RATING PREDICTION

Submitted by

MIENGANDHA SINHA

ACKNOWLEDGMENT

I would like to express my gratitude to the Company FlipRobo Technology to give this project to me. In making of this project I hereby used to take help from the references from some websites and also from which is given by the company as sample documentation and details related project and professional guided me a lot in the project and the other previous projects helped me and guided me with completion of the project.

INTRODUCTION

- Rating prediction

We all know ratings and reviews are important. While they've only been around for about two decades, it's hard to imagine shopping without them. According to consumer research we conducted of 30,000+ global shopping, the majority (**88%**) of shopping use reviews to discover and evaluate products and on based on these the customers easy to find things. The ability to successfully decide whether a review will be helpful to other customers and thus give the product more exposure is vital to companies that support these reviews, e-commercial sites like Amazon, Flipkart, Snapdeal, etc. In this research project, show a new approach to enhance the accuracy of the rating prediction by using machine learning methods the training performance of our model changes as we change the training method, the dataset used for training and the features used in the model. It help to build an application which can predict the rating by seeing the review.

- Conceptual Background of the Domain Problem

A primary objective of this project is to predict the ratings of various products by using some attributes that are highly correlated with a rate and reviews through machine learning models. So they are looking for new machine learning models from new data to help in to predict ratings for the reviews which were written in the past and they don’t have a rating. So, we have to build an application which can predict the rating by seeing the review by valuating the model. By doing some research on this project, are able to trained the models and predicting things make the previous background to work efficiently.

• Review of Project

From the collecting data phase to the model building phase gives important variables in the used to predict rating. There all types of reviews and ratings in the datasets gives various information. And then the model building do all the data visualizations, data pre-processing steps, evaluating the model, data cleaning and selecting the best model for the project.

• Motivation for the Problem Undertaken

Here, the datasets have the total of 198 entries with 3 rows, no null values, EDA has to be performed to see whether it gain or loss in the variable and its compare to the price among every aspects, to build machine learning models, to determine the optimal values of Hyper parameters and the selection of the best model, by predicting of the value can help to the clients and for the further change in the market from the new data.

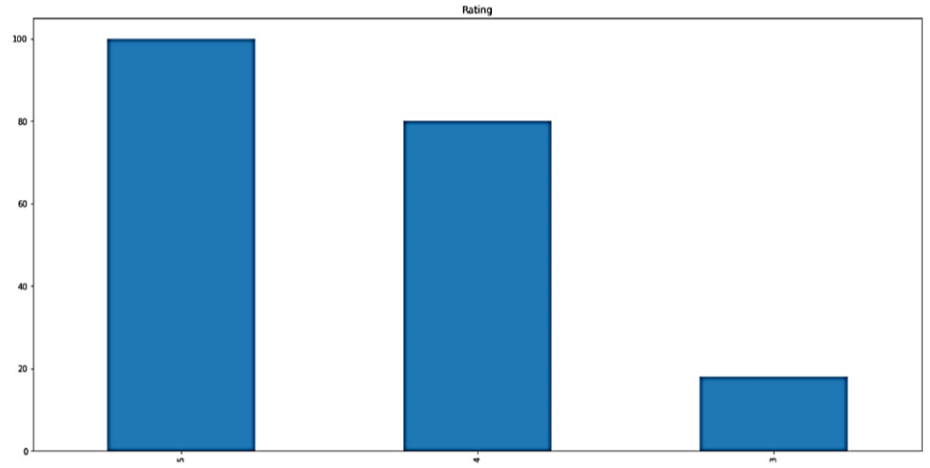
ANALYTICAL PROBLEM FRAMING

- Mathematical/Analytical Modelling of the Problem

In this project, mathematical/analytical modelling are used. Checking the null values found that having no null values in the datasets, the type of data frame is in pandas, data frame info tells that int64(2 variable), object(1 variable), by using the data visualization they are:

