Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

<u>Διπλωματική Εργασία</u>

Μέθοδοι Μηχανικής Μάθησης σε προβλήματα ταξινόμησης και η εφαρμογή τους σε τραπεζικά δεδομένα

της

Παυλίδου Ναταλίας του Ιωάννη

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Ιούνιος 2022

Ευχαριστίες

Τελειώνοντας το «ταξίδι» των σπουδών μου και με την συγγραφή της παρούσας διπλωματικής εργασίας, νιώθω την ανάγκη να ευχαριστήσω όλους όσους με το δικό τους τρόπο βοήθησαν και υποστήριξαν αυτή μου την προσπάθεια.

Κατά κύριο λόγο, οφείλω να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου κ. Κωνσταντάρα Ιωάννη για τις πολύτιμες γνώσεις και συμβουλές του, για την ενθάρρυνση και σημαντική καθοδήγησή του προκειμένου να περατωθεί η συγγραφή της εργασίας καθώς και για την ευκαιρία που μου έδωσε να μελετήσω σε βάθος ένα θέμα που με ενδιαφέρει πολύ.

Θα ήθελα να ευχαριστήσω επίσης όλους τους καθηγητές του ΠΜΣ για τις γνώσεις που μου προσέφεραν όλα τα χρόνια των μεταπτυχιακών σπουδών μου.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και όλους τους ανθρώπους που βρίσκονταν δίπλα μου και μου συμπαραστάθηκαν όλο αυτό το διάστημα. Ένα ξεχωριστό ευχαριστώ αξίζει στον Φώτη που με στήριζε, με ενθάρρυνε και μου έδινε τη δύναμη να συνεχίσω να εξελίσσομαι.

Περίληψη

Η παρούσα διπλωματική εργασία προσεγγίζει την επίλυση του προβλήματος της δυαδικής ταξινόμησης με τη χρήση μεθόδων Μηχανικής Μάθησης. Πιο συγκεκριμένα, ο στόχος είναι η δημιουργία ενός βέλτιστου μοντέλου που θα εκπαιδεύεται σωστά στα τραπεζικά δεδομένα και θα είναι σε θέση να προβλέπει αν κάποιο άτομο που προσεγγίζεται κατά τη διάρκεια μιας καμπάνιας μάρκετινγκ θα προχωρήσει σε προθεσμιακή κατάθεση, με βάση τα δεδομένα που έχουν συλλεχθεί γι αυτόν από προηγούμενες καμπάνιες. Το σύνολο των τραπεζικών δεδομένων περιέχει δεδομένα που έχουν συλλεχθεί από μία Πορτογαλική τράπεζα. Πρόκειται για ένα μη ισορροπημένο σύνολο δεδομένων καθώς τα άτομα που έχουν πραγματοποιήσει προθεσμιακή κατάθεση είναι πολύ λιγότερα από τα άτομα που δεν έχουν πραγματοποιήσει. Επιπλέον, μεγάλης σημασίας είναι και η μελέτη της ικανότητας του μοντέλου να προσαρμοστεί σε νέα σύνολα δεδομένων και να κάνει σωστές προβλέψεις σε αυτά. Για την επιλογή του βέλτιστου μοντέλου γίνεται σύγκριση της αποτελεσματικότητας δύο αλγορίθμων με τη χρήση κατάλληλων μετρικών. Τέλος μελετάται η επίδραση ορισμένων μεταβλητών στο τελικό αποτέλεσμα.

Λέξεις κλειδιά: δυαδική ταξινόμηση, Μηχανική Μάθηση, τραπεζικά δεδομένα, μη ισορροπημένο σύνολο, μετρικές, επίδραση μεταβλητών

Abstract

The present thesis approaches the resolution of the problem of binary classification with the application of Machine Learning methods. More specific, the aim is the creation of an optimum model which is going to be trained correctly in bank data in order to predict if one person, who is approached by a marketing campaign, will proceed to term deposit according to the data which have been collected for this person from previous campaigns. The bank dataset contains data which have been collected from a Portuguese bank. It is an imbalanced dataset because the people who have done a term deposit are few in comparison to the people that have not. Moreover, crucial is the study of the capability of the algorithm to be adjusted to new datasets and to make correct predictions on them. For the election of the optimum model, the effectiveness of two algorithms is being compared with the usage of qualified metrics. In conclusion, its being studied the effect of specific variables on the final result.

Keywords: binary classification, Machine Learning, bank data, imbalanced dataset, metrics, variables effect

Περιεχόμενα

1 Εισαγωγή

1.1 Αντικείμενο της Εργασίας 1

1.2 Βιβλιογραφική επισκόπιση 2

1.3 Σχεδιασμός της μελέτης..... 3

1.4 Εργαλεία 4

1.5 Δομή Εργασίας 5

2 Μηχανική Μάθηση

2.1 Μέθοδοι Μηχανικής Μάθησης 7

2.2 Αλγόριθμοι Μηχανικής Μάθησης 10

3 Μη ισορροπημένα σύνολα δεδομένων

3.1 Θεωρητικό υπόβαθρο..... 15

3.2 Μετρικές αξιολόγησης..... 16

3.3 Εφαρμογές στην Επιχειρησιακή Αναλυτική..... 23

3.4 Διαχείριση μη ισορροπημένων συνόλων δεδομένων..... 25

4 Περίπτωση Μελέτης

4.1 Δεδομένα..... 31

4.2 Μεθοδολογία προ-επεξεργασίας.....32

4.3 Περιγραφική Στατιστική..... 33

4.4 Σύγκριση..... 54

4.5 Προ-επεξεργασία δεδομένων..... 68

4.5.1 Αφαίρεση ακραίων τιμών (Outliers) 68

4.5.2 Προ-επεξεργασία ελλειπουσών τιμών με χρήση cross-tabulation..... 69

5 Ταξινόμηση και Αποτελέσματα

5.1 Πρώτη Περίπτωση Μελέτης..... 76

5.1.1 Περίπτωση της ταξινόμησης σε μη ισορροπημένο σύνολο με τη χρήση του συμβατικού αλγορίθμου Logistic Regression..... 77

5.1.2 Περίπτωση της ταξινόμησης σε ισορροπημένο σύνολο δεδομένων..... 79

5.1.3 Περίπτωση της ταξινόμησης σε μη ισορροπημένο σύνολο με τη χρήση τροποποίησης του συμβατικού αλγορίθμου..... 85

5.2. Δεύτερη Περίπτωση Μελέτης..... 89

5.3 Σημαντικότητα μεταβλητών..... 91

6 Συμπεράσματα και Προτάσεις

6.1 Συμπεράσματα..... 93

6.2 Περιορισμοί και αδυναμίες..... 95

6.3 Προτάσεις για μελλοντική έρευνα..... 95

Βιβλιογραφία

Κατάλογος Σχημάτων

Σχήμα 1.1: Η διαδικασία CRISP-DM..... 4

Σχήμα 2.1: Γραφική παράσταση Λογιστικής Παλινδρόμησης..... 11

Σχήμα 2.2: Γραφική αναπαράσταση του αλγορίθμου Support Vector Machines..... 12

Σχήμα 2.3: Οπτική απεικόνιση της μεροληψίας και της διακύμανσης..... 13

Σχήμα 2.4: Bias-Variance trade-off..... 14

Σχήμα 3.1: Σχηματική αναπαράσταση της ROC καμπύλης..... 20

Σχήμα 3.2: Σύγκριση ROC καμπυλών..... 21

Σχήμα 3.3: Σχηματική αναπαράσταση της Precision-Recall καμπύλης..... 22

Σχήμα 3.4: TomekLink εφαρμογή..... 26

Σχήμα 3.5: Γραφική αναπαράσταση των τεχνικών υποδειγματοληψίας και υπερδειγματοληψίας 27

Σχήμα 3.6: Μέθοδοι διαχείρισης ισορροπημένων και μη ισορροπημένων συνόλων δεδομένων με σκοπό την ταξινόμηση..... 30

Σχήμα 4.1: Age Count Distribution (histogram) 34

Σχήμα 4.2: Age Distribution (boxplot) 35

Σχήμα 4.3: Job Count Distribution (barplot) 36

Σχήμα 4.4: Marital Count Distribution (barplot) 37

Σχήμα 4.5: Education Count Distribution (barplot) 38

Σχήμα 4.6: Default, Housing, Loan Count Distribution (barplots) 39

Σχήμα 4.7: Contact, Months, Day of Week Count Distribution (barplots) 39

Σχήμα 4.8: Campaign Count Distribution (barplot) 41

Σχήμα 4.9: Pdays Count Distribution (barplot) 42

Σχήμα 4.10: Previous Count Distribution (barplot) 44

Σχήμα 4.11: Poutcome Count Distribution (barplot) 45

Σχήμα 4.12: Poutcome (propotional pie chart) 45

Σχήμα 4.13: y Distribution (barplot) 48

Σχήμα 4.14: duration Distribution (histogram) 49

Σχήμα 4.15: Consumer Price Index Distribution (histogram) 50

Σχήμα 4.16: Consumer Confidence Index Distribution (histogram) 51

Σχήμα 4.17: Employment Variation Rate Distribution (histogram) 52

Σχήμα 4.18: Euribor3m Distribution (histogram) 53

Σχήμα 4.19: Nr.Employed Distribution (histogram) 54

Σχήμα 4.20: Μέσες τιμές των ποσοτικών μεταβλητών για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y..... 54

Σχήμα 4.21: Κατανομή τιμών της μεταβλητής age για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y..... 55

Σχήμα 4.22: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής job..... 55

Σχήμα 4.23: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής marital..... 56

Σχήμα 4.24: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής education..... 56

Σχήμα 4.25: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής default..... 57

Σχήμα 4.26: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής housing..... 58

Σχήμα 4.27: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής loan..... 58

Σχήμα 4.28: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής contact..... 59

Σχήμα 4.29: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής month..... 60

Σχήμα 4.30: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής day of week..... 60

Σχήμα 4.31: Κατανομή τιμών των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής campaign..... 61

Σχήμα 4.32: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής campaign..... 62

Σχήμα 4.33: Κατανομή τιμών της μεταβλητής duration για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y..... 62

Σχήμα 4.34: Κατανομή τιμών των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής previous..... 63

Σχήμα 4.35: Κατανομή τιμών των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής pdays..... 63

Σχήμα 4.36: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής poutcome..... 64

Σχήμα 4.37: Η σχέση μεταξύ του αριθμού και της διάρκειας των κλήσεων (scatter plot με βάση τις τιμές της y) 65

Σχήμα 4.38: Κατανομή τιμών της μεταβλητής Cons.price.idx για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y..... 65

Σχήμα 4.39: Κατανομή τιμών της μεταβλητής Cons.conf.idx για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y..... 66

Σχήμα 4.40: Κατανομή τιμών της μεταβλητής Emp.var.rate για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y..... 66

Σχήμα 4.41: Κατανομή τιμών της μεταβλητής Euribor3m για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y..... 67

Σχήμα 4.42: Κατανομή τιμών της μεταβλητής Nr.employed για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y..... 67

Σχήμα 4.43: Age Distribution μετά την αφαίρεση των outliers (boxplot) 68

Σχήμα 4.44: Κατανομή τιμών της μεταβλητής age για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y μετά την αφαίρεση των outliers..... 69

Σχήμα 4.45: y Distribution μετά την αφαίρεση των outliers..... 69

Σχήμα 4.46: Πίνακας συνάφειας των μεταβλητών «job» και «education» 71

Σχήμα 4.47: Πίνακας συνάφειας μετά τις αντικαταστάσεις των αγνώστων τιμών της μεταβλητής «education» 72

Σχήμα 4.48: Πίνακας συνάφειας μεταξύ των μεταβλητών «job» και «marital»..... 73

Σχήμα 4.49: Πίνακας συνάφειας μεταξύ των μεταβλητών «housing» και «default» 74

Σχήμα 4.50: Πίνακας συνάφειας μεταξύ των μεταβλητών «housing» και «loan».... 74

Σχήμα 5.1: Το πλήθος των παρατηρήσεων για κάθε τάξη στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου..... 76

Σχήμα 5.2: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης..... 78

Σχήμα 5.3: Precision-Recall καμπύλη για το μοντέλο της Λογ. Παλινδρόμησης..... 78

Σχήμα 5.4: Το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης πριν και μετά τη μέθοδο της τυχαίας υποδειγματοληψίας..... 79

Σχήμα 5.5: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υποδειγματοληψίας..... 80

Σχήμα 5.6: Precision-Recall καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υποδειγματοληψίας ... 81

Σχήμα 5.7: Το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης πριν και μετά τη μέθοδο της τυχαίας υπερδειγματοληψίας..... 81

Σχήμα 5.8: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υπερδειγματοληψίας..... 82

Σχήμα 5.9: Precision-Recall καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υπερδειγματοληψίας.. 83

Σχήμα 5.10: Το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης πριν και μετά τη μέθοδο SMOTE..... 84

Σχήμα 5.11: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της SMOTE..... 85

Σχήμα 5.12: Precision-Recall καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της SMOTE..... 85

Σχήμα 5.13: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε μη ισορροπημένο σύνολο με την εφαρμογή του Weighted αλγορίθμου..... 87

Σχήμα 5.14: Precision-Recall καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε μη ισορροπημένο σύνολο με την εφαρμογή του Weighted αλγορίθμου 87

Σχήμα 5.15: Feature Importance Σκορ του αλγορίθμου Weighted Logistic Regression 92

Κατάλογος Πινάκων

Πίνακας 3.1: Πίνακας Σύγχυσης..... 16

Πίνακας 4.1: Οι μεταβλητές του συνόλου δεδομένων..... 31

Πίνακας 4.2: Περιγραφικά στοιχεία για τη μεταβλητή «Age» 33

Πίνακας 4.3: Πίνακας συχνοτήτων της μεταβλητής «job» 35

Πίνακας 4.4: Πίνακας συχνοτήτων της μεταβλητής «marital» 37

Πίνακας 4.5: Πίνακας συχνοτήτων της μεταβλητής «education» 37

Πίνακας 4.6: Περιγραφικά στοιχεία για τη μεταβλητή «campaign» 40

Πίνακας 4.7: Πίνακας συχνοτήτων της μεταβλητής «pdays» 41

Πίνακας 4.8: Πίνακας συχνοτήτων της μεταβλητής «previous» 43

Πίνακας 4.9: Πίνακας συχνοτήτων της μεταβλητής «poutcome» 44

Πίνακας 4.10: Πίνακας συνάφειας ανάμεσα στις μεταβλητές «previous» και «poutcome» 45

Πίνακας 4.11: Πίνακας συνάφειας ανάμεσα στις μεταβλητές «pdays» και «poutcome» 46

Πίνακας 4.12: Περιγραφικά στοιχεία για τη μεταβλητή «duration» 48

Πίνακας 4.13: Περιγραφικά στοιχεία για τη μεταβλητή «consumer price index».... 49

Πίνακας 4.14: Περιγραφικά στοιχεία για τη μεταβλητή «consumer confidence index» 50

Πίνακας 4.15: Περιγραφικά στοιχεία για τη μεταβλητή «employment variation rate» 51

Πίνακας 4.16: Περιγραφικά στοιχεία για τη μεταβλητή «euribor3m» 52

Πίνακας 4.17: Περιγραφικά στοιχεία για τη μεταβλητή «Nr.employed» 53

Πίνακας 4.18: Πίνακας ποσοστών θετικής και αρνητικής ανταπόκρισης για τις τιμές της μεταβλητής «month» 59

Πίνακας 4.19: Πίνακας ποσοστών θετικής και αρνητικής ανταπόκρισης για τις 6 πρώτες τιμές της μεταβλητής «campaign» 61

Πίνακας 4.20: Πίνακας ποσοστών θετικής και αρνητικής ανταπόκρισης για τις τιμές της μεταβλητής «poutcome» 64

Πίνακας 4.21: Πίνακας πλήθους «unknown values» πριν την απομάκρυνση των outliers και μετά..... 70

Πίνακας 4.22: Πίνακας πλήθους «unknown values» μετά την εφαρμογή του cross-tabulation και επικρατούσα τιμή για κάθε μεταβλητή..... 73

Πίνακας 5.1: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης.... 78

Πίνακας 5.2: Μετρικές αξιολόγησης για το μοντέλο της Λογ. Παλινδρόμησης..... 79

Πίνακας 5.3: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υποδειγματοληψίας..... 80

Πίνακας 5.4: Μετρικές αξιολόγησης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υποδειγματοληψίας.... 81

Πίνακας 5.5: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υπερδειγματοληψίας..... 83

Πίνακας 5.6: Μετρικές αξιολόγησης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υπερδειγματοληψίας.. 83

Πίνακας 5.7: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της μεθόδου SMOTE..... 85

Πίνακας 5.8: Μετρικές αξιολόγησης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της μεθόδου SMOTE 85

Πίνακας 5.9: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης σε μη ισορροπημένο σύνολο με την εφαρμογή του Weighted αλγορίθμου..... 87

Πίνακας 5.10: Μετρικές αξιολόγησης για το μοντέλο της Λογιστικής Παλινδρόμησης σε μη ισορροπημένο σύνολο με την εφαρμογή του Weighted αλγορίθμου 87

Πίνακας 5.11: Σύγκριση των στοιχείων TP,FN,FP και TN για όλα τα μοντέλα της Λογιστικής Παλινδρόμησης..... 88

Πίνακας 5.12: Σύγκριση των μετρικών αξιολόγησης για όλα τα μοντέλα της Λογιστικής Παλινδρόμησης..... 89

Πίνακας 5.13: Σύγκριση των στοιχείων TP,FN,FP και TN για όλα τα μοντέλα SVM... 89

Πίνακας 5.14: Σύγκριση των μετρικών αξιολόγησης για όλα τα μοντέλα SVM..... 90

Πίνακας 5.15: Σκορ σημαντικότητας των μεταβλητών με βάση το βέλτιστο μοντέλο (Weighted Logistic Regression) 91

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της Εργασίας

Η ανάπτυξη της τεχνολογίας, των μεθόδων αλλά και των υπολογιστικών δυνατοτήτων που λαμβάνουν χώρα από τις αρχές του 21ου αιώνα, παρέχουν τη δυνατότητα βαθύτερης παρατήρησης και πρόβλεψης των κινήσεων των τραπεζών χρησιμοποιώντας πολύπλοκα μαθηματικά μοντέλα ικανά να δεχτούν μεγάλο αριθμό παραμέτρων. Η χρήση της Τεχνητής Νοημοσύνης επιτρέπει ένα νέο τρόπο μελέτης και πρόβλεψης των εξελίξεων του τραπεζικού τομέα. Επιλέξαμε η μελέτη να γίνει στον τομέα της τραπεζικής, αλλά σαφώς η ίδια μεθοδολογία θα μπορούσε να εφαρμοσθεί και σε πολλούς άλλους τομείς όπου υπάρχει έντονα η ανάγκη, όπως στον ιατρικό τομέα και στον χρηματοοικονομικό.

Η αυξημένη εφαρμογή καμπανιών μάρκετινγκ τα τελευταία χρόνια παράγει δεδομένα τα οποία χρειάζεται να αναλυθούν για να υποστηρίξουν τη διαδικασία της λήψης των αποφάσεων. Η δυσκολία ανάλυσης αυτών των πολύπλοκων δεδομένων συνέβαλλε στην ανάπτυξη των τεχνικών Επιχειρηματικής Ευφυΐας, η οποία στοχεύει στην εξαγωγή χρήσιμης γνώσης για την υποστήριξη της λήψης αποφάσεων.

Αναλογιζόμενοι την οικονομική κρίση στην Ευρώπη και στην Ελλάδα τα τελευταία χρόνια, μία συνέπειά της για τις τράπεζες ήταν οι περιορισμοί πιστώσεων και ο πλήρης έλεγχος διακίνησης κεφαλαίων, τα γνωστά σε όλους capital controls με σκοπό να αποτρέψουν την πλήρη κατάρρευση του τραπεζικού συστήματος. Εκείνη η εποχή έγινε αρωγός για την εφαρμογή στον τραπεζικό τομέα νέων συστηματικών διαδικασιών για την δημιουργία τραπεζικών προϊόντων και την ανάπτυξη καμπανιών που θα αποφέρουν περισσότερα έσοδα στις τράπεζες. Για το λόγο αυτό, θεωρήσαμε αναγκαίο να εξετάσουμε τον τραπεζικό τομέα γενικότερα, δίνοντας έμφαση στην αναγκαιότητα πρόβλεψης τρόπων προσέγγισης νέων ή δυνητικών πελατών που θέλουν να δημιουργήσουν νέες προθεσμιακές καταθέσεις σε μία μετά των capital controls εποχή.

Το ερώτημα που τίθεται επομένως είναι: Μπορούμε να εξάγουμε γνώση από τα διαθέσιμα δεδομένα παλαιότερων καμπανιών για να βελτιώσουμε την αποτελεσματικότητα των μελλοντικών καμπανιών μάρκετινγκ;

Στην παρούσα εργασία κάνουμε μία προσπάθεια να απαντήσουμε στο ερώτημα αυτό, χρησιμοποιώντας δεδομένα από μία άμεση καμπάνια μάρκετινγκ που έγινε από μία Πορτογαλική τράπεζα τη χρονική περίοδο Μάιος 2008 έως Νοέμβριος 2010. Προσπαθούμε να κατανοήσουμε καλύτερα τα χαρακτηριστικά των παλιών πελατών καθώς και ποιες είναι οι μεταβλητές που επηρεάζουν περισσότερο την απόφασή τους.

Η εργασία παρέχει το απαραίτητο θεωρητικό υπόβαθρο και αναλύει κάποιους από τους δημοφιλέστερους αλγορίθμους μηχανικής μάθησης που χρησιμοποιούνται σε προβλήματα ταξινόμησης. Η διερεύνηση του προβλήματος εστιάζει σε δύο βασικούς αλγορίθμους, τη Λογιστική Παλινδρόμηση (Logistic Regression - LR) και τις Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM) και προσπαθεί να προβλέψει την απάντηση των ατόμων που προσεγγίζονται από μία τράπεζα κατά τη διάρκεια μίας καμπάνιας μάρκετινγκ ως προς το αν θα πραγματοποιήσουν προθεσμιακή κατάθεση ή όχι.

Ακόμη, καθώς το σύνολο δεδομένων είναι μη ισορροπημένο παρουσιάζονται διαφορετικές προσεγγίσεις για την προεπεξεργασία των δεδομένων. Η πρώτη προσέγγιση αφορά τη χρησιμοποίηση αλγορίθμων στη συμβατική τους μορφή, η δεύτερη την εφαρμογή αλγορίθμων μετά την εξισορρόπηση του συνόλου δεδομένων και η τρίτη αφορά τροποποιήσεις των συμβατικών αλγορίθμων οι οποίοι ακολουθούν την cost-sensitive δομή.

1.2 Βιβλιογραφική επισκόπιση

Στο άρθρο 'Enhancing Bank Direct Marketing through Data Mining' οι Sérgio Moro, Raul M. S. Laureano και Paulo Cortez (2012) χρησιμοποίησαν επίσης την μεθοδολογία CRISP-DM που χρησιμοποιήθηκε και στην παρούσα εργασία αλλά ως αλγορίθμους εφάρμοσαν τους Naïve Bayes, Decision Trees και Support Vector Machines και τους

σύγκριναν υπολογίζοντας τις μετρικές AUC (Area Under the ROC Curve) και ALIFT (Area Under the Lift Curve) από όπου προέκυψε ο βέλτιστος αλγόριθμος να είναι ο Support Vector Machine.

Στο άρθρο 'Predicting Bank Marketing Campaign Success using Machine Learning' οι ερευνητές Chaitra Hegde, Aakash Kaku και Neelang Parghi (2014) επικεντρώθηκαν στο μεγαλύτερο σύνολο δεδομένων των 41.118 παρατηρήσεων που συγκεντρώθηκε από την Πορτογαλική τράπεζα κατά τη διάρκεια της καμπάνιας μάρκετινγκ και είναι επίσης διαθέσιμο στο UCI Machine Learning Repository. Κατά τη διάρκεια αυτής της μελέτης εφαρμόστηκαν οι αλγόριθμοι Decision Trees, Random Forests, Gradient Boosted Trees και AdaBoost. Προέκυψε ως βέλτιστος αλγόριθμος ταξινόμησης ο AdaBoost που έδωσε AUC-score=0,8036.

Οι Md. Rashid Farooqi και Naiyar Iqbal (2019) στο άρθρο τους 'Performance Evaluation for Competency of Bank Telemarketing Prediction using DataMining Techniques' χρησιμοποίησαν τους αλγορίθμους Decision tree, Sequential Minimal Optimization, Artificial Neural Network, Naive Bayes και k-Nearest Neighbour για την δυαδική ταξινόμηση και κατέληξαν βάσει των μετρικών Accuracy και ROC Curve ότι ο βέλτιστος αλγόριθμος είναι τα Δέντρα Αποφάσεων.

Στην βιβλιογραφία που βασίστηκε η συγκεκριμένη μελέτη γίνεται πληθώρα εφαρμογών διαφόρων αλγορίθμων Επιβλεπόμενης Μηχανικής Μάθησης και αξιολογούνται τα αποτελέσματα των αλγορίθμων με βάση μετρικές όπως Accuracy, Precision, Recall, f1-score κα. Η μεθοδολογία που χρησιμοποιείται πιο συχνά είναι η CRISP-DM, η οποία περιγράφεται και στην πορεία.

1.3 Σχεδιασμός της μελέτης

Προκειμένου να απαντηθούν τα ερευνητικά ερωτήματα και να εξαχθούν αξιόπιστα συμπεράσματα θα πρέπει να επιλεγεί το κατάλληλο είδος ερευνητικής διαδικασίας και να σχεδιαστεί η μελέτη με σαφή και συγκεκριμένα βήματα. Για τις ανάγκες της παρούσας εργασίας κρίθηκε ως πιο κατάλληλη η ερευνητική μεθοδολογία CRISP-DM (CRoss-Industry Standard Process for Data Mining) η οποία εφαρμόζεται ευρέως στην Εξόρυξη Δεδομένων. Η μέθοδος αυτή παρέχει μια διαδικασία έξι σταδίων:

1. κατανόηση των επιχειρήσεων

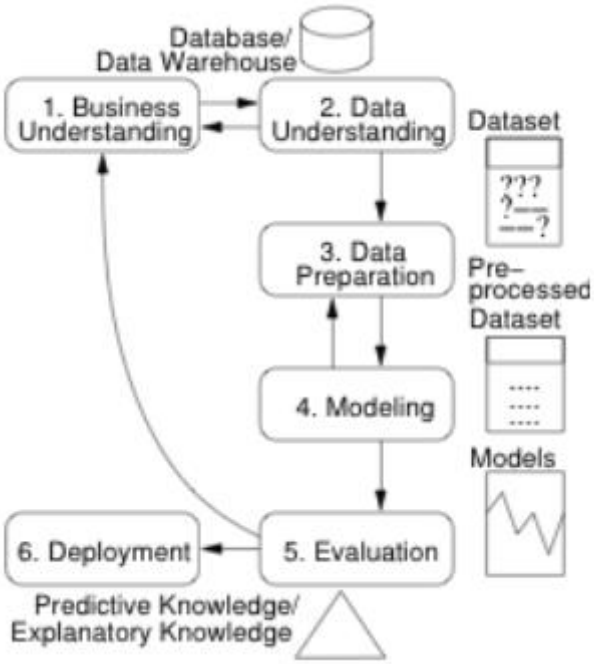
2. κατανόηση δεδομένων

3. προετοιμασία δεδομένων

4. μοντελοποίηση

5. αξιολόγηση

6. ανάπτυξη



Σχήμα 1.1: Η διαδικασία CRISP-DM

Το κεφάλαιο 4 πραγματεύεται τα στάδια 1,2 και 3 ενώ το κεφάλαιο 5 τα υπόλοιπα 3 στάδια.

1.4 Εργαλεία

Η επεξεργασία των δεδομένων έγινε στο περιβάλλον Spyder και υλοποιήθηκε σε γλώσσα προγραμματισμού Python. Η Python είναι μια γλώσσα προγραμματισμού γενικής χρήσης, πάρα πολύ υψηλού επιπέδου, απλή και εύκολη στην εκμάθηση,

ισχυρή, δυναμική, αποδοτική, παραγωγική και επεκτάσιμη. Είναι κατάλληλη και για αρχάριους και για έμπειρους προγραμματιστές.

Στην παρούσα διπλωματική χρησιμοποιήθηκαν οι βιβλιοθήκες:

• pandas: για την ανάλυση των δεδομένων

• imblearn: για τη διαχείριση της μη ισορροπίας του σετ δεδομένων

• matplotlib: για την οπτικοποίηση των δεδομένων και το σχεδιασμό διαγραμμάτων

• seaborn: για την οπτικοποίηση των δεδομένων και το σχεδιασμό διαγραμμάτων

•scikit learn: για την εφαρμογή αλγορίθμων Μηχανικής Μάθησης

1.5 Δομή Εργασίας

Η εργασία οργανώνεται σε 6 κεφάλαια:

Στο κεφάλαιο 1 παρουσιάζεται το αντικείμενο της εργασίας, γίνεται μία βιβλιογραφική ανασκόπιση του θέματος και περιγράφεται το θεωρητικό υπόβαθρο της μεθοδολογίας με βάση την οποία θα υλοποιήσουμε την περίπτωση μελέτης. Ακόμη, αναφέρονται τα εργαλεία που θα χρησιμοποιήσουμε καθώς και η διάρθρωση των κεφαλαίων της.

Στο κεφάλαιο 2 θα εξηγήσουμε περιληπτικά την έννοια της Μηχανικής Μάθησης και τη σημασία της. Θα γίνει αναφορά στις μεθόδους Μηχανικής Μάθησης και στους πιο δημοφιλείς αλγορίθμους που χρησιμοποιούνται στο μεγαλύτερο μέρος των εφαρμογών της.

Στο κεφάλαιο 3 αναλύουμε σημαντικές τεχνικές και μεθόδους, οι οποίες πρόκειται να χρησιμοποιηθούν στη συνέχεια της εργασίας. Συγκεκριμένα, περιγράφεται η έννοια του μη ισορροπημένου συνόλου δεδομένων, οι εφαρμογές τους σε επιχειρηματικά προβλήματα, οι μέθοδοι διαχείρισης των συνόλων αυτών και οι βασικές τεχνικές εξισορροπησής τους. Ακόμη, δίνονται οι βασικότερες μετρικές απόδοσης για τη διαδικασία αξιολόγησης ενός μοντέλου Μηχανικής Μάθησης.

Στο κεφάλαιο 4 περιγράφεται η μεθοδολογική προσέγγιση της εργασίας, δηλαδή ποια δεδομένα χρησιμοποιήθηκαν, τα περιγραφικά στατιστικά του δείγματος, η μεθοδολογία προ-επεξεργασίας των δεδομένων καθώς και γίνεται μία σύγκριση των 2 κατηγοριών του δείγματος σύμφωνα με την εξαρτημένη μεταβλητή που χρησιμοποιείται στην ταξινόμηση που ακολουθεί στο επόμενο κεφάλαιο.

Στο κεφάλαιο 5 θα γίνει προσπάθεια ταξινόμησης των δεδομένων του δείγματος μέσω διαφόρων αλγορίθμων Μηχανικής Μάθησης. Για κάθε μοντέλο που εκπαιδεύεται, δίνονται και σχολιάζονται τα αντίστοιχα διαγράμματα και οι μετρικές αξιολόγησης του, ώστε να καταλήξουμε σε γενικά συμπεράσματα. Η μελέτη επικεντρώνεται στη σύγκριση μεταξύ 2 αλγορίθμων ταξινόμησης ενώ ταυτόχρονα μελετάται η δομή του συνόλου δεδομένων όταν αυτό είναι ή δεν είναι ισορροπημένο.

Στο κεφάλαιο 6 καταγράφουμε τα γενικά συμπεράσματα από τη μελέτη που υλοποιήσαμε και παρουσιάζουμε μελλοντικές επεκτάσεις της.

Κεφάλαιο 2

Μηχανική Μάθηση

2.1 Μέθοδοι Μηχανικής Μάθησης: Ο όρος Μηχανική Μάθηση αναφέρεται στην ανάπτυξη και τη χρήση υπολογιστικών συστημάτων ικανών να "μαθαίνουν" και να προσαρμόζονται χωρίς την ύπαρξη προκαθορισμένων εντολών. Στην ενότητα αυτή θα αναλύσουμε τις υποκατηγορίες της Μηχανικής Μάθησης, οι οποίες είναι:

- **Επιβλεπόμενη Μάθηση (Supervised Learning):** Σε κάθε πρόβλημα επιβλεπόμενης μάθησης διακρίνουμε δύο είδη μεταβλητών: τις ανεξάρτητες ή ελεγχόμενες ή επεξηγηματικές ή μεταβλητές εισόδου (independent, predictor, input, explanatory variables) και τις εξαρτημένες ή απόκρισης ή μεταβλητές εξόδου (dependent, response variables, target variable). Η μέθοδος αυτή χρησιμοποιεί labeled datasets δηλαδή σύνολα δεδομένων που για κάθε παρατήρηση είναι γνωστή η ταμπέλα του ('label') δηλαδή η τιμή της εξαρτημένης μεταβλητής. Χρησιμοποιείται για την εκπαίδευση των μοντέλων με σκοπό να προβλέψει τις τιμές της εξαρτημένης μεταβλητής (μεταβλητή-στόχος) βάσει των τιμών των ανεξάρτητων μεταβλητών. Οι τιμές της εξαρτημένης μεταβλητής είναι γνωστές για το σύνολο δεδομένων, και η διαδικασία στοχεύει στην εκπαίδευση του αλγορίθμου στις τιμές αυτές και στη γενίκευσή του σε σύνολα δεδομένων με άγνωστη και ζητούμενη την τιμή της εξαρτημένης μεταβλητής. Συνήθως χρησιμοποιείται για:
    - **Παλινδρόμηση (Regression):** Πρόκειται για την τεχνική που εξετάζει τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με σκοπό την πρόβλεψη των τιμών της μιας, μέσω των τιμών της άλλης (ή των άλλων). Σε πειραματικές έρευνες, ανεξάρτητη μεταβλητή X είναι εκείνη την οποία μπορούμε να ελέγξουμε, δηλαδή, να καθορίσουμε τις τιμές της ενώ εξαρτημένη μεταβλητή Y είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις ανεξάρτητες μεταβλητές. Για παράδειγμα στις εταιρίες η παλινδρόμηση χρησιμοποιείται για να προβλέψει το ύψος της διαφημιστικής

καμπάνιας των προϊόντων ανάλογα με τη ζήτησή τους και τις πωλήσεις τους.

