Review Rating Prediction

Submitted by:

PAVANKUMAR BOLL

ACKNOWLEDGMENT

I would like to express my special thanks to flip robo technlogies for given me such a wonderfull of apportunity to explore and insite work on a data science projects.

And and also thanks to Datatraiend institute's mentors to motivate and understand the concepts of Data science. I have got lot of help from their resourses and concept to complete these project.

Link that helps me to complete this project

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

INTRODUCTION

- ## Business Problem Framing

  We have a client who has a website where people write different reviews for technical

  products. Now 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. Now they want 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.

- ## Conceptual Background of the Domain Problem

  Data is gather from different different e-commerce web site’s. Data is scrap by using selenium. Selenium is one of best testing tool that i’ve used for to scrap data from the e-commerce website’s html page. Data have the three columns i.e Rating Heading, Rating and Review. So i have a scrap a data from such as amazon and flipkart. I have scrap review and rating approx 200 rows both positive and negative review and also i have handle to scrap close to equal number of ratings like 1 ,2,3,4,5 star from each products. Product like laptops, monitots, headphones, smart wathches, mobiles, routers and home theaters.

- Motivation for the Problem Undertaken

  Motivation of this project is to determine the rating form out of 5 star on the basis of review written by user. This also help’s to determine the product quality and type of review that is towards to positive or negative. On the basis of rating we can conclude the quality of product if product has much number of 5 star rating’s means product was awesome if average rating is 3 or 4 star means product has some positive and some negative wich has to improve or rectifies those issues.

Analytical Problem Framing

- Data Sources and their formats

  In this project i've scrap almost 1 lack 14 thousand rows. So i have took a sample data upto 35 thousand rows and worked on it. This sample data can be inference to eniter data. While scrappind the review some of rows got null values or got null values in rating column. I have drop those columns which has null or having string like "-". Firstly i have created different different csv files for each product with respect to web sites. Created on data frame in which i have merged all the data sets and created one sigle data.

- Data Preprocessing Done

  In this project each data is in text formate machine can understand only numbers. So for this i have used tool kit of NLP (Natual Language Processing) that is NLTK. In the part of data cleaning i have removed stopwords, punctuation, emogis, numbers extra spaces, words length have less than 2, and removed email addresses etc. Because this type of data is not use full for the prediction rating this type of words are increasing the length of the text. After cleaning the data i have stem each words to give the base words and remove plural's and tenses. For stemming the words i have used Lemmatization package which is in NLTK. After all these steps i have converted text to numbers by using TFIDF and Count Vectorizer

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

- ## Data Inputs- Logic- Output Relationships

  Review’s are in text formate and in any certain language’s. Based on words present in the review rating get predicted. Suppose in review word’s like excellent great awesome this type of words represents product was really good and it will give a raitng 4or 5 star. Same as neagtive words will give rating 1 or 2 star.

- ## Hardware and Software Requirements and Tools Used

  Tools and libraries that i have used to solve project:

  - Software: Anaconda, Jupyter Notebook, Python3, Google Colab.
  - Libraries: Numpy, Pandas, Matplotlib, Seaborn, Sklearn.

Model/s Development and Evaluation

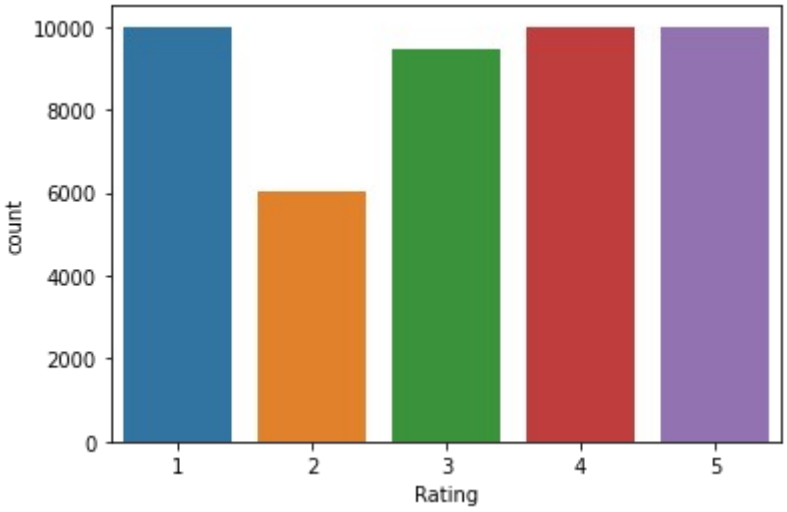
- Identification of possible problem-solving approaches (methods)