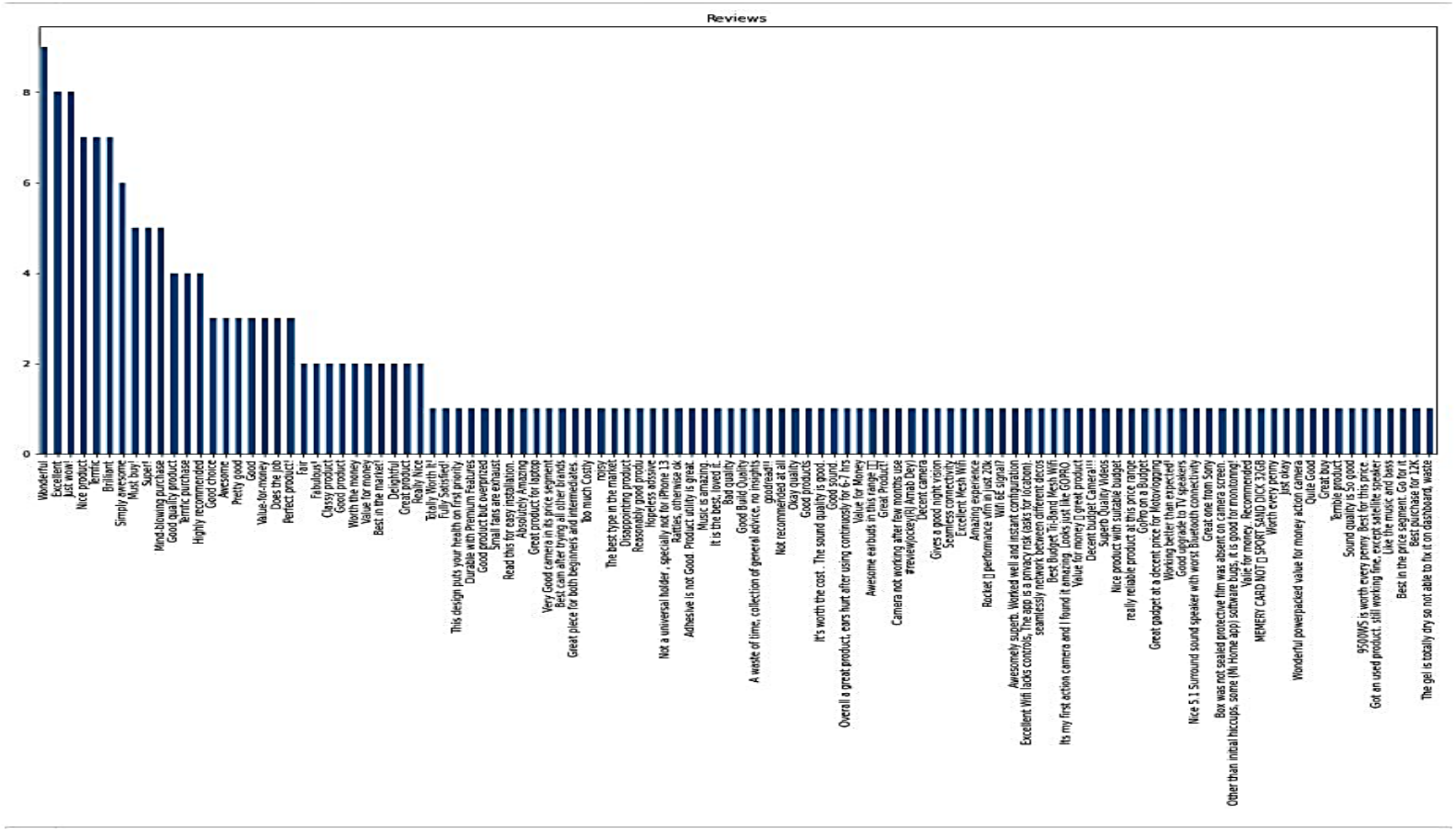
1. It is about the rating of the product:

Five Star rating has the highest accuracy among other stars

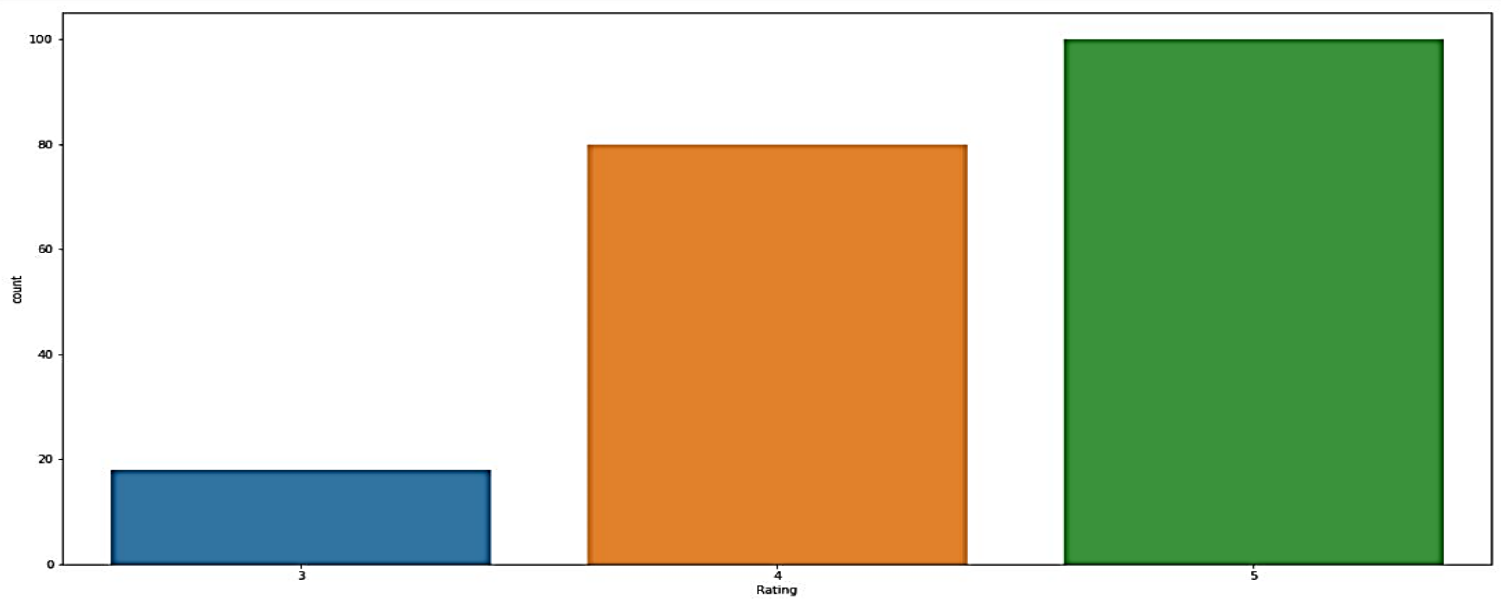


2. It is about the reviews of products:

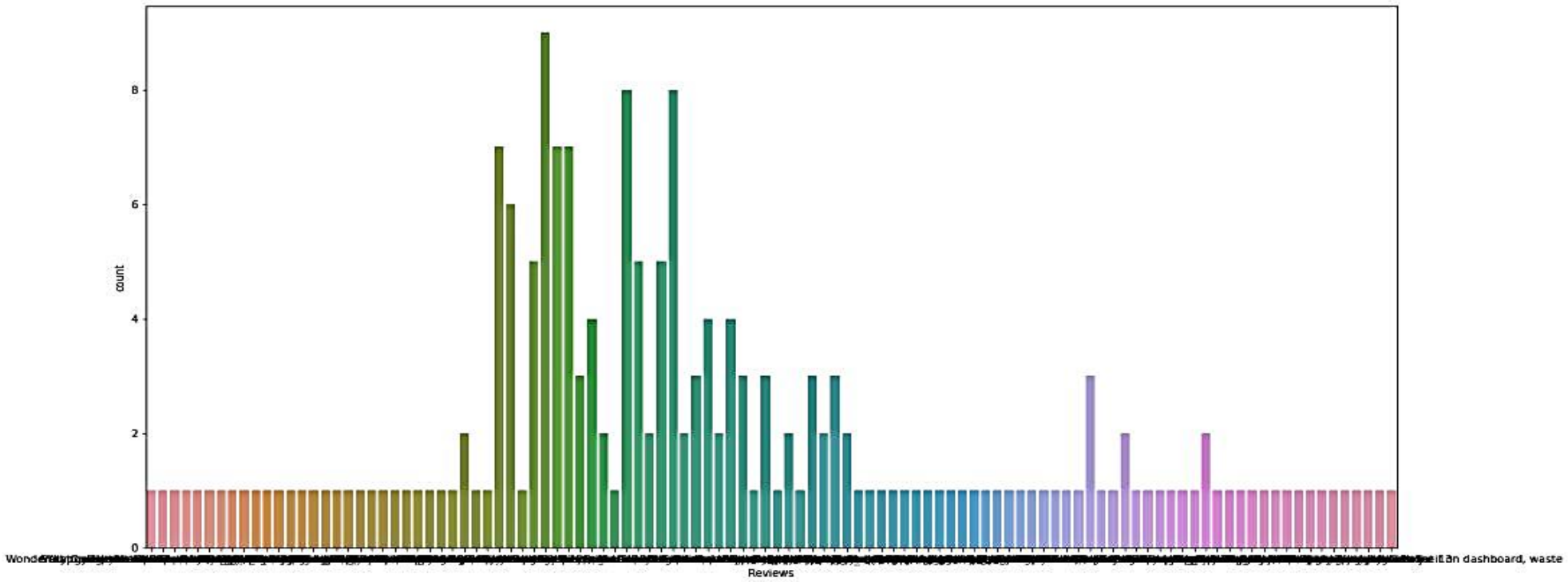
Positive review is on high as compared to others



