CS410 Project Presentation

Gargi Deb, Sudhir Koundinya Nagesh, Ambarish Tripathi,
Gayathri Coimbatore Ramachandran

Agenda

1. Problem Statement
2. Solution
3. Exploratory Data Analysis
4. Feature Extraction
5. Classification Models
6. Evaluation (Confusion Matrix / AUC-ROC Curves)
7. Anomaly Detection
8. Streamlit
9. Self Evaluation
10. Demo

Problem Statement

In the past few years, there has been an increase of fake reviews being made across many sites. With the rise of E-Commerce, many buyers make their decision of buying a product or purchasing a service based on the reviews that they read.

There is a need to segregate the fake reviews from genuine ones.

Solution

To reduce the number of fake reviews, they must be first detected.

Using a text classification algorithm, this project is intended to classify fake and genuine reviews.

We are using the dataset : https://osf.io/tyue9/ and

- Python (Pandas, Sklearn, Matplotlib libraries)
- Streamlit

Exploratory Data Analysis

- Load data using Pandas into DataFrame
- Check,
  - ❏ number of unique labels
  - ❏ number of records based on unique label, rating and category
  - ❏ if any column has null value
  - ❏ if any correlation exists between features
  - ❏ whether there is data imbalance

Feature Extraction

- Word count
- Character count
- Capital letters count
- Digit count
- Punctuation
- Capital letters ratio
- Digit ratio
- Punctuation ratio
- Sentiment Score

features

label	word_count	character_count	capital_letters_count	digit_count	punctuation_count	capital_ratio	digit_ratio	punctuation_ratio	sentiment_score	rating
1	13	75	4	0	5	0.053333333	0	0.066666667	0.9593	5
1	16	80	1	0	3	0.0125	0	0.0375	0.891	5
1	14	67	2	0	2	0.029850746	0	0.029850746	0.7906	5
1	17	81	2	0	2	0.024691358	0	0.024691358	0.7463	1
1	18	85	3	0	2	0.035294118	0	0.023529412	0.7397	5
1	8	44	36	0	1	0.818181818	0	0.022727273	0	3
1	19	89	2	0	1	0.02247191	0	0.011235955	0.7506	5
1	17	85	2	0	1	0.023529412	0	0.011764706	0.9169	3
1	17	81	2	0	3	0.024691358	0	0.037037037	0.7087	5
1	16	74	2	0	1	0.027027027	0	0.013513514	0.858	5

Classification Models

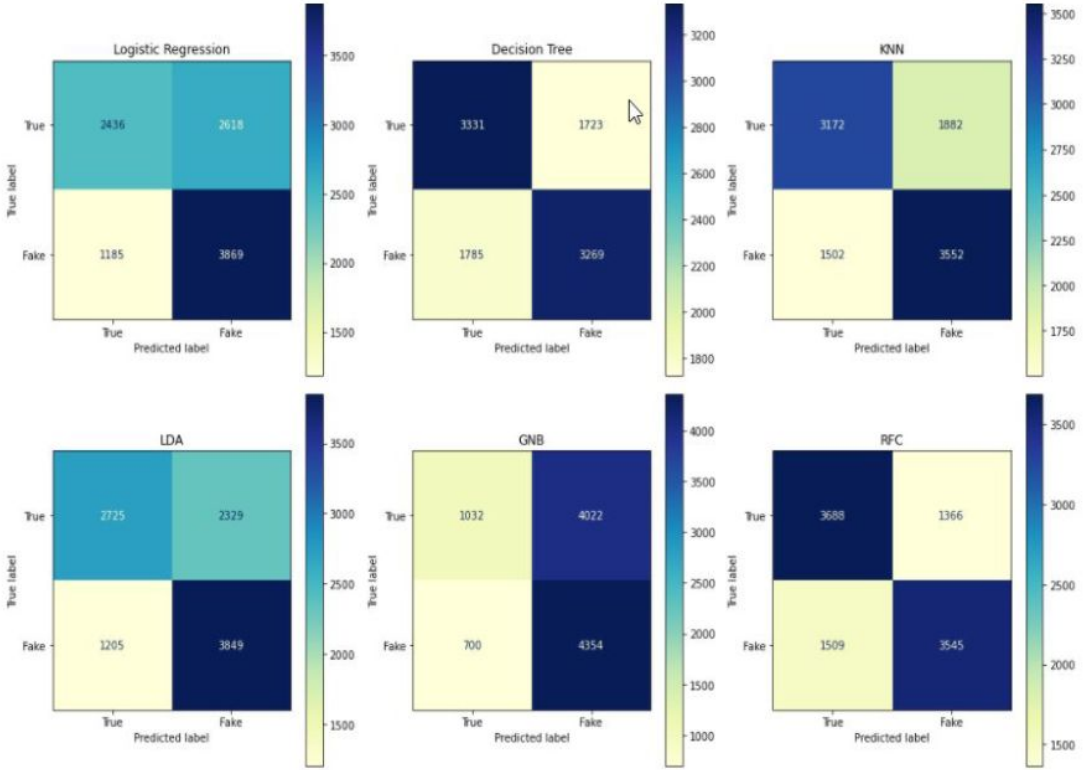
- LogisticRegression
- DecisionTreeClassifier
- KNeighborsClassifier
- LinearDiscriminantAnalysis
- GaussianNB
- RandomForestClassifier

Evaluation – Confusion Matrix

- Illustrates how well different models perform to classify correct and incorrect predictions for each of the target classes under each classifiers. It plots – True Positives, False Negatives, False Positives and True Negatives for each of the models.

