|  |  |
| --- | --- |
| Flip Robo Technologies | LinkedInRate Prediction Project   |  | | --- | | Submitted By,  Poovarasi Vijayan | |

**Acknowledgement:**

I wish to express my sincere thanks to the following companies, without whom I would have not got opportunity to work on this project; Data Trained Institute and Flip Robo Technology-Bangalore.

**Introduction:**

**Business Problem Facing:**

A website has a forum for writing technical reviews of products and consists of repository of reviews for technical products. Now they are adding a new feature to their website i.e. The reviewer will have to add stars(rating) as well with the review. The rating is out 5 stars and it only has 5 options available 1 star, 2 stars, 3 stars, 4 stars, 5 stars. Now they want to predict ratings for the reviews which were written in the past and they don’t have a rating. An application to predict the rating by seeing the review is required to be built. Therefore, a predictive model to accurately predict a user’s rating based on input review is required to be made.

**Concept Background of the Domain Problem:**

In today’s web-based world, virtually everyone is reading online reviews. In fact, 91% of people read them and 84% trust as much as they would a personal recommendation. The effects of reviews are measurable, too. Negative reviews can carry as much weight as positive ones. One study found that 82% of those who read online reviews specifically seek out negative reviews.

That may be sound alarming – this states that only emphasizes that negative reviews aren’t going unnoticed – but there are sone benefits: Research indicates that users spend five times as long on sites when interacting with negative reviews, with an 85% increase in conversion rate.

Customers like to see lots of reviews. A single review with few positive words makes up an opinion, but a few dozen that say the same thing make a consensus.

The more reviews, the better, and one study found that consumers want to see at least 40 reviews to justify trusting an average star rating. However, a few reviews are still better than no reviews.

With the vast array of review sites and the level of trust most consumers have in reviews, it’s a safe assumption that virtually everyone considering your products, no matter your target demographic, industry, or market, is reading online reviews before making a purchase.

**Review of Literature:**

“Review-Based Rating Prediction” by Tal Hadad was reviewed and studied to gain insights into the importance of contextual information of user sentiments in determining the rating of products, the role of natural language processing tools and techniques in identifying the user sentiments towards various products based on their reviews and ratings It is learnt that Contextual information about a user’s opinion of a product can be explicit or implicit and can be inferred in different ways such as user score ratings and textual reviews. A user may express in his review(s), their satisfaction / dissatisfaction with a product, based on its quality, features performance, and monetary worth and they may then give the product a rating score based on their opinion of it. These reviews have contextual data based on users’ experiences with the products and their opinions of them. The user ratings have a strong correlation with the contextual data carried in their reviews. Thus, comparing the similarity in the reviews with the similarity in the scores based on those reviews can be a basis for predicting user ratings based on the context, inference and semantics of their reviews.

**Motivation for Problem Undertaken:**

Having lived in India for so many years if there is one thing that I had been taking for granted,

it’s that housing and rental prices continue to rise. Since the housing crisis of 2008, housing

prices have recovered remarkably well, especially in major housing markets. However, in the

4th quarter of 2016, I was surprised to read that Bombay housing prices had fallen the most in

the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first

time there was a decline since Q1 of 2017. The decline has been partly attributed to political

uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency

among customers and also the comparison can be made easy through this model. If customer

finds the price of house at some given website higher than the price predicted by the model, so

he can reject that house.

Having lived in India for so many years if there is one thing that I had been taking for granted,

it’s that housing and rental prices continue to rise. Since the housing crisis of 2008, housing

prices have recovered remarkably well, especially in major housing markets. However, in the

4th quarter of 2016, I was surprised to read that Bombay housing prices had fallen the most in

the last 4 years. In fact, median resale prices for condos and coops fell 6.3%, marking the first

time there was a decline since Q1 of 2017. The decline has been partly attributed to political

uncertainty domestically and abroad and the 2014 election. So, to maintain the transparency

among customers and also the comparison can be made easy through this model. If customer

finds the price of house at some given website higher than the price predicted by the model, so

he can reject that house.

