Proposal

1. Domain Background

I entered the "Porto Seguro's Safe Driver Prediction" Kaggle contest for this capstone project. Nothing ruins the thrill of buying a brand-new car more quickly than seeing your new insurance bill. The sting's even more painful when you know you're a good driver. It doesn't seem fair that you have to pay so much if you've been cautious on the road for years. Porto Seguro is one of Brazil's largest auto and homeowner insurance companies. Inaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones.

This is the typical problem in the actuarial science field, where actuaries use different modeling techniques and thousands of variables to predict the probability for certain customer to file a claim. Most of the models developed by actuaries are supervised learning models, because the historical target variable is known. (claim or no claim)

In this competition, I was challenged to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year. While Porto Seguro has used machine learning for the past 20 years and a lot of papers have been published. For example, Smite. K. A discusses this issue in his paper "*An analysis of customer retention and insurance claim patterns using data mining: A case study*". However, they're looking to explore new, more powerful methods now. A more accurate prediction will allow them to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.

The detailed information about this data competition can be found online (https://www.kaggle.com/c/porto-seguro-safe-driver-prediction)

2. Problem Statement

For this capstone project, I will predict the probability that an auto insurance policy holder files a claim in the next year with Porto Seguro's historical data. This is a classification problem and the final output for each record will be 0 or 1. "0" means there is no claim in the next year and "1" means there is a claim in the next year. Below is a clear and rigorous statement of the problem:

- I was asked to develop a statistic model to learn from Porto Seguro's historical claim history. This model will be a supervising learning model (most likely logistic regression model) and the final accuracy will be measured by normalized Gini index. This model will be acceptable only if the calculated normalized Gini index on the testing data is over 0.2 and the results will be better if more data is provided.

3. Datasets and Inputs

Two datasets are provided by Porto Seguro on Kaggle: train.csv and test.csv. Those data are extracted from Porto Seguro's claim history.

Smith, K. A., et al. "An Analysis of Customer Retention and Insurance Claim Patterns Using Data Mining: A Case Study." The Journal of the Operational Research Society, vol. 51, no. 5, 2000, pp. 532–541. JSTOR, JSTOR, www.jstor.org/stable/254184.

- train.csv contains the training data, where each row corresponds to a policy holder, and the target columns signifies that a claim was filed.
- test.csv contains the test data.

In the train and test data, features that belong to similar groupings are tagged as such in the feature names (e.g., ind, reg, car, calc). Values of -1 indicate that the feature was missing from the observation. The target columns signify whether or not a claim was filed for that policy holder.

The specific 59 potential variables I will be using are as follows:

ps_ind_01 ps_ind_02_cat ps_ind_03 ps_ind_04_cat ps_ind_05_cat ps_ind_06_bin
ps_ind_07_bin ps_ind_08_bin ps_ind_09_bin ps_ind_10_bin ps_ind_11_bin ps_ind_12_bin
ps_ind_13_bin ps_ind_14 ps_ind_15 ps_ind_16_bin ps_ind_17_bin ps_ind_18_bin
ps_reg_01 ps_reg_02 ps_reg_03 ps_car_01_cat ps_car_02_cat ps_car_03_cat
ps_car_04_cat ps_car_05_cat ps_car_06_cat ps_car_07_cat ps_car_08_cat ps_car_09_cat
ps_car_10_cat ps_car_11_cat ps_car_11 ps_car_12 ps_car_13 ps_car_14
ps_car_15 ps_calc_01 ps_calc_02 ps_calc_03 ps_calc_04 ps_calc_05
ps_calc_06 ps_calc_07 ps_calc_08 ps_calc_09 ps_calc_10 ps_calc_11
ps_calc_12 ps_calc_13 ps_calc_14 ps_calc_15_bin ps_calc_16_bin ps_calc_17_bin
ps_calc_18_bin ps_calc_19_bin ps_calc_20_bin

Feature names include the postfix “bin” to indicate binary features and “cat” to indicate categorical features. Features without these designations are either continuous or ordinal. There are 595212 records in the training sample and the data is not balanced. (Only 21694 insurers filed claims, which is less than 4% of the total training sample)

Although there are over 500,000 observations, the training dataset itself is not very big. I will read the training dataset provided in excel as a panda dataframe and further divides it into two parts. One for training and other one for validation.

4. Solution Statement

Since the target variable is a binary variable (claim or no claim). I plan to mainly use a logistic regression model and try out different set of explanatory variables provided by Porto Seguro. The final solution will be a logistic regression model with the best model fitting performance.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Before I start to fit the models, I will perform the following pre-processing steps:

- Checking for missing values
- Normalizing Numerical Features
- Creating dummies for categorical variables
- Using recursive feature elimination method to select
- Checking correlation to avoid colinearity issue

Smith, K. A., et al. “An Analysis of Customer Retention and Insurance Claim Patterns Using Data Mining: A Case Study.” The Journal of the Operational Research Society, vol. 51, no. 5, 2000, pp. 532–541. JSTOR, JSTOR, www.jstor.org/stable/254184.

Meanwhile, I will also test some more complicated models, such as random forest, GaussianNB, support vector machine, AdaBoostClassifier.

To summarize, a clear solution statement is as follows:

- I will build a supervise learning model with the highest normalized Gini index on validation dataset, which is most likely to correctly predict if certain customer will file a claim in one year based on Porto Seguro historical claim history. This supervise learning model may be a logistic regression model, random forest model, Guassian NB model, SVM model, or Adaboost classifier model.

