Machine Learning Engineer Nanodegree

Capstone Proposal

Kalle Bylin
November 18th, 2017

Proposal

Domain Background

Most people consider life to be greatly unpredictable. In the middle of it all great effort is dedicated to maximizing utilities and reduce costs. Still, the unpredictability and uncertainty of most events often generates unexpected or unwanted outcomes. These outcomes can be caused by natural disasters, accidents, changing trends, premeditated harm, etc. Insurance is a means of protection against risks and was first used by Chinese and Babylonian traders several millennia ago [1] to spread risks from an individual to larger communities. In other words, the unfavorable event is not necessarily eliminated, but its effects are less critical to the victim because he is part of a community in which everyone is contributing expecting to receive help in case of an unfavorable outcome.

Today insurance companies are the entities focused on creating these communities with pooled resources. This pooling must consider the nature of the risk that is being hedged against. On one side, the cost of being insured has to be low enough for the insured to consider it beneficial. On the other side, the insurance fee must be high enough to cover the losses that may actually happen. Many insurance companies are now taking a data-centric approach to establish these fees. The overview of Kaggle's Porto Seguro Competition states that Porto Seguro, a Brazilian insurance company, has been using machine learning for the past 20 years [2].

Problem Statement

Being able to predict a claim, and the unwanted events, before they happen allows the insurance company to set accurate fees. This particular project focuses on claims made for automobile accidents. In this case, the risk of an accident is also highly influenced by the driver. For this reason, Porto Seguro, one of the largest automobile and homeowner insurance companies in Brazil, wants to be able to predict the probability that a driver will file an auto insurance claim in the next year. This prediction is used to set insurance fees.

This is a supervised binary classification problem. Each client is predicted to file a claim or not and by the end of each year, the company can compare the prediction for each client to what really happened and assess the accuracy of the model.

A quite simple but not very useful approach would be to each year calculate the ratio of people out of all the clients that initiated an insurance claim and use this as the probability of any driver initiating a claim. On top of this, machine learning techniques can be used to take into account the importance and value of each feature to generate a more accurate prediction for each individual driver.

In short, we will be using the data provided by Porto Seguro consisting of a training set and a test set. This data was collected during the period of a year and each row is a customer. The columns are anonymized features related to each customer. This input will be used to train a machine learning model using a specific evaluation metric (which will be explained in more detail further on in this proposal). The objective is for this model to output the probability for each client filing a claim or not in the coming year.

Datasets and Inputs

The dataset for this competition has been provided by Porto Seguro on Kaggle and has been anonymized to avoid revealing personal information of its clients and to prevent competitors from taking advantage of their data. It has been collected by the company and has already been through a certain degree of cleaning and preprocessing. The dataset can be found here: https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data

The data has already been split into a training and a test set. The training set has a total of 595.212 rows and 59 columns. Out of those, one column is for the targets indicating if each client filed a claim or not. The target variable for the test set is stored on Kaggle's servers and cannot be used locally. The features have been grouped according to similarity and this grouping can be seen in the names of each feature. "Car" means that the feature is related to the car, "ind" that it is related to the driver or individual, "reg" that it is related to the region and "calc" that it is a calculated feature.

Additionally, each feature name also indicates if it is binary ("bin"), categorical ("cat"), or continuous (the rest). Not knowing exactly what each feature represents is an added challenge. Still, it is relevant and appropriate for this problem because we do know that they describe different aspects of the car, the driver, the population and the geographical region. The calculated features are also assumed to have been created to add important information. The rules of the competition prohibit using external data. It could be that the calculated fields also already contain data collected from other logical sources.

For this project, the data will be stored locally in the project folder as .csv files and will be loaded directly into the working environment with Pandas.

Solution Statement

This dataset has labels available for training and which we would want to predict when using the model on new data. For that reason, supervised machine learning techniques are going to be used with the target variable as the labels for training. The problem is also binary so different methods like logistic regression or decision trees can be used to create a binary classification model. Ensemble methods like random forest can also be used to train a classifier on this data.

The data has already been through a certain amount of preprocessing and the anonymization makes it more challenging. But to achieve a viable result, it is important to study the features and try different transformations or methods of feature engineering to make it easier for the models to learn.

In short, the training data will be analyzed and studied to understand it better. With this information the data will be preprocessed to make it more suitable for the machine learning methods mentioned above. This data contains information about the cars, drivers, region, etc. and will at this point be used as input to train different models. After training, each model will be used to generate predictions of validation data and then of the given test set by the Kaggle competition. The model with the highest score for the chosen evaluation metrics can then be used to predict future claims. The score provided by Kaggle will also be included in the analysis.

Benchmark Model

The company behind the competition does not want to share their benchmark model or score as they consider this information part of their competitive edge. For that reason, this project will be starting off with a Naïve model as a benchmark in which all the company's clients are treated the same.

This model will be evaluated with the same metric described in the next section which will also be used to evaluate the final model. In other words, predictions will be made using the Naïve model which will be compared to the target variable and evaluated according to the chosen metric. For example, if the metric is accuracy then the Naïve model is going to evaluated according to how many of the predictions are correct with respect to the target variable.

Evaluation Metrics

The accuracy could be a first and relevant evaluation metric. This is a supervised classification problem, so it is relatively easy to calculate how many of the drivers were

classified correctly. Still, in this context it is not the best metric because, given the nature of the industry, there should be significantly more drivers that don't file claims than drivers that do. A very good accuracy could be achieved by just classifying every driver as "not filing a claim in the next year".

