C111 Machine Learning - Project Proposal

Emma Allen

November 2023

1 Project Overview

My project consists of classifying a data set of emails as spam and non spam. This will be achieved by utilising five machine learning models. These models will be compared to conclude which model is best for this particular data set.

The models that will be used are (using sklearn module):

1. Logistic Regression
2. Decision Trees
3. K Nearest Neighbours
4. Support Vector Machine
5. Artificial Neural Network

In addition, for the artificial neural network, i will implement it in two ways, firstly with sklearn and secondly with PyTorch

2 Data

Data Link: https://archive.ics.uci.edu/dataset/94/spambase

This data set contains 4601 instances and 57 features.

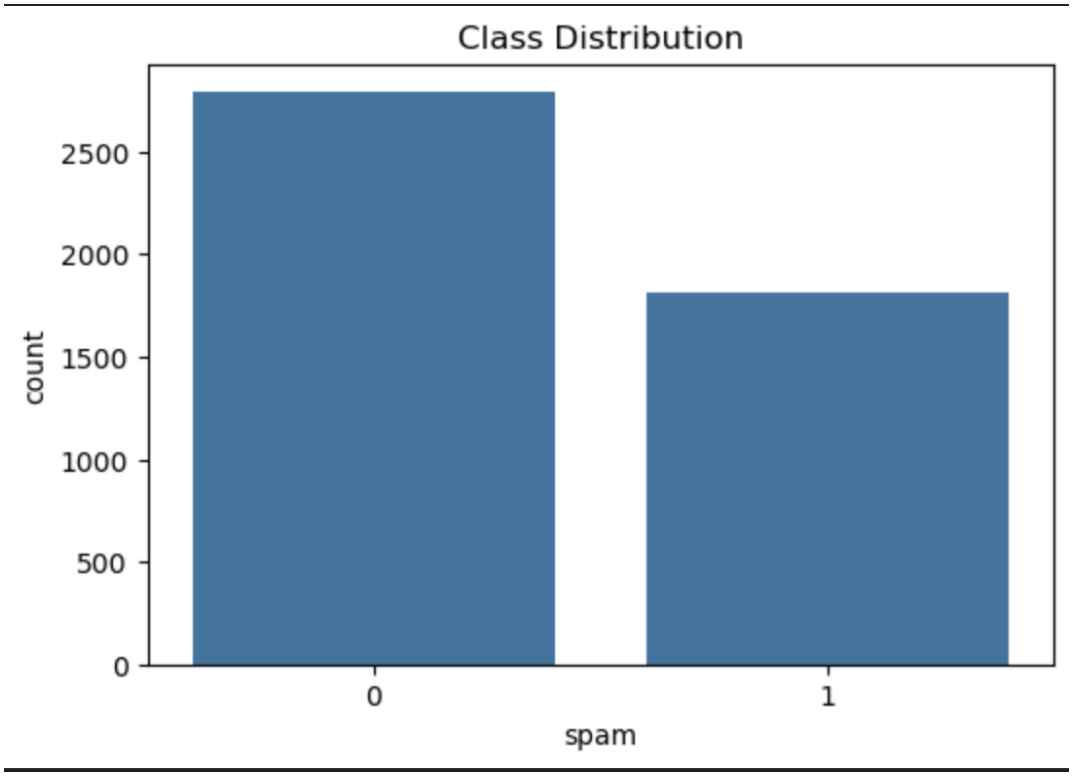
Each instance represents an email. The spam emails comes from the author's postmaster and individuals who have filed spam. the non-spam emails come from a collection of work and personal emails. The word 'George' and the area code '650' are indicators of non-spam.

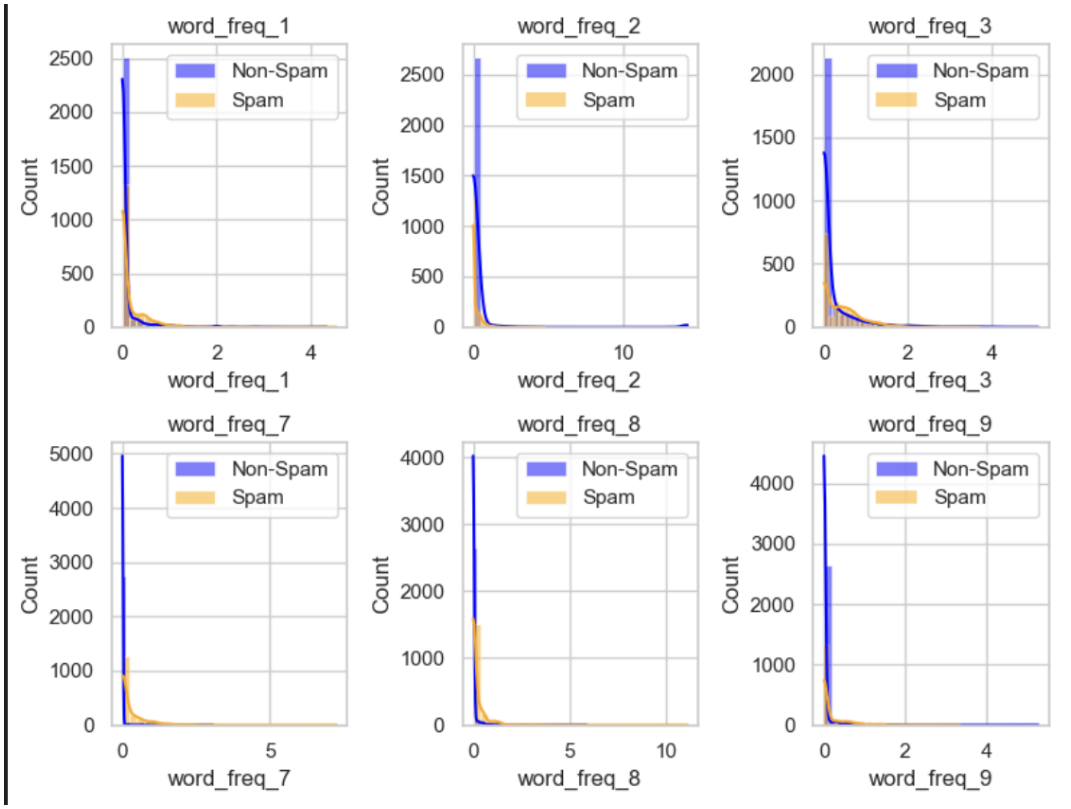
A typical performance for this data set is around 7% mis-classification error

