Ashesi Machine Learning Midsem

Aaron Amarh Ashitey

Department of Computer Science and Information Systems, Ashesi University

[25-26_SEM1_CS 452_A] - Machine Learning

Dr. Isaac Nyantakyi

November 04, 2025

Ashesi Machine Learning Midsem

Introduction

In this brief report, I present my process and findings for my Midsemester Machine Learning exam. For the project, I went through an elaborate process of exploring the data, processing the data, finding a perfect model for the data, and making predictions. In subsequent paragraphs, I will delve into the details of the several aspects of my journey that made achieving a final RMSE of 0.5854 possible.

Data Analysis and Insights

This section of the work started with some extraction of some general information from the data. This information included the data types of the entries, some value counts of the columns, and some general statistics. I needed to extract this information as part of my overall goal of making sure the data was good enough to start my model training and avoid any unforeseen results. I then identified the target column, looked out for missing values and duplicates so I could take them out. Fortunately, there were none of such in the dataset. The dataset, I must say, was quite pristine.

After confirming that the dataset was pristine and hence did not need me to do much cleaning, I had to explore it more to make it inform my model choice and feature choices. I plotted the distribution of the features to see if they were gaussian so that I could effectively use a linear model. However, they were mostly skewed. From this point, I started to realise that I would need a model that did not need the model to be gaussian. I, however, still standardized the data so I could try linear models.

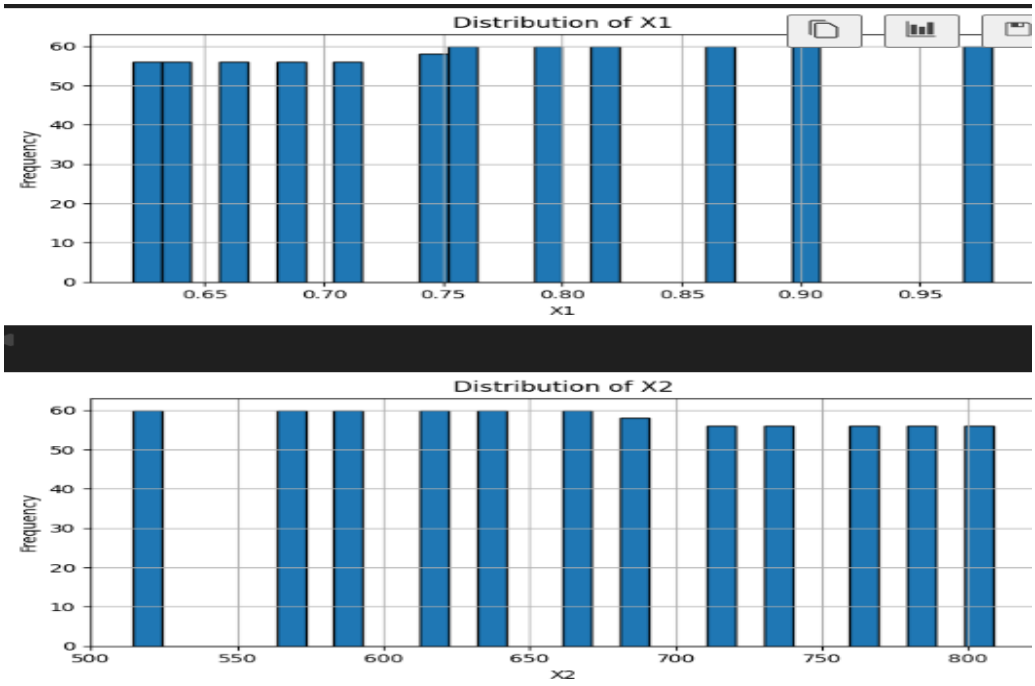

Figure 1: some examples of the distributions of the dataset features

After gaining this insight that informed my model choice, I needed to see if all features were necessary for my model so I did the correlation heatmap to help eliminate the irrelevant features. I realized that X4 and X2 had the least correlation with the target variable. At the same time, X2 had a high correlation with X4 while X1 had a high correlation with X5. However, since X1 and X5 had a high correlation with the target variable, I maintained them and took out X2 and X4. Taking them out did not change the RMSE significantly. Nonetheless, it made my model train faster, so I was happy I took them out eventually.

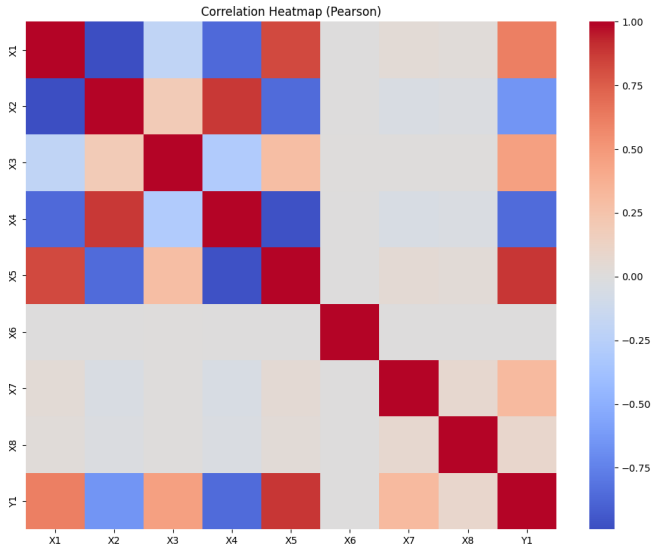

Figure 2: correlation heatmap

Model Selection and Justification

Due to the skewedness of the dataset and the multicollinearity of some features, linear models were not going to work, so I settled for tree-based ensemble models. However, just to show this, I started off with some linear models, some tree-based models and some neural networks to see which gives the lowest RMSE. The top 5 were used to make an ensemble. I borrowed the concept of complementary and Kalman filters here. So what this seeks to achieve is to assign higher weights to models that are doing better and let the others complement with lower weights.

I then used an optimizer called ‘scipy.optimize.minimize’ to find the perfect blend of their predictions. This process iteratively nudged the weights of each model to find the combination yielding the lowest possible validation RMSE. This final model is justified because the optimized weights allow individual model errors to cancel each other out, resulting in a single superior model that is more robust and accurate than any component alone.

After this superior model, the RMSE from my validation set was 0.395199. This proved that having the iterations and constantly adjusting weights was a very good approach as compared to using the individual models. However, on the website I had 0.5854 meaning there was still some slight overfitting. The RMSE did not move down despite several changes to the parameters.

```
Ensemble weights:
  GradientBoosting    : 0.8164
  ExtraTrees          : 0.0756
  RandomForest        : 0.0000
  MLP                 : 0.1079
  Ridge               : 0.0000
```

Figure 3: the weights of the different models in the ensemble

From the results attained, it is obvious that the dataset warranted non-linear models as the only linear model that made it to the top 5 ended up with a weight of 0.