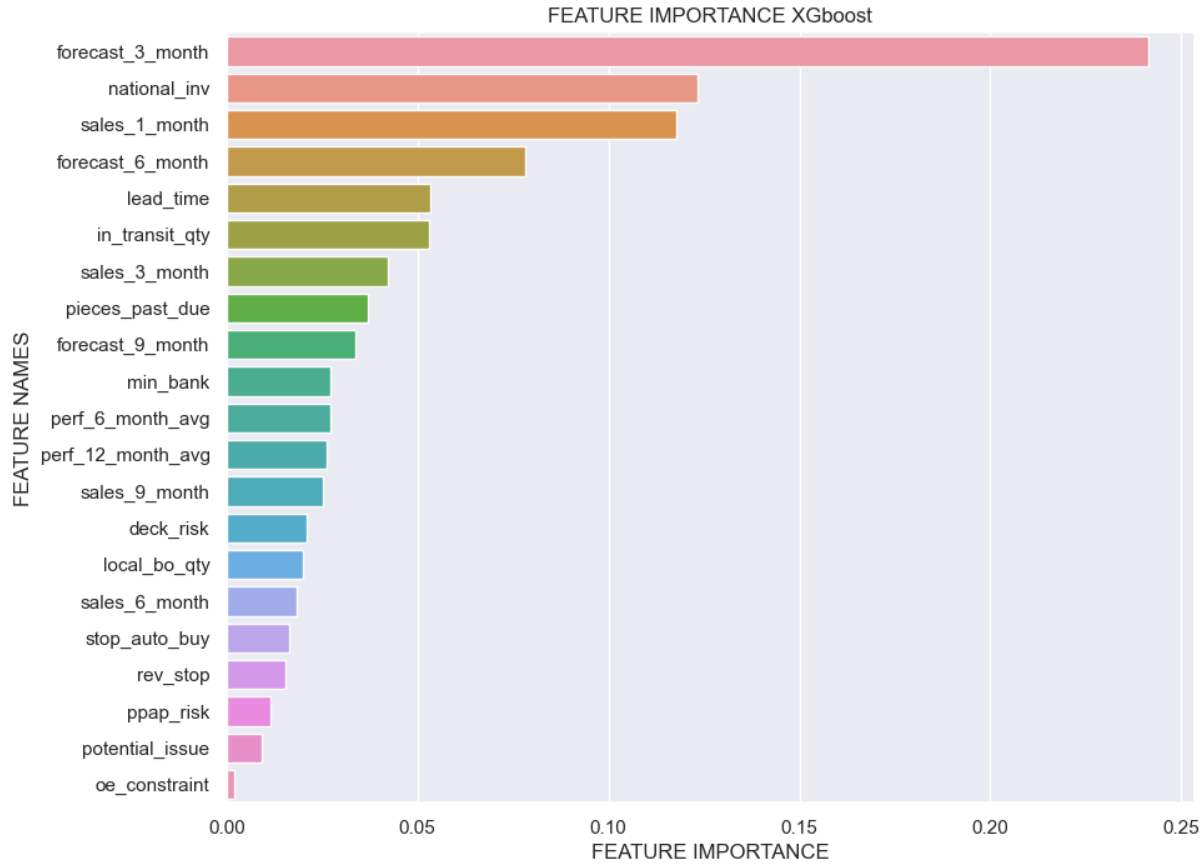

Confusion matrix @0.50
Πραγματικό label
0
1
383013
178
2302
494
350000
300000
250000
200000
150000
100000
50000
0
1
Predicted label

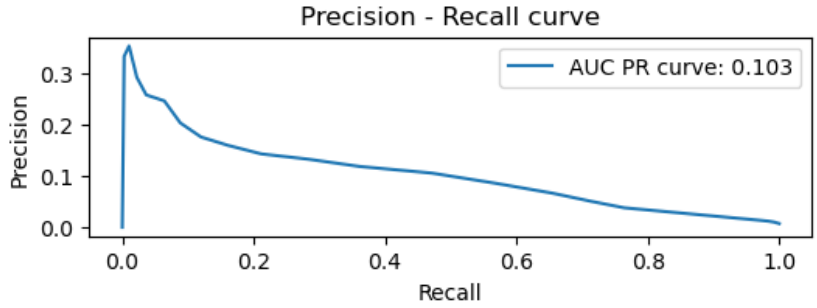
Figure 31: XG Boost σε δεύτερο σετ δεδομένων

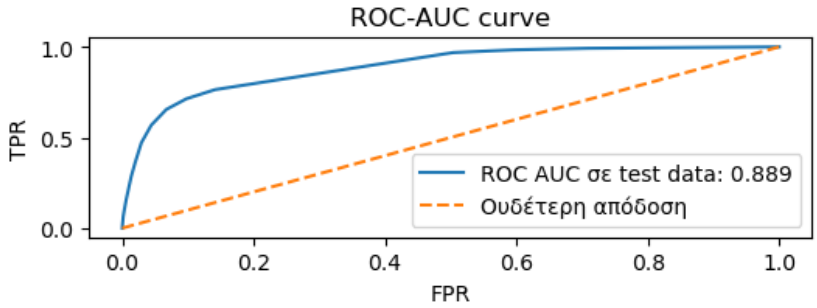



FEATURE IMPORTANCE XGboost
forecast_3_month
national_inv
sales_1_month
forecast_6_month
lead_time
in_transit_qty
sales_3_month
pieces_past_due
forecast_9_month
min_bank
perf_6_month_avg
perf_12_month_avg
sales_9_month
deck_risk
local_bo_qty
sales_6_month
stop_auto_buy
rev_stop
ppap_risk
potential_issue
oe_constraint
FEATURE NAMES
0.00
0.05
0.10
0.15
0.20
0.25
FEATURE IMPORTANCE

