Τεχνικές Εξόρυξης Δεδομένων

Εαρινό Εξάμηνο 2022-2023

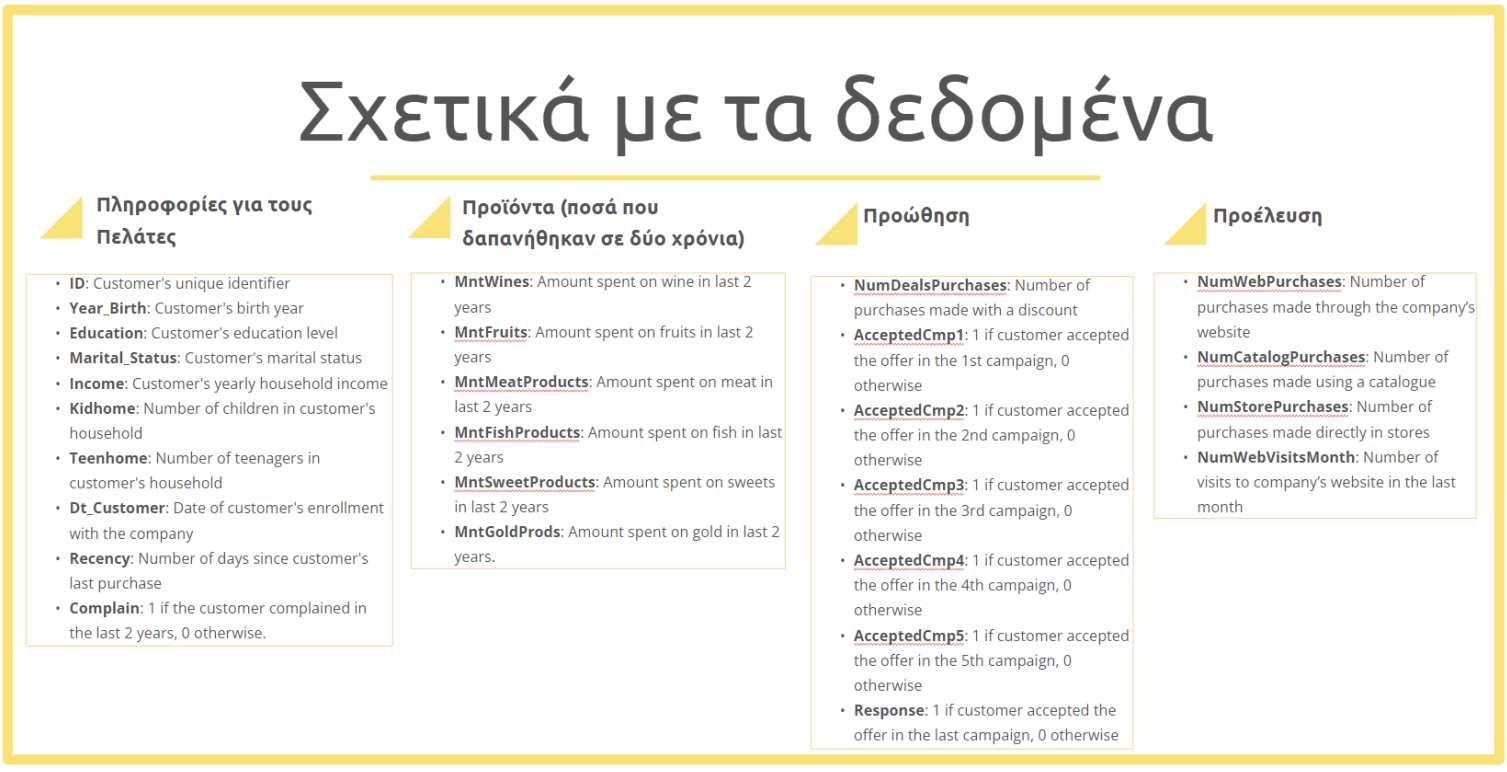
1η Άσκηση - Ατομική ή ομαδική εργασία (2 Ατόμων)

Σκοπός της εργασίας

Σκοπός της εργασίας είναι η εξοικείωσή σας με τα βασικά στάδια της διαδικασίας που ακολουθούνται για την εφαρμογή τεχνικών εξόρυξης δεδομένων, ήτοι: προ-επεξεργασία / καθαρισμός, μετατροπή, εφαρμογή τεχνικών εξόρυξης δεδομένων και αξιολόγηση. Η υλοποίηση θα γίνει στην γλώσσα προγραμματισμού Python με την χρήση των εργαλείων/βιβλιοθηκών: jupyter notebook, pandas και SciKit Learn.

