FLIP ROBO

Submitted By,
Tinu Shiby

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my mentor Mr. Kesav Bensal, Flip Robo Technologies for his inspiration and guidance at every stage of the project. The blessing, help and guidance given by him will certainly help me in achieving better things in future.

I wish to express my profound gratitude to the DataTrained family for providing an opportunity to undertake this internship.

Lastly, I thank the almighty and my parents for their every small support and encouragement which helped me to complete this project successfully.

INTRODUCTION

All the E-Commerce websites rely greatly on customer satisfaction, and customer's used to express their opinions by writing reviews on whatever product they purchased. Now the company wants to predict the ratings from the review.Company offers a rating out of 5 stars.The motive behind such an initiative is that the decision to purchase a product depends on the reviews and ratings.

Motivation for the project is that many product reviews are just textual, making it difficult to identify so its better to have a star rating. Getting an overall sense of textual review could in turn improve consumer experience.

DATA COLLECTION

The details of cars are collected from the website of flipkart scrapped the details of almost 20077 products.

The following items are scrapped:

- Reviews
- Ratings

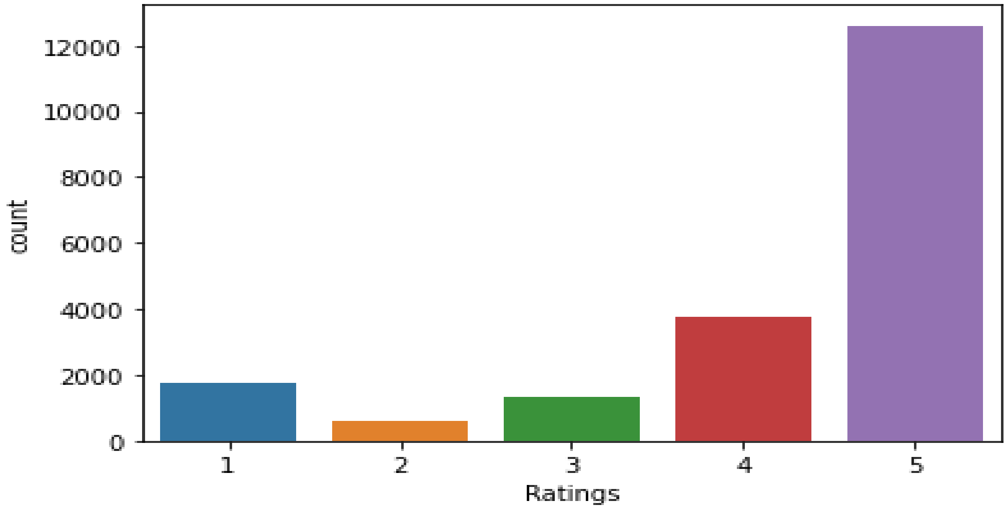
Dataframe was created and converted to csv file for model building.

DATA ANALYSIS

The goal is to find out how price varies according to the variables. To start with the data analysis, i have imported the following:

- Pandas
- Matplotlib
- Seaborn
- Train test split
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- SVC
- XGBoost Classifier
- GridSearchCV
- Cross-validation score
- Classification report

Plotted a graph which represents the count of each ratings



We can observe that the dataset is imbalanced.

DATA PREPROCESSING

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.
Dataset doesn’t have any null values.So started the data cleaning. Below are the steps taken to clean the data:

- Converting to lower case
- Removal of Special characters
- Removal of numbers
- Removal of hyperlinks
- Removal of whitespaces

Then with the help of WordNet plotted the graph for each rating.

Converted the text to vector by TF-IDF method.

Applied SMOTE technique to make the data balanced.

Applied GridSearchCV to find out the best parameters of Random Forest Classifier.

Through pickle saved the model.

BUILDING MACHINE LEARNING MODELS

As the problem is a classification problem I have used the following models:

- Logistic Regression
- Decision tree Classifier
- Random forest Classifier
- SVC
- XGBoost Classifier

The accuracy and cross-validation score of each model is listed down:

Model	Accuracy	Cross-Validation score	Difference
Logistic Regression	78.21	70.75	7.46
Decision tree	86.23	78.46	7.77
Random Forest	89.58	82.94	6.64
SVC	87.94	81.00	7.94
XGBoost	81.82	74.98	6.84

Random Forest Classifier is selected as the best model on the basis of accuracy as well as it has the least difference.

CONCLUSION

- To predict the ratings, used Natural Language Processing toolkit as well as Machine Learning Algorithms and selected Random Forest Classifier as the best model
- Few challenges faced while doing the project is the class imbalance and presence of a lot of text data. With SMOTE technique and TF-IDF method solved that two challenges
- As with any project there is room for improvement here. We couldn’t reach to the goal of maximum accuracy in the project, we did end up creating a system that can with some improvement and deep learning algorithms get very close to that goal.