MLP - Multilayer Perceptron NN

Figure 32: MLP σε πρώτο σετ δεδομένων





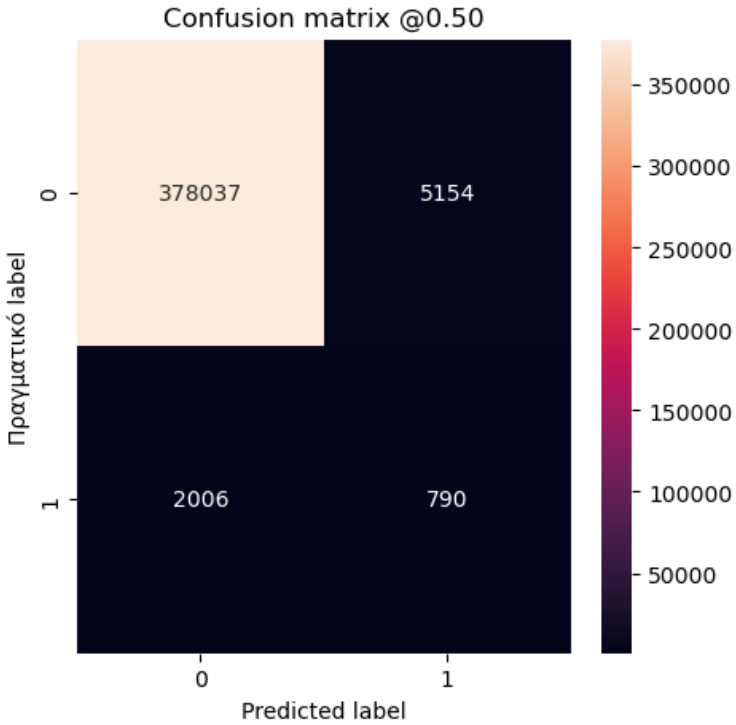


Figure 33: Feature Importance MLP

