

Email spam classifier project

Submitted by:

PALLAVI BAJPAI

**ACKNOWLEDGMENT**

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and references. I would like to extend my sincere thanks to all of them.

I am highly indebted to my institution i.e. Data Trained and the mentors and chat support team for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of Flip and work for their kind co-operation and encouragement which help me in completion of this project.

**INTRODUCTION**

* Business Problem Framing

Understanding the problem is a crucial first step in solving any machine learning problem. In this article, we will explore and understand the process of classifying emails as spam or not spam. This is called Spam Detection, and it is a binary classification problem.

The reason to do this is simple: by detecting unsolicited and unwanted emails, we can prevent spam messages from creeping into the user’s inbox, thereby improving user experience.

* Conceptual Background of the Domain Problem

A good understanding of programming concepts along with some mathematic basic concepts like statistics , probability are very helpful. Thorough understanding of machine learning and the different models is also very important to solve this problem.

* Review of Literature

Considerable about of online research is done in order to understand the problem and requirement of ML in solving this problem.

* Motivation for the Problem Undertaken

The reason to do this is simple: by detecting unsolicited and unwanted emails, we can prevent spam messages from creeping into the user’s inbox, thereby improving user experience.

* Data Sources and their formats

A collection of 5573 rows SMS spam messages was manually extracted from the Grumble text Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.

A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.

* Data Preprocessing Done
* Fair amount of data cleaning process was also involved. This includes checking for space and null values and then dealing with them. Also our available data has 8 columns.
* Duplicate data was removed.
* Outliers were removed but to an extent to maintain 7-8% of data loss.
* Data was imbalanced, so resampling was done.
* State the set of assumptions (if any) related to the problem under consideration

None

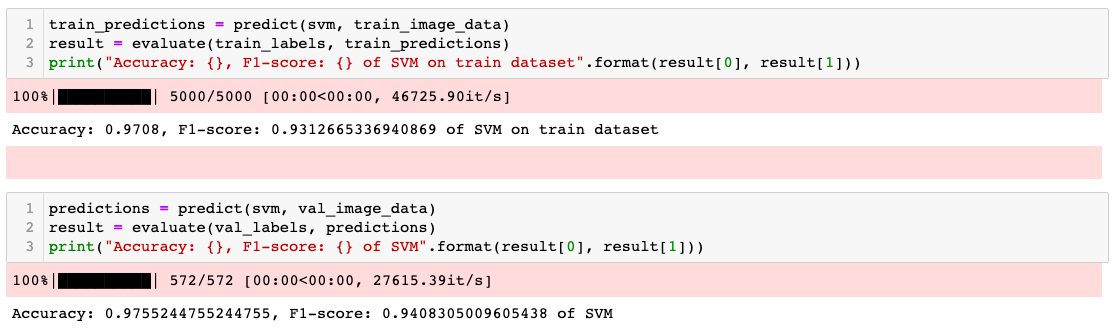
* Hardware and Software Requirements and Tools Used
* Python code was written in Jupiter notebook. Below are the libraries needed in the process.
* Pandas, numpy, matplotlib.pyplot, sklearn.preprocessing, sklearn.model\_selection.cross\_val\_score, sklearn.linear\_model, sklearn.linear.svm, sklearn.ensemble.GradientBoostingRegressor, sklearn.metrics.confusion\_matrix,accuracy\_score, sklearn.metrics.f1\_score, scipy.stats, seabor.

