O2 - Study case - YELP Dataset

Sentiment Analysis

J. Korecek

O2

Prague, October 2020

Dataset

The dataset[1] consist of 5 files with around 10 GB.

- **business** - Contains business data including location data, attributes, and categories.
- **review** - Contains full review text data including the user_id that wrote the review and the business_id the review is written for.
- **user** - User data including the user's friend mapping and all the metadata associated with the user.
- **checking** - Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions
- **tip** - Contains photo data including the caption and classification (one of "food", "drink", "menu", "inside" or "outside").

[1] https://www.yelp.com/dataset/documentation/main

Sentiment analysis

Loading, reprocessing etc...

Sentiment analysis is **classification** problem. We use *Logistic Regression, Random Forest and XGBoost* to find model predicting positive or negative sentiment.

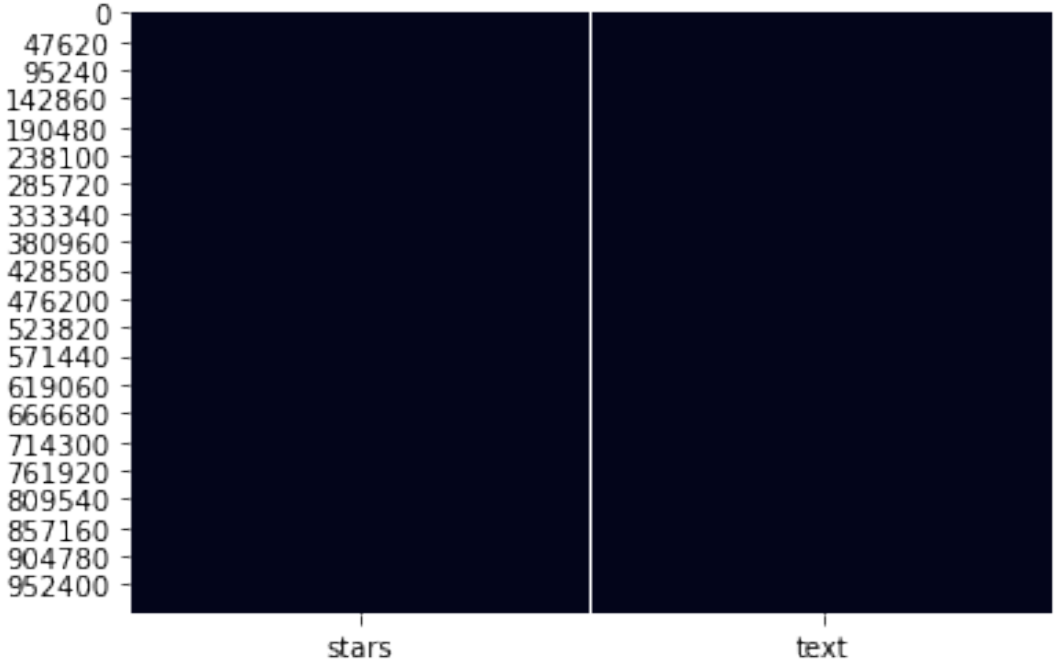
Size

Since there are hardware restrictions, we use 1 mil. rows of the data. Selecting two columns *stars, text*. Stars has values from 1 - 5 indicating , 5 as best and 1 as worst.

