

Email Classification

Submitted by:

Amey Takle

**INTRODUCTION**

* Business Problem Framing

The task is to classify the given set of messages in spam and not spam email.

* Conceptual Background of the Domain Problem

This is a binary classification problem where 0 stands for email being genuine and 1 stands for email being a spam

Visualizing the target will give us an idea of bias of the data

We will test diff algorithms to check for better accuracy and metrics

* Motivation for the Problem Undertaken

Objective behind this project is to help Service Providers determine what amount of consumers will pay back the credit in a given amount of time.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The target variable (dependent) is label which tells us whether the email is spam or not a spam.

Other columns are actual messages and subjects of the email.

* Data Preprocessing Done

Email Addresses, web addresses, monetary figures, phonenumbers in the messages were replaced with corresponding strings.

Stopwords were also cleaned.

* Data Inputs- Logic- Output Relationships

Target cant be said balanced but is a fairly distributed and no regularization seems required here.

Some Null values present but that can be ignored as they are very low.

* Hardware and Software Requirements and Tools Used

Hardware used: system memory 8GB, Processor: 5th gen core i7

Model is developed on Jupyter Notebook

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

As this is a binary classification

for improved accuracy I have employed classification algorithms like Random Forest Classifier ,Decision tree classifier.

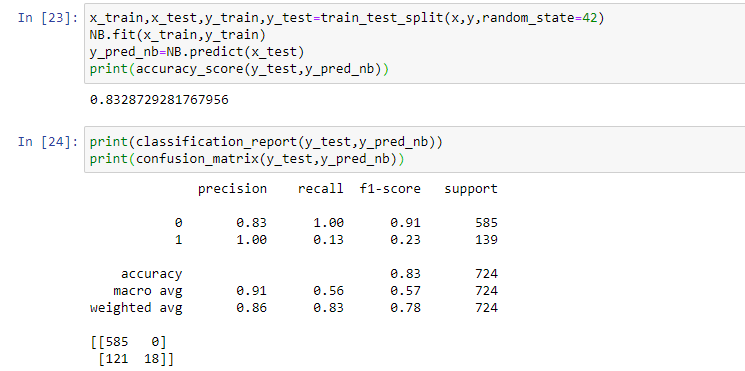
Other classification algos like KNN Classifiers also can be used.

* Testing of Identified Approaches (Algorithms)

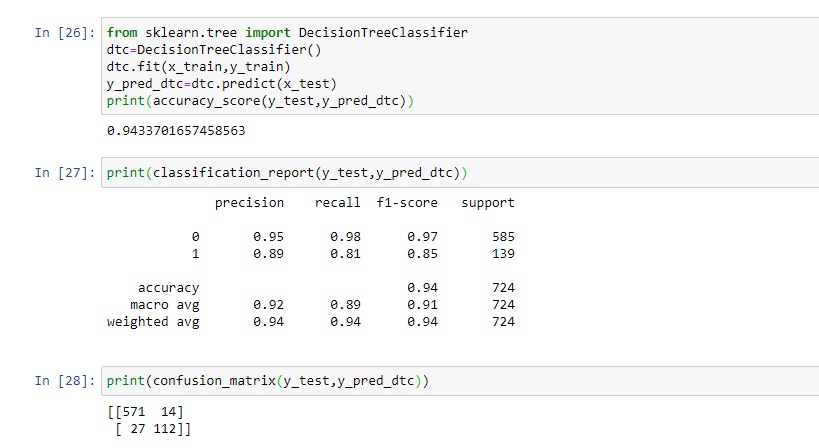
Algorithms used here were

1. NaiveBayes
2. RandomForestClassifier
3. DecisionTreeClassifier

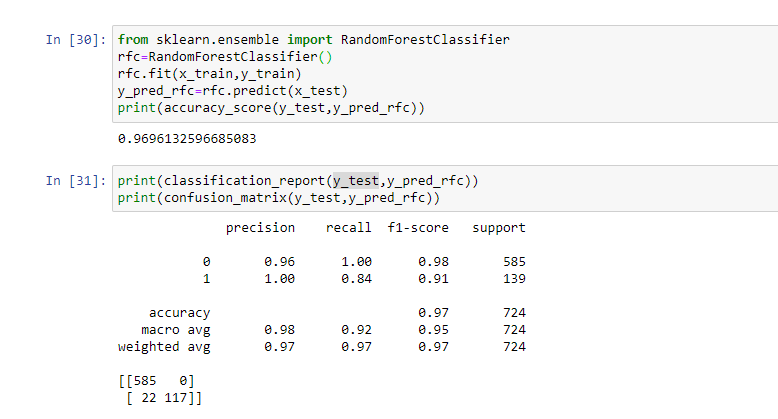
* Run and Evaluate selected models
  + NaiveBayes