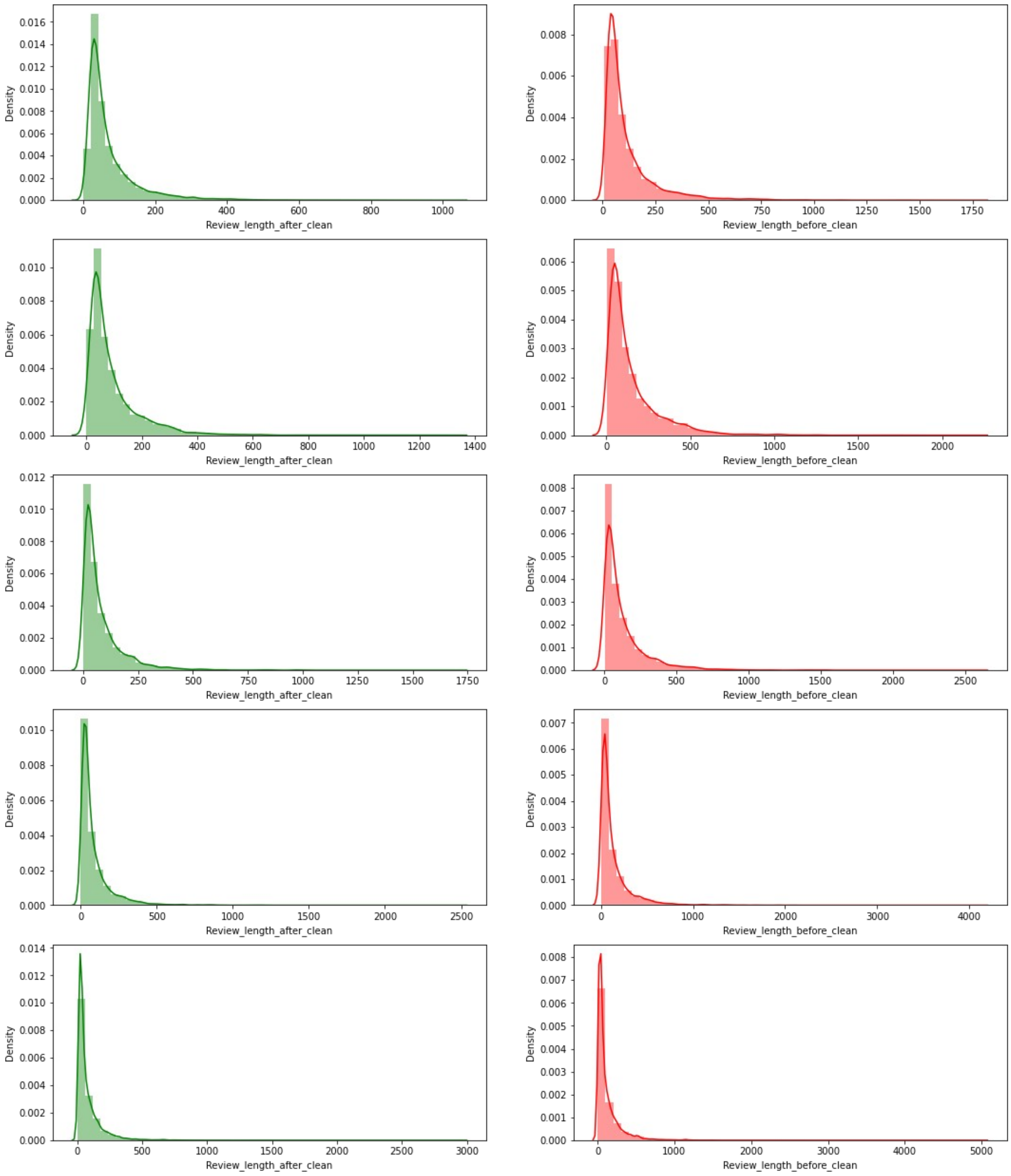
  The data set contain lable data and target data lable has cleaned by some EDA steps. Target data is in the form of classified i.e (Rating). By seeing the our target column we can conclude that this problem comes under supervised and Classification Problem. Target column having in factor . For more testing SVC, Decision Tress Classifier, KNN algorithms have used. Rather testing on one by one algorithm and choosing best alogithm and hyper tuning that best model. I have hyper tunned all the algorithm and added into pipeline so i will get best algorithm along with the best parameters.

