Intro To Data Science Project Two

Stephanie Ajah; Ernest N. Frimpong

Executive Summary

We predicted car insurance uptake using 1613 features on the insurance dataset. Using random forests and logistic classification with a lasso penalty, the AUCs for both models are 0.67 and 0.65 respectively. Also, the amounts earned on revenue per person on each model were 18 and 16 cents respectively on Mihai's leaderboard. Below, we summarize the data pre-processing section, detail the models used and compare the two curves with the ROC curves.

Data Pre-Processing

On the training data, we had 5 duplicated rows, so we took those out leaving 101,886 rows. We proceeded to drop some rows based on variables that had erroneous values or abnormal values relative to the median for the variable and for the car industry in general. For instance, rows with '***year of birth'*** showing as '9999' were taken out and commute distances to work over 600km were taken out as people most certainly do not travel that far just to get to work.

We proceeded to take out variables that had over 80% values missing such as ***'years as principal driver'***. For the remaining variables with missing values, we fill with the median based on groups in other variables. We regroup ***'vehicle make'*** variable and frequency encode ***'vehicle model'*** before obtaining dummies. After creating dummies, we standardize our variables.

Modelling & Model Tuning

Model One

Initially, I trained a RandomForestClassifier using its default configuration and evaluated it based on the validation dataset. However, this approach yielded a poor balance between false positives and false negatives, as the model was overly focused on optimizing accuracy, earning only 2 cents in revenue on the class leaderboard. After some research, I discovered that the default threshold for classification in Random Forest is 0.5, which may not always be optimal for a given dataset. After recognizing this limitation, I decided to revisit my approach and fine-tune the threshold.

I split the dataset into training and validation sets using an 80-20 ratio and retrained the RandomForestClassifier on the training data. To refine the model's decision-making, I calculated False Positive Rate (FPR) and False Negative Rate (FNR) using the predicted probabilities for the positive class. I tested these rates across 100 thresholds evenly spaced between 0 and 1. By plotting and interpolating the FPR and FNR values, I found the threshold where the two rates intersected, achieving a balance between false positives and false negatives. This optimal threshold was 0.21.

I applied the threshold of 0.21 to the test data, leading to a significant improvement over the initial model's performance. It went from a revenue of 2 cents to 18 cents!

Finally, I generated a Receiver Operating Characteristic (ROC) curve to evaluate overall model performance, using the Area Under the Curve (AUC) as a key metric.

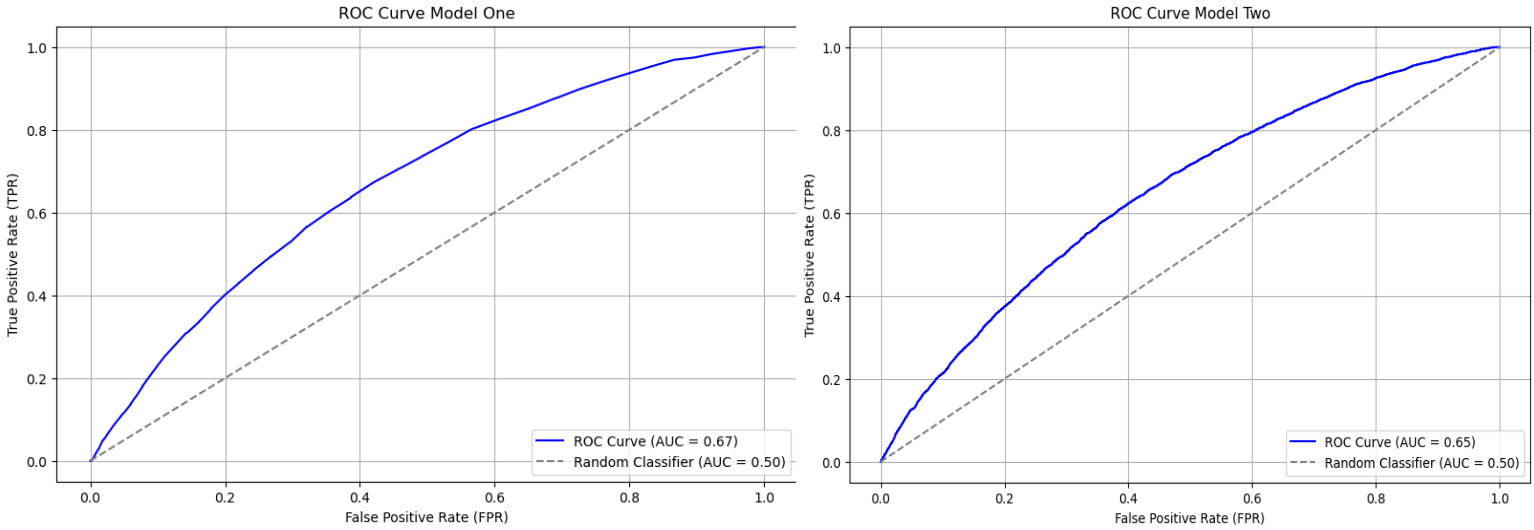
Model Two

I use two approaches. For model Two Approach A, I use principal component analysis (PCA) to reduce the dimensions and then apply logistic classification with a lasso regularization penalty. For approach B, I use the logistic classification and lasso penalty but without PCA. The approach with better results on the Mihai's test data is taken as the final model.

Approach A: I take an explained variance of 90% which was 32 components as a guide to create a range of component numbers (32 to 42) to cross-validate with logistic regression. That determined the optimal component of 35 which I used to create my PCA dataset. The PCA dataset was used to perform cross-validated logistic regression with a lasso penalty while optimizing on a range of thresholds. The optimal threshold of 0.23 was selected and resulted in revenue of 13 cents per person on Mihai's leaderboard.

Approach B: Using the original training dataset with 1613 columns, a cross-validated logistic regression with lasso penalty was done over a range of thresholds. On Mihai's leaderboard, revenue of 16 cents per person was obtained using an optimal threshold of 0.23. Approach B is used as the final model for Model Two.

ROC Curves For Both Models



Model's two ROC curve though very similar to the ROC curve for random forest in model one, the AUC for model two is slightly lower (0.65) than model one (0.67). Model one (random forest classification) outperforms model two (logistic classification).