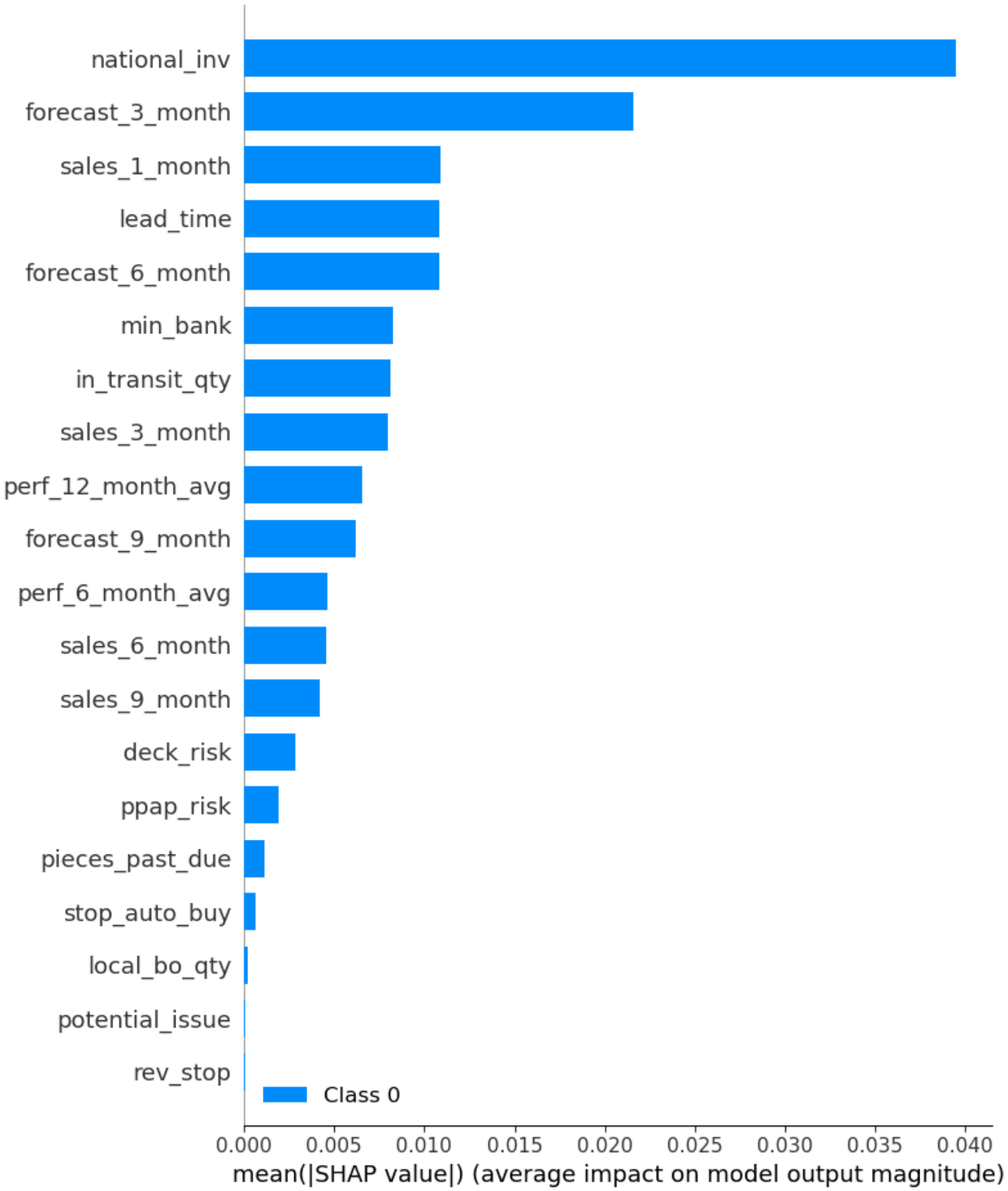


Figure 34: MLP σε δεύτερο σετ δεδομένων

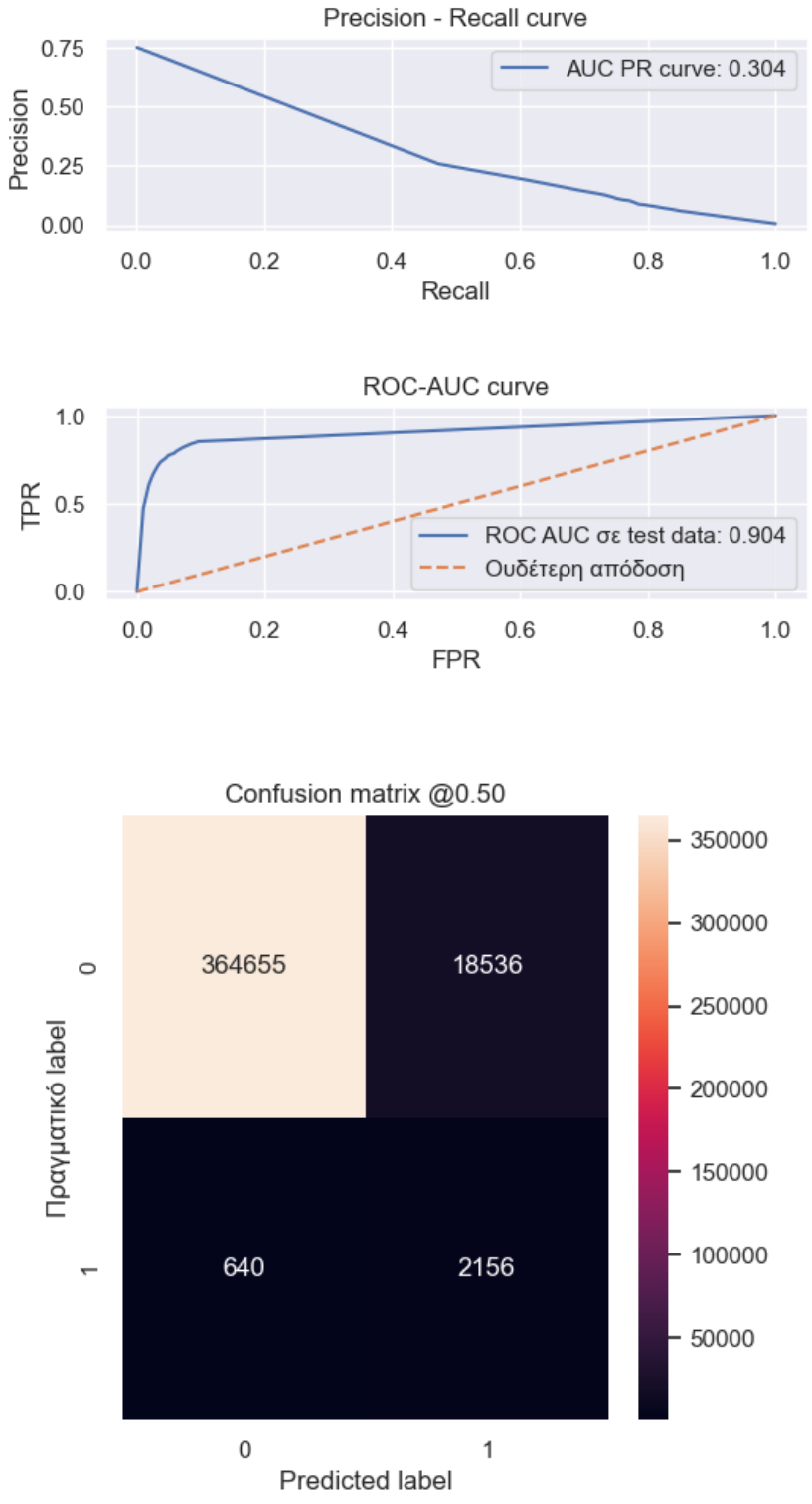
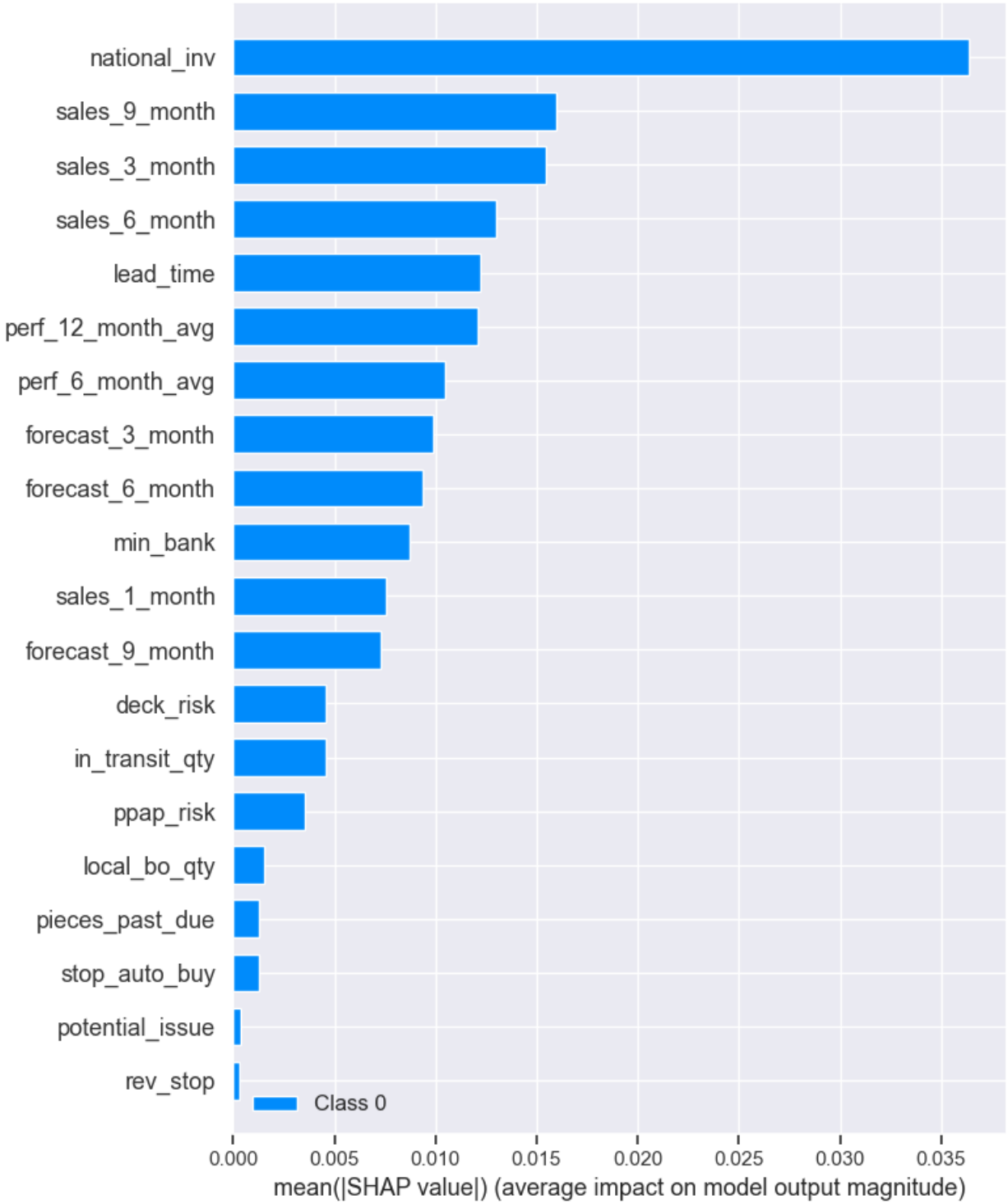


