ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Αναφορά 2ης υποχρεωτικής εργασίας

ΟΛΓΑ ΣΑΡΙΔΟΥ(ICS23078) – ΑΠΟΣΤΟΛΟΣ ΣΙΝΙΟΡΗΣ(ICS24123)

ΜΑΘΗΜΑ: BIG DATA
ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ: ΔΡ. ΑΣΗΜΙΝΑ ΔΗΜΑΡΑ

Περιεχόμενα

Θέμα 1ο 2
Ερώτημα 2ο – Έλεγχος Ελλιπών Τιμών 2
Ερώτημα 3ο - Επιλογή στρατηγικής καθαρισμού των missing values 2
Ερώτημα 4ο – Σχολιασμός πελατών 2
Θέμα 2ο 3
Ερώτημα 2.1 – Churn ανά τύπο συμβολαίου 3
Ερώτημα 2.2 – Churn ανά πλήθος υπηρεσιών 3
Ερώτημα 2.3 – Churn και μηνιαία χρέωση 4
Ερώτημα 2.3 (Προαιρετικό) – Churn ανά contract type 4
Ερώτημα 3ο – Churn ανά χώρα 5
Θέμα 3ο 5
Ερώτημα 3ο – Σχολιασμός στρατηγικής 5
1. Scatter Plot 6
2. Feature Importance 7
Βιβλιογραφικές αναφορές 8

Θέμα 1ο

Ερώτημα 2ο – Έλεγχος Ελλιπών Τιμών

Τα πέντε πεδία εστίασης (AGE, TENURE_MONTHS, MONTHLY_CHARGES, TOTAL_CHARGES, CHURN) έχουν την ίδια ποσοτική επίδραση στο DataFrame, με 300 ελλιπείς τιμές το καθένα (3%). Ωστόσο, το πεδίο CHURN είναι αυτό το οποίο επηρεάζει περισσότερο από όλα την ανάλυση των δεδομένων. Αυτή η μεταβλητή είναι η μεταβλητή στόχος. Η έλλειψη τιμής της σε κάποια γραμμή του DataFrame καθιστά τη γραμμή αυτή ανώφελη.

Ερώτημα 3ο - Επιλογή στρατηγικής καθαρισμού των missing values

Επιλέξαμε τη στρατηγική της Διαμέσου (Median) για την αντικατάσταση των ελλιπών τιμών στα αριθμητικά πεδία (AGE, TENURE_MONTHS, MONTHLY_CHARGES, TOTAL_CHARGES). Η Διάμεσος είναι η ασφαλέστερη μέθοδος και η πιο ανθεκτική σε ακραίες τιμές (outliers), διασφαλίζοντας ότι η αντικατάσταση (imputation) δεν θα παρασυρθεί από αυτές τις ακραίες τιμές. Για το πεδίο CHURN (μεταβλητή στόχος), θα αφαιρέσουμε τις γραμμές όπου CHURN ελλιπείς ή NULL, καθώς αυτό καθιστά τη γραμμή αυτή μη χρησιμοποιήσιμη για την εκπαίδευση του επιβλεπόμενου μοντέλου.

Ερώτημα 4ο – Σχολιασμός πελατών

Οι πελάτες που παραμένουν στην εταιρεία (churn = 0) έχουν μέση διάρκεια παραμονής 38.34 μήνες και μέση χρέωση 36.26. Εν αντιθέση, οι πελάτες που αποχωρούν (churn = 1) έχουν μέση διάρκεια παραμονής 33.58 μήνες και μέση χρέωση 37.02. Παρατηρούμε ότι η αποχώρηση είναι πιο πιθανή σε πελάτες που βρίσκονται στην εταιρεία για μικρότερο χρονικό διάστημα και με μεγαλύτερες μηνιαίες χρεώσεις. Ως εκ τούτου, δημιουργείται ένα υψηλό ρίσκο αποχώρησης (churn risk) για τους προαναφερθέντες πελάτες.

