RATINGS PREDICTION

Submitted by:

MOHAMMED MINHAJ

ACKNOWLEDGMENT

Wikipedia, Youtube.com, google.com. The data were collected from the used car website cardheko.com. Referred in study materials of machine learning which I am studying.

INTRODUCTION

- ## Business Problem Framing

  A client have a website which people write the review of the products like technical products. They are decided to make the people to write the ratings of the products. The products demands in market with the ratings and review of the particular products.

- ## Conceptual Background of the Domain Problem

  Predict the ratings using the feature review.

- ## Review of Literature

  Reviews of products gives a better understanding of the products that we wish to buy. For example, a movie rating and review make it's performance in box office. As well as in the products in e-commerce websites also demands according to the review and ratings. The customer will understand about the product who bought before. And customer can give the review and ratings among the products. If the product have the ratings four out of five it is good product.

- ## Motivation for the Problem Undertaken

  As I mentioned a review of product will helps the customer to buy the products and evaluate the quality of the products. It helps the customer whether buy the product or not. Ratings make simple understanding of the quality of product other than review. So it helps customers who buys from e-commerce websites. We all are buying different types of products from e-commerce. The first thing we observe about the product in websites are the review and ratings. So this project motivates me to help the customers to evaluate quality of the product by ratings and review.

Analytical Problem Framing

- Mathematical/ Analytical Modelling of the Problem

  In this project, dataset have only two entries review and ratings of products. Rating is given according to the review. And also the data were collected from e-commerce websites. Collected the review and ratings of the products like technical products.

- Data Sources and their formats

  Data were collected from the e-commerce websites like amazon and flipkart.com. I have collected around 21000 data. It have review and ratings and also have the review title. Later I dropped that column which is not needed for the prediction.

```
1 df.head()
```

	Review Title	Ratings	Review Summary
0	Moderate	4	If you can affort few thousands extra you can ...
1	Best in the market!	5	Laptop is amazing and sleek. Good for day-to-d...
2	Mind-blowing purchase	5	I was confused between asus and lenovo s145 la...
3	Simply awesome	5	Got this for 61k. Great value for money. Best ...
4	Value-for-money	4	Laptop is good with good configuration in this...

  The review contain the products like mobile, laptops, home theatre, printers, routers, monitors. The majority of collected from the flipkart.com.

- Data Pre-processing Done

  First of all I dropped the column the review title from the dataset. And some NLP pre-processing done in the column review. Create new feature of the length of the review. Remove the punctuations, remove the stops words from review. And also converts it into lower case. And at last vectorize (tfidfvectorization) the review. Our target variable Ratings have ‘None’ data, I removed it with mode

function. Here I used mode because, the data type of the column is categorical features.

- Data Inputs- Logic- Output Relationships

  Review is the only feature for the prediction of rating. It impacts the rating. That is review of the product affects the rating of the products.

- State the set of assumptions (if any) related to the problem under consideration

  Review and rating helps the customer or people like who are buying the products from ecommerce website to evaluate the quality of the products. Whether the product is branded or a copy of the brand. This all solved by checking the review.

- Hardware and Software Requirements and Tools Used

  The dataset have only used 660 kb of memory is used. Pandas were used for importing dataset. Numpy is imported for dealing with numbers. Matplotlib and seaborn were imported for visualization techniques for better analysing the data. For vectorization technique feature extraction.text imported for this. From model selection module cross validation score, grid search cv for tuning parameters and train test split for splitting model into training and testing. And some metrics accuracy score which evaluates the performance of the algorithm and classification report which shows recall, precision etcetera for evaluating performance of the algorithm. And the modules for the algorithms. Naïve bayes for multinomialNB, SVM module for linear SVC and ensemble for random forest.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Our objective in this project is to predict the ratings prediction. Here I used one feature to predict the ratings, that is review summary. The rating is given according to the review.