- **Ταξινόμηση (Classification):** Το αντικείμενο της ταξινόμησης είναι ο διαχωρισμός των δεδομένων σε κλάσεις ή υποκατηγορίες όσον αφορά μία μεταβλητή-στόχο. Ο κύριος στόχος της είναι να αναγνωριστεί η κλάση στην οποία ανήκει μία νέα παρατήρηση βάσει των χαρακτηριστικών της που αποτελούν τις ανεξάρτητες μεταβλητές του προβλήματος. Ο καθορισμός της κλάσης στην οποία ανήκει μία νέα παρατήρηση γίνεται είτε απευθείας με την ανάθεση της τιμής της κλάσης είτε μέσω υπολογισμού της πιθανότητας του αν θα ανήκει στη συγκεκριμένη κατηγορία. Οι τεχνικές της ταξινόμησης εφαρμόζονται είτε σε προβλήματα πολλών κλάσεων (Multi-Class Classification) είτε πιο συχνά σε προβλήματα που οι πιθανές κλάσεις είναι δύο με την ταξινόμηση να ονομάζεται δυαδική (Binary Classification). Για παράδειγμα, στον τραπεζικό τομέα χρησιμοποιείται για να προσδιοριστεί εάν ένας πελάτης που ζητά δάνειο θα είναι σε θέση να το αποπληρώσει χρησιμοποιώντας ως κύριο παράγοντα το οικονομικό ιστορικό του πελάτη.

Ουσιαστικά, η κυριότερη διαφορά της παλινδρόμησης από την ταξινόμηση είναι ότι η μεταβλητή στόχος (target variable) είναι συνεχής στην περίπτωση της παλινδρόμησης ενώ είναι διακριτή στην περίπτωση της ταξινόμισης.

- **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning):** Χρησιμοποιεί unlabeled datasets δηλαδή δεν υπάρχει εξαρτημένη μεταβλητή (μεταβλητή-στόχος) και ο αλγόριθμος κατασκευάζει ένα μοντέλο για κάποιο σύνολο εισόδων χωρίς να γνωρίζει τις επιθυμητές εξόδους. Ο σκοπός του μοντέλου είναι η εύρεση μοτίβων στα δεδομένα και χρησιμοποιείται με τους ακόλουθους τρόπους:

  - **Ομαδοποίηση (Clustering):** Είναι η διαδικασία εύρεσης ομοιοτήτων μέσα σε ένα σύνολο δεδομένων προκειμένου να διαχωριστεί σε ομάδες ώστε τα σημεία που ανήκουν στην ίδια ομάδα να μοιάζουν όσο το δυνατόν περισσότερο μεταξύ τους, ενώ τα σημεία που ανήκουν σε διαφορετικές ομάδες να διαφέρουν όσο περισσότερο γίνεται. Όπως

αναφέρθηκε και προηγουμένως, ο αλγόριθμος δεν γνωρίζει τις επιθυμητές εξόδους δηλαδή στην συγκεκριμένη περίπτωση τις ομάδες στις οποίες χωρίζονται τα δεδομένα αλλά προσπαθεί να ομαδοποιήσει τα δεδομένα βάσει ομοιοτήτων και διαφορών (πχ Customer Segmentation).

- **Ανάλυση συσχέτισης (Association Analysis):** Είναι η διαδικασία εύρεσης γεγονότων που σχετίζονται και συμβαίνουν μαζί. Χρησιμοποιείται συχνά από τις εταιρίες που τις ενδιαφέρει η ανάλυση καλαθιού (market basket analysis). Οι εταιρίες δηλαδή αναλύουν τις αγοραστικές συνήθειες πελατών βρίσκοντας συσχετίσεις μεταξύ διαφορετικών προϊόντων τα οποία οι πελάτες τοποθετούν στα καλάθια αγορών τους. Η ανακάλυψη αυτών των συσχετίσεων μπορεί να βοηθήσει τους εμπόρους να αναπτύξουν κατάλληλες στρατηγικές μάρκετινγκ για να αυξήσουν τις πωλήσεις τους.
- **Συστήματα Συστάσεων (Recommendation Systems):** Είναι η διαδικασία κατά την οποία προτείνονται προϊόντα ή υπηρεσίες στους πελάτες, τα οποία βασίζονται στην ομοιότητά τους με άλλους πελάτες.
- **Εντοπισμό ανωμαλιών (Anomaly detection):** Είναι η διαδικασία ανίχνευσης των παρατηρήσεων, γεγονότων ή προϊόντων τα οποία διαφέρουν σημαντικά από το αναμενόμενο ή την πλειοψηφία των δεδομένων. Στη βιβλιογραφία συναντάται συχνά και ως Outlier Detection.

- **Ημι-επιβλεπόμενη Μάθηση (Semi-supervised Learning):** σε αυτήν την περίπτωση έχουμε labeled αλλά και unlabeled δεδομένα, επιτρέποντας έτσι να γίνει η εκπαίδευση σε πιο ρεαλιστικές συνθήκες, όπου για κάποιες από τις παρατηρήσεις οι επιθυμητές έξοδοι είναι γνωστές και για κάποιες άλλες παρατηρήσεις οι επιθυμητές έξοδοι είναι άγνωστες. Πολύ συχνά ερχόμαστε αντιμέτωποι με το πρόβλημα της ελλιπούς πληροφορίας στα δεδομένα μας και με τη χρήση αυτής της μεθόδου αντιμετωπίζεται το πρόβλημα και βελτιστοποιείται η διαδικασία μάθησης.

- **Ενισχυτική Μάθηση (Reinforcement Learning):** πρόκειται για μια προσέγγιση που βασίζεται στη μάθηση βέλτιστης συμπεριφοράς ενός πράκτορα μέσω της αλληλεπίδρασής του με το περιβάλλον. Αποτελεί ίσως τη μορφή μάθησης που χρησιμοποιείται περισσότερο και στον πραγματικό κόσμο, αφού αφορά στη μάθηση μέσω trial and error, δηλαδή τον τρόπο με τον οποίο μαθαίνουν συνήθως τα έμβια όντα. Ουσιαστικά υπάρχει ποινή όταν το μηχάνημα κάνει κάτι λάθος και επιβράβευση όταν το κάνει σωστά.

2.2 Αλγόριθμοι Μηχανικής Μάθησης

΄Οπως προαναφέρθηκε, σε μεθόδους μηχανικής μάθησης με επίβλεψη, ο αλγόριθμος χρησιμοποιεί δεδομένα εισόδου για τα οποία είναι γνωστή η επιθυμητή έξοδος, εκπαιδεύεται με βάση αυτά και τελικά μπορεί να προβλέπει και να ταξινομεί τα νέα δεδομένα εισόδου, στις γνωστές επιθυμητές εξόδους.

Παραδείγματα αλγορίθμων Μηχανικής Μάθησης αυτής της κατηγορίας είναι η Γραμμική Παλινδρόμηση (Linear Regression), η Λογιστική Παλινδρόμηση (Logistic Regression), οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), τα Δέντρα Απόφασης (Decision Trees) και τα Νευρωνικά Δίκτυα (Neural Networks). Οι αλγόριθμοι Επιβλεπόμενης Μάθησης χρησιμοποιούνται τόσο σε προβλήματα Ταξινόμησης όσο και σε προβλήματα Παλινδρόμησης. Η ενότητα αυτή επικεντρώνεται στους 2 αλγορίθμους επιβλεπόμενης μάθησης που θα χρησιμοποιηθούν στην εργασία για ταξινόμηση.

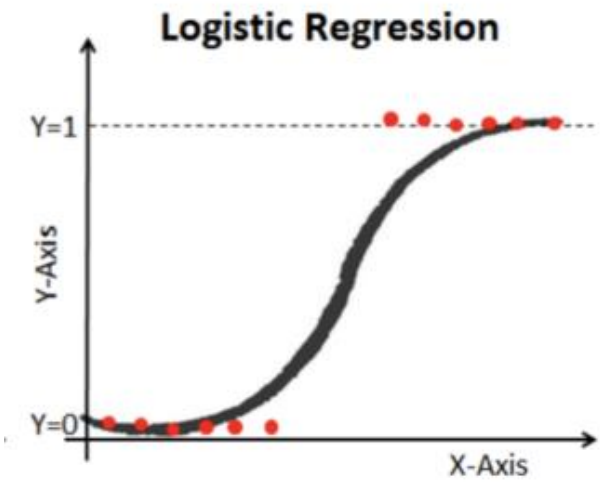
Λογιστική Παλινδρόμηση:

Για ένα γνωστό δείγμα εισόδου x, μπορεί να υπολογιστεί η πιθανότητα το δείγμα αυτό να ανήκει σε μία από τις δύο κατηγορίες της εξαρτημένης μεταβλητής. Η πιθανότητα αυτή κυμαίνεται μεταξύ 0 και 1 και επιπλέον υπάρχει μια τιμή κατωφλίου που αναλόγως με αυτή, η εκτιμώμενη πιθανότητα και επομένως και τα δείγματα ταξινομούνται στην ανάλογη τάξη.

Η συνάρτησή της δίνεται από τον τύπο: $f(x) = \frac{e^x}{1+e^x}$

Όπου x είναι η μεταβλητή εισόδου και f(x) η πιθανότητα να ανήκει στην μία κλάση.

Ο αλγόριθμος της Λογιστικής Παλινδρόμησης βασίζεται πάνω στο μοντέλο Γραμμικής Παλινδρόμησης, αλλά προσαρμόζει τα δεδομένα εκπαίδευσης πάνω στην σιγμοειδή συνάρτηση (sigmoid function) όπως φαίνεται και στο παρακάτω σχήμα.



Σχήμα 2.1: Γραφική παράσταση Λογιστικής Παλινδρόμησης

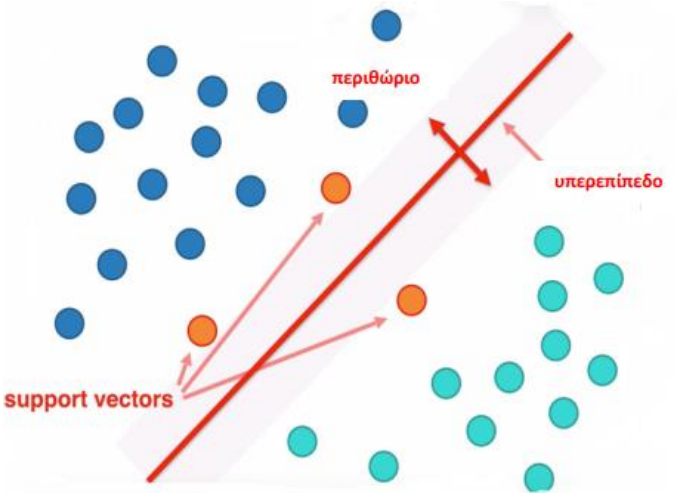
Η σημαντικότερη διαφορά μεταξύ της Λογιστικής και της Γραμμικής Παλινδρόμησης οφείλεται στη διαφορετική μορφή της εξαρτημένης μεταβλητής, η οποία στην πρώτη περίπτωση μπορεί να είναι κατηγορική ή αριθμητική, ενώ στη δεύτερη είναι αποκλειστικά ποσοτική.

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines-SVM)

Ο ταξινομητής SVM προσπαθεί να βρει ένα γραμμικό ή μη γραμμικό υπερεπίπεδο απόφασης το οποίο να διαχωρίζει το σύνολο εκπαίδευσης με τέτοιο τρόπο ώστε τα σημεία που ανήκουν στην ίδια τάξη να βρίσκονται στη ίδια πλευρά του υπερεπιπέδου. Για να το πετύχει αυτό προβάλλει όλα τα διαθέσιμα σημεία από το σύνολο εκπαίδευσης σε αυτόν τον χώρο, που μπορεί να είναι δύο ή περισσοτέρων διαστάσεων, και στη συνέχεια προσπαθεί να βρει το βέλτιστο υπερεπίπεδο, δηλαδή αυτό που διαχωρίζει βέλτιστα τα σημεία με βάση την τάξη τους. Μεταξύ όλων των πιθανών υπερεπιπέδων αναζητά εκείνο για το οποίο η απόσταση από το κοντινότερο

σημείο είναι μέγιστη, δηλαδή αναζητά το υπερεπίπεδο που μεγιστοποιεί το περιθώριο (maximal margin hyperplane).

Στην περίπτωση που έχει γίνει προβολή των σημείων στο διδιάστατο χώρο, το υπερεπίπεδο είναι ευθεία, η οποία μεγιστοποιεί το περιθώριο και βρίσκεται όσο πιο μακρυά γίνεται από τα διανύσματα υποστήριξης.



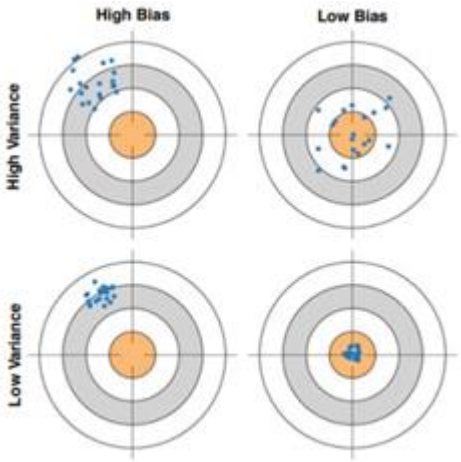
Σχήμα 2.2: Γραφική αναπαράσταση του αλγορίθμου Support Vector Machines

Μεροληψία και Διακύμανση

Με τη χρήση των αλγορίθμων Μηχανικής Μάθησης με Επίβλεψη γίνεται προσπάθεια να εκτιμηθούν οι έξοδοι των δεδομένων εκπαίδευσης. Στην περίπτωση που οι εκτιμήσεις των εξόδων διαφέρουν σε μεγάλο βαθμό από τις πραγματικές τιμές των labels, το μοντέλο χαρακτηρίζεται από υψηλή μεροληψία ενώ στην περίπτωση που οι εκτιμήσεις των εξόδων διαφέρουν σε μεγάλο βαθμό από τη μέση τιμή των εκτιμήσεων τότε το μοντέλο χαρακτηρίζεται από υψηλή διακύμανση. Πιο συγκεκριμένα, η μεροληψία ισοδυναμεί με την αναμενόμενη απόκλιση από την πραγματική τιμή της συνάρτησης ή της παραμέτρου που επιθυμούμε να εκτιμήσουμε ενώ η διακύμανση από την άλλη πλευρά, παρέχει ένα μέτρο της απόκλισης από τη μέση τιμή του εκτιμητή που είναι πιθανό να προκληθεί από την εκάστοτε δειγματοληψία των δεδομένων.

Το επόμενο σχήμα διευκρινίζει τις έννοιες της διακύμανσης και της μεροληψίας. Το κέντρο του στόχου αναπαριστά την τέλεια πρόβλεψη των πραγματικών τιμών των εξόδων ενώ τα σημεία μακρυά του κέντρου αναπαριστούν προβλέψεις με σφάλματα.

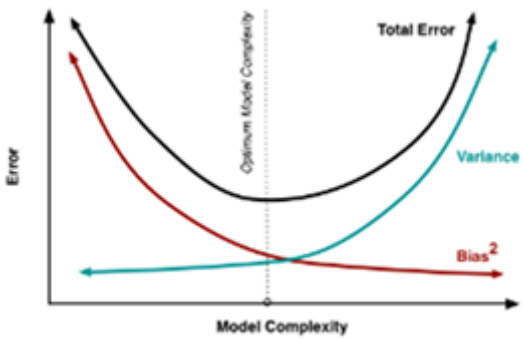
Για παράδειγμα, στο πάνω αριστερά σχήμα, οι προβλέψεις των τιμών των εξόδων απέχουν πολύ από τις πραγματικές τιμές στο κέντρο άρα υπάρχει μεγάλη μεροληψία. Επίσης, τα σημεία-προβλέψεις είναι διασκορπισμένα και απέχουν αρκετά από τη μέση τιμή των προβλέψεων οπότε υπάρχει μεγάλη διακύμανση.



Σχήμα 2.3: Οπτική απεικόνιση της μεροληψίας και της διακύμανσης

Στην κάτω δεξιά περίπτωση αναπαρίσταται η επίτευξη του στόχου των αλγορίθμων, δηλαδή επιτυγχάνεται ακριβής πρόβλεψη των πραγματικών τιμών με μεγάλη αξιοπιστία καθώς όλες οι προβλέψεις του αλγορίθμου είναι και κοντά στις πραγματικές τιμές αλλά και απέχουν ελάχιστα μεταξύ τους (μικρή μεροληψία και μικρή διακύμανση).

Στη στατιστική και τη μηχανική μάθηση, το δίλλημα της μεροληψίας-διακύμανσης είναι η σύγκρουση στην προσπάθεια ταυτόχρονης ελαχιστοποίησης αυτών των δύο αντικρουόμενων πηγών σφάλματος. Η μεροληψία και η διακύμανση μετρούν δύο διαφορετικές πηγές σφάλματος σε έναν εκτιμητή και είναι σημαντικό να γίνει ένα trade-off μεταξύ αυτών (bias-variance trade-off), καθώς η σχέση μεταξύ της μεροληψίας και της διακύμανσης είναι άρρηκτα συνδεδεμένη με τις έννοιες της υποπροσαρμογής και της υπερπροσαρμογής.



Σχήμα 2.4: Bias-Variance trade-off

Υπερπροσαρμογή (Overfitting)

Ένα μοντέλο που χαρακτηρίζεται από μεγάλη διακύμανση στην απόδοση από dataset σε dataset και χαμηλή μεροληψία συχνά υπερεκπαιδεύεται. Αυτό σημαίνει ότι τα δεδομένα εκπαίδευονται μόνο για το συγκεκριμένο σύνολο εκπαίδευσης, απομνημονεύουν τα μοτίβα που υπάρχουν σε αυτό και αδυνατούν να γενικευτούν για unseen datasets. Αυτό το πρόβλημα συναντάται όταν το μοντέλο είναι πάρα πολύ πολύπλοκο για τα δεδομένα ή όταν υπάρχουν πολλές μεταβλητές και λίγα δεδομένα.

Υποπροσαρμογή (Underfitting)

Από την άλλη ένα μοντέλο που χαρακτηρίζεται από μεγάλη μεροληψία και χαμηλή διακύμανση σημαίνει ότι ο αλγόριθμος δεν είναι ικανός να μοντελοποιήσει τα δεδομένα εκπαίδευσης, αλλά ούτε και να γενικεύσει σε νέα δεδομένα. Αυτό συμβαίνει στην περίπτωση που ο αλγόριθμος είναι ιδιαίτερα απλός ή όταν οι μεταβλητές εισόδου είναι λίγες ή ελάχιστα επιδρούν στην εκτίμηση της εξόδου.

Κεφάλαιο 3

Μη ισορροπημένα σύνολα δεδομένων

3.1 Θεωρητικό υπόβαθρο

Σε πολλά προβλήματα δυαδικής ταξινόμησης παρατηρείται ότι τα περισσότερα δεδομένα ανήκουν σε μία συγκεκριμένη κλάση (η οποία ονομάζεται πλειοψηφική) ενώ λιγότερα ανήκουν στην δεύτερη κλάση (μειοψηφική). Σε αυτές τις περιπτώσεις το πρόβλημα λέγεται μη ισορροπημένο (imbalanced). Στην προσπάθεια να εφαρμοστεί κάποιο μοντέλο ταξινόμησης και να αξιολογηθεί με διάφορες μετρικές για την καταλληλότητά του, παρατηρείται ότι δίνει φτωχά αποτελέσματα για τη μειοψηφική κλάση. Αυτό συμβαίνει επειδή τα κλασικά μοντέλα αποκτούν λίγη πληροφορία για την μειοψηφική κλάση λόγω του μεγέθους της, με συνέπεια να είναι δύσκολο να ανιχνεύσουν δομές και να δώσουν ουσιώδεις προβλέψεις γι αυτήν. Αντίθετα, δίνουν καλά αποτελέσματα για την πλειοψηφική κλάση καθώς το μοντέλο μεροληπτεί προς την κλάση αυτή. Βέβαια σε πολλά προβλήματα όπως και στο συγκεκριμένο, ο πρωταρχικός στόχος είναι η πρόβλεψη για την μειοψηφική κλάση άρα υπάρχει μεγάλο κόστος από τη λανθασμένη ταξινόμηση (misclassification cost).

Ένα εύρος αλγορίθμων ταξινόμησης, όπως η Λογιστική Παλινδρόμιση, τα Δέντρα Αποφάσεων, τα Νευρωνικά Δίκτυα, οι Μηχανές Διανυσματικής Υποστήριξης έχουν αναπτυχθεί και εφαρμοστεί σε πολλά προβλήματα διαφόρων πεδίων. Οι παραπάνω συμβατικοί αλγόριθμοι υποθέτουν μια σχετικά ισορροπημένη κατανομή των δεδομένων όσων αφορά την εξαρτημένη μεταβλητή. Γενικά, οι συνήθεις μέθοδοι ταξινόμησης εστιάζουν στην εύρεση του ταξινομητή με το ελάχιστο Error Rate. Αυτοί οι αλγόριθμοι αγνοούν τη διαφορά στα κόστη λανθασμένης ταξινόμησης. Πιο συγκεκριμένα, υποθέτουν ότι η λανθασμένη ταξινόμηση μέσα στο ίδιο σύνολο δεδομένων έχει ίδια κόστη είτε πρόκειται για false positive είτε για false negative, ενώ σε εφαρμογές του πραγματικού κόσμου αυτή η υπόθεση δεν ισχύει και οι διαφορές ανάμεσα στα κόστη από τη λανθασμένη ταξινόμηση μπορεί να είναι πολύ μεγάλες.

3.2 Μετρικές αξιολόγησης

Για την αξιολόγηση των μεθόδων ταξινόμησης, χρησιμοποιείται ένα πλήθος μετρικών. Η επιλογή της κατάλληλης μετρικής απόδοσης είναι μεγάλης σημασίας και συνήθως εξαρτάται από την κάθε περίπτωση που εξετάζουμε και το πρόβλημα που έχουμε να αξιολογήσουμε. Για παράδειγμα στην παλινδρόμηση θα προτιμηθούν διαφορετικές μετρικές από αυτές που χρησιμοποιούνται στην ταξινόμηση. Σε αυτήν την ενότητα θα επικεντρωθούμε στην περίπτωση της ταξινόμησης και θα αναλύσουμε τις βασικές μετρικές που προτείνονται για αυτή. Κάποιες από αυτές προτιμώνται αναλόγως με το αν έχουμε ισορροπημένο ή μη ισορροπημένο σύνολο δεδομένων, αλλά και με βάση το πρόβλημα που έχουμε να αντιμετωπίσουμε.

Οι μετρικές αυτές υπολογίζονται με τη χρήση του πίνακα σύγχυσης. Ο πίνακας σύγχυσης είναι ένας πίνακας που αποτελείται από τις πραγματικές τιμές των δεδομένων και τις τιμές πρόβλεψης που προκύπτουν από κάθε μοντέλο. Στην περίπτωση της δυαδικής ταξινόμησης, γίνεται η σύμβαση ότι υπάρχουν δύο περιπτώσεις, στις οποίες συνέβη ή δεν συνέβη ένα γεγονός, και οι οποίες θα αντιστοιχούν σε “positive” (κλάση 1) και “negative” (κλάση 0) αντίστοιχα. Η μορφή του δίνεται:

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Πίνακας 3.1: Πίνακας Σύγχυσης

Τα στοιχεία του πίνακα είναι:

- True Positive (TP): οι παρατηρήσεις που ανήκουν στην θετική κλάση και κατηγοριοποιούνται σωστά ως θετικά.
- False Positive (FP): οι παρατηρήσεις που ανήκουν στην αρνητική κλάση και κατηγοριοποιούνται λανθασμένα ως θετικά.

- False Negative (FN): οι παρατηρήσεις που ανήκουν στην θετική κλάση και κατηγοριοποιούνται λανθασμένα ως αρνητικά.
- True Negative (TN): οι παρατηρήσεις που ανήκουν στην αρνητική κλάση και κατηγοριοποιούνται σωστά ως αρνητικά.

Τα διαγώνια στοιχεία του πίνακα σύγχυσης υποδηλώνουν τις σωστά ταξινομημένες παρατηρήσεις ενώ τα μη διαγώνια στοιχεία υποδηλώνουν τις λανθασμένα ταξινομημένες παρατηρήσεις. Σαφώς, ο πίνακας σύγχυσης με την εφαρμογή ενός ιδανικού αλγορίθμου ταξινόμησης είναι διαγώνιος δηλαδή περιέχει μηδενικές τιμές στα μη διαγώνια στοιχεία του (FN = FP = 0) και μη μηδενικές τιμές στα στοιχεία της κύριας διαγωνίου του (TP και TN).

Οι πιο συχνές μετρικές αξιολόγησης της δυαδικής ταξινόμησης που μπορούν να υπολογισθούν μέσω των τεσσάρων προαναφερόμενων στοιχείων του πίνακα σύγχυσης είναι:

- Accuracy (Ορθότητα) = $\frac{TP+TN}{TP+FP+FN+TN}$: είναι το ποσοστό των σωστά κατηγοριοποιημένων παρατηρήσεων.
- Error Rate = 1 - Accuracy = $\frac{FP+FN}{TP+FP+FN+TN}$: είναι το ποσοστό των λάθος κατηγοριοποιημένων παρατηρήσεων.

Με τη χρήση συμβατικών αλγορίθμων μηχανικής μάθησης σε μη ισορροπημένα σύνολα δεδομένων, η χρήση του Accuracy και του Error Rate ως μετρικές αξιολόγησης των αποτελεσμάτων του μοντέλου δεν είναι κατάλληλη. Ένα παράδειγμα για να κατανοήσει κάποιος αυτό είναι το επόμενο. Εάν σε ένα δοσμένο σύνολο δεδομένων, η μειοψηφική κλάση είναι το 5% και η πλειοψηφική κλάση το 95%, μία αφελής προσέγγιση είναι η κατηγοριοποίηση όλων των παρατηρήσεων στην πλειοψηφική κλάση, τακτική που θα έδινε 95% Accuracy. Αν και το ποσοστό αυτό είναι εξαιρετικό, η τακτική που ακολουθήθηκε στο παράδειγμα, αναγνωρίζει το 0% της μειοψηφικής κλάσης που σε ορισμένες περιπτώσεις αποτελεί και το σημαντικότερο εγχείρημα. Σε τέτοιες περιπτώσεις, η μετρική Accuracy δεν παρέχει επαρκείς πληροφορίες για την λειτουργικότητα του ταξινομητή.

Παρατηρώντας τον πίνακα σύγχυσης, η πάνω σειρά αντιστοιχεί στις θετικές περιπτώσεις του συνόλου δεδομένων και η κάτω σειρά στις αρνητικές περιπτώσεις. Ωστόσο, η αναλογία των 2 γραμμών είναι αντιπροσωπευτική της κατανομής των κλάσεων του συνόλου δεδομένων και οποιαδήποτε μετρική χρησιμοποιεί τιμές από τις 2 γραμμές είναι ευαίσθητη σε περιπτώσεις μη ισορροπημένων συνόλων. Σύμφωνα με τους τύπους με τους οποίους ορίζονται οι μετρικές Accuracy και Error Rate, παρατηρείται ότι χρησιμοποιούν τιμές από διαφορετικές γραμμές του πίνακα σύγχυσης άρα είναι ευαίσθητες στις κατανομές των δεδομένων. Σε μη ισορροπημένα σύνολα δεδομένων χρησιμοποιούνται περισσότερο οι μετρικές Precision και Recall που ορίζονται στη συνέχεια. Η μετρική Precision είναι και αυτή ευαίσθητη στην κατανομή των δεδομένων σε αντίθεση με τη Recall. Παρόλα αυτά αν χρησιμοποιηθούν σωστά μπορούν να αξιολογήσουν αποτελεσματικά την επίδοση του αλγορίθμου σε μη ισορροπημένα σύνολα δεδομένων.

- Recall (Ανάκληση) = Sensitivity = True Positive Rate = $\frac{TP}{TP+FN}$: είναι το ποσοστό των θετικών περιπτώσεων που κατηγοριοποιήθηκαν σωστά.
- Precision (Ακρίβεια) = Positive Predictive Value = $\frac{TP}{TP+FP}$: είναι το ποσοστό των σωστών θετικών προβλέψεων που έγιναν σε σχέση με το σύνολο των θετικών προβλέψεων (οι οποίες μπορεί να είναι είτε λάθος, είτε σωστές).

Ο κύριος στόχος της μάθησης από μη ισοσσοπημένα σύνολα δεδομένων είναι η βελτίωση του Recall χωρίς την μείωση του Precision. Αυτό βέβαια δεν είναι δυνατό καθώς κατά την αύξηση του αριθμού των True Positives για την μειοψηφική κλάση (με συνέπεια την αύξηση του Recall), ο αριθμός των False Positives αυξάνεται επίσης, γεγονός που μειώνει το Precision. Γι αυτό το λόγο χρησιμοποιείται συχνά η μετρική F1-score που συνδυάζει τα trade-offs των Recall και Precision.

- F1-score = $\frac{2*Precision*Recall}{Precision+Recall}$: είναι ο αρμονικός μέσος των Recall και Precision. Ως γνωστών ο αρμονικός μέσος 2 αριθμών είναι πιο κοντά στον μικρότερο εκ των 2. Οπότε μία υψηλή τιμή δείχνει ότι και το Recall και το Precision έχουν υψηλές τιμές.

- Specificity = True Negative Rate = $\frac{TN}{TN+FP}$: είναι το ποσοστό των αρνητικών περιπτώσεων που κατηγοριοποιήθηκαν σωστά.
- False Positive Rate = 1 – Specificity = $\frac{FP}{FP+TN}$: είναι το ποσοστό των αρνητικών περιπτώσεων που κατηγοριοποιήθηκαν λανθασμένα.
- False Negative Rate = $\frac{FN}{FN+TP}$: είναι το ποσοστό των θετικών περιπτώσεων που κατηγοριοποιήθηκαν λανθασμένα.
- Negative Predictive Value = $\frac{TN}{TN+FN}$: είναι το ποσοστό των σωστών αρνητικών προβλέψεων που έγιναν σε σχέση με το σύνολο των αρνητικών προβλέψεων (οι οποίες μπορεί να είναι είτε λάθος, είτε σωστές).
- $G - mean = \sqrt{Recall * Specificity} = \sqrt{TPrate * TNrate}$: είναι ο γεωμετρικός μέσος των Recall και Specificity.

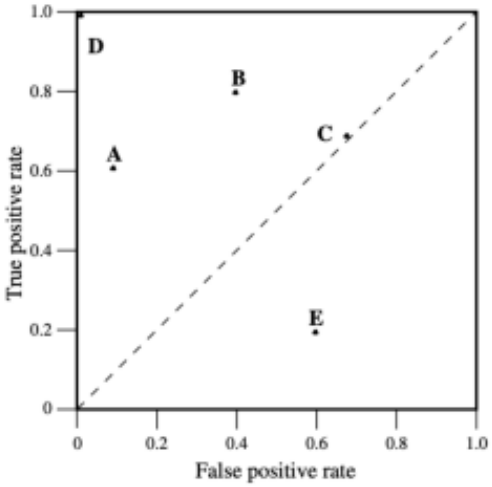
Ο σκοπός ενός ταξινομητή είναι να ελαχιστοποιήσει το False Positive Rate και το False Negative Rate ή ισοδύναμα να μεγιστοποιήσει το True Positive Rate και το True Negative Rate. Δυστυχώς, στα προβλήματα του πραγματικού κόσμου υπάρχει ένα trade-off ανάμεσα στα False Positive Rate και False Negative Rate.

Περισσότερο κατάλληλες για την αξιολόγηση των αλγορίθμων κατηγοριοποίησης κρίνονται η καμπύλη Λειτουργικών Χαρακτηριστικών (Receiver Operating Characteristics) και η καμπύλη Ακρίβεια-Ανάκληση (Precision-Recall).

Receiver Operating Characteristics (ROC) Curve

Κάποια πλεονεκτήματα της χρήσης της ROC καμπύλης σε μη ισορροπημένα σύνολα δεδομένων είναι ότι δεν λαμβάνει υπόψιν την κατανομή των κλάσεων και το κόστος λανθασμένης ταξινόμησης, είναι ανεξάρτητη της αναλογίας θετικών και αρνητικών περιπτώσεων μέσα στο δείγμα και δεν μεροληπτεί εναντίον της μειοψηφικής τάξης.

Ένας ταξινομητής που το αποτέλεσμα που εξάγει είναι το label της κλάσης και είναι διακριτό, παράγει ένα ζευγάρι (FPrate, TPrate) που αντιστοιχεί σε ένα σημείο του διδιάστατου γραφήματος. Στο σχήμα που ακολουθεί φαίνεται μια ROC καμπύλη με πέντε ταξινομητές A ώς E από τους οποίους ο καθένας θα έχει ένα ζευγάρι TPR-FPR το οποίο θα αντιστοιχεί σε ένα μοναδικό σημείο στο ROC διάγραμμα.



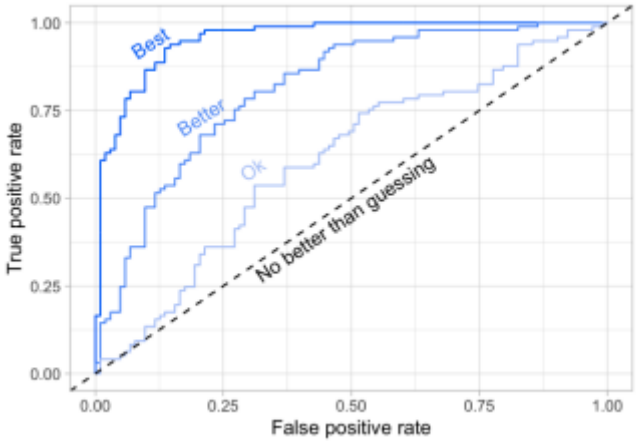
Σχήμα 3.1: Σχηματική αναπαράσταση της ROC καμπύλης

Μερικά σημεία ενδιαφέροντος στο διάγραμμα είναι:

- Το σημείο (0,0) δηλώνει πως όλες οι παρατηρήσεις θα ταξινομούνται στην αρνητική κλάση άρα ο ταξινομητής δεν σημειώνει κανένα FP σφάλμα και ταυτόχρονα δεν κερδίζει κανένα TP.
- Το σημείο (1, 1) όπου όλα τα δείγματα θα ταξινομούνται στη θετική κλάση που σημαίνει ότι ο ταξινομητής δεν σημειώνει κανένα FN σφάλμα και δεν κερδίζει κανένα TN.
- Το σημείο (0, 1) αναπαριστά την τέλεια ταξινόμηση που επιτυγχάνεται όταν τα σφάλματα FP και FN είναι μηδενικά και όλες οι παρατηρήσεις ταξινομούνται σωστά. Στο διάγραμμα ο ταξινομητής D είναι σε αυτή τη θέση.

