Project:

Spam Filter using ML Models

Introduction and Data Acquisition

Problem Statement. The SMS box can be spammed by messages from marketing campaigns to fraudsters. This makes the account holder susceptible to being conned or miss an important message.

Objective. The goal of this project is to build a spam filter that can effectively categorise a message as either 'SPAM' or 'NON SPAM'.

Dataset Source. We will be using the SMS Spam Collection Dataset (UCI).
https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

Data Preparation

Data Import. Import dataset from UCI repository.

Data Cleaning & Preprocessing. include the following steps
- Case conversion (lower case)
- Remove digits, punctuations and stop words

Document Term Matrix. is a mathematical matrix representation of document
- Bag of Words (BOW): frequency for each word in document.
- Term Frequency - Inverse Document Frequency (TF-IDF) : wt for each word in the document.

Descriptive Statistics

Dataset Description.

- ○ Total no. of observation: 5572
- ○ No. of spam messages: 747
- ○ No. of non-spam messages: 4825

Additional Statistics.

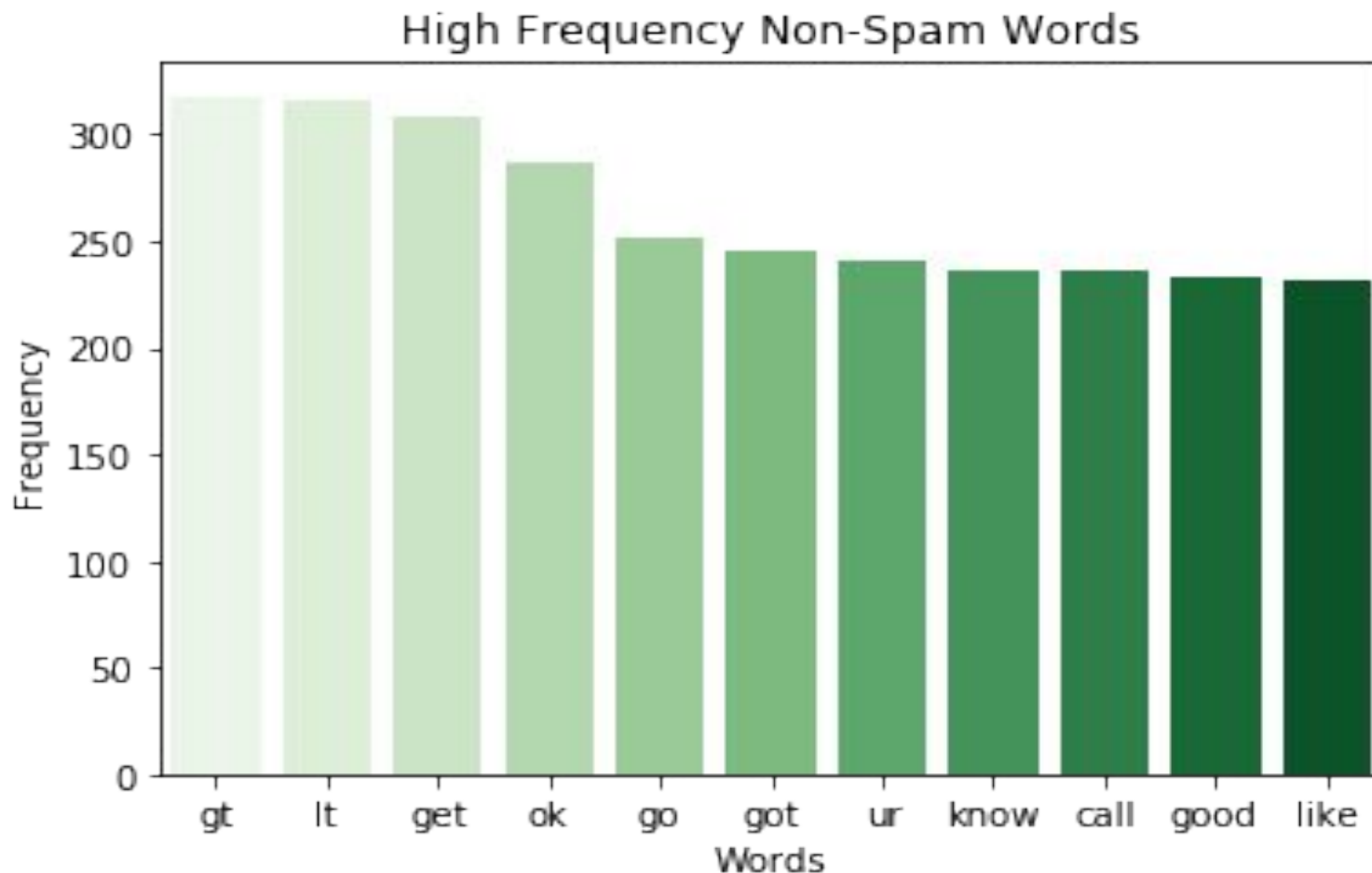| Label	Message Length (median)	Contains Digit % (No of docs containing digit / total number of docs)
Non-Spam	52	15.5%
Spam	150	94.78%

Exploratory Data Analysis

Most Frequent Words.