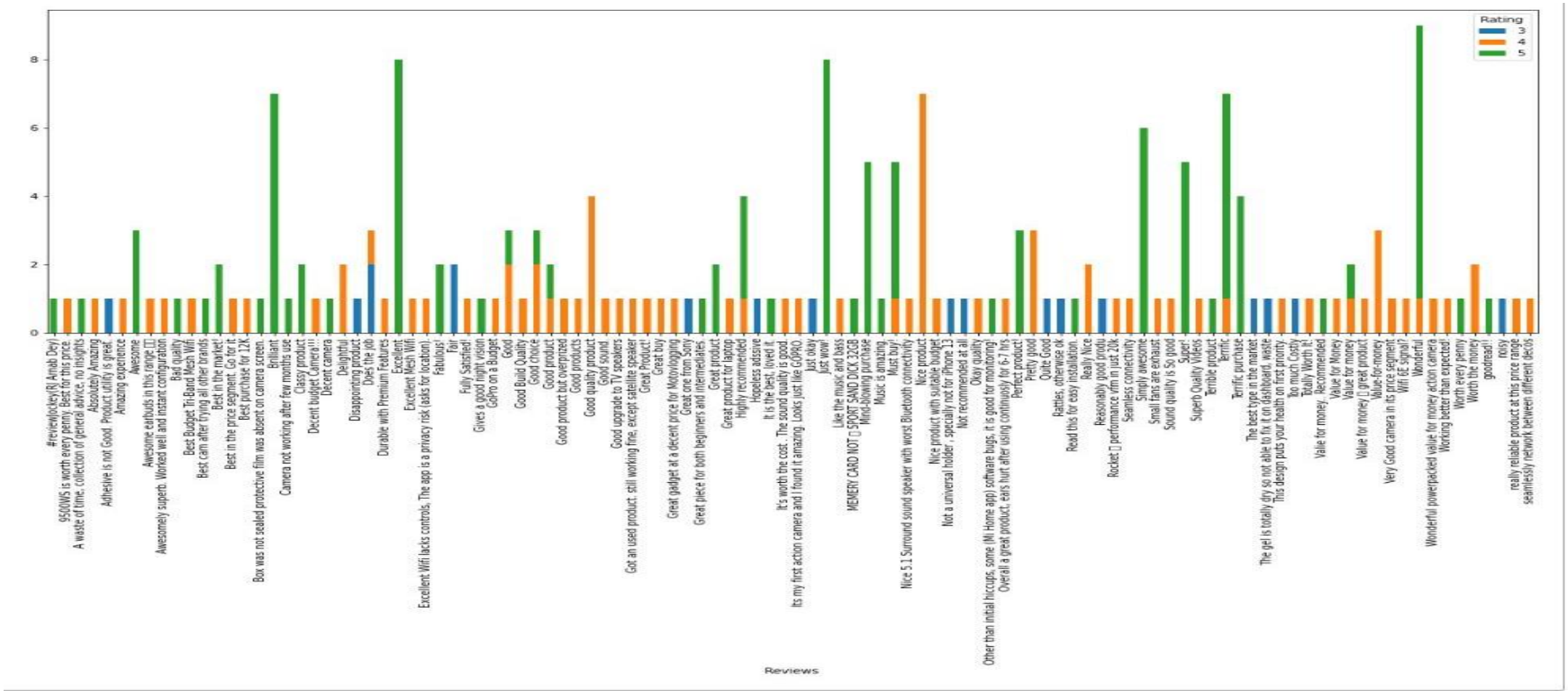
3. The count plot of the ratings of product:



4. The count plot of the reviews of product:



5.


