30 MAYIS 2017

ISTANBUL TECHNICAL UNIVERSITY

BLG 454E LEARNING FROM DATA

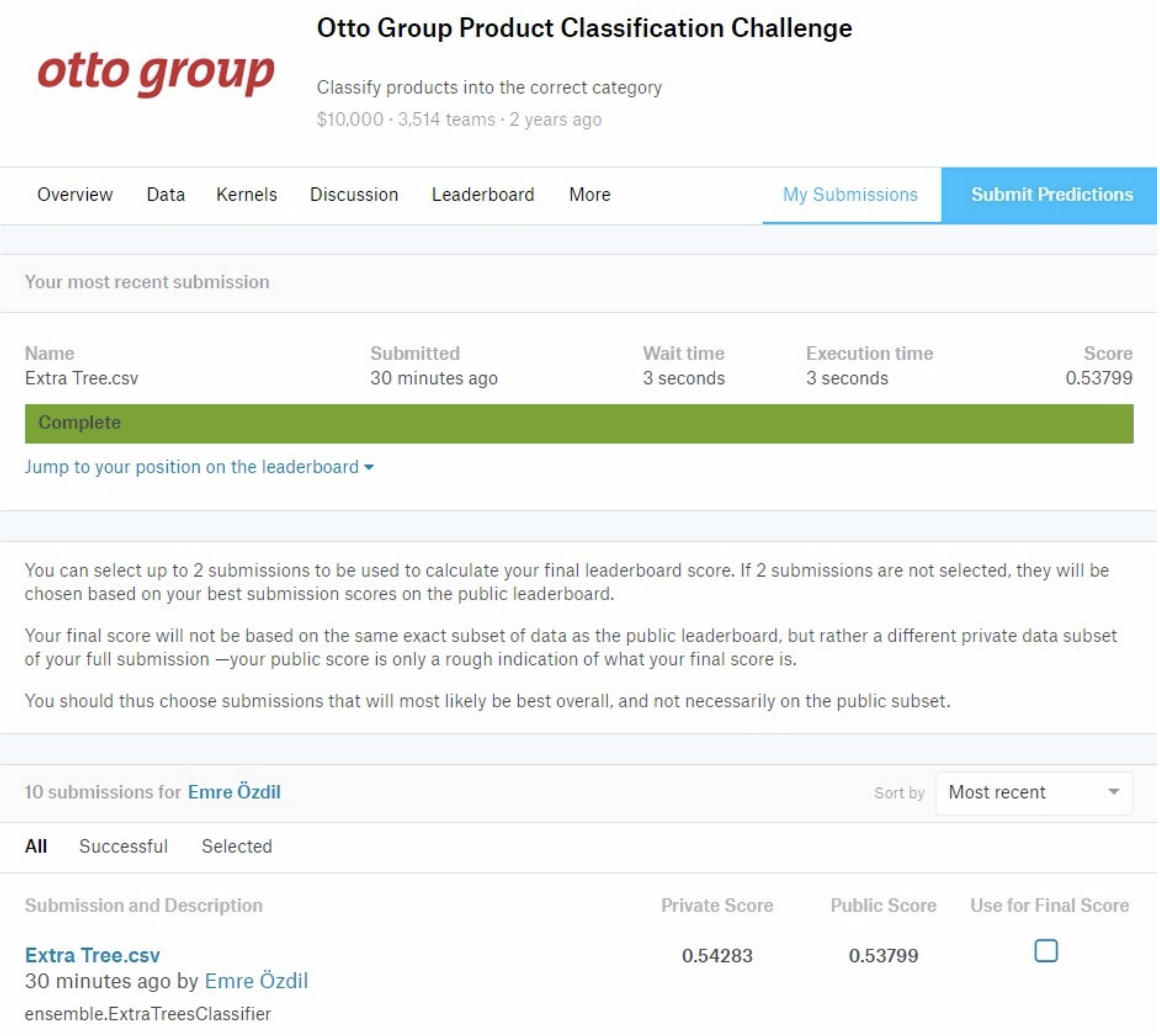
TERM PROJECT

TEAM: BEE MEISTER
EMRE ÖZDİL – 150120138
MERVE ECEVİT – 150140115

Kaggle Name

Our final submission is done from https://www.kaggle.com/emreozdil this account.