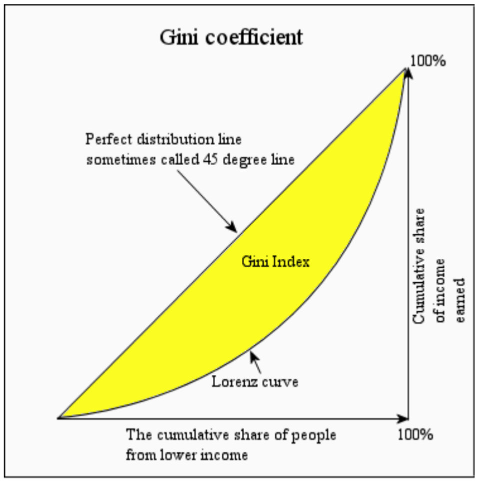
5. Benchmark Model

Since this project is a Kaggle contest, I can submit my results after I develop my model. There is a leaderboard on the Kaggle website, where over 3500 teams have reported their model performance on the hold-out testing dataset. Currently, the best result has a Gini score of 0.29. I plan to keep revising my model until I get similar performance.

Also, I will train a simplified random forest model as a benchmark. I will just train this benchmark model on all original explanation variables without any pre-processing steps. In the end, my final solution will most likely outperform this benchmark regression model.

6. Evaluation Metrics

My submissions will be evaluated using the Normalized Gini Coefficient.


Gini coefficient
100%
Perfect distribution line
sometimes called 45 degree line
Gini Index
Cumulative share of income earned
Lorenz curve
The cumulative share of people from lower income
100%

During scoring, observations are sorted from the largest to the smallest predictions. Predictions are only used for ordering observations; therefore, the relative magnitude of the predictions are not used during scoring. The scoring algorithm then compares the cumulative proportion of positive class observations to a theoretical uniform proportion. The Gini coefficient is defined as a ratio of the areas on the Lorenz curve diagram. If the area between the line of perfect equality and Lorenz curve is A, and the area under the Lorenz curve is B, then the Gini coefficient is A/(A+B). Since A+B = 0.5, the Gini coefficient, G = 2A = 1-2B. If the Lorenz curve is represented by the unction Y = L(X), the value of B can be found with integration and:

Smith, K. A., et al. “An Analysis of Customer Retention and Insurance Claim Patterns Using Data Mining: A Case Study.” The Journal of the Operational Research Society, vol. 51, no. 5, 2000, pp. 532–541. JSTOR, JSTOR, www.jstor.org/stable/254184.

$$G = 1 - 2\int_0^1 l(x)Dx$$

The Gini Coefficient ranges from approximately 0 for random guessing, to approximately 0.5 for a perfect score. The theoretical maximum for the discrete calculation is (1 - frac_pos) / 2.

The Normalized Gini Coefficient adjusts the score by the theoretical maximum so that the maximum score is 1.

7. Project Design

I will start with a high-level summary of my project design. As the first step, I will take a look at the dataset provided by Porto Seguro and make certain adjustments (e.g. remove missing values). Then, I will try out different kinds of supervising learning models such as logistic regression model to see which one gives me the best fitting performance on the validation dataset. After having several candidate models, I will then check the model performance from different angles. (e.g. Backtesting, sensitivity analysis, benchmarking, and regression diagnostics). In the end, I will select the modeling approach with the best fitting performance and further tweak it. More details about my project design are provided below:

Logistic regression remains the most commonly applied regression technique used with binary target variables. (e.g. “0/1”, “yes/no”). After picking logistic regression as my main approach, I will then evaluate the explanatory variable and target variable in the training data. This process involves both the business rationale check and statistical tests. (e.g. correlation). The business information provided in the data set is very limited, I will try my best to extract as many useful insights as possible. The pre-processing steps I plan to take are below:

- Checking for missing values: Only 3 features have missing values, so I will just use feature reduction, which will not remove too much useful information
- Normalizing Numerical Features
- Creating dummies for categorical variables
- Using recursive feature elimination method to select
- Checking correlation to avoid collinearity issue

Once the initial inspection is done, I will start to write code script to build the logistic regression model. Variables will be picked using stepwise selection method. (Keep removing the variable that contributes the least). After having my preliminary model, I will then check the model performance from different angles to further improve the model accuracy. (e.g. Backtesting, sensitivity analysis, benchmarking, and regression diagnostics).

Although I picked logistic regression as my main approach, I will also try out other advanced statistical models. For example, I will build some challenger models using GaussianNB/ Adaboost/ Random Forest techniques. I will use grid search technique to tune my supervised model.

I will spend most of the time on the logistic regression by going through all the pre-processing steps as discussed above, trying different sets of the explanatory variables, and tuning the model with grid

Smith, K. A., et al. “An Analysis of Customer Retention and Insurance Claim Patterns Using Data Mining: A Case Study.” The Journal of the Operational Research Society, vol. 51, no. 5, 2000, pp. 532–541. JSTOR, JSTOR, www.jstor.org/stable/254184.

search technique. For all the other supervised learning models, I will mostly reply on the grid search technique and less effort will be spent.

Smith, K. A., et al. "An Analysis of Customer Retention and Insurance Claim Patterns Using Data Mining: A Case Study." The Journal of the Operational Research Society, vol. 51, no. 5, 2000, pp. 532–541. JSTOR, JSTOR, www.jstor.org/stable/254184.