Rating
3
4
5
0
2
4
6
8
Reviews
#reviewlockey(R) Amab Dey)
9500WS is worth every penny. Best for this price
A waste of time, collection of general advice, no insights
Absolutely Amazing
Adhesive is not Good. Product utility is great.
Amazing experience
Awesome
Awesome earbuds in this range
Awesomely superb. Worked well and instant configuration
Bad quality
Best Budget Tri-Band Mesh Wifi
Best cam after trying all other brands
Best in the market!
Best in the price segment. Go for it.
Best purchase for 12K
Box was not sealed protective film was absent on camera screen
Brilliant
Camera not working after few months use
Classy product
Decent budget Camera !!!
Decent camera
Delightful
Disappointing product
Does the job
Durable with Premium Features
Excellent
Excellent Mesh Wifi
Excellent Wifi lacks controls. The app is a privacy risk (asks for location).
Fabulous!
Fair
Fully Satisfied!
Gives a good night vision
GoPro on a Budget
Good
Good Build Quality
Good choice
Good product
Good product but overprized
Good products
Good quality product
Good sound
Good upgrade to TV speakers
Got an used product, still working fine, except satellite speaker
Great Product!
Great buy
Great gadget at a decent price for Motovlogging
Great one from Sony
Great piece for both beginners and intermediates
Great product
Great product for laptop
Highly recommended
Hopeless adssive
It is the best, loved it.
It's worth the cost. The sound quality is good
Its my first action camera and I found it amazing. Looks just like GOPRO.
Just okay
Just wow!
Like the music and bass
MEMERY CARD NOT SPORT SAND DICK 32GB
Mind-blowing purchase
Music is amazing
Must buy!
Nice 5.1 Surround sound speaker with worst Bluetooth connectivity
Nice product
Nice product with suitable budget
Not a universal holder , specially not for iPhone 13
Not recommended at all
Okay quality
Other than initial hiccups, some (Mi Home app) software bugs, it is good for monitoring!
Overall a great product, ears hurt after using continuously for 6-7 hrs
Perfect product!
Pretty good
Quite Good
Ratties, otherwise ok
Read this for easy installation
Really Nice
Reasonably good produ
Rocket performance vfm in just 20k
Seamless connectivity
Simply awesome
Small fans are exhaust
Sound quality is So good
Super!
Superb Quality Videos
Terrible product
Terrific
Terrific purchase
The best type in the market
The gel is totally dry so not able to fix it on dashboard, waste
This design puts your health on first priority
Too much Costly
Totally Worth it!
Valie for money. Recommended
Value for Money
Value for money
Value for money great product
Value-for-money
Very Good camera in its price segment
Wifi bE signal?
Wonderful
Wonderful powerpacked value for money action camera
Working better than expected!
Worth every penny
Worth the money
goodread!!
noisy
really reliable product at this price range
seamlessly network between different decos

- Data Sources and their formats

The data sources and their formats are from .csv file.

```
1 df=pd.read_csv('prediction excel sheet.csv')
2 df.head()
```

	Unnamed: 0	Rating	Reviews
0	0	4	Very Good camera in its price segment
1	1	4	Wonderful powerpacked value for money action c...
2	2	4	Great gadget at a decent price for Motovlogging
3	3	4	GoPro on a Budget
4	4	4	really reliable product at this price range

- Data Preprocessing Done

The steps followed for the cleaning of the data is Label Encoder after then importing preprocessing there transform the target columns into features then lastly it has to set/fir for the data frame.

Data Preprocessing

```
1 from sklearn import preprocessing
2 features=df.drop(['Unnamed: 0','Reviews'],axis=1)
3 target=df['Rating']
4 col_names=list(features.columns)
5 scaler=preprocessing.StandardScaler()
6 features=scaler.fit_transform(features)
7 features=pd.DataFrame(features,columns=col_names)
8 features.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Rating	198.0	-3.364312e-16	1.002535	-2.170608	-0.635678	0.899252	0.899252	0.899252

- Data Inputs- Logic- Output Relationships

  The relationships between inputs and outputs can be studied extracting weights of the trained model. Regression is that relationships between them can be blocky or highly structured based on the training data. It requires the data scientist to train the algorithm with both labeled inputs and desired outputs.

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

  Presumptions are by using regression label encoding, classifier, selection of the best models, confusion matrix that it means the relationship between the dependent and independent variables look fairly linear. Thus, our linearity assumption is satisfied.

- Hardware and Software Requirements and Tools Used
  By importing many libraries are

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import pickle
from sklearn.datasets import make_classification
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score
from matplotlib import pyplot

import warnings
warnings.filterwarnings('ignore')
```