- Testing of Identified Approaches (Algorithms)
  a. Linear SVC
  b. Random Forest Classifier
  c. Multinomial Naïve Bayes

- Run and Evaluate selected models

MultinomialNB

Model Building

[36]:

```
MNB=MultinomialNB()

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=45)

MNB.fit(x_train,y_train)

predgn=MNB.predict(x_test)

print("Report=",classification_report(y_test,predgn))

print("accuracy=", accuracy_score(y_test,predgn))

```

```
Report=               precision    recall  f1-score   support

           3       0.97      0.68      0.80       365
           4       1.00      0.91      0.95      1322
           5       0.94      1.00      0.97      3595

    accuracy                           0.95      5282
   macro avg       0.97      0.86      0.91      5282
weighted avg       0.96      0.95      0.95      5282

accuracy= 0.9547519878833776
```

In above picture we can observe that the 25% of data were used for testing and 75% for training at random state 45. The accuracy score is 95% which gives good performance. Precision, recall are also good.

LinearSVC

In [37]:

```
#Linear Svc
ls=LinearSVC()

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=45)

ls.fit(x_train,y_train)

predls=ls.predict(x_test)

print("Report=",classification_report(y_test,predls))

print("accuracy=", accuracy_score(y_test,predls))
```

```
Report=               precision    recall  f1-score   support

           3       0.90      0.98      0.94       365
           4       0.99      0.97      0.98      1322
           5       1.00      0.99      1.00      3595

    accuracy                           0.99      5282
   macro avg       0.96      0.98      0.97      5282
weighted avg       0.99      0.99      0.99      5282

accuracy= 0.9869367663763726
```

In above picture we can observe that the 25% of data were used for testing and 75% for training at random state 45. The accuracy score is 98% which gives good performance. Precision, recall are also good.

Random Forest

```
#random forest
rfc=RandomForestClassifier()

x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=45)

rfc.fit(x_train,y_train)

predrf=rfc.predict(x_test)

print("Report=",classification_report(y_test,predrf))

print("accuracy=", accuracy_score(y_test,predrf))
```

```
Report=               precision    recall  f1-score   support

           3       0.90      0.98      0.94       365
           4       0.99      0.98      0.99      1322
           5       1.00      1.00      1.00      3595

    accuracy                           0.99      5282
   macro avg       0.96      0.98      0.97      5282
weighted avg       0.99      0.99      0.99      5282

accuracy= 0.9901552442256721
```