Figure 35: Feature Importance MLP



LSTM - Long Short-Term Memory NN

Figure 36: LSTM σε πρώτο σετ δεδομένων

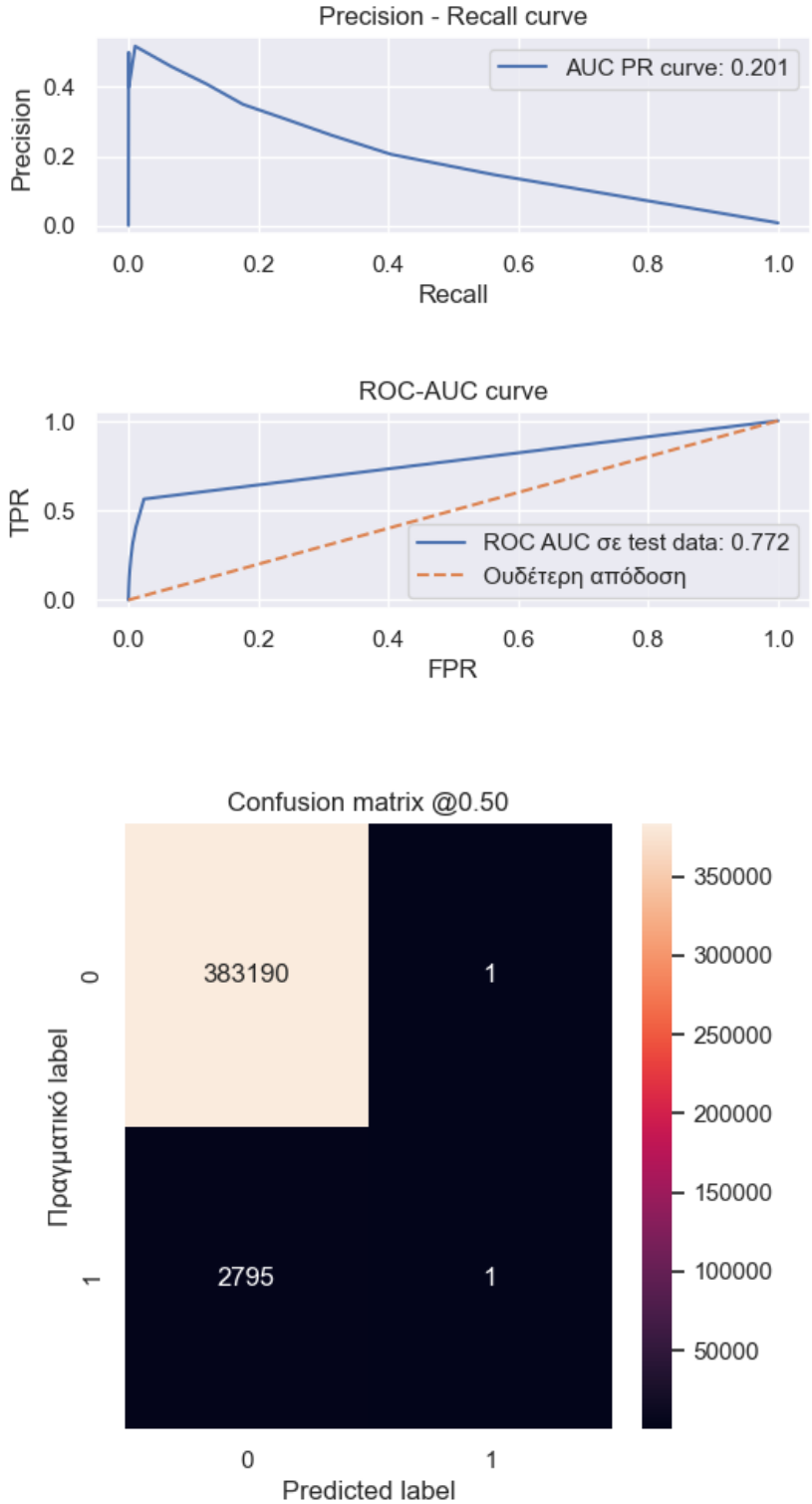
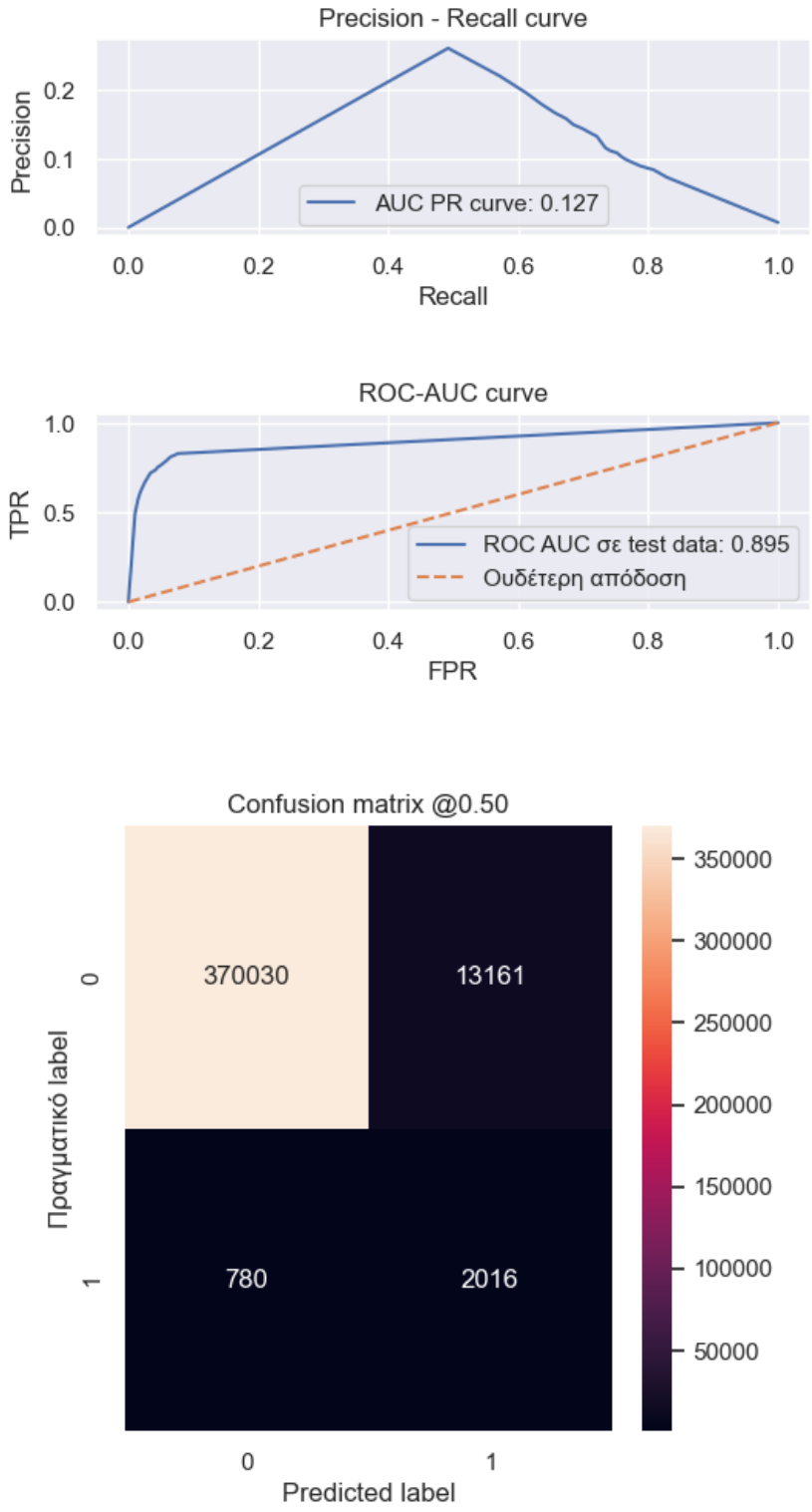


Figure 37: LSTM σε δεύτερο σετ δεδομένων



Αποτελέσματα

Ακολουθεί πίνακας που συγκρίνει τα true positives, τα false positives, τα macro F1 score, τα Precision-Recall AUC score και τα Receiver Operating Characteristic score όλων των μοντέλων. Τα αποτελέσματα αφορούν το evaluation που έχει γίνει στα test σετ. Τα συνολικά backorders είναι 2796.

Μοντέλα	True Positives	False Positives	Macro F1	PR-AUC	ROC-AUC
Μοντέλα με χρήση weights					
Logistic Regression	2	17	0.5	0.056	0.501
Random Forest	2444	56628	0.5	0.175	0.934
XGBoost	494	178	0.64	0.398	0.838
MLP NN	790	5154	0.59	0.103	0.889
LSTM NN	1	1	0.5	0.201	0.772
Log Transformation + Smote Tomek					
Logistic Regression	2437	55797	0.5	0.102	0.932
Random Forest	2400	50751	0.51	0.186	0.933
XGBoost	689	1566	0.63	0.197	0.858
MLP NN	2156	18536	0.58	0.304	0.904
LSTM NN	2016	13161	0.60	0.127	0.895
Καλύτερο Μοντέλο: XGBoost					

Όπως είναι αναμενόμενο το μοντέλο της λογιστικής παλινδρόμησης δεν ήταν κατάλληλο για την φύση του προβλήματός μας. Κατάφερε να διακρίνει μόνο 2 backorder και είχε ROC-AUC score 0.5 και πολύ κακό PR score. Με Smote Tomek κατάφερε να σημειώσει ένα εντυπωσιακό αριθμό True Positives (2437) και υψηλό ROC score 0.932. Όμως είχε τεράστιο συγκριτικά αριθμό λάθος προβλέψεων False Positives 55797.

Το Random Forest μοντέλο αποδίδει καλυτέρα με ROC score 0.934 και ταυτόχρονα έχει εντοπίσει σχεδόν όλο τον αριθμό των backorders. Παρόλα αυτά δεν μπορεί να θεωρηθεί ιδανικό μοντέλο, καθώς έχει τοποθετήσει ένα μεγάλο πλήθος παρατηρήσεων λανθασμένα ως backorders (False Positives). Να σημειωθεί ότι έγινε χρήση balanced subsample ώστε το δένδρο να εκπαιδευτεί το ίδιο καλά και στις 2 κλάσεις. Με χρήση Smote Tomek παρατηρούμε ελάχιστη βελτίωση στα αποτελέσματα, αλλά τίποτα αξιοσημείωτο.