Ratings are an important metric in e-commerce application to determine a product’s quality, consumer demand, worth and profitability. The sentiment of a user towards a product is reflected in their rating score and their review of the product. This helps determine how the product is perceived by the consumers and in turn gives an idea about the acceptance of the product by the consumers. There is a strong positive correlation between the rating of a product and its consumer demand. Therefore, it is necessary to build a predictive model which can, with good accuracy predict what rating a user might give a particular product based on the user review. This helps understand user sentiment towards a product and determine the product’s worth and acceptance by consumers.

**Analytical Problem Framing:**

**Mathematical/Analytical Modelling of the Problem:**

Effective Machine Learning model needs to be created which predicts the ratings based on reviews. So ‘ratings’ is the target variable, here Classification problem with NLP is used to predict the ratings.

Web Scraping is done to collect data from [www.flipkart.com](http://www.flipkart.com)

First, the analysis is started with importing the data, this dataset contains Null values as there are no null values, continued with data cleaning and pre-processing.

Once this is all done then data is ready for modelling. Modelling is done with Logistic Regression, Random Forest Classifier, Ada Boost Classifier, Gradient Boosting Classifier, Decision Tree Classifier, Extra Tree Classifer . From this the best model is identified and using Cross Validation technique is checked for overfitting and underfitting and Hypertuning is done to increase accuracy. Then, Finally the best model is saved.

**Data Source and their Format:**

The data is scrapped using selenium webscraping which is stored in xlsx format.

* Data contains 31480 entries each having 2 variables.
* Here the dataset is divided into independent and target variable.

This project is more about exploration, feature engineering and classification that can be done on this data. Since the data set is huge and includes multi classification of ratings, we can do good amount of data exploration and derive some interesting features using the Review column available. The model is built so that it can predict Ratings of the reviewer.

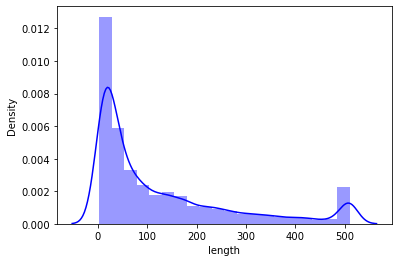
**Data Pre-Processing:**

Columns: Unnamed: 0(just a series of numbers) was dropped since it doesn't contribute to building a good model for predicting the target variable values.  
• The train and test dataset contents were then converted into lowercase.  
• Punctuations, unnecessary characters etc were removed, currency symbols, phone numbers, web URLs, email addresses etc were replaced with single words  
• Tokens that contributed nothing to semantics of the messages were removed as Stop words. Finally retained tokens were lemmatized using WordNetLemmatizer().  
• The string lengths of original comments and the cleaned comments were then compared.

**Data Input Logic Output Relations:**

The comment tokens so vectorised using TfidVectorizer are input and the corresponding rating is predicted based on their context as output by classification models.

**Data Visualization:**

****

This data set contains upto 500 words with most lies below 100.



**Key Observations:**

1. These visualizations shows most occuring words in 5,4,3,2,1 star ratings.
2. From the graphs above the following observations are made:  
   Reviews corresponding to 5.0 rating frequently carry words like: ‘good’, ’excellent’, ’quality’, ‘value money’ etc indicating very high customer satisfaction and high quality product.
3. Reviews corresponding to 4.0 rating frequently carry words like: ‘good’, ’great’, ’performance’, ’features’, ‘quality’, ’price’ etc indicating high customer satisfaction and good quality product.
4. Reviews corresponding to 3.0 rating frequently carry words  
   like: ‘good’, ’well’, ’average, ’quality’, ’issue’ etc indicating  
   customer dissatisfaction and average to below average  
   product quality.
5. Reviews corresponding to 2.0 rating frequently carry words like: ‘problem’, ‘bad’, ’issues’, ’waste money’, ’poor quality’ etc indicating high customer dissatisfaction and below average product quality.
6. Reviews corresponding to 1.0 rating frequently carry words like: ‘stopped’, ’working’, ’cheap’, ’return’, ’issue’, ’waste money’, ’poor quality’,’ customer care’, ’bad’, ‘poor build quality’ etc indicate very high customer dissatisfaction and poor quality product.