- Testing of Identified Approaches (Algorithms)

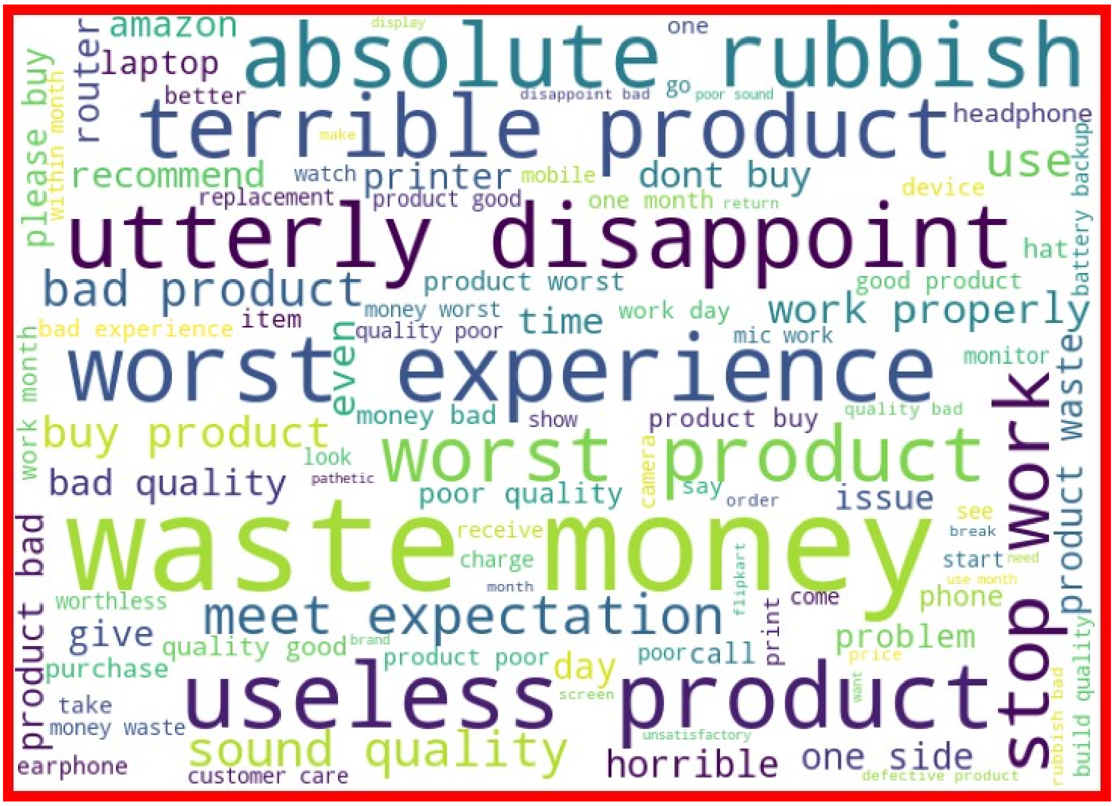
  For testing how good model is generated. I have splitted data set into two parts (70% 30%) and each part is splited into further two part that are lable data and target data. First 70% splited data is used for training the model and another rest of 30% used for Prediction testing the model how good model is predicted in terms of percentage.

