ΠΑΠΑΔΟΠΟΥΛΟΣ ΝΙΚΟΛΑΟΣ 4140

ΤΡΙΤΟ ΣΕΤ ΑΣΚΗΣΕΩΝ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ
ΑΝΑΦΟΡΑ 1ΗΣ ΑΣΚΗΣΗΣ

ΒΗΜΑ 1

Τα index για τα clusters μου έχουν ως εξής:

- 0 - Beauty & Spas
- 1 - Shopping
- 2 - Bars

Παρατηρήσεις για τις μεθόδους και την αποτελεσματικότητά τους:

Logistic regression:

Το classifier score είναι πολύ ψηλό, που σημαίνει ότι οι περισσότερες προβλέψεις γενικά είναι σωστές. Απ'το confusion matrix βλέπουμε ότι από τις (307 + 14) = 321 επιχειρήσεις που ανήκουν στο cluster 0, οι 307 από αυτές προβλέφθηκαν να ανήκουν στο cluster 0. Αυτό μας ρίχνει λίγο το recall score από τη μονάδα, όμως έχουμε precision score σχεδόν 1, καθώς από τις προβλέψεις για το cluster 0, μόλις μία έχει να κάνει με αντικείμενο που δεν ανήκει στο cluster 0. Το cluster 2 έχει και αυτό πολύ καλά scores, τα οποία πέφτουν λίγο από τη μονάδα επειδή από τις επιχειρήσεις που ανήκουν σε αυτή την ομάδα δεν προβλέφθηκε ότι ανήκουν οι 13, ενώ έγιναν 14 λάθος κατηγοριοποιήσεις σε αυτή την ομάδα από άλλα είδη επιχειρήσεων. Τέλος, μόνο μία από τις επιχειρήσεις που ανήκουν στο cluster 3 δεν κατηγοριοποιήθηκαν σε αυτό, το οποίο δίνει πολύ ψηλό recall score γι'αυτή την κλάση.

Οι καλές επιδόσεις του αλγόριθμου για όλες τις κλάσεις αποτυπώνονται στα αρκετά καλά average precision & recall scores. Φυσικά, το f1 score που συνδυάζει τα παραπάνω με τον τύπο f1 = (2 * recall * precision) / (recall + precision) κυμαίνεται σε ανάλογα επίπεδα.

SVM:

Για αυτή τη μέθοδο έχουμε παρόμοια καλές επιδόσεις. Για το cluster 0, οι 308 από τις 321 επιχειρήσεις που ανήκουν σε αυτό γίνονται εν τέλει clustered σε αυτό, με τις υπόλοιπες 13 να πάνε στο 2ο cluster. Έχουμε μόλις δύο λάθος κατηγοριοποιήσεις σε cluster 0, το οποίο οδηγεί σε precision score σχεδόν ίσο με τη μονάδα. Όσον αφορά άλλα δύο clusters, παρατηρούμε γενικά παρομοίως καλό clustering. Το

υψηλότερο recall score το έχει το cluster 3, το οποίο σημαίνει ότι οι επιχειρήσεις που ανήκουν σε αυτό κατατάχθηκαν εν τέλει στο ίδιο cluster σε βαθμό μεγαλύτερο από κάθε άλλο cluster. Τα average precision και recall scores είναι σχεδόν τα ίδια με αυτά της μεθόδου logistic regression. Το ίδιο ισχύει και για το f1-score, το οποίο πέφτει μονάχα κατά περίπου 1%.

K-NN:

Περαιτέρω 1% πτώση παρατηρείται στο score της μεθόδου “K-nearest-neighbors”. Τα precision και recall scores της είναι σχεδόν ίδια μεταξύ τους, με ό,τι αυτό συνεπάγεται όπως έχω εξηγήσει προηγουμένως.

Naive Bayes:

Μακράν η χειρότερη απόδοση από τις 5 μεθόδους, με average f1 score στο 0.85 περίπου. Ιδιαίτερο πρόβλημα φαίνεται, από το confusion matrix, να υπάρχει στο cluster 1, αφού υπερβολικά πολλές επιχειρήσεις αυτού του cluster αποφασίζεται να ανήκουν στο 2.

Decision Trees:

Η μέθοδος αυτή έχει τη δεύτερη χειρότερη απόδοση από τις μεθόδους που δοκιμάζουμε. Η συνολική ακρίβεια πέφτει στο 0.92%. Ο κύριος παράγοντας στον οποίο οφείλεται το γεγονός αυτό είναι το clustering για το cluster 1: οι μετρικές precision και recall πέφτουν κάτω από το 90%, το οποίο σημαίνει ότι αρκετά στοιχεία που στην πραγματικότητα ανήκουν σε αυτό το cluster δεν κατηγοριοποιούνται σε αυτό μέσω της μεθόδου, ενώ παράλληλα αρκετά στοιχεία από άλλα clusters κατηγοριοποιούνται σε αυτό λανθασμένα. Αυτό επηρεάζει αρνητικά τα average precision,recall και f1 score φυσικά.

5 - fold cross validation:

Η μέθοδος αυτή κάνει shuffle τα δεδομένα και τα χωρίζει σε 5 ίσα set, δοκιμάζοντας το κάθε set ως test set. Αυτό την καθιστά ως πολύ καλό μέτρο αξιολόγησης των μεθόδων.

Η μέθοδος logistic regression αξιολογείται σχεδόν το ίδιο καλά με προηγουμένως, πέφτοντας μόνο κατά 1% στην επίδοση. Αυτό την καθιστά αρκετά καλή. Το ίδιο ακριβώς ισχύει και για τη μέθοδο K-nn, η οποία συνολικά είναι κατά 1% λιγότερο αποδοτική από την LR.

Η D-tree μέθοδος λαμβάνει σχεδόν την ίδια αξιολόγηση με αυτή τη μέθοδο.

Τέλος, κατά 1% σε αξιολόγηση πέφτει και η SVM σε αυτή τη διαδικασία, διατηρώντας, ωστόσο, πολύ καλά ποσοστά.

Μπορούμε να εξάγουμε το συμπέρασμα ότι οι δύο πιο αποδοτικές μέθοδοι για τη συγκεκριμένη classification εφαρμογή είναι οι SVM και Logistic regression.

ΒΗΜΑ 2

	TF-IDF VECTORISER	GOOGLE VECTORISER
Logistic Regression	0.976	0.956
SVM	0.975	0.962
K-NN	0.963	0.968
Naive Bayes	0.855	0.944
Decision Trees	0.919	0.897

Παραπάνω παραθέτω τα classifier scores που πετυχαίνει κάθε μέθοδος με τα vectors της αναπαράστασης TF-IDF και του μοντέλου της google. Τα αποτελέσματα δεν είναι αυτά που περίμενα, καθώς δύο από τις τρεις μεθόδους (LR, SVM, D-Trees) έπεσαν λίγο σε απόδοση, ενώ μονάχα η μέθοδος Naive Bayes (σημαντικά) και η K-NN (αμοιδρά) βοηθήθηκαν. Θα περίμενε κανείς, ίσως, γενική βελτίωση από ένα μοντέλο που του έχει ασκηθεί training σε τέτοιο όγκο δεδομένων που του παρέχει η Google.