* Decision Tree Classifier:



* RandomForestClassifier



* Key Metrics for success in solving problem under consideration

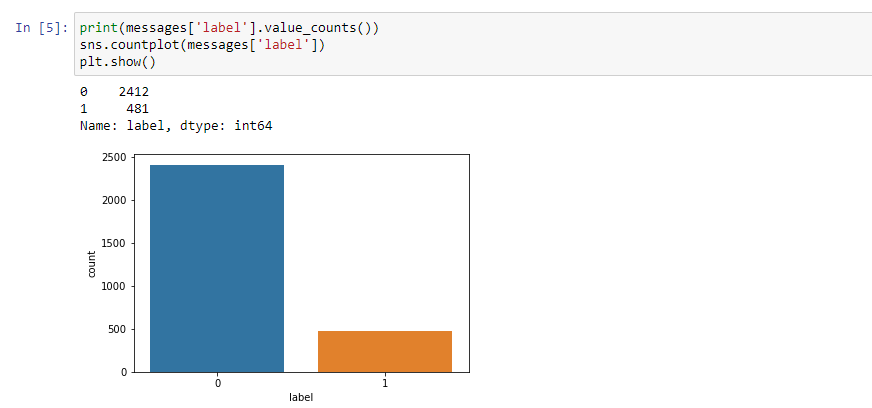
This being a classification problem, the key metrics that are employed here are Accuracy Score, Confusion Matrix.

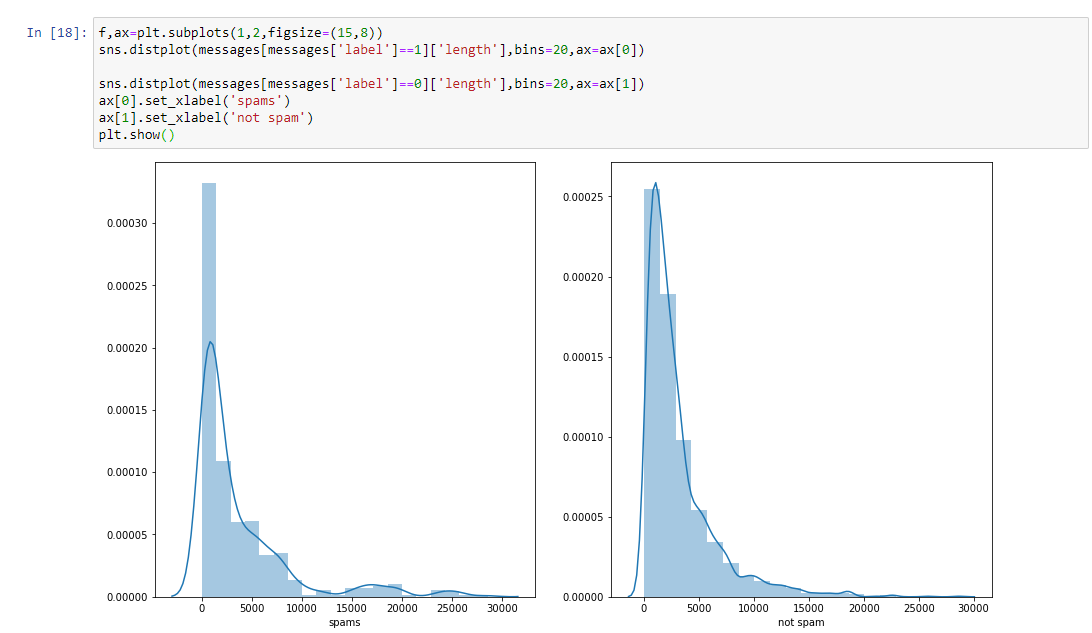
Classification Report is also employed to get metrics like precision, recall, f1-score.