- Visualizations





Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0
200
400
600
800
1000
Review_length_after_clean
Density
0.000
0.002
0.004
0.006
0.008
0
250
500
750
1000
1250
1500
1750
Review_length_before_clean
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
200
400
600
800
1000
1200
1400
Review_length_after_clean
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0
500
1000
1500
2000
Review_length_before_clean
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0
250
500
750
1000
1250
1500
1750
Review_length_after_clean
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0
500
1000
1500
2000
2500
Review_length_before_clean
Density
0.000
0.002
0.004
0.006
0.008
0.010
0
500
1000
1500
2000
2500
Review_length_after_clean
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0
1000
2000
3000
4000
Review_length_before_clean
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0
500
1000
1500
2000
2500
3000
Review_length_after_clean
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0
1000
2000
3000
4000
5000
Review_length_before_clean



absolute rubbish
terrible product
utterly disappoint
worst experience
worst product
waste money
useless product
stop work
meet expectation
sound quality
bad product
work properly
buy product
bad quality
poor quality
dont buy
amazon
laptop
router
please buy
recommend
printer
use
headphone
one
better
go
disappoint bad
poor sound
make
watch
mobile
device
replacement
product good
one month
return
within month
hat
battery backup
good product
product worst
time
work day
money worst
item
bad experience
quality poor
mic work
even
monitor
work month
money bad
show
product buy
quality bad
product waste
look
camera
pathetic
say
order
issue
see
receive
break
product bad
charge
start
need
flipkart
month
use month
worthless
come
phone
print
give
quality good
brand
product poor
problem
purchase
poor
call
price
day
take
screen
want
build quality
money waste
unsatisfactory
bad
rubbish
horrible
one side
earphone
customer care
defective product
display


use
buy
good
bad quality
product
better product
expect better
slightly disappoint
amazon
worst product
replacement
product bad
make
speaker
option
feel
laptop
good product
thing
bass
recommend
quality poor
slow
please
low
provide
think
camera
really
range
average
still
router
great
lot
app
say
purchase
issue
printer
work
problem
moderate
feature
mobile
screen
sound quality
quality good
performance
try
worth
bad product
come
day
battery
waste money
watch
headphone
order
show
signal
receive
need
look
price
display
month
poor
start
money
first
call
less
speed
device
one
sound
return
bluetooth
phone
monitor
product good
give
high
stop work
work properly
replace
nice
time
print
wifi
hour
connect
much
worst
best
even
charge
way better
review


connect
poor
less
nice
think
much
thing
best
backup
average
product
good
app
hour
great
good nice
speed
slow
mobile
problem
screen
good product
satisfy
feature
battery
buy
good good
fair
range
provide
wifi
purchase
watch
give
bad
fair good
okay good
feel
nice product
sound quality
price
option
use
overall
price point
speaker
device
good
quality good
make
nice good
say
even
job good
worst
headphone
receive
laptop
battery life
job
better
work
value money
low
okay
product
range good
print
come
phone
good range
printer
price range
pro
disappoint
camera
day
issue
take
time
charge
decent product
good decent
bass
sound
well
call
good sound
look
need
one
design
show
monitor
game
router
expect
display
build quality
sometime
product nice
performance

CONCLUSION

- Learning Outcomes of the Study in respect of Data Science

  I have learn some inbuild funtion of python at the time of data cleaning. Object columns to Integer are done by using python map function over on lambda function. The main challenging is that to run all model once and get the best model name with best parameters also with a AUC ROC curve plot. This is done by using pipeline. This will helps me to get out from overfitting and underfitting.