Ένας ταξινομητής είναι καλύτερος από έναν άλλον αν το αντίστοιχο σημείο του στον ROC χώρο είναι πιο κοντά στο σημείο D(0,1) από ότι το άλλο σημείο. Στην περίπτωση που το αντίστοιχο σημείο του ταξινομητή είναι πάνω στην διαγώνιο του σχήματος (όπως το σημείο C), ο ταξινομητής αυτός παρέχει μία τυχαία υπόθεση για τα labels των κλάσεων. Ακόμη, αν το σημείο βρίσκεται στο κάτω δεξιά τρίγωνο του σχήματος, τότε ο ταξινομητής λειτουργεί χειρότερα από την τυχαία υπόθεση (όπως το σημείο E). Παρόλα αυτά, αυτός ο ταξινομητής δε σημαίνει ότι δεν παρέχει χρήσιμες πληροφορίες, αλλά κάνει λανθασμένη ταξινόμηση και προβλέπει ανάποδα τις κλάσεις (τα TP γίνονται FN και τα FP γίνονται TN).

Στην περίπτωση που ο ταξινομητής εξάγει ως αποτέλεσμα μια συνεχή ποσοτική τιμή που αναπαριστά την πιθανότητα μία παρατήρηση να ανήκει σε μία κλάση πρόβλεψης, η πρόβλεψη της κλάσης αλλάζει όσο ποικίλει η τιμή του επιλεγμένου κατωφλιού. Επομένως, κάθε κατώφλι ορίζει ένα ζευγάρι TPR-FPR οπότε για τις διάφορες τιμές του παράγεται μία σειρά από σημεία στον ROC χώρο και άρα σχηματίζεται μία καμπύλη.



Σχήμα 3.2: Σύγκριση ROC καμπυλών

Η καμπύλη παρουσιάζεται σε ένα διδιάστατο διάγραμμα όπου ο άξονας x αναπαριστά το FPrate ενώ ο άξονας y το TPrate. Αναπαριστά τα trade-offs ανάμεσα στα benefits (True Positives) και τα costs (False Positives). Βάση του διαγράμματος, αναγνωρίζεται ότι ένας ταξινομητής δεν έχει την ικανότητα να αυξήσει τον αριθμό των true positives χωρίς να αυξήσει τον αριθμό των false positives.

Για την σύγκριση αρκετών μοντέλων κατηγοριοποίησης με τη χρήση των ROC καμπυλών τους, το μοντέλο που η καμπύλη του κυριαρχεί των υπολοίπων σε όλο το χώρο, είναι και το βέλτιστο. Επομένως, όσο πιο υψηλή και κοντά στο σημείο (0, 1) είναι η ROC καμπύλη, τόσο καλύτερα μπορεί το μοντέλο να ξεχωρίσει τα χαρακτηριστικά των δειγμάτων των δύο κλάσεων. Μία ακόμη χρήσιμη και αξιόπιστη μετρική είναι η τιμή AUC, η οποία ισούται με την περιοχή κάτω από την ROC καμπύλη. Ουσιαστικά ισούται με την πιθανότητα ένας δυαδικός ταξινομητής να κατατάξει μία τυχαία θετική παρατήρηση υψηλότερα σε σχέση με μία τυχαία αρνητική παρατήρηση.

Δηλαδή, εκφράζει την ικανότητα του ταξινομητή να διακρίνει μεταξύ θετικών και αρνητικών τάξεων. Η τιμή του AUC υπολογίζεται κατά προσέγγιση από την παρακάτω σχέση:

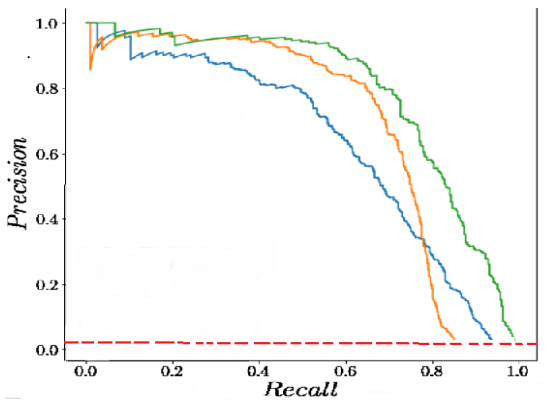
$$\text{AUC} = \frac{1+TPrate-FPrate}{2} = \frac{TPrate+TNrate}{2} = \frac{Sensitivity+Specificity}{2}$$

Στην περίπτωση που δεν κυριαρχεί μία ROC καμπύλη έναντι των υπολοίπων στο διάγραμμα, για την σύγκριση ταξινομητών κατά μέσο όρο μπορεί να χρησιμοποιηθεί η τιμή AUC. Η baseline τιμή AUC για τις καμπύλες ROC είναι 0.5 και αντιστοιχεί σε έναν ταξινομητή που πραγματοποιεί τυχαία ταξινόμηση.

Αν και η ROC καμπύλη παρέχει έναν χρήσιμο τρόπο για την αξιολόγηση της επίδοσης του ταξινομητή, ωστόσο έχει και κάποιους περιορισμούς. Στην περίπτωση των μη ισορροπημένων δεδομένων, παρατηρείται ότι η ROC καμπύλη παρέχει μία πολύ αισιόδοξη οπτική της επίδοσης του αλγορίθμου. Σε αυτήν την περίπτωση η Precision-Recall καμπύλη αξιολογεί περισσότερο αποδοτικά τον ταξινομητή.

Precision-Recall Curve (PRC)

Η καμπύλη Precision-Recall Curve χρησιμοποιείται σε εφαρμογές όπου έχει σημασία η ταξινόμηση των θετικών δειγμάτων και δείχνει το trade-off μεταξύ Recall και Precision για διαφορετικά κατώφλια. Μία μεγάλη περιοχή κάτω από την καμπύλη αναπαριστά υψηλό recall και υψηλό precision, όπου το υψηλό recall σχετίζεται με μικρό False Negative και το υψηλό precision σχετίζεται με χαμηλό False Positive. Ο στόχος ενός ταξινομητή είναι να επιτευχθούν μικρές τιμές για τα False Positives και False Negatives, άρα μεγάλο Recall και Precision.



Σχήμα 3.3: Σχηματική αναπαράσταση της Precision-Recall καμπύλης

Η baseline της PRC δεν είναι σταθερή αλλά καθορίζεται από την αναλογία των θετικών παρατηρήσεων στο σύνολο του δείγματος. Αν θεωρηθεί το πλήθος των παρατηρήσεων της θετικής κλάσης ως P και το πλήθος των παρατηρήσεων της αρνητικής κλάσης ως N τότε η ευθεία baseline στο σχήμα προκύπτει ως εξής: y = P/(P + N). Επομένως στην περίπτωση του ισορροπημένου συνόλου δεδομένων προκύπτει η ευθεία y = 0.5, αλλά στην περίπτωση ενός μη-ισορροπημένου συνόλου δεδομένων που η αναλογία των P:N θα είναι 1 : 10, η baseline προκύπτει να είναι $y = \frac{1}{11} = 0.09$.

Σε περιπτώσεις μη ισορροπημένων συνόλων δεδομένων, η ROC καμπύλη αδυνατεί να περιγράψει γραφικά την απώλεια στην απόδοση, καθώς δεν εμφανίζει ρητά τις μεγάλες διαφορές στην τιμή των Precision και Recall. Το προτιμότερο θα ήταν επομένως ειδικά σε περιπτώσεις με μη ισορροπημένα σύνολα δεδομένων να ελέγχουμε και τις δύο καμπύλες ROC και PRC καθώς η κάθε μια περιέχει διαφορετική πληροφορία για το μοντέλο, προσφέροντας μια πιο ολοκληρωμένη εικόνα της απόδοσής του.

3.3 Εφαρμογές στην Επιχειρησιακή Αναλυτική

Τα προβλήματα ταξινόμησης μη ισορροπημένων συνόλων δεδομένων έχουν ποικίλες εφαρμογές στον τομέα των επιχειρήσεων. Οι κυριότερες εφαρμογές είναι:

- **Ανίχνευση απάτης (Fraud detection):** Με την ανάπτυξη της τεχνολογίας και των επικοινωνιών, έχει παρατηρηθεί αύξηση των περιστατικών ηλεκτρονικής απάτης με αποτέλεσμα την απώλεια εκατομμυρίων δολαρίων κάθε χρόνο. Η εύρεση λύσεων γι αυτό το πρόβλημα είναι πολύ σημαντική για πολλές επιχειρήσεις. Είναι σύνηθες στις βάσεις δεδομένων των επιχειρήσεων να υπάρχουν συναλλαγές στις οποίες οι έμπιστοι χρήστες είναι πολλοί περισσότεροι από ότι οι δόλιοι. Οι οργανισμοί προσπαθούν να αναγνωρίσουν την απάτη, ελέγχοντας ύποπτα μοτίβο συναλλαγών που υπάρχουν στις βάσεις δεδομένων τους. Υπάρχουν διάφορα είδη απάτης όπως τραπεζικές απάτες, ασφαλιστικές απάτες, email phishing, κλπ.

- **Διαχείριση Πελατειακών Σχέσεων (Customer Relationship Management-CRM):** Η αναγνώριση των εν δυνάμει πελατών είναι υψίστης σημασίας για τις πωλήσεις, τα κέρδη και την ανάπτυξη κάθε εταιρίας. Πολλές εταιρίες κατά τη διάρκεια καμπανιών προσέγγισης των πελατών που κύριο στόχο είχαν την προώθηση ενός προϊόντος ή μιας υπηρεσίας, έχουν συλλέξει δεδομένα που αφορούν την αγορά ή όχι του προϊόντος καθώς και πληροφορίες για τον πελάτη. Οι εταιρίες αξιοποιούν τα δεδομένα αυτά, προσπαθούν να ανταποκριθούν στις ανάγκες και στις επιθυμίες των πελατών με συστηματικό τρόπο με σκοπό να δημιουργήσουν πιστούς πελάτες, να μειώσουν την δαπανηρή επικοινωνία με λάθος άτομα και να αυξήσουν τα κέρδη τους.
- **Πρόβλεψη ανατροπής (Churn prediction):** Πρόκειται για την τάση των πελατών να παύουν να χρησιμοποιούν ένα προϊόν ή μία υπηρεσία από μία εταιρία ή έναν οργανισμό κατά τη διάρκεια ενός συγκεκριμένου διαστήματος.
- **Ιατρική Διάγνωση (Medical diagnosis):** Πρόκειται για το πρόβλημα της διάγνωσης μίας ασθένειας όπου τα περιστατικά ασθένειας είναι σπανιότερα σε σύγκριση με τον γενικό πληθυσμό. Ο στόχος σε αυτήν την περίπτωση είναι η αναγνώριση των ατόμων με την ασθένεια.

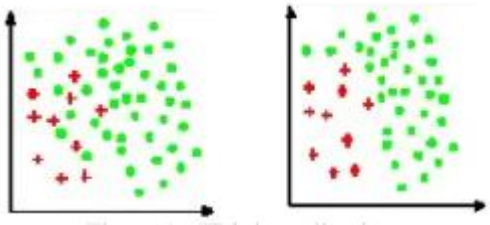
Συγκεκριμένα, στο παράδειγμα της ιατρικής διάγνωσης μίας ασθένειας, αν θεωρηθεί ως positive κλάση το να έχει κάποιος την ασθένεια και ως negative κλάση να είναι υγιής, το κόστος μιας false positive πρόβλεψης είναι μικρότερο και περιορίζεται σε επιπρόσθετους ιατρικούς ελέγχους στους οποίους θα υποβληθεί το άτομο, ενώ το κόστος μιας false negative πρόβλεψης είναι σημαντικότερο καθώς ο ασθενής θεωρείται υγιής, δεν ερευνάται παραπάνω η κατάστασή του για να βρεθεί η κατάλληλη θεραπεία και μπορεί να χάσει τη ζωή του. Όμοια, στην ανίχνευση απάτης στις τραπεζικές συναλλαγές, αν θεωρηθεί ως positive κλάση το να είναι η συναλλαγή δόλια, το κόστος μιας false positive πρόβλεψης μπορεί να προκαλέσει περαιτέρω άγονη έρευνα αυτών των συναλλαγών, ενώ μία false negative πρόβλεψη έχει ως αποτέλεσμα την πραγματοποίηση της δόλιας συναλλαγής. Αντίθετα, στην περίπτωση της διαχείρισης σχέσεων πελατών, αν θεωρηθεί ως positive κλάση το να αγοράσει το προϊόν ο πελάτης, μία false negative πρόβλεψη μπορεί να προκαλέσει επιπρόσθετα κόστη προώθησης των προϊόντων στον πελάτη ενώ το κόστος από μία false positive

πρόβλεψη (ότι το άτομο θα αγοράσει το προϊόν ενώ δεν θα το αγοράσει) είναι σημαντικότερο καθώς σημαίνει απώλεια του πελάτη και επομένως λιγότερα κέρδη καθώς η εταιρία θεωρεί ότι είναι ήδη πελάτης της και δεν θα προχωρήσει σε κάποια καμπάνια μάρκετινγκ.

3.4 Διαχείριση μη ισορροπημένων συνόλων δεδομένων

Υπάρχουν αρκετές τεχνικές διαχείρισης μη ισορροπημένων συνόλων δεδομένων για τις οποίες θα γίνει μία σύντομη αναφορά:

1) **Μέθοδοι δειγματοληψίας:** Αυτές οι μέθοδοι χειρίζονται την άνιση κατανομή των κλάσεων στο σύνολο εκπαίδευσης με δειγματοληψία είτε της πλειοψηφικής είτε της μειοψηφικής κλάσης.
    i) **Υποδειγματοληψία** (Under-Sampling): Κάποιες τεχνικές είναι:
        - **Random under-sampling:** Με αυτή τη μέθοδο επιλέγονται τυχαία παρατηρήσεις από την πλειοψηφική κλάση και απομακρύνονται μέχρι ο αριθμός των παρατηρήσεων στις 2 κλάσεις να είναι ίδιος. Το βασικό και προφανές πρόβλημα που προκύπτει με αυτή την τεχνική είναι ότι υπάρχει η σημαντική απώλεια της πληροφορίας που παρέχουν οι παρατηρήσεις που απομακρύνονται από το δείγμα.
        - **Tomek links:** Ο αλγόριθμος λειτουργεί με την παρακάτω διαδικασία. Έστω 2 παρατηρήσεις που ανήκουν σε διαφορετικές κλάσεις (πλειοψηφική και μειοψηφική). Το ζεύγος των 2 σημείων λέγεται Tomek Link αν δεν υπάρχει άλλο σημείο ανάμεσά τους με μικρότερη απόσταση. Αν 2 σημεία αποτελούν Tomek Link τότε είτε το ένα είναι θόρυβος είτε και τα 2 είναι όρια κλάσεων. Η μέθοδος μπορεί να χρησιμοποιηθεί είτε ως under-sampling μέθοδος είτε σαν μέθοδος data cleaning. Στην πρώτη περίπτωση απομακρύνεται η παρατήρηση που ανήκει στην πλειοψηφική κλάση ενώ στην δεύτερη περίπτωση απομακρύνεται όλο το Tomek Link ζεύγος.



Σχήμα 3.4: TomekLink εφαρμογή

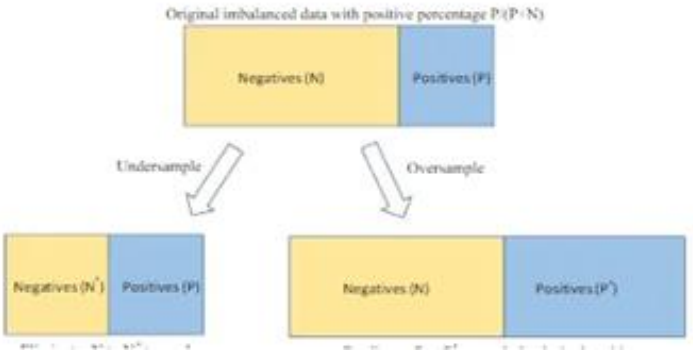
- **Edited Nearest Neighbour (ENN):** Ο αλγόριθμος αυτός απομακρύνει οποιαδήποτε παρατήρηση της οποίας η κλάση είναι διαφορετική από την κλάση τουλάχιστον των 2 εκ των 3 κοντινότερων γειτόνων. Για ένα binary classification πρόβλημα, ο αλγόριθμος μπορεί να περιγραφεί ως ακολούθως: Για κάθε παρατήρηση Ei του συνόλου εκμάθησης, βρίσκονται οι 3 κοντινότεροι του γείτονες. Αν το Ei ανήκει στην πλειοψηφική κλάση και η κατηγοριοποίηση των 3 κοντινότερων γειτόνων δίνει αντίθετο αποτέλεσμα του πραγματικού, τότε το Ei απομακρύνεται. Αν το Ei ανήκει στην μειοψηφική κλάση και οι 3 κοντινότεροι γείτονές του το κατηγοριοποιούν λανθασμένα τότε απομακρύνονται οι κοντινότεροι γείτονες που ανήκουν στην πλειοψηφική κλάση.

ii) **Υπερδειγματοληψία** (Over-Sampling): Κάποιες τεχνικές είναι:

- **Random over-sampling :** Με αυτή τη μέθοδο επιλέγονται τυχαία παρατηρήσεις από τη μειοψηφική κλάση και με επαναδειγματοληψία δημιουργούνται αντίγραφα αυτών μέχρι ο αριθμός των νέων παρατηρήσεων μαζί με τον αριθμό των αρχικών παρατηρήσεων να είναι σε πλήθος ίδιος με τον αριθμό της πλειοψηφικής κλάσης. Ένα μειονέκτημα της μεθόδου αυτής είναι ότι αυξάνει το μέγεθος του συνόλου εκμάθησης με αποτέλεσμα να αυξάνεται και ο χρόνος που απαιτείται για την εφαρμογή των αλγορίθμων. Το κυριότερο μειονέκτημα είναι βέβαια ότι με την εισαγωγή αντιγράφων των παρατηρήσεων της μειοψηφικής κλάσης εισάγεται μεροληψία και μπορεί να οδηγήσει σε υπερπροσαρμογή. Επιπλέον, δεν βοηθάει στην επίλυση του προβλήματος έλλειψης δεδομένων για την

μειοψηφική κλάση καθώς δεν εισάγει νέες παρατηρήσεις αλλά μόνο ακριβή αντίγραφα των υπαρχόντων.

- **SMOTE (Synthetic Minority Over-sampling Technique*):*** Δημιουργεί συνθετικά δεδομένα της μειοψηφικής κλάσης με παρεμβολή των ήδη υπαρχόντων (interpolation). Για κάθε παρατήρηση της μειοψηφικής κλάσης βρίσκονται οι k κοντινότεροι γείτονες και επιλέγονται τυχαία κάποιοι από αυτούς σύμφωνα με το ρυθμό του oversampling. Μία νέα συνθετική παρατήρηση δημιουργείται κατά μήκος της γραμμής που ενώνει την αρχική παρατήρηση και κάθε έναν από τους κοντινότερους γείτονες που έχουν επιλεγεί.



Σχήμα 3.5: Γραφική αναπαράσταση των τεχνικών υποδειγματοληψίας και υπερδειγματοληψίας

Μελέτες έχουν δείξει ότι η random over-sampling δεν βελτιώνει την εκμάθηση του μοντέλου ενώ η random under-sampling φαίνεται να έχει ένα πλεονέκτημα σε σύγκριση με την random over-sampling. Κάποιες έρευνες έχουν συνδυάσει τις 2 αυτές προσεγγίσεις, την random over-sampling για τη βελτίωση του accuracy της κατηγοριοποίησης και την random under-sampling για τη μείωση του μεγέθους του συνόλου εκπαίδευσης. Τα μειονεκτήματα των 2 μεθόδων δειγματοληψιών μπορούν να αντιμετωπιστούν με τη χρήση πιο προχωρημένων στρατηγικών δειγματοληψίας όπως αυτών που αναφέρθηκαν παραπάνω.

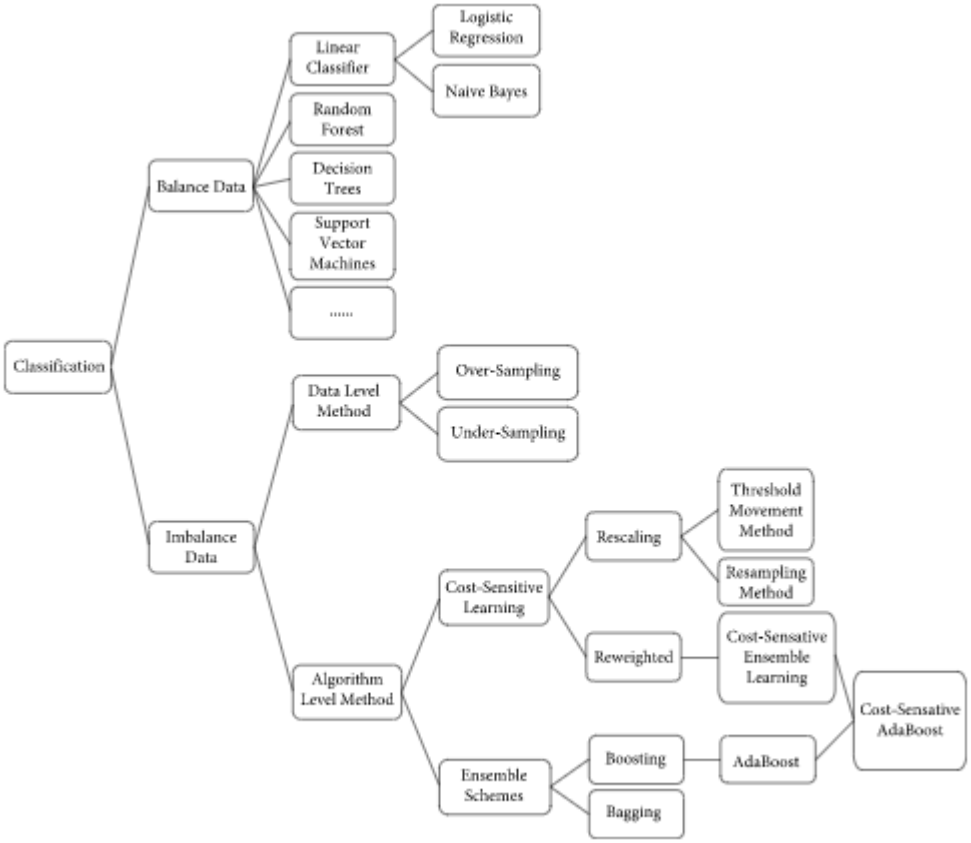
2) **Συλλογική Μάθηση (Ensemble Learning)**: Η Συλλογική Μηχανική Μάθηση είναι μία τεχνική κατά την οποία συνδυάζεται ένα πλήθος αδύναμων ταξινομητών (weak

classifiers) με στόχο την δημιουργία ενός ισχυρού ταξινομητή για την επίτευξη ενός καλύτερου αποτελέσματος. Υπάρχουν δύο είδη συλλογική μάθησης, το bagging στο οποίο οι αδύναμοι ταξινομητές δημιουργούνται παράλληλα και δεν έχουν εξάρτηση μεταξύ τους, και το boosting στο οποίο οι αδύναμοι ταξινομητές δημιουργούνται διαδοχικά και έχουν ισχυρή εξάρτηση μεταξύ τους.

- **Bagging** (***Bootstrap AGGregatING***): Είναι μία μέθοδος που συνδυάζει τις προβλέψεις των ταξινομητών που εφαρμόζονται σε διαφορετικά σύνολα εκμάθησης που προκύπτουν με τη δειγματοληψία Bootstrap. Η τεχνική bagging βασίζεται στην απλή ιδέα ότι σε κάθε επανάληψη δημιουργούνται bootstrap αντίγραφα από το σύνολο m δειγμάτων εκπαίδευσης και κάθε ταξινομητής εκπαιδεύεται στο συγκεκριμένο bootstrap σύνολο. Στο τέλος ο κάθε ταξινομητής επιστρέφει τη δική του πρόβλεψη και όλες μαζί συνδυάζονται. Στην περίπτωση της κατηγοριοποιήσης το αποτέλεσμα λαμβάνεται με πλειοψηφική απόφαση (majority voting) —δηλαδή η τιμή με τις περισσότερες ψήφους επιλέγεται ως η μεταβλητή απόκρισης— ενώ στην περίπτωση της παλινδρόμησης το αποτέλεσμα προκύπτει από τον μέσο όρο τους (averaging vote). Με αυτόν τον τρόπο, εξομαλύνονται οι αποκλίσεις, μειώνεται η διακύμανση της πρόβλεψης και πετυχαίνεται καλύτερη ακρίβεια. Αυτού του είδους το η Συλλογική Μάθηση λειτουργεί παράλληλα και ανεξάρτητα, αφού ο κάθε ταξινομητής εκπαιδεύεται σε ένα bootstrap σύνολο δεδομένων και τελικά το αποτέλεσμα προκύπτει από το voting ή το averaging τους.
- **Boosting:** Πρόκειται για μια επαναληπτική διαδικασία που δημιουργεί έναν ισχυρό ταξινομητή, από ένα σύνολο επιμέρους αδύναμων ταξινομητών, οι οποίοι αν λειτουργούσαν ανεξάρτητα δεν θα έδιναν καθόλου καλή πρόβλεψη. Αναλυτικά, η διαδικασία boosting έχει ως εξής: αρχικά στην πρώτη επανάληψη ο αρχικός ταξινομητής εκπαιδεύεται με ίσα βάρη για όλα τα δεδομένα του αρχικού συνόλου εκπαίδευσης. Στη συνέχεια όλα τα βάρη προσαρμόζονται με βάση το αν οι παρατηρήσεις ταξινομήθηκαν σωστά στην πρώτη επανάληψη, με τέτοιο τρόπο ώστε σε κάθε επόμενη επανάληψη τα βάρη των παρατηρήσεων που

ταξινομήθηκαν λανθασμένα να αυξάνονται ενώ να μειώνονται τα βάρη των παρατηρήσεων που ταξινομήθηκαν σωστά. Με αυτόν τον τρόπο αναπροσαρμόζονται τα βάρη των παρατηρήσεων με λανθασμένες προβλέψεις έτσι ώστε ο ταξινομητής να τείνει να επικεντρωθεί σε αυτές. Πράγματι στις επόμενες ταξινόμησεις, το μοντέλο θα εστιάζει στα στοιχεία που έχουν ταξινομηθεί εσφαλμένα. Αυτή η διαδικασία θα επαναλαμβάνεται έως ότου το ολικό σφάλμα γίνει αρκετά μικρό. Βέβαια, σε imbalanced datasets είναι γνωστό ότι οι παρατηρήσεις που ανήκουν στην μειοψηφική κλάση είναι περισσότερο εκτεθειμένες σε σφάλμα από ότι οι παρατηρήσεις της πλειοψηφικής κλάσης. Γι αυτό είναι εύλογο να θεωρηθεί ότι η μέθοδος boosting βελτιώνει την απόδοση της κατηγοριοποίησης καθώς αυξάνει τα βάρη των παρατηρήσεων της μειοψηφικής κλάσης.

3) **Cost-sensitive Learning:** Αναθέτει διαφορετικά βάρη στις κλάσεις όταν συμβαίνει λανθασμένη ταξινόμηση. Γενικά, οι συνήθεις μέθοδοι ταξινόμησης εστιάζουν στην εύρεση του ταξινομητή με το ελάχιστο Error Rate e=FP+FN, χωρίς να υπάρχει περιορισμός στα ποσοστά FP και FN στο Error ή ισοδύναμα στοχεύουν στην εύρεση υψηλού Accuracy. Οι συνήθεις μέθοδοι αυτοί ονομάζονται Cost-insensitive classifiers και υποθέτουν ότι το κόστος των false-negative και false-positive προβλέψεων είναι το ίδιο. Σε πραγματικά προβλήματα, αυτή η υπόθεση δεν ισχύει όπως αναλύθηκε και στην ενότητα 3.3. Παρόλα αυτά, οι Cost-sensitive Learning μέθοδοι προσπαθούν να βρουν τον ταξινομητή με το ελάχιστο κόστος λανθασμένης ταξινόμησης. Πολλοί δημοφιλείς αλγόριθμοι ταξινόμησης μπορούν να προσαρμοστούν κάτω από αυτή τη δομή. Για παράδειγμα, τροποποιώντας τον αλγόριθμο SVM προκύπτει ο αλγόριθμος Weighted SVM, τροποποιώντας τον αλγόριθμο Logistic Regression προκύπτει ο Weighted Logistic Regression κτλ.



Σχήμα 3.6: Μέθοδοι διαχείρισης ισορροπημένων και μη ισορροπημένων συνόλων δεδομένων με σκοπό την ταξινόμηση

Κεφάλαιο 4

Περίπτωση Μελέτης

4.1 Δεδομένα

Το αρχείο δεδομένων πάρθηκε από το UCI Machine Learning Repository και πρόκειται για το bank-additional.csv. Περιέχει 4.119 παρατηρήσεις και 21 μεταβλητές, εκ των οποίων οι 20 είναι επεξηγηματικές μεταβλητές (ανεξάρτητες) και η 1 είναι μεταβλητή απόκρισης (εξαρτημένη). Ουσιαστικά προέρχεται από ένα μεγαλύτερο σύνολο δεδομένων γνωστό ως Bank Marketing Dataset, το οποίο συλλέχθηκε κατά τη διάρκεια μια άμεσης καμπάνιας μάρκετινγκ μιας Πορτογαλικής τράπεζας. Η καμπάνια αυτή ήταν βασισμένη σε τηλεφωνήματα. Κάθε εν δυνάμει πελάτης προσεγγίστηκε μέσω περισσότερων του ενός τηλεφωνήματος για να συλλέξει η τράπεζα την απόφαση του. Στο κεφάλαιο αυτό θα ασχοληθούμε με την περιγραφή, την οπτικοποίηση, την προεπεξεργασία των δεδομένων και τη σύγκριση των 2 κλάσεων και στο επόμενο κεφάλαιο θα προχωρήσουμε σε ταξινόμηση ώστε να βρεθεί αν ο κάθε ερωτώμενος θα πραγματοποιήσει προθεσμιακή κατάθεση ή όχι. Μετά την πραγματοποίηση της κλήσης ανάμεσα στον εκπρόσωπο της τράπεζας και στον πελάτη, η διάρκεια του τηλεφωνήματος είναι γνωστή όπως και η θετική ή η αρνητική του απάντηση στην καμπάνια. Γι αυτό το λόγο, η μεταβλητή «duration» θα απομακρυνθεί από το σύνολο δεδομένων μας ώστε το μοντέλο ταξινόμησης να δίνει ρεαλιστικά αποτελέσματα πρόβλεψης χρησιμοποιώντας τα ιστορικά δεδομένα που είναι συγκεντρωμένα για τον κάθε πελάτη.

Συγκεντρωτικά, οι 21 μεταβλητές του συνόλου είναι:

Κατηγορία	Μεταβλητή	Είδος
Personal Client Information	1.Age	Ανεξάρτητη
	2.Job	Ανεξάρτητη
	3.Marital	Ανεξάρτητη

	4.Education	Ανεξάρτητη
Bank Client Information	5.Default	Ανεξάρτητη
	6.Housing	Ανεξάρτητη
	7.Loan	Ανεξάρτητη
Contact Information	8.Contact	Ανεξάρτητη
	9.Month	Ανεξάρτητη
	10.Day of Week	Ανεξάρτητη
	11.Duration	Ανεξάρτητη
	12.Campaign	Ανεξάρτητη
History Information	13.Pdays	Ανεξάρτητη
	14.Previous	Ανεξάρτητη
	15.Poutcome	Ανεξάρτητη
Social and economic indicators	16.Emp var rate	Ανεξάρτητη
	17.Cons price idx	Ανεξάρτητη
	18.Cons conf idx	Ανεξάρτητη
	19.Euribor3m	Ανεξάρτητη
	20.Nr employed	Ανεξάρτητη
Response variable	21.y	Εξαρτημένη

Πίνακας 4.1: Οι μεταβλητές του συνόλου δεδομένων

4.2 Μεθοδολογία προ-επεξεργασίας

Όσον αφορά την προ-επεξεργασία των δεδομένων, ένα σημαντικό ζήτημα είναι η διαχείριση των ελλειπουσών τιμών (missing values). Στις κατηγορικές μεταβλητές παρατηρείται ότι υπάρχουν τιμές ως «unknown». Ένας τρόπος διαχείρισης αυτών είναι να απομακρυνθούν οι παρατηρήσεις που έχουν ως τιμή την «unknown» σε

κάποια μεταβλητή, αυτό όμως θα οδηγούσε σε πολύ μεγάλη μείωση του πλήθος των παρατηρήσεων-εγγραφών και θα επηρέαζε τα αποτελέσματα μας. Ένας άλλος τρόπος διαχείρισης είναι να γίνει imputation των ελλειπουσών τιμών με κάποιο περιγραφικό μέτρο όπως η επικρατούσα τιμή στην περίπτωση των κατηγορικών μεταβλητών ή με την μέση τιμή ή τη διάμεσο στην περίπτωση των ποσοτικών μεταβλητών. Στην συγκεκριμένη έρευνα θα χρησιμοποιηθεί η μέθοδος του imputation των ελλειπουσών τιμών με τη χρήση μεταβλητών από τις οποίες φαίνεται να υπάρχει κάποιου είδους συσχέτιση. Θα χρησιμοποιηθεί δηλαδή η μέθοδος cross-tabulation ανάμεσα σε κατηγορικές μεταβλητές που έχουν κάποια εξάρτηση και οι τιμές της μίας μεταβλητής επηρεάζουν τις τιμές της άλλης. Για παράδειγμα, το cross-tabulation μεταξύ των «job» και «education» μπορεί να χρησιμοποιηθεί βάσει της υπόθεσης ότι η δουλειά ενός ατόμου επηρεάζεται από την εκπαίδευσή του. Επομένως, οι ελλειπούσες τιμές ενός ατόμου στην μεταβλητή «job» μπορούν να βρεθούν παρατηρώντας τη δουλειά των περισσότερων ατόμων του δείγματος που έχουν την ίδια εκπαίδευση με αυτόν. Στην περίπτωση που δεν φαίνεται να υπάρχει κάποια λογική συσχέτιση ανάμεσα σε άλλες μεταβλητές, χρησιμοποιείται η επικρατούσα μεταβλητή κάθε μεταβλητής για το imputation των missing values.

Άλλο ένα σημαντικό θέμα που προκύπτει είναι η διαχείριση των ακραίων τιμών (outliers) του δείγματος. Όσον αφορά τη μεταβλητή «age» υπάρχουν outliers τα οποία θα βρεθούν και θα αφαιρεθούν από το δείγμα μας.