**Tools Used:**

1. Python 3.8
2. Numpy
3. Pandas
4. Matplotlib
5. Seaborn
6. Data Science
7. Machine Learning

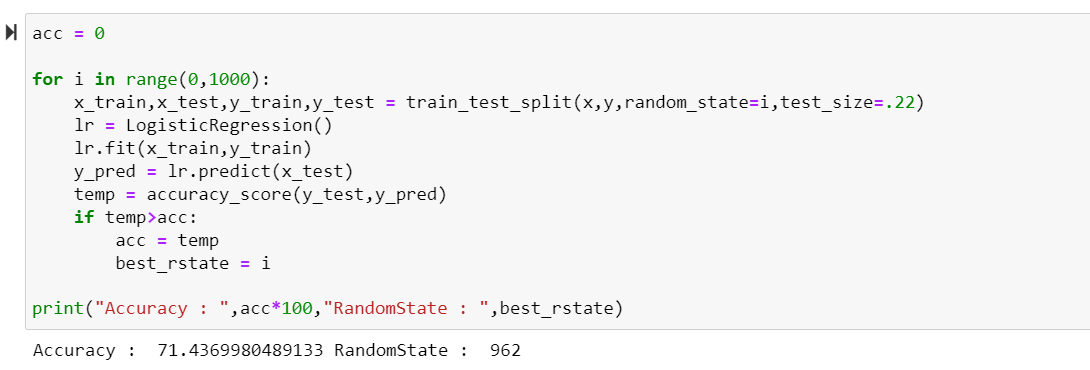
**Model/s Development and Evaluation:**

**Test of Identified Approaches:**

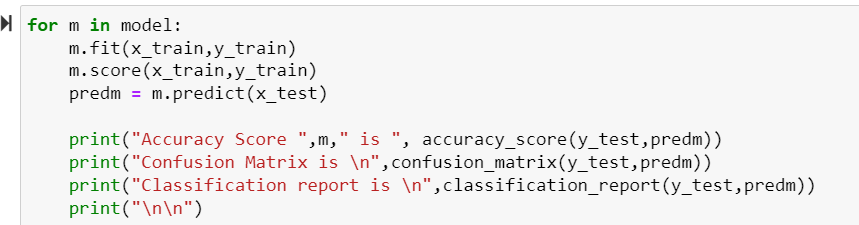
The approaches followed in this project are

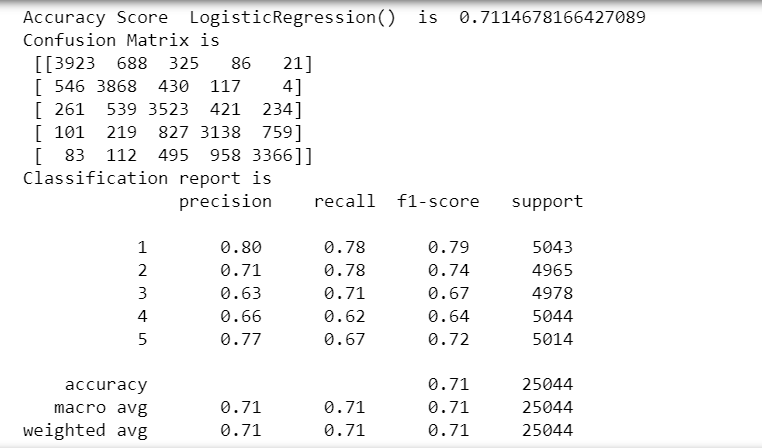
1. Find the best random state of a model.
2. Use all the other models to find accuracy score.
3. Then find Cross Validation of all models to find the best accuracy, which is the least difference between r2 score and cv score.
4. With the best model accuracy, we need to do hyper tuning using GridSearchCV.
5. Then finally saving the best model and by keeping this we need to predict the ratings.

