Εξόρυξη δεδομένων και γνώσης

Χρήση τεχνικών εξόρυξης δεδομένων για πρόβλεψη πωλήσεων σε μικρομεσαίες επιχειρήσεις

Αντωνακάκης Βασίλειος Παναγιώτης 2119226

Καλαϊτζίδης Τριαντάφυλλος 2119211

Σαραφίδης Θεμιστοκλής 2118087

Εισαγωγή

Λόγω του μεγέθους και της σύνθεσής της, η Ελληνική οικονομία απαρτίζεται σχεδόν ολοκληρωτικά από μικρομεσαίες επιχειρήσεις (96,3% του συνόλου των επιχειρήσεων σύμφωνα με στοιχεία της EUROSTAT). Μεγάλο μέρος αυτού του ποσοστού αποτελούν μικρές επιχειρήσεις οι οποίες δραστηριοποιούνται στον χώρο των πωλήσεων προϊόντων. Συνήθως οικογενειακής διαχείρισης και συνοικιακής εμβέλειας, αυτές οι επιχειρήσεις αδυνατούν να έχουν πρόσβαση σε τεχνογνωσία και μέσα που θα μπορέσουν να βελτιώσουν και την αποδοτικότητά τους αυξάνοντας παράλληλα και το περιθώριο κέρδους τους. Οι λόγοι αυτής της αδυναμίας συνήθως είναι οικονομικής αλλά μπορεί να συμβάλουν και δευτερεύοντες λόγοι όπως ελλιπής ενημέρωση για τις νέες τεχνολογίες, ακαμψία στους τρόπους λειτουργίας της επιχείρησης ή και ανεπαρκής επαφή της επιχείρησης με την απαραίτητη ενσωμάτωση τεχνολογικών καινοτομιών στην επιχειρηματική πραγματικότητα του 21ου αιώνα. Στα πλαίσια αυτής της εργασίας θα εξετάσουμε την εύκολη εφαρμογή τεχνικών πρόβλεψης πωλήσεων, ένα πολύ κρίσιμο βήμα για μία μοντέρνα επιχείρηση. Ο λόγος για τον οποίο είναι σημαντική η σωστή και ακριβής πρόβλεψη πωλήσεων είναι διότι πάνω σε αυτές τις προβλέψεις μπορούν να βασιστούν περαιτέρω σημαντικές αποφάσεις οικονομικών, επιχειρηματικής στρατηγικής, προώθησης αλλά και προμηθειών κάθε μια εκ των οποίων περιέχει και κάποιο κόστος για την ίδια την επιχείρηση. Επομένως αν μπορεί να προβλεφθεί μια κάποια ασφάλεια η έκβαση μίας απόφασης με βάση τα ήδη υπάρχοντα δεδομένα, οι επιχειρήσεις μπορούν να λάβουν πιο γρήγορα και αποδοτικά αποφάσεις εξοικονομώντας χρήματα και πόρους που στην πορεία μπορούν να βοηθήσουν στην μεγέθυνση και ενδεχόμενη επέκταση της επιχείρησης.

Απαιτήσεις για την υλοποίηση του συστήματος προβλέψεων

Για την επιτυχία της λύσης του προβλήματος που εντοπίσαμε χρειαζόμαστε προηγούμενα δεδομένα πωλήσεων από την επιχείρηση που εξυπηρετούμε. Χρησιμοποιώντας αυτά μπορούμε να προβλέψουμε τις μελλοντικές κινήσεις της αγοράς, επιτρέποντας έτσι στην επιχείρηση να κάνει πιο ενημερωμένες κινήσεις.

Πιο συγκεκριμένα χρειαζόμαστε μια βάση δεδομένων στην οποία η επιχείρηση θα αποθηκεύει το ιστορικό των παραγγελιών του κάθε προϊόντος ξεχωριστά σε

συνάρτηση με την ημέρα και την ημερομηνία την οποία πουλήθηκε. Το αποτέλεσμα που θα παράξουμε θα λαμβάνει υπόψη όλες αυτές τις καταγραφές και θα προβλέπει ξεχωριστά ποια προϊόντα θα πουληθούν σε οποιαδήποτε ημερομηνία καθώς τη ποσότητα αυτών.