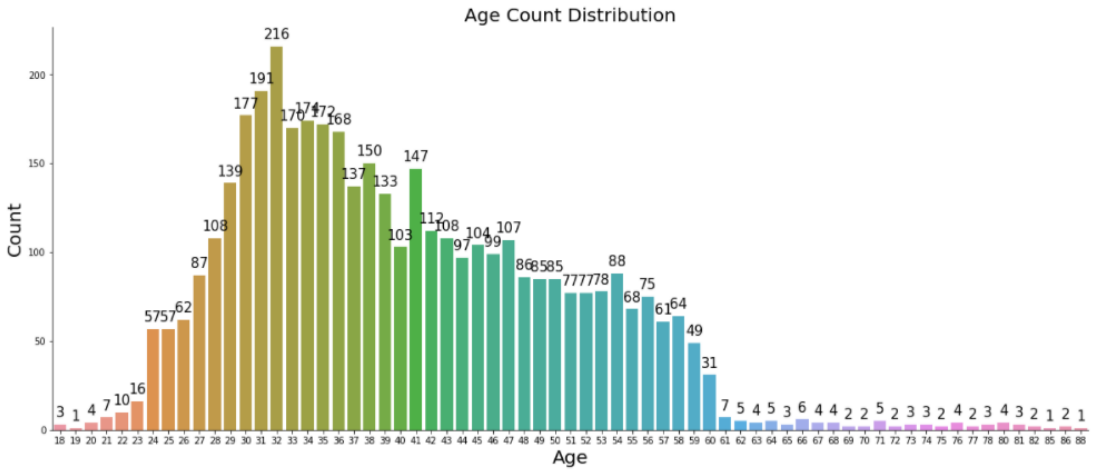
4.3 Περιγραφική στατιστική

Παρακάτω παρουσιάζεται το ιστόγραμμα της μεταβλητής «Age». Στον πληθυσμό φαίνεται η μικρότερη ηλικία να είναι τα 18 έτη και μεγαλύτερη ηλικία τα 88 έτη. Είναι εμφανές από το διάγραμμα ότι υπάρχει δεξιά ασυμμετρία των τιμών με επικρατούσα τιμή τα 32 έτη, διάμεση τιμή τα 38 έτη και μέση τιμή 40,11.

Περιγραφικά μέτρα	**Τιμή**
Πλήθος	4.119

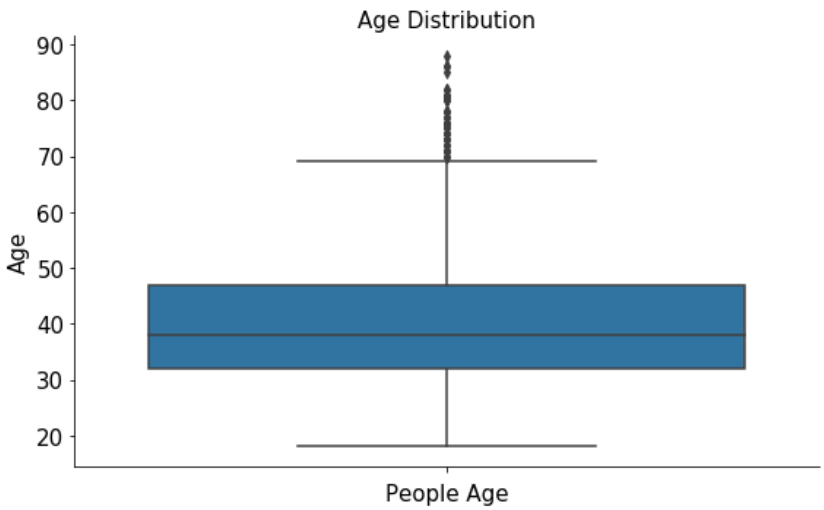
Μέσος όρος	40,11
Τυπική απόκλιση	10,31
Ελάχιστο στοιχείο	18
1° τεταρτημόριο	32
2° τεταρτημόριο	38
3° τεταρτημόριο	47
Μέγιστο στοιχείο	88

Πίνακας 4.2: Περιγραφικά στοιχεία για τη μεταβλητή «Age»



Σχήμα 4.1: Age Count Distribution (histogram)

Δίνεται το θηκόγραμμα από το οποίο παρατηρείται ότι υπάρχουν ακραίες τιμές οι οποίες στην πορεία θα αφαιρεθούν από την ανάλυση.



Σχήμα 4.2: Age Distribution (boxplot)

Από τον πίνακα με τα μέτρα θέσης και διασποράς, οι τιμές των τεταρτημορίων είναι Q1=32, Q2=38, Q3=47. Το ενδοτεταρτημοριακό εύρος υπολογίζεται να είναι IQ=Q3-Q1=47-32=15. Οι ακραίες τιμές είναι αυτές που είναι μεγαλύτερες από Q3+1,5*IQ=47+1,5*15=69,5. Το πλήθος των outliers είναι 39 στα 4.119 άτομα, με το ποσοστό επί του γενικού συνόλου να ισούται με 0,9%.

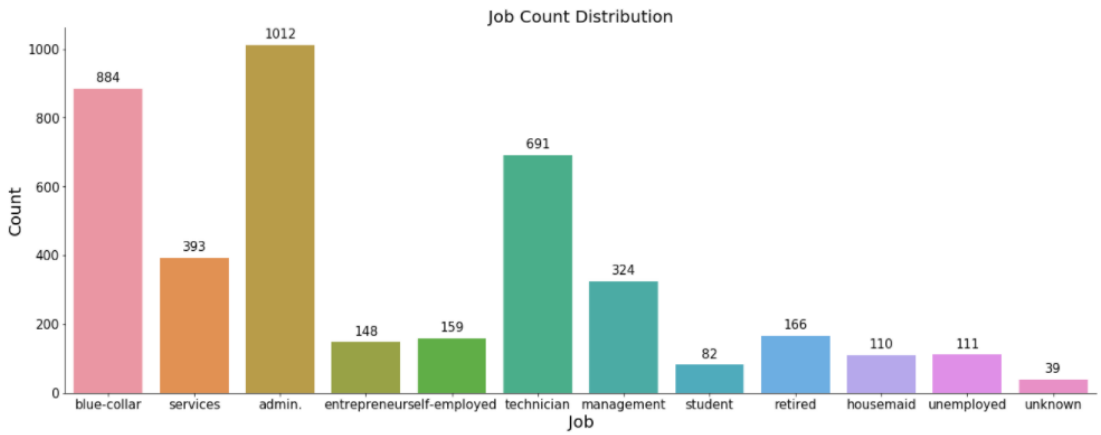
Δουλειά

Υπάρχουν 12 διαφορετικές τιμές για τη μεταβλητή «Job». Η μεταβλητή αυτή είναι κατηγορική και έχει επικρατούσα τιμή την τιμή admin με συχνότητα 1.012, το οποίο σημαίνει ότι το δείγμα αποτελείται από 1.012 άτομα που απασχολούνται στη διοίκηση (admin). Τη δεύτερη μεγαλύτερη συχνότητα έχουν τα άτομα που απασχολούνται σε χειρωνακτικά επαγγέλματα (blue-collar) και οι τεχνίτες (technician).

Επάγγελμα	**Πλήθος**
admin	1.012
blue-collar	884
technician	691
services	393

management	324
retired	166
self-employed	159
entrepreneur	148
unemployed	111
housemaid	110
student	82
unknown	39
Σύνολο	4.119

Πίνακας 4.3: Πίνακας συχνοτήτων της μεταβλητής «job»



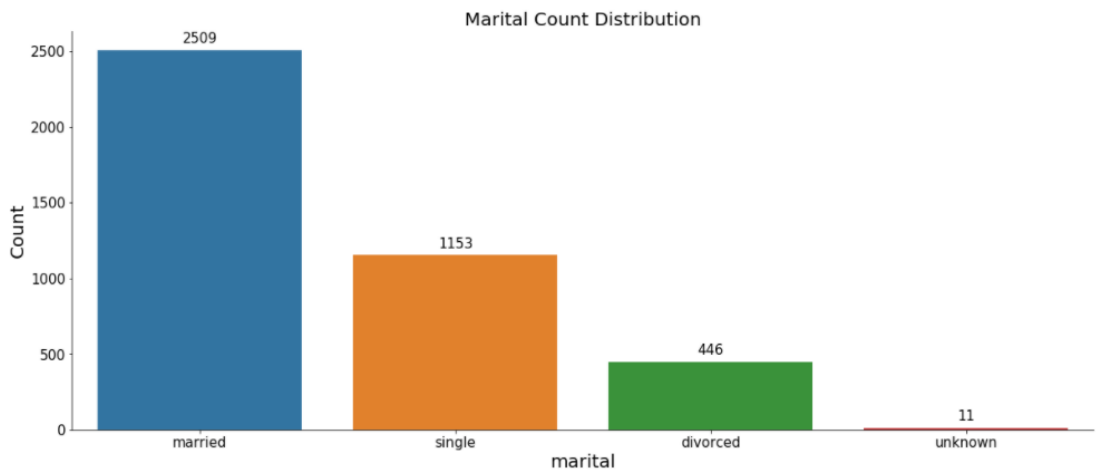
Σχήμα 4.3: Job Count Distribution (barplot)

Γάμος

Η μεταβλητή «marital» μπορεί να πάρει 4 δυνατές τιμές. Η πλειοψηφία του δείγματος αποτελείται από παντρεμένους σε πλήθος 2.509, ακολουθούν οι ανύπαντροι με πλήθος 1.153, οι διαζευγμένοι και οι χήροι με πλήθος 446 (κατηγορία divorced) και υπάρχουν 11 άγνωστες τιμές.

Κατάσταση γάμου	Πλήθος
married	2.509
single	1.153
divorced	446
unknown	11
Σύνολο	4.119

Πίνακας 4.4: Πίνακας συχνοτήτων της μεταβλητής «marital»



Σχήμα 4.4: Marital Count Distribution (barplot)

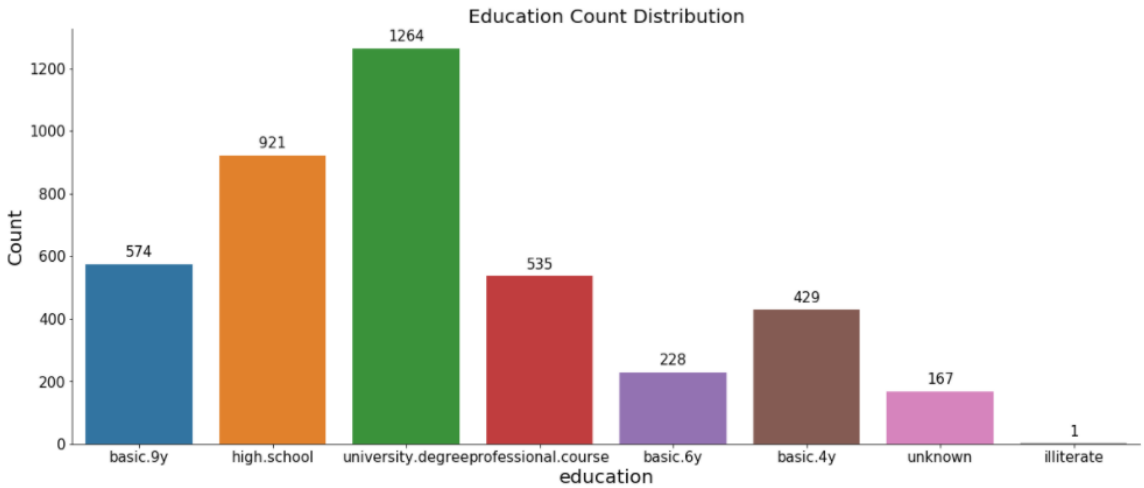
Εκπαίδευση

Η πλειοψηφία των ατόμων του δείγματος αποτελείται από αποφοίτους πανεπιστημίου σε πλήθος 1.264, ακολουθούν οι απόφοιτοι Λυκείου.

Εκπαίδευση	Πλήθος
University degree	1.264
High school	921
Basic 9y	574
Professional cource	535

Basic 4y	429
Basic 6y	228
unknown	167
illiterate	1
Σύνολο	4.119

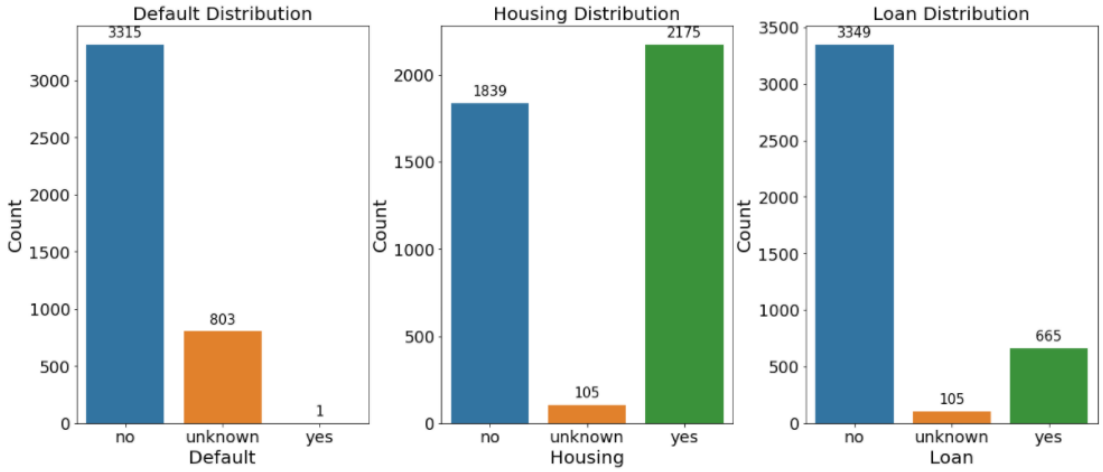
Πίνακας 4.5: Πίνακας συχνοτήτων της μεταβλητής «education»



Σχήμα 4.5: Education Count Distribution (barplot)

Default, Housing, Loan

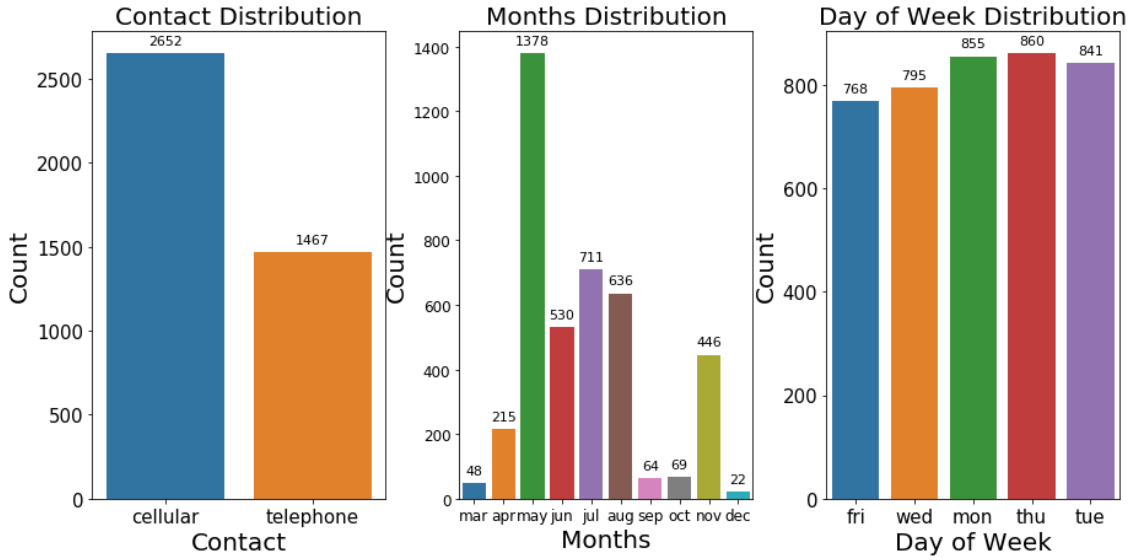
Όσον αφορά τις μεταβλητές «default», «housing», «loan», παρατίθενται τα παρακάτω ραβδογράμματα. Οι μεταβλητές αυτές αναφέρονται στην ύπαρξη ή όχι αθέτησης, στεγαστικού δανείου και καταναλωτικού δανείου από τη μεριά του πελάτη. Από τα ραβδογράμματα συμπεραίνεται ότι οι περισσότεροι πελάτες δεν έχουν κάποια αθέτηση, έχουν στεγαστικό δάνειο αλλά όχι προσωπικό δάνειο.



Σχήμα 4.6: Default, Housing, Loan Count Distribution (barplots)

Contact, Month, Day

Τα παρακάτω ραβδογράμματα αφορούν το τρόπο προσέγγισης του πελάτη από την τράπεζα, το μήνα και την ημέρα της εβδομάδας που έγινε η προσέγγιση. Είναι προφανές ότι οι περισσότεροι πελάτες έχουν προσεγγιστεί από την τράπεζα με χρήση κινητού τηλεφώνου το μήνα Μάϊο ενώ δεν φαίνεται να υπάρχει προτίμηση όσον αφορά την ημέρα προσέγγισης των πελατών.



Σχήμα 4.7: Contact, Months, Day of Week Count Distribution (barplots)

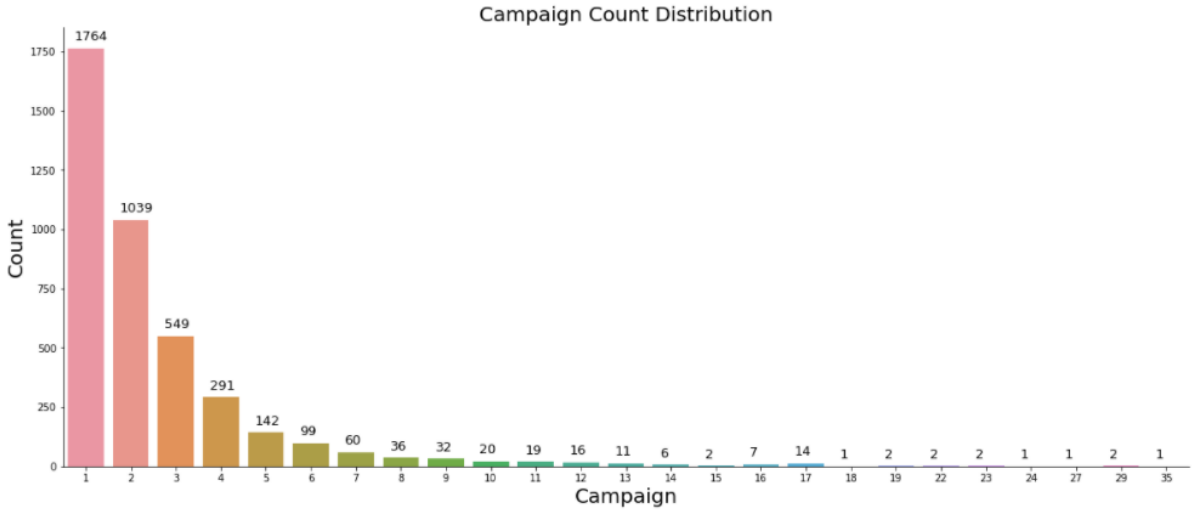
Campaign

Πρόκειται για μία ποσοτική μεταβλητή που εκφράζει το πλήθος των επαφών που έγιναν στον πελάτη κατά τη διάρκεια αυτής της καμπάνιας. Από τον πίνακα με τα περιγραφικά στοιχεία που ακολουθεί, προκύπτει η διάμεσος να είναι 2 το οποίο σημαίνει ότι το 50% των ατόμων του δείγματος έχει προσεγγιστεί με το πολύ 2 επαφές.

Περιγραφικά μέτρα	**Τιμή**
Πλήθος	4.119
Μέσος όρος	2,53
Τυπική απόκλιση	2,56
Ελάχιστο στοιχείο	1
1° τεταρτημόριο	1
2° τεταρτημόριο	2
3° τεταρτημόριο	3
Μέγιστο στοιχείο	35

Πίνακας 4.6: Περιγραφικά στοιχεία για τη μεταβλητή «campaign»

Βάση του ιστογράμματος, οι τιμές της μεταβλητής αυτής είναι εμφανές ότι ακολουθούν δεξιά ασυμμετρία. Από το σχήμα συμπεραίνεται ακόμη ότι οι περισσότεροι από τους πελάτες προσεγγίστηκαν 1, 2 ή 3 φορές κατά τη διάρκεια της καμπάνιας, γεγονός που επιβεβαιώνεται και από το 3° τεταρτημόριο του πίνακα που αφορά την κατανομή του 75% των παρατηρήσεων.



Σχήμα 4.8: Campaign Count Distribution (barplot)

Pdays

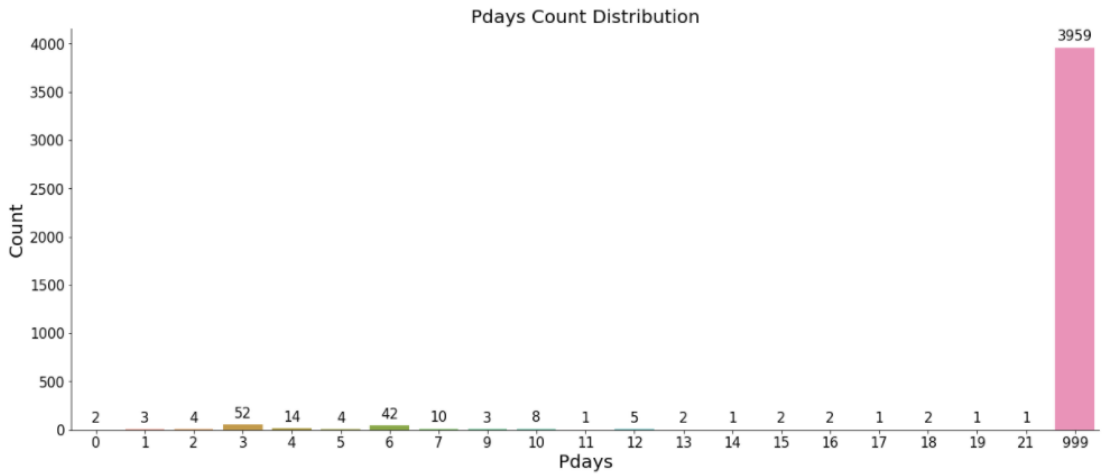
Πρόκειται για τον αριθμό των ημερών που πέρασαν από την προσέγγιση του πελάτη από προηγούμενη καμπάνια. Κατά συνέπεια, υπάρχουν 2 άτομα που προσεγγίστηκαν πριν 0 μέρες, 3 άτομα που προσεγγίστηκαν πριν 1 ημέρα, 4 άτομα που προσεγγίστηκαν πριν 2 μέρες κλπ.

Pdays	**Πλήθος**
0	2
1	3
2	4
3	52
4	14
5	4
6	42
7	10
9	3
10	8

11	1
12	5
13	2
14	1
15	2
16	2
17	1
18	2
19	1
21	1
999	3.959
Σύνολο	4.119

Πίνακας 4.7: Πίνακας συχνοτήτων της μεταβλητής «pdays»

Από τον πίνακα συχνοτήτων και το παρακάτω ραβδόγραμμα, είναι εμφανές ότι η επικρατούσα τιμή για την μεταβλητή αυτή είναι το 999 που αφορά τα άτομα που δεν έχουν προσεγγιστεί ξανά.



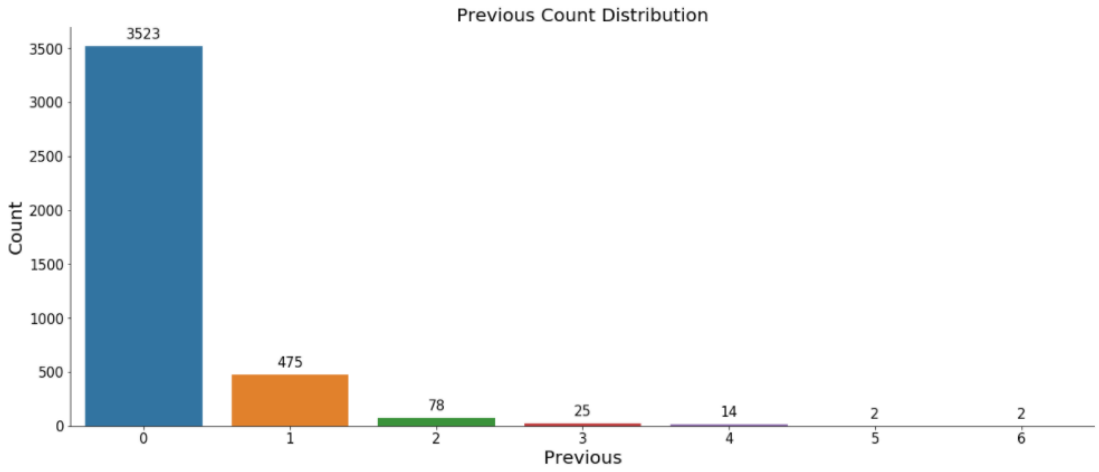
Σχήμα 4.9: Pdays Count Distribution (barplot)

Previous

Η μεταβλητή αυτή αφορά το πλήθος των φορών που ο κάθε πελάτης προσεγγίστηκε πριν από αυτήν την καμπάνια. Όπως είναι εμφανές 3.523 άτομα δεν έχουν προσεγγιστεί ξανά από προηγούμενη καμπάνια, 475 άτομα έχουν προσεγγιστεί 1 φορά από προηγούμενη καμπάνια, 78 άτομα έχουν προσεγγιστεί 2 φορές κτλ. Συνοψίζοντας, από τα 4119 άτομα τα 3.523 δεν είχαν προσεγγιστεί ξανά ενώ τα υπόλοιπα 4119-3523=596 είχαν προσεγγιστεί τουλάχιστον μία φορά από προηγούμενη καμπάνια.

Previous	**Πλήθος**
0	3.523
1	475
2	78
3	25
4	14
5	2
6	2
Σύνολο	4.119

Πίνακας 4.8: Πίνακας συχνοτήτων της μεταβλητής «previous»



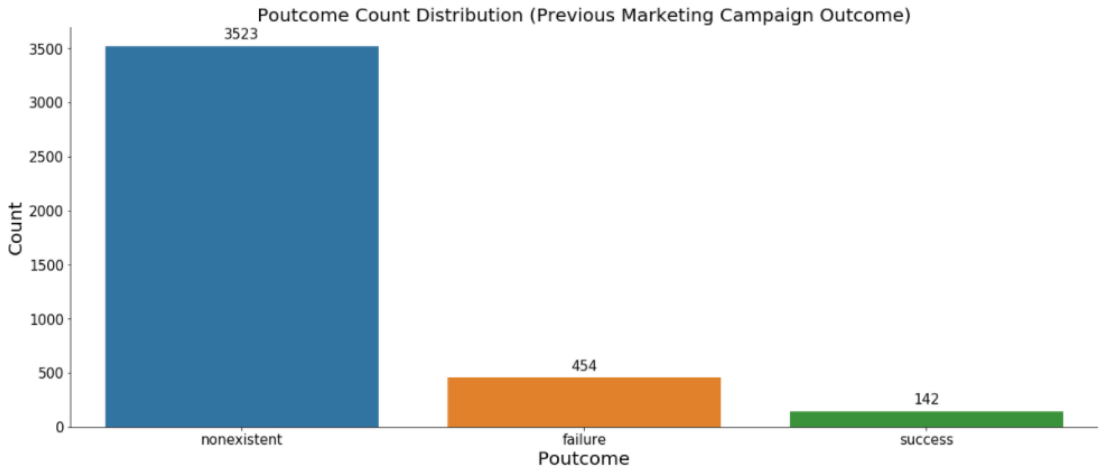
Σχήμα 4.10: Previous Count Distribution (barplot)

Poutcome

Πρόκειται για μια κατηγορική μεταβλητή που αφορά το αποτέλεσμα που καταγράφηκε από την προηγούμενη καμπάνια προσέγγισης των πελατών. Από τα 4.119 άτομα, τα 3.523 δεν έχουν προσεγγιστεί ξανά και γι αυτό το αποτέλεσμά τους έχει καταγραφεί ως 'nonexistent' ενώ 454 άτομα είχαν απαντήσει αρνητικά στην προηγούμενη καμπάνια και μόλις 142 είχαν απαντήσει θετικά.

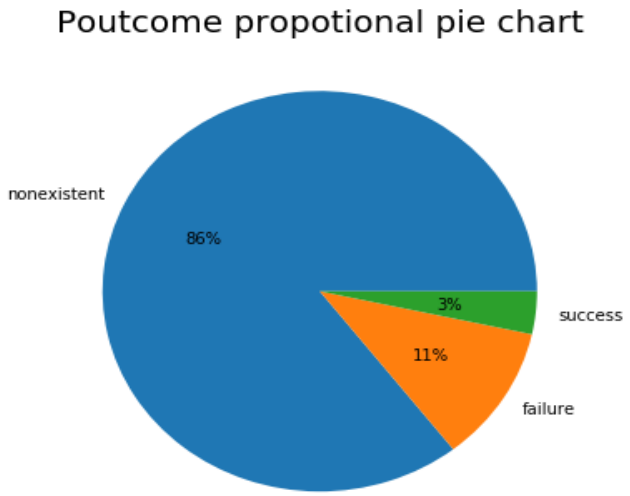
Poutcome	**Πλήθος**
nonexistent	3.523
failure	454
success	142
Σύνολο	4.119

Πίνακας 4.9: Πίνακας συχνοτήτων της μεταβλητής «poutcome»



Σχήμα 4.11: Poutcome Count Distribution (barplot)

Το κυκλικό διάγραμμα με τα ποσοστά δίνεται παρακάτω και παρατηρείται ότι το 86% των ατόμων του δείγματος δεν έχει προσεγγιστεί ξανά.



Σχήμα 4.12: Poutcome (propotional pie chart)

Πίνακας συνάφειας ανάμεσα στις μεταβλητές «previous» και «poutcome»

Poutcome /previous	failure	nonexistent	success	Σύνολο
0	0	3.523	0	3.523

1	384	0	91	475
2	53	0	25	78
3	9	0	16	25
4	8	0	6	14
5	0	0	2	2
6	0	0	2	2
Σύνολο	454	3.523	142	4.119

Πίνακας 4.10: Πίνακας συνάφειας ανάμεσα στις μεταβλητές «previous» και «poutcome»

Όπως φαίνεται από τον παραπάνω πίνακα, από τα 3.523 άτομα που δεν έχουν προσεγγιστεί ξανά όλα έχουν ως Poutcome την τιμή «nonexistent», από τα 475 άτομα που έχουν προσεγγιστεί 1 φορά, τα 384 δεν είχαν προβεί σε προθεσμιακή κατάθεση και τα υπόλοιπα 91 είχαν προβεί.

Πίνακας συνάφειας ανάμεσα στις μεταβλητές «pdays» και «poutcome»

Poutcome /pdays	**failure**	**nonexistent**	**success**	**Σύνολο**
0	0	0	2	2
1	0	0	3	3
2	0	0	4	4
3	2	0	50	52
4	0	0	14	14
5	0	0	4	4
6	3	0	39	42

7	1	0	9	10
9	2	0	1	3
10	2	0	6	8
11	0	0	1	1
12	1	0	4	5
13	1	0	1	2
14	1	0	0	1
15	1	0	1	2
16	0	0	2	2
17	1	0	0	1
18	2	0	0	2
19	0	0	1	1
21	1	0	0	1
999	436	3.523	0	3.959
Σύνολο	454	3.523	142	4.119

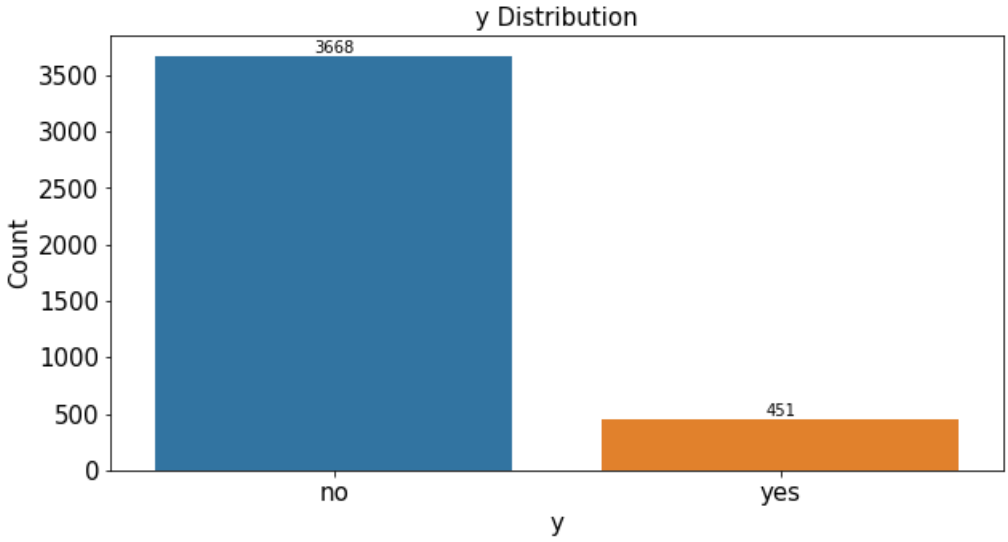
Πίνακας 4.11: Πίνακας συνάφειας ανάμεσα στις μεταβλητές «pdays» και «poutcome»

Η τιμή 999 στη μεταβλητή «pdays» αναπαριστά την περίπτωση που το άτομο δεν έχει προσεγγιστεί ξανά. Όμως αν και η τιμή αυτή υπάρχει 3.959 φορές, στις 3.523 από αυτές αντιστοιχεί η τιμή 'nonexistent' για τη μεταβλητή «poutcome» και στις υπόλοιπες 436 εμφανίζεται η τιμή 'failure'. Πιθανώς να προκύπτει από κάποιο λάθος κατά την καταγραφή των αποτελεσμάτων.

Y (Προθεσμιακή κατάθεση)

Πρόκειται για μία δίτιμη κατηγορική μεταβλητή που παίρνει τις τιμές 'yes' και 'no', ανάλογα αν το άτομο προχωράει σε προθεσμιακή κατάθεση ή όχι και αποτελεί την εξαρτημένη μεταβλητή στο πρόβλημα ταξινόμησης. Το ραβδόγραμμα δείχνει ότι από

τα 4.119 άτομα του δείγματος, τα 451 προχώρησαν σε προθεσμιακή κατάθεση ενώ τα 3.668 δεν προχώρησαν.



Σχήμα 4.13: y Distribution (barplot)

Το ποσοστό επιτυχίας της τωρινής καμπάνιας είναι 451/4.119=0,11=11%.

