Ratings Project

Submitted by:

ABHISHEK PAI

ACKNOWLEDGMENT

The background information relating to the project was been provided by fliprobo as a part of the internship phase.

The data was collected from various websites to aid this project.

Related guidance was been provided by fliprobo for the completeion of this project

INTRODUCTION

- Business Problem Framing
  To predict ratings for the reviews which were written in the past and that don't have a rating. To solve this problem building an application that can predict the rating by seeing the review.
- Conceptual Background of the Domain Problem
  A client who has a website where people write different reviews for technical products. 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.
- Review of Literature
  There is not much research performed as the Data and related information was provided by the source itself, which was been taken into consideration based on the information given by Flip Robo.
- Motivation for the Problem Undertaken
  The Project was assigned by flip Robo as part of the internship phase for better understanding the concept and getting the idea of the industry.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  After importing data various analyses were performed which had univariate, bivariate, and multivariate analysis.

  Univariate analysis: Univariate analysis is the simplest form of analyzing data. It doesn't deal with causes or relationships and its major purpose is to describe; It takes data, summarizes that data, and finds patterns in the data.

  Bivariate analysis: Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables, to determine the empirical relationship between them. Bivariate analysis can help test simple hypotheses of association.

  Multivariate analysis: Multivariate statistics is a subdivision of statistics encompassing the simultaneous observation and analysis of more than one outcome variable. Multivariate statistics concerns understanding the different aims and backgrounds of each of the different forms of multivariate analysis, and how they relate to each other.

- Data Sources and their formats

  After loading the data, the information of data was been checked and a five-row sample was been observed.

- Data Pre-processing Done

The entire data was in form of CSV and was a mixture of numbers, and objects. The output variable is information in a pattern of 1,2,3,4,5 each having its significant meaning. The output was based on the data which was provided by a source on the behavioural pattern of the entity. The object part was been converted and extracted to perform ML

- Hardware and Software Requirements and Tools Used

  The system with a 16 core processor was been used,
  The operating system was Windows 10,
  Anaconda 3 was been used for performing ML
  Libraries:

```
import pandas as pd
import selenium
from selenium import webdriver
import time
from selenium.common.exceptions import
StaleElementReferenceException, NoSuchElementException
import urllib
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings # Ignores any warning
warnings.filterwarnings("ignore")
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer

nltk.download('stopwords')
import wordcloud
from wordcloud import WordCloud
import re
from pylab import rcParams
```

Model/s Development and Evaluation

.

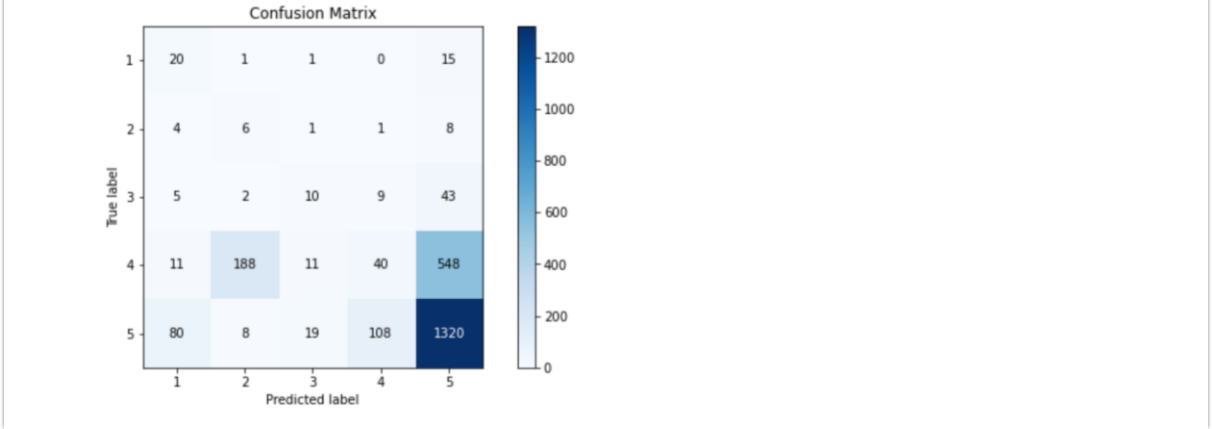
- Testing of Identified Approaches (Algorithms)

  Random Forest Classifier

  .

