Machine Learning Engineer Nanodegree

Capstone Proposal

Mohamad Tarbin
February 13th , 2018

Proposal

Domain Background

Banks consider the denial of credit applications to risky customers an important priority in order to avoid undesirable decisions and consequences, like granting a customer a credit limit increase to risky customers. For example, a customer who has skipped making the minimum payment for several months. In such case, the credit card should be set to default after the customer has failed to make a payment for 6 months in a row. A credit default is a credit status applied when a customer fails to make the minimum payment for 6 months.

Risk Management is important field in Banking, and when we are talking about credit cards, credit risk management will be the first line of defense for any credit decision. Performing credit risk management correctly allows banks to minimize the risk associated with credit cards, such as avoiding the increase of credit limits for risky customers. For example: Helping a customer when a crisis or natural disaster occurred. The decision should go into Risk Analyst. I have been working In the credit risk Management Field for over 2 years, so this subject is of great interest and relevance to me.

A research has been done by Department of Economics The Ohio State University that is show using data it's and machine learning will give ability to find complex pattern in the user behavior and find new aspects of credit card behavior(AN EMPIRICIAL INVESTIGATION OF CREDIT CARD DEFAULT)

Problem Statement

The problem finding the patterns in customer behaviors that help predicting if a customer have a chance of getting his credit card defaulted before the 6 months period. By applying the right algorithm and parameters we can get a prediction on the 5^{th} month or even 4^{th} month to know if the customer have high chance of getting his credit card defaulted

Datasets and Inputs

The dataset that will be used for this project is the “Default of Credit Card Clients Dataset”. It contains information about 30,000 customers which include general information like Age, marital status, sex, education, etc. in addition to more relevant data, like amount of credit available, payment status of the past 6 months, amount of most recent payment, the target for our model will be to know if a customer credit card will be defaulted. This data was obtained from Donald Bren School of Information & Computer Sciences. link

Solution Statement

The Goal is to train a model with the capability to minimize the risk by making predictions based on the dataset. The model predicts whether a late-paying customer will only be late for a limited duration (1 or 2 months), or if this customer is predicted not to make a payment at all. This makes it a classification problem.

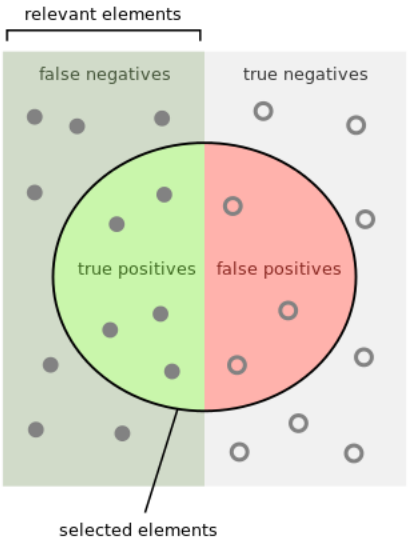
Benchmark Model

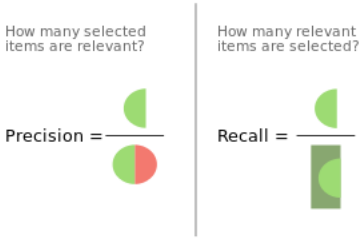
The model that been developed by (Mahyar M. Moghadam) Payment Default Prediction is what been taking as Benchmark model. With the current model we will be aiming at 80% accuracy

Evaluation Metrics

To validate the model and determine its performance, the actual information from the data will be compared with the predictions, and the Recall and precision will be analyzed. Those two are calculated by looking at four things:

- True positive : accounts got defaulted and model predict them as defaulted
- False positive : accounts did not defaulted and model predict them as defaulted
- True negative : accounts did not defaulted and model predict them as not defaulted
- False negative : accounts got defaulted and model predict them as not defaulted





$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Images Source : Wikipedia

Project Design

Data Preprocessing : Will check the data and prepare it by removing data input errors, missing values. Also checking for outliers and removing 10% - 25% or the leading and tailing data narrow the outliers

Data Processing : Will visualize the data and check the correlation and try to find pattern first visually. After that one of the supervised learning Algorithm will be use. It is yet uncertain whether supervised learning will be employed.

Training and Testing : will split the data and do 80% training and 20% for testing and to ensure the model is robust and we getting a result we can trust a cross validation will be applied.