High Frequency Spam Words

Exploratory Data Analysis

Most Frequent Words.

Exploratory Data Analysis

Word Cloud.

Spam Message

Non Spam Message
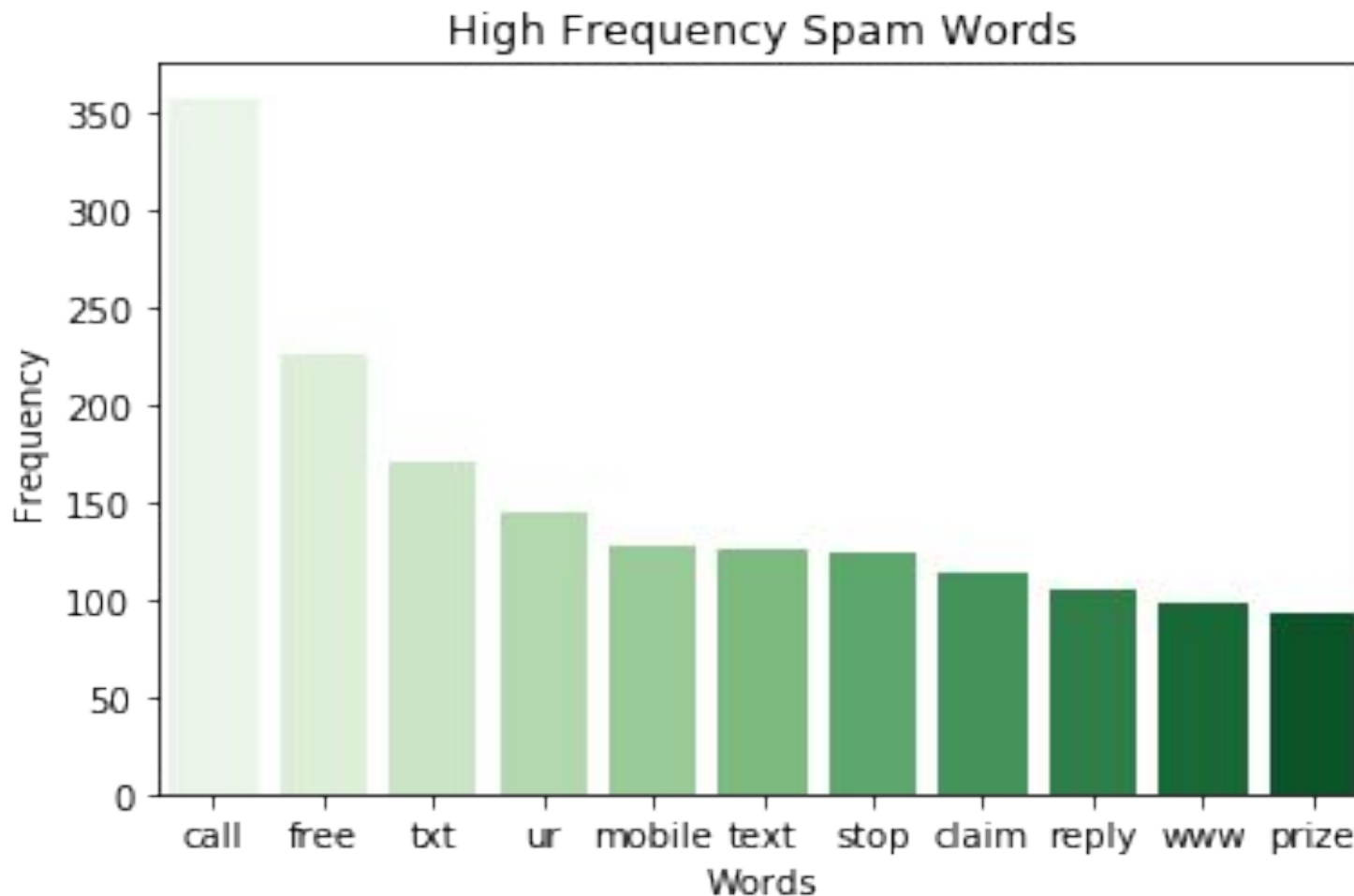
Exploratory Data Analysis

Message Length.

