Review Predection

Submitted By –

Ganesh Kumbhar

Acknowledgement

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I respect and thank Ms, Khusboo Garg for providing me an opportunity to do the project work in FlipRobo Technologies and giving us all support and guidance which made me complete the project duly. I am extremely thankful to him for providing such a nice support and guidance.

I owe my deep gratitude to everyone ho guided me.

- *FlipRobo Technologies*
- *DataTrained Academy*
- *Krush Naik*
- *Code With Harry*
- *Medium.com*

Some mentors helped me with their research work mentioned as follow -

Using Naïve Bayes Model and Natural Language Processing for Classifying

Messages on Online Forum

I am thankful to and fortunate enough to get constant encouragement, support and guidance, which helped me in successfully completing this project work.

INTRODUCTION

Business Problem Framing

You were recently hired in a start up company and you were asked to build a system to identify spam emails.

Perform all necessary actions not only limited to,

1. Data Preparation
2. Building word dictionary
3. Feature extraction
4. Training classifiers
5. Testing
6. Performance evaluation using multiple metrics (Confusion matrix, f1 score, roc, auc)

Conceptual Background of the Domain Problem

Natural Language processing or NLP is a subset of Artificial Intelligence (AI), where it is basically responsible for the understanding of human language by a machine or a robot.

One of the important subtopics in NLP is Natural Language Understanding (NLU) and the reason is that it is used to understand the structure and meaning of human language, and then with the help of computer science transform this linguistic knowledge into algorithms of Rules-based machine learning that can solve specific problems and perform desired tasks.

Review of Literature

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.

Motivation for the Problem Undertaken

This is a Natural Language Processing Problem and this will lead us to create highly communicative machines using human-native language.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

- The Problem is of Classification.
- The Data consists of 3 features in the dataset.
- There are more than 22K samples in the dataset.
- Use appropriate algorithms to build up the model.
- The data is in English language which also consists of numbers and special characters.
- The dirty data should be cleaned in order to retrieve meaning from the data
- There are 62 records with missing values in the dataset.
- Also create word dictionaries and wordcloud for further and future analytics.
- The target classes has 20:80 ratio for spam: ham
- The target has 5 classes only, it is a binary classification problem. Using appropriate metrics for scoring and evaluations .

Data Sources and their formats

We scrapped the databy using web scrapping technique

Data Preprocessing

The Data pre-processing done is as follows:

1. Removing Stop words from the data.
2. Removing punctuations and other special characters from the records
3. Some more granular cleaning for treating hyphen and underscore joined words.
4. Removing the words which are less than 3 letters in length 5. Perform Stemming using PorterStemmer class from sklearn library
6. Perform Lemmatizing using WordNet class from sklearn library
7. Further, we remove all the words which do not convey any meaning in the context of the English Language
8. Vectorize the data using tf-idf Vectorise

Data Inputs- Logic- Output Relationships

Data is fed in the form of a Pandas data frame to the model. The data is the vectorised meaningful words of the records. For the output we get the predicted label value of the record, that is whether the document is likely to be the same email or not. The output results in a binary value either 1 or 0 respectively.

Hardware and Software Requirements and Tools Used

Software

- ➔ Jupyter Notebook (Python 3.8)
- ➔ Microsoft Excel
- ➔ Microsoft Word

Hardware

- ➔ Processor - Intel i5 9th Gen
- ➔ RAM - 8 GB
- ➔ Graphic Memory - 4Gb , Nvidia 1060

Libraries

- ➔ Pandas
- ➔ Numpy
- ➔ Matplotlib
- ➔ Seaborn
- ➔ Scipy
- ➔ Sklearn
- ➔ NLTK

Model/s Development and Evaluation

Testing of Identified Approaches.

- DecisionTreeClassifier
- GaussianNB
- LogisticRegression
- KNeighborsClassifier

- GaussianNB
- RandomForestClassifier
- AdaBoostClassifier
- GradientBoostingClassifier
- BaggingClassifier
- ExtraTreesClassifier

Run and Evaluate selected models

First we used ML Models

In [76]:
```
#ML Models
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB

#model selection
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score

#metrics
from sklearn.metrics import accuracy_score,confusion_matrix,classification_report,roc_curve,roc_auc_score,f1_score
```

Then we made a for loop to test and print the result of both the two models

In result we got all the values of matrices then with our choice of matrix, we selected the model

we see that KNeighborsClassifier has accuracy score = 96%, f1 score = 96% , auc roc score = 94% and average cross validation score = 96% but this can further be increased to we will first see all ensemble technique

Then we used different ensemble bagging and boosting method

```
from sklearn.ensemble import
RandomForestClassifier,
AdaBoostClassifier ,
GradientBoostingClassifier,
BaggingClassifier,
ExtraTreesClassifier
```