Θέμα 2ο

Ερώτημα 2.1 – Churn ανά τύπο συμβολαίου

Οι πελάτες με Month-to-Month συμβόλαιο φαίνεται πράγματι να έχουν υψηλότερο churn_rate (ρυθμό αποχώρησης) σε σχέση με τους υπόλοιπους πελάτες που έχουν συμβόλαιο τύπου One/Two Year. Συγκεκριμένα, οι Month-to-Month πελάτες έχουν churn_rate = 53.21%, οι One-Year πελάτες έχουν churn_rate = 20.61% και οι Two-Year πελάτες έχουν churn_rate = 13.65%.

CONTRACT_TYPE	total_customers	cust_churn1	churn_rate
Month-to-month	5326	2834	53.21
One year	2440	503	20.61
Two year	1934	264	13.65

Ερώτημα 2.2 – Churn ανά πλήθος υπηρεσιών

Πράγματι, οι πελάτες που συνδυάζουν περισσότερες από 3 υπηρεσίες έχουν το χαμηλότερο churn_rate = 25.12%. Ακολουθούν αυτοί με 2 υπηρεσίες με churn_rate = 37.38%. Οι πελάτες που δεν συνδυάζουν καμία υπηρεσία έχουν churn_rate = 43.23%, ενώ αυτοί που έχουν μόνο μία υπηρεσία έχουν churn_rate = 47.98% (το υψηλότερο μέχρι στιγμής).

NUM_SERVICES	churn_rate
0	43.23
1	47.98
2	37.38
3	25.12

Ερώτημα 2.3 – Churn και μηνιαία χρέωση

Οι πελάτες με churn που προσεγγίζει το 1 (δηλ. τείνουν να αποχωρήσουν) φαίνεται ότι έχουν υψηλότερες χρεώσεις κατά μέσο όρο (avg_monthly_charges). Με βάση τα αποτελέσματα, 37.02 χρηματικές μονάδες δαπανούνται από τους πελάτες που έχουν churn κοντά στο 1, ενώ οι πελάτες που έχουν churn κοντά στο 0 δαπανούν 36.26 χρηματικές μονάδες κατά μέσο όρο.

CHURN	avg_monthly_charges
0	36.26
1	37.02

Ερώτημα 2.3 (Προαιρετικό) – Churn ανά contract type

Αν και το παραπάνω ερώτημα εξήγαγε ως συμπέρασμα, πως οι πελάτες που το ποσοστό churn τους προσεγγίζει το 1 έχουν υψηλότερες μέσες μηνιαίες χρεώσεις, εδώ φαίνεται πως ανά τύπο συμβολαίου, όσοι προσεγγίζουν το 1, έχουν ελαφρώς χαμηλότερες μέσες μηνιαίες χρεώσεις σε σχέση με τους πελάτες, όπου το ποσοστό churn τους προσεγγίζει το 0.

CONTRACT_TYPE	CHURN	avg_monthly_charges
Month-to-month	0	41.21
Month-to-month	1	38.48
One year	0	35.25
One year	1	33.44
Two year	0	30.04
Two year	1	28.12

Ερώτημα 3ο – Churn ανά χώρα

Όπως, προκύπτει από τα αποτελέσματα, το μεγαλύτερο ποσοστό churn το έχει η Ιταλία(συγκεκριμένα 40.78%). Ακολουθούν η Γερμανία, η Ελλάδα, το Ηνωμένο Βασίλειο και η Ισπανία. Άξιο αναφοράς, είναι ότι αν και η Ελλάδα έχει τους περισσότερους πελάτες, είναι τρίτη σε ποσοστό churn, ενώ η Ιταλία που έχει λιγότερους πελάτες, έχει υψηλότερο ποσοστό churn.

COUNTRY	total_customers	churn_rate
IT	1003	40.78
DE	1198	37.65
GR	3732	37.03
UK	1560	36.15
ES	1050	36.1

Θέμα 3ο

Ερώτημα 3ο – Σχολιασμός στρατηγικής

Η τιμή του RMSE (7.93) υποδηλώνει ότι το μοντέλο πέφτει έξω κατά μέσο όρο περίπου 8€ ανά λογαριασμό, ένα σφάλμα που κρίνεται λογικό και «επιχειρησιακά υποφερτό», δεδομένου ότι ο μέσος όρος λογαριασμών είναι 36.54€