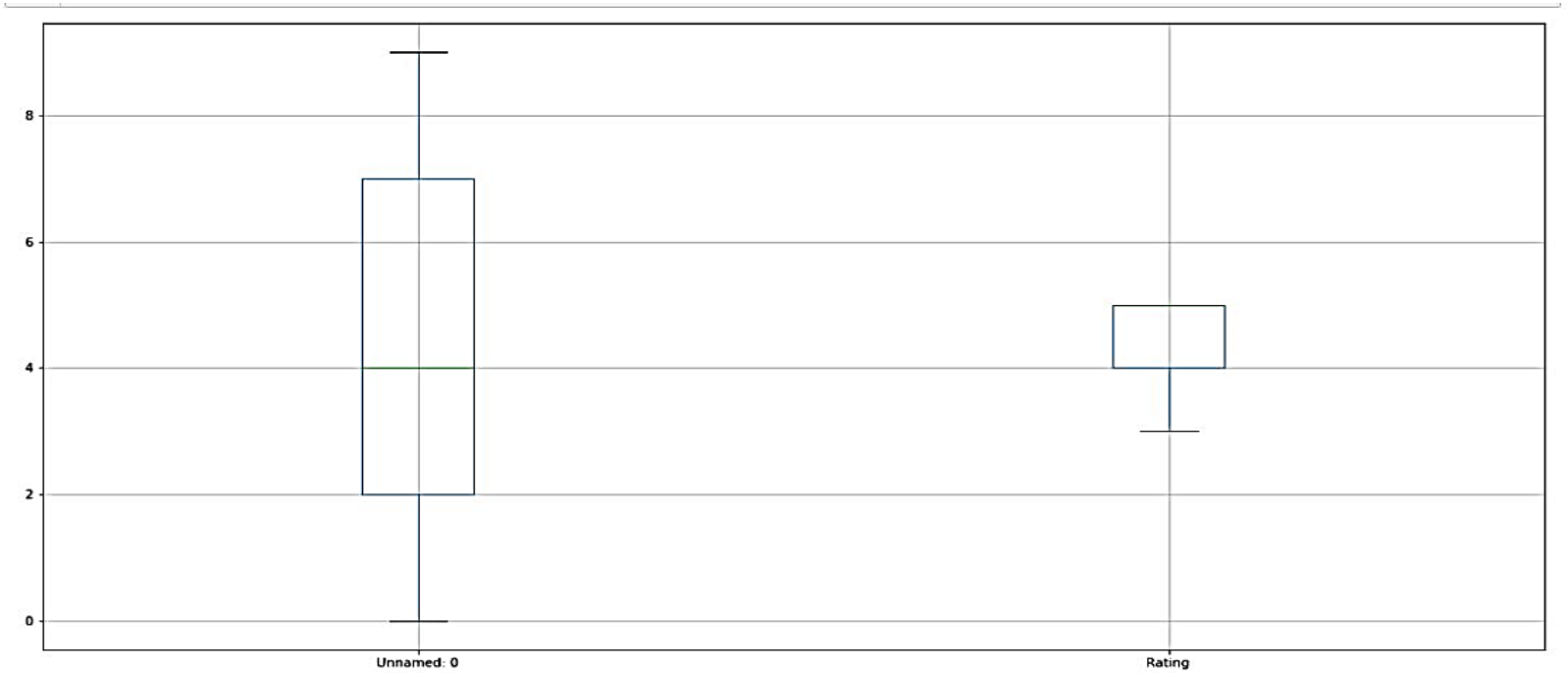
MODEL/s DEVELOPMENT AND EVALUATION

- Identification of possible problem-solving approaches(methods)

  The collection and interpretation of data in order to uncover patterns and trends. It is a component of data analytics. Statistical analysis can be used in situations like gathering research interpretations, statistical modelling or designing surveys and studies. The approaches/methods of identification are descriptive and inferential statistics which are describes as the properties of sample and population data, and inferential statistics which uses those properties to test hypotheses and draw efficient conclusions in terms of outputs.

- Testing of Identified Approaches(Algorithms)

There is no outliers.



