MACHINE LEARNING ENGINEER NANODEGREE

Capstone Proposal

OCTOBER 28, 2017

ANSHUL PRAKASH

Proposal

With predictive analytics and machine learning being the buzzword today, each organization wants to know their customer better to provide them the best products and services at the right price. This understanding of customers involves not only getting insights from the past activities but also to predict the future activities as accurately as possible. This is of utmost important when it comes to pricing auto insurance as predicting probable future claims can be extremely helpful in pricing the products accordingly. However, this power of pricing using the predictive power of machine learning algorithm comes with a caveat as an error in predicting can be really damaging specially when the algorithm predicts that a customer will make a claim and that does not happen. This results in an overpriced insurance bill for a customer who may have a safe driving record.

So, as my capstone project for Machine Learning Engineer Nanodegree program I plan to build a model to predict the probability that a driver will initiate an auto insurance claim in next year. This is an ongoing competition on Kaggle with the title “Porto Seguro’s Safe Driver Prediction”. [1]

Domain Background

Auto insurance protects a user against monetary loss if they have an accident. It is a contract between the user and the insurance company. The user agrees to pay the premium and the insurance company agrees to pay losses as defined in the policy. Predictive analytics is used in appraising and controlling risk in underwriting, pricing, rating, claims, marketing and reserving in insurance sector.

In underwriting, predictive analytics enable better risk assessment and classification which leads to better pricing. Pricing analytics can be the “tipping point’ for consumers in choosing insurers and products. They also lead to better profitability for insurers as they target desirable customer segments and support real time, dynamic pricing.

The use of predictive analytics has become more common in the insurance industry in recent years. Predictive analytics for insurance entails the use of special technology to sift through and analyze historical data and consumer trends in effort to project future behavior. By studying the behavioral tendencies of varying demographics under differing sets of environmental circumstances, companies can learn what products those people might be inclined to buy, how best to reach them and most importantly what should be the price of the product so that it is a win-win situation for both the user and the company.

Porto Seguro is one of Brazil’s largest auto and homeowner insurance companies. Porto Seguro has used machine learning for the past 20 years to offer the right product to right customer.

Problem Statement

Inaccuracies in car insurance company’s claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones. The challenge is to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year.

A more accurate prediction will allow Porto Seguro to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.

Datasets and Inputs

The dataset [2] provided on Kaggle consists of both training and testing data. In training data each row corresponds to a policy holder, and the target columns signifies that a claim was filed.

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). In addition, 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. Values of -1 indicate that the feature was missing from the observation. The target column signifies whether or not a claim was filed for that policy holder.

Solution Statement

The solution will comprise of a model that will use all the relevant features present in the training data to predict the probability that a driver will initiate an auto insurance claim in next year. These predictions can be used by Porto Seguro as inputs to their pricing algorithms to give the most accurate estimate for the price of product.

Benchmark Model

Currently the best model in the public leaderboard on Kaggle has score of 0.29. The maximum possible score is 0.5 and its calculation is described in the next section.

Evaluation Metrics

Submissions for this challenge on Kaggle are evaluated using the Normalized Gini Coefficient.

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 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.

Project Design

First, data exploration will be done in order to better understand the data provided. Description for all the features present in the training data will be sought and if there are any abnormalities in data then they will be addressed appropriately.

Next, Exploratory data analysis will be done using various visualization libraries of python. This will give some idea about which features are relevant and what is their degree of effect on the target variable.

If possible, feature engineering will be attempted so as to give the best optimal model.

Then modelling will be done first using some basic models such as logistic regression and decision trees. Then some advanced techniques such as random forest, XGBoost will be attempted based on the score achieved.

Target will be to be in the top 50 of the public leaderboard or a score greater than 0.30.

References

[1] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction

[2] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data