**Predicting fine food reviews from Amazon dataset**

**Team details:**

Chitra Bhanu Barla, Red ID: 822056424

Swetha Shivraj, Red ID: 819683404

**Goal of the project:**

To perform sentiment classification and categorize the Amazon fine food reviews as positive or negative.

**Dataset:**

The dataset has 5000 reviews comprising of 10 columns, out of which score, summary and text are of interest to us.

The original dataset is extracted from the snap website here: <https://snap.stanford.edu/data/web-FineFoods.html>

**Running the project**

Open Anaconda navigator and launch Jupyter notebook. Use Apache toree - scala kernel. Load the project file titled – ‘AmazonFineFoodReviews\_Final\_Project.ipynb’.

Select Cell 🡪 Run all

The data file (Reviews\_5000.csv) is included in the same directory as the notebook.

**Results**

Out of all the models, Random Forest classifier gave the maximum accuracy and precision. Including the “text” feature helped to increase the accuracy for logistic regression, but for other models it led to overfitting. Overall Random Forest classifier gave the best results.

Certain kernels were posted in kaggle, but our project gave far better accuracies, since we have used many text preprocessing techniques which helped to improve the accuracy.

Please find below the comparison table of different prediction models with parameters – Accuracy, Precision and Recall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Logistic regression** | **Logistic regression with Text** | **Random Forest** | **Naïve Bayes** |
| **Accuracy** | 0.84 | 0.87 | 0.915 | 0.87 |
| **Precision** | 0.47 | 0.66 | 0.87 | 0.57 |
| **Recall** | 0.57 | 0.44 | 0.50 | 0.37 |