Περιγραφή

Η **Ανάλυση Προσωπικότητας Πελατών** είναι μια λεπτομερής ανάλυση των ιδανικών πελατών μιας εταιρείας. Βοηθά μια επιχείρηση να κατανοήσει καλύτερα τους πελάτες της και διευκολύνει την τροποποίηση των προϊόντων σύμφωνα με τις συγκεκριμένες ανάγκες, συμπεριφορές και ανησυχίες των διαφόρων τύπων πελατών.Για παράδειγμα, αντί να δαπανά χρήματα για την προώθηση ενός νέου προϊόντος σε κάθε πελάτη στη βάση δεδομένων της επιχείρησης, μια επιχείρηση μπορεί να αναλύσει ποια κατηγορία πελατών είναι πιο πιθανό να αγοράσει το προϊόν και στη συνέχεια να προωθήσει το προϊόν μόνο σε αυτό τη συγκεκριμένη κατηγορία.



Ζητούμενα

1. **Προεπεξεργασία/Καθάρισμα**: Ελέγξτε αν υπάρχουν τιμές που λείπουν στα δεδομένα και χειριστείτε τις ανάλογα, μετατρέψτε στήλες που αφορούν ημερομηνίες σε DateTime objects και ελέγξτε αν υπάρχουν και κάποια χαρακτηριστικά dtype: object τα οποία μπορείτε να κωδικοποιήσετε/μετατρέψετε σε αριθμητικές τιμές **(5%)**.

2. Εκτυπώστε τις **μοναδικές τιμές** στα κατηγορικά χαρακτηριστικά **Marital_Status** και **Education** για να αποκτήσετε μια πιο σαφή εικόνα των δεδομένων. Αλλάξτε τις τιμές [Alone,Absurd,YOLO] των Marital_Status με την τιμή 'Single'. Χρησιμοποιήστε όποιο τύπο γραφήματος θέλετε για να παρουσιάσετε το πλήθος των τιμών σε κάθε κατηγορία. **(5%)**

3. **Δημιουργία νέων χαρακτηριστικών**: **(10%)**
   Α. Δημιουργία ενός χαρακτηριστικού ("Customer_For") που αντιπροσωπεύει τον αριθμό των ημερών που οι πελάτες άρχισαν να ψωνίζουν στο κατάστημα σε σχέση με την τελευταία καταγεγραμμένη ημερομηνία (Recency).
   Β. Εξαγωγή της ηλικίας "**Age**" ενός πελάτη με βάση το "**Year_Birth**" που υποδεικνύει το έτος γέννησης του αντίστοιχου ατόμου.
   Γ. Δημιουργήστε ένα άλλο χαρακτηριστικό "**Spent**" που υποδεικνύει το συνολικό ποσό που ξόδεψε ο πελάτης σε όλες τις κατηγορίες σε διάστημα δύο ετών.
   Δ. Δημιουργήστε ένα χαρακτηριστικό "**Παιδιά**" για να δηλώσετε το σύνολο των παιδιών σε ένα νοικοκυριό, δηλαδή τα παιδιά και τους εφήβους.
   Ε. Για να αποκτήσετε περαιτέρω σαφήνεια του νοικοκυριού, δημιουργήστε ένα χαρακτηριστικό με ένδειξη "**Family_Size**" που δείχνει το συνολικό αριθμό των ατόμων σε ένα νοικοκυριό
   Στ. Δημιουργήστε ένα χαρακτηριστικό "**Is_Parent**" που δηλώνει αν ένας πελάτης είναι και γονιός
   Ζ. Δημιουργήστε ένα άλλο χαρακτηριστικό "**Living_With**" χρησιμοποιώντας το "**Marital_Status**" για να εξάγετε την κατάσταση διαβίωσης των ζευγαριών. Συγκεκριμένα το χαρακτηριστικό αυτό πρέπει να έχει δύο τιμές, "Partner" και "Alone".
   Η. Δημιουργήστε τη στήλη "Age Group" χρησιμοποιώντας τη στήλη "Age", η οποία να ομαδοποιεί τις ηλικίες στις παρακάτω τιμές "21-30", "31-40", "41-50", "51-60", "61-70", "71-80", ">80".