Προεπεξεργασία των δεδομένων

Κατά την συλλογή των δεδομένων θα πρέπει να εφαρμοστούν κάποιοι κανόνες μορφοποίησής τους έτσι ώστε να καταστεί δυνατή η χρήση τους από τους αλγόριθμους. Αρχικά οι πωλήσεις των προϊόντων θα καταγράφονται ως κωδικοί στην βάση δεδομένων. Μαζί θα καταγράφεται και η μέρα της αγοράς σε μορφή αριθμού από το 1 έως το 30-31 αναπαριστώντας κάθε μια από τις μέρες του μήνα καθώς και ο μήνας επίσης σε αριθμητική τιμή από 1 έως 12. Στην συνέχεια θα πρέπει να υπολογίζεται το πλήθος πωλήσεων του συγκεκριμένου αντικειμένου για κάθε μέρα. Στην συνέχεια, με την χρήση της πολλαπλής γραμμικής παλινδρόμησης θα μπορέσουμε να βρούμε την συσχέτιση του κάθε προϊόντος με την ημέρα πώλησης. Μέσω των αποτελεσμάτων θα μπορέσουμε να προβλέψουμε προσεγγιστικά τις πωλήσεις για την εβδομάδα.

Τεχνική εξόρυξης δεδομένων

Στη συνέχεια θα χρησιμοποιήσουμε την τελειοποιημένη Βάση δεδομένων για την παραγωγή γνώσης. Ο αλγόριθμος που θα χρησιμοποιηθεί είναι η Πολλαπλή Γραμμική Παλινδρόμηση (Multiple Linear Regression) ξεχωριστά για το κάθε προϊόν. Θα εφαρμόσουμε την τεχνική στη γλώσσα προγραμματισμού Python χρησιμοποιώντας τα πακέτα: Matplotlib, Numpy, Pandas και Sklearn.

Λίγα λόγια για τη τεχνική: Η Πολλαπλή Γραμμική Παλινδρόμηση είναι μια τεχνική στην οποία λαμβάνουμε υπόψη μας μια σειρά από χαρακτηριστικά (features) και μια εξαρτημένη μεταβλητή (target) για να παράξουμε μια γραμμή που εκφράζει τη γενική κίνηση της εξαρτημένης μεταβλητής εν συναρτήσει των ανεξάρτητων. Ο ορισμός της Γραμμής Πολλαπλής Γραμμικής Παλινδρόμησης P για χαρακτηριστικά $x_1, x_2, \dots, x_p$ ειναι: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$. Το τελικό αποτέλεσμα εκφράζει το πως η μέση τιμή μ αλλάζει εν συναρτήσει των χαρακτηριστικών. Έτσι, μπορούμε να μοντελοποιήσουμε την εμπορική συμπεριφορά του προϊόντος και να προβλέψουμε πόσες πωλήσεις θα έχει κατά μέσο όρο στην ίδια ημερομηνία.

Χαρακτηριστικά και Στόχος (Features and Target): Όπως προαναφέραμε τα χαρακτηριστικά της Βάσης μας είναι τα εξής: Ημέρα (Day), Μήνας (Month) και οι πωλήσεις του κάθε προϊόντος ξεχωριστά, τα οποία ευελπιστούμε και να

προβλέψουμε. Για τον σκοπό αυτόν, θα πρέπει να εφαρμόσουμε την τεχνική ξεχωριστά για κάθε προϊόν καθώς το κάθε προϊόν έχει διαφορετικές τάσεις πωλήσεων ανάλογα με την ημέρα και τον μήνα. Στη συνέχεια, αφού η εφαρμογή είναι η ίδια σε κάθε προϊόν, θα εξετάσουμε την εφαρμογή της τεχνικής για το προϊόν “Americano”.

Εφαρμογή:

Θα ξεκινήσουμε εισάγοντας τις απαραίτητες βιβλιοθήκες για την λειτουργία της εφαρμογής μας:

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import linear_model, metrics
import csv
import pandas as pd
```

Μετά θα πρέπει να εισάγουμε την βάση δεδομένων την οποία θα επεξεργαστούμε:

```
data = pd.read_csv('Bakery Sum Sales.csv')
```

Πριν εφαρμόσουμε την τεχνική, πρέπει να διαχωρίσουμε τα δεδομένα που θα χρησιμοποιήσουμε όπως αναφέραμε πιο πριν. Δηλαδή σε Χαρακτηριστικά και Στόχο (που σε αυτή την περίπτωση είναι οι πωλήσεις του “Americano”).

