State Farm Data Scientist Opening Pre-Employment Assessment

Junzhou Lin

During this take home assessment, I used two modeling approaches. One is the Random forest, and another one is the Logistic regression.

The advantage of the Random forest is the model generally less affected by outliers than the Logistic regression model and it allow a reduction in overfitting which is caused by having too many features. However, it sacrificed speed for accuracy. The model takes much longer time to run than the Logistic regression, and also it's hard to explain to a non-tech person.

Logistic regression is a high bias low variance model. The advantage is that it's stable, simple to implement and easy to explain to a non-tech person by using coefficient or plot. However, the disadvantage is it might not be as accurate as other complex models.

Even though both Logistic regression model and Random forest model have very similar accuracy in this case, but consider the dataset given is very imbalanced, therefore, looking at the accuracy score might not be the best choice. By looking at the AUC, we can see that Random forest model (0.7350) has a higher AUC area than Logistic regression model (0.6953). Plus, I didn't deal with outliers nor feature importance because I am uncertain about the definition of each columns. Therefore, I prefer the Random forest model over Logistic regression model, and I believe the Random forest model will perform better on the test set because it's more tolerated towards outliers and feature overfitting.

My guess for AUC will be the same as the average AUC I got from the 5 fold cross-validation AUC score. Random forest will have a 0.7350 AUC, and a Logistic Regression will have a 0.6953 AUC.

It's hard to explain what's going on in the model to a non-tech person just by telling numbers or scores. I always believe plot is the best to illustrate ideas. Therefore, we could plot the ROC, AUC, confusion matrix. If we don't have the access to any of the classification matric, I'll plot the partial dependence plots on those importance features to access how any one features affects the model's decision, then take those features and start plotting them individually against the prediction results, while keeping all other features constant. By comparing the before and after plots for each model, it can give the business partner a general sense of which model is performing better at distinguish and predict class pattern.