Customer Churn με Apache Spark

Ονοματεπώνυμο 1: *Συμεών Αλμανίδης (ics23083)*
Ονοματεπώνυμο 2: *Ευστράτιος Μαυρίγκος (ics23117)*

Εκδόσεις εργαλείων:

- Python 3.12.12
- Spark 4.0.1

Υλικό (hardware) που χρησιμοποιήθηκε για την έρευνα:

Processor	Intel(R) Core(TM) i3-8100 CPU @ 3.60GHz 3.60 GHz
Installed RAM	8.00 GB
Storage	250GB
Graphics Card	NVIDIA GeForce GT 1030 (2 GB)

Θέμα 1: Φόρτωση, Έλεγχος και Καθαρισμός Δεδομένων

Δημιουργούμε ένα DataFrame στο Spark από το αρχείο `telecom_churn_10k.csv`, εμφανίζοντας τις πρώτες 10 γραμμές και το σχήμα του. Εξετάζουμε τις ελλιπείς τιμές στις στήλες AGE, TENURE_MONTHS, MONTHLY_CHARGES, TOTAL_CHARGES και CHURN, καταγράφοντας πόσες υπάρχουν. Τέλος, εφαρμόζουμε στρατηγική καθαρισμού, απομακρύνοντας τις γραμμές με κενά στο CHURN και συμπληρώνοντας τα αριθμητικά πεδία με μέσο όρο, διασφαλίζοντας την ακεραιότητα των δεδομένων. Η χρήση του μέσου όρου επιτρέπει τη διατήρηση της συνολικής κατανομής των δεδομένων, αποφεύγοντας ακραίες τιμές που θα μπορούσαν να επηρεάσουν την ανάλυση. Επίσης, αυτή η προσέγγιση είναι απλή και εύκολα κατανοητή, διευκολύνοντας την εκτέλεση στατιστικών αναλύσεων.

Θέμα 2: Ανάλυση με Spark SQL & Επιβεβαίωση Επιχειρησιακών Μοτίβων

Churn ανά τύπο συμβολαίου

CONTRACT_TYPE	TOTAL_CUSTOMERS	CHURN_COUNT	CHURN_RATE_PERCENT
Month-to-month	5326	2834	53.21
One year	2440	503	20.61
Two year	1934	264	13.65

Churn ανά πλήθος υπηρεσιών

NUM_SERVICES	TOTAL	CHURN_RATE_PERCENT
0	229	43.23
1	2149	47.98
2	5152	37.38
3	2170	25.12

Μέση Μηνιαία Χρέωση ανά Churn Status

CHURN	AVG_MONTHLY_BILL
0.0	36.28
1.0	37.04

Top-5 Χώρες με υψηλότερο ποσοστό Churn (>100 πελάτες)

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

Έλεγχος των Hints του προβλήματος

Churn Rate για Παλιούς πελάτες με Full Services

CUSTOMER_PROFILE	COUNT	CHURN_RATE
Others	8207	39.64
Full Package & Loyal	1493	23.31

Churn Rate για Mobile Only + Month-to-month

RISK_PROFILE	COUNT	CHURN_RATE
Others	8578	32.91
Mobile Only & Monthly	1122	69.34

Μέσος όρος παραπόνων/κλήσεων στους Churners

CHURN	AVG_COMPLAINTS	AVG_SUPPORT_CALLS
0.0	0.46	1.46
1.0	0.56	1.61

Η ανάλυση των δεδομένων επιβεβαιώνει ότι οι πελάτες χωρίς ισχυρές δεσμεύσεις τείνουν να αποχωρούν ευκολότερα. Αυτό αποτυπώνεται στα συμβόλαια Month-to-month (53.21% churn), ειδικά όταν δεν πλαισιώνονται από άλλες υπηρεσίες (69.34% churn σε Mobile-only). Αντίθετα, η στρατηγική διάθεσης πολλαπλών υπηρεσιών (bundling) φαίνεται να ενισχύει την πιστότητα, καθώς οι πελάτες με πλήρες πακέτο (Internet + Mobile + TV) εμφανίζουν σημαντικά χαμηλότερο ρυθμό αποχώρησης (25.12%). Τέλος, το κόστος φαίνεται να παίζει επίσης έναν ρόλο, καθώς οι αποχωρήσαντες επιβαρύνονται με ελαφρώς υψηλότερα πάγια (37.04€ έναντι 36.28€).

Θέμα 3: Πρόβλεψη Μηνιαίας Χρέωσης με Decision Tree για δυναμική τιμολόγηση

Για την πρόβλεψη της μηνιαίας χρέωσης, το μοντέλο χρησιμοποίησε τα εξής χαρακτηριστικά:

- Ηλικία (AGE)
- Διάρκεια παραμονής σε μήνες (TENURE_MONTHS)

- Αριθμός παραπόνων (NUM_COMPLAINTS)
- Αριθμός κλήσεων υποστήριξης (SUPPORT_CALLS)
- Ύπαρξη Internet (HAS_INTERNET)
- Ύπαρξη Mobile (HAS_MOBILE)
- Ύπαρξη TV (HAS_TV)
- Συνολικός αριθμός υπηρεσιών (NUM_SERVICES)
- Τύπος συμβολαίου (CONTRACT_TYPE)
- Μέθοδος πληρωμής (PAYMENT_METHOD)
- Χώρα (COUNTRY)

Τιμές Μετρικών Αξιολόγησης:

- RMSE (Root Mean Squared Error): 7.88€
- R^2 (Coefficient of Determination): 0.49

Ο μέσος όρος της μηνιαίας χρέωσης στο dataset είναι 36.56€. Με ένα RMSE περίπου 7.88€, το ποσοστό σφάλματος είναι περίπου 21.5% (7.88 / 36.56 * 100). Αυτό το επίπεδο σφάλματος θεωρείται μέτριο προς ικανοποιητικό για ένα αρχικό μοντέλο, αλλά όχι αρκετά ακριβές για εφαρμογές που απαιτούν αυστηρή τιμολόγηση. Η τιμή R^2 = 0.49 υποδηλώνει ότι το μοντέλο εξηγεί περίπου το 49% της διακύμανσης των μηνιαίων χρεώσεων, αφήνοντας ένα σημαντικό ποσοστό ανεξήγητο, κάτι που επιβεβαιώνει την ανάγκη για περαιτέρω βελτίωση.