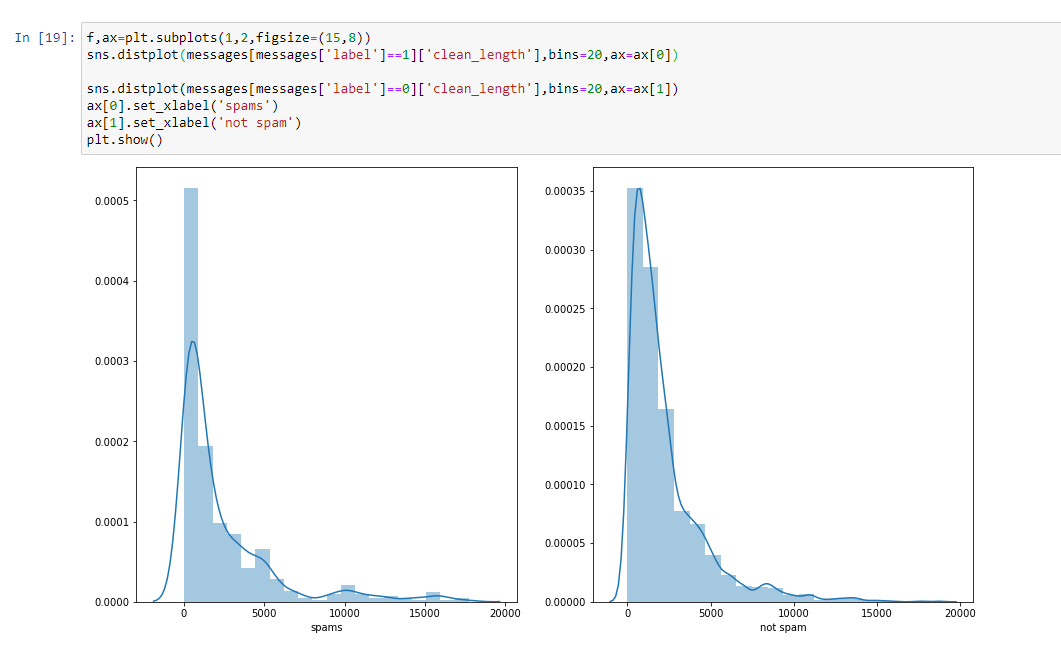
* **Visualizations**

Count plot done to see the count of each label

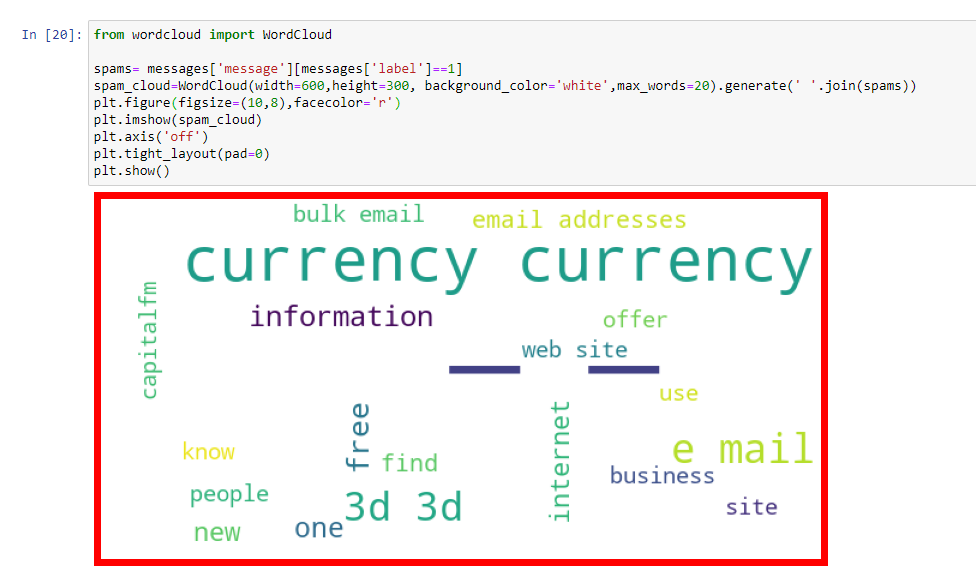
Primary visualizations employed here are distplots to see distribution of lengths of messages before and after cleaning