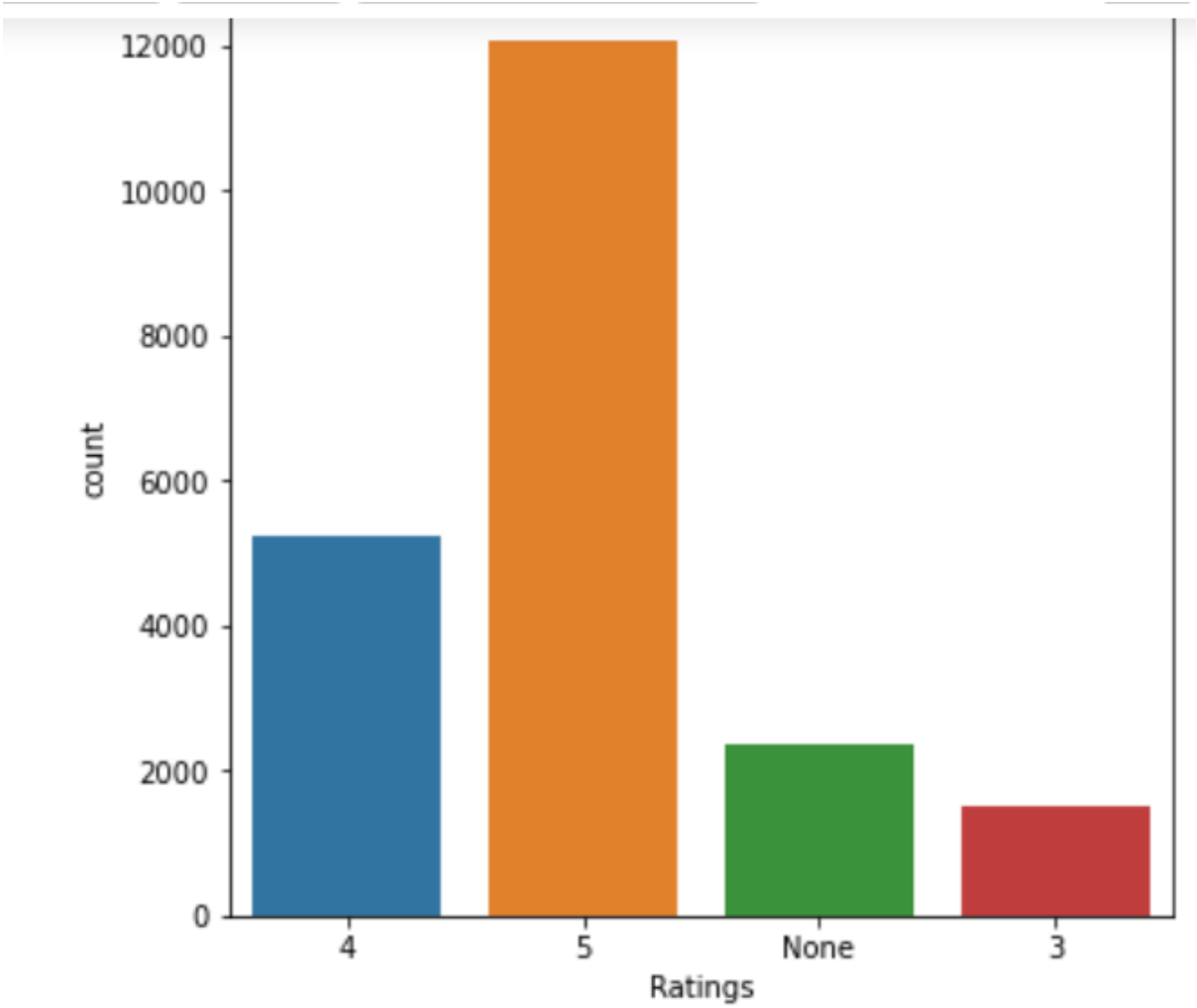
In above picture we can observe that the 25% of data were used for testing and 75% for training at random state 45. The accuracy score

is 99% which gives good performance. Precision, recall are also good.