Again we used for loop to test different models and get the result

```
#testing Different Models
model=[RandomForestClassifier(), AdaBoostClassifier() , GradientBoostingClassifier(), BaggingClassifier(), ExtraTreesClassifier(]

for i in model:
    i.fit(x_train,y_train)
    pred=i.predict(x_test)
    print(i)
    #Printing Model Score
    print('Model Score =', i.score(x_train,y_train))
    print('\n')

    #Printing Accuracy Score
    acc = accuracy_score(y_test,pred)
    print('Accuracy Score = ', acc )
    print('\n')

    #Printing Confusion Matrix
    print('Counfusion Matix')
    print(confusion_matrix(y_test,pred))
    print('\n')

    #Printing Classification Report
    print('Classification Report')
    print(classification_report(y_test,pred))
    print('\n')

    #Printing AUC ROC Score
    roc_auc = roc_auc_score(y_test,pred)
    print('Auc Roc Score =', roc_auc )
    print('\n')

    #Printing Cross Validation Score
    cross = cross_val_score(i,x,y,cv=5).mean()
    print('Cross Val Score =' , cross )
    print('\n')
    print('-------------------------------------------------------------------------------------------------------------------')
```

We got list of result with all the ensemble technique
we see that extra tree classifier has best accuracy score, f1 score, auc_roc score ann cross val score

so we will process with **Adaboost Classifier**

Key Metrics for success in solving problem under consideration

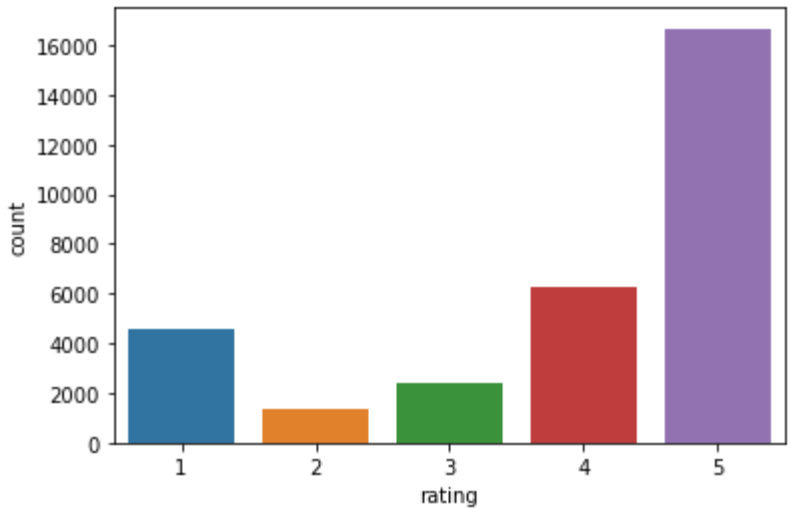
1. **Accuracy Score** - Dataset is balanced so first we will look at accuracy score shows best result when it comes to balance the dataset. 2. **F1 - Score** - in this data set the target will decide the message is spam or not, Hence 0 and 1, and as both zero and one is important to us therefore recall and precision what will be our preferred metric and as we all know that , it combines precision and recall into one metric by calculating the harmonic mean between those two.and preferred metric is F1 score.
3. **AUC ROC** - We can see a healthy ROC curve, pushed towards the

top-left side both for positive and negative classes. It is not clear which one performs better across the board as with FPR = 0.15 positive class is higher and starting from FPR= 0.15 the negative class is above.In order to get one number that tells us how good our curve is, we can calculate the Area Under the ROC Curve, or ROC AUC score. The more top-left your curve is the higher the area and hence higher ROC AUC score.

4. **Cross Validation Score** - to check if our model is overfitting or not we use cross validation score, higher the cross validation score higher the cross validation score means the model is not overfitting.

