Machine Learning for Classification

Data Mining Module Assignment

ANLY-533

Date: November 26th, 2017

Author: Adam Sampson

Introduction: The goal of this analysis is to build a model to predict whether or not a driver insured with Porto Seguro will file an insurance claim. The data used was provided through a Kaggle competition and was provided as-is with no explanation regarding variable meanings. Multiple deliverables are provided including a submission to the Kaggle Competition, this report, the code used to generate the model, and an additional set of binary scores for the test set. All of the above are available on GitHub.

<https://github.com/adam-sampson/Kaggle-Competition-Porto-Seguros-Safe-Driver-Prediction>

# The Business Problem:

With insurance claims it is highly desired to determine information about the risk being undertaken when insuring customers. The most basic of understanding is being able to predict what percentage of customers will file a claim in each year, and the amount that will be spent closing out claims. This information is the base level required to maintain a profitable business that takes in at least as much money as it spends.

However, more information is desired in order to maintain competitive edge and to undercut competitors by enticing low risk customers with competitive pricing, while charging higher rates for higher risk customers. The better a model is at differentiating between low risk and high risk customers, the better it is able to drive profit margins and competitiveness in the market.

# Overview of the Data:

The data provided was scrubbed to withhold identifying information. All of the information was provided either as binary (bin) data, categorical data (cat), ordinal data, or interval data. The binary and categorical variables are marked, but it is not possible to be totally sure of whether the remaining data is interval or ordinal. This is additionally compounded by the fact that many variables appear to have undergone some sort of transformation or normalization procedure prior to being revealed.

The data contains information about a few different types of information:

* Information about the individual (ind)
* Information about the car they drive (car)
* Information from their registration (reg)
* Calculations by Porto Seguro (calc)

The data is provided in 2 sets. The first is the training set and includes 59 variables/columns including the result columns “target”. This set has 595,212 rows. The second set of data is the competition test set which has 58 variables (it does not have “target”) and has 892,816 rows.

# Data Preparation Methods:

Much of the review performed on the data was perfomed in the file Sampson\_Explore\_Data.R (see GitHub).

The first step taken was to review the data to determine which variables should be treated as categorical an which should be treated as interval. It was decided that for the sake of speed ordinal values should be treated at interval values. In order to validate this decision a model was made with 20,000 randomly selected data points to compare a logistic regression with them as categories to a logistic regression with them as interval data. The change did not have a significant improvement on the results, but did drastically change the modeling time.

The next step was to review the variables to determine which ones appear to make good variables. This was done visually through graphing the possible values of a variable with two bars. The first bar indicates how often that variable outcome value results in a target of 0, and the second bar indicates how often that variable outcome value results in a target of 1. Bars that are evenly matched are equally likely to results in a target outcome, but bars that are heavily mismatched are likely to be predictive. This graph also shows another piece of information on the same time. The bars for each possible value of the variable indicate how often that particular values occurs in the dataset. In other words, we are looking for highly uneven sets of bars that are large compared to the rest of the options in the data.

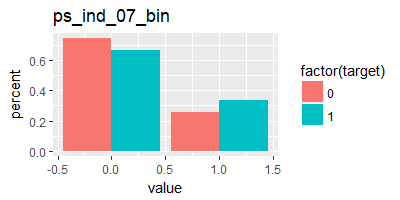


Figure - Example Plot – Red and Teal are mismatched indicating possible prediction

The next step in preparing the data was to deal with missing values indicated by -1 in the data. After reviewing the data a two-step process was decided upon for dealing with these values. First, any missing values that accounted for more than 0.5% of the values for the variable were annotated by creating a new binary column that samply indicated whether the value was missing or not (I.E. ps\_reg\_03\_miss). This was done due to the fact that many of the missing values seem to have predictive value when reviewing the graphs from the step above. Next, the missing values in the main variable column were replaced with either the mean or the mode depending on the data type. Binary or Categorical missing values were replaced with the mode. Interval or Ordinal missing values were replaced with the mean. Ordinal was handled this way because the models are treating the value as interval in my analysis. Otherwise the median would have been used.

Next, the data was converted to appropriate data types. Binary and Categorical variables were converted to factors. Ordinal and Interval data was left in either integer or numeric data types.

Finally, a validation set was randomly sampled from the training set. The set used to train the models was 80% of the total training data, and a validation set of 20% of this total training data was held in reserve to evaluate the models.

# Modeling Methods:

The first step in modeling was to get a baseline performance. This was done by first randomly sampling 20,000 data points out of the full data set in order to speed up modeling times. Then, with no feature selection, only conversion of binary and categorical to factors, models were run for logistic regression, naïve bayes, c5.0, and neural net were run. The best performing model (based on the validation set) was the logistic regression with an Area Under the Curve (AUC) of 0.5594.

Next, modeling was performed with an aggressive feature selection attempt from the visual indicators. Variables that did not appear relevant were removed, and values within categorical variables were grouped in an attempt to decrease the sparseness of the data being used in the model. This model was completely ineffective and either match or lagged behind the baseline model using all variables.

The second attempt at feature selection simply removed all variables/columns which were identified in visual inspection to be poor indicators of the target variable. Once again, logistic regression was the best predictor with an AUC of 0.5611.

A third attempt at features selection was performed using the results of the logistic regression in attempt two to remove additional variables. This attempt reduced simplified the model, but also reduced the AUC for all models.

A fourth attempt at feature selection was perfomed using the results of the logistic regression from step one. Variables that were not significant in this regression were removed from the model. This resulted in similar results to the second attempt, but with an AUC of approximately 0.58 which was not as good as the second attempt.

Next, the full data set was used to create logistic regression models using feature selections 2 and 4. This improved the models significantly. However, feature selection method 2 remained the best with an AUC of 0.6183, compared to feature selection method 4 which resulted in an AUC of 0.6055. This was a significant improvement compared to using 20,000 randomly selected rows, which is to be expected due to the fact that more information was included when using the full data set.

# Evaluation of Results:

The best model created was a logistic regression using feature selection number two. The AUC was 0.6183 which shows that the model is a fair improvement over simply predicting the target using random guessing.

# Deployment Recommendations:

A linear regression model using the code from the Functions.R function “performFeatSelection2(in.dt)” creates a useful model for predicting whether or not a customer is likely to make a claim. However, this model generates a significant number of false positives. This means that the results need further evaluation by subject matter experts to determine whether or not it makes good business sense to use these predictions to increase rates for risky drivers, and to lower rates for less risky drivers. Due to the lack of labelling of the data it is not possible to begin the next step of identifying traits that this insurance company should look for.