Sentiment analysis

Descriptive statistics

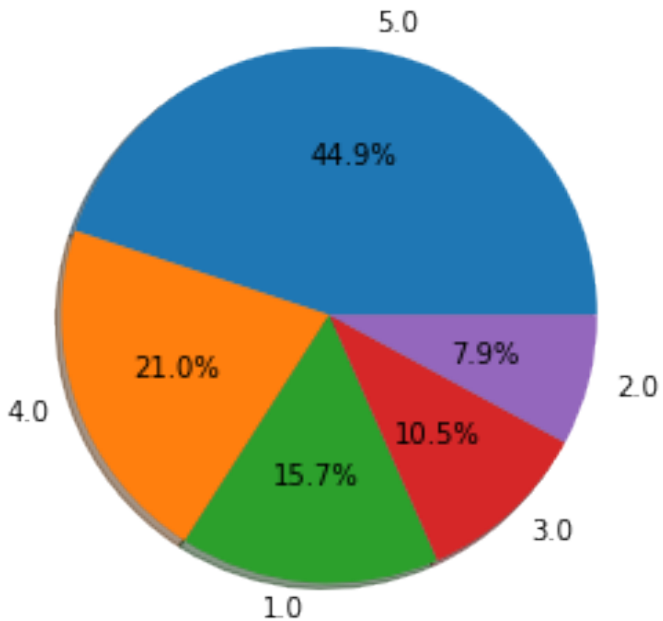
The out layers are not present as well as no missing values.



Sentiment analysis

Distribution

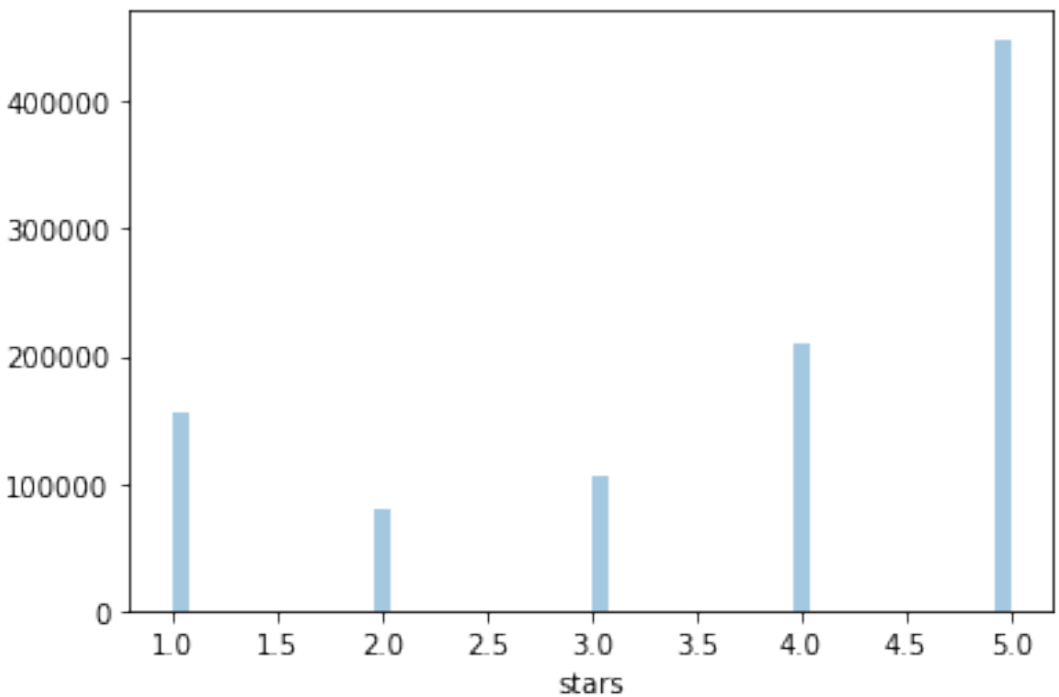
Stars distribution - pie plot



Sentiment analysis

Distribution

Stars distribution - histogram



Sentiment analysis

Labeling,reprocessing

We choose 4,5 stars as label 1 - positive and 0 as negative label. Since that gives approximately around 50 % distribution of the data.
Processioning is done on following basis :

- **cleaning** - extract common weird parts as hashtags etc... and removing punctuation's
- **stopwords** - remove common stopwords

Sentiment analysis

Modeling

Data are split to train and test in 70 % and 30 %.