- Run and Evaluate selected models

	precision	recall	f1-score	support
1	0.17	0.54	0.25	37
2	0.03	0.30	0.05	20
3	0.24	0.14	0.18	69
4	0.25	0.05	0.08	798
5	0.68	0.86	0.76	1535
accuracy			0.57	2459
macro avg	0.27	0.38	0.27	2459
weighted avg	0.52	0.57	0.51	2459

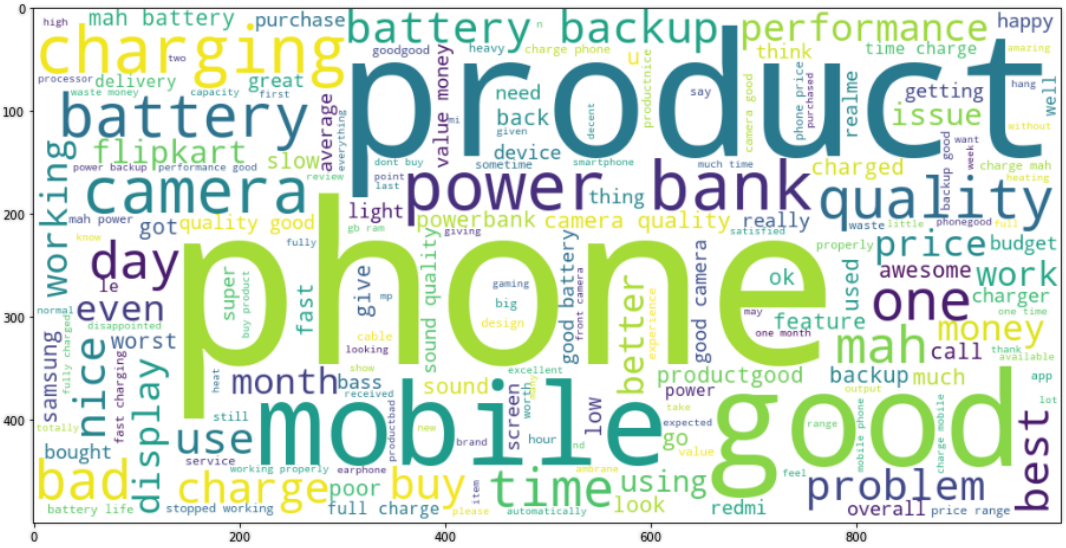


- Key Metrics for success in solving problem under consideration

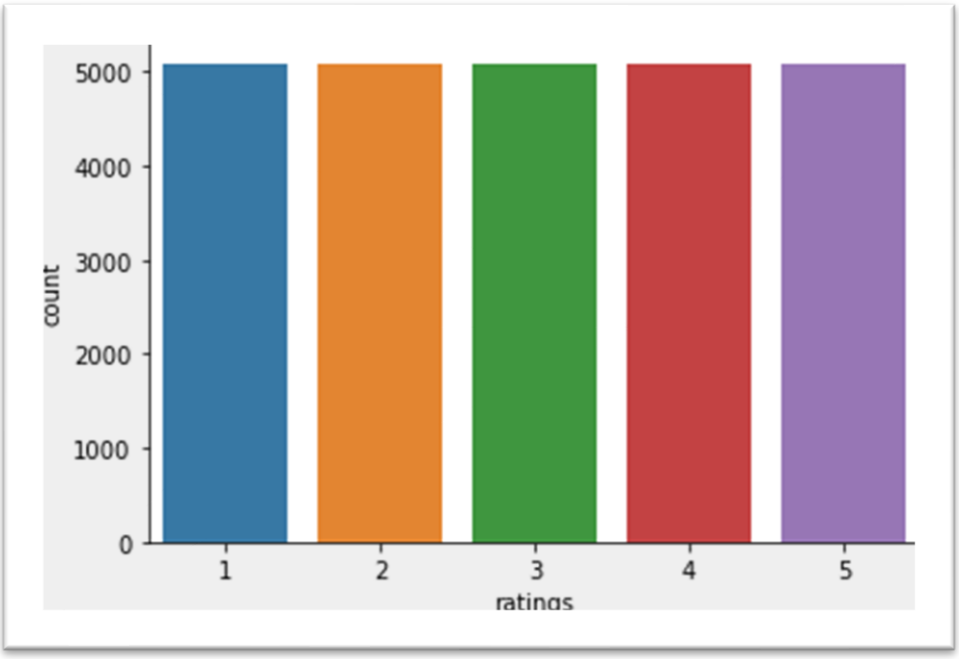
  Accuracy score is **used** to evaluate the performance of a regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable.

- Visualizations

  Bag of words



  Balanced dataset



- Interpretation of the Results

  1 model has been used

  The random forest has performed better after gridsearch cv.

  The finalized model is Random Forest.

  .

CONCLUSION

- Key Findings and Conclusions of the Study
- The project is based on doing sentiment analysis on review and providing related ratings. random forest classifier has been used which provided 57 % accuracy and as per tests, the model provides a fair accuracy. Although the accuracy can be improved by adding a lot of data to help the model understand the ratings based on reviews there is a huge scope of improvement.
- Learning Outcomes of the Study in respect of Data Science

  Adding more data can help to increase the accuracy.

- Limitations of this work and Scope for Future Work

  There is a lot of scopes, more tweaks in a model can help to get better results.