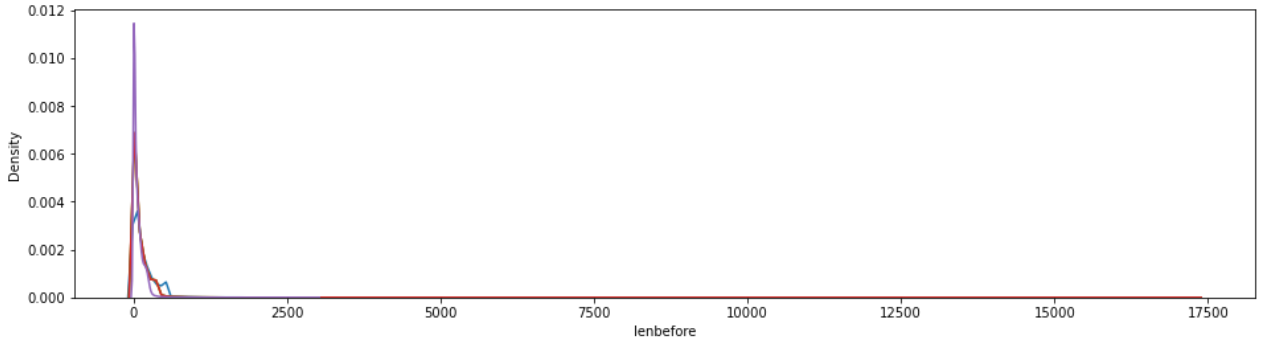
Visualizations

Label -

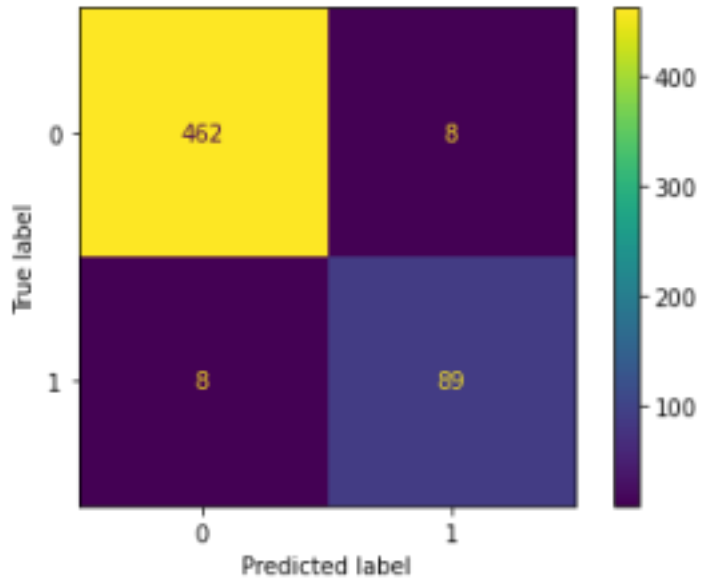


Imbalance of the target classes

KDE plot of length of records after various processing step wise



Confusion Matrix on the test data



Interpretation of the Results.

When we see the result of matrices we see that the we see that extra tree classifier has best accuracy score, f1 score, auc_roc score ann cross val score

This is our final model with

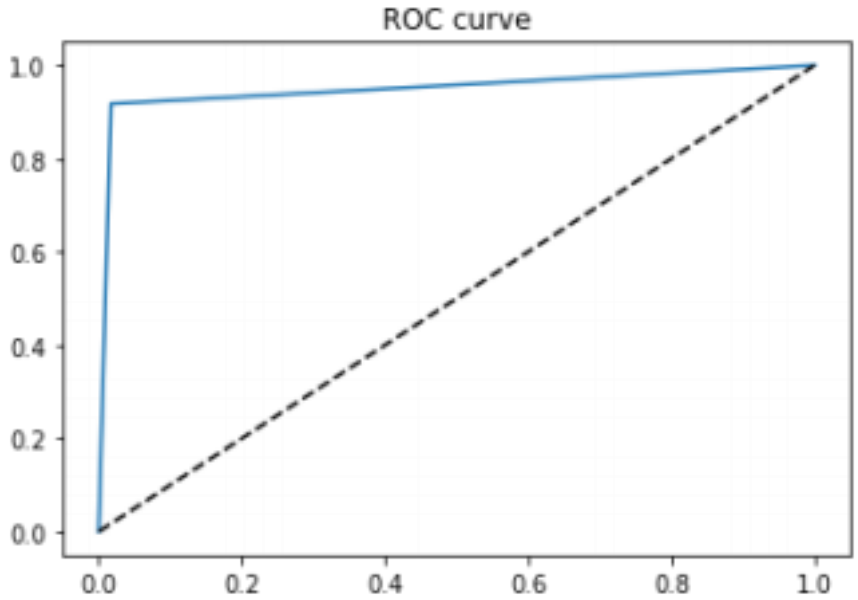
Accuracy = 97%

F1-Score = 97%

Auc Roc Score = 95%

Cross Val Score = 98%

and sharp AUC_ROC_Curve



CONCLUSION

Key Findings and Conclusions of the Study

1. NLP gets hard as humans are not used to typing as proper grammar these years.
2. Sweet spots should be found between whether to pick stemming or lemmatization or both.

3. Adaboost algorithms are quicker than rest of the algorithms.

Learning Outcomes of the Study in respect of Data Science

1. Almost 90 percent of the time is spent of data cleaning and data modelling.
2. You do not get a Gaussian distribution in real-word problem. NLP becomes difficult due to sloppy use of language by humans
3. This also created issues while teaching machines may take a long time to converge on a Huge dataset like this.

Limitations of this work and Scope for Future Work

1. More data is always appreciated
2. The model could be integrated with any email app used by the Data Analysts and Developers to predicit rating.
3. The model could be placed into a Continuous Integration and Continuous Deployment for an online training environment.