BRYAN CARAWAY

bryan_caraway@baylor.edu
github.com/Bryancaraway
linkedin.com/in/bdcaraway
210.204.5115

June 19, 2022

Executive Summary
State Farm
Data Scientist
Assignment

For this task, two classification models were developed: a generalized linear model (GLM) and a non GLM. The GLM is a logistic regressor, and the non GLM is a feed-forward deep neural network (DNN). The logistic regression model is easy to deploy, requires no hyper-parameter tuning, and is quick to train, however it requires extensive feature engineering for it to perform well, i.e. predictors should be uncorrelated, cannot have missing values, and have decent predictive power by themselves. By contrast, A DNN is more complex to develop, may require time-consuming hyperparameter tuning, and takes more time to train. An advantage of DNNs is their ability to automatically engineer features in the hidden layers from linear and non-linear combinations of the input predictors, thus reducing the need for manually constructed features. Additionally, the complexity of DNNs allows flexibility for the user to tailor the model to the task. e.g. label imbalance, regularization, and feature selection.

The area under the curve (AUC) is estimated for the test data set by taking the average of the KFold validation data set AUCs, where K $= 5$. The estimated test AUCs both demonstrate acceptable discrimination, and are 0.7623 and 0.7723 for the GLM and non GlM, respectively. By this metric, the DNN should perform better on the test set, albeit marginally (about 1%).

Test AUC Estimates	
GLM	DNN
0.7623	0.7723

In non-technical terms, the DNN will, on average, correctly distinguish about one additional observation out of one hundred when compared to the GLM. Dependent on the task, the improvement can be quantified in terms of cost savings, profit, customers retained, customers acquisition, etc.