- Key Metrics for success in solving problem under consideration

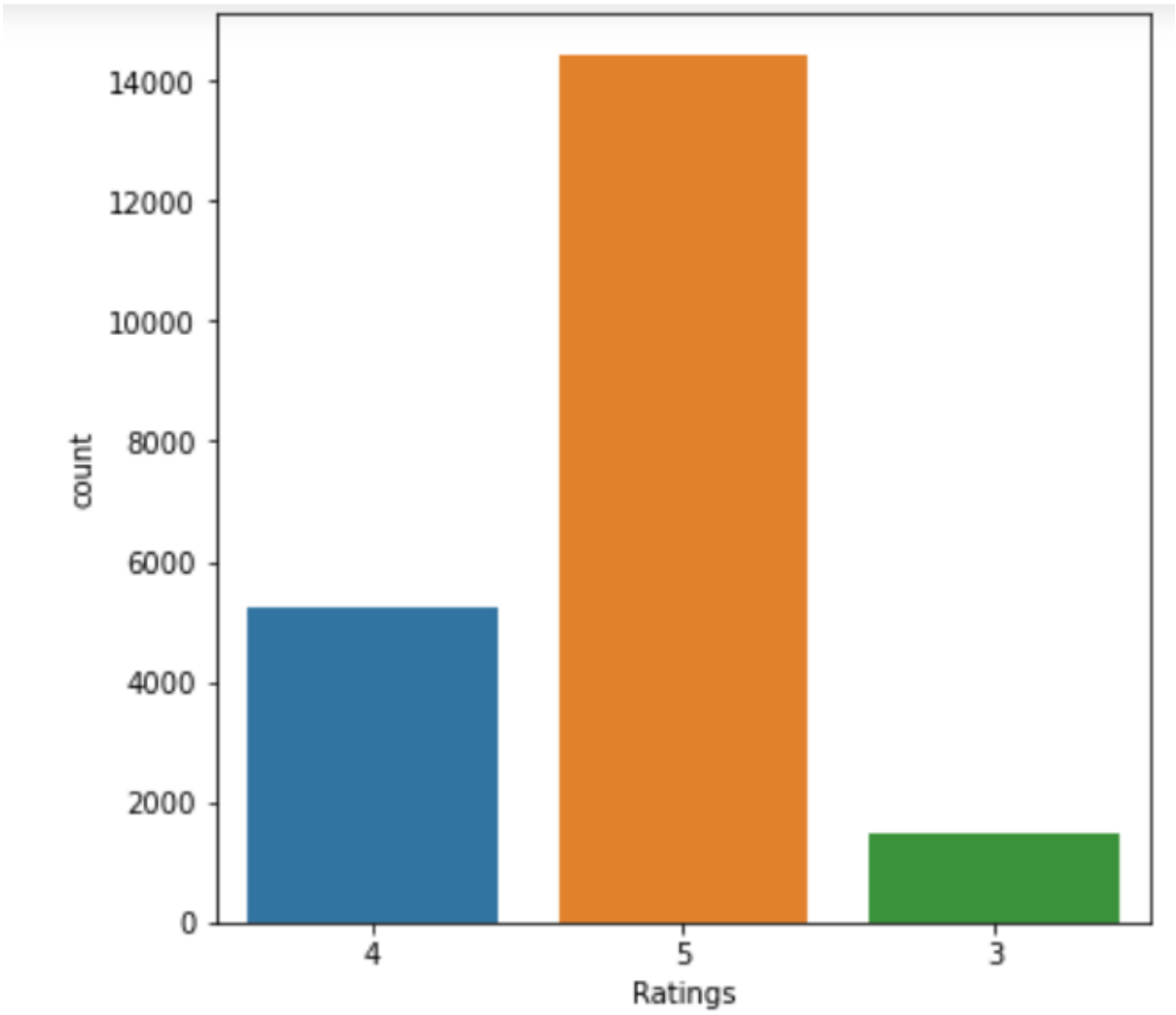
  Accuracy score gives the performance of the algorithm. Here for this project I used metrics like classification report, accuracy score. Classification report shows the precision, recall and f1 score. Which evaluates the performance of the algorithm. For this project I used three different model which gives better accuracy.

- Visualizations




```
5        12062
4         5221
None      2354
3         1491
Name: Ratings, dtype: int64
```

This the visualization of the ratings which shows the none value in the ratings.




```
5    14416
4     5221
3     1491
Name: Ratings, dtype: int64
```

Visualization after treating the ‘none’ values.

- Interpretation of the Results

  After all this observation of the project, I understand theta ratings is related to review. Ratings and review evaluates the quality of the products.

CONCLUSION

- Key Findings and Conclusions of the Study

  As I mentioned the review and ratings plays a major role in how it demands in the market. As a customer we can benefits it. We can evaluate the performance of product by checking the review and rating.

- Learning Outcomes of the Study in respect of Data Science

  Visualization gives a basic understanding of data. How it is related among other features and how features related to target. A cleaned data gives better performance. Here for this project I used three different models. All models gives better performance. Random forest gives the result in little bit slow. But among other algorithms it performs well and I select it as the final model.

- Limitations of this work and Scope for Future Work

  I am working on a project which have the pre-processing techniques used in NLP for first time. It helps me to learn more about it. We want to use multiple algorithms for a project that which we don’t know which have better performance. We need more data for prediction, more data gives better performance. So for the predictions we want to collect more data. It gives better preformnace.