Machine Learning Engineer Nanodegree

Capstone Project

Mohamad Tarbin
Feb 24th, 2018

I. Definition

Project Overview:

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 5th month or even 4th month to know if the customer have high chance of getting his credit card defaulted

Metrics

To validate and check the health of our project we need we use the accuracy score to check how well the model is doing and that's by checking the following : - 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 those will let us know how well is the model we develop is doing.

II. Analysis

Data Exploration

- 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. Data look healthy no messing record or error input has been found.

Exploratory Visualization

Mathplot library has been use in Data Visualization, it was necessary to understand the data and understand how each feature contribute to the Algorithm.. some feature was useful more than other, some features can be replace or combine under one feature because of some correlation specially the Pay and Bill amount features... for most of the part the the plotted diagram will present one feature's values count in the dataset when looking over all customer.. against looking at defaulting customers to see how the values of that feature contribute on the final result (customer credit card being default)

Algorithms and Techniques

The dataset target is customer's credit card being default or not so it was a supervised learning machine learning problem. and to find the right algorithm I had some candidate and decide to try they all to see how well they do with this dataset. the algorithms were (Random Forest , Gradient Boosting, Ada Boost, Decision Tree) and the results was really close between 81% and 82% and that what I was targeting for the accuracy so I decide to go with Ada Boost after reading this article (https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/) after that i start dropping the data the relevant for one months and check how the algorithm doing so create a loop to remove a month (6th month for example) then train the model and check the accuracy and the model did really good and the score was not impacted much til the 2nd month and the due the data in those column was correlated

Benchmark

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

III. Methodology

Data Preprocessing

after downloading the data from Kaggle I used Panada to load it into a data frame.. after that i did data where it does have record for Amount of the bill(BILL_AMT) and how much the was the customer payment on that bill (PAY_AMT) was following indexing from 1 but the payment status it start from 0 and jump to 2.. so i rename the column to 1 to match the patten... i think it was just a mistake while collecting the data maybe.

changing the column values in gender to be 1 or 0 instead of 1 or 2

Implementation

the library for that i use was Sklearn I like it because it cover most of the machine learning algorithm and it's well documented and simple to use and the most important.. powerful! with Sklearn it's simple as creating the Classifier Object then calling fit() method with the training data, call score() with the testing data and we get the score ... before that i did split the data for training , testing were i kept 25% of the data for testing and 75 to train the model.. and as i mention before after selecting the model i start dropping data to see how well the model doing if i training with 2 month worse of data and it was great ! Grid Search has been use to search for the best parameters that make the model predictions even better.

Refinement

I did set the The number of trees in the Ada boosting to be 15 tree and that has been done by setting the parameter n_estimators and that's cause a jump about 1-2% in the score. other than that the ADA Boosting take care of the rest

IV. Results

(approx. 2-3 pages)

Model Evaluation and Validation

The Model has been trained with 22'500 records and been test with 7500 records that the model did not seen before and the accuracy was 82%..for Proof on Concept i think this is really good. for production scenarios we can have a lot of improvement specially with bigger dataset we might even go with Neural Network but for the purpose of this project the result can be trusted

Justification

After all the model get a close result to the benchmark I did establish earlier we the accuracy also was around 82% and the model did the job or helping in predicting the customer that might have higher chance that their credit card getting defaulted from only 2 month worth of data

V. Conclusion

Free-Form Visualization

Feature Importances was implemented it let you see how the decision tree or the forest of the tree taking the features. which one is more important than the other so this help us understand what is more important to the model. here is the list of the features sorted by importance and the importance score:

1. feature 5 LIMIT_BAL (0.180000)
2. feature 10 SEX (0.160000)
3. feature 8 EDUCATION (0.140000)
4. feature 9 MARRIAGE (0.120000)
5. feature 7 AGE (0.100000)
6. feature 2 PAY_1 (0.080000)
7. feature 0 PAY_2 (0.080000)
8. feature 4 BILL_AMT1 (0.060000)
9. feature 3 BILL_AMT2 (0.040000)
10. feature 6 PAY_AMT1 (0.020000)
11. feature 1 PAY_AMT2 (0.020000)

Reflection and Improvement :

The model did did meet my expectation and did solve the problem and for sure.. there is always a room for improvement so more data will always be great to use