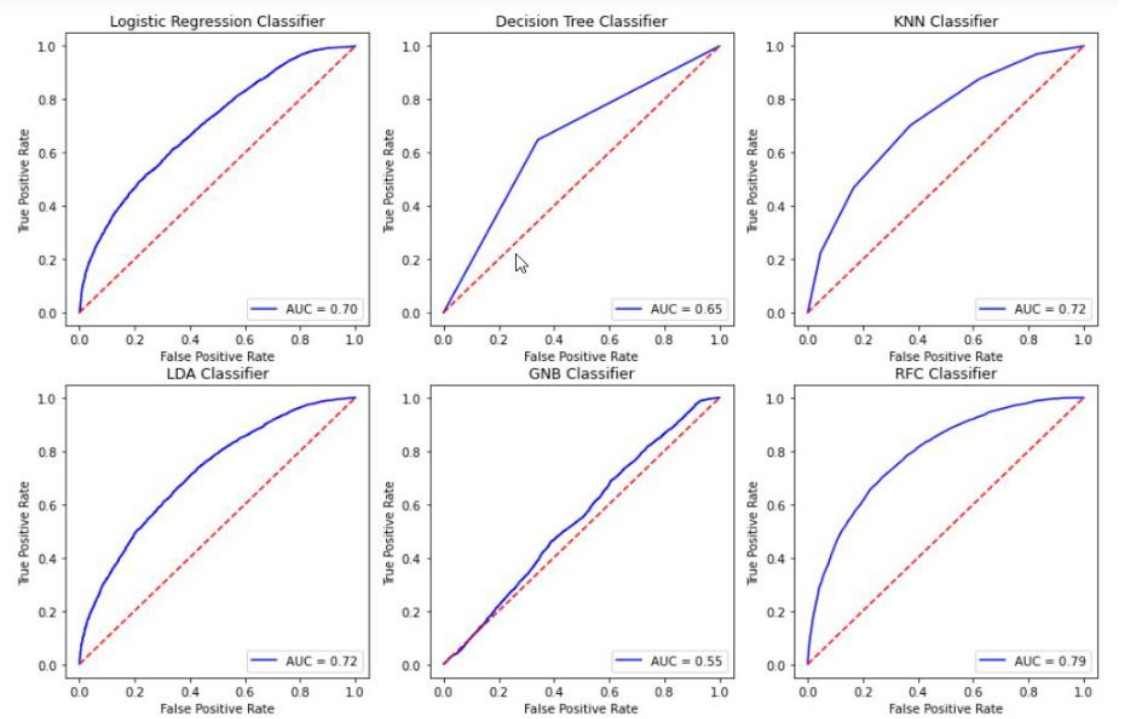
- Precision, recall, F score and Accuracy values can be computed using different models.

    - RFC Classifier Accuracy : 0.71
    - K-NN Classifier Accuracy:0.67
    - Decision Tree Classifier Accuracy: 0.65
    - LDA Classifier Accuracy: 0.65
    - Logistic regression Classifier Accuracy: 0.62
    - GNB classifier Accuracy: 0.53