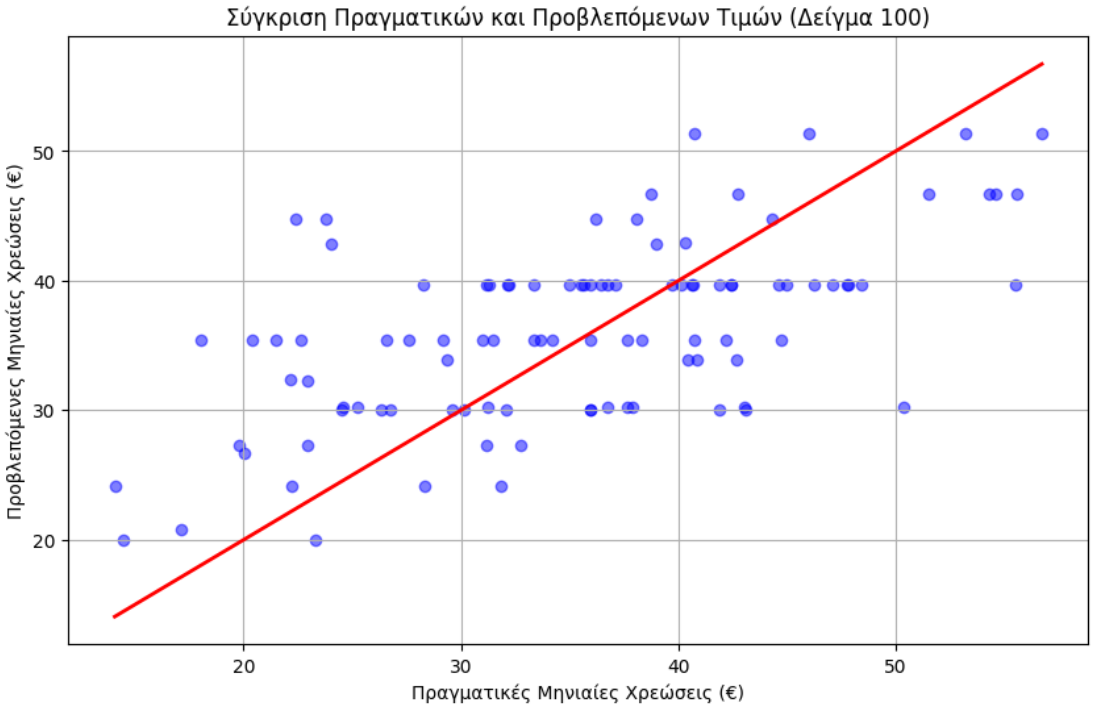
Η τιμή του R^2 (0.49) δείχνει ότι το μοντέλο εξηγεί σχεδόν το ήμισυ (49%) της διακύμανσης των χρεώσεων και αυτό συνεπάγεται ότι τα χαρακτηριστικά (feature_cols) που επιλέξαμε είναι καθοριστικά αλλά όχι τα μοναδικά.

Το μοντέλο είναι «χονδρικά χρήσιμο» για την υποστήριξη της τιμολογιακής πολιτικής και τον σχεδιασμό προσφορών, καθώς μπορεί να προβλέψει με επάρκεια την κατηγορία τιμής στην οποία θα ανήκει ένας πελάτης. Οι προβλέψεις παρέχουν μια αξιόπιστη βάση για σενάρια "what-if" και για την εκτίμηση της αναμενόμενης μηνιαίας χρέωσης βάσει του προφίλ του πελάτη.