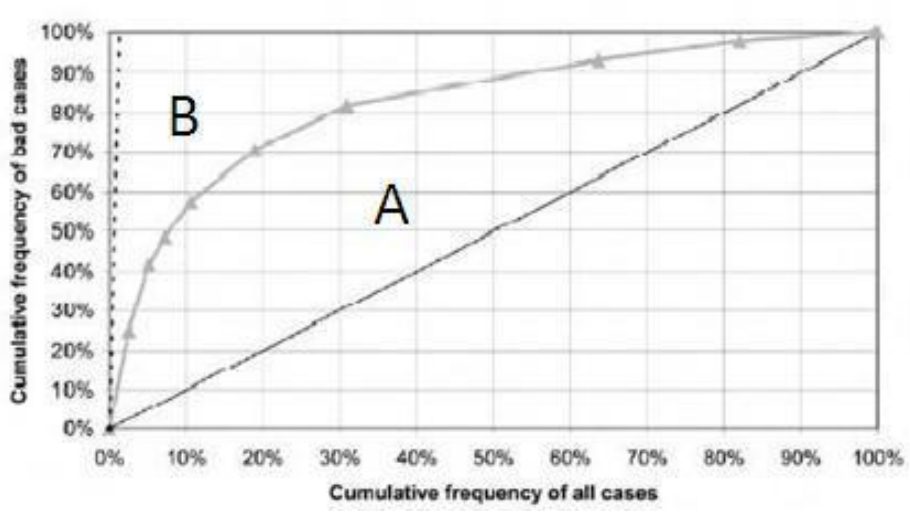
A better evaluation metric would be the F-beta score that takes into account precision and recall. In this case the F-beta is mentioned instead of simply the F1 score because it can be important for the insurance company to have better recall. In other words, it is important to evaluate out of the drivers that did file a claim how many were classified as such, to better adjust the insurance fees.

It is also important to mention that the competition evaluates submissions using the Normalized Gini Coefficient. The Gini coefficient is best known for and is usually used to assess inequality. But it is also used to, for example, evaluate the predictive power of credit scoring tools [3]. A good credit score is supposed to assign high scores to safe applicants and low scores to riskier applicants. The Gini Coefficient is used to evaluate how well it does that. It is a scale of predictive power from 0 to 1 in which a lower score means less predictive power and is equivalent to a random prediction, like using a coin toss. A Gini Coefficient of 1 indicates perfect predictive power. For credit scores it perfectly identifies who will repay and who will default.

There are several ways to calculate the Gini Coefficient, but a simple way it can be calculated using the ROC curve as the ratio of the area between the ROC curve and the diagonal line, and the area above the ROC curve. It can also be calculated with this formula:

$$Gini = 2 * AUC - 1$$

Being AUC the area under the ROC curve. The next image shows a more intuitive graphical representation.



Figur 1 Source: https://thealphastrategist.wordpress.com/2014/08/11/an-insiders-guide-to-predictive-modeling-comparison/

Following the calculation stated above it is calculated in this image as:

Gini = A / (A + B)

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

Project Design

The objective of this project is to generate a model that can predict if a driver insured by Porto Seguro is likely to file a claim or not in the next year. This model will be chosen with empirical evidence and will require a certain degree of experimentation. The data provided by Porto Seguro will be used to train different models which will be assessed with respect to a given evaluation metric. The model with the highest score on unseen data (the test set) will be presented as a tool for Porto Seguro to better understand its clients, predict claims and improve its insurance fees.

The data for this project is available on https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data. The first step is to download the data and set up a directory for the project. A virtual environment can also be created specifically for this project to install the necessary dependencies and libraries.

With this set up the first step is to load the data and perform some initial data exploration to understand it better. It is useful to study the size, shape and content of the data. We already know that it is anonymized, but it is still worth checking out the different features and which type they are.

After a general look at the data, the target variable is going to be analyzed in more detail to know exactly what it is we want to predict and how to do it. As mentioned above, due to the nature of the industry, accidents are not exceptional events and thus we would expect to see many more drivers that don't file claims than those that do. Imbalanced data can often make it more difficult for models to accurately learn, so at this point it might be recommended to apply techniques like sampling to make the classes more balanced.

After this, the features can be explored in more depth. A few tools that can help us understand the data better could be a heatmap showing correlation between the features or pairplots to gain a better understanding of the distribution of each feature and its relation to other features.

At this point, it is often the case that data is highly skewed and not normally distributed. It can be helpful to apply non-linear scaling like the Box-Cox test to transform the data into a form which is more suitable for training.

After gaining a deeper understanding of the data we can start constructing models that can learn to predict the desired variable. First, we can construct a Naïve model to use as a benchmark as stated above. Given that this is a binary classification we can then try a logistic regression and then more advanced models and compare if there is any improvement in the predictive power of the model.

In order to improve models, we can also try applying techniques like grid search with different parameters and cross-validation. This is useful to save time instead of manually trying out different parameters for the model.

Once this is finished it is important to analyze the results of the different models and choose the one with the best performance given the chosen evaluation metrics. This is the model that will be presented to be used to predict future claims by the insurance company. A last step is to also present possible recommendations and improvements to the model.

Sources:

1. https://en.wikipedia.org/wiki/History_of_insurance
2. https://www.kaggle.com/c/porto-seguro-safe-driver-prediction
3. https://www.eflglobal.com/every-lender-has-a-gini-coefficient-so-what-exactly-is-it/