Το XGBoost μοντέλο απόδωσε ακόμα καλύτερα, με macro F1 score 0.64 και ROC score 0.838. Κατάφερε να βρει ένα σεβαστό ποσοστό των backorders κρατώντας ταυτόχρονα τα False Positives πολύ χαμηλά (178). Με Smote Tomek μπορούμε να πούμε έπεσε λίγο η επίδοση του, γιατί παρόλο που σημείωσε καλύτερο ROC score 0.858 και 689 True Positives, τα False Positives αυξήθηκαν από 178 σε 1566.

Στην περίπτωση των νευρωνικών δικτύων είχαμε ανάμεικτα αποτελέσματα. Το MLP με default weights για τις κλάσεις δεν κατάφερε να διακρίνει τα backorders. Με χρήση weight 1 προς 50 για την θετική κλάση, σημείωσε macro F1 score 0.59 και ROC score 0.889. Εντόπισε 790 από τα backorders, αλλά χαρακτήρισε λάθος 5154 προϊόντα ως backorders. Με Smote Tomek η επίδοση έμεινε αναλογικά σε παρόμοια επίπεδα, βρίσκοντας τριπλάσιο αριθμό True Positives αλλά και False Positives.

Από την άλλη, το LSTM μοντέλο δεν απέδωσε καθόλου καλά στο πρώτο σετ δεδομένων, αλλά με Smote Tomek resampling κατάφερε να αποδώσει λίγο καλύτερα από το MLP, με 0.6 macro F1 score, 0.895 ROC score, 2016 TP και 13161 FP.

Ως καλύτερο μοντέλο θα χαρακτηρίζαμε το XGBoost, το οποίο είναι γνωστό ότι ανταποκρίνεται με επιτυχία σε προβλήματα ανισότητας μεταξύ κλάσεων, γεγονός που αντικατοπτρίζεται και στα αποτελέσματα. Κατάφερε να ξεπεράσει σε απόδοση και το Random Forest και τα νευρωνικά δίκτυα. Η περιπλοκότατα και η φύση του «black box» των νευρωνικών δικτύων δεν είναι πάντα προτιμότερη για την επίλυση προβλημάτων, καθώς εκπαιδεύονται με πάρα πολλές παραμέτρους και συχνά το αποτέλεσμα είναι απρόβλεπτο. Επομένως απαιτούν αρκετό πειραματισμό και tuning των παραμέτρων μέχρι να ρυθμιστούν ιδανικά για το σετ των δεδομένων μας.

Σαν γενικό συμπέρασμα μπορούμε να πούμε ότι στην περίπτωση που είχαμε κάνει χρήση της υβριδικής τεχνικής resampling Smote Tomek, δεν παρατηρήσαμε γενική βελτίωση στην απόδοση των μοντέλων, γιατί παρά το γεγονός πως πλέον δεν υπάρχει η ανισότητα μεταξύ των κλάσεων, είναι πολύ πιο εύκολο να κάνουν τα μοντέλα overfit προς την θετική κλάση, η οποία δεν είναι πια σπάνια. Εκτός αυτού, κάναμε χρήση weights κατά την διάρκεια του training στο πρώτο dataset, οπότε είχαμε αντιμετωπίσει έτσι σε κάποιο βαθμό την ανισότητα.

Μία τεχνική που αξίζει να δοκιμαστεί σύμφωνα με την Google σε περίπτωση imbalanced data, είναι να γίνει downsampled και ταυτόχρονα upweight η αρνητική κλάση (σε αντίστοιχο βαθμό). Έτσι μπορούμε να πετύχουμε ταχύτερη σύγκλιση. Κατά τη διάρκεια της εκπαίδευσης, βλέπουμε πιο συχνά την τάξη της μειοψηφίας, κάτι που θα βοηθήσει το μοντέλο να συγκλίνει πιο γρήγορα. Επιπρόσθετα, εξασφαλίζουμε καλύτερη βαθμονόμηση. Το ανοδικό βάρος διασφαλίζει ότι το μοντέλο μας εξακολουθεί να είναι βαθμονομημένο και τα αποτελέσματα μπορούν ακόμα να ερμηνευθούν ως πιθανότητες.

Docker και Containers

Το Docker παρέχει τη δυνατότητα συσκευασίας και εκτέλεσης εφαρμογών σε ένα απομονωμένο περιβάλλον που ονομάζεται κοντέινερ. Είναι δυνατή η εκτέλεση πολλών κοντέινερ ταυτόχρονα σε έναν συγκεκριμένο κεντρικό υπολογιστή.

Τα κοντέινερ απλοποιούν την ανάπτυξη και την παράδοση κατανεμημένων εφαρμογών. Γίνονται ολοένα και πιο δημοφιλείς καθώς οι οργανισμοί στρέφονται προς την εγγενή cloud ανάπτυξη και τα υβριδικά multicloud περιβάλλοντα. Είναι ελαφριά και περιέχουν όλα όσα χρειάζονται για την εκτέλεση της εφαρμογής, χωρίς ο χρήστης να χρειάζεται να επέμβει στον υπολογιστή του, καθώς όλο το περιβάλλον εκτέλεσης είναι ήδη ρυθμισμένο και λειτουργεί ανεξάρτητα από το περιβάλλον εργασίας του.

Χρήση Docker για:

- Γρήγορη, συνεπής παράδοση των εφαρμογών: Βελτιστοποιεί τον κύκλο ζωής του development επιτρέποντας στους προγραμματιστές να εργάζονται σε τυποποιημένα περιβάλλοντα χρησιμοποιώντας τοπικά κοντέινερ που παρέχουν τις εφαρμογές και τις υπηρεσίες. Τα containers είναι ιδανικά για ροές εργασίας συνεχούς παράδοσης (CI / CD).
- Η πλατφόρμα με βάση τα κοντέινερ του Docker επιτρέπει εξαιρετικά φορητούς φόρτους εργασίας. Τα κοντέινερ Docker μπορούν να εκτελούνται σε τοπικό φορητό υπολογιστή ενός προγραμματιστή, σε φυσικές ή εικονικές μηχανές (vm) σε ένα κέντρο δεδομένων, σε παρόχους cloud ή σε ένα συνδυασμό περιβαλλόντων. Η φορητότητα του και η ελαφριά φύση του διευκολύνουν επίσης τη δυναμική διαχείριση του φόρτου εργασίας, κλιμακώνοντας ή καταστρέφοντας εφαρμογές και υπηρεσίες, όπως υπαγορεύουν οι επιχειρηματικές ανάγκες, σε σχεδόν πραγματικό χρόνο.
- Εκτελεί περισσότερους φόρτους εργασίας στο ίδιο hardware: Είναι ελαφρύ και γρήγορο. Παρέχει μια βιώσιμη, οικονομικά εναλλακτική λύση για virtual machines. Είναι ιδανικό για περιβάλλοντα υψηλής πυκνότητας και για μικρές και μεσαίες εφαρμογές όπου υπάρχουν αρκετές απαιτήσεις αλλά ελάχιστοι πόροι.