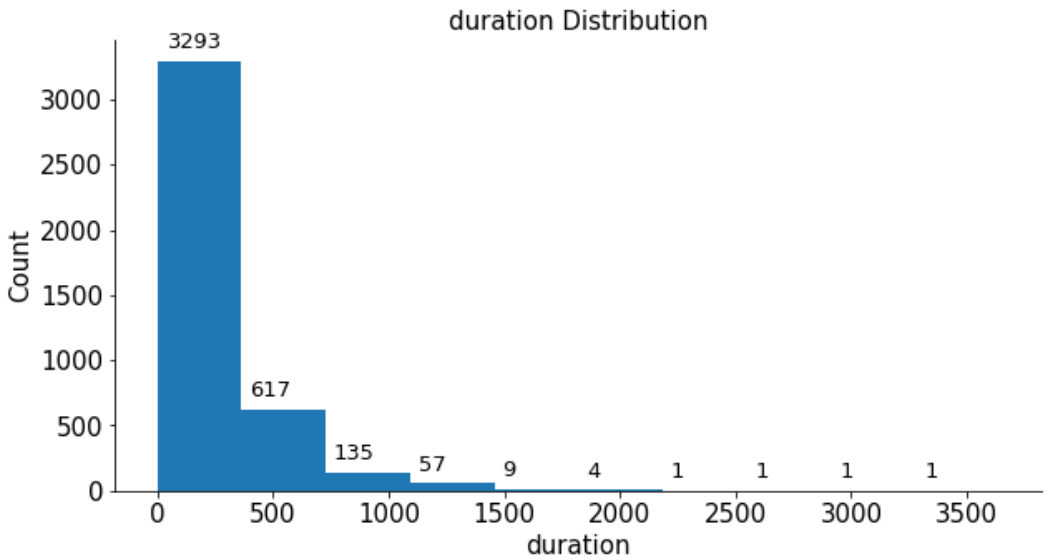
Διάρκεια

Η μεταβλητή «duration» είναι μία ποσοτική μεταβλητή που εκφράζει τη διάρκεια (σε δευτερόλεπτα) της κλήσης ανάμεσα στον πελάτη και την τράπεζα κατά την προσέγγισή του. Είναι συνεχής μεταβλητή και φαίνεται να παίρνει πολλές τιμές. Γι αυτόν τον λόγο γίνεται ομαδοποίηση των τιμών σε 10 κλάσεις. Από το ιστόγραμμα που ακολουθεί φαίνεται να υπάρχει δεξιά ασυμμετρία, με τις περισσότερες τιμές της να ανήκουν στην 1^{η} κλάση. Στον πίνακα που ακολουθεί παρουσιάζονται κάποια περιγραφικά στοιχεία της μεταβλητής αυτής.

Περιγραφικά μέτρα	**Τιμή**
Πλήθος	4.119
Μέσος όρος	256,78
Τυπική απόκλιση	254,70
Ελάχιστο στοιχείο	0

1° τεταρτημόριο	103
2° τεταρτημόριο	181
3° τεταρτημόριο	317
Μέγιστο στοιχείο	3.643

Πίνακας 4.12: Περιγραφικά στοιχεία για τη μεταβλητή «duration»



Σχήμα 4.14: duration Distribution (histogram)

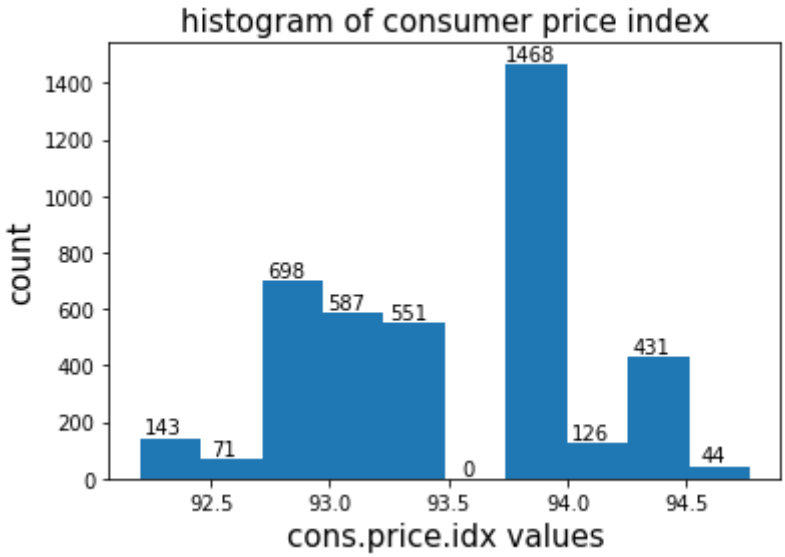
Consumer Price Index

Είναι ένας δείκτης της μέσης αλλαγής των τιμών που πληρώνουν οι καταναλωτές για μια ποσότητα προϊόντων και υπηρεσιών (διατροφή, κατοικία, μεταφορές, ιατρική περίθαλψη, εκπαίδευση, ψυχαγωγία). Για την ποσοτική αυτή μεταβλητή δίνεται ο πίνακας με τα σημαντικότερα περιγραφικά στοιχεία καθώς και το ιστόγραμμα της.

Περιγραφικά μέτρα	**Τιμή**
Πλήθος	4.119
Μέσος όρος	93,57
Τυπική απόκλιση	0,57

Ελάχιστο στοιχείο	92,2
1ο τεταρτημόριο	93,07
2ο τεταρτημόριο	93,75
3ο τεταρτημόριο	93,99
Μέγιστο στοιχείο	94,76

Πίνακας 4.13: Περιγραφικά στοιχεία για τη μεταβλητή «consumer price index»



Σχήμα 4.15: Consumer Price Index Distribution (histogram)

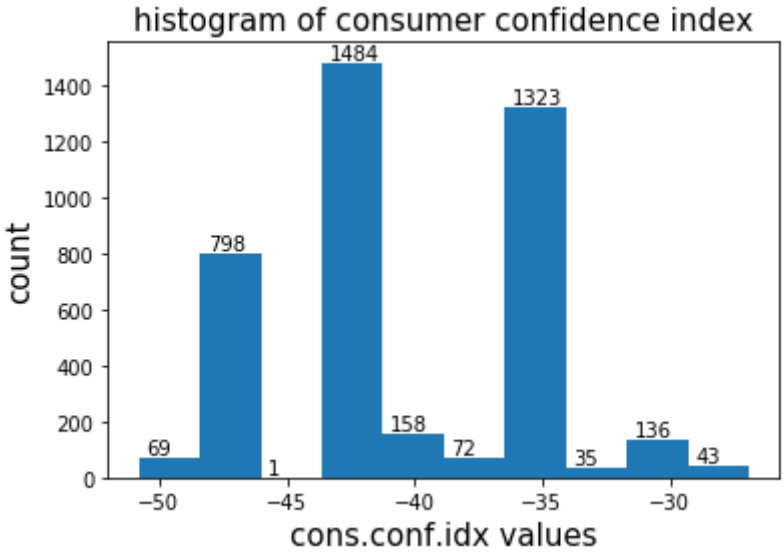
Consumer Confidence Index

Είναι ένας δείκτης της συμπεριφοράς των ανθρώπων σχετικά με τις τωρινές και μελλοντικές οικονομικές συνθήκες. Δείχνει πόσο αισιόδοξοι ή απαισιόδοξοι είναι οι καταναλωτές σε σχέση με την οικονομία και την ικανότητα να βρουν δουλειά.

Περιγραφικά μέτρα	**Τιμή**
Πλήθος	4.119
Μέσος όρος	-40,5
Τυπική απόκλιση	4,6
Ελάχιστο στοιχείο	-50,8

1° τεταρτημόριο	-42,7
2° τεταρτημόριο	-41,8
3° τεταρτημόριο	-36,4
Μέγιστο στοιχείο	-26,9

Πίνακας 4.14: Περιγραφικά στοιχεία για τη μεταβλητή «consumer confidence index»



Σχήμα 4.16: Consumer Confidence Index Distribution (histogram)

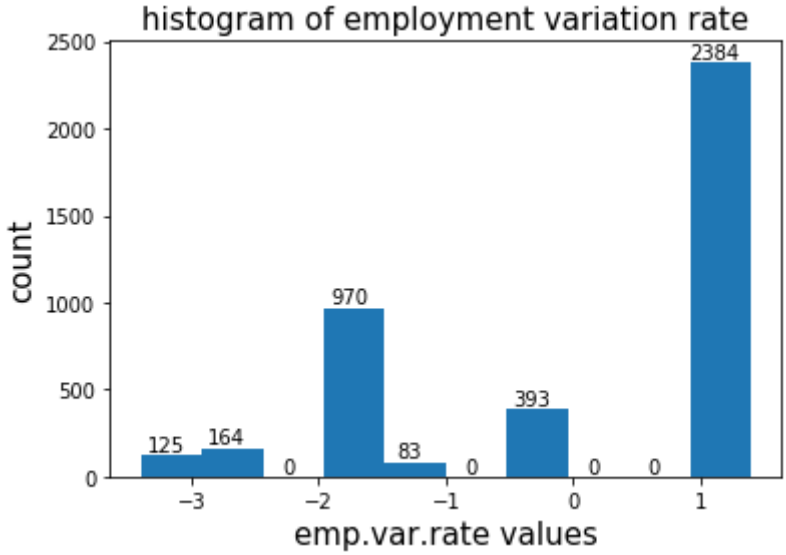
Employment Variation Rate

Ο δείκτης αυτός δείχνει την μεταβολή του πλήθους των ατόμων που προσλαμβάνονται/απολύονται λόγω των αλλαγών της οικονομίας.

Περιγραφικά μέτρα	**Τιμή**
Πλήθος	4.119
Μέσος όρος	0,08
Τυπική απόκλιση	1,56
Ελάχιστο στοιχείο	-3,4
1° τεταρτημόριο	-1,8
2° τεταρτημόριο	1,1

3° τεταρτημόριο	1,4
Μέγιστο στοιχείο	1,4

Πίνακας 4.15: Περιγραφικά στοιχεία για τη μεταβλητή «employment variation rate»



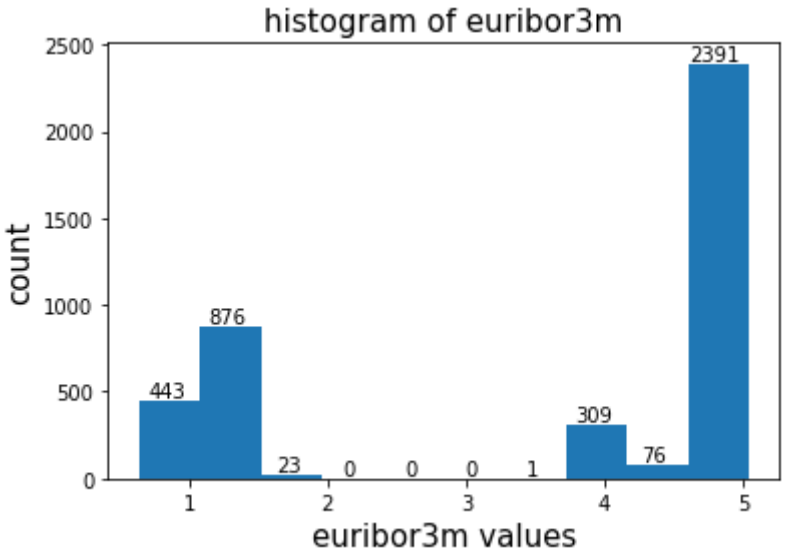
Σχήμα 4.17: Employment Variation Rate Distribution (histogram)

Euribor3m

Είναι το μέσο interest rate με το οποίο οι Ευρωπαϊκές τράπεζες δανείζουν η μία την άλλη.

Περιγραφικά μέτρα	**Τιμή**
Πλήθος	4.119
Μέσος όρος	3,62
Τυπική απόκλιση	1,73
Ελάχιστο στοιχείο	0,63
1° τεταρτημόριο	1,33
2° τεταρτημόριο	4,85
3° τεταρτημόριο	4,96
Μέγιστο στοιχείο	5,04

Πίνακας 4.16: Περιγραφικά στοιχεία για τη μεταβλητή «euribor3m»



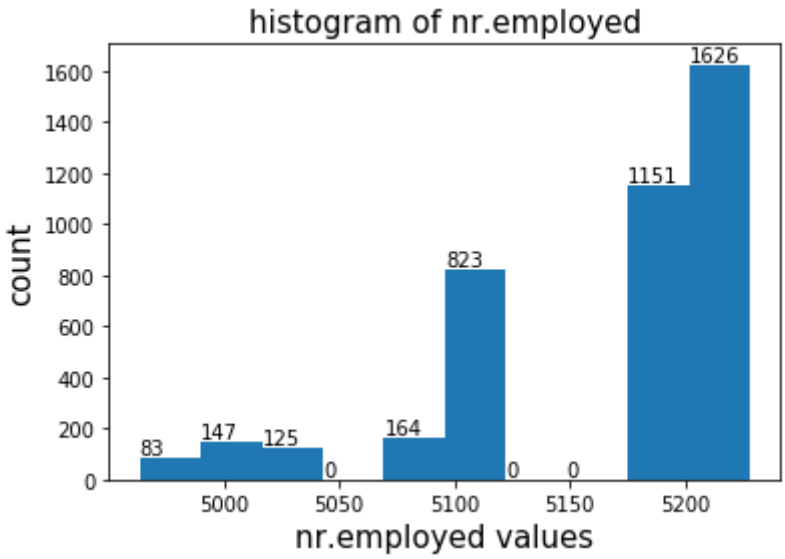
Σχήμα 4.18: Euribor3m Distribution (histogram)

Number of employees

Πρόκειται για μία ποσοτική μεταβλητή που εκφράζει τον μέσο αριθμό των εργαζόμενων ατόμων.

Περιγραφικά μέτρα	**Τιμή**
Πλήθος	4.119
Μέσος όρος	5.166,48
Τυπική απόκλιση	73,66
Ελάχιστο στοιχείο	4.963,6
1^{ο} τεταρτημόριο	5.099,1
2^{ο} τεταρτημόριο	5.191
3^{ο} τεταρτημόριο	5.228,1
Μέγιστο στοιχείο	5.228,1

Πίνακας 4.17: Περιγραφικά στοιχεία για τη μεταβλητή «Nr.employed»



Σχήμα 4.19: Nr.Employed Distribution (histogram)

4.4 Σύγκριση

Σε αυτό το σημείο έχει ενδιαφέρον να γίνει σύγκριση των ατόμων που προχωρούν σε προθεσμιακή κατάθεση και αυτών που δεν προχωρούν. Όπως είχε αναφερθεί προηγουμένως από τα 4.119 άτομα, μόλις τα 451 έχουν απαντήσει θετικά στην καμπάνια και έχουν προχωρήσει σε προθεσμιακή κατάθεση ενώ οι υπόλοιποι 3.668 έχουν απαντήσει αρνητικά στην καμπάνια.

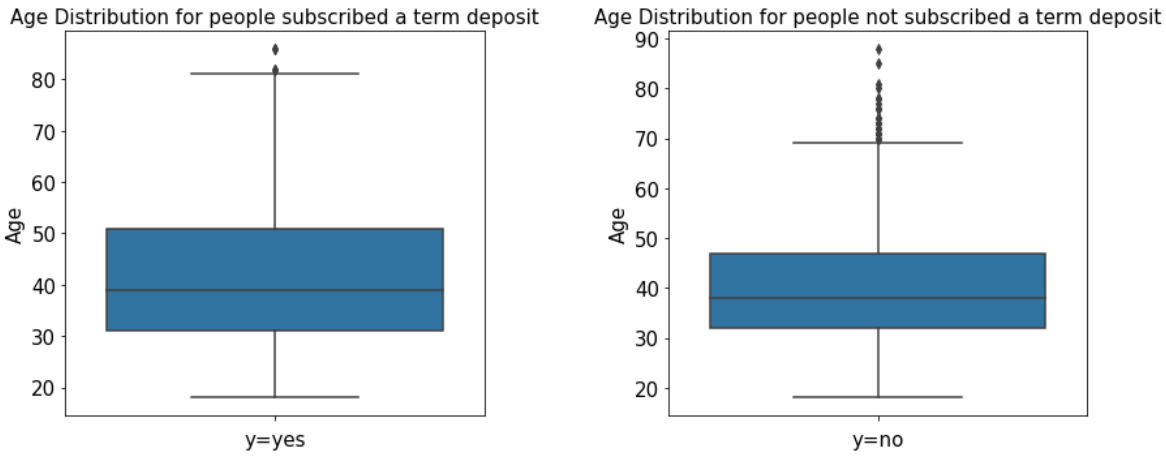
Ο πίνακας που ακολουθεί περιέχει τις μέσες τιμές όλων των ποσοτικών μεταβλητών για τις 2 κατηγορίες ατόμων και βοηθάει σε μία πρώτη σύγκριση. Στην πορεία, γίνεται πιο αναλυτικά η σύγκριση των 2 ομάδων για κάθε μεταβλητή.

y	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
no	39.895311	219.40976	2.605780	982.763086	0.141767	0.240185	93.599677	-40.586723	3.802826	5175.502072
yes	41.889135	560.78714	1.980044	778.722838	0.585366	-1.177384	93.417268	-39.786475	2.145448	5093.118625

Σχήμα 4.20: Μέσες τιμές των ποσοτικών μεταβλητών για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y

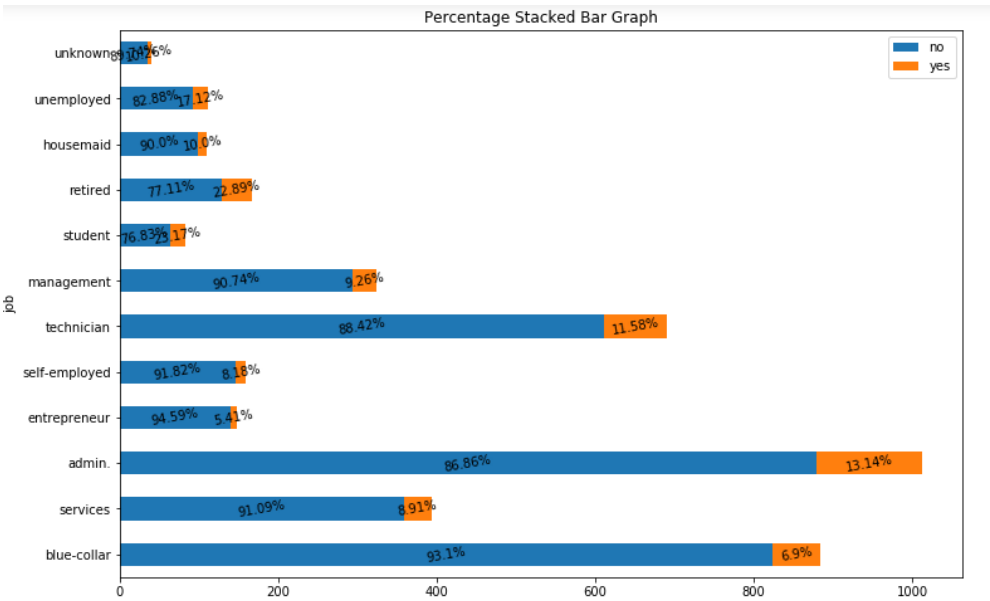
Αρχικά γίνεται σύγκριση των ατόμων των 2 δειγμάτων όσον αφορά τις προσωπικές πληροφορίες που έχουν συγκεντρωθεί με τη χρήση των μεταβλητών «age», «job»,

«marital» και «education». Συγκρίνοντας τις τιμές της ηλικίας για τα άτομα που απαντούν θετικά ή αρνητικά, δεν παρατηρείται διαφορά.



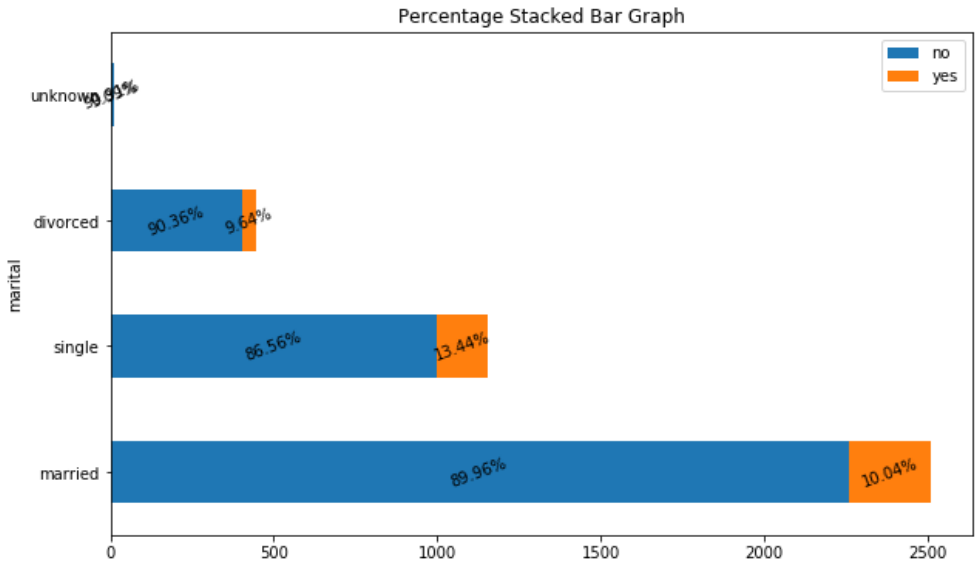
Σχήμα 4.21: Κατανομή τιμών της μεταβλητής age για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y

Όσον αφορά τη δουλειά των ατόμων που απαρτίζουν τα 2 δείγματα, παρατηρείται ότι οι συνταξιούχοι, τα άτομα που απασχολούνται σε θέσεις διοίκησης αλλά και οι τεχνίτες έχουν απαντήσει θετικά σε μεγαλύτερο ποσοστό στην καμπάνια (22,89%, 13,14% και 11,58% αντίστοιχα).



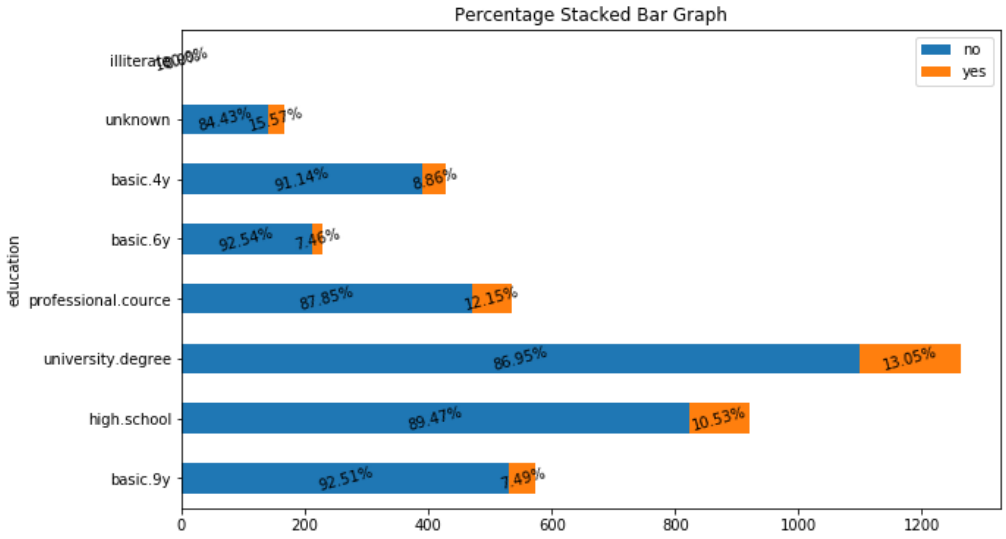
Σχήμα 4.22: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής job

Όσον αφορά τη μεταβλητή «marital», οι ανύπαντροι σε σχέση με τους παντρεμένους, τους διαζευγμένους και τους χήρους είναι περισσότερο διατεθειμένοι να καταθέσουν.



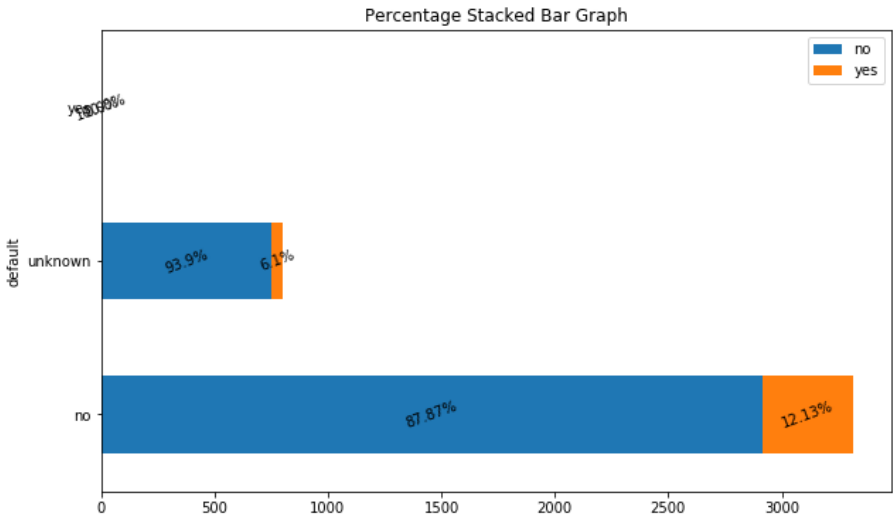
Σχήμα 4.23: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής marital

Άτομα με καλό υπόβαθρο εκπαίδευσης όπως απόφοιτοι Πανεπιστημίου ή με μαθήματα εξειδίκευσης είναι επίσης πιθανότερο να προβούν σε προθεσμιακή κατάθεση.

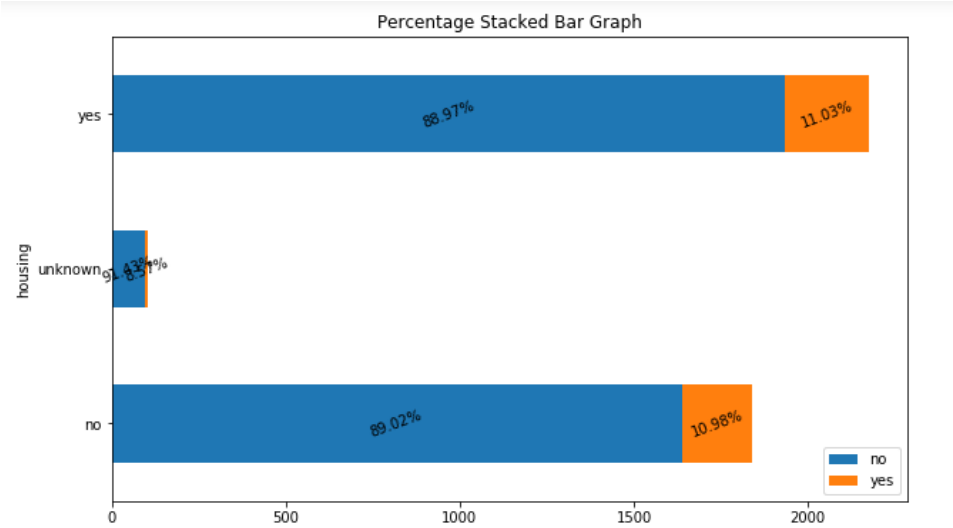


Σχήμα 4.24: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής education

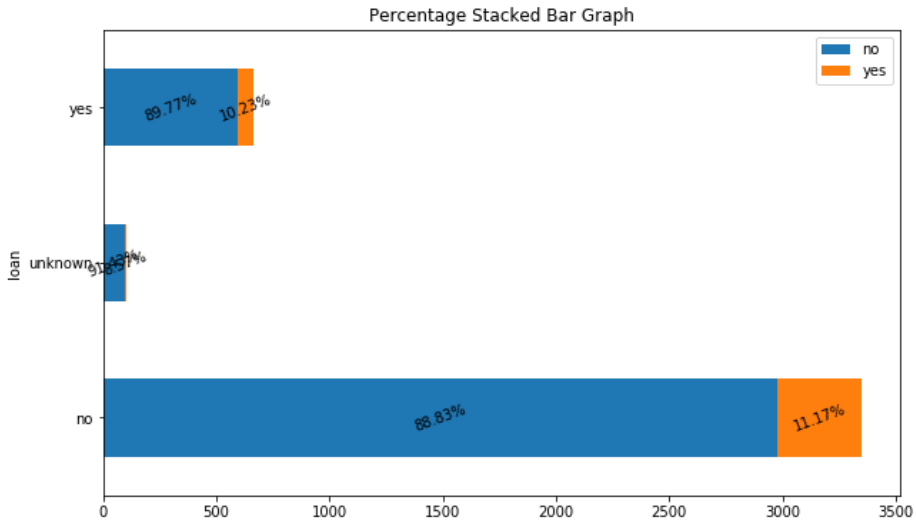
Όσον αφορά τις τραπεζικές πληροφορίες που έχουν συγκεντρωθεί για τον κάθε πελάτη μέσω των μεταβλητών «default», «housing», «loan» δίνονται τα παρακάτω συσσωρευμένα ραβδογράμματα. Από αυτά παρατηρείται ότι τα άτομα τα οποία δεν έχουν κάποια αθέτηση προχωρούν σε μεγαλύτερο ποσοστό σε προθεσμιακή κατάθεση ενώ δεν φαίνεται να υπάρχει κάποια διαφορά ανάμεσα σε αυτούς που προχωρούν σε προθεσμιακή κατάθεση και σε αυτούς που δεν προχωρούν ανάλογα για τις μεταβλητές «housing» και «loan».



Σχήμα 4.25: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής default

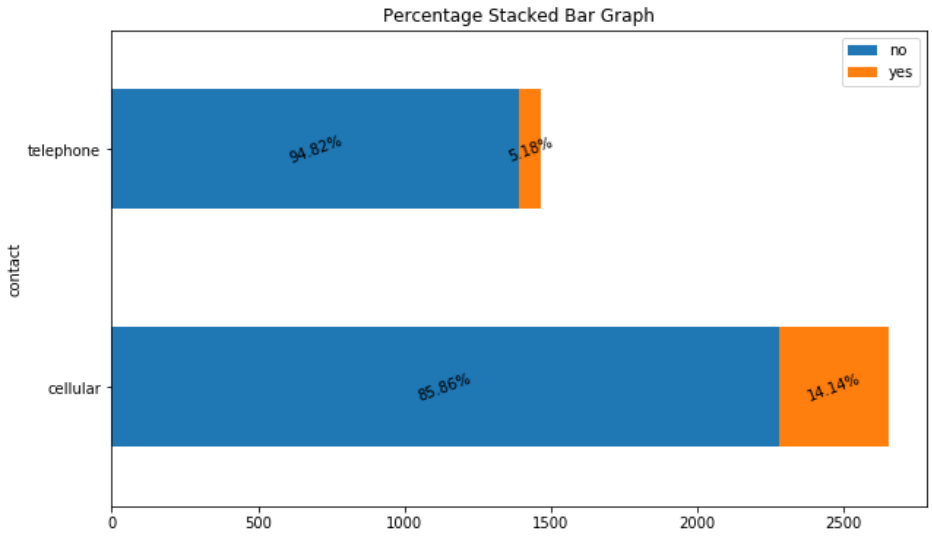


Σχήμα 4.26: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής housing



Σχήμα 4.27: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής loan

Τα ακόλουθα σχήματα δείχνουν την σύγκριση των 2 δειγμάτων σε σχέση με τις πληροφορίες επαφής που συγκεντρώθηκαν μέσω των μεταβλητών «contact», «months», «day of week», «campaign» και «duration». Φαίνεται να υπάρχει διαφορά αυτών που προχωρούν σε προθεσμιακή βάσει του μέσου επαφής και να απαντούν θετικά στην καμπάνια άτομα που προσεγγίζονται μέσω κινητού τηλεφώνου.

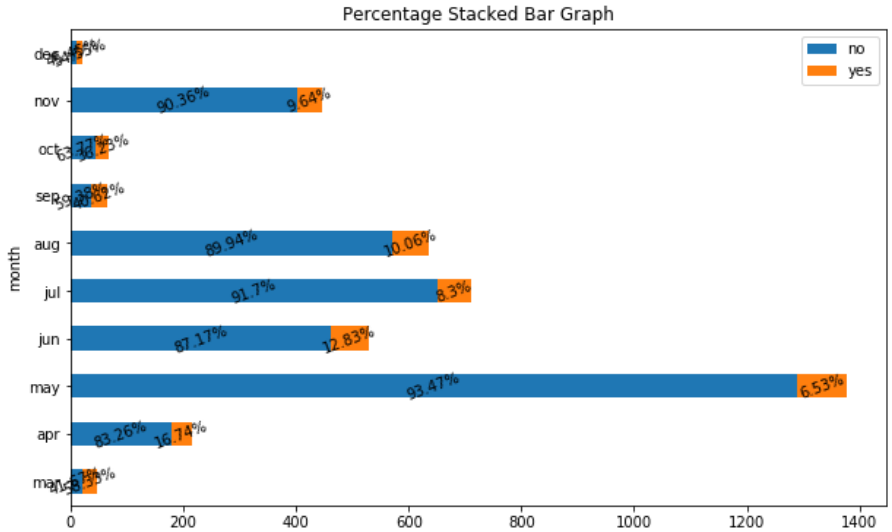


Σχήμα 4.28: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής contact

Όσον αφορά την επιτυχία της καμπάνιας σε σχέση με τον μήνα που έχει πραγματοποιηθεί δίνεται ο πίνακας με τα ποσοστά για κάθε μήνα και το συσσωρευμένο ραβδόγραμμα. Από αυτά τα στοιχεία προκύπτει ότι τους μήνες Μάρτιο, Απρίλιο, Ιούνιο, Σεπτέμβριο, Οκτώβριο και Δεκέμβριο τα άτομα απαντούν περισσότερο θετικά στην καμπάνια μάρκετινγκ της τράπεζας.

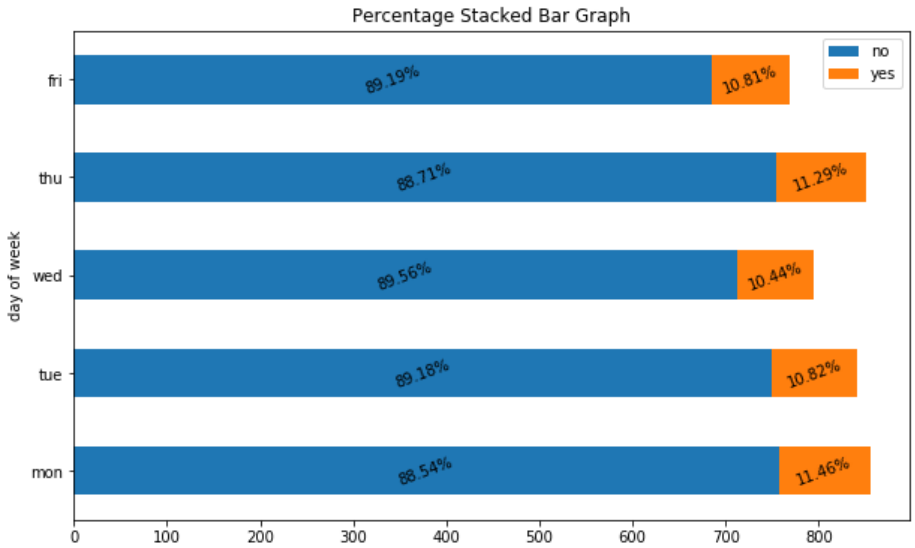
	no	yes
mar	41.666667	58.333333
apr	83.255814	16.744186
may	93.468795	6.531205
jun	87.169811	12.830189
jul	91.701828	8.298172
aug	89.937107	10.062893
sep	59.375000	40.625000
oct	63.768116	36.231884
nov	90.358744	9.641256
dec	45.454545	54.545455

Πίνακας 4.18: Πίνακας ποσοστών θετικής και αρνητικής ανταπόκρισης για τις τιμές της μεταβλητής «month»



Σχήμα 4.29: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής month

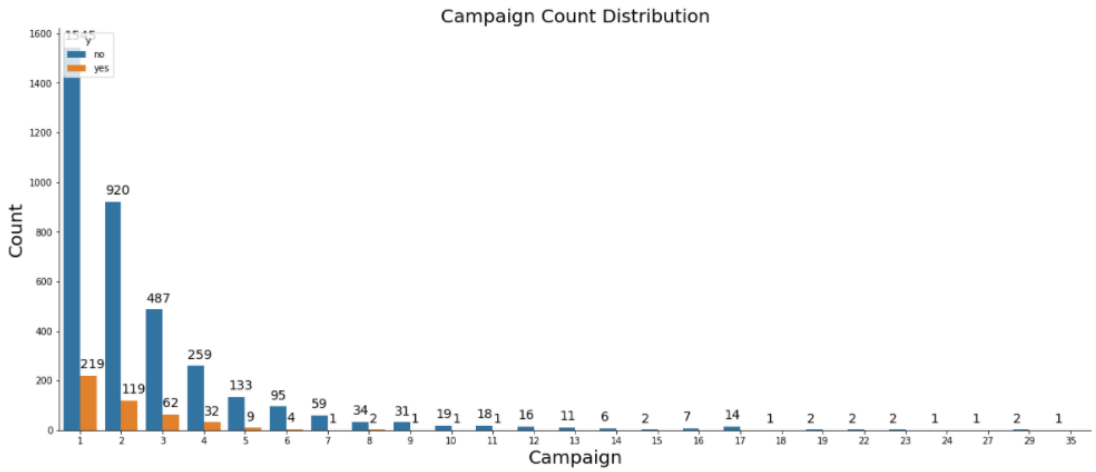
Όσον αφορά την ύπαρξη ή όχι διαφοράς ανάμεσα στα άτομα που απαντούν θετικά και αρνητικά στην καμπάνια ανάλογα με τη μέρα που αυτή πραγματοποιείται, το σχήμα που ακολουθεί δείχνει ότι δεν υπάρχει διαφορά.