Further constraint

Count vector will take only words with frequency higher than 100 (included)

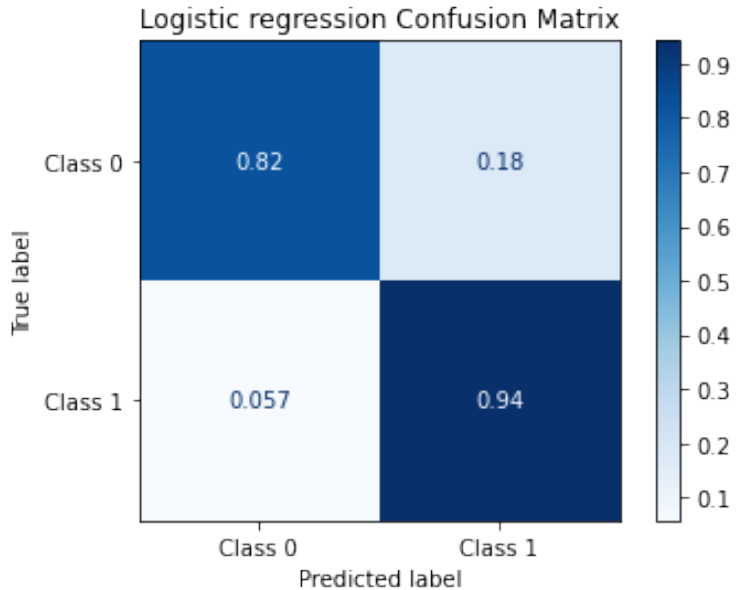
Counter vector has 9625 different words after that.

Sentiment analysis

Modeling - Logistic regression

The Logistic regression is classical and one of the oldest approach in the the classification. For first model we tried LG with 500 iteration.

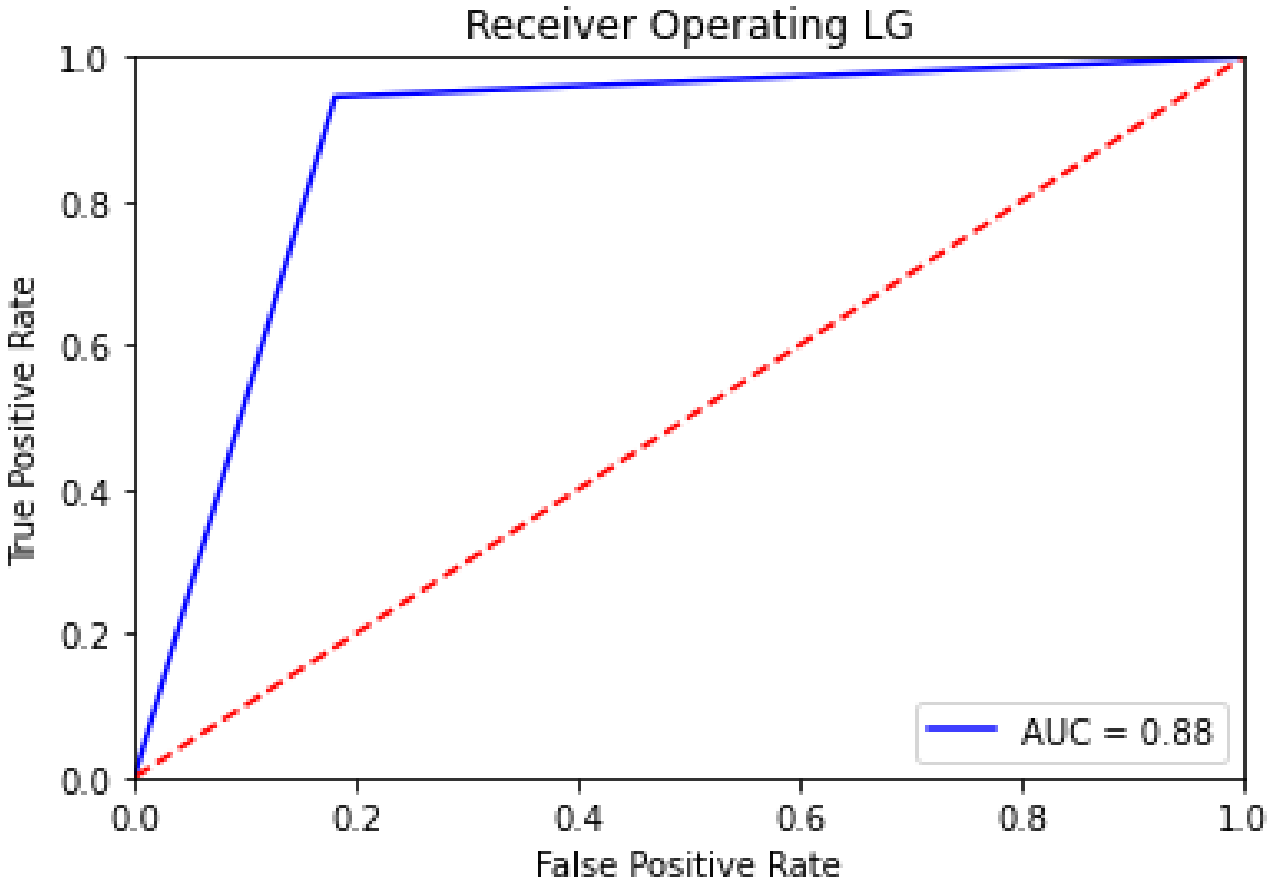
- **Accuracy** - 0.9020
- **MSE** - 0.09802