**Model/s Development and Evaluation**

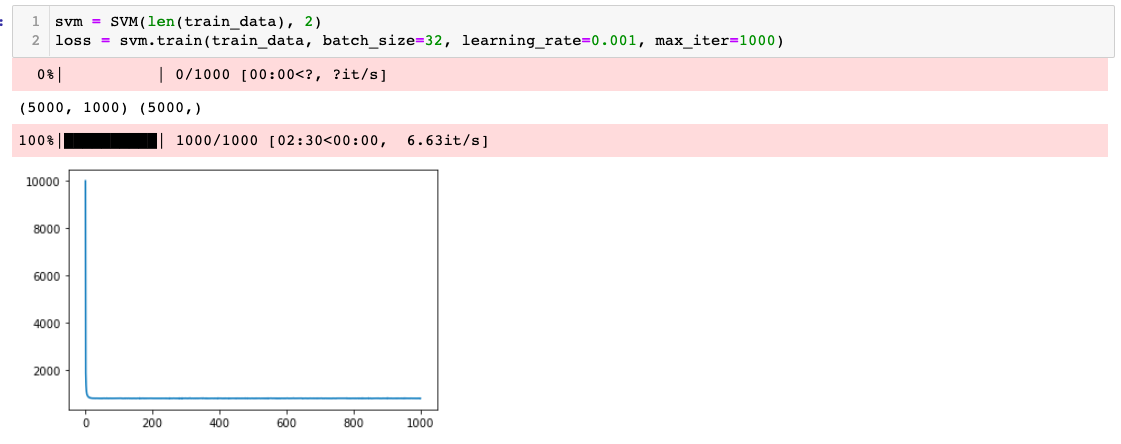
* Identification of possible problem-solving approaches (methods)
* Messages were sliced into a set of keywords.
* Sliced keywords were then analysed to be present into a spam or non-spam classification.
* Testing of Identified Approaches (Algorithms)

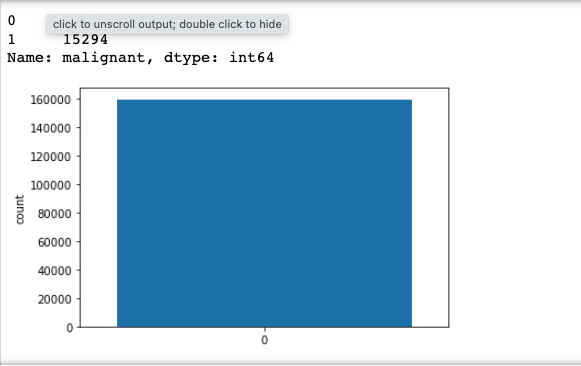
Support vector classifier algorithm was used for training and testing of data

* Run and Evaluate selected models

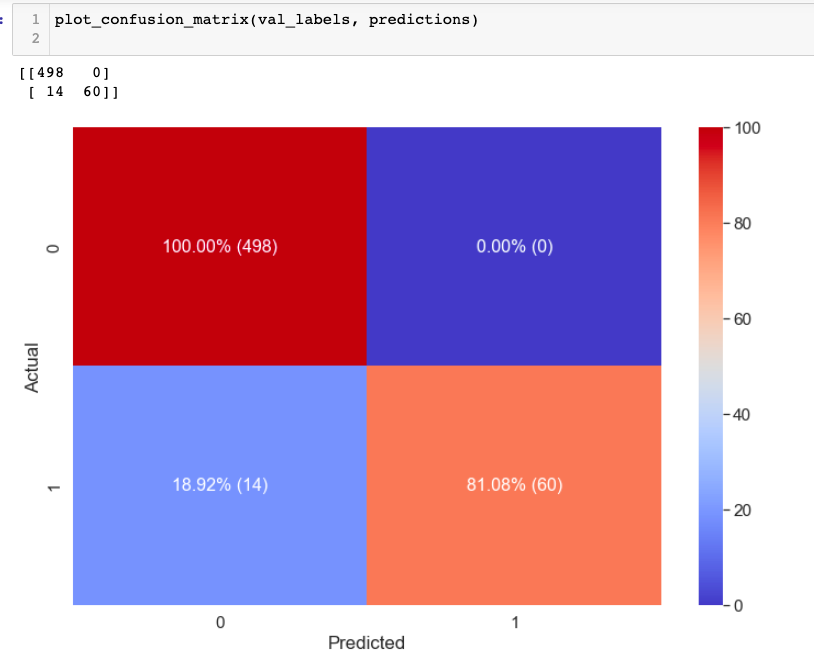


* Key Metrics for success in solving problem under consideration
* F1 score was the key metric used to finalize the model.
* Visualizations





* Interpretation of the Results



**CONCLUSION**

* Key Findings and Conclusions of the Study

Accuracy score was high in support vector classifier hence we trained the model with the same.

* Learning Outcomes of the Study in respect of Data Science

F1 score was high in SVM classifier

* Limitations of this work and Scope for Future Work

None.