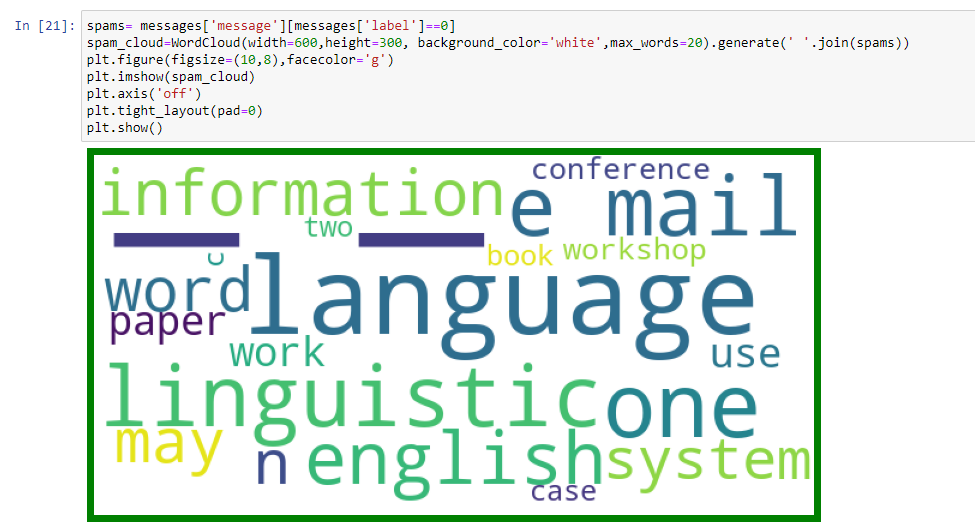






Further word clouds were visualized to get an idea of loud words in the messages:





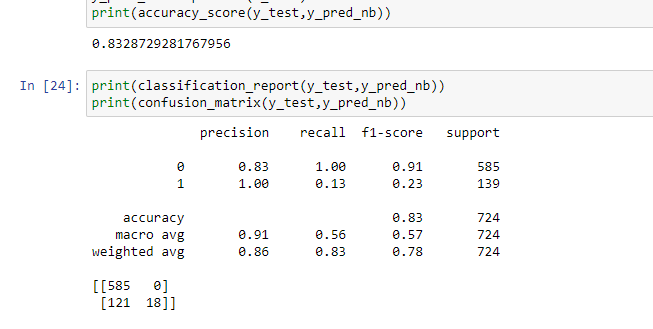
* Interpretation of the Visualizations
  + Count plot done to see the count of each label

In the dataset we have 2412 genuine messages and 481 spam messages.

* + Primary visualizations employed here are distplots to see distribution of lengths of messages before and after cleaning
  + Further word clouds were visualized to get an idea of loud words in the messages

Observing wordclouds we can clearly see some common words in spam mail which represents currency, free, bulk, etc.

* **Results**
  + NB:

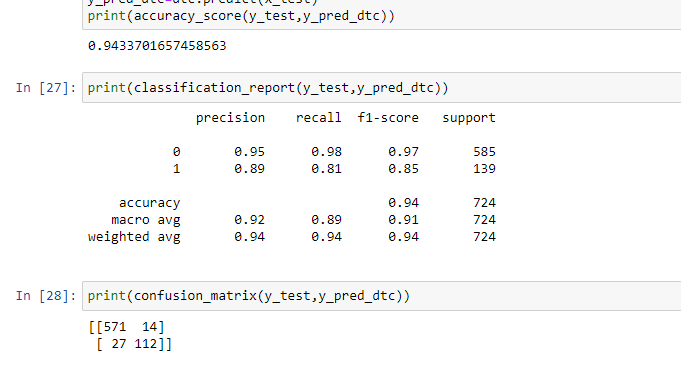


Accuracy : 0.8328

F1-score: 0.91 for 0

0.23 for 1

* DTC:

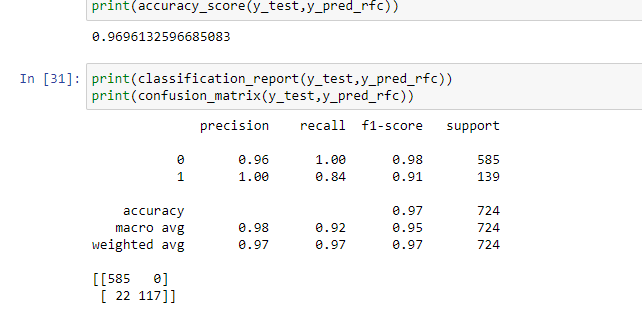


Accuracy : 0.9433

F1-score: 0.97 for 0

0.85 for 1

* RFC:



Accuracy : 0.9696

F1-score: 0.98 for 0

0.91 for 1

**CONCLUSION**

* Key Findings and Conclusions of the Study
* Cleaning the stopwords from the document is very important when dealing with text.
* Loud words help to distinguish the spam better.
* Different classification models must be employed for better accuracy.