**Run and evaluate of selected model:**

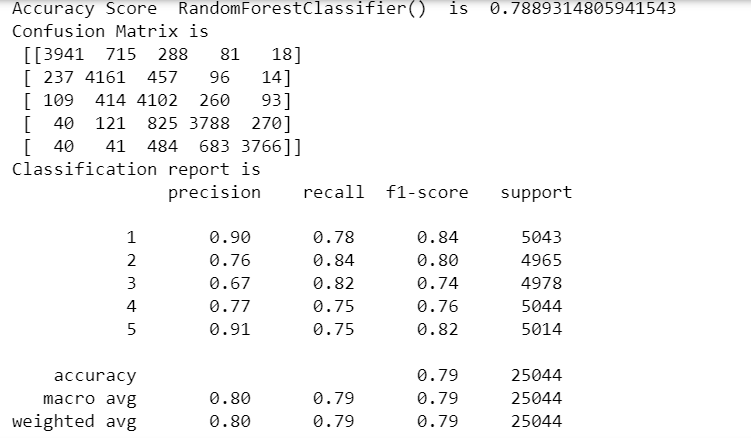
****

We got best random state 962 with accuracy 71%

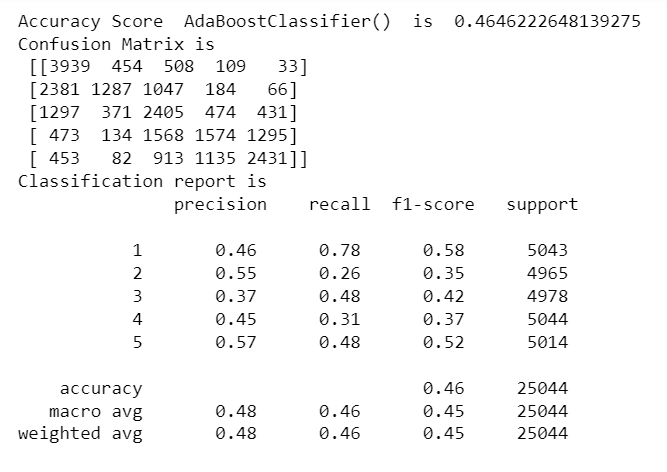
****

****

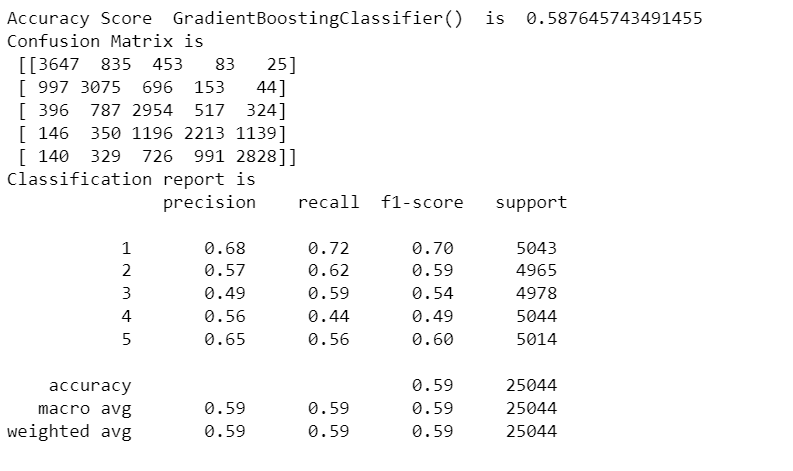
Logistic Regression gives 71.11% accuracy

****

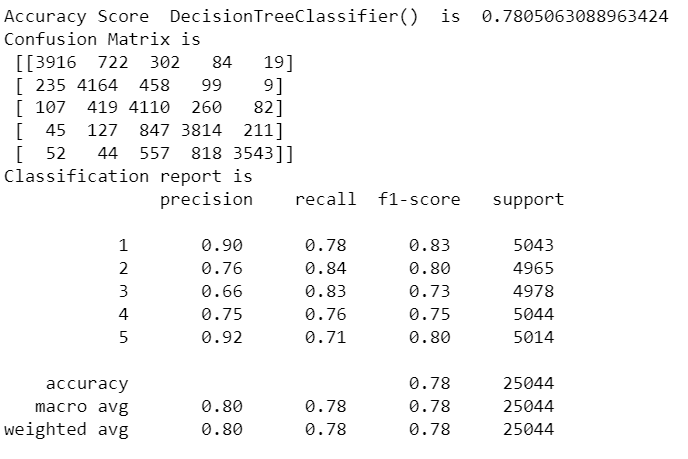
RandomForestClassifier gives 78.8% accuracy