```
# defining feature matrix(X) and response vector(y)
X = data[['Day','Month']]
y = data['americano']
```

Για να μπορέσουμε να κρίνουμε το πόσο επιτυχημένο ήταν το εγχείρημά μας θα υιοθετήσουμε μια τεχνική που λέγεται Train/Test Split. Θα διαχωρίσουμε δηλαδή την Βάση Δεδομένων μας σε δυο κομμάτια. Ένα κομμάτι (το 60% των εισαγωγών) για την εκπαίδευση του μοντέλου στο οποίο η Μεταβλητή Στόχου (Target Variable) είναι ορατή και ένα κομμάτι (το 40% των εισαγωγών) για την δοκιμή του μοντέλου στο οποίο η Μεταβλητή Στόχου δεν είναι ορατή. Με αυτόν τον τρόπο επιτυγχάνουμε μια πρακτική δοκιμασία του μοντέλου μας ώστε να ξέρουμε πόσο καλά μπορεί να λειτουργήσει σε πραγματικές καταστάσεις.

```
# splitting X and y into training and testing sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.4, random_state=1)
```

Καλούμε το μοντέλο μας και εισάγουμε τα δεδομένα εκπαίδευσης.

```
# create linear regression object
reg = linear_model.LinearRegression()

# train the model using the training sets
reg.fit(X_train, y_train)
```

Και για να τελειοποιήσουμε το μοντέλο μας, κάνουμε μια σειρά από μετρήσεις.

```
# regression coefficients
print('Coefficients: ', reg.coef_)

# variance score: 1 means perfect prediction
print('Variance score: {}'.format(reg.score(X_test, y_test)))

r_sq = reg.score(X_train, y_train)
print(f"coefficient of determination: {r_sq}")
```

Αποτελέσματα:

Coefficients: [-0.00072963 0.32523976]

Variance score: -0.20838884822708348

coefficient of determination: 0.06232276367254075

Παράδειγμα πρόβλεψης:

```
Enter a Date value for prediction
Day (1-30): 25
Month (1-12): 8
Predicted value:
[1.51822343]
```

Πλεονεκτήματα για τις επιχειρήσεις

Έχοντας αναφέρει παραπάνω την πολλαπλή γραμμική παλινδρόμηση θα πάρουμε αρχικά παραδείγματα πάνω σε αυτή και έπειτα γενικά παραδείγματα. Σκεφτόμαστε έστω την μικρο-μεσαία επιχείρηση μας. Εφαρμόζοντας την παλινδρόμηση θα μπορούσαμε εύκολα να εντοπίσουμε συσχετίσεις στα δεδομένα, που στη συνεχεία

θα μετατρέψουμε σε χρήσιμες πληροφορίες για την ανάληψη σημαντικών αποφάσεων. Βλέποντας σε ποια περιοχή κάθε προϊόν τις επιχείρησης πουλάει παραπάνω, θα μπορούσαμε να αυξήσουμε τα κέρδη. Παρόμοιος και για τις περιοχές που δεν υπάρχει αρκετή ζήτηση σε κάποια από τα προϊόντα η τις υπηρεσίες θα μπορούσαμε να αποφύγουμε ζημιά. Όλα αυτά με τη χρησιμοποίηση της μεθόδου παλινδρόμησης.

Ένα άλλο παράδειγμα, εφαρμόζοντας μια από τις πολλές τεχνικές εξόρυξης δεδομένων, θα ήταν εντοπισμός προϊόντων η υπηρεσιών που αγοράζονται μαζί. Δηλαδή των εντοπισμό κάποιων μοτίβων που με το μάτι δεν εντοπίζονται εύκολα. Υπάρχουν πολλά τέτοια μοντέλα που είτε κατηγοριοποιούν τα δεδομένα τις επιχείρησης είτε τα ομαδοποιούν. Το κομμάτι που μας ενδιαφέρει είναι πως μέσα από τον εντοπισμό αυτών των μοτίβων θα μπορούσαμε εύκολα χρησιμοποιήσουμε αυτή την γνώση για ένα στοχευμένο επιθετικό μάρκετινγκ.

Ιδιαίτερη προσοχή κατά τον εντοπισμό μοτίβων και συσχετίσεων μεταξύ μεταβλητών καθώς πάντα υπάρχει η περίπτωση λάθους λόγο των έκτοπων τιμών (outliers). Εύκολα θα μπορούσαμε να οδηγηθούμε σε λάθος αποτελέσματα, πιο συγκεκριμένα η μέθοδος της πολλαπλής γραμμικής παλινδρόμησης είναι αρκετά ευαίσθητη προς τις ακραίες τιμές, αλλά με ένα σωστό καθαρισμό των δεδομένων και μια μικρή προεπεξεργασία θα οδηγήσει σε πιο ορθά αποτελέσματα.

Τέλος, αν ακολουθήσουμε τακτικές μειώσεις των διαστάσεων των δεδομένων, θα οδηγηθούμε με μειωμένα δεδομένα. Άρα μείωσης του αποθηκευτικού χώρου που χρειάζονται και χρόνου για τη συντήρηση τους αλλά και οργάνωσης τους. Εύκολα αντιλαμβανόμαστε πως έτσι μειώνονται τα έξοδα για αποθήκευση, πόροι για την επεξεργασία τους και εξαγωγή πληροφορίας. Πράγματα πολύ χρήσιμα για μια μικρό μεσαία επιχείρηση καθώς δεν υπάρχει αφθονία στο κεφάλαιο και ο χρόνος θα πρέπει να εστιάζεται στην εξέλιξη των μελλοντικών παροχών και διατήρησης ενός ανταγωνιστικού πνεύματος.

Πηγές

Αρχικό dataset προ προεπεξεργασίας:

https://www.kaggle.com/datasets/hosubjeong/bakery-sales?select=Bakery+Sales.csv

Αποθετήριο εργασίας στο GitHub:

https://github.com/VassilisAntonakakis/dataMiningProject

Θεωρητικό υπόβαθρο για την πρόβλεψη πωλήσεων:

https://www.theseus.fi/bitstream/handle/10024/106191/Haataja_Timo.pdf?sequence=1&isAllowed=y

http://repository.library.teimes.gr/xmlui/bitstream/handle/123456789/7398/%ce%97%20%ce%a3%ce%97%ce%9c%ce%91%ce%a3%ce%99%ce%91%20%ce%a4%ce%97%ce%a3%20%ce%a0%ce%a1%ce%9f%ce%92%ce%9b%ce%95%ce%a8%ce%97%ce%a3%20%ce%a0%ce%a9%ce%9b%ce%97%ce%a3%ce%95%ce%a9%ce%9d%20%ce%93%ce%99%ce%91%20%ce%a4%ce%91%20%ce%a3%ce%a4%ce%95%ce%9b%ce%95%ce%a7%ce%97%20%ce%95%ce%a0%ce%99%ce%a7%ce%95%ce%99%ce%a1%ce%97%ce%a3%ce%95%ce%a9%ce%9d..pdf?sequence=1&isAllowed=y

http://repository.library.teiwest.gr/xmlui/bitstream/handle/123456789/3231/%CE%9C%CE%99%CE%9A%CE%A1%CE%9F%CE%9C%CE%95%CE%A3%CE%91%CE%99%CE%95%CE%A3%20%CE%95%CE%A0%CE%99%CE%A7%CE%95%CE%99%CE%A1%CE%97%CE%A3%CE%95%CE%99%CE%A3%20%CE%A3%CE%A4%CE%97%CE%9D%20%CE%95%CE%9B%CE%9B%CE%91%CE%94%CE%91%20%CE%9A%CE%91%CE%99%20%CE%A3%CE%A4%CE%97%CE%9D%20%CE%95%CE%A5%CE%A1%CE%A9%CE%A0%CE%91%CE%99%CE%9A%CE%97%20%CE%95%CE%9D%CE%A9%CE%A3%CE%97%20%CE%93%CE%95%CE%9D%CE%99%CE%9A%CE%9F%CE%A4%CE%95%CE%A1%CE%91%20%CE%9A%CE%91%CE%99%20%CE%97%20%CE%95%CE%A0%CE%99%CE%94%CE%A1%CE%91%CE%A3%CE%97%20%CE%A4%CE%97%CF%82%20%CE%9F%CE%99%CE%9A%CE%9F%CE%9D%CE%9F%CE%9C%CE%99%CE%9A%CE%97%CE%A3%20%CE%9A%CE%A1%CE%99%CE%A3%CE%97%CE%A3%20%CE%A3%CE%A4%CE%97%CE%9D%20%CE%95%CE%A0%CE%99%CE%94%CE%9F%CE%A3%CE%97%20%CE%A4%CE%9F%CE%A5%CE%A3.pdf?sequence=1&isAllowed=y