- Run and Evaluate selected models

  DATA SCALING

```
1  from sklearn.preprocessing import StandardScaler
2  # Data Scaling Formula z=(x-mean)/std
3  scaler = StandardScaler()
4  X_scaled=scaler.fit_transform(X)
5  X_scaled
```

```
array([[-1.523999933,  1.397445572],
       [-1.17024567,  1.498367644],
       [-0.816649202, -0.351867749],
       [-0.462273837, -0.789195794],
       [-0.108984711,  1.700211477],
       [ 0.244768944,  0.287304653],
       [ 0.598522626 ,  0.926476784],
       [ 0.952276251, -1.226524094],
       [ 1.306029994 ,  1.330164444],
       [ 1.659783563, -0.049101741],
       [-1.523999933, -1.529289844],
       [-1.17024567,  1.733852111],
       [-0.816649202, -0.957398981],
       [-0.462273837, -1.596571117],
       [-0.108984711,  1.431086361],
       [ 0.244768944,  0.724632951],
       [ 0.598522626 , -1.697493031],
       [ 0.952276251, -0.991039621],
```

GRADIENT BOOSTING CLASSIFIER:

Accuracy score: Train result: 94.2%

Test result: 63.3%

```
==========Train Result==========
Accuracy Score: 94.20289855072464
_____
CLASSIFICATION REPORT :
                3         4         5  accuracy  macro avg  weighted avg
precision  0.937500  0.962963  0.926471  0.942029   0.942311      0.943007
recall     1.000000  0.896552  0.969231  0.942029   0.955261      0.942029
f1-score   0.967742  0.928571  0.947368  0.942029   0.947894      0.941683
support   15.000000 58.000000 65.000000  0.942029 138.000000    138.000000
_____
Cofusion Matrix:
[[15  0  0]
 [ 1 52  5]
 [ 0  2 63]]
```

```
==========Test Result==========
Accuracy: 63.33333333333333
_____
CLASSIFICATION REPORT :
                3         4         5  accuracy  macro avg  weighted avg
precision  0.0  0.526316  0.777778  0.633333   0.434698      0.646686
recall     0.0  0.454545  0.800000  0.633333   0.418182      0.633333
f1-score   0.0  0.487805  0.788732  0.633333   0.425512      0.638956
support    3.0 22.000000 35.000000  0.633333  60.000000     60.000000
_____
Confusion Matrix:
[[ 0  2  1]
 [ 5 10  7]
 [ 0  7 28]]
```

```
# hyperparameter tuning
from sklearn.model_selection import GridSearchCV
grid_param = {
    'max_depth' : range(4,8),
    'min_sample_split' : range(2,8,2),
    'learning_rate' : np.arange(0.1,0.3)
}
```

Selecting the best model

```
:
print('The accuracy of the KNN Model is 0.6166666666666667')
print('The accuracy of the Random Forest Model is  0.65')
```

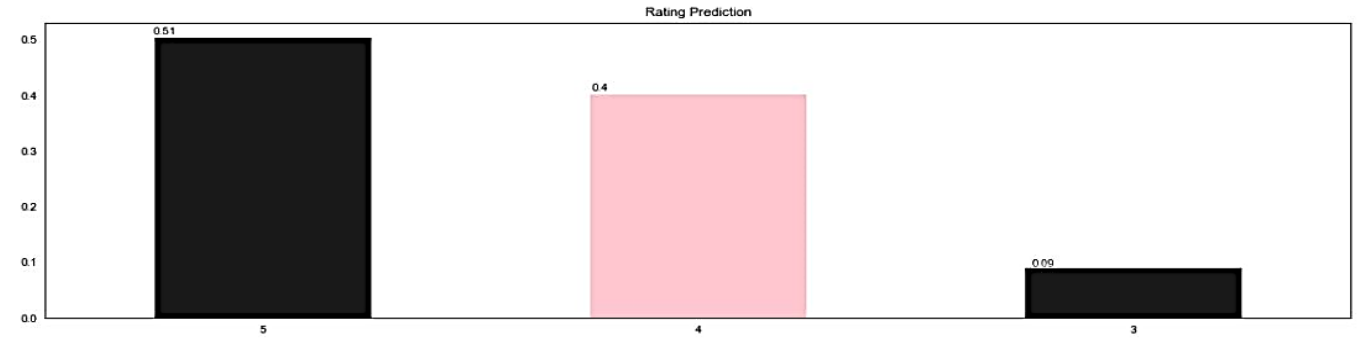
```
The accuracy of the KNN Model is 0.6166666666666667
The accuracy of the Random Forest Model is  0.65
```

- Visualizations

RESAMPLING THE LABEL

```
plt.figure(figsize=(20,5))
ax=df.Rating.value_counts(normalize=True).plot(kind='bar', color=['black', 'pink'], alpha=0.9, rot=0)
plt.title('Rating Prediction')
for i in ax.patches:
    ax.annotate(str(round(i.get_height(),2)),(i.get_x() * 1.01, i.get_height() * 1.01))

plt.show()
```