****

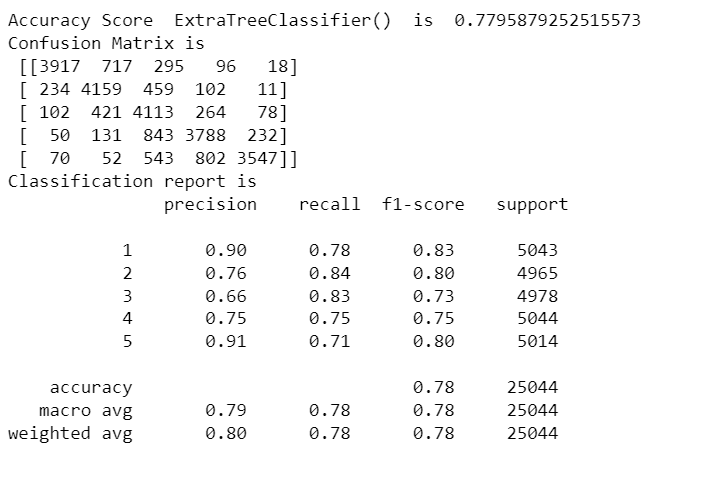
AdaBoostClassifier() give 46.46% accuracy score

****

GradientBoostingClassifier gives 58.78% accuracy score

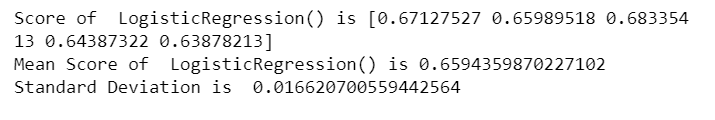
****

DecisionTreeClassifier gives 78.05% accuracy

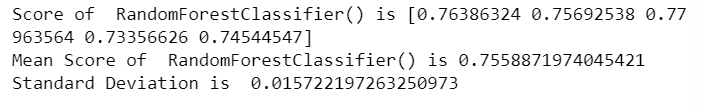
****

ExtraTreeClassifier gives 77.95% accuracy

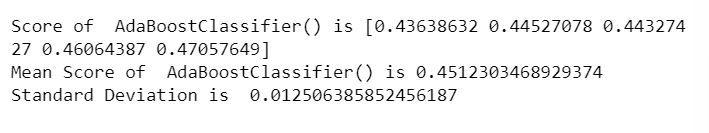
**Cross Validation:**

****

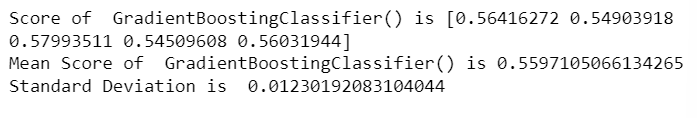
CV Score of LogisticRegression is 0.6594

****

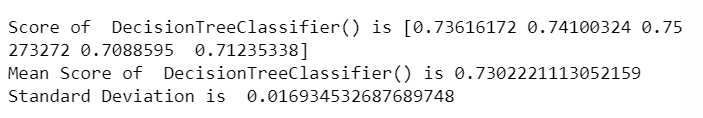
CV Score of RandomForestClassifier is 0.733

****

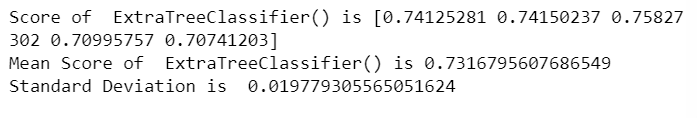
CV Score of AdaBoostClassifier is 0.4512