Spam Message   **Non Spam Message**
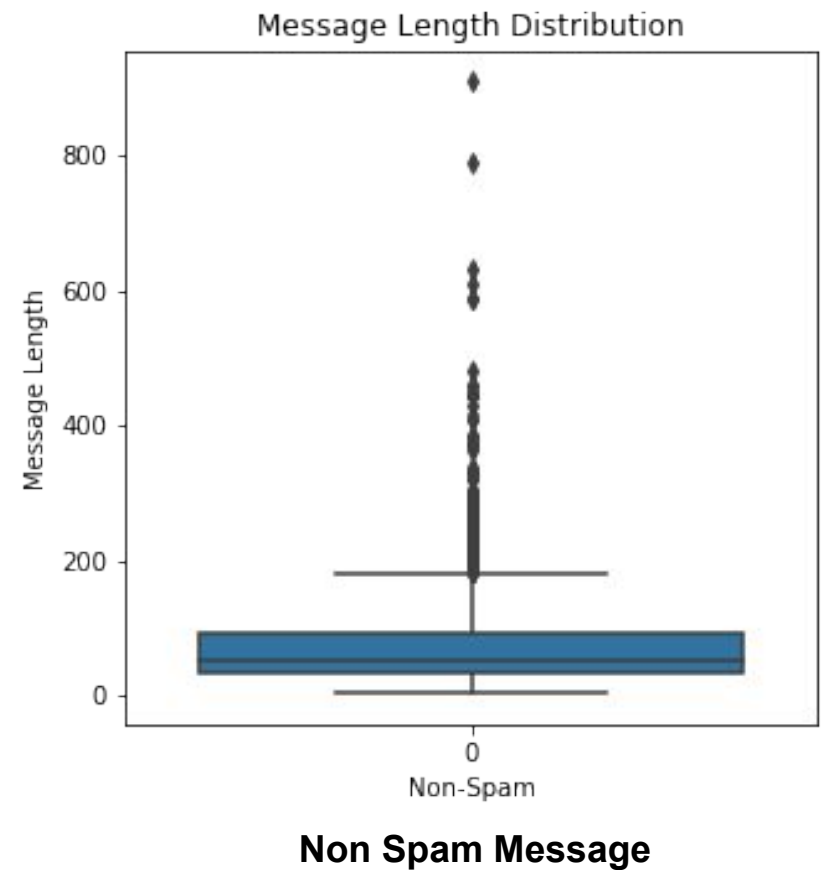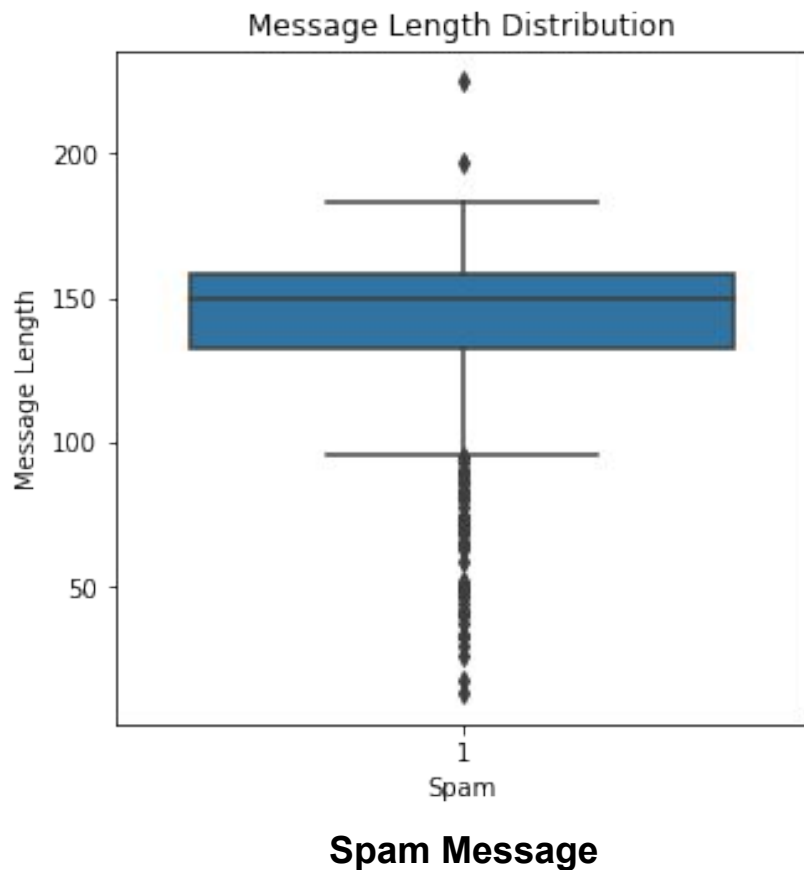
Model Building and Selection

Approach.

