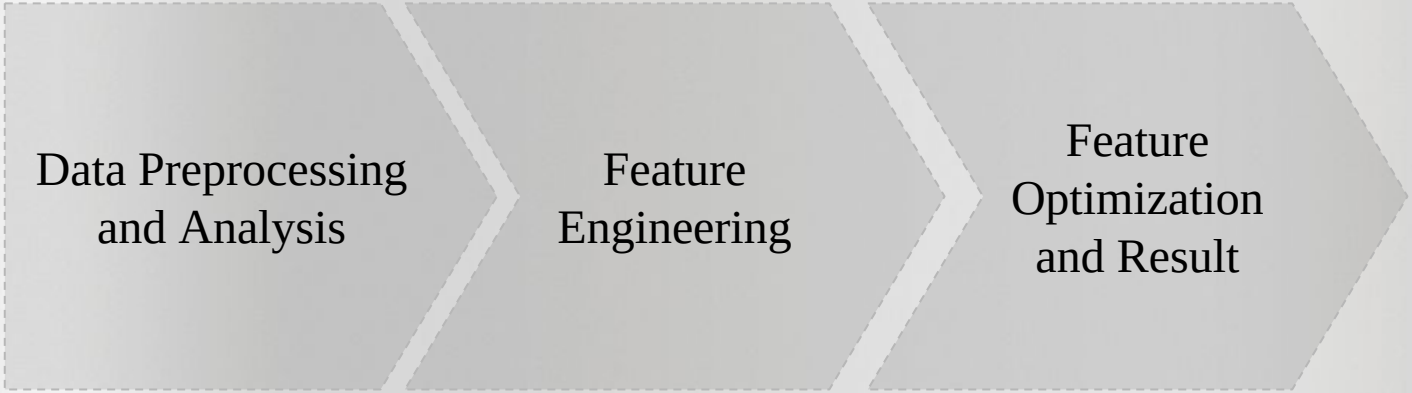
Anomaly Detection Challenge

Challenge 2: Review Anomaly Detection

Team: Abracadata
Ishmeet Kaur(03677735)
Mustika Rizki Fitriyanti(03667399)


Data Preprocessing and Analysis
Feature Engineering
Feature Optimization and Result

1.
Data Preprocessing & Analysis

1.1 About the Data

Aim: Classify Fake and Non fake Reviews

Problem Type:Binary Classification

Dataset

- Training Set: 2908 (1 Mising Review Content in training, 5 in reviewer and none in hotel)
- Test Set: 2950 (1 missing data entry of the hotel and 4 missing data entries of the reviewer)
-

Number of Features: 33 when combined with reviewer and hotel dataset

1.2 Preprocessing

- **Reading the dataset:**

  Read by pandas library and removed seperators in review content by regex.

- **Replacement of the Missing Values:**

  Could not be ignored , missing values in test set.

  Initially, all the missing values were replaced by a null value.

  Replacement by mean, median , mode to increase the accuracy.

2.
Feature Engineering

2.1 Textual Mining(Review Content)

- **Bag of Words**
  - learns a vocabulary ,models by the occurrence of word.
  - Tokenization of the dataset was done .
  - Stemming of the words was also done.

- **Tf-Idf**
  - Increases proportionally to the frequency , reduces importance of frequent words.
  - Dense matrix of the Tf-Idf matrix was compared with the test set.

- **N-Grams:**
  - set of n contiguous words appearing together in the document
  - Implemented unigram, bigram and trigrams as features of the training set.

2.2 Behavioural Analysis (I)

- **Maximum Number of Reviews**
  - ReviewCount / date-yelpJoinDate)
- **Percentage of Positive Reviews**
  - Number of ratings per reviewer in 4,5 /reviewCount.
- **Review Length**
  - length of the review content.
- **Absolute and Expected Rating Deviation**
  - Abs Rating: Difference of the rating of the reviewer and the hotel. -
  - Expected Rating : the standard deviation of all the ratings of a reviewer

2.2 Behavioural Analysis (II)

➢ **Maximum Content Similarity(MCS)**
- maximum value (cosine distance of the reviews of all the reviewers.)

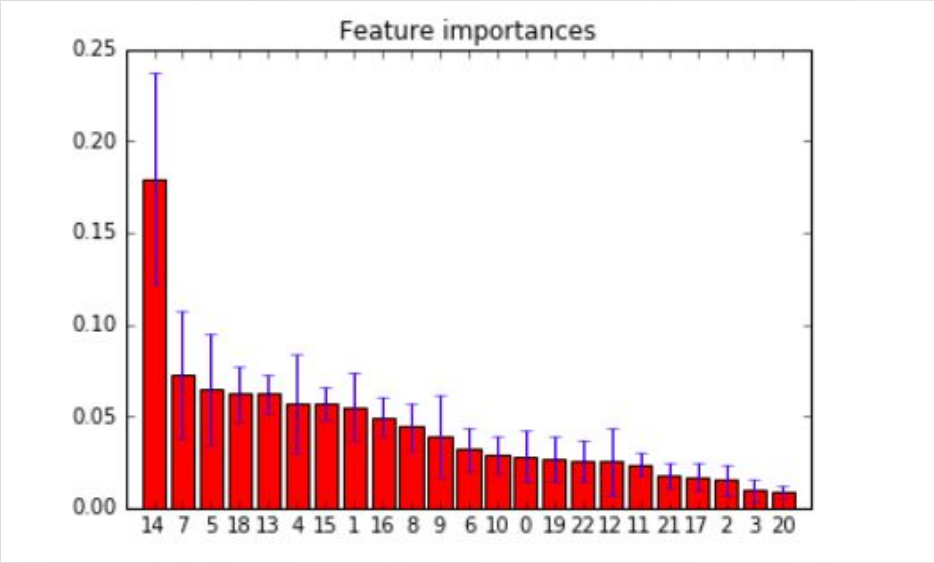
➢ **Mean Rating of reviewer by location of the hotel:**
- The mean rating of the reviewer was taken grouped by the location of the hotel as the hotels in one location had similar ratings .

