Machine Learning – Project Assignment

Alvin Tran

1. INTRODUCTION

Machine learning is about taking data from everyday life and using algorithms to learn more about the data that is not so obvious at the surface. It is important to note that some algorithms will perform better on the dataset than others. To add on to that, performance can be improved by preprocessing the data beforehand and picking parameters that makes sense for the algorithms with regard to the dataset.

For this project, two datasets were provided in which we were to apply three machine learning algorithms per dataset. With each algorithm, we were to tune and optimize the algorithms we used so that the algorithms can perform the best it can on the dataset with high and meaningful metric scores.

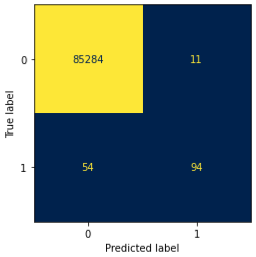
One of the datasets is the Credit Card Fraud Detection dataset. The dataset requires machine learning algorithms for classification problems. Classification problems are problems where a certain value of the dataset can be predicted and categorized into two options, 0 or 1, or in this case with the credit card dataset, not fraud or fraud. I have not finished using all three of my chosen algorithms for this dataset, so the results remain inconclusive. The algorithms I have chosen and completed are logistic regression with imbalanced learn and without, and support vector machines without imbalanced learning. Based off the best model for logistic regression, the best accuracy score, precision score, recall score, and f1 score I have obtained is 0.9768, 0.0627, 0.8885, and 0.1171 respectively. Based off of the support vector machines, the scores are 0.9992, 0.9659, 0.5743, and 0.7203 respectively.

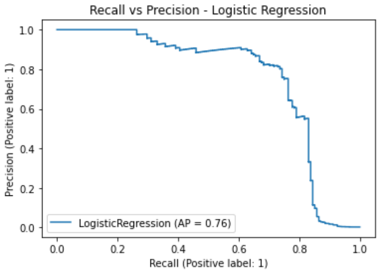
The other dataset is the Energy Efficiency dataset. Unlike the credit card dataset, this dataset requires machine learning algorithms for regression problems. Regression problems are problems where the values of the dataset can be predicted as any value using the existing data and, in this case, with the energy efficiency dataset, values of heating load and cooling load can be predicted based off of data in the dataset. I have not finished using all three of my chosen algorithms for this dataset, so the results remain inconclusive. The algorithms I have chosen and the only one I completed is linear regression. Just based on linear regression, the mean squared error and r2 score for heating load predictions is 7.4127 and 0.9289 respectively while the mean squared error and r2 score for cooling load predictions is 11.0915 and 0.8787 respectively.

2. Credit Card Fraud Detection Dataset

The Credit Card Fraud Detection dataset contains data such as the time in seconds between each transaction and the first one, numerical values from PCA transformations, the amount of each transaction, and if that transaction was fraudulent or not. This dataset is considered extremely unbalanced because the fraudulent category only consists of 492 instances out of 284,807 instances having a percentage of 0.172%. In the case of unbalanced datasets, we have to use imbalanced learning methods so that we can tell if the metric data from the machine learning algorithms on the dataset is truly accurate.

One of the machine learning algorithms I have chosen is logistic regression. Without any imbalanced learning methods, the algorithm returns an accuracy of 0.9992. Although it may sound good, we will need to look at the precision score, recall score, and f1 score to determine if the algorithm’s performance was good because the dataset is imbalanced. The meaning of each of the scores is that the precision score tells how good the algorithm is at predicting true positives out of all the instances that were labeled positive while the recall score tells how good the algorithm is at predicting true positives out of the instances that were labeled true positive and the instances that were labeled false negative, and the f1 score is the weighted average of both the precision and recall score. The precision score came out to be 0.8952 which is pretty decent while the recall score is only 0.6351 which is a bad result. The f1 score came out to be 0.7430.





In the case of this dataset, we would want a higher recall score because we do not want false negatives because we are working credit card fraud and we do not want the case where a credit card is fraud but label it not fraud.