Machine Learning Report

The next phase of this project is to apply machine learning techniques to build predictive models on the Caravan Insurance data set. The following are an overview of the steps I took to analyze my data.

First, I narrow down the variables I am to model on by the P-value I did for the chi-square test. I narrow down from 85 variables to 28 variables, which saves time on building and to tune the hyper-parameters.

My starting point is for this section of this project is continuing from the result obtained from Statistical Inference. From our chi-square test, we may narrow down what variables I think could be more relevant. The selection criteria include the variable if its chi-square p-value is less than 5%, which means this variable statistically can have a meaningful difference from the average.

The data set has already split into training and test data set in the previous section of this project. I will build models from the training data set and test the models' performances using the testing data set.

Second, I list down possible models for my data. Since my target variable is categorical, I felt CART, Random Forest, GBM, and Logistic Regression seem to be the best possible fits. The problem is a classification problem. Initially, I took the models as is and apply them without tweaking the hyperparameters.

The third step is to look at how the model performs.

To evaluate how my model is performing, I turned to use the confusion matrix initially. The three metrics I am most interested in the confusion matrix are accuracy, precision, and recall. Accuracy shows how much of the sample classified correctly, percentage-wise. This metric shows the model doing pretty well in both training and test data, but different look into precision or recall indicates that the model's failure at predicting true positives.

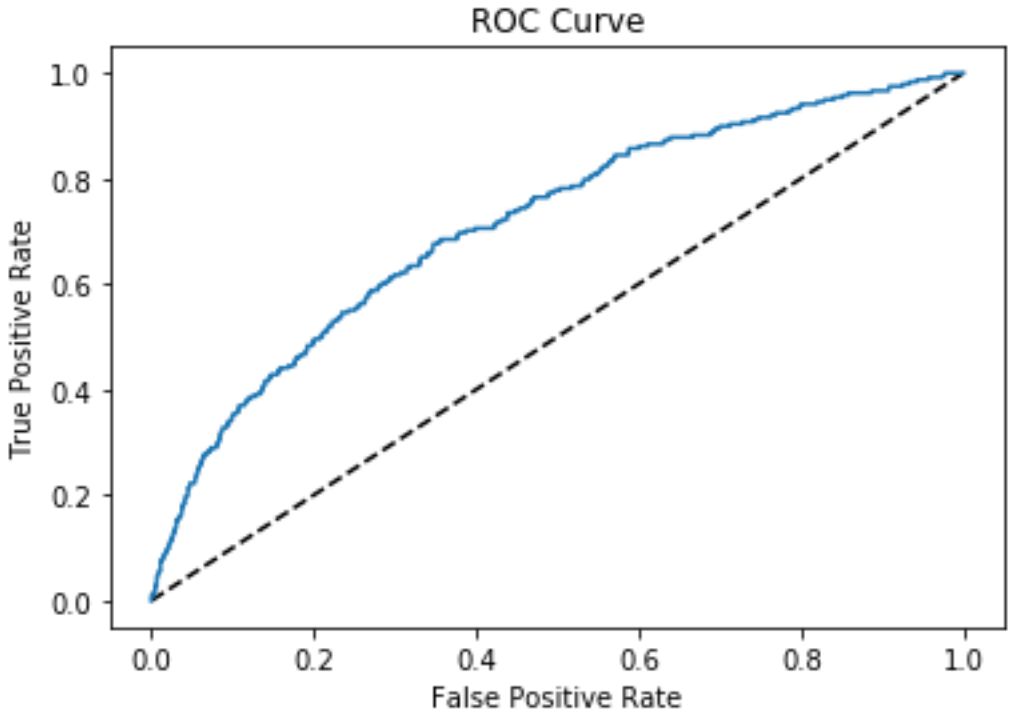
The following is an example of that from my logistic regression model:

```
      0  1
0  5470  4
1   346  2

 accuracy:      0.94
 precision:     0.333
 recall:        0.006

      0  1
0  3759  3
1   235  3
```

```
accuracy:       0.94
precision:      0.5
recall:         0.013

CV Score: 0.9395
```

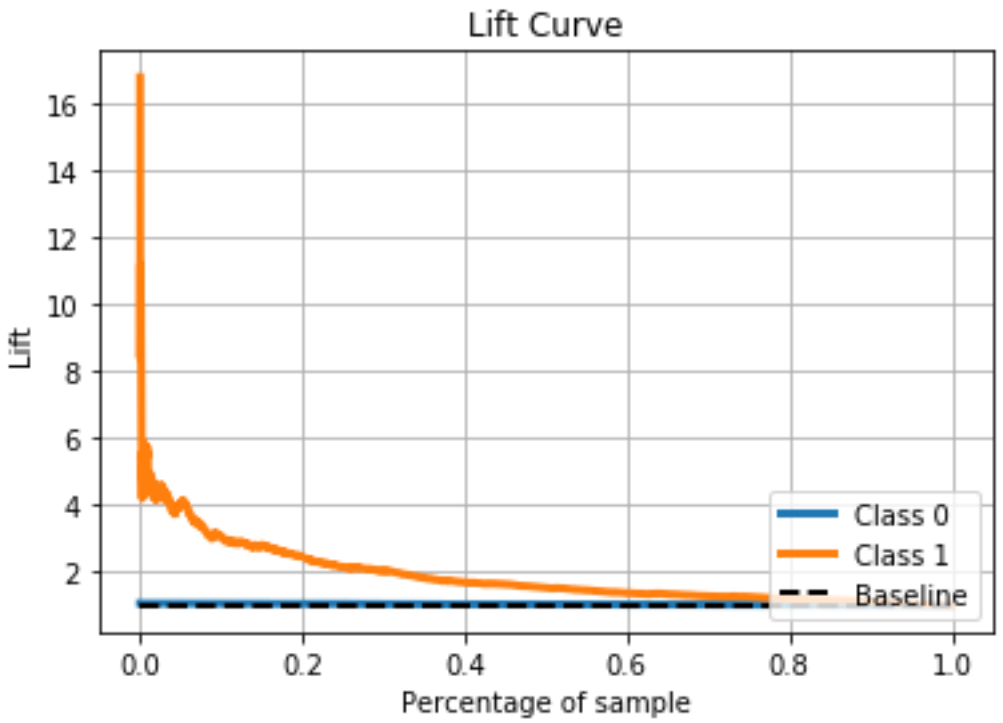
This phenomenon is due to an imbalance in the data. We see a significant proportion of the sample population does not have caravan insurance than those who do. Therefore, the models generally are great at predict who wouldn't buy caravan insurance and poorly at those who might be interested.

Another perspective is to look at the ROC and AUC. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.



I also observe the lift and cumulative gains chart. Lift charts measure the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model. This helps us to see which propotion of the population only needed to be contracted for soliciting caravan insurance. Ideally, the implication of these measures is to find the most gain for the model with the least amount of customer needed to be contacted, thus getting the most bang for your buck.

i.e Lift chart for Logistic Regression



The fourth step is to fine-tune the models using grid search or random search on the hyperparameters. I mainly use grid search try to get some improvements of these models. Unfortunately, the gain is minimal so it might not be worth the time spent to further optimize the model. Perhaps other projects could have different results. Also depending on the ML algorithm, the time spent to do grid search may not be worthwhile as these process could take a along time just to yield a minor improvement.

From this exercise we are able to find someway to model our data. However the result are far from ideal. The best performing model has an AUC of 0.713, meaning that there a lot of space for the model, if possible to improve. If I can devote more time, I will attempt to address the imbalance in the data between positive and negative response in our target variable.