**MACHINE LEARNING**

Ιπποκράτης Κοτσάνης

ΑΜ: 131

**ΤΕΛΙΚΗ ΕΡΓΑΣΙΑ – REPORT**

* Αρχικά φορτώνουμε το csv dataset Walmart cleaned, το οποίο μετατρέπουμε σε pandas Dataframe, αφού έχουμε φορτώσει πρώτα τις απαραίτητες βιβλιοθήκες που θα χρησιμοποιήσουμε.

Text

Description automatically generated

* Στην συνέχεια θα εξερευνήσουμε τα δεδομένα μας που έχουμε διαθέσιμα στο Dataset. (Data exploration)

Text

Description automatically generated

Πιο συγκεκριμένα με βάση τον τύπο που δόθηκε στην εκφώνηση το κατάστημα Store που θα μελετήσουμε είναι: 131 mod 45 + 1 = 42.

Επομένως θα περιορίσουμε το dataset στο Store 42 και στην συνέχει με την βοήθεια της εντολής “iloc” θα διαγράψουμε το πρώτο column του dataframe, το οποίο είναι unnamed.

Στην συνέχεια θα υπολογίσουμε τον correlation matrix για να δούμε ποια χαρακτηριστικά είναι τα πιο σημαντικά στο dataset μας. Τα χαρακτηριστικά που είναι ασήμαντα σε επόμενο βήμα θα τα διαγράψουμε.

Παρακάτω βλέπουμε τον correlation matrix:

A picture containing chart

Description automatically generated

Παρατηρούμε ότι τα χαρακτηριστικά isHoliday, Dept και Weekly\_Sales έχουν το πιο υψηλό correlation.

* Θα συνεχίσουμε με την διαδικασία που λέγεται Data Manipulation.

Στην διαδικασία αυτή θα κάνουμε τα εξής βήματα:

1. Θα εξερευνήσουμε και θα αφαιρέσουμε τις null τιμές.
2. Θα φτιάξουμε 2 columns για το year και τον αριθμό της εβδομάδας, παίρνοντας τις πληροφορίες από το column Date.
3. Θα αφαιρέσουμε τα χαρακτηριστικά με μικρό correlation.

Text

Description automatically generated

A picture containing table

Description automatically generated

Όπως μπορούμε να δούμε δεν υπάρχουν δεδομένα με null τιμές, εκτός από το Next week, το οποίο ακόμα δεν έχει συμπληρωθεί.

Παρακάτω θα βρούμε τα χαρακτηριστικά που θα έχουμε ως είσοδο (inputs) και ως στόχο (targets).

Text

Description automatically generated

Αφού βρούμε τα inputs και targets θα χωρίσουμε τα χαρακτηριστικά μας σε αριθμητικά (numeric) και αυτά που περιέχουν κάποια κατηγορία (categorical). Τα numeric columns θα τα συμπληρώσουμε όπου χρειάζεται και θα κάνουμε scale. Aυτό το κάνουμε διότι μπορούμε να χρησιμοποιήσουμε μόνο numeric data για να κάνουμε train τα μοντέλα μας.

Κατά την εκτέλεση, παρατηρούμε ότι δεν έχουμε categorical columns. Στην περίπτωση που είχαμε θα έπρεπε να κάνουμε one hot encoding το οποίο υπάρχει ως σχόλιο στο project.

Text

Description automatically generated

Στην παραπάνω εικόνα βλέπουμε τον χωρισμό του dataset σε 3 departments. Επιλέγουμε τα τμήματα 1, 10 και 11.

Η μεταβλητή select\_department χρησιμοποιείται για την επιλογή του department, κάθε φορά που τρέχει ο αλγόριθμος.

Εισάγουμε στο column Next week τις τιμές του Weekly\_sales της επόμενης εβδομάδας.

Στην συνέχεια χωρίζουμε το dataset του department που έχει επιλεχθεί σε training set για την χρονιά 2010, validation set για την χρονιά 2011 και testing set για την χρονιά 2012.

* Διαδικασία Training και Models selection.

Text

Description automatically generated

Αφού χωρίσουμε τα δεδομένα μας θα χρησιμοποιήσουμε 2 μοντέλα. Ανάλογα με τα αποτελέσματα του καθενός θα χρησιμοποιήσουμε αυτό με τα καλύτερα. Τα μοντέλα είναι Decision Tree και Random Forest.

1. Decision Tree

Ένα δέντρο αποφάσεων είναι ένα εργαλείο υποστήριξης αποφάσεων που χρησιμοποιεί ένα μοντέλο αποφάσεων που μοιάζει με δέντρο και τις πιθανές συνέπειές τους, συμπεριλαμβανομένων των αποτελεσμάτων τυχαίων συμβάντων, του κόστους πόρων και της χρησιμότητας. Είναι ένας τρόπος εμφάνισης ενός αλγόριθμου που περιέχει μόνο εντολές ελέγχου υπό όρους.

Text

Description automatically generated

Kάνουμε train το μοντέλο μας κατά τα γνωστά και στην συνέχεια εκτυπώνουμε το decision tree, καθώς και ένα γράφημα με την σημαντικότητα των χαρακτηριστικών.

Decision tree visualization:

Diagram

Description automatically generated

Chart, bar chart

Description automatically generated

Τα χαρακτηριστικά Weekly\_Sales, Week και unemployment είναι τα πιο σημαντικά.

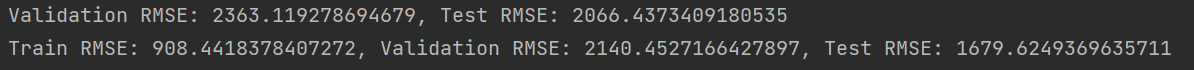
1. Random Forest

Τα τυχαία δάση είναι μια μέθοδος εκμάθησης συνόλου για ταξινόμηση, παλινδρόμηση και άλλες εργασίες που λειτουργεί κατασκευάζοντας ένα πλήθος δέντρων αποφάσεων κατά το χρόνο εκπαίδευσης. Για εργασίες ταξινόμησης, η έξοδος του τυχαίου δάσους είναι η κλάση που επιλέγεται από τα περισσότερα δέντρα. Για εργασίες παλινδρόμησης, επιστρέφεται ο μέσος όρος ή ο μέσος όρος πρόβλεψης των μεμονωμένων δέντρων. Για να δημιουργήσουμε το μοντέλο του τυχαίου δάσους, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση RandomForestRegressor.

Text

Description automatically generated

Υπολογίζοντας με την βοήθεια της μετρικής RMSE τα score συμπεραίνουμε ότι το μοντέλο Random forest είναι καλύτερο.



* Hyperparameter tuning

Ας ορίσουμε μια βοηθητική συνάρτηση test\_params που μπορεί να ελέγξει τη δεδομένη τιμή μιας ή περισσότερων υπερπαραμέτρων. Για αυτό το νέο τυχαίο μοντέλο, θα χρησιμοποιήσουμε έναν αριθμό εκτιμητών 100.

Ας ορίσουμε επίσης μια βοηθητική συνάρτηση για τον έλεγχο και τη σχεδίαση διαφορετικών τιμών μιας μεμονωμένης παραμέτρου. (test\_param\_and\_plot)

Επίσης φτιάχνουμε 2 ειδών συναρτήσεις μία για το validation set (2011) και μία για το testing set (2012).

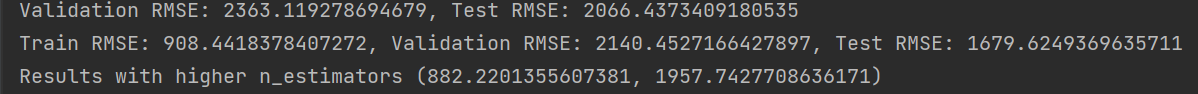
Text

Description automatically generated

Text

Description automatically generated

Αρχικά αυξάνοντας τους εκτιμητές μπορούμε να δούμε καλύτερα αποτελέσματα. Στην αρχή το μοντέλο μας έτρεξε με 10 και τώρα με 100.



Παράμετρος: min\_samples\_leaf

Chart, line chart

Description automatically generated

Text

Description automatically generated

Εδώ, μπορούμε να δούμε πώς αυξάνεται το RMSE με την παράμετρο min\_samples\_leaf, οπότε θα χρησιμοποιήσουμε την προεπιλεγμένη τιμή (1).

Παράμετρος: max\_leaf\_nodes

Chart, line chart

Description automatically generated

Text

Description automatically generated

Το RMSE μειώνεται με την παράμετρο max\_leaf\_nodes, επομένως θα χρησιμοποιήσουμε την προεπιλεγμένη τιμή (none).

Παράμετρος: max\_depth

Chart

Description automatically generated

Text

Description automatically generated

Το RMSE μειώνεται με την παράμετρο max\_depth, επομένως θα χρησιμοποιήσουμε την προεπιλεγμένη τιμή (none).

O κώδικας παρουσιάζεται παρακάτω:

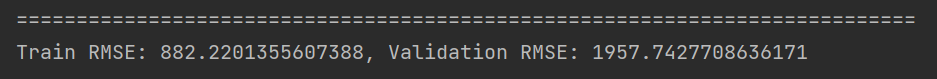
Text

Description automatically generated

* Training το καλύτερο μοντέλο

Text

Description automatically generated



Και στην συνέχεια βρίσκουμε τα πιο σημαντικά χαρακτηριστικά.

Chart, bar chart

Description automatically generated

Τέλος είμαστε έτοιμοι να κάνουμε τα predictions μας με βάση το validation set, δηλαδή για το 2011.

Text

Description automatically generated

Για λόγους στατιστικής ανάλυσης αλλά και συμπερασμάτων τα predictions τα εισάγουμε σε ένα νέο column που λέγεται “Next week predictions”.

Ομοίως δουλεύουμε και για το testing set, προκειμένου να κάνουμε predictions για το 2012.

Text

Description automatically generated

* Results – comments

Για κάθε department, όπως δείξαμε γίνεται πρόβλεψη για το 2011 και για το 2012. Ας δούμε, λοιπόν, πώς τα πήγαμε. Στα γραφήματα παρακάτω παρουσιάζονται οι πραγματικές τιμές (μπλε) μαζί με τις predictions (κόκκινο) για κάθε department.

Department: 1

Graphical user interface, chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Department: 10

Graphical user interface, chart

Description automatically generated

Graphical user interface, chart

Description automatically generated

Department: 11

Graphical user interface, chart, histogram

Description automatically generated

Graphical user interface, chart

Description automatically generated

Όπως μπορούμε να δούμε από τα παραπάνω γραφήματα οι προβλέψεις για τα τμήματα 1 και 11 είναι αρκετά καλές, ενώ για το τμήμα 10 ικανοποιητικές με περιθώρια βελτίωσης. Γενικά το μοντέλο μας θα μπορούσε κανείς να πει ότι πήγε σε γενικές γραμμές αρκετά καλά.

* CSV

Text

Description automatically generated

Δημιουργούμε 2 csv για το κάθε department, ένα για το έτος 2011 και ένα για το έτος 2012.