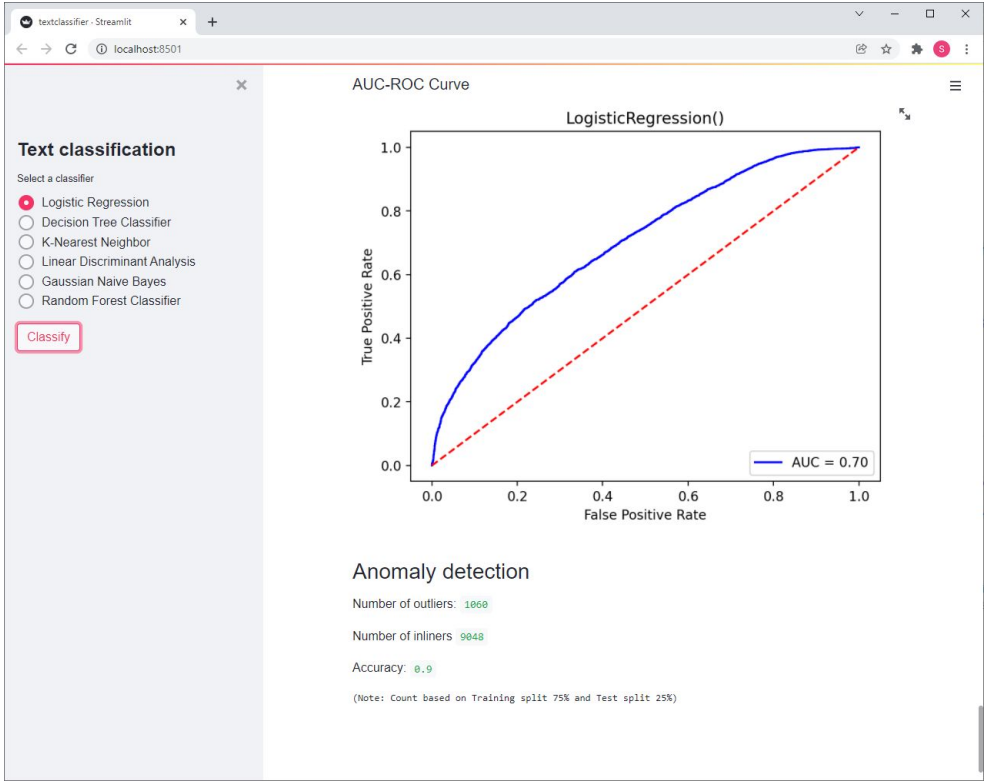
Evaluation – AUC-ROC Curves

- AUC–ROC curve highlights how well a model can classify between different classes.
- The higher the AUC, the better the model at predicting true positives and negatives.
- Models listed in decreasing order of their performance (AUC values):
    - Random Forest Classifier
    - K-Nearest Neighbors
    - Linear Discriminant Analysis
    - Logistic Regression
    - Decision Tree
    - Gaussian Naïve Bayes



Anomaly Detection

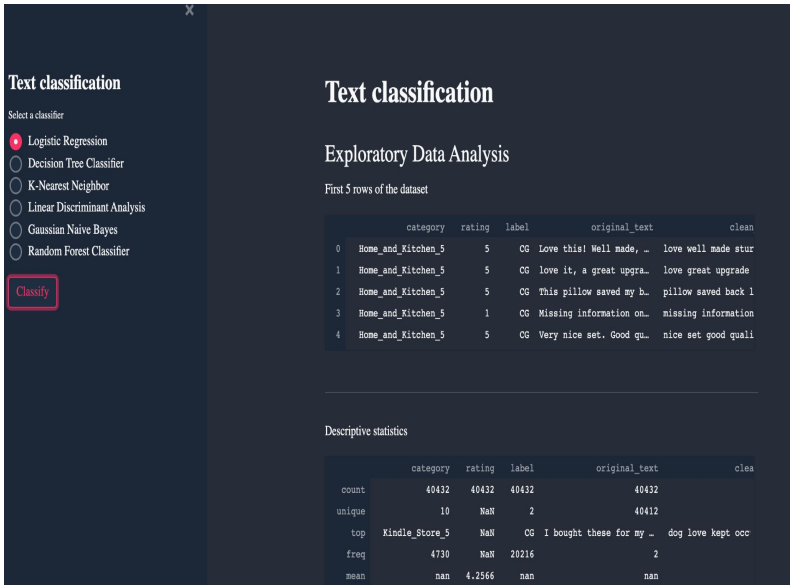
- Isolation Forest with 10% contamination
- Count based on Train/Test split - 75%/25%



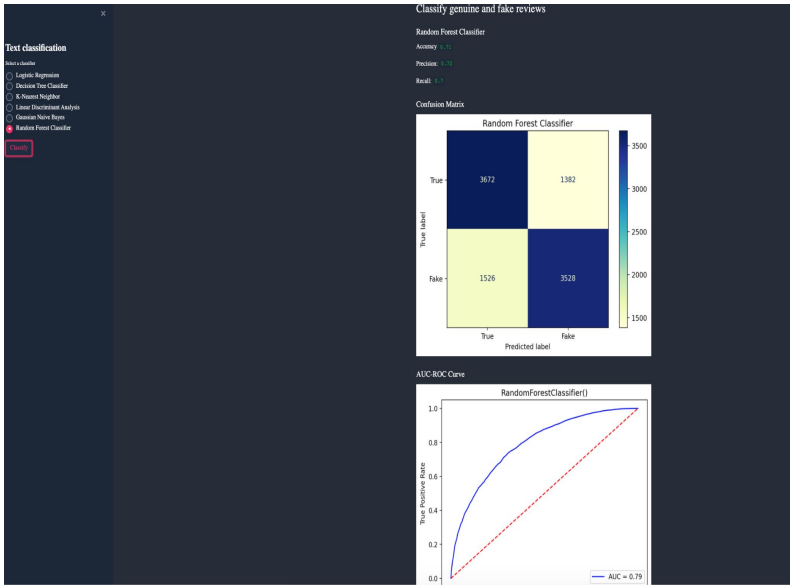
Streamlit

Select a model and click on classify button to view the EDA and model analysis results.

EDA



Model analysis



Self Evaluation

All the below mentioned planned tasks are completed. The outcome meets the expected results.

- Identification of the dataset
- Exploratory Data Analysis
- Clean, Parse and fix class imbalance
- Place the data into proper data structures
- Develop model on train dataset
- Evaluate the model
- Anomaly / Outlier Detection