- ○ Bag of Words (BOW)
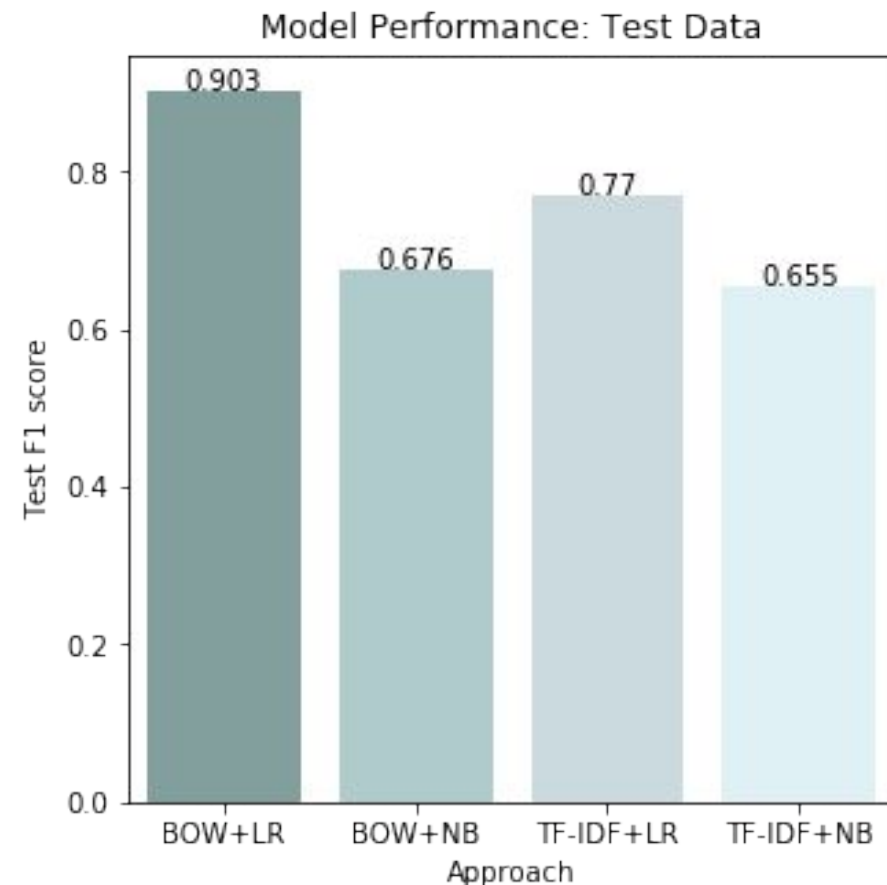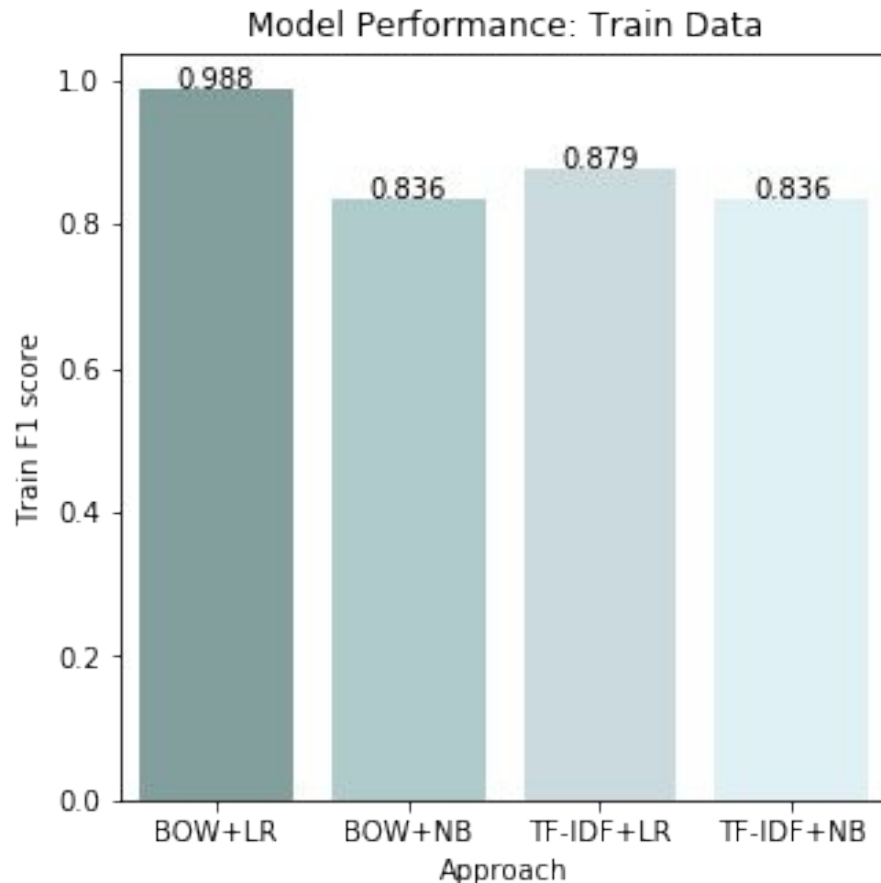- ○ Term Frequency - Inverse Document Frequency (TF-IDF)