Οπτικοποίηση αποτελεσμάτων με Scatter Plot και Feature Importance

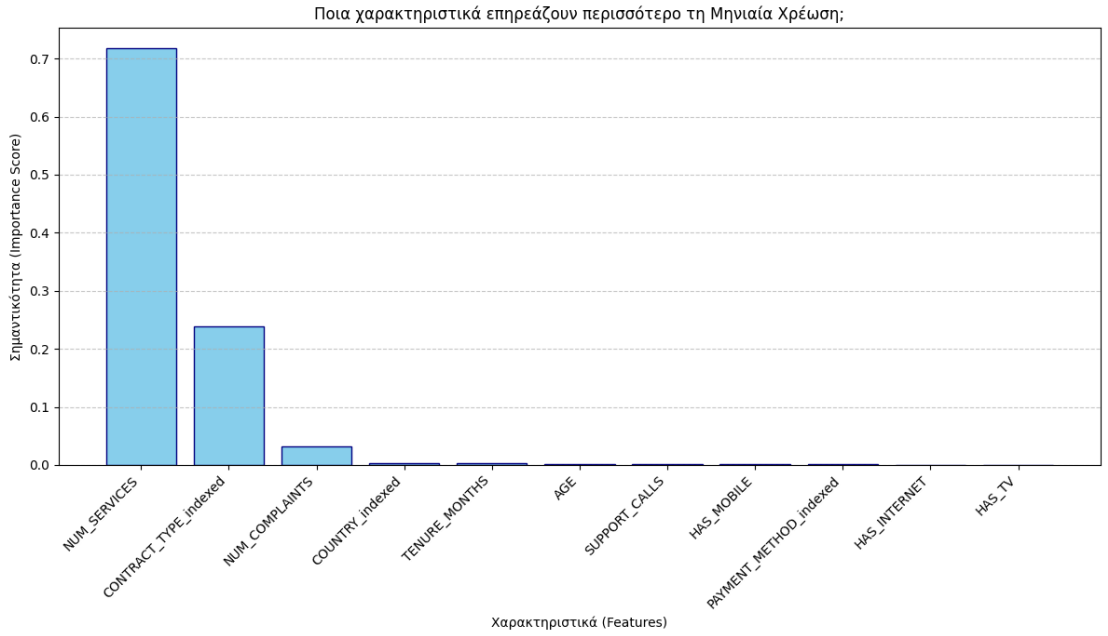
1. Scatter Plot

Στο παρακάτω διάγραμμα βλέπουμε τη σύγκριση των τιμών που προέβλεψε το μοντέλο σε σχέση με τις πραγματικές τιμές των λογαριασμών. Η κόκκινη γραμμή αντιπροσωπεύει την τέλεια πρόβλεψη. Παρατηρούμε ότι οι μπλε κουκίδες ακολουθούν τη γενική πορεία της γραμμής, γεγονός που αποδεικνύει πως το μοντέλο έχει αντιληφθεί τη γενική τάση. Ωστόσο, η διασπορά των σημείων γύρω από τη γραμμή εξηγεί το RMSE των 7.93€, υποδεικνύοντας την ύπαρξη αποκλίσεων, τις οποίες το μοντέλο δεν μπορεί να εκμηδενίσει πλήρως.



2. Feature Importance

Το συγκεκριμένο γράφημα αποκαλύπτει ποιοι παράγοντες επιβαρύνουν περισσότερο στην πρόβλεψη της μηνιαίας χρέωσης. Είναι ξεκάθαρο πως ο αριθμός των υπηρεσιών (NUM_SERVICES) είναι το κυρίαρχο χαρακτηριστικό, ακολουθούμενο από τον τύπο του συμβολαίου (CONTRACT_TYPE). Αυτό σημαίνει ότι το μοντέλο βασίζεται κυρίως στο πόσες και τι είδους παροχές έχει ένας πελάτης για να υπολογίσει το κόστος. Αντιθέτως, χαρακτηριστικά όπως η ηλικία (AGE) ή η χώρα (COUNTRY) έχουν ελάχιστη έως και μηδενική επίδραση στο τελικό αποτέλεσμα.



Βιβλιογραφικές αναφορές

I. https://spark.apache.org/sql/
II. https://www.geeksforgeeks.org/sql/sql-conversion-function/
III. https://www.geeksforgeeks.org/sql/sql-having-clause-with-examples/
IV. https://www.geeksforgeeks.org/sql/conditional-summation-in-sql/
V. https://spark.apache.org/docs/latest/ml-classification-regression.html#decision-trees
VI. https://spark.apache.org/docs/latest/ml-pipeline.html
VII. https://www.datacamp.com/tutorial/rmse
VIII. https://www.datacamp.com/tutorial/r-squared
IX. https://www.w3schools.com/python/matplotlib_plotting.asp
X. Διαφάνειες και κώδικας μαθήματος