Σχήμα 4.30: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής day of week

Στην προσπάθεια να εξετάσουμε αν υπάρχει διαφορά στα 2 δείγματα σε σχέση με τη μεταβλητή «campaign» δίνουμε το παρακάτω ραβδόγραμμα από το οποίο είναι εμφανές ότι όταν το πλήθος των φορών που προσεγγίζεται ένα άτομο είναι

μεγαλύτερο από 6 τότε το πλήθος αυτών που ανταποκρίνονται θετικά στην καμπάνια είναι μικρό.

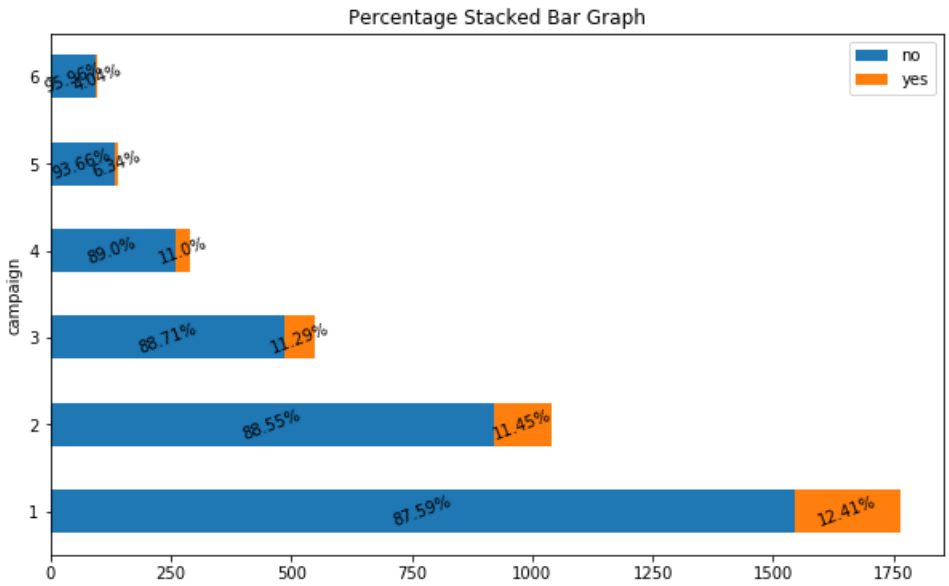


Σχήμα 4.31: Κατανομή τιμών των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής campaign

Ο πίνακας με τα ποσοστά ανταπόκρισης των ατόμων για τις πρώτες 6 τιμές της μεταβλητής «campaign» παρουσιάζονται στον επόμενο πίνακα αλλά και στο stacked bar που ακολουθεί. Είναι εμφανές ότι τα άτομα που προσεγγίζονται έως 4 φορές είναι πιο πιθανό να ανταποκριθούν θετικά.

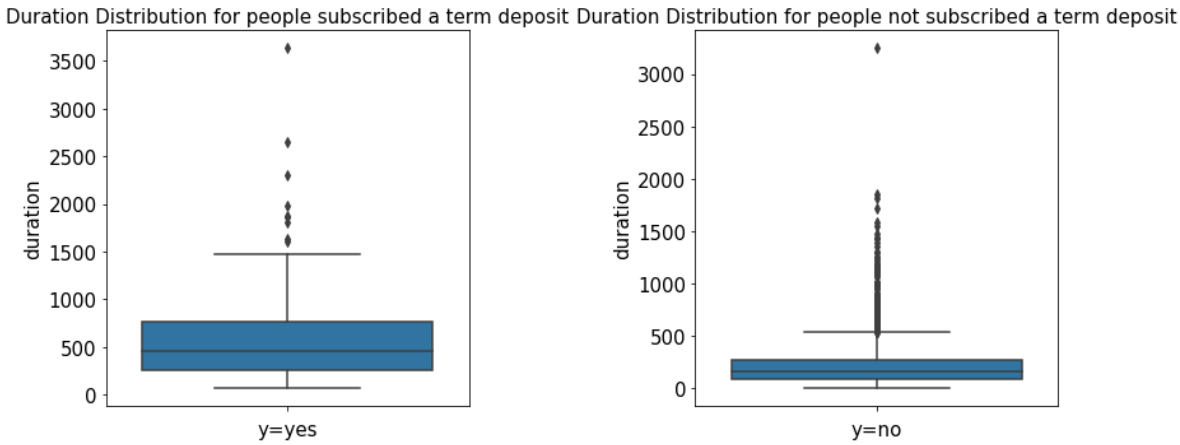
	no	yes
1	87.585034	12.414966
2	88.546679	11.453321
3	88.706740	11.293260
4	89.003436	10.996564
5	93.661972	6.338028
6	95.959596	4.040404

Πίνακας 4.19: Πίνακας ποσοστών θετικής και αρνητικής ανταπόκρισης για τις 6 πρώτες τιμές της μεταβλητής «campaign»



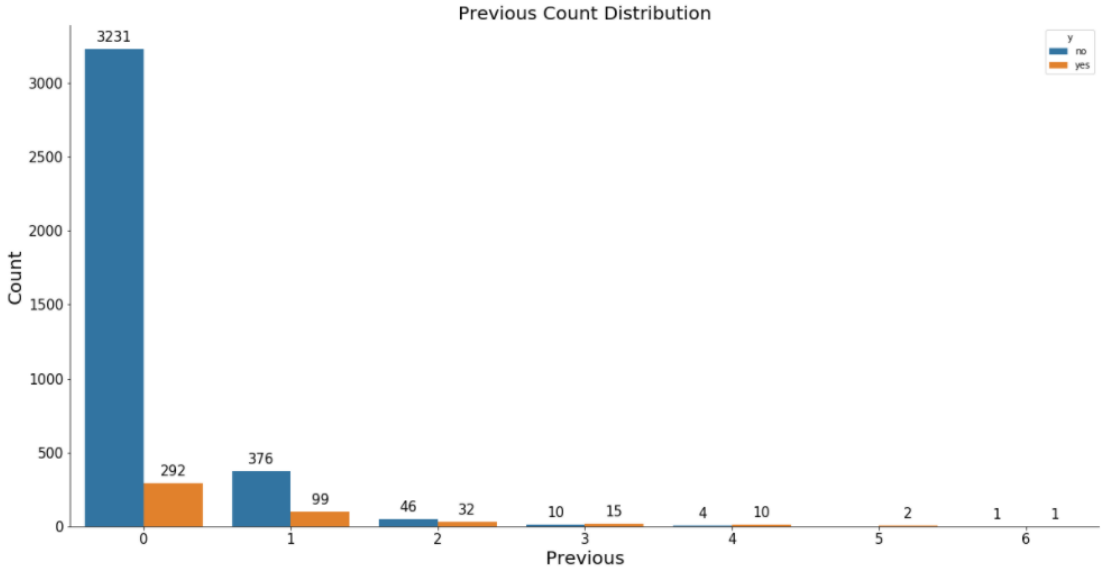
Σχήμα 4.32: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής campaign

Σημαντική είναι η σύγκριση της διάρκειας της κλήσης για τα άτομα που τελικά προχώρησαν σε προθεσμιακή κατάθεση (y='yes') και για τα άτομα που δεν προχώρησαν (y='no'). Από τα θηκογράμματα που ακολουθούν παρατηρούμε ότι η διάρκεια της κλήσης φαίνεται να είναι μικρότερη στα άτομα που δεν κάνουν προθεσμιακή κατάθεση και μεγαλύτερη σε όσους κάνουν.

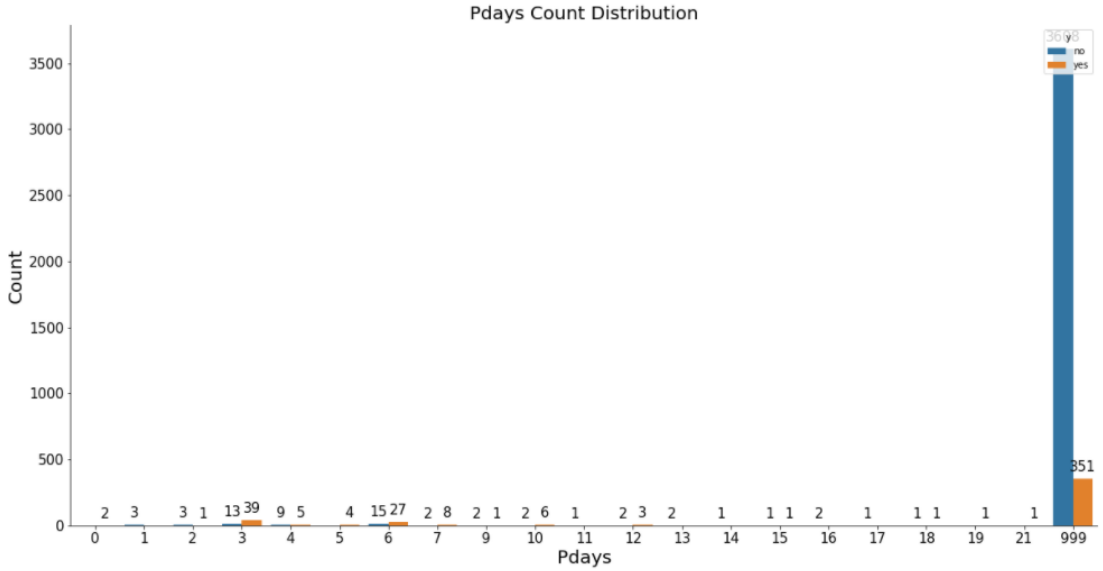


Σχήμα 4.33: Κατανομή τιμών της μεταβλητής duration για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y

Στη συνέχεια γίνεται σύγκριση των 2 δειγμάτων όσον αφορά τις μεταβλητές που περιέχουν πληροφορίες σχετικά με το ιστορικό προσέγγισης των πελατών όπως η «previous», «pdays» και η «poutcome».



Σχήμα 4.34: Κατανομή τιμών των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής previous



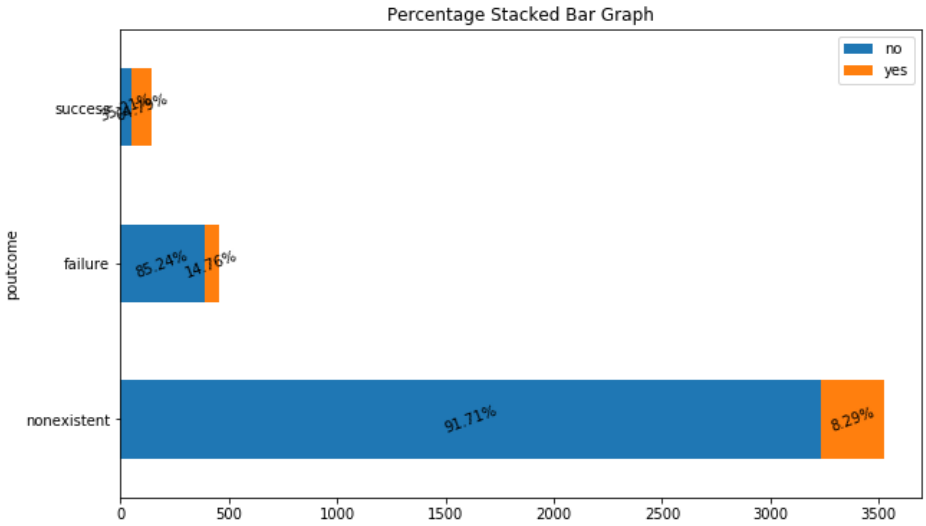
Σχήμα 4.35: Κατανομή τιμών των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής pdays

Διαφορά φαίνεται να υπάρχει ανάμεσα στο άτομα που έχουν προχωρήσει σε προθεσμιακή κατάθεση και στα άτομα που δεν έχουν προχωρήσει όσον αφορά τη μεταβλητή «poutcome». Είναι εμφανές ότι το μεγαλύτερο μέρος των ατόμων που

είχαν προβεί σε προθεσμιακή κατάθεση σε προηγούμενη καμπάνια απαντάει θετικά και στην τωρινή καμπάνια της τράπεζας.

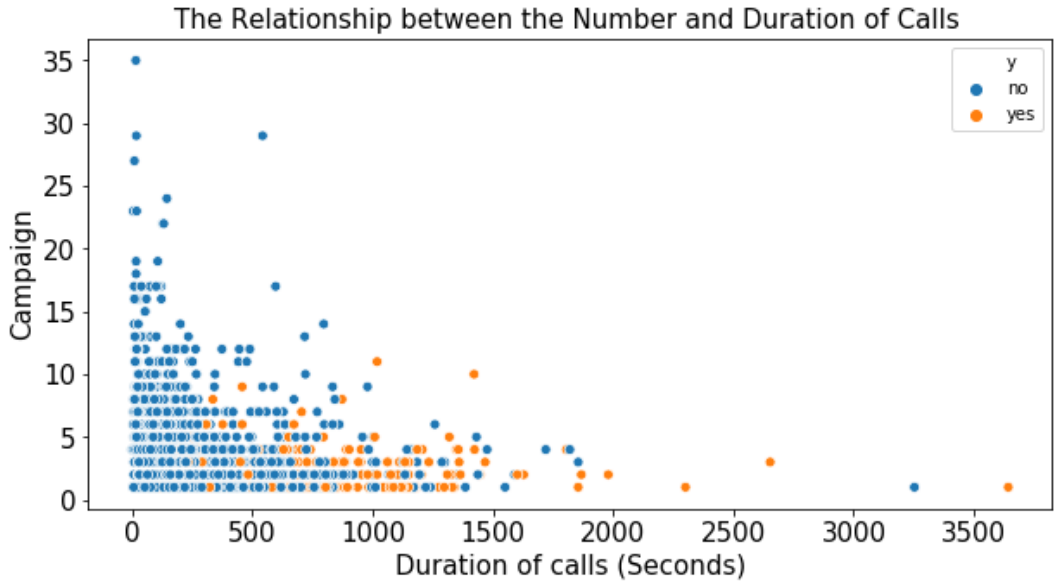
	no	yes
nonexistent	91.711609	8.288391
failure	85.242291	14.757709
success	35.211268	64.788732

Πίνακας 4.20: Πίνακας ποσοστών θετικής και αρνητικής ανταπόκρισης για τις τιμές της μεταβλητής «poutcome»



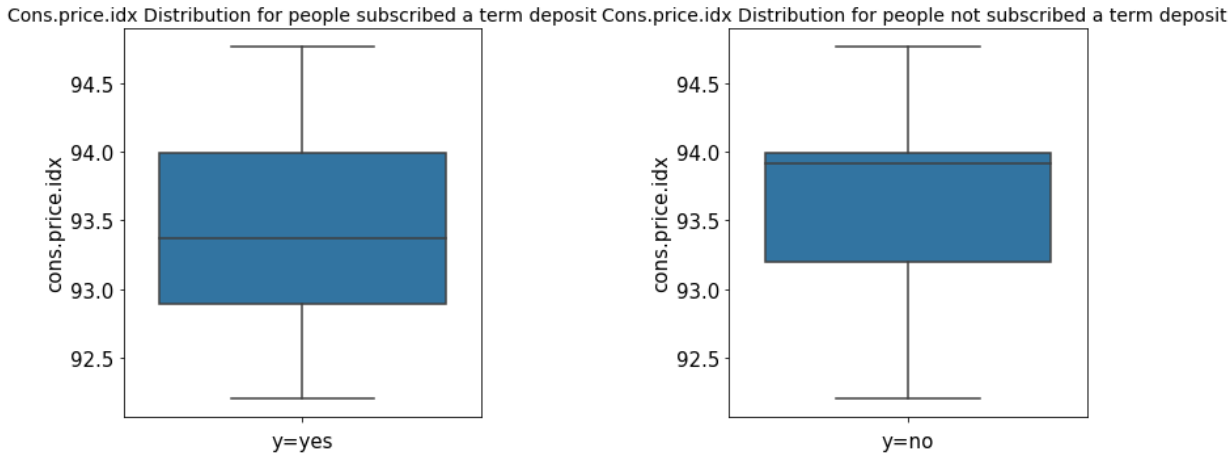
Σχήμα 4.36: Ποσοστιαία κατανομή των 2 κατηγοριών της εξαρτημένης μεταβλητής y ανά τιμή της μεταβλητής poutcome

Από το διάγραμμα διασποράς προκύπτει ότι τα άτομα που προβαίνουν σε προθεσμιακή, προσεγγίζονται λιγότερες φορές κατά τη διάρκεια της καμπάνιας και η κλήση φαίνεται να διαρκεί περισσότερη ώρα.



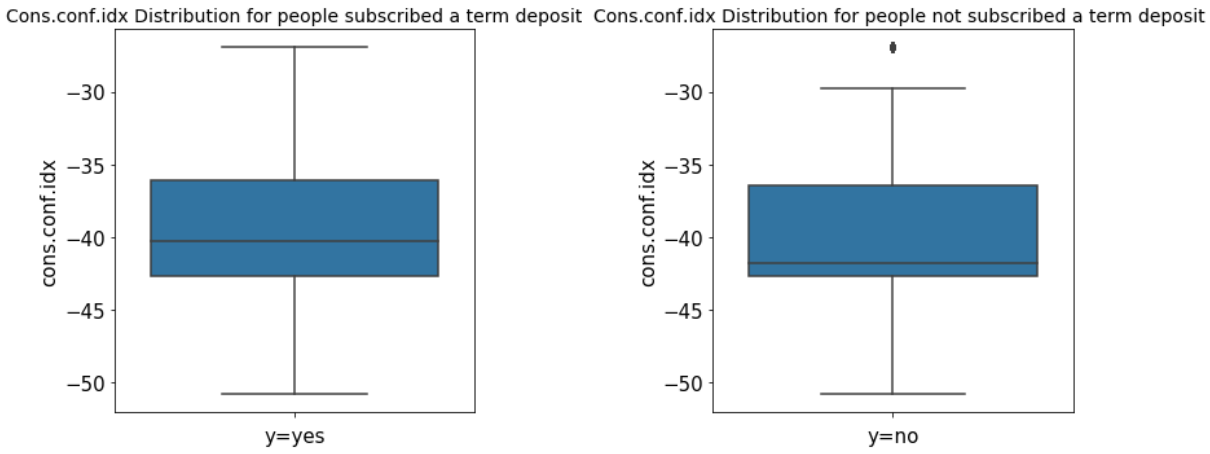
Σχήμα 4.37: Η σχέση μεταξύ του αριθμού και της διάρκειας των κλήσεων (scatter plot με βάση τις τιμές της y)

Σε αυτό το σημείο θα γίνει σύγκριση των τιμών των μεταβλητών που αφορούν οικονομικούς δείκτες ανάμεσα στα άτομα που προχωρούν σε προθεσμιακή και στα άτομα που δεν προχωρούν. Από το θηκόγραμμα στα αριστερά φαίνεται τα άτομα που προχωρούν σε προθεσμιακή να έχουν ελάχιστα μικρότερες τιμές στη μεταβλητή «cons.price.idx».



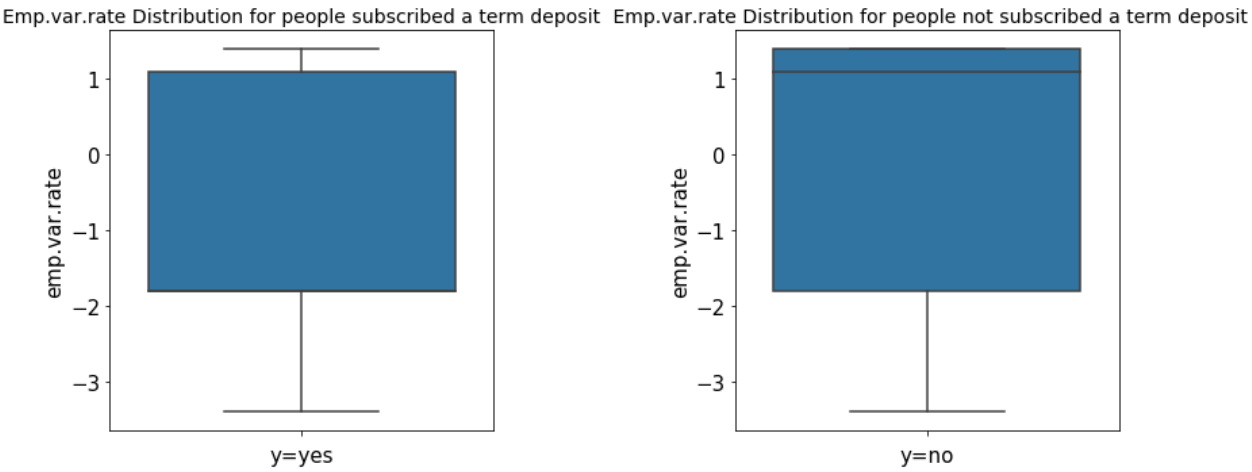
Σχήμα 4.38: Κατανομή τιμών της μεταβλητής Cons.price.idx για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y

Όσον αφορά τη σύγκριση τιμών της μεταβλητής «cons.conf.idx», φαίνεται τα άτομα που προβαίνουν σε προθεσμιακή να έχουν ελάχιστα μεγαλύτερες τιμές από τα άτομα που δεν προβαίνουν.



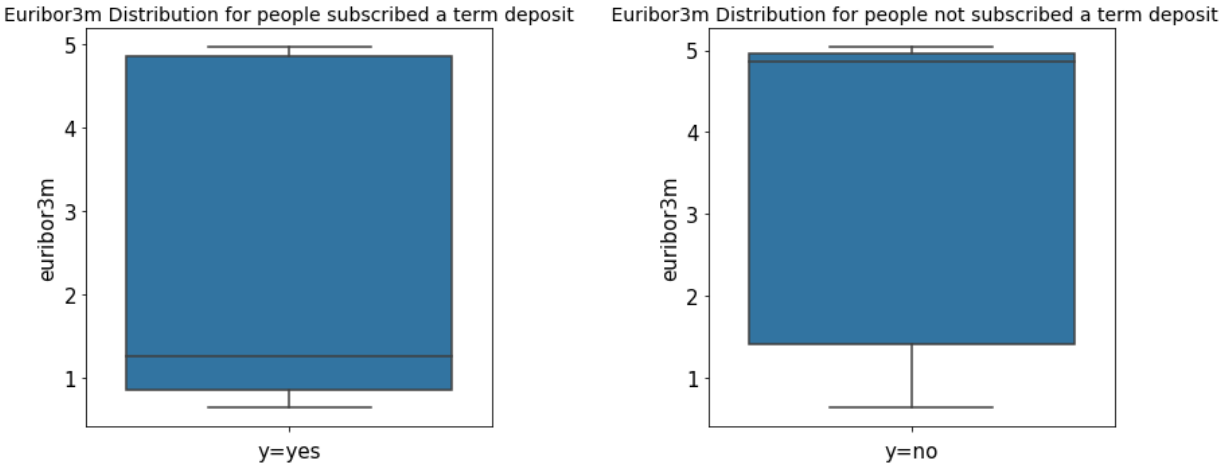
Σχήμα 4.39: Κατανομή τιμών της μεταβλητής Cons.conf.idx για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y

Εστιάζοντας στα θηκογράμματα που ακολουθούν, είναι εμφανής η διαφορά που υπάρχει ανάμεσα στα άτομα που προχωρούν σε προθεσμιακή κατάθεση και στα άτομα που δεν προχωρούν όσον αφορά τις τιμές της μεταβλητής «emp.var.rate». Η διάμεσος στο αριστερό θηκόγραμμα είναι ίση με -1,8 και ταυτίζεται με το κάτω άκρο του ορθογωνίου ενώ η διάμεσος του δεξιού θηκογράμματος είναι 1,1. Αυτό σημαίνει ότι το 50% των ατόμων που έχουν κάνει προθεσμιακή κατάθεση έχουν τιμές emp.var.rate κάτω από -1,8 ενώ το 50% αυτών που δεν έχουν κάνει προθεσμιακή έχουν τιμές κάτω από 1.1. Δηλαδή, όσοι κάνουν προθεσμιακή φαίνεται να έχουν μικρότερες τιμές σε αυτήν την μεταβλητή.



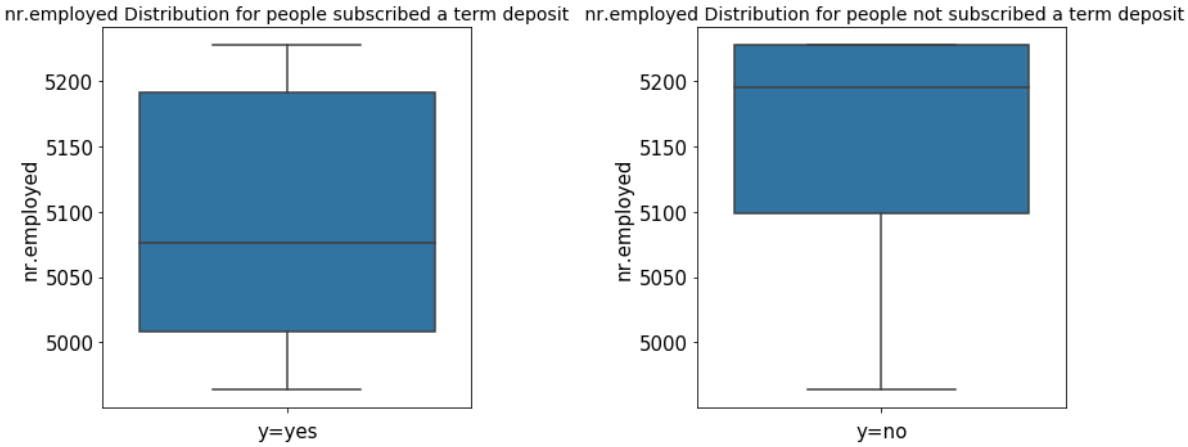
Σχήμα 4.40: Κατανομή τιμών της μεταβλητής Emp.var.rate για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y

Συγκρίνοντας τα θηκογράμματα που ακολουθούν παρατηρείται η διαφορά που υπάρχει στις διαμέσους των 2 δειγμάτων και συμπεραίνεται ότι τα άτομα που προχωρούν σε προθεσμιακή κατάθεση έχουν χαμηλότερες τιμές στην μεταβλητή «Euribor3m» από τα άτομα που δεν προχωρούν σε προθεσμιακή.



Σχήμα 4.41: Κατανομή τιμών της μεταβλητής Euribor3m για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y

Από τα ακόλουθα θηκογράμματα φαίνεται η σύγκριση των τιμών της μεταβλητής «nr.employed» γι αυτούς που προβαίνουν σε προθεσμιακή και γι αυτούς που δεν προβαίνουν. Συγκεκριμένα, τα άτομα που προβαίνουν σε προθεσμιακή κατάθεση έχουν μικρότερες τιμές στη μεταβλητή αυτή από τα άτομα που δεν προβαίνουν.



Σχήμα 4.42: Κατανομή τιμών της μεταβλητής Nr.employed για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y

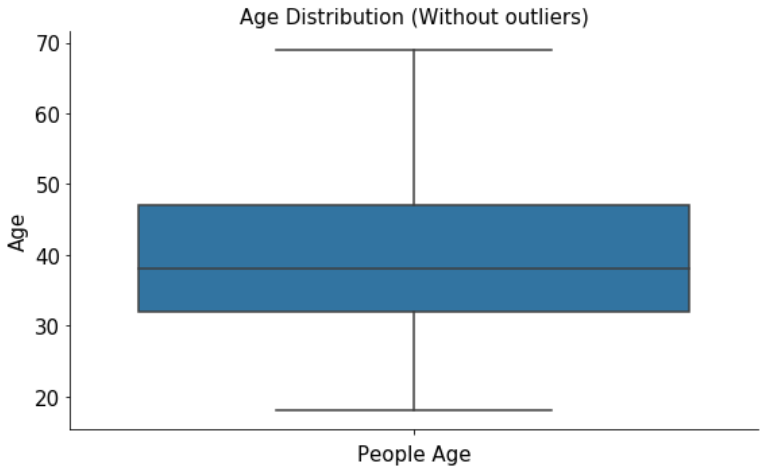
Συνοψίζοντας, από την σύγκριση που προηγήθηκε έγινε μία πρώτη διερεύνηση των πιο σημαντικών μεταβλητών στην έκφανση του αποτελέσματος της τωρινής

καμπάνιας. Τα παραπάνω θα ελεγχθούν και με τη χρήση των 'Features Importance' με το μοντέλο ταξινόμησης που θα αποδειχθεί πιο αποδοτικό στο κεφάλαιο 5.

4.5 Προ-επεξεργασία δεδομένων

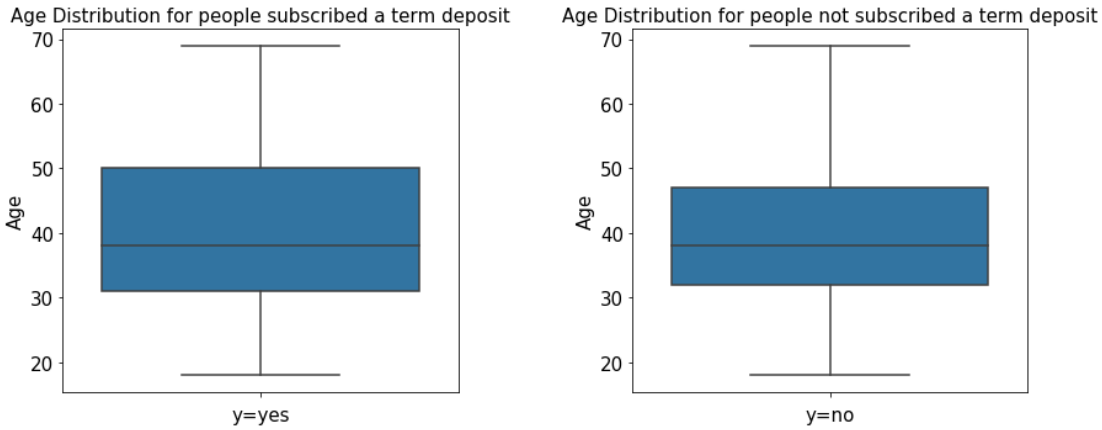
4.5.1 Αφαίρεση ακραίων τιμών (Outliers)

Όπως είχε αναφερθεί και προηγουμένως όσον αφορά τη μεταβλητή «age», οι μεγαλύτερες τιμές από τα 69,5 έτη είναι outliers. Συνολικά υπάρχουν 39 outliers στα 4.119 άτομα, οι οποίες και αφαιρούνται. Μετά την αφαίρεση των ακραίων τιμών το δείγμα αποτελείται από 4.080 άτομα και το θηκόγραμμα αυτών παρουσιάζεται παρακάτω.



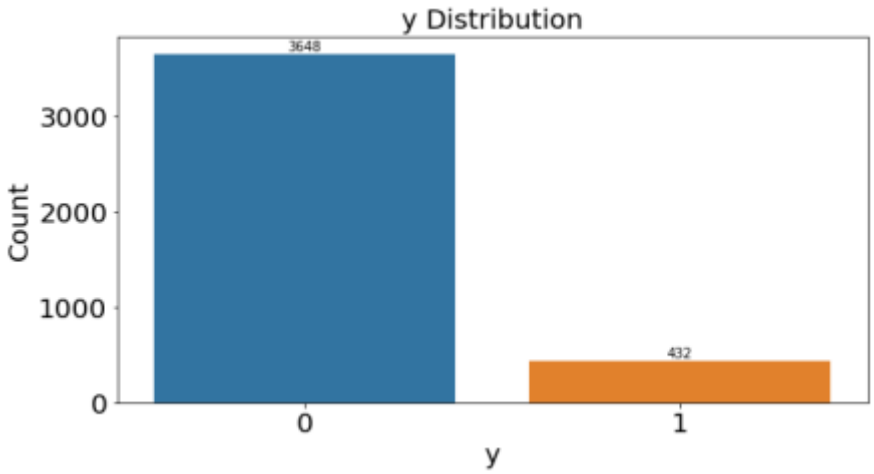
Σχήμα 4.43: Age Distribution μετά την αφαίρεση των outliers (boxplot)

Σε αυτό το σημείο παρατίθονται τα θηκογράμματα γι αυτούς που έκαναν προθεσμιακή κατάθεση (y='yes') και γι αυτούς που δεν έκαναν (y='no').



Σχήμα 4.44: Κατανομή τιμών της μεταβλητής age για τις 2 κατηγορίες της εξαρτημένης μεταβλητής y μετά την αφαίρεση των outliers

Από τα 4.080 άτομα, οι 432 απάντησαν θετικά στην καμπάνια και οι 3.648 απάντησαν αρνητικά.



Σχήμα 4.45: y Distribution μετά την αφαίρεση των outliers

Το ποσοστό επιτυχίας της καμπάνιας μετά την αφαίρεση των outliers είναι 432/4.080=10,5% και επομένως δεν έχει μεταβληθεί ιδιαίτερα σε σχέση με πριν (11%).

