Intermediate Project Report

This project involves analyzing credit data and attempting to identify instances of credit fraud within that data. The main purpose of this project is to review methods used by other people in identifying fraud and repurposing them to use with a smaller number of features. The reason for doing this is because, for privacy reasons, it is difficult to get large amounts of data with known features, the goal of the project is to try to find the minimum number of features needed to get a realistic chance of identifying fraud.

The data used in this project is from Kaggle.com and can be found at https://www.kaggle.com/mlg-ulb/creditcardfraud. This data includes 31 different variables to use, with only Time and Amount being identified. The rest of the features given by that data are labeled V1 through V28 and have no description or name to be identified with. Specifically, that dataset contains transactions on September 2013 in Europe, and has 492 identified frauds out of the 284,804 posted transactions, with the identified frauds being labeled by the 'Class' features.

So far, for the project, I have read the papers used for reference, and identified the models that will be compared in the actual project, with those models being an unsupervised Generative Adversarial Network outlined in https://arxiv.org/pdf/1904.10604.pdf and a supervised AdaBoost model. The GAN model highlighted in the link is different from the traditional model since it also defines a "score function A(x) that measures how anomalous an example x is, based on a convex combination of a reconstruction loss L_G and a discriminator-based loss L_D". With L_G in this case meaning how well the encoder and generator reconstruct the input, and L_D meaning the confidence of the discriminator. In the case of the AdaBoost model, the model is similar to a regressive model, except that it is used for classification, and it adds a weight to each instance in the initial training dataset. For this, the initial weight is calculated as x_i = 1/n where x_i is the ith instance and n is the number of training instances. The reason these models were chosen is because, in each of their respective papers, these models were found to be the best at detecting credit fraud in their respective datasets. However, neither of these models were compared against each other in either of the papers mentioned, so the initial comparison with the given number of features before reducing them will also be recorded in the final report.

What remains to be done for the rest of the project is to see if any remaining models are likely prospects to be reanalyzed with a smaller number of features, and to implement and record the data from these models while reducing the number of features after training each one.