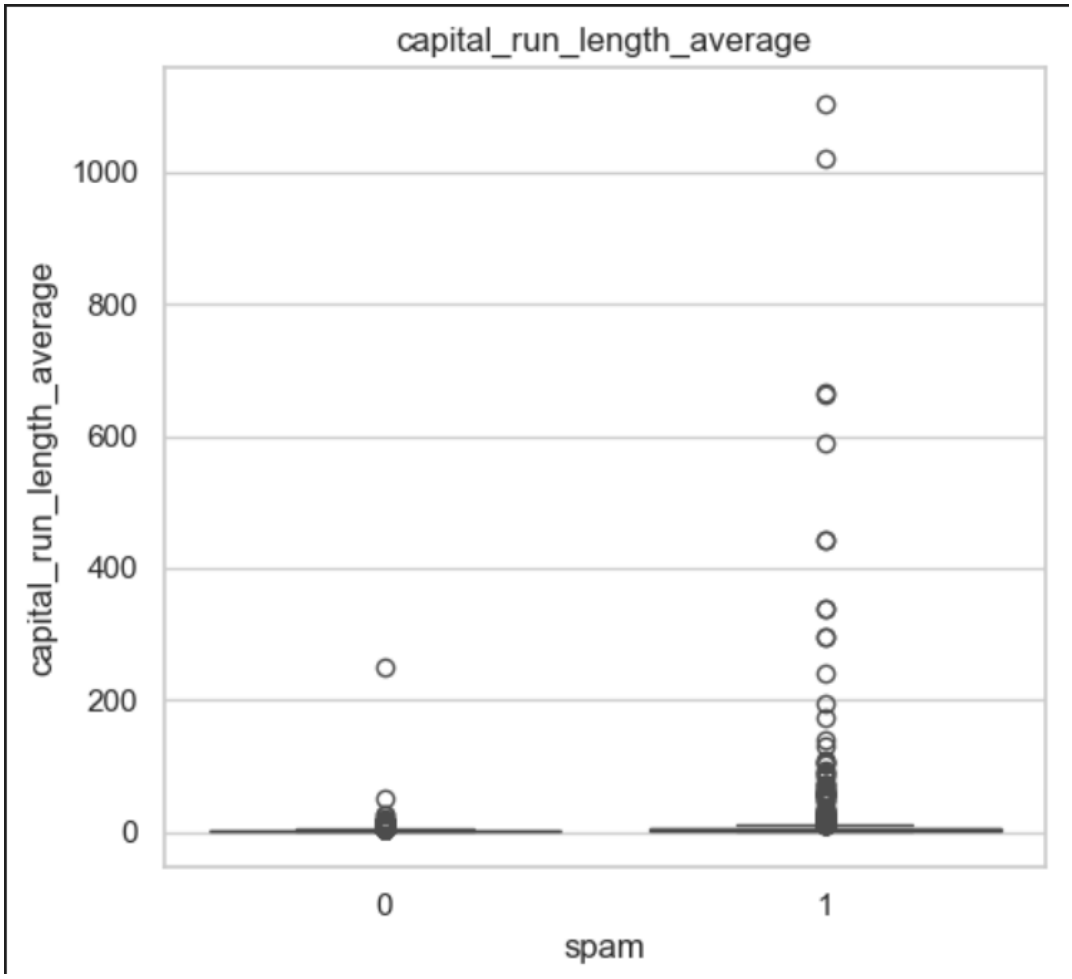
3 Data Visualisation

To commence this project, I have visualised the data to show:

1. The distribution of the spam class
2. The distribution of word frequency attributes for spam and non-spam
3. The distribution of capital run length attributes for spam and non-spam


Class Distribution
count
2500
2000
1500
1000
500
0
0
1
spam


word_freq_1
word_freq_2
word_freq_3
word_freq_7
word_freq_8
word_freq_9
Non-Spam
Spam
Count


capital_run_length_average
capital_run_length_average
1000
800
600
400
200
0
0
1
spam