4.5.2 Προ-επεξεργασία ελλειπουσών τιμών με χρήση cross-tabulation

Οι ελλείπουσες τιμές του συνόλου δεδομένων είναι αυτές που παίρνουν την τιμή «unknown» και υπάρχουν στις μεταβλητές «job», «marital», «education», «default», «housing» και «loan». Το πλήθος των τιμών αυτών για κάθε μεταβλητή έχει

συγκεντρωθεί στον επόμενο πίνακα και αφορά το αρχικό σύνολο δεδομένων των 4.119 παρατηρήσεων και το σύνολο δεδομένων που προκύπτει μετά την απομάκρυνση των outliers που περιέχει 4.080 παρατηρήσεις.

Μεταβλητή	**Πλήθος "unknown values" (αρχικό dataset)**	**Πλήθος "unknown values" (τελικό dataset)**
job	39	39
marital	11	11
education	167	164
default	803	796
housing	105	105
loan	105	105

Πίνακας 4.21: Πίνακας πλήθους «unknown values» πριν την απομάκρυνση των outliers και μετά

Για τη διαχείριση των unknown values χρησιμοποιούμε τον πίνακα συνάφειας μεταξύ των μεταβλητών «job» και «education». Στη μεταβλητή «job» υπάρχουν 39 άγνωστες τιμές και στη μεταβλητή «education» υπάρχουν 164 άγνωστες τιμές. Όσον αφορά την 1η γραμμή του πίνακα παρατηρείται ότι υπάρχουν 24 άτομα με δουλειά admin και άγνωστη τιμή στην εκπαίδευση. Παρατηρείται όμως στην ίδια γραμμή ότι οι περισσότεροι που κάνουν αυτό το επάγγελμα είναι απόφοιτοι πανεπιστημίου σε πλήθος 567 και οι αμέσως επόμενοι είναι απόφοιτοι Δευτεροβάθμιας εκπαίδευσης σε πλήθος 311. Οπότε ίσως μπορεί να γίνει αντικατάσταση αυτών των 24 άγνωστων τιμών με 'university.degree'. Όμοια από τη 2η γραμμή του πίνακα παρατηρείται ότι υπάρχουν 42 άτομα που ασχολούνται σε χειρωνακτικά επαγγέλματα αλλά έχουν άγνωστη τιμή στην εκπαίδευση, όμως οι περισσότεροι με αυτό το επάγγελμα έχουν ως εκπαίδευση basic.9y. Οπότε μπορεί να γίνει αντικατάσταση αυτών των 42 αγνώστων τιμών με 'basic.9y'. Οι επόμενες 7 άγνωστες τιμές εκπαίδευσης για τα άτομα με επάγγελμα 'entrepreneur' μπορούν να αντικατασταθούν με την τιμή

‘university degree’ στην εκπαίδευση. Όμοια οι επόμενες 3 άγνωστες τιμές στην εκπαίδευση με τιμή στη δουλειά ‘housemaid’, μπορούν να αντικατασταθούν με την τιμή ‘basic4y’ που είναι η τιμή με τη μεγαλύτερη συχνότητα στην 4η γραμμή του πίνακα κτλ.

education job	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree	unknown	All
admin.	8	20	44	311	0	38	567	24	1012
blue-collar	221	152	324	89	0	49	6	42	883
entrepreneur	18	5	23	17	0	14	64	7	148
housemaid	48	9	5	11	0	8	22	3	106
management	13	8	20	41	0	7	220	14	323
retired	44	6	10	20	1	25	24	3	133
self-employed	11	2	28	15	0	12	87	4	159
services	16	12	56	254	0	25	18	12	393
student	2	0	5	35	0	8	12	20	82
technician	7	10	34	95	0	330	196	19	691
unemployed	13	3	18	23	0	14	37	3	111
unknown	8	1	6	6	0	2	3	13	39
All	409	228	573	917	1	532	1256	164	4080

Σχήμα 4.46: Πίνακας συνάφειας των μεταβλητών «job» και «education»

Ο πίνακας συνάφειας μετά τις αντικαταστάσεις των αγνώστων τιμών της μεταβλητής «education» με τον τρόπο που περιγράφηκε παραπάνω γίνεται

education job	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree	unknown	All
admin.	8	20	44	311	0	38	591	0	1012
blue-collar	221	152	366	89	0	49	6	0	883
entrepreneur	18	5	23	17	0	14	71	0	148
housemaid	51	9	5	11	0	8	22	0	106
management	13	8	20	41	0	7	234	0	323
retired	47	6	10	20	1	25	24	0	133
self-employed	11	2	28	15	0	12	91	0	159
services	16	12	56	266	0	25	18	0	393
student	2	0	5	55	0	8	12	0	82
technician	7	10	34	95	0	349	196	0	691
unemployed	13	3	18	23	0	14	40	0	111
unknown	8	1	6	6	0	2	3	13	39
All	415	228	615	949	1	551	1308	13	4080

Σχήμα 4.47: Πίνακας συνάφειας μετά τις αντικαταστάσεις των αγνώστων τιμών της μεταβλητής «education»

Παρόμοια διαδικασία θα εφαρμοστεί για να αντικατασταθούν οι άγνωστες τιμές της μεταβλητής «job». Τα άτομα με εκπαίδευση 'basic.4y' και άγνωστη τιμή στην μεταβλητή «job» είναι 8 σε πλήθος και οι άγνωστες τιμές μπορούν να αντικατασταθούν με 'blue-collar' γιατί αυτή η τιμή έχει τη μεγαλύτερη συχνότητα στην 1η στήλη(221). Με παρόμοια λογική και για τα άτομα με εκπαίδευση 'basic.6y' και 'basic.9y' και άγνωστη τιμή στην δουλειά, μπορούν να αντικατασταθούν με 'blue-collar'.

Με την αντικατάσταση των άγνωστων τιμών της μεταβλητής «job» προκύπτει ο παρακάτω πίνακας συνάφειας. Παρατηρούμε ότι υπάρχουν 13 παρατηρήσεις για τις οποίες δεν μπορούμε να προβούμε σε κάποια αντικατάσταση γιατί έχουν άγνωστη τιμή και στη μεταβλητή «job» και στην «education».

education job	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree	unknown	All
admin.	8	20	44	317	0	38	594	0	1021
blue-collar	229	153	372	89	0	49	6	0	898
entrepreneur	18	5	23	17	0	14	71	0	148
housemaid	51	9	5	11	0	8	22	0	106
management	13	8	20	41	0	7	234	0	323
retired	47	6	10	20	1	25	24	0	133
self-employed	11	2	28	15	0	12	91	0	159
services	16	12	56	266	0	25	18	0	393
student	2	0	5	55	0	8	12	0	82
technician	7	10	34	95	0	351	196	0	693
unemployed	13	3	18	23	0	14	40	0	111
unknown	0	0	0	0	0	0	0	13	13
All	415	228	615	949	1	551	1308	13	4080

Σχήμα 4.47: Πίνακας συνάφειας μετά τις αντικαταστάσεις των αγνώστων τιμών της μεταβλητής «job»

Συνοψίζοντας, μετά την αντικατάσταση των 'Unknown values' μέσω του cross-tabulation, δίνεται ο πίνακας που περιέχει τις μεταβλητές που συνεχίζουν να έχουν άγνωστες τιμές, το πλήθος αυτών και την επικρατούσα τιμή κάθε μεταβλητής:

Μεταβλητή	**Πλήθος "unknown values"**	**Επικρατούσα τιμή**
job	13	'admin'

marital	11	‘married’
education	13	‘university.degree’
default	796	‘no’
housing	105	‘yes’
loan	105	‘no’

Πίνακας 4.22: Πίνακας πλήθους «unknown values» μετά την εφαρμογή του cross-tabulation και επικρατούσα τιμή για κάθε μεταβλητή

Για την διαχείριση των υπολοίπων αγνώστων τιμών, δίνονται κάποιοι ακόμα πίνακες συνάφειας, όμως δεν φαίνεται να υπάρχει κάποια πιθανή συσχέτιση μεταξύ των μεταβλητών εκτός της συσχέτισης που αναφέρθηκε προηγουμένως ανάμεσα στις μεταβλητές «job» και «education». Για να αντικατασταθούν οι άγνωστες τιμές που υπάρχουν θα χρησιμοποιηθεί η επικρατούσα τιμή κάθε μεταβλητής.

marital job	divorced	married	single	unknown	All
admin.	114	516	388	3	1021
blue-collar	65	645	187	1	898
entrepreneur	22	107	19	0	148
housemaid	16	75	15	0	106
management	42	232	49	0	323
retired	24	97	9	3	133
self-employed	16	94	48	1	159
services	48	244	100	1	393
student	0	3	79	0	82
technician	78	386	228	1	693
unemployed	11	71	28	1	111
unknown	0	12	1	0	13
All	436	2482	1151	11	4080

Σχήμα 4.48: Πίνακας συνάφειας μεταξύ των μεταβλητών «job» και «marital»

Από την τελευταία στήλη του παραπάνω πίνακα παρατηρείται ότι η επικρατούσα τιμή της μεταβλητής «job» είναι η «admin» καθώς με συχνότητα 1.021 ενώ από την

τελευταία γραμμή ότι η επικρατούσα τιμή της «marital» είναι η τιμή «married» με συχνότητα 2.482.

default housing	no	unknown	yes	All
no	1458	363	1	1822
unknown	81	24	0	105
yes	1744	409	0	2153
All	3283	796	1	4080

Σχήμα 4.49: Πίνακας συνάφειας μεταξύ των μεταβλητών «housing» και «default»

Όμοια, παρατηρείται ότι η επικρατούσα τιμή της «housing» είναι τα άτομα που κατέχουν στεγαστικό δάνειο με συχνότητα 2.153 ενώ της μεταβλητής «default» είναι τα άτομα που δεν έχουν κάποια αθέτηση με συχνότητα 3.283.

Ακόμη, από τον παρακάτω πίνακα παρατηρείται ότι η επικρατούσα τιμή της «loan» είναι τα άτομα που δεν έχουν καταναλωτικό δάνειο με συχνότητα 3.316.

loan housing	no	unknown	yes	All
no	1568	0	254	1822
unknown	0	105	0	105
yes	1748	0	405	2153
All	3316	105	659	4080

Σχήμα 4.50: Πίνακας συνάφειας μεταξύ των μεταβλητών «housing» και «loan»

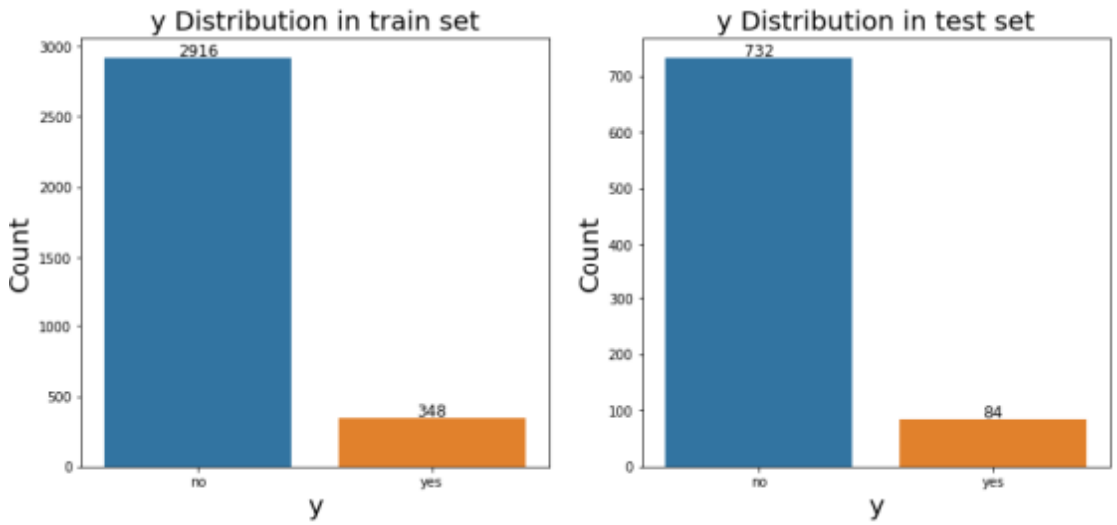
Όπως αναφέρθηκε και προηγουμένως για τη διαχείριση αυτών των άγνωστων τιμών στο σύνολο δεδομένων μας, χρησιμοποιούμε την επικρατούσα τιμή κάθε μεταβλητής. Επομένως, οι 13 άγνωστες τιμές της μεταβλητής «job» αντικαθίστανται όλες ως «admin», οι 11 άγνωστες τιμές της «marital» αντικαθίστανται με «married», οι 13 άγνωστες τιμές της «education» με «university.degree», οι 796 της «default» με «no», οι 105 της «housing» με «yes» και οι 105 της «loan» με «no».

Κεφάλαιο 5

Ταξινόμηση και Αποτελέσματα

Έχοντας κατανοήσει και προετοιμάσει κατάλληλα τα δεδομένα επόμενο βήμα είναι η επιλογή του κατάλληλου μοντέλου που θα χρησιμοποιηθεί και η ανάπτυξή του. Τα μοντέλα προς υλοποίηση βασίζονται στις τεχνικές Μηχανικής Μάθησης. Με τα δεδομένα και το μοντέλο έτοιμα, επόμενο βήμα είναι η εκπαίδευση του μοντέλου στα δεδομένα μας. Γι αυτό το λόγο το σύνολο των δεδομένων χωρίστηκε σε δύο υποσύνολα, το ένα με το 80% των δεδομένων για εκπαίδευση (σύνολο εκπαίδευσης) και το άλλο με το υπόλοιπο 20% για να δοκιμάσει το αναπτυγμένο μοντέλο (σύνολο ελέγχου). Με το διαχωρισμό αυτό προκύπτει το σύνολο εκπαίδευσης να έχει 3.264 παρατηρήσεις και το σύνολο ελέγχου 816 παρατηρήσεις. Τέλος, παρουσιάζονται τα αποτελέσματα τα οποία προέκυψαν από την διεξαγωγή των πειραμάτων με στόχο να συγκριθούν οι διαφορετικές μέθοδοι.

Από το γεγονός ότι από τα 4.080 άτομα του δείγματος, οι 432 απάντησαν θετικά στην καμπάνια και οι 3.648 απάντησαν αρνητικά, είναι προφανές ότι έχουμε να διαχειριστούμε ένα μη ισορροπημένο σύνολο δεδομένων. Δηλαδή έχουμε ένα σύνολο του οποίου το πλήθος των παρατηρήσεων δεν είναι το ίδιο για όλες τις κλάσεις και επομένως μερικές από τις τάξεις δεν “εκπροσωπούνται” επαρκώς. Όμοια, αυτή η άνιση κατανομή των παρατηρήσεων στις τάξεις παραμένει και στο σύνολο εκπαίδευσης στο οποίο παρατηρείται ότι υπάρχουν 348 άτομα που απάντησαν θετικά και 2.916 άτομα που απάντησαν αρνητικά αλλά και στο σύνολο ελέγχου που υπάρχουν 84 άτομα που απάντησαν θετικά και 732 που απάντησαν αρνητικά.



Σχήμα 5.1: Το πλήθος των παρατηρήσεων για κάθε τάξη στο σύνολο εκπαίδευσης και στο σύνολο ελέγχου

5.1 Πρώτη Περίπτωση Μελέτης

Στην πρώτη περίπτωση μελέτης θα γίνει η σύγκριση των αλγορίθμων ταξινόμησης με βάση το μοντέλο της Λογιστικής Παλινδρόμισης στο σύνολο δεδομένων. Η διαδικασία της σύγκρισης γίνεται σε 3 περιπτώσεις. Στην 1η περίπτωση, θα εφαρμοστεί ο συμβατικός αλγόριθμος με τις προκαθορισμένες παραμέτρους στο μη ισορροπημένο σύνολο δεδομένων. Στην 2η περίπτωση, θα μελετηθεί η απόδοση του ίδιου αλγορίθμου στο ισορροπημένο σύνολο δεδομένων μετά την απαραίτητη δειγματοληψία. Με στόχο να γίνει η εξισορρόπηση της κατανομής των δεδομένων θα χρησιμοποιήσουμε τη μέθοδο Random Under-sampling, την Random over-sampling και τη SMOTE. Γνωρίζουμε ότι όλες τις τεχνικές δειγματοληψίας για εξισορρόπηση δεδομένων, πρέπει να τις εφαρμόσουμε στο σύνολο εκπαίδευσης, διότι δεν πρέπει να διακινδυνέψουμε να έχουμε τεχνητά δείγματα στο σύνολο ελέγχου. Είναι μια μέθοδος για καλύτερη εκπαίδευση του μοντέλου και σε καμία περίπτωση δεν θα έπρεπε να εφαρμοστεί σε όλο το σύνολο δεδομένων μας. Στην 3η περίπτωση θα εφαρμοστεί μία τροποποιημένη μορφή του συμβατικού αλγορίθμου της Λογιστικής Παλινδρόμισης, που ονομάζεται cost-sensitive Logistic Regression (ή αλλιώς Weighed Logistic Regression) που προκύπτει με την πρόσθεση βαρών.

5.1.1 Περίπτωση της ταξινόμησης σε μη ισορροπημένο σύνολο με τη χρήση του συμβατικού αλγορίθμου Logistic Regression

Το πρώτο μοντέλο που εφαρμόζεται στο σύνολο δεδομένων είναι η Λογιστική Παλινδρόμιση. Παρουσιάζεται αρχικά, ο πίνακας σύγχυσης που δίνει True-Positive=18, False-Negative=66, False-Positive =12 και True-Negative=720:

Πραγματικότητα vs πρόβλεψη	**y='yes'**	**y='no'**	**Σύνολο**
y='yes'	18	66	84
y='no'	12	720	732
Σύνολο	30	786	816

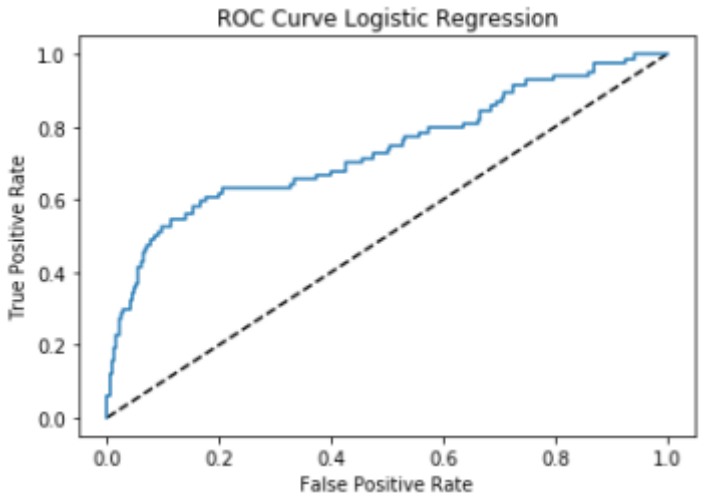
Πίνακας 5.1: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης

Στον επόμενο πίνακα δίνονται κάποιες μετρικές αξιολόγησης του μοντέλου από τις οποίες παρατηρείται ότι

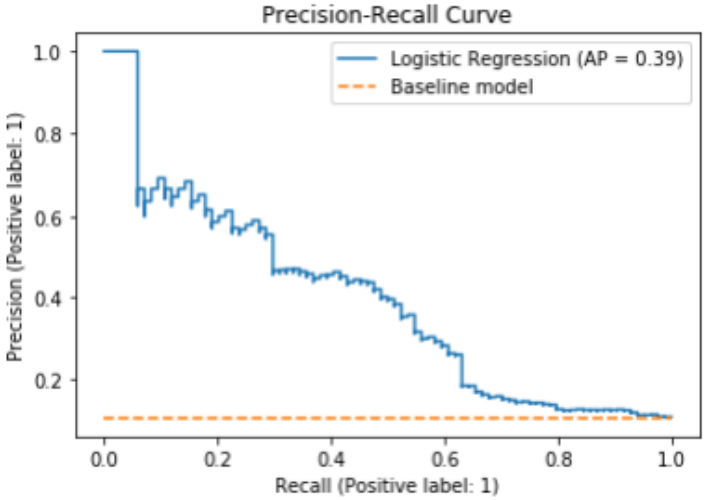
- Το μοντέλο εμφανίζει καλές τιμές για Accuracy και AUC (Accuracy= 0.90 και AUC-score =0.60).
- Για την πλειοψηφική κλάση που απαρτίζεται από τα άτομα που δεν προχωρούν σε προθεσμιακή κατάθεση (y='no'=0), οι μετρικές precision, recall, f1-score λαμβάνουν αρκετά μεγάλες τιμές
- Για την μειοψηφική κλάση που απαρτίζεται από τα άτομα που προχωρούν σε προθεσμιακή κατάθεση (y='yes'=1), οι μετρικές δεν λαμβάνουν μεγάλες τιμές. Πράγματι περιμέναμε αυτό το αποτέλεσμα στις μετρικές, καθώς το σετ δεδομένων όντας μη ισορροπημένο, τείνει να έχει bias δηλαδή να μεροληπτεί προς την τάξη ¨0' και συγκεκριμένα στα True Negatives.

	precision	recall	f1-score	support
Κλάση 0	0,92	0,98	0,95	732
Κλάση 1	0,60	0,21	0,32	84
Accuracy			0,90	816
Macro avg	0,76	AUC=0,60	0,63	816
Weighted avg	0,88	0,90	0,88	816

Πίνακας 5.2: Μετρικές αξιολόγησης για το μοντέλο της Λογ. Παλινδρόμησης



Σχήμα 5.2: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης

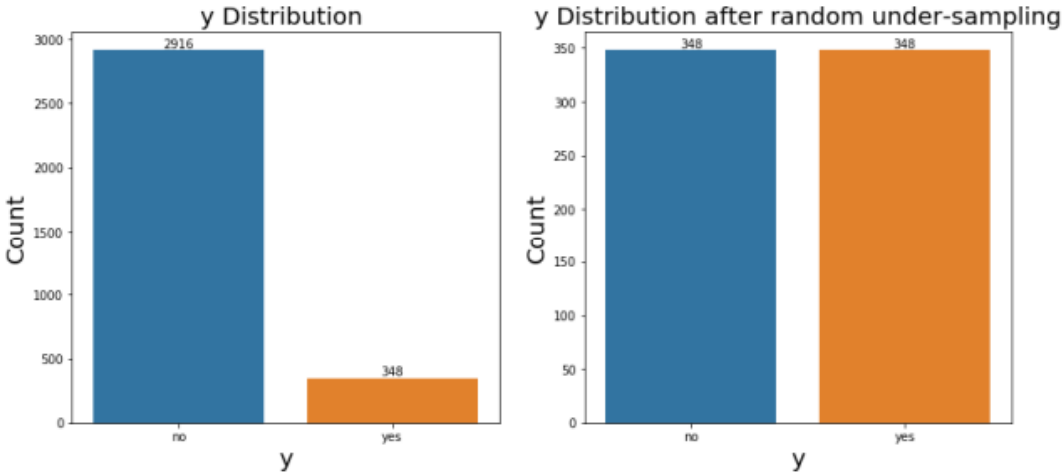


Σχήμα 5.3: Precision-Recall καμπύλη για το μοντέλο της Λογ. Παλινδρόμησης

5.1.2 Περίπτωση της ταξινόμησης σε ισορροπημένο σύνολο δεδομένων

1η υποπερίπτωση: Random Under-Sampling

Εφαρμόζουμε την απαραίτητη προεπεξεργασία των δεδομένων, η οποία έγκειται κυρίως στην μετατροπή του σε ισορροπημένο μέσω της μεθόδου της τυχαίας υποδειγματοληψίας. Εφαρμόζουμε την τεχνική μέσω της συνάρτησης RandomUnderSampler(), που παρέχεται έτοιμη στην γλώσσα Python μέσω του πακέτου imblearn.under_sampling. Η εφαρμογή της τεχνικής γίνεται στο σύνολο εκπαίδευσης και επομένως θα καταλήξουμε με ισάριθμα δείγματα σε κάθε τάξη, συγκεκριμένα 348 σε πλήθος στην κάθε μία. Η κατανομή των δειγμάτων στις δύο τάξεις πριν και μετά φαίνεται στη συνέχεια.



Σχήμα 5.4: Το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης πριν και μετά τη μέθοδο της τυχαίας υποδειγματοληψίας

Ο πίνακας σύγχυσης είναι

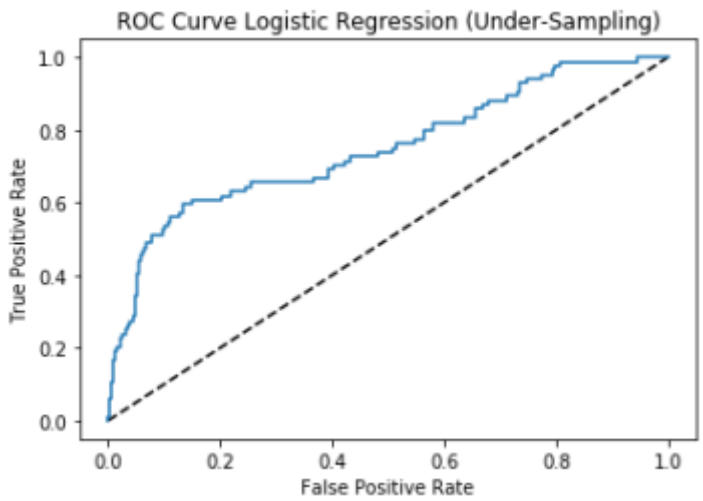
Πραγματικότητα vs πρόβλεψη	**y='yes'**	**y='no'**	**Σύνολο**
y='yes'	51	33	84
y='no'	136	596	732
Σύνολο	187	629	816

Πίνακας 5.3: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υποδειγματοληψίας

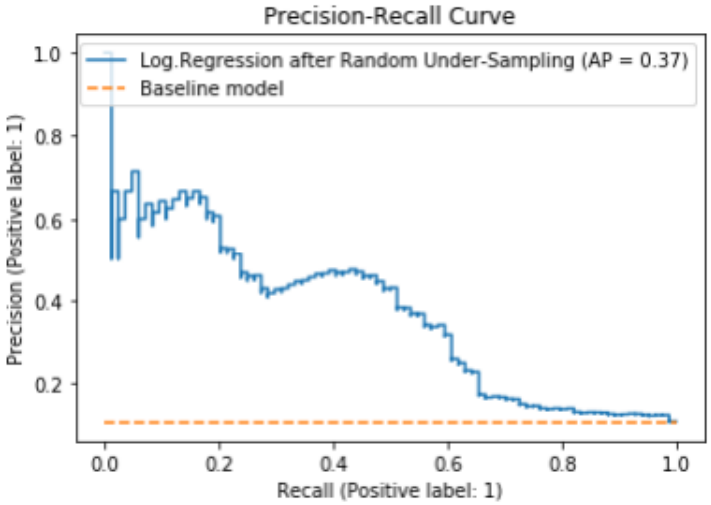
Το classification report είναι

	precision	recall	f1-score	support
Κλάση 0	0,95	0,81	0,88	732
Κλάση 1	0,27	0,61	0,38	84
accuracy			0,79	816
Macro avg	0,61	AUC=0,71	0,63	816
Weighted avg	0,88	0,79	0,82	816

Πίνακας 5.4: Μετρικές αξιολόγησης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υποδειγματοληψίας



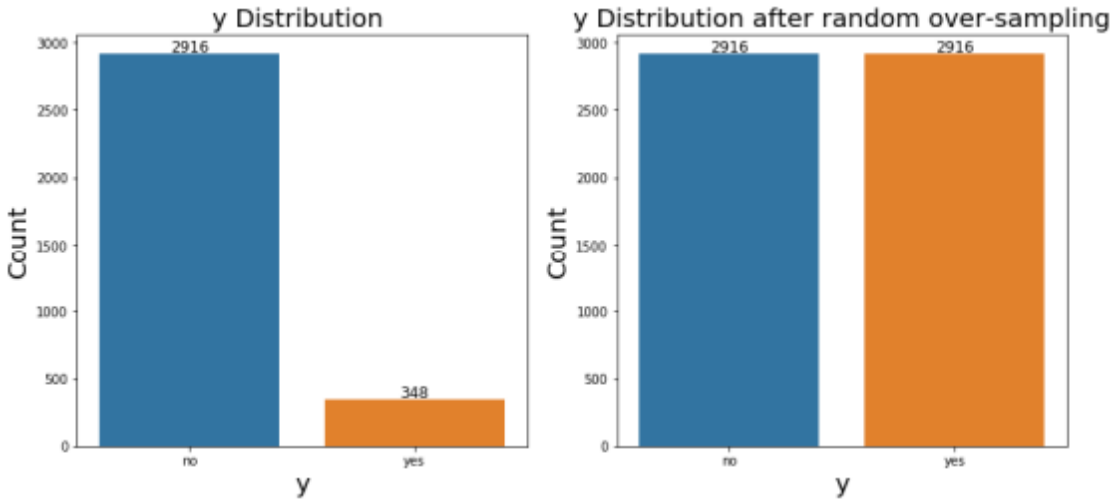
Σχήμα 5.5: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υποδειγματοληψίας



Σχήμα 5.6: Precision-Recall καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υποδειγματοληψίας

2η υποπερίπτωση: Random Over-Sampling

Για τη μετατροπή του συνόλου εκπαίδευσης σε ισορροπημένο μέσω της μεθόδου της τυχαίας υπερδειγματοληψίας, εφαρμόζουμε την τεχνική μέσω της συνάρτησης RandomOverSampler(), που παρέχεται έτοιμη στην γλώσσα Python μέσω του πακέτου imblearn.over_sampling. Μετά την εφαρμοφή την μεθόδου θα καταλήξουμε με ισάριθμα δείγματα σε κάθε τάξη, συγκεκριμένα 2.916 σε πλήθος στην κάθε μία. Η κατανομή των δειγμάτων στις δύο τάξεις πριν και μετά φαίνεται στη συνέχεια.



Σχήμα 5.7: Το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης πριν και μετά τη μέθοδο της τυχαίας υπερδειγματοληψίας

Ο πίνακας σύγχυσης είναι

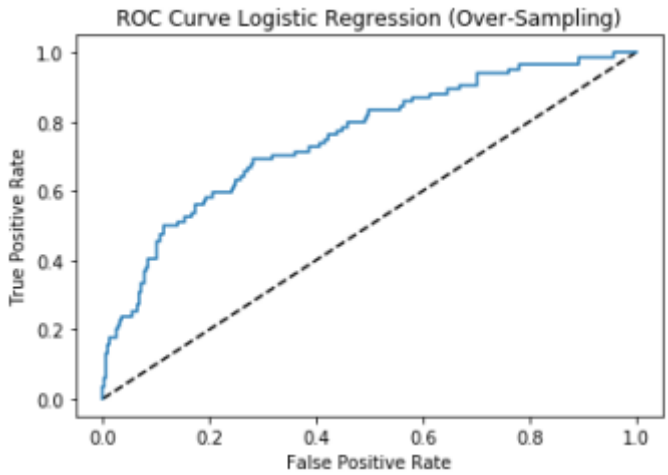
Πραγματικότητα vs πρόβλεψη	y='yes'	y='no'	Σύνολο
y='yes'	52	32	84
y='no'	171	561	732
Σύνολο	223	593	816

Πίνακας 5.5: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υπερδειγματοληψίας

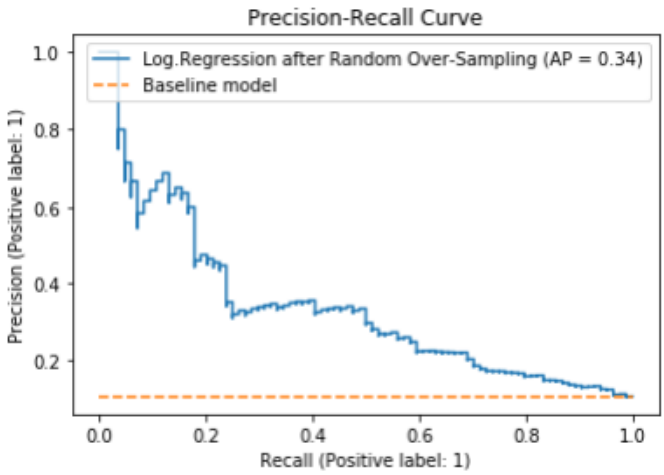
Το classification report είναι

	precision	recall	f1-score	support
Κλάση 0	0.95	0.77	0.85	732
Κλάση 1	0.23	0.62	0.34	84
accuracy			0.75	816
Macro avg	0.59	AUC=0.69	0.59	816
Weighted avg	0.87	0.75	0.79	816

Πίνακας 5.6: Μετρικές αξιολόγησης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υπερδειγματοληψίας



Σχήμα 5.8: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υπερδειγματοληψίας

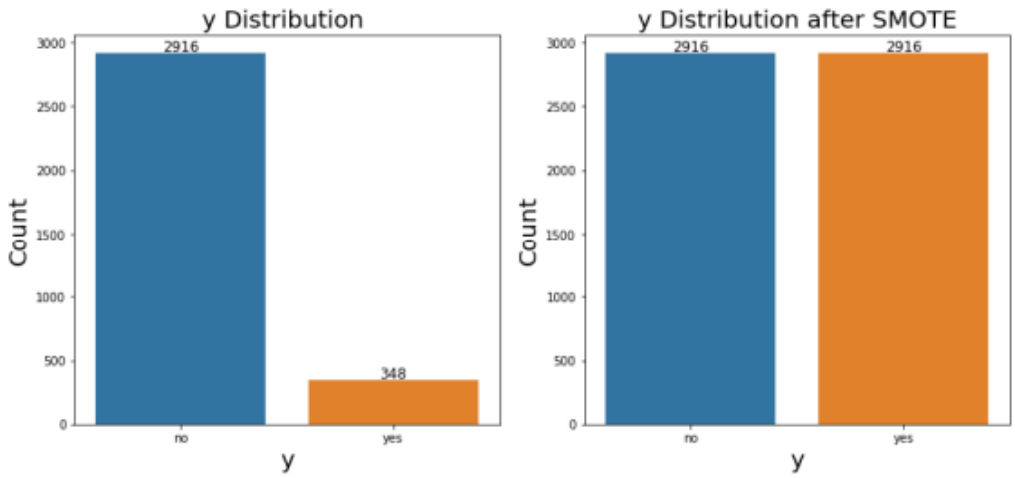


Σχήμα 5.9: Precision-Recall καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της τυχαίας υπερδειγματοληψίας

Συγκρίνοντας τα αποτελέσματα των 2 αλγορίθμων με την εφαρμογή των 2 μεθόδων δειγματοληψιών, παρατηρείται ότι η εφαρμογή της Λογιστικής Παλινδρόμησης μετά την εφαρμογή της μεθόδου random under-sampling οδηγεί σε λίγο καλύτερα αποτελέσματα από ότι η random over-sampling. Παρόλα αυτά καμία από τις 2 μεθόδους δειγματοληψίας δεν υπερέχει σημαντικά της άλλης.

