University of St Andrews

Knowledge Discovery and Datamining

Practical 2

Report

Author and team name:
190030150

Professor:
Steve Drasco

April 17, 2020

1 Introduction

The goal of this practical is to predict the probability that a driver will initiate an auto insurance claim in the next year. The data includes a training and test set. First the data will be explored and cleaned. After cleaning the data some models will be fitted and evaluated. The models will then be uploaded to the Kaggle leaderboard where a score is calculated based on how well the model performs.

2 Data Exploration

The data includes binary features, categorical features and continuous features. The first step was to pick the features to use for model fitting. The features with missing values were explored and two features with the most missing values were removed since they have more than 40 percent of their values missing. A correlation plot of all the features were also plotted. It was discovered that all features with the name „calc“ in it were not correlated heavily with any other features. Those features were removed. The next step was to deal with the rest of the missing values. I used three seperate methods and evaluated which one gave the best results by fitting the baseline model to each dataset. The first method was to fill in the missing values with the median value for continuous features and the most common value for categorical features. The second method was to make a seperate category for missing values for categorical data. The third method was to interpolate the continuous feature with the values being treated as being equally spaced. Finally dummy variables were created for all categorical features.

3 Model fitting and evaluation

The models will be evaluated using cross validation where the ROC AUC score will be used as an evaluation metric. The ROC curve plots the false positive rate against the true positive rate and The ROC AUC score is calculated as the area under the ROC curve. The closer the score is to 1 the better. For cross validation the data will be split into five folds where one of the folds acts as a validation set and the rest as a training set.

The first model we fit was a logistic regression. This model was used as the baseline model since it is quite fast. The model was first fit on three datasets to pick the best method for dealing with missing values. The mean of the ROC scores indicated that method 2 gave the best results. That model had a mean ROC score of 0.629 and will be our first model. The model was fit on the unseen test dataset and uploaded to Kaggle. The score calculated was 0.258.

The random forest classifier is next up. It constructs multiple of decision trees on random subsets of the data. The output is the class that is most commonly picked by the individual decision trees. It was picked since it is relatively fast and performed well on the last practical. The model was fitted on the training set first with 200 estimators and 6 as the maximum depth. The ROC AUC score of that model was 0.625. The model was then improved by finetuning the hyperparameters max depth and number of estimators. Due to limited computing power few values were selected and a randomized grid search was used. The best random forest model returned had 200 estimators and 10 as maximum depth. The mean ROC score for that model was 0.631. The model was fit on the unseen test data and uploaded to Kaggle where it got a score of 0.256. It performs slightly worse on the unseen test dataset than the previous logistic regression model.

The final model used was Adaboost. This model first trains a simple decision tree and uses it to make predictions. It then resamples the data with increased weights of the instances where predictions are wrong. With that data it fits another decision tree and updates the weights again which makes the model gradually better. The model was fitted with 200 estimators and a learning rate of 0.5. The model was then improved by finetuning the hyperparameters using randomized gridsearch. After finetuning the same model was returned as the best model. The model had a mean ROC score of 0.634. The model was fitted on the unseen test dataset and uploaded to Kaggle. The Adaboost algorithm got the best score which was 0.267 which is an improvement and my best result.

4 Conclusion

The models that were fitted did reasonably well and I am happy with the results. However there is always room for improvement. More time could be spent on data exploration and feature engineering. Finding a better way to deal with missing values might improve the results. For example predicting the missing value instead of filling them with the median and the mode. Including interactions might also help. By combining features together and adding the best ones to the dataset could improve the score. Experimenting with more models, finetuning the previous models more or combining models could also yield improved results. Three models were tried in this project and many more exist. Training a neural net or a support vector machine are some of the possibilities.