The first three text mining features gave an accuracy of 51% and the other behavioral features provided an accuracy of 82.3% (Kaggle private score).

3.
Feature Weighting

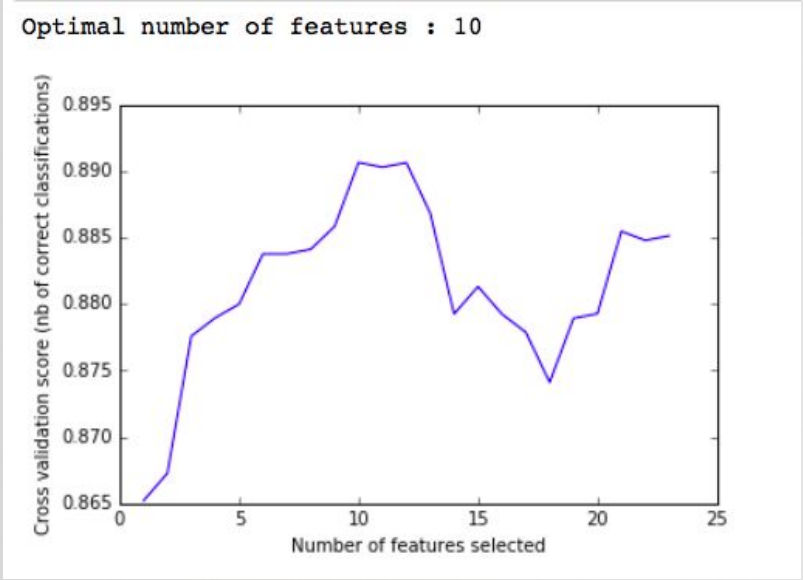
3.1 Random Forest Feature Weighting

- Use 5-10 features based on the weight into the classifiers.
- The highest accuracy resulted by 5 features



3.2 RFECV

- Feature ranking with recursive feature elimination and cross-validated selection of the best number of features

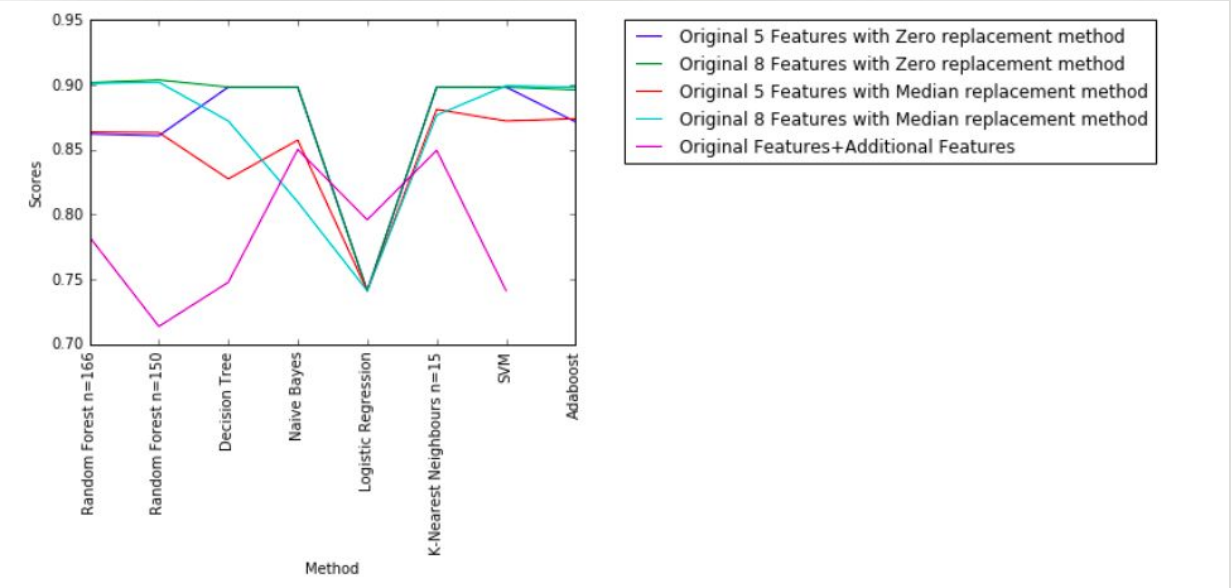


4.
Result Summary

4.1 Method

- Naive Bayes Classifier
- Decision Tree Classifier
- K- nearest Neighbors
- Random Forest
- Logistic Regression
- AdaboostClassifier
- Support Vector Machine

4.2 The accuracy score



4.3 Kaggle Score

Method	Public Score	Private Score
All features + Behavioural Features	0.79288	0.82313
5 Features median with KNN (k=15)	0.80821	0.82172
8 Features zero with Random Forest(with n=166)	0.82066	0.81004

5. Learning

- Importance of Feature Engineering and Behavioural Analysis.
- Applied and in the process learnt about different models
- Too many features (specially related introduced noise in the dataset)
- The more the number of features , the better is a wrong assumption.

Thank you for your attention :)