4. Ελέγξτε αν υπάρχουν **ακραίες τιμές** στα χαρακτηριστικά και διαγράψτε τις από τα δεδομένα. **(5%)**

5. Στη συνέχεια, εξετάστε τη **συσχέτιση μεταξύ των χαρακτηριστικών** με ένα heatmap διάγραμμα. (Εξαιρώντας τα κατηγορικά χαρακτηριστικά σε αυτό το σημείο) **(5%)**

6. Ερωτήματα που θα απαντηθούν με **γραφήματα** - επιλέξτε 10 από τα παρακάτω. **(20%)**
    1. Σε ποιά κατηγορία Marital_Status ανήκει το μεγαλύτερο ποσοστό των πελατών της εταιρείας;
    2. Πόσοι πελάτες έχουν κάνει Complain ;
    3. Σχέση μεταξύ του αριθμού των αγορών **Spent** και της οικογενειακής κατάστασης.
    4. Η σχέση μεταξύ του αριθμού των αγορών **Spent** και του αριθμού των παιδιών και του μεγέθους της οικογένειας.
    5. Τι σχέση έχει η ηλικία **Age Group** με το χαρακτηριστικό **Spent** των αγορών;
    6. Τι σχέση έχει το εισόδημα **Income** με το χαρακτηριστικό **Spent** των αγορών;
    7. Ποια είναι η σχέση μεταξύ της εκπαίδευσης και του εισοδήματος;
    8. Ποια είναι η σχέση μεταξύ του εισοδήματος και του μεγέθους της οικογένειας;
    9. Ποια είναι η σχέση μεταξύ του εισοδήματος και του αριθμού των παιδιών;
    10. Ποια είναι η σχέση μεταξύ του εισοδήματος και του **Living_With**;
    11. Ποια είναι η σχέση μεταξύ του εισοδήματος και του αριθμού **Spent** των αγορών;
    12. Ποια είναι η σχέση μεταξύ του αριθμού των αγορών από τον δικτυακό τόπο και του αριθμού των επισκέψεων στον δικτυακό τόπο;
    13. Ποια είναι τελικά το ποσοστό των πελατών που αποδέχονται όλες τις προσφορές από το κατάστημα;
    14. Σχεδιάστε το ιστόγραμμα για τη στήλη NumDealsPurchases.
    15. Σχεδιάστε το ιστόγραμμα για τη στήλη Income.
    16. Σχεδιάστε το ιστόγραμμα για τη στήλη Kidhome.
    17. Σχεδιάστε το ιστόγραμμα για τη στήλη Family_Size.
    18. Οι πελάτες με μεταπτυχιακό τίτλο σπουδών ξοδεύουν περισσότερα χρήματα για κρασί;

7. **Principal component analysis (PCA) (25%)**: Σε αυτό το πρόβλημα, υπάρχουν πολλοί παράγοντες βάσει των οποίων γίνεται μία ταξινόμηση. Αυτοί οι παράγοντες είναι βασικά χαρακτηριστικά ή γνωρίσματα. Όσο μεγαλύτερος είναι ο αριθμός των χαρακτηριστικών, τόσο πιο δύσκολη είναι η εργασία. Πολλά από αυτά τα χαρακτηριστικά είναι συσχετισμένα και, ως εκ τούτου, περιττά. Αυτός είναι ο λόγος για τον οποίο θα πραγματοποιήσετε **μείωση της διάστασης** στα επιλεγμένα χαρακτηριστικά. Η μείωση της διάστασης είναι η διαδικασία μείωσης του αριθμού των υπό εξέταση τυχαίων μεταβλητών, και έχει ώς αποτέλεσμα την απόκτηση ενός συνόλου κύριων μεταβλητών.
Οι μεταβλητές στο σύνολο δεδομένων που αφορούν κατηγορίες και όχι αριθμητικές τιμές, μετά και τις προσθήκες χαρακτηριστικών που έγιναν στα προηγούμενα ερωτήματα είναι οι παρακάτω ['Education','Marital_Status','Living_With']. Για αυτές τις μεταβλητές θα χρησιμοποιήσετε τον **LabelEncoder()** ώστε να μετατραπούν σε αριθμητικά δεδομένα (η διαδικασία ονομάζεται **one hot encoding**).