Kubernetes

Οι Kubernetes είναι μια φορητή, επεκτάσιμη πλατφόρμα ανοιχτού κώδικα για τη διαχείριση φορτίων εργασίας και υπηρεσιών με εμπορευματοκιβώτια, που διευκολύνει τόσο τη δηλωτική ρύθμιση όσο και τον αυτοματισμό. Έχει ένα μεγάλο, ταχέως αναπτυσσόμενο οικοσύστημα και οι υπηρεσίες, η υποστήριξη και τα εργαλεία Kubernetes είναι ευρέως διαθέσιμα. Έχουν πολλές χρήσεις στα containers, όπως π.χ. την επανεκκίνηση των κοντέινερ που αποτυγχάνουν, την αντικατάσταση τους και τον τερματισμό λειτουργίας όσων δεν ανταποκρίνονται στον έλεγχο που καθορίζεται από το χρήστη.

Ακόμη βοηθούν σε:

- Ανακάλυψη υπηρεσίας και εξισορρόπηση φορτίου: Οι Kubernetes μπορούν να εκθέσουν ένα κοντέινερ χρησιμοποιώντας το όνομα DNS ή χρησιμοποιώντας τη δική του διεύθυνση IP. Εάν η κίνηση προς ένα κοντέινερ είναι υψηλή, οι Kubernetes μπορούν να φορτώσουν την ισορροπία και να διανείμουν την κυκλοφορία του δικτύου έτσι ώστε το deployment να είναι σταθερό.
- Επιτρέπουν την αυτόματη προσαρμογή ενός συστήματος αποθήκευσης της επιλογής, όπως τοπικές αποθήκες, δημόσιους παρόχους cloud και άλλα.

Notebook Docker Image

Ένας πολύ εξυπηρετικός τρόπος για την εκτέλεση και το μοίρασμα κώδικα που περιλαμβάνει πληθώρα βιβλιοθηκών είναι η δημιουργία ενός Image του συνολικού περιβάλλοντος εργασίας. Έτσι εξασφαλίζεται η συμβατότητα με όλους τους χρήστες. Για παράδειγμα, μπορούμε να φτιάξουμε ένα Docker Image του Jupyter Notebook της εργασίας, που θα περιλαμβάνει τον κώδικα και όλα τα μοντέλα που έχουμε εκπαιδεύσει και θα είναι έτοιμο για τρέξιμο.

Αυτό που θα χρειαστεί να κάνουμε ουσιαστικά είναι να δημιουργήσουμε ένα dockerfile, το οποίο θα περιέχει τις οδηγίες για την δημιουργία του image μας. Αρχικά ορίζουμε την τελευταία έκδοση των ubuntu ως το λειτουργικό στο οποίο θα τρέχει το Notebook και εγκαθιστούμε την έκδοση 3.10 της python. Στην συνέχεια ορίζουμε το work directory και κάνουμε copy τους φακέλους που θέλουμε να περιλαμβάνει. Στην περίπτωσή μας, κάνουμε αντιγραφή το Notebook αρχείο, τα δεδομένα, τα εκπαιδευμένα μοντέλα και το yml αρχείο με το conda environment που εργαστήκαμε. Έπειτα κάνουμε pip install όλες τις βιβλιοθήκες που χρειάζονται και συγκεκριμένα τις εκδόσεις τις οποίες έχουμε εγκατεστημένες στο conda περιβάλλον. Τέλος, ορίζουμε την CMD εντολή, η οποία καθορίζει τις εντολές που πρέπει να εκτελεστούν κατά την εκκίνηση του κοντέινερ, όπως π.χ. το container να τρέχει στο port 8888.

Αφού κάνουμε build το image μπορούμε να τρέξουμε το κοντέινερ με την εντολή:
docker run -it -p 8888:8888 me2052-backorder-prediction-thesis

Κβαντικοί Υπολογιστές

Στη φυσική, ένα κβάντο είναι η μικρότερη δυνατή διακριτή μονάδα οποιασδήποτε φυσικής ιδιότητας. Συνήθως αναφέρεται σε ιδιότητες ατομικών ή υποατομικών σωματιδίων, όπως ηλεκτρόνια, νετρίνα και φωτόνια. Κβαντικοί υπολογιστές είναι ουσιαστικά οι υπολογιστές οι οποίοι αξιοποιούν τα φαινόμενα της κβαντικής φυσικής για την λειτουργία τους. Μερικά από αυτά τα φαινόμενα είναι τα Superposition, entanglement και Quantum interference.
Σε αντίθεση με τους κλασικούς digital υπολογιστές οι οποίοι λειτουργούν με βάση τα bits ως μονάδα πληροφοριών, τα οποία είναι δυαδικά και μπορούν να κρατήσουν μόνο μια θέση (0 ή 1), οι κβαντικοί υπολογιστές βασίζονται στα qubits. Τα qubits μπορούν να κρατήσουν ένα superposition όλων των πιθανών θέσεων, δηλαδή ένα συνδυασμό όλων των θέσεων.
Entanglement είναι η ικανότητα των κβαντικών σωματιδίων να συσχετίζουν τα αποτελέσματα των μετρήσεών τους μεταξύ τους. Όταν τα qubits είναι entangled, σχηματίζουν ένα ενιαίο σύστημα και επηρεάζουν το ένα το άλλο. Μπορούμε να χρησιμοποιήσουμε τις μετρήσεις από ένα qubit για να βγάλουμε συμπεράσματα για τα άλλα. Προσθέτοντας και εμπλέκοντας περισσότερα qubits σε ένα σύστημα, οι κβαντικοί υπολογιστές μπορούν να υπολογίσουν εκθετικά περισσότερες πληροφορίες και να λύσουν πιο περίπλοκα προβλήματα.
Quantum interference είναι η εγγενής συμπεριφορά ενός qubit, λόγω superposition, να επηρεάζει την πιθανότητα κατάρρευσής του. Οι κβαντικοί υπολογιστές έχουν σχεδιαστεί και κατασκευαστεί για να μειώνουν όσο το δυνατόν περισσότερο τις παρεμβολές και να διασφαλίζουν τα πιο ακριβή αποτελέσματα. Για παράδειγμα, η Microsoft χρησιμοποιεί τοπολογικά qubits, τα οποία σταθεροποιούνται χειραγωγώντας τη δομή τους και περιβάλλοντάς τα με χημικές ενώσεις που τα προστατεύουν από εξωτερικές παρεμβολές.

Ένας κβαντικός υπολογιστής έχει τρία κύρια μέρη: Μια περιοχή που στεγάζει τα qubits, μια μέθοδο για τη μεταφορά σημάτων στα qubits και έναν κλασικό υπολογιστή για την εκτέλεση ενός προγράμματος και την αποστολή οδηγιών. Υπάρχουν διάφοροι τρόποι στέγασης των qubit, όπως η μονάδα που φιλοξενεί τα qubit να διατηρείται σε θερμοκρασία ακριβώς πάνω από το απόλυτο μηδέν για να μεγιστοποιηθεί η συνοχή τους και να μειωθούν οι παρεμβολές. Άλλος τρόπος είναι η χρήση θαλάμου κενού για την ελαχιστοποίηση των κραδασμών και τη σταθεροποίηση των qubits. Τα σήματα μπορούν να σταλούν στα qubits χρησιμοποιώντας μια ποικιλία μεθόδων, συμπεριλαμβανομένων των μικροκυμάτων, του laser και της τάσης.