- **Accuracy - Train** - 0.9087

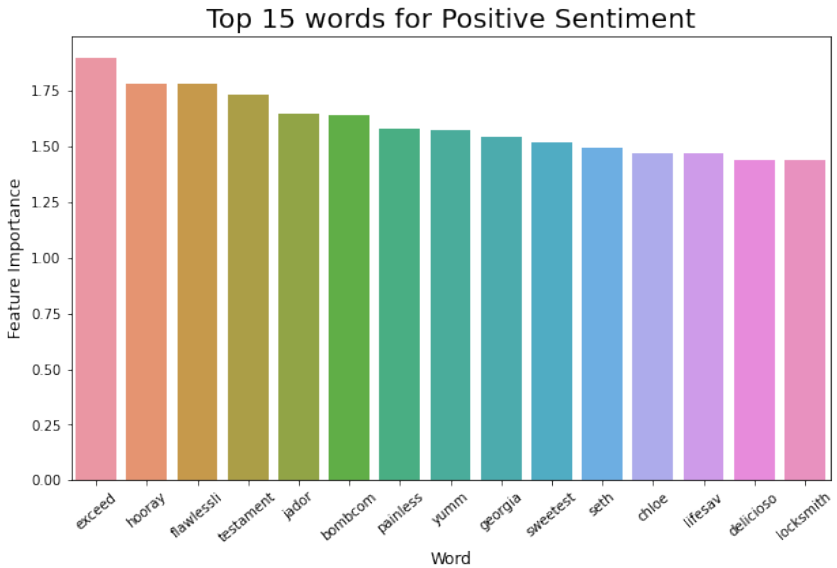
Sentiment analysis

Modeling - Logistic regression - ROC



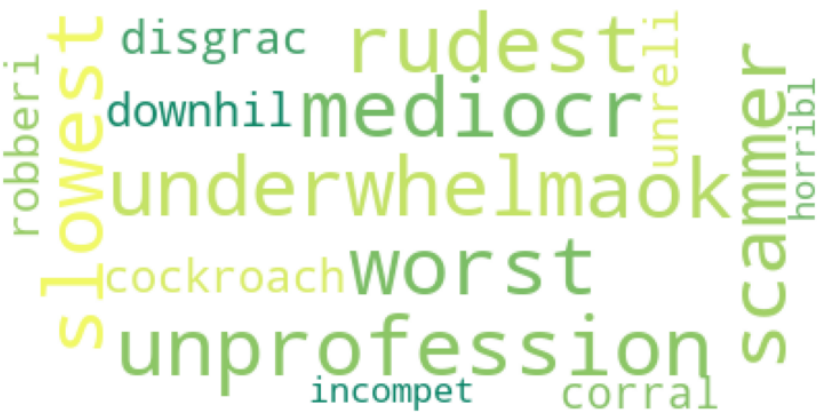
Sentiment analysis

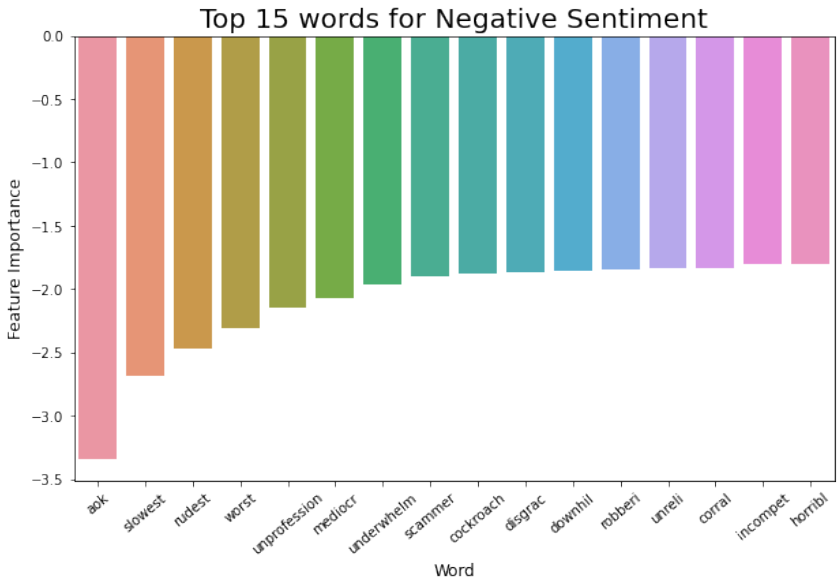
Modeling - Logistic regression - Positive



