Rating Prediction Project

Submitted By:-
Mohit Meel

Acknowledgement

- The project allocated by the company contains the steps or process how to complete the project as it contains the file of sample documentation and problem statement which shows the process to do the project.
- The data sources used is the flipkart website from where I have scrape the ratings and reviews on different electronic/technical products and the tool used to scrape the data is web scraping selenium.
- The references and professionals that are helping with the project is our SME who is always ready to help us to complete the project whenever we need.

Introduction

- The problem statement of the project is to build a model or an application for ascertaining the ratings of the reviews given by the customers. Now a days, the customers buying the product from online websites give their reviews and also ratings after using the products. The reviews and ratings given by the customers can be good or bad in terms of working of the product, these reviews and ratings helps the company to understand the response of the product in the market. The bad reviews and ratings given, the company fix it and make it better than earlier.
- The data collected for the project is purely primary as it is scraped from the most famous websites which is flipkart. The data is showing ratings and reviews given by the customers of the different technical products. The data has 29974 rows and 2 columns. The ratings column has 1,2,3,4,5 which means that the customers are rating the products good and bad out of 5.
- The aim to work with the project is to know the ratings based on the customer reviews. Different reviews has different ratings so it is necessary to understand that reviews given whether good and bad which means to scale the reviews out of 5.

Analytical Problem Framing

- The data sources is the primary as it is collected from flipkart websites. The data is in csv file. It has 29974 rows and 2 columns that are reviews and ratings. The data is object type.
  The technique used is web scraping selenium to scrape reviews and ratings.
- The data is first analyzed and understand by writing the code such as head, tail, dtypes, columns, info, describe, Corr, shape, skew, is null, then the data cleaning process is done in which the reviews have first converted into lower case next the replaced/removed the email address, URL with web address, money symbols, phone numbers and numbers if any after that the ignoring of punctuation, stop words is done next using word net lemmatize next using the tf-idf vectorizer for converting the text into vectors.
- The relationship between the input and output is that the input (that is reviews which is in the text format) and the output (that is ratings are scaled from 1 to 5) is highly correlated as these reviews will be taken to ascertain whether these are scaled out of 5. The output variables are 1,2,3,4,5.
- The only assumption I have taken that the target column of is showing class imbalance so I have removing class imbalance using smote technique and the datatype of data is object therefore I have not removed outliers and skewness present in it.
- The libraries that are used is pandas as seaborn, matplotlib, warnings, NumPy, word tokenize, reg exp tokenize, stop words, wordnet, string, Word Net Lemmatize, Porter Stemmer.

Model Development and Evaluation

- The approaches I have made for solving the problem is classification approach as the prediction variable is discrete which means from range 1 to 5, any number can be the output. I have send the 70% of data for training and 30% for testing and random state is 45.
- The models which I have taken for the training and testing is logistic regression, decision tree classifiers, SVC(poly,rbf,linear), random forest classifiers, ada boost classifiers, gradient boosting classifiers.
- The models listed above has the accuracy score of 92%, 92.5%,(92.3% , 92.7% , 92.2%), 92.8% , 91.1% and 91.8% respectively. The best which comes out from all listed above models is the random forest classifiers with an accuracy score of 92.8%.
- The key metrics used are accuracy score, confusion matrix, classification report, roc auc score, roc curve and grid search cv for the best model.
- The visualization part done on the project is the heatmap for knowing null values and correlation between the input and output variable, the boxplot for knowing the outliers present, the kde-plot, hist-plot for checking the skewness in the dataset and the roc curve plot showing the roc auc score of random forest classifiers.
- The results come up from visualization is that the outliers are present and skewness is also present but can't be removed because of object datatype if removed then data will become biased.
  The roc curve is exactly it should show according to the accuracy score achieved from random forest classifier

Conclusion

- The key findings and observations from the problem is that the company wants to know about the survival of the product in the market by examining the reviews and ratings given by the customers on the purchased product and also making the product better than earlier based on the reviews get by the customers so that sales and customers can be increased.
- The learnings from the project is how to use the different NLP techniques with the text columns. The reviews and ratings given by customers, in what way they are helpful to the company. The algorithm which stands best is the random forest classifier based on achieving highest accuracy score among all algorithms used. There were no challenges that I have faced in doing the project.
- The limitations of the solution given is somewhere difficult to understand the reviews whether they are in favour of the product or against. This is because the ratings range is 1 to 5 so according to me there should be another column too in which there should be two options good or bad.