Μερικές από τις χρήσεις και περιοχές εφαρμογής κβαντικών υπολογιστών που έχουν τη δυνατότητα να έχουν μεγάλο αντίκτυπο είναι:

- Quantum machine learning: Η εκπαίδευση μοντέλων μηχανικής εκμάθησης έχει υψηλό υπολογιστικό κόστος και αυτό έχει εμποδίσει το εύρος και την ανάπτυξη του πεδίου. Για να επιταχυνθεί η πρόοδος σε αυτόν τον τομέα, αναπτύσσονται τρόποι για γραφτεί και να εφαρμοστεί κβαντικό λογισμικό που επιτρέπει την ταχύτερη μηχανική εκμάθηση.
- Quantum simulation: Λειτουργούν εξαιρετικά καλά για τη μοντελοποίηση άλλων κβαντικών συστημάτων επειδή χρησιμοποιούν κβαντικά φαινόμενα στους υπολογισμούς τους. Αυτό σημαίνει ότι μπορούν να χειριστούν την πολυπλοκότητα και την ασάφεια των συστημάτων που θα υπερφόρτωναν τους κλασικούς υπολογιστές. Παραδείγματα περιλαμβάνουν τη φωτοσύνθεση, την υπεραγωγιμότητα και τους πολύπλοκους μοριακούς σχηματισμούς.

- Optimization: Η βελτιστοποίηση είναι η διαδικασία εύρεσης της βέλτιστης λύσης σε ένα πρόβλημα, δεδομένων του επιθυμητού αποτελέσματος και των περιορισμών του. Στην επιστήμη και τη βιομηχανία, οι κρίσιμες αποφάσεις λαμβάνονται με βάση παράγοντες όπως το κόστος, η ποιότητα και ο χρόνος παραγωγής. Εκτελώντας αλγόριθμους βελτιστοποίησης εμπνευσμένους από κβαντικά στοιχεία σε κλασικούς υπολογιστές, μπορούμε να βρούμε λύσεις που προηγουμένως ήταν αδύνατες. Αυτό μας βοηθά να βρούμε καλύτερους τρόπους διαχείρισης πολύπλοκων συστημάτων, όπως οι ροές κυκλοφορίας, οι παραδόσεις πακέτων και η αποθήκευση ενέργειας.
- Cryptography

Παρόλο που οι κβαντικοί υπολογιστές βρίσκονται ακόμα σε πειραματικό στάδιο λειτουργίας και έχουν να αντιμετωπιστούν αρκετά προβλήματα για να μπορούν να θεωρηθούν αποδοτικοί (ένα από τα μεγαλύτερα είναι η τεράστια ευαισθησία που έχουν στο θόρυβο), έχουν καταφέρει να φτάσουν στο σημείο του «Quantum supremacy», δηλαδή να εκτελέσουν μια διαδικασία που ένας κλασικός υπολογιστής δεν μπορεί. Καθώς οι κλίμακες του κβαντικού hardware και οι κβαντικοί αλγόριθμοι προχωρούν, για πολλά μεγάλα, σημαντικά προβλήματα θα βρεθούν λύσεις. Αυτήν την στιγμή υπάρχει ένας αγώνας μεταξύ εταιριών κολοσσούς στον τομέα της τεχνολογίας, όπως η IBM, η Microsoft, η Google, κινεζικών εταιριών, κ.α., για το ποια θα καταφέρει να φτιάξει τον καλύτερο κβαντικό υπολογιστή και να επικρατήσει/αποκτήσει προβάδισμα στον χώρο. Για παράδειγμα, η IBM έχει ήδη πάνω από 20 κβαντικά συστήματα διαθέσιμα στο cloud στην Αμερική και το 2024 σκοπεύει να ανοίξει Quantum data center στην Γερμανία.

Επίλογος

Ο τομέας της ανάλυσης δεδομένων, της μηχανικής μάθησης και γενικότερα της τεχνικής νοημοσύνης, αναπτύσσεται χρόνια και συνεχίζει να βελτιώνεται και να εξελίσσεται συνεχώς. Τα εργαλεία που υπάρχουν πλέον στην διάθεση μας είναι ποικίλα, γεγονός που απαιτεί και περισσότερο χρόνο για την εξοικείωση μας με αυτά.
Η επανάσταση που υπήρξε στον τομέα της βιομηχανίας με τον ερχομό των ηλεκτρονικών υπολογιστών ήταν συγκλονιστική και το μεταβατικό στάδιο που διανύουμε αυτήν την στιγμή με την ραγδαία πρόοδο στον τομέα της τεχνικής νοημοσύνης αλλά και τον ερχομό των κβαντικών υπολογιστών θα οδηγήσει σε μια αντίστοιχης σημασίας επανάσταση ίσως και μεγαλύτερη. Προβλήματα σε όλους τους τομείς της βιομηχανίας θα αντιμετωπίζονται ευκολότερα και με μεγαλύτερη αποτελεσματικότητα και οι διαδικασίες θα είναι γρηγορότερες, αποδοτικότερες και πολλαπλές.

Στην περίπτωσή μας, αναλύσαμε ένα πρόβλημα κατηγοριοποίησης με πολύ άνισα δεδομένα, που αφορά ένα από τα σημαντικότερα προβλήματα που προσπαθούν να αντιμετωπίσουν οι εταιρίες αυτήν την στιγμή, το backorder prediction. Το αντιμετωπίσαμε με δύο διαφορετικές προσεγγίσεις όσον αφορά τα δεδομένα. Μία με όσον το δυνατό λιγότερη παρέμβαση στα πρωτότυπα δεδομένα και μία με δεδομένα που έχουν υποστεί μεγάλη επεξεργασία.

Με τις προαναφερόμενες τεχνολογίες που θα επηρεάσουν σε μεγάλο βαθμό όλο τον κλάδο των logistics, οι προβλέψεις τον backorders θα γίνουν πολύ πιο ακριβείς. Αυτό θα συμβεί γιατί εκτός από την εξέλιξη των μεθόδων και της υπολογιστής ισχύς που θα έχουμε στην διάθεσή μας για την διαδικασία κατασκευής καλυτέρων μοντέλων πρόβλεψης, θα έχουμε ταυτόχρονη βελτίωση σε υποδομές, διαχείριση αποθεμάτων, γρηγορότερη και φθηνότερη παραγωγή και μεταφορά κλπ.. Γεγονότα που θα μειώσουν τα backorders, αλλά και το οικονομικό penalty στην περίπτωση ύπαρξης αυτών, καθώς όλο το δίκτυο θα είναι πιο αποδοτικό.

Hardware και Software

Hardware

1. Gigabyte Aorus Master, Intel i9-10900K CPU @ 3.70 GHz, 32GB RAM, 1000GB SSD, NVIDIA RTX 3080 Ti GPU 12 GB RAM | Για εκπαίδευση των μοντέλων
2. Lenovo mini-PC, Intel i7-12700T CPU @ 1.40 GHz, 16GB RAM, 512GB SSD | Για λοιπές εργασίες

Operating Systems

1. Windows 10 Pro 22H2 64-bit
2. Windows 11 Pro 21H2 64-bit

Εργαλεία που χρησιμοποιήθηκαν

- Azure Data Studio
- PyCharm Professional Edition
- Anaconda3/Jupyter Notebook
- Notepad++
- Microsoft Azure
- Docker