ML Models.

- ○ Logistic Regression (LR)
- ○ Naive Bayes Classifier (NB)

Performance Metrics. F1 score for train-test 70:30

Model Building and Selection

Model Performance.

Model Performance: Train Data — Train F1 score: BOW+LR 0.988, BOW+NB 0.836, TF-IDF+LR 0.879, TF-IDF+NB 0.836

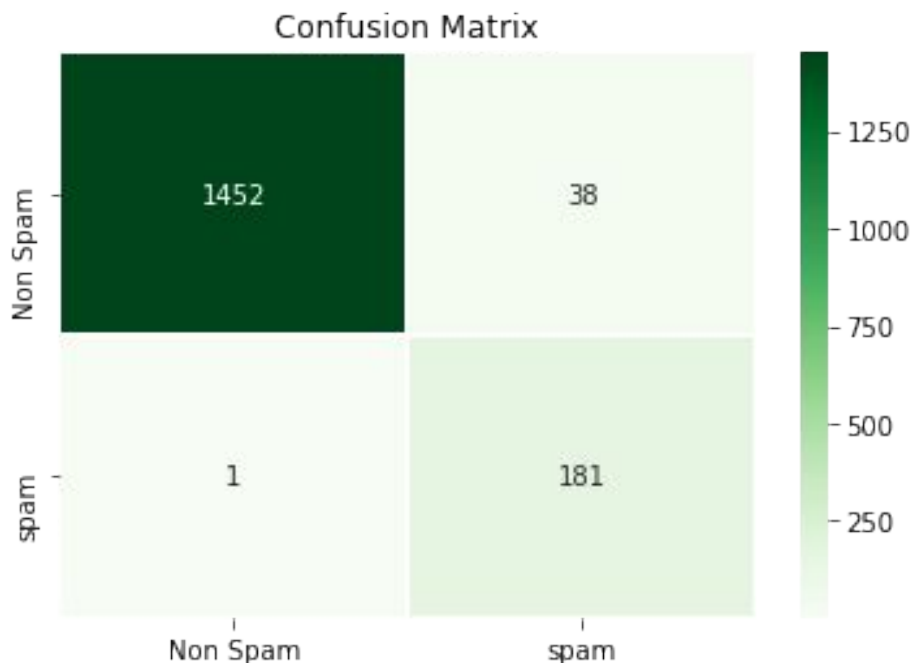Model Performance: Test Data — Test F1 score: BOW+LR 0.903, BOW+NB 0.676, TF-IDF+LR 0.77, TF-IDF+NB 0.655

Logistic Regression with Bag of Words approach gives a higher F1 score for both train and test data.

Model Interpretation and Results

Best Performing Model. Logistic Regression with Bag of Words approach

Confusion Matrix.

True Positive. 181
False Negative. 38

False Positive. 1
True Negative. 1452

Precision. 0.99 Recall. 0.82
F1 Score. 0.90

Model Interpretation and Results

Model Coefficients. for words in BOW vocabulary

Word	Model Coefficient
claim	1.8
ringtone	1.7
urgent	1.45
prize	1.38
win	1.37

Word	Model Coefficient
cash	1.36
free	1.33
chat	1.28
freephone	1.11
adult	1.00

Conclusion

1. Logistic Regression model with BOW approach performs well in the task for classifying message in spam and non-spam categories.

2. Messages containing words such as claim, free, prize, win, cash has higher probability of being spam message.