Στη συνέχεια δημιουργήστε ένα αντίγραφο του dataframe το οποίο θα περιέχει όλες τις αριθμητικές στήλες και διαγράψτε τις στήλες που σχετίζονται με προσφορές και προωθητικές ενέργειες, δηλαδή τις ['*AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2', 'Complain', 'Response*'].
Έτσι, τα δεδομένα που έχουν προκύψει περιέχουν χαρακτηριστικά διαφόρων διαστάσεων και διακυμάνσεων. Οι διαφορετικές διακυμάνσεις των χαρακτηριστικών των δεδομένων επηρεάζουν αρνητικά τη μοντελοποίηση ενός συνόλου δεδομένων. Η λύση είναι να γίνει αυτό που ονομάζεται *Standardization* έτσι ώστε κάθε στήλη/χαρακτηριστικό/μεταβλητή να έχει μ = 0 και σ = 1.
Τέλος χρησιμοποιήστε την μέθοδο συμπίεσης Principal Component Analysis (PCA) για να μειώσετε τις διαστάσεις σε n_components=3. Σχεδιάστε την (τρισδιάστατη) προβολή του αποτελέσματος.

8. **Υλοποίηση Συσταδοποίησης (Clustering) (25%)**

Βήματα

- Μέθοδος ELBOW για τον προσδιορισμό του αριθμού των συστάδων που πρέπει να σχηματιστούν
- Συσταδοποίηση μέσω συσσωρευτικής συσταδοποίησης (Agglomerative και K-Means)
- Εμφάνιση των σχηματιζόμενων συστάδων μέσω διαγράμματος (πχ scatter plot).

9. Προφίλ των πελατών (bonus)

Προσπαθήστε μέσα από διαγράμματα να σκιαγραφήσετε το προφίλ των συστάδων που σχηματίζονται ώστε να καταλήξετε σε ένα συμπέρασμα σχετικά με το ποιος είναι ο “σημαντικός” πελάτης και ποιος χρειάζεται περισσότερη προσοχή από την ομάδα μάρκετινγκ του καταστήματος.

Για να πετύχετε αυτό σχεδιάστε μερικά από τα χαρακτηριστικά που είναι ενδεικτικά των προσωπικών χαρακτηριστικών του πελάτη υπό το πρίσμα της συστάδας στην οποία βρίσκονται (πχ Ηλικία, Is_Parent, Family_Size κτλ). Τέλος για κάθε ένα από τα clusters συγκεντρώστε τα βασικά του χαρακτηριστικά. Πχ

Cluster 0:	**Cluster 1:**
Ξοδεύουν τα λιγότερα	Ξοδεύουν περισσότερα
Έχουν το μικρότερο εισόδημα	Έχουν μεγαλύτερο εισόδημα
Έχουν εφήβους στο σπίτι	Οι περισσότεροι δεν ειναι γονείς
Είναι πιο μεγάλοι σε ηλικία	Συμμετείχε ενεργά και στις 6 προωθητικές ενέργειες

Παραδοτέο:

Η εργασία μπορεί να εκπονηθεί ατομικά ή σε ομάδες **2 ατόμων**.

Θα ανεβάσετε στο eclass ένα φάκελο της μορφής sdixxxx. (όπου sdi το ΑΜ ενός εκ των ατόμων της ομάδας). Ο φάκελος sdixxxx της εργασίας είναι ο φάκελος στον οποίο θα έχετε ΜΟΝΟ τον κώδικά σας έτσι όπως περιγράφεται στη συνέχεια (δηλαδή δεν θα παραδώσετε εκ νέου τα δεδομένα εκπαίδευσης/δοκιμής).

Το παραδοτέο σας πρέπει να περιέχει **ΥΠΟΧΡΕΩΤΙΚΑ** ένα **Python notebook** με το οποίο θα μπορεί κάποιος να τρέξει την εργασία σας βήμα-βήμα. Στο notebook μπορείτε σε όποια σημεία κρίνετε απαραίτητο να εισάγετε **visualizations** με τον τρόπο που εξηγήσαμε στα φροντιστήρια ώστε να παρουσιάσετε και με ωραίο τρόπο τα αποτελέσματά σας. **Το notebook αποτελεί και την ολοκληρωμένη αναφορά** για την εργασία σας (δεν θα παραδώσετε τίποτα σε doc, pdf) , σχεδιάστε το με προσοχή, να θυμάστε να γράψετε μία περιγραφή σε κάθε βήμα για το τι ποιό ερώτημα απαντάται σε κάθε κελί.