3η υποπερίπτωση:Synthetic Minority Over-sampling Technique(Smote)

Όπως αναφέρθηκε και προηγουμένως χωρίζοντας το σύνολο δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου, παρατηρείται ότι το train set, στο οποίο καλείται ο αλγόριθμος να εφαρμοστεί, είναι μη ισορροπημένο περιέχοντας 348 άτομα που έχουν προβεί σε προθεσμιακή κατάθεση και σε 2.916 που δεν έχουν προβεί. Με την εφαρμογή της μεθόδου SMOTE δημιουργούνται νέα συνθετικά δεδομένα ώστε το σύνολο εκπαίδευσης να γίνει ισορροπημένο. Η κατανομή των δειγμάτων στις δύο τάξεις πριν και μετά την εφαρμογή της SMOTE φαίνεται στη συνέχεια.



Σχήμα 5.10: Το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης πριν και μετά τη μέθοδο SMOTE

Ο πίνακας σύγχυσης είναι

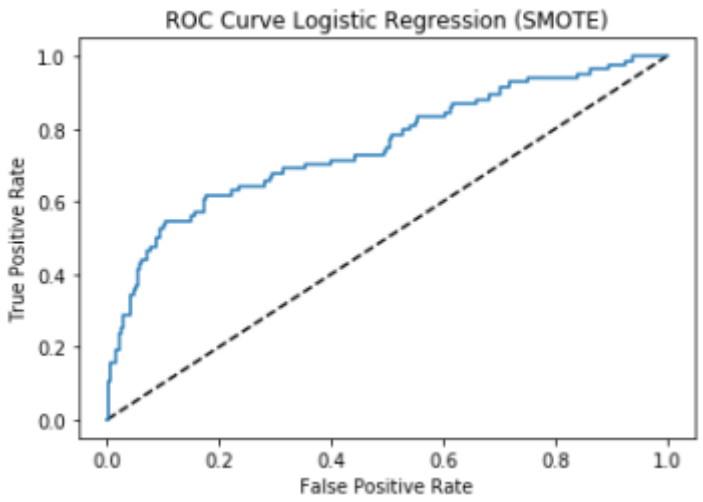
Πραγματικότητα vs πρόβλεψη	**y='yes'**	**y='no'**	**Σύνολο**
y='yes'	48	36	84
y='no'	123	609	732
Σύνολο	171	645	816

Πίνακας 5.7: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της μεθόδου SMOTE

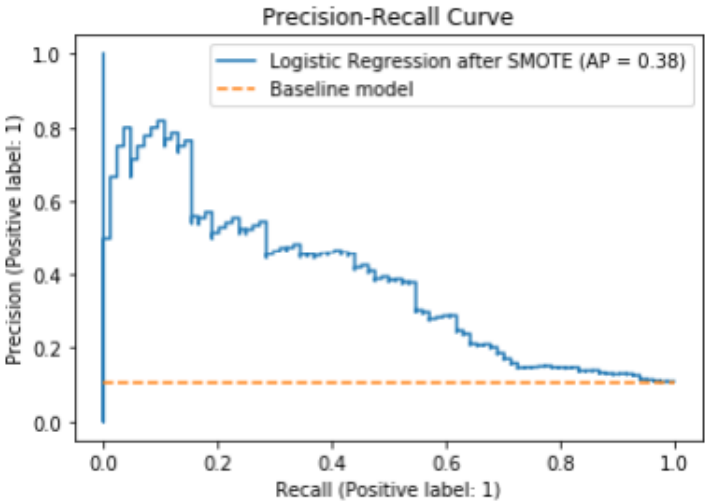
Το classification report είναι

	precision	**recall**	**f1-score**	**support**
Κλάση 0	0,94	0,83	0,88	732
Κλάση 1	0,28	0,57	0,38	84
accuracy			0,81	816
Macro avg	0,61	AUC=0,70	0,63	816
Weighted avg	0,88	0,81	0,83	816

Πίνακας 5.8: Μετρικές αξιολόγησης για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της μεθόδου SMOTE



Σχήμα 5.11: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της SMOTE



Σχήμα 5.12: Precision-Recall καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε ισορροπημένο σύνολο μετά την εφαρμογή της SMOTE

5.1.3 Περίπτωση της ταξινόμησης σε μη ισορροπημένο σύνολο με τη χρήση τροποποίησης του συμβατικού αλγορίθμου

Weighted Logistic Regression (class_weight='balanced')

Ο πίνακας σύγχυσης είναι

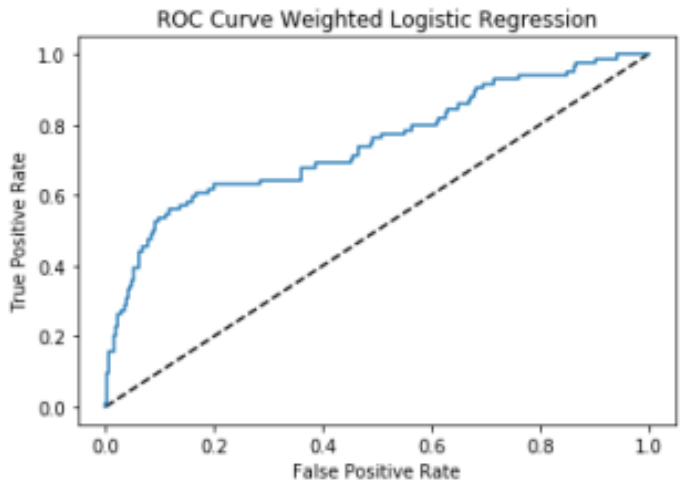
Πραγματικότητα vs πρόβλεψη	**y='yes'**	**y='no'**	**Σύνολο**
y='yes'	51	33	84
y='no'	127	605	732
Σύνολο	178	638	816

Πίνακας 5.9: Πίνακας σύγχυσης για το μοντέλο της Λογιστικής Παλινδρόμησης σε μη ισορροπημένο σύνολο με την εφαρμογή του Weighted αλγορίθμου

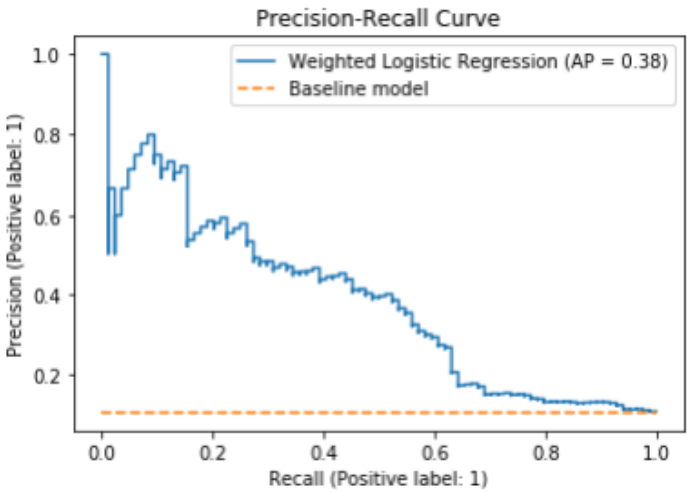
Παρατηρείται ότι για την πλειοψηφική κλάση μειώθηκαν λίγο οι τιμές των μετρικών αξιολόγησης ενώ για την μειοψηφική, η τιμή του recall αυξήθηκε σε 0.61, του precision μειώθηκε σε 0.29 ενώ του F1-score αυξήθηκε σε 0.39. Γενικότερα, το accuracy μειώθηκε λίγο (από 0.90 σε 0.80) ενώ το AUC score αυξήθηκε (από 0.60 σε 0.72).

	precision	**recall**	**f1-score**	**support**
Κλάση 0	0,95	0,83	0,58	732
Κλάση 1	0,29	0,61	0,39	84
accuracy			0,80	816
Macro avg	0,62	AUC=0,72	0,64	816
Weighted avg	0,88	0,80	0,83	816

Πίνακας 5.10: Μετρικές αξιολόγησης για το μοντέλο της Λογιστικής Παλινδρόμησης σε μη ισορροπημένο σύνολο με την εφαρμογή του Weighted αλγορίθμου



Σχήμα 5.13: ROC καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε μη ισορροπημένο σύνολο με την εφαρμογή του Weighted αλγορίθμου



Σχήμα 5.14: Precision-Recall καμπύλη για το μοντέλο της Λογιστικής Παλινδρόμησης σε μη ισορροπημένο σύνολο με την εφαρμογή του Weighted αλγορίθμου

Συνοψίζοντας, όλα τα μοντέλα βρίσκονται πολύ κοντά από άποψη απόδοσης όπως φαίνεται και στον πίνακα που ακολουθεί:

Μοντέλο Μηχανικής Μάθησης	**TP**	**FN**	**FP**	**TN**
L.R.	18	66	12	720
L.R. Random Under-	51	33	136	596

sampling				
L.R. Random Over-sampling	52	32	171	561
L.R. SMOTE	48	36	123	609
Weighted L.R.	51	33	127	605

Πίνακας 5.11: Σύγκριση των στοιχείων TP,FN,FP και TN για όλα τα μοντέλα της Λογιστικής Παλινδρόμησης

Από τον παραπάνω πίνακα φαίνεται ότι τα μοντέλα αποδίδουν σχεδόν όμοια. Παρόλα αυτά το Weighted Logistic Regression φαίνεται να εμφανίζει το καλύτερο συνδυασμό των TP, TN, FP, FN διότι είναι μέσα στα τρία μοντέλα με υψηλότερο πλήθος True Positives και ταυτόχρονα έχει το χαμηλότερο πλήθος False Positives ανάμεσα σε αυτά τα τρία. Γεγονός που το καθιστά ιδιαίτερα πετυχημένο αν αναλογιστούμε το ζητούμενο που έχουμε να λύσουμε, δηλαδή τη δημιουργία ενός μοντέλου που να ταξινομεί επιτυχώς τις νέες positive παρατηρήσεις.

Μοντέλο Μηχανικής Μάθησης	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC ROC**	**AUC PRC**
L.R.	0.90	0.60	0.21	0.32	0.60	0.39
L.R. Random Under-sampling	0.79	0.27	0.61	0.38	0.71	0.37
L.R. Random Over-sampling	0.77	0.24	0.60	0.34	0.69	0.34
L.R. SMOTE	0.81	0.28	0.57	0.38	0.70	0.38

Weighted L.R.	0.80	0,29	0,61	0,39	0.72	0.38

Πίνακας 5.12: Σύγκριση των μετρικών αξιολόγησης για για όλα τα μοντέλα της Λογιστικής Παλινδρόμησης

Από τον παραπάνω πίνακα παρατηρούμε ότι όταν το σύνολο δεδομένων είναι ισορροπημένο είτε μετά την εφαρμογή υπερδειγματοληψίας είτε υποδειγματοληψίας, το μοντέλο μπορεί και προβλέπει καλύτερα τις θετικές παρατηρήσεις (υψηλότερο Recall). Οι πιο σημαντικές μετρικές αξιολόγησης, όπως το f1-score, το AUC ROC και το AUC PRC βελτιώνονται χωρίς όμως να αυξάνονται πάρα πολύ. Ακόμη, συγκρίνοντας τα αποτελέσματα των μετρικών για όλα τα μοντέλα είτε έχει γίνει εξισορρόπηση του δείγματος είτε όχι, παρατηρείται ότι υψηλότερες τιμές στις σημαντικότερες μετρικές δίνει το μοντέλο Weighted Logistic Regression.

5.2. Δεύτερη Περίπτωση Μελέτης

Στην δεύτερη περίπτωση μελέτης θα γίνει η σύγκριση των αλγορίθμων ταξινόμησης με βάση το μοντέλο Support Vector Machine στο σύνολο δεδομένων. Η διαδικασία της σύγκρισης γίνεται όπως και προηγουμένως σε 3 περιπτώσεις. Στην 1^{η} περίπτωση, θα εφαρμοστεί ο συμβατικός αλγόριθμος με τις προκαθορισμένες παραμέτρους στο μη ισορροπημένο σύνολο δεδομένων, στην 2^{η} περίπτωση, θα μελετηθεί η απόδοση του ίδιου αλγορίθμου στο ισορροπημένο σύνολο δεδομένων μετά την απαραίτητη δειγματοληψία ενώ στην 3^{η} περίπτωση θα εφαρμοστεί μία τροποποιημένη μορφή του συμβατικού αλγορίθμου Support Vector Machine, που ονομάζεται Cost-Sensitive SVM (ή αλλιώς Weighed SVM) που προκύπτει με την πρόσθεση βαρών.

Συνοψίζοντας, όλα τα μοντέλα βρίσκονται πολύ κοντά από άποψη απόδοσης όπως φαίνεται και στον πίνακα που ακολουθεί:

Μοντέλο Μηχανικής Μάθησης	**TP**	**FN**	**FP**	**TN**
SVM	16	68	11	721
SVM Random Under-	51	33	152	580

sampling				
SVM Random Over-sampling	50	34	149	583
SVM SMOTE	37	47	101	631
Weighted SVM	48	36	125	607

Πίνακας 5.13: Σύγκριση των στοιχείων TP,FN,FP και TN για όλα τα μοντέλα SVM

Από τον παραπάνω πίνακα φαίνεται ότι τα μοντέλα αποδίδουν σχεδόν όμοια. Παρόλα αυτά το Weighted Logistic Regression φαίνεται να εμφανίζει το καλύτερο συνδυασμό των TP, TN, FP, FN διότι είναι μέσα στα τρία μοντέλα με υψηλότερο πλήθος True Positives και ταυτόχρονα έχει το χαμηλότερο πλήθος False Positives ανάμεσα σε αυτά τα τρία.

Μοντέλο Μηχανικής Μάθησης	**Accuracy**	**Precision**	**Recall**	**F1-score**	**AUC ROC**	**AUC PRC**
SVM	0,90	0,59	0,19	0,29	0,59	0,39
SVM Random Under-sampling	0,77	0,25	0,61	0,36	0,70	0,27
SVM Random Over-sampling	0,78	0,25	0,60	0,35	0,70	0,28
SVM SMOTE	0,82	0,27	0,44	0,33	0,65	0,26
Weighted SVM	0,80	0,28	0,57	0,37	0,70	0,30

Πίνακας 5.14: Σύγκριση των μετρικών αξιολόγησης για όλα τα μοντέλα SVM

Συγκρίνοντας τώρα τα αποτελέσματα των μετρικών για όλα τα μοντέλα είτε έχει γίνει εξισορρόπηση του δείγματος είτε όχι, επίσης παρατηρείται ότι υψηλότερες τιμές στις σημαντικότερες μετρικές f1-score, AUC ROC και AUC PRC δίνει το μοντέλο Weighted SVM.

5.3 Σημαντικότητα μεταβλητών

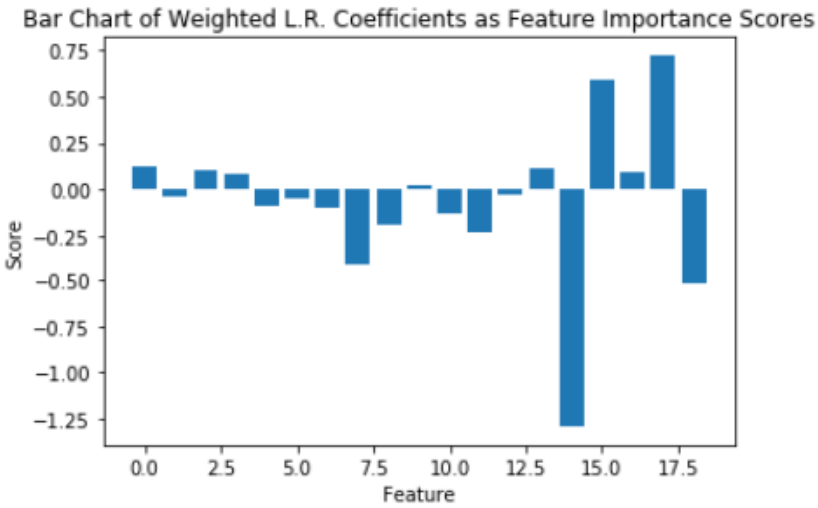
Ως σημαντικότητα μίας μεταβλητής αναφέρουμε την τεχνική ανάθεσης ενός σκορ στις μεταβλητές εισόδου που βασίζεται στο πόσο χρήσιμες είναι για την πρόβλεψη της εξαρτημένης μεταβλητής. Θα προσπαθήσουμε να ερμηνεύσουμε το μοντέλο που αποδίδει καλύτερα και να υπολογίσουμε τη σημαντικότητα των μεταβλητών εισόδου του, να μελετήσουμε πως αυτές επηρεάζουν την πρόβλεψη και να ανακαλύψουμε ποια από τα χαρακτηριστικά του συμβάλλουν περισσότερο. Ο πίνακας που ακολουθεί περιέχει τις σημαντικότητες των μεταβλητών για το μοντέλο που προέκυψε ως πιο αποδοτικό από την ανάλυση, δηλαδή το μοντέλο Weighted Logistic Regression. Με κόκκινο σημειώνονται οι τιμές των σημαντικότερων μεταβλητών και με πορτοκαλί οι αμέσως επόμενες.

Χαρακτηριστικό	**Ονομασία**	**Σκορ**
Χαρακτηριστικό 0	age	0.11935
Χαρακτηριστικό 1	job	-0.03760
Χαρακτηριστικό 2	marital	0.10201
Χαρακτηριστικό 3	education	0.08355
Χαρακτηριστικό 4	Default	-0.08556
Χαρακτηριστικό 5	Housing	-0.05451
Χαρακτηριστικό 6	Loan	-0.09775
Χαρακτηριστικό 7	Contact	-0.41979
Χαρακτηριστικό 8	Month	-0.19393
Χαρακτηριστικό 9	Day of Week	0.01951
Χαρακτηριστικό 10	Campaign	-0.12653

Χαρακτηριστικό 11	Pdays	-0.23515
Χαρακτηριστικό 12	Previous	-0.02854
Χαρακτηριστικό 13	Poutcome	0.11509
Χαρακτηριστικό 14	Emp var rate	-1.28949
Χαρακτηριστικό 15	Cons price idx	0.59105
Χαρακτηριστικό 16	Cons conf idx	0.08895
Χαρακτηριστικό 17	Euribor3m	0.72111
Χαρακτηριστικό 18	Nr employed	-0.52131

Πίνακας 5.15: Σκορ σημαντικότητας των μεταβλητών με βάση το βέλτιστο μοντέλο (Weighted Logistic Regression)

Είναι φανερό ότι οι οι μεταβλητές με τους υψηλότερους συντελεστές στο μοντέλο είναι: emp.var.rate, euribor3m, cons.price.idx, nr.employed, contact, pdays, month, campaign και poutcome. Αυτές οι μεταβλητές παρουσιάζουν μεγάλη επίδραση στην αποτελεσματικότητα του μοντέλου και άρα η γνώση τους είναι απαραίτητη για την στόχευση του σωστού group πελατών για την επιτυχία των καμπανιών μάρκετινγκ.



Σχήμα 5.15: Feature Importance Σκορ του αλγορίθμου Weighted Logistic Regression

Κεφάλαιο 6

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ

6.1 Συμπεράσματα

Η παρούσα εργασία αποτέλεσε μια προσπάθεια προσέγγισης του προβλήματος της δυαδικής ταξινόμησης σε τραπεζικά δεδομένα. Βασιζόμενοι σε μη ισορροπημένα δεδομένα που συλλέχθηκαν από μία τράπεζα και είναι διαθέσιμα στο UCI Machine Learning Repository προσπαθήσαμε να προβλέψουμε την απάντηση των ερωτώμενων για την πραγματοποιήση προθεσμιακής κατάθεσης ή όχι.

Αρχικά, έγινε μελέτη μιας διαφορετικής μεθόδου προεπεξεργασίας των δεδομένων που αφορά την διαχείριση των «unknown values» με τη χρήση cross-tabulation. Αναλυτικότερα, κατά την εκπόνηση της εργασίας μελετήθηκαν 2 αλγόριθμοι και χρησιμοποιήθηκαν διάφορες μέθοδοι δειγματοληψίας στην προσπάθεια μετατροπής του μη ισορροπημένου συνόλου δεδομένων σε ισορροπημένο.

Όταν εφαρμόζουμε μία καμπάνια μάρκετινγκ λόγω των δαπανών που προκύπτουν θα πρέπει να είμαστε όσο μπορούμε ακριβείς στη γνώση του πελατειακού μας στόχου όσον αφορά ποιά άτομα θα πρέπει να καλέσουμε για να φέρουμε νέους πελάτες και νέα έσοδα.

Είναι εμφανές δηλαδή ότι το επίπεδο της οικονομίας μιας χώρας έχει ιδιαίτερα σημαντικό ρόλο στην απόφαση του ατόμου αν θα πραγματοποιήσει προθεσμιακή κατάθεση. Άρα, θα πρέπει οι συγκεκριμένες καμπάνιες να προγραμματίζονται σε περιόδους οικονομικής ευημερίας και η τράπεζα πρέπει να εστιάσει στην προσέγγιση πελατών με κατάλληλες τιμές στους κοινωνικοοικονομικούς δείκτες.

Το ρόλο αυτό αναλαμβάνει το τμήμα μάρκετινγκ σε συνεργασία με τους οικονομολόγους έτσι ώστε να διερευνήσουν δείκτες όπως employment variation rate, Euribor3m, cons.price.idx, nr.employed κλπ που μπορούν να δείξουν την ύπαρξη οικονομικής ευημερίας ή όχι και να δώσουν στην τράπεζα την πληροφορία των κατάλληλων ατόμων για να προσεγγίσουν.

Η μεταβλητή *'duration'* (διάρκεια τηλεφωνικής κλήσης) δεν χρησιμοποιήθηκε στο μοντέλο πρόβλεψης για διάφορους λόγους που προαναφέραμε. Ωστόσο η συσχέτισή της με την εξαρτημένη μεταβλητή είναι μεγάλη, γεγονός που δείχνει ότι διαδραματίζει πολύ σημαντικό ρόλο στο αποτέλεσμα. Όσο περισσότερο διαρκεί η κλήση τόσο πιθανότερο είναι να προβεί σε προθεσμιακή κατάθεση ο πιθανός πελάτης. Για το λόγο αυτό, θα πρέπει να υιοθετηθούν τρόποι επιμήκυνσης της κλήσης και να βρεθούν τεχνικές ανάμειξης του πελάτη στη συζήτηση.

Το πλήθος των κλήσεων στο οποίο έχει προχωρήσει η τράπεζα για την προσέγγιση των πελατών είναι πολύ σημαντικός παράγοντας. Έχει γίνει εμφανές ότι όσο μεγαλύτερο είναι το πλήθος των κλήσεων σε έναν πελάτη τόσο πιο δύσκολο είναι να απαντήσει θετικά.

Ακόμη, παρατηρήθηκε ότι τους μήνες Μάρτιο, Απρίλιο, Ιούνιο, Σεπτέμβριο, Οκτώβριο και Δεκέμβριο φαίνεται οι πελάτες να έχουν περισσότερο θετική ανταπόκριση. Η τράπεζα μπορεί να εφαρμόσει τις καμπάνιες μάρκετινγκ αυτούς τους μήνες για να έχει μεγαλύτερη πιθανότητα να υπάρξει θετική απάντηση στην καμπάνια της.

Η τράπεζα θα πρέπει να δώσει βαρύτητα στην βάση δεδομένων των πελατών της για να μπορέσει να βελτιώσει τις τραπεζικές της υπηρεσίες. Για παράδειγμα, η οικογενειακή κατάσταση και η επαγγελματική ιδιότητα αποκαλύπτουν σε ποιο στάδιο ζωής βρίσκονται ενώ η δανειοληπτική τους κατάσταση μπορούν να αποτελούν δείκτη γενικού ρίσκου. Έχοντας αυτές τις πληροφορίες, η τράπεζα μπορεί να εκτιμήσει την πιθανότητα ο πελάτης να κάνει μια επένδυση και με αυτόν τον τρόπο να ικανοποιήσει την ζήτηση για τραπεζικές υπηρεσίες με στόχο τον σωστό πελάτη στη σωστή χρονική στιγμή.

Από την ανάλυση παρατηρήθηκε ότι συνταξιούχοι, άτομα που απασχολούνται στη διοίκηση και οι τεχνίτες είναι πιο πιθανό να πραγματοποιήσουν προθεσμιακή κατάθεση.

Επιπρόσθετα, όσον αφορά το υπόβαθρο εκπαίδευσης είναι πιο πιθανό να προβούν σε προθεσμιακή κατάθεση άτομα που είναι απόφοιτοι Πανεπιστημίου ή με μαθήματα εξειδίκευσης.

6.2 Περιορισμοί και αδυναμίες

Μετά την εφαρμογή του Γενικού Κανονισμού Προστασίας Δεδομένων (GDPR - 2016/679) από την ΕΕ που ισχύει από τις 25/05/2018 με άμεση εφαρμογή σε όλα τα Κράτη-Μέλη δημιουργεί ένα μεγαλύτερο εμπόδιο στην διαθεσιμότητα πληροφοριών, γεγονός που δυσχεραίνει την αναζήτηση δεδομένων και ιδιαίτερα των τραπεζικών δεδομένων που μας ενδιαφέρει. Γι αυτό το λόγο επιλέχθηκε το σύνολο δεδομένων που είναι διαθέσιμο στο UCI Machine Learning Repository και έχει μελετηθεί αρκετά τα τελευταία χρόνια.

Ασφαλώς σε κάθε μελέτη εκτός από τη στρατηγική που θα επιλέξουμε, τεράστιο ρόλο παίζει και η ποιότητα των δεδομένων. Τα αποτέλεσματα των μοντέλων που λάβαμε δεν είναι ιδιαίτερα εντυπωσιακά, αλλά μας βοήθησαν να κατανοήσουμε σε βάθος τη διαδικασία εκπαίδευσης ισορροπημένων και μη ισορροπημένων δεδομένων, αλλά και να συγκρίνουμε πολύ σημαντικούς αλγορίθμους ταξινόμησης.

Σε αυτό το σημείο, είναι σημαντικό να αναφέρουμε πως η βελτίωση στην απόδοση που αποσκοπούμε να επιτύχουμε σε ένα πρόβλημα Μηχανικής Μάθησης είναι άρρηκτα συνδεδεμένη με το πεδίο εφαρμογής του προβλήματος. Εν προκειμένω, στην μελέτη των τραπεζικών δεδομένων καταφέραμε να βελτιώσουμε τις σημαντικότες μετρικές απόδοσης των μοντέλων με την εφαρμογή του τροποποιημένου αλγορίθμου.

Τα αποτελέσματα που προκύπτουν από την παρούσα εργασία εγείρουν το ενδιαφέρον σχετικά με την απόδοση των δύο μοντέλων και της συγκεκριμένης προεπεξεργασίας των δεδομένων για να διερευνηθεί περαιτέρω το πρόβλημα της δυαδικής ταξινόμησης.

6.3 Προτάσεις για μελλοντική έρευνα

Κατά την εκπόνηση της εργασίας προσπαθήσαμε να ερμηνεύσουμε και να κατανοήσουμε με ποιο τρόπο προκύπτουν τα αποτελέσματα του τελικού μοντέλου, δίνοντας βαθμό σημαντικότητας στα χαρακτηριστικά και αναλύοντας ποια από αυτά συμβάλλουν περισσότερο.

Καθώς όλες οι μετρικές αξιολόγησης βελτιώθηκαν με την εφαρμογή του τροποποιημένου αλγορίθμου, καταλήγουμε ότι είναι σημαντική η γνώση και η εφαρμογή από τους αρμόδιους των Cost-sensitive αλγορίθμων. Ακόμη είναι προτιμότερο να γίνεται επιλογή των χαρακτηριστικών που επιδρούν περισσότερο στην πρόβλεψη του αποτελέσματος και τα υπόλοιπα χαρακτηριστικά προτείνεται να απομακρυνθούν.

Το πρόβλημα της δυαδικής ταξινόμησης σε τραπεζικά δεδομένα είναι ένα ευρύ θέμα με πολλές δυνατότητες επέκτασης. Μία από τις μελλοντικές επεκτάσεις πάνω στο συγκεκριμένο ζήτημα είναι η εφαρμογή μοντέλων που χρησιμοποιούν ως μεταβλητές εισόδου ένα υποσύνολο των αρχικών μεταβλητών, που έχει αποδειχτεί ότι συμβάλλουν περισσότερο στην πρόβλεψη του αποτελέσματος.

Υπάρχει μία πρωτοφανής ευκαιρία αξιοποίησης όλης αυτής της πληροφορίας με στόχο τη βελτίωση του τραπεζικού συστήματος. Η εξόρυξη δεδομένων και η βαθιά μάθηση συνήθως απαιτούν ένα τεράστιο πλήθος δεδομένων το οποίο δεν είναι σχεδόν ποτέ διαθέσιμο λόγω της προστασίας τους. Η σύγχρονη τεχνολογία της Μηχανικής Μάθησης, μπορεί να ξεπεράσει αυτό το εμπόδιο με στόχο την προσέγγιση των κατάλληλων πελατών και άρα την αύξηση των κερδών των τραπεζών.

Βιβλιογραφία

Chaitra Hegde, Aakash Kaku, Neelang Parghi, (2014) Predicting Bank Marketing Campaign Success using Machine Learning, https://aakashrkaku.github.io/

Sérgio Moro, Raul M. S. Laureano, Paulo Cortez, (2012) Enhancing Bank Direct Marketing through Data Mining, Academia Edu

Md. Rashid Farooqi, Naiyar Iqbal, (2019) Performance Evaluation for Competency of Bank Telemarketing Prediction using Data Mining Techniques, International Journal of Recent Technology and Engineering, Academia Edu

Sérgio Moro, Raul M. S. Laureano, Paulo Cortez, (2011) Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology https://repositorium.sdum.uminho.pt

Olatunji Razaq Apampa, (2016) Evaluation of classification and ensemble algorithms for bank customer marketing response prediction, Journal of International Technology and Information Management, Academia Edu

Talayeh Razzaghi, Andrea Otero, Petros Xanthopoulos, (2014) Imbalanced Classification for Business Analytics, Academia Edu

Mohamed Bekkar and Dr. Taklit Akrouf Alitouche, (2013) Imbalanced data learning approaches review, International Journal of Data Mining & Knowledge Management Process, Academia Edu

Gustavo E. A. P. A. Batista, Ronaldo C. Prati, Maria Carolina Monard, (2004) A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, https://www.researchgate.net/

Ahmad Alzahrani and Samira Sadaoui, (2019) Instance-incremental Classification of Imbalanced Bidding Fraud Data, Academia Edu

Yanmin Sun, Mohamed S. Kamel, Andrew K.C. Wong, Yang Wang, (2007) Cost-sensitive boosting for classification of imbalanced data, Elsevier, Academia Edu

Honghan Bei, Yajie Wang, Zhaonuo Ren, Shuo Jiang, Keran Li and Wenyang Wang, (2021) A Statistical Approach to Cost-Sensitive AdaBoost for Imbalanced Data Classification, Hindawi

Haibo He, Edwardo A. Garcia, (2009) Learning from Imbalanced Data, IEEE Transactions on knowledge and data engineering, Academia Edu

Yan Wang, Xuelei Sherry Ni, (2019) Predicting class-impalanced business risk using resampling, regularization, and model emsempling algorithms, International Journal of Managing Information Technology, Academia Edu

Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, (2006) Handling imbalanced datasets: A review, Academia Edu

Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, Nik Nairan Abdullah, (2014) An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets, Academia Edu

Daniel Grzonka, Grazyna Suchacka, Barbara Borowik, (2016) Application of selected supervised classification methods to bank marketing campaign, Academia Edu

Victoria López, Alberto Fernández, Jose G. Moreno-Torres, Francisco Herrera, (2012) Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics, Elsevier, Academia Edu

Dech Thammasiri, Dursun Delen, Phayung Meesad, Nihat Kasap, (2014) A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, Elsevier, Academia Edu

Iain Brown, Christophe Mues, (2011) An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Elsevier, Academia Edu

Izabela Anna Wowczko, (2014) Class Imbalance Dilemma, Academia Edu

Nitesh V Chawla, (2005) Data Mining for Imbalanced Datasets: An Overview, Academia Edu

Gourab Nath, Riddhi Luthra, Rahul Chellani, (2019) A Review On Class Imbalanced Correction Techniques: A Case Of Credit Card Default Prediction On A Highly Imbalanced Dataset, Academia Edu

Ana Palacios, Krzysztof Trawinski, Oscar Cordon, Luciano Sanchez (2014) Cost-Sensitive Learning of Fuzzy Rules for Imbalanced Classification Problems Using FURIA, Academia Edu

Avinash Sharma, Priyanka Singh, (2019) Analysis of Imbalanced Classification Algorithms: A Perspective View, International Journal of Trend in Scientific Research and Development, Academia Edu

Tom Fawcett, (2005) An introduction to ROC analysis, Elsevier, Academia Edu

Joshua Brady and Donald R. Brockmeier, Bias-Variance Tradeoff: A Property-Casualty Modeler's Perspective, https://www.casact.org/

Jesse Davis, Mark Goadrich, (2006) The Relationship Between Precision-Recall and ROC Curves, https://pages.cs.wisc.edu

Kinga Włodarczyk, Kingsley Success Ikani, (2019) Data Analysis of a Portuguese Marketing Campaign using Bank Marketing data Set, https://www.researchgate.net/

Γεώργιος Παπαδόπουλος, σημειώσεις μαθήματος, Γεωπονικό Πανεπιστήμιο Αθηνών

Ιωάννα Στυλιανού, (2021) Διπλωματική εργασία με θέμα «Αποτίμηση Ρίσκου με Χρήση Τεχνικών Μηχανικής Μάθησης για Εφαρμογήστις Επιστήμες Υγείας»

Μπατσιάκα Ι. Μαρία, (2021) Διπλωματική εργασία με θέμα «Σύγκριση Μοντέλων Μηχανικής Μάθησης για την Πρόβλεψη της Κυκλοφοριακής Ταχύτητας σε Αστικές Περιοχές με Καιρικά Μοτίβα»

Κοφτερού Δήμητρα, (2022) Διπλωματική εργασία με θέμα «Πρόβλεψη Τιμής Μετοχών με Χρήση Τεχνικών Μηχανικής Μάθησης»

Κωνσταντινίδης Στέργιος, (2022) Διπλωματική εργασία με θέμα «Αναγνώριση Είδους Μουσικής με Τεχνικές Μηχανικής και Βαθιάς Μάθησης»

Κοντιζάς Ιωάννης, (2021) Διπλωματική εργασία με θέμα «Χρήση Μεθόδων Μηχανικής Μάθησης για την Πρόβλεψη της Ευστάθειας κατά τον Αποκεντρωμένο Έλεγχο σε Έξυπνα Δίκτυα»

Αλέξανδρος Βαρκόπουλος, (2021) Διπλωματική εργασία με θέμα «Μέθοδοι Τεχνητής Νοημοσύνης για την ασφάλεια δικτυών υπολογιστών»

Μάρκος Άγγελος, σημειώσεις μαθήματος «Εξόρυξη Δεδομένων και Προχωρημένες Τεχνικές Προβλεπτικής Αναλυτικής»