Πηγές

1) Kostiantyn Bokhan, “Machine learning in supply chain: 8 use cases that will impress you”, n-ix, November 03 2020, https://www.n-ix.com/machine-learning-supply-chain-use-cases/
2) Schwab Klaus, “The Fourth Industrial Revolution”, Foreign Affairs, 2015
3) Schwab Klaus, “foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution, weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-andhow-to respond”, December 2015
4) McCulloch, Warren. and Pitts, Walter**,** “A logical calculus of the ideas immanent in nervous activity”. Bulletin of Mathematical Biophysics, 1943
5) Werbos, Paul J., “The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting”, New York: John Wiley & Sons, 1994
6) “Machine Learning”, Wikipedia, https://en.wikipedia.org/wiki/Machine_learning
7) “Supply Chain”, Wikipedia, https://en.wikipedia.org/wiki/Supply_chain
8) I-Scoop, The Fourth Industrial Revolution: guide Industry 4.0, 2016
9) Πετράκης Παναγιώτης, Περιγραφή και Εισαγωγική Ανάλυση της 4ης Βιομηχανικής Επανάστασης, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
10) Παναγόπουλος Δημ., Μηχανική Μάθηση (Machine Learning), Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
11) Παναγόπουλος Δημ., Τεχνητή Νοημοσύνη (Artificial Intelligence), Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
12) Srinivasa Raja, “Backorder Prediction”, Analytics Vidhya, Mar 4, 2021
13) Chinmay Dalvi, “Backorder Prediction using Machine Learning”, Medium, Jun 15, 2021
14) Samir Saci, “Machine Learning for Retail Demand Forecasting”, towardsdatascience, Aug 21 2020
15) Mahya Seyedan & Fereshteh Mafakheri, “Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities”, journalofbigdata/springeropen, 25 July 2020
16) Réal André Carbonneau, Rustam Vahidov, Kevin Laframboise, “Machine Learning-Based Demand Forecasting in Supply Chains”, ResearchGate, January 2007
17) Jason Brownlee, ” Undersampling Algorithms for Imbalanced Classification”, Machine Learning Mastery, January 20 2020
18) IBM Cloud Education, “What is Artificial Intelligence (AI)?”, IBM, 3 June 2020, https://www.ibm.com/cloud/learn/what-is-artificial-intelligence
19) Tatko, Ryan W., White, Christopher R., Wiliams, Geoffrey M., Myers, Robin K., “Scientific and Technical Report Backorder Prediction for defense logistics agency research and development”, IBM, 26 February 2021
20) B.J. Copeland, “artificial intelligence”, Britannica, https://www.britannica.com/technology/artificial-intelligence
21) “What is supply chain management?”, IBM, https://www.ibm.com/topics/supply-chain-management
22) “Survey: 82% of Americans Scared That Supply Chain Issues Will Ruin Their Life Plans”, Oracle, September 29 2021, https://www.oracle.com/news/announcement/survey-of-americans-scared-of-supply-chain-issues-2021-09-29/
23) “What Is Deep Learning?”, MathWorks, https://www.mathworks.com/discovery/deep-learning.html

24) Jason Brownlee, ”How to Choose an Activation Function for Deep Learning”, Machine Learning Mastery, January 18 2021, https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/
25) Niklas Donges, “Random Forest Algorithm: A Complete Guide”, builtin, July 22 2021, https://builtin.com/data-science/random-forest-algorithm
26) “Unsupervised Learning”, IBM, 21 September 2020, https://www.ibm.com/cloud/learn/unsupervised-learning
27) Vihar Kurama, “Gradient Boosting in Classification: Not a Black Box Anymore!”, PaperspaceBlog, https://blog.paperspace.com/gradient-boosting-for-classification/
28) Saban Adana, Sedat Cevikparmak, Hasan Celik, Hasan Uvet, “Predicting Backorders Using Machine Learning Techniques”, ResearchGate, November 2019, https://www.researchgate.net/publication/339954197_Predicting_Backorders_Using_Machine_Learning_Techniques (not used yet)
29) Yıldırım Demir, Masoud M. Hassan, “Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data”, Bitlis Eren Üniversitesi Fen Bilimleri Dergisi, July 2021, https://www.researchgate.net/publication/354606592_Combination_of_PCA_with_SMOTE_Oversampling_for_Classification_of_High-Dimensional_Imbalanced_Data
30) Niwratti Kasture, “Why Hyper parameter tuning is important for your model?”, Analytics Vidhya, Nov 16, 2020, https://medium.com/analytics-vidhya/why-hyper-parameter-tuning-is-important-for-your-model-1ff4c8f145d3
31) Ibomoiye Domor Mienye, Yanxia Sun, “Performance analysis of cost-sensitive learning methods with application to imbalanced medical data”, ScienceDirect, 5 August 2021, https://www.sciencedirect.com/science/article/pii/S235291482100174X
32) Yugesh Verma, “Why Data Scaling is important in Machine Learning & How to effectively do it”, Developers Corner, August 29 2021, https://analyticsindiamag.com/why-data-scaling-is-important-in-machine-learning-how-to-effectively-do-it/
33) “Data Mining Process: Models, Process Steps & Challenges Involved”, Software Testing Help, July 16 2022, https://www.softwaretestinghelp.com/data-mining-process/
34) Hairani, Anthony Anggrawan, Dadang Priyanto, “Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link”, ResearchGate, March 2023, https://www.researchgate.net/publication/368993605_Improvement_Performance_of_the_Random_Forest_Method_on_Unbalanced_Diabetes_Data_Classification_Using_Smote-Tomek_Link
35) “What are recurrent neural networks?”, IBM, https://www.ibm.com/topics/recurrent-neural-networks
36) “What are convolutional neural networks?”, IMB, https://www.ibm.com/topics/convolutional-neural-networks
37) Prasad Ostwal, “Principal Component Analysis Visualization”, Jan 20, 2019, https://ostwalprasad.github.io/machine-learning/PCA-using-python.html
38) Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals, “UNDERSTANDING DEEP LEARNING REQUIRES RE-THINKING GENERALIZATION”, Google Brain, 26 Feb 2017, https://arxiv.org/pdf/1611.03530.pdf
39) Paula Villasante Soriano, Cansu Kebabci, “What is a Principal Component Analysis (PCA)?”, Statistics Globe, https://statisticsglobe.com/principal-component-analysis-pca

40) Jason Brownlee, “How to Calibrate Probabilities for Imbalanced Classification”, Machine Learning Mastery, 26 Feb 2020, https://machinelearningmastery.com/probability-calibration-for-imbalanced-classification/
41) Jason Brownlee, “ROC Curves and Precision-Recall Curves for Imbalanced Classification”, Machine Learning Mastery, 16 Sep 2020, https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/
42) Enes Zvornicanin, “Differences Between Bidirectional and Unidirectional LSTM”, Baeldung, June 8 2023, https://www.baeldung.com/cs/bidirectional-vs-unidirectional-lstm
43) Rukshima Dabare, Kok Wai Wong, Polychronis Koutsakis, Mohd Fairuz Shiratuddin, “A Study of the Effect of Dropout on Imbalanced Data Classification using Deep Neural Networks”, School of Engineering and Information Technology Murdoch University, Oct 2018, https://www.jmest.org/wp-content/uploads/JMESTN42352707.pdf
44) Charles Elkan, “The Foundations of Cost-Sensitive Learning”, Department of Computer Science and Engineering 0114 University of California, https://eva.fing.edu.uy/pluginfile.php/63457/mod_resource/content/1/Elkan_2001_The_foundations_of_cost-sensitive_learning.pdf
45) “What is quantum computing?”, IBM, https://www.ibm.com/topics/quantum-computing
46) “Introduction to quantum computing”, Azure, https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-quantum-computing/?cdn=disable#introduction