USING THE HEATMAP FOR THE DATASET

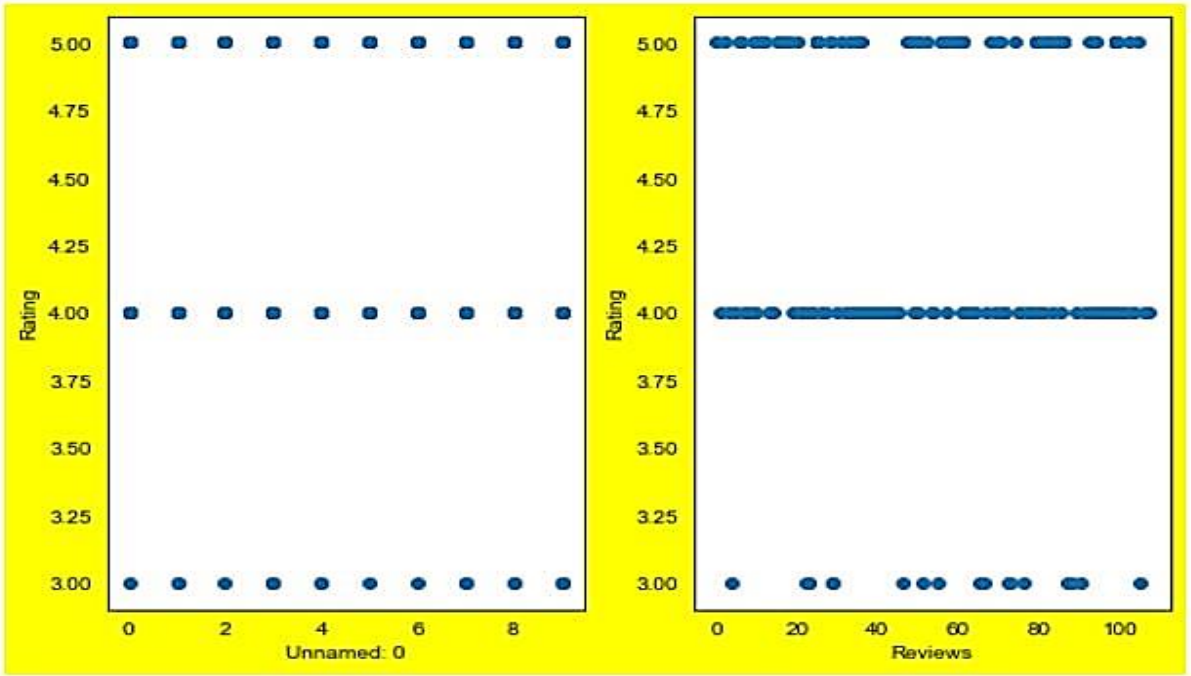
```
plt.figure(figsize=(20,5))
sns.heatmap(df.isnull(),cbar=False,cmap='twilight')
```

```
<AxesSubplot:>
```

VISUALIZING THE RELATIONSHIPS

```
# Visualizing relationship
plt.figure(figsize=(15,10), facecolor= 'yellow')
plotnumber = 1

for column in X:
    if plotnumber<=8 :
        ax = plt.subplot(2,4,plotnumber)
        plt.scatter(X[column],y)
        plt.xlabel(column,fontsize=10)
        plt.ylabel('Rating',fontsize=10)
    plotnumber+=1
plt.tight_layout()
```



The relationship between the dependent and independent variables look good in linear. Thus, our linearity assumption is satisfied.

- Interpretation of the Results

The results were interpreted from the visualizations, preprocessing and modelling:

1. The purpose of this case is to understand the rise of e-commerce which brought a significant rise in the importance of customer reviews and evaluate used to predict rating and to develop a strategy that utilizes data mining techniques towards ratings prediction.
2. There are two main methods to approach this problem. The first one is based on review text content analysis and uses the principles of natural language process (the NLP method). This method lacks the insights that can be drawn from the relationship between costumers and items.
3. The model of the independent variables and dependent variables are exactly vary with the variables.
4. It can accordingly manipulating the strategy of the areas that will yield high returns as it make easier for the clients.
5. By visualizations there are many things to be noted when it will according to work each other
6. By preprocessing the data it means that from the help of label encoder helps the dataset column to transform to fit another column in to it.

CONCLUSION

In conclusion, combining the formerly know data about each user' similarity to other users with the sentiment analysis of the rating and reviews itself, does help to improve the model prediction of rate the user's review, will get the purpose.

THANK YOU