****

CV Score of GradientBoostingClassifier is 0.5597

****

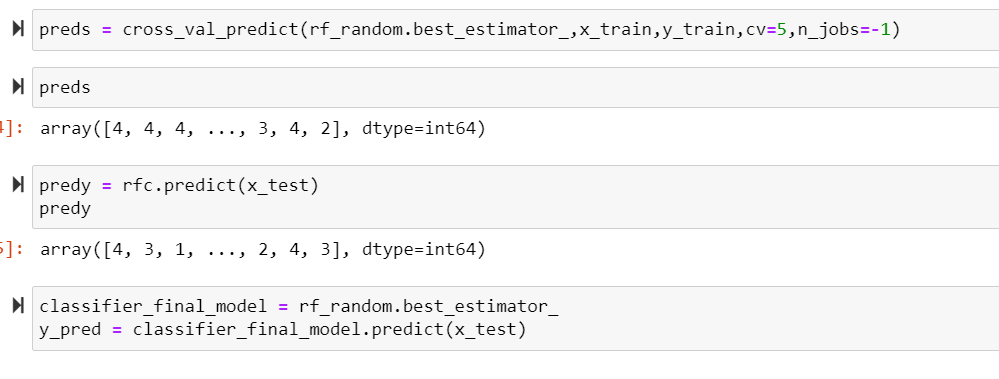
CV Score of DecisionTreeClassifier is 0.7302

****

CV Score of ExtraTreeClassifier is 0.7316

**HyperTuning:**

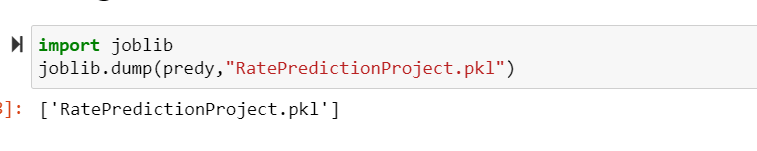
****

****

After Hypertuning we have 78% accuracy score which is improved.

**Saving the best model:**

The best model is saved using pickle.

****

**Key Metrics for success in solving problem under consideration:**

Some of key metrics for the evaluation of this project are

1. AUC ROC Curve
2. Cross\_validation
3. Hypertuning

**AUC ROC Curve:**

AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). Figure 5. AUC (Area under the ROC Curve). AUC provides an aggregate measure of performance across all possible classification thresholds.

**Cross Validation:**

Cross-validation aids in determining the model's overfitting and underfitting. The model is constructed to run on several subsets of the dataset in cross validation, resulting in numerous measurements of the model. If we fold the data five times, it will be separated into five parts, each representing 20% of the whole dataset. During the Cross-validation, the first part (20%) of the 5 parts will be left out as a holdout set for validation, while the rest of the data will be utilised for training. We'll acquire the initial estimate of the dataset's model quality this way.

Further rounds are produced in the same way for the second 20% of the dataset, which is kept as a holdout set while the remaining four portions are utilised for training data during the process. We'll acquire the second estimate of the dataset's model quality this way. During the cross-validation procedure, these stages are repeated to obtain the remaining estimate of model quality.

**Hypertuning:**

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

**Conclusion:**

**Key Findings and Conclusions of the Study:**

In this project data is collected for reviews and ratings for different products from [www.flipkart.com](http://www.flipkart.com). Then the Ratings is detected from commercial websites on a scale of 1 to 5 on the basis of the reviews given by the users. By using of natural language processing and machine learning algorithms the problem is solved. Then checked frequently occurring words in our data as well as rarely occurring words. After all these steps built function is used to train and test different algorithms and using various evaluation metrics I have selected Random Forest Classifier for our final model. Finally, by doing hyperparameter tuning we got optimum parameters for the final model.

**Learning Outcomes of the Study in respect of Data Science:**

Data cleaning was a very important step in removing null values from the dataset. Visualising data helped identify class composition of label column and the most frequently occurring words in reviews corresponding to each of the rating scores. The various data pre-processing and feature engineering steps in the project lent cognizance to various efficient methods for processing textual data. The NLTK suite is very useful in pre-processing text-based data and building classification models.

**Limitations of this work and Scope for Future Work:**

First drawback is scrapping the data as it is a fluctuating process. Followed by raw data which is not in format to analyse. Also, this study will not cover all Regression algorithms instead, it is focused on the chosen algorithm, starting from the basic ensemble techniques to the advanced ones.

**References:**

A few external references have been used to complete this project successfully. Below are the sources

1. <https://scikit-learn.org/>
2. <https://github.com/>
3. <https://analyticsvidhya.com/>