Sentiment analysis

Modeling - Logistic regression - Negative



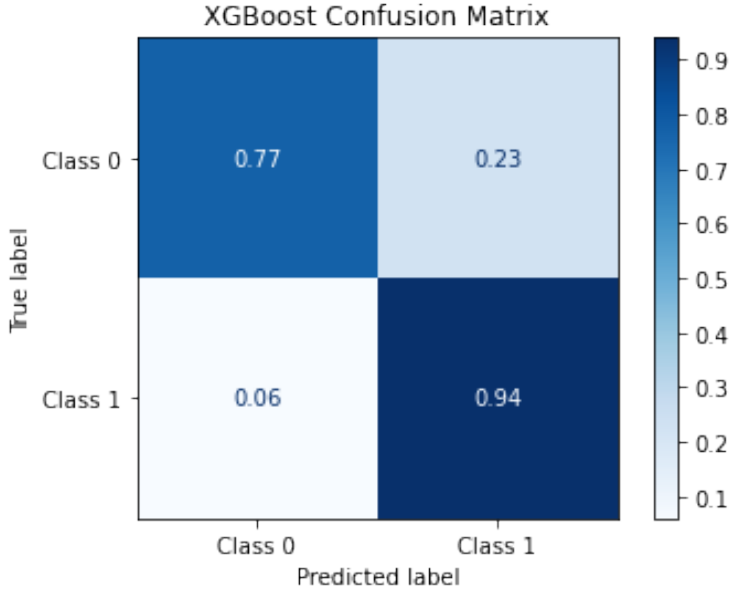


Sentiment analysis

Modeling - XGBoost

XGBoost get it's fame at Kaggle, where was winning competition. it is based on clever approach to penalizing trees.

