**Milestone Three**

DAT 690

Marek Krawczyk

9/27/2022

**Modeling**:

Creation of the first model is rudimentary to gauging the performance of the models against each other. In the original assessment the usage of Random Forest model was implemented for factor analysis of the variables. Although rudimentary, the Random Forest classification gives us a good insight into correlation of our variables, without performing any VIF testing, or breaking down the data into PCA.

To gauge the performance of Logistic Regression on our data, a simple linear model was used.

FIGURE 1: Creation of a Model

Graphical user interface, text, application, email

Description automatically generated

After establishing our methodology, it is important to split the data into Train, and Test respectively. It is important to set a random state so that our results in splitting will always be the same. Fitting the model comes next, with fitment of X\_train, and y\_train.

**Evaluation:**

Taking apart the first iteration of our model is necessary to see how it can be improved. The idea is to get the model in a general form, and improve the methodology and format of the variables, in order to get better results in the final product.

The confusion matrix allows us to dig into the statistic of correct classification. With this tool we are able to see what our model struggles with the most. In this case we are able to predict non-defaults very well as 159 instances were correct. However the model is pretty inaccurate when it comes to predicting Negatives as it failed to predict correct label 35 times, shown as False Negatives – 35. It is nearly at 50% predicting Negatives, which should be improved in the final model.

FIGURE 2: Confusion MatrixChart, treemap chart

Description automatically generated

The model itself has a Macro Accuracy of 78.8%. Which is a pretty good starting point for our analysis. Eliminating the non-correlative variables didn’t hinder our model, as we had similar success rate with our Random Forest preliminary model.

FIGURE 3: Macro Accuracy

Graphical user interface, text

Description automatically generated with medium confidence

Another metric that can help us to understand what exactly is going on is Receiver Operating Characteristic curve. It allows to take a look at the difference between sensitivity and specificity. The AUC score of 80%, which strikes a median between perfect classifier of 1, and worthless classifier at .5.

FIGURE 3: ROC curve

Chart, line chart

Description automatically generated

**Moving On:**

It seems that our classification is not yet good enough to call off the development of the project. What is most probably a hindrance to our Logistic Regression model is the data set format. We have not create any dummy variables to mirror their categorical meanings.   
In the future iterations, One Hot Encoding or Dummy variables will take place of the categorical variables, so the model has a better time understanding the features listed.