Semester Project

William Huibregtse[1], Joshua Baker[1], Chris East[1]

Abstract
Include abstract here – A summary of your work

Keywords
Keyword1 — Synergy — Keyword3

[1] *Computer Science, School of Informatics , Computing and Engineering, Indiana University, Bloomington, IN, USA*

Contents

1. Problem and Data Description

First we want to get a general idea of our data set and get a deeper understanding of the underlying structure.
There are 59 named features or variables for our data set.
With 892816 observations for training and 595212 for test
There are no duplicate observations.
Features that belong to similar groupings are given certain feature names.
- Ind: related to individual or driver
- Reg: related to geographical region
- Car: related to car being insured
- Calc: are calculated features done by Proto themselves

Postfix descriptors describes the features data type.
- Bin: Binary (1 or 0)
- Cat: Categorical *Note: the dataset has the categorical data already convert into factors and then integers
- All other variables are either integer or numeric

As stated the Data Types are numeric and integer, with integer being the predominant type 49 to 10.
Missing values are represented by -1.
In total, there are 13 variables with missing values.
There is Target feature which denotes the binary classification for that observation. This feature is the feature we are trying to learn/predict for the test data.
There is an ID feature which is an anonymized identities of insured drivers.

Porto Seguro's Safe Driver Prediction has 59 variables and 1.3 million observations, which qualifies as a good candidate for reducing overall dimensions of the data to significantly increase the speed of analysis techniques at the cost of more upfront data processing.

There are only 21694 cases of classification 1, which is 3.64 percent of the observations in the training data set, showing significant skew in the expected class towards a "0" prediction.

2. Data Preprocessing & Exploratory Data Analysis

2.1 Feature Engineering

As even missing data can be significant, a new feature was added to the data set. This feature was the count of missing values for each entry before these missing values were processed. This technique allows the retention of the potentially useful information provided by the missing values.

2.2 Handling Missing Values

After observing the summaries of each variable in our data sets, it was clear that variables ps-car-03-cat and ps-car-05-cat contained mostly missing values for each data set. Because we later used a missing value replacement techique to process the data, applying this techique to variables with mostly missing values may have overfitted and affected the outcome of future analysis, thus both because of this possible overfitting and that useful information may be captured in the engineered feature of missing value counts, these columns were removed before proceeding with our NA replacement technique.

After the columns with mostly missing were removed, we replaced missing values with the mean of it's variable using the R package "mice".

2.3 Exploratory Data Analysis

To reduce the complexity of future analysis techniques, we used the prcomp function from base R to perform PCA dimensionality reduction on our data. After the transformation, the first 16 principle components represented over 99 percent of the the data's variance. Using this as a reasonable cut-off, we proceeded forward using only these first 16 principle components.

Note: To perform this transformation correctly, the test and train data must be transformed together and split after the dimensions have been reduced.

3. Algorithm and Methodology

Once the data was processed, we used a NaiveBayes algorithm (from base R) to form our model.

4. Experiments and Results

Upon completing the first predictions against the autograding tool on Kaggle, different sizes of highest variance principle components were tested to see the effect on overall GINI score. A jump from 0.125 to 0.20 was achieved by including the first 16 principle components versus originally using just 5 (99 versus 95 percent of total variance represented in each sample, respectively).

5. Summary and Conclusions

Acknowledgments