FLIP ROBO

EMAIL SPAM CLASSIFIER

Submitted by:

Dattatraya Panda

ACKNOWLEDGMENT

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Mohd Kashifas well as Flip Robo Technologies who gave me the opportunity to do this project

on Surprise Housing Price Prediction, which also helped me in doing lots of research where I came to know about so many new things.

Also, I have utilized a few external resources that helped me to complete the project. I ensured that I learn from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below.

1) *https://www.google.com/*
2) *https://www.youtube.com/*
3) *https://scikit-learn.org/stable/user_guide.html*
4) *https://github.com/*
5) *https://www.kaggle.com/*
6) *https://medium.com/*
7) *https://towardsdatascience.com/*
8) *https://www.analyticsvidhya.com/*

End-To-End machine learning project in Data Science for beginners.

I am going to write about a complete end-to-end project for EMAIL SPAM CLASSIFIER which should serve as a guiding path for many Data Science aspirants.

I have written down all the techniques in the form of sub-topics that I will be explaining one by one. And those sub-topics are as follows:
# 1. Data cleaning

# 2. EDA

# 3. Text Preprocessing

# 4. Model building

# 5. Evaluation

# 6. Improvement

-Context

The SMS Spam Collection is a set of SMS tagged messages that have been collected for SMS Spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam.

Content

What is a Spam Filtering?

Spam Detector is used to detect unwanted, malicious and virus infected texts and helps to separate them from the nonspam texts. It uses a binary type of classification containing the labels such as ‘**ham’** (nonspam) and **spam**. Application of this can be seen in Google Mail (GMAIL) where it segregates the spam emails in order to prevent them from getting into the user’s inbox.

The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

This corpus has been collected from free or free for research sources at the Internet:

-> A collection of 5573 rows SMS spam messages was manually extracted from the Grumbletext 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.

Hardware & Software Requirements & Tools Used:

Hardware used:

- Processor: Core i5 -10300H CPU @ 2.50GHz
- RAM: 8 GB
- Operating System: 64-bit
- ROM/SSD: 512 TB SSD
- Graphics: NVIDIA GeForce GTX 1650 Ti

Software requirement:

- Anaconda Navigator - Jupyter Notebook

Libraries Used:

- Numpy
- Pandas
- Matplotlib

Seaborn