Kaggle Score



Otto Group Product Classification Challenge

Classify products into the correct category

$10,000 · 3,514 teams · 2 years ago

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
Extra Tree.csv	30 minutes ago	3 seconds	3 seconds	0.53799

Complete

Jump to your position on the leaderboard ▾

You can select up to 2 submissions to be used to calculate your final leaderboard score. If 2 submissions are not selected, they will be chosen based on your best submission scores on the public leaderboard.

Your final score will not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission —your public score is only a rough indication of what your final score is.

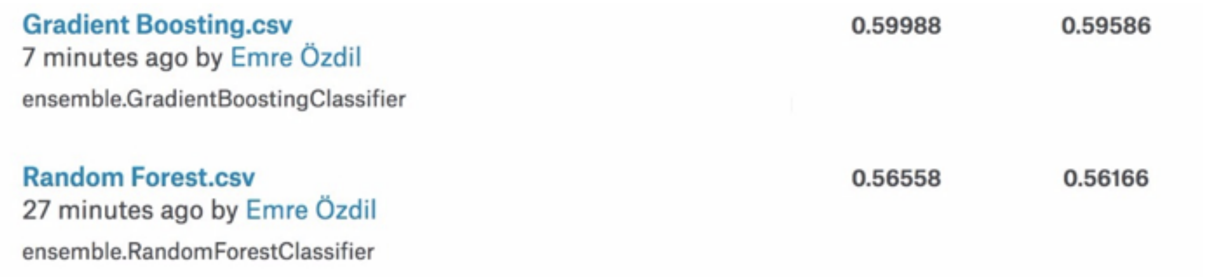
You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

10 submissions for Emre Özdil   Sort by Most recent ▾

All   Successful   Selected

Submission and Description	Private Score	Public Score	Use for Final Score
Extra Tree.csv 30 minutes ago by Emre Özdil ensemble.ExtraTreesClassifier	0.54283	0.53799	□

As can be seen from above, our best Kaggle score is 0.53799 and it is submitted on 31 May. Our rank is 1703. The other methods that we have used can be seen in below:



Gradient Boosting.csv 7 minutes ago by Emre Özdil ensemble.GradientBoostingClassifier	0.59988	0.59586
Random Forest.csv 27 minutes ago by Emre Özdil ensemble.RandomForestClassifier	0.56558	0.56166

Dataset Description

In the given dataset, each row is a single product. There are 93 numerical features to represent counts of different events. There are nine class for all products. Each target category represents one of our most important product categories. The products for the training and testing sets are selected randomly.

File descriptions

- trainData.csv - the training datasets
- testData.csv - the test datasets
- sampleSubmission.csv – an example of submission file

Data fields

- id - unique id for product
- feat_1, feat_2, ..., feat_93 - features of a product
- target - class of the product

Methods Used

We have built a predictive model which is able to distinguish between 9 product categories. We tried different methods. However, Extra Trees Classifier is the best rank that we have get. We used ensemble methods which combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability over a single estimator. There are two types of ensemble methods which are averaging and boosting methods. Random Forest Classifier and Extra Trees Classifier are a kind of averaging method. Gradient Boosting is a kind of boosting method. The main difference between averaging and boosting method is the building style of base estimators. In averaging method, several estimators are created and the average of their predictions is found. In boosting method, base estimators are built sequentially and one of them tries to reduce the bias of the combined estimator.

Experiment Results

The results of the different classifier methods are as following:

- Extra Trees -> 0.53799
- Random Forest -> 0.56166
- Gradient Boosting -> 0.59586

Discussion

We have tried Extra Trees Classifier, Random Forest Classifier and Gradient Boosting Classifier. We took our best ranking in Extra Trees Classifier which is 1703.