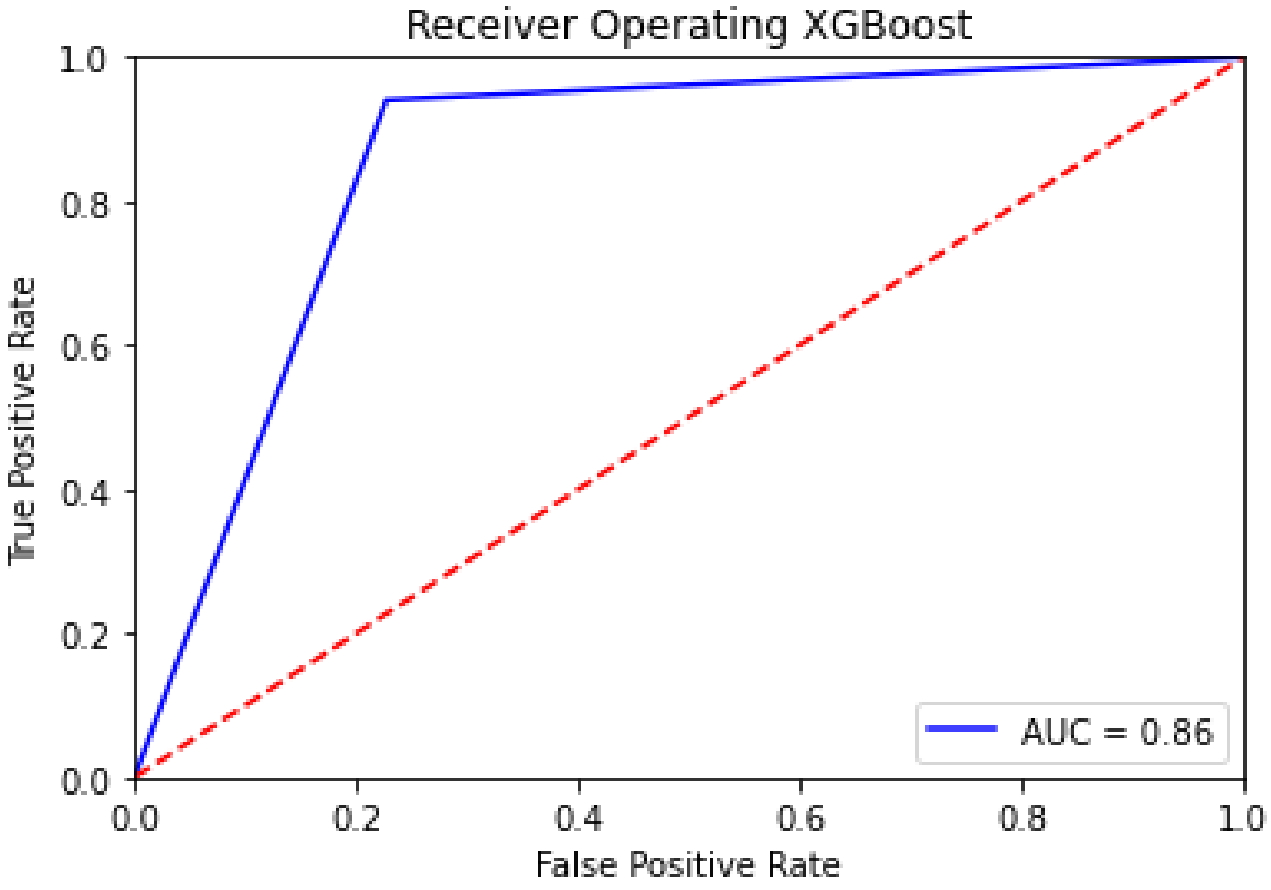
- **Accuracy** -0.8831
- **MSE** - 0.1168



- **Accuracy - Train** - 0.9087

Sentiment analysis

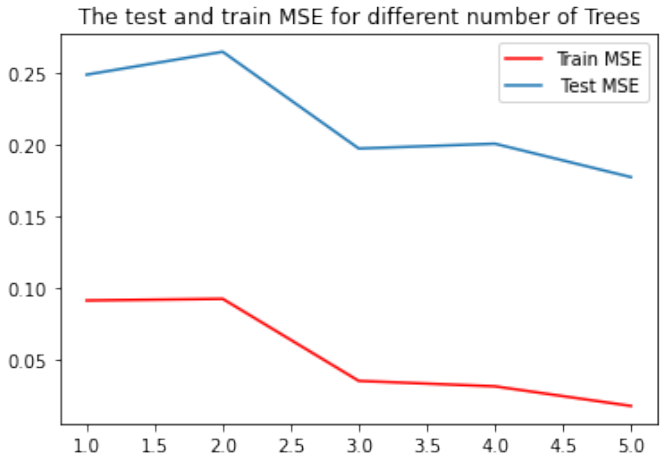
Modeling - XGBoost - ROC

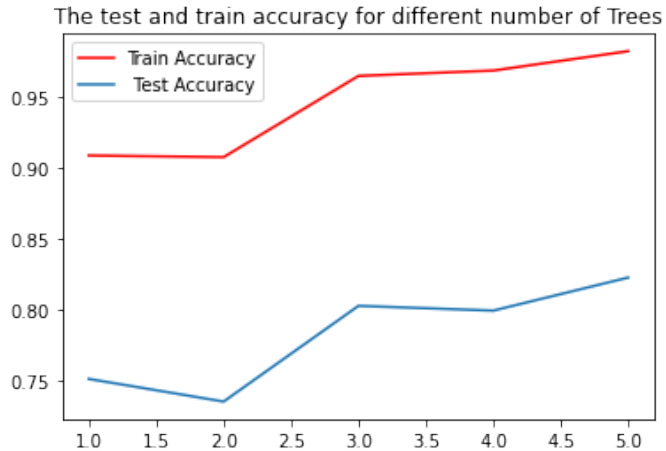


Sentiment analysis

Modeling - Random Forest

Random Forest we need to find correct depth first.



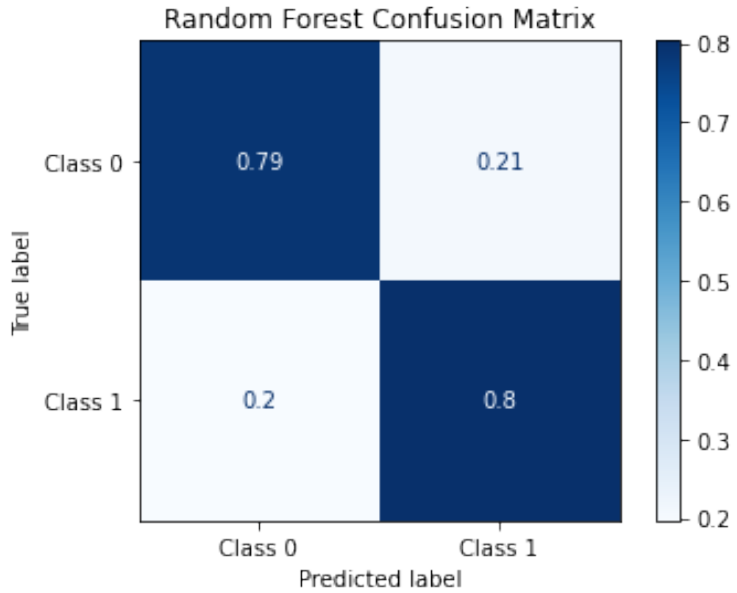


Train MSE is lowest at 5 as well as accuracy is highest at 5.

Sentiment analysis

Modeling - Random Forest - 5 depth trees

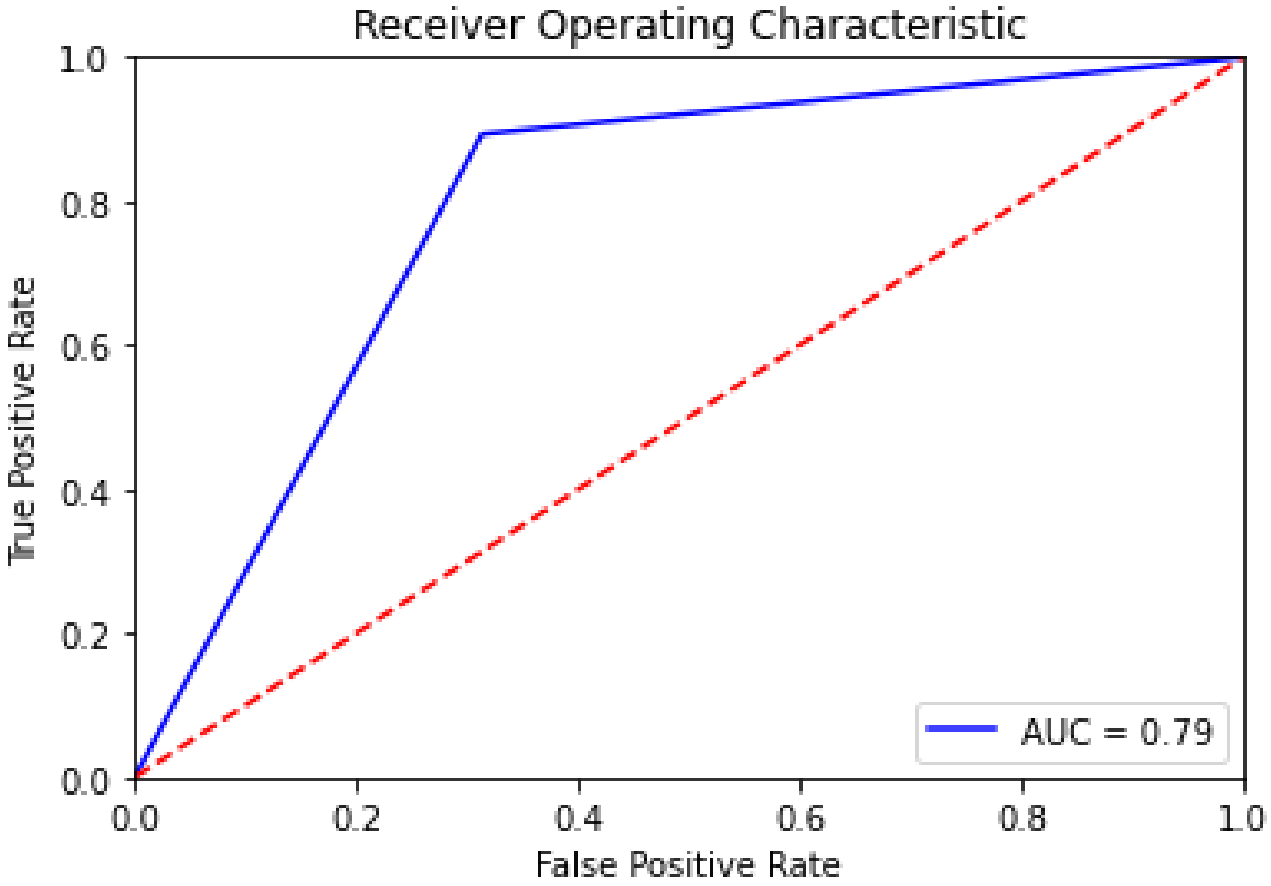
- **Accuracy** - 0.8225
- **MSE** - 0.17749



- **Accuracy - Train** - 0.9823

Sentiment analysis

Modeling - Randim Forest - ROC



Sentiment analysis

Modeling - Summary

Regressor/Test	LG	XGBoost	Random Forrest
Accuracy(Test)	0.9020	0.8831	0.82049
MSE(Test)	0.09802	0.1168	0.1795
Accuracy - train set	0.9087	0.8904	0.9821

- We can see that Random forest highly over fit model. Accuracy Train much higher then Accuracy
- Logistic Regression perform best well rounded .
- In case, we have better machine, we could tune XGBoost.

Sentiment analysis

Modeling - Grid search

Lastly, I tried pipeline the Logistic Regression with regularization L1 and L2, in math know s L2 and L1 - Manhattan Norm. We get following results: L1 penalty with parameter weight 1. Accuracy improved 0.90479.
Possible improvement - using multi-class regression, that is have more classifications like neutral as well.

Sentiment analysis

Codes

- GitHub - main script
- Kaggle - main script
- Kaggle - utils script
- Kaggle - utils test script