**Ερώτημα 2)**

Εδώ μας ζητείται με βάση τα σχόλια που έχουν κάνει πελάτες της Amazon να προβλέψουμε τη βαθμολογία που έχουν βάλει σε αυτό το προϊόν χρησιμοποιώντας έναν Random Forest.

Τα score κυμαίνονται από 1 έως 5 οπότε έχουμε 5 κολάσεις.

Τα σχόλια θα τα μετατρέψουμε σε διανύσματα χρησιμοποιώντας το word2vec μοντέλο της Google το οποίο όμως εφαρμόζεται σε μεμονωμένες λέξεις για αυτό εμείς θα υπολογίσουμε τον μέσο ορό για όλο το σχόλιο.

Text

Description automatically generatedΑρχικά βλέπουμε τη δομή των δεδομένων μας

Στα οποία όπως παρατηρήσαμε υπάρχουν κάποιες δίπλες καταχωρήσεις ~700 τις οποίες και αφαιρούμε.

Η κατανομή είναι:

Chart, histogram

Description automatically generated

Βλέπουμε ότι ακολουθούν τη J-κατανομή όπως σχεδόν όλα τα dataset με κριτικές προϊόντων.

Τα δεδομένα μας είναι πάρα πολύ ανισόρροπα με ~30k για τη κλάση 5 και ~3κ για τη κλάση 2 για αυτό πρέπει να τα εξισορροπήσουμε “κάπου στη μέση” προκειμένου κάθε κλάση να έχει παρόμοιο αριθμό δειγμάτων για να γίνει καλυτέρα το training.

Αρχικά χωρίζουμε τα δεδομένα σε train – test καθώς δεν μας επηρεάζει εάν στο testing τα δεδομένα είναι imbalanced.

Στα train data έχουμε τα εξής πλήθη κλάσεων.

A screenshot of a computer

Description automatically generated with low confidence

Οπότε επιλέγουμε να κάνουμε resampling σε όλες τις κολάσεις στα 5000 δείγματα.

Αυτό σημαίνει ότι για τις κλάσεις 4,5 θα απορρίψουμε δείγματα ενώ για τις υπόλοιπες θα χρησιμοποιήσουμε κάποια δείγματα και 2η φορά.

Συγκεκριμένα για τη κλάση 2 θα υπερδιπλασιάσουμε το πλήθος της καθώς είναι πολύ μικρή σε σύγκλιση με τις υπόλοιπες.

Συνολικά θα έχουμε 25000 training samples.

Για να μετατρέψουμε τα κείμενα σε διανύσματα ακολουθούμε 2 βήματα:

- Σε κάθε σχόλιο αφαιρούμε τα σημεία στίξης, κάνουμε lowercase τα γράμματα και σπάμε τις λέξεις σε μια λίστα.

- Κάθε λέξη την κάνουμε διάνυσμα με το word2vex και στο τέλος υπολογίζουμε τον μέσο ορό .

Τα δεδομένα μας έχουν διάσταση (n\_samples, 300).

Στον Random Forest Classifier επιλέγουμε 400 δέντρα και σαν κριτήριο κατηγοριοποίησης την εντροπία και μετρά το train παίρνουμε τα αποτελέσματα.

Στα test data φτιάχνουμε το confusion matrix

Calendar

Description automatically generated

Table, calendar

Description automatically generatedΥπολογίζουμε και τα precision/recall/f1-score

Το μοντέλο παρουσιάζει 73.5% accuracy και το f1 score κυμαίνεται από 0.571 μέχρι 0.879 για τις μεταβλητές.

Παρατηρούμε επίσης στη κλάση 5 ότι αρκετά predictions αντιστοιχίζονται στην 4, λογικό τα σχόλια μοιάζουν αρκετά αλλά αρκετά παραδόξως αντιστοιχίζονται στην κλάση 1

Ο κώδικας βρίσκεται εδώ: <https://colab.research.google.com/drive/1p4S_Ml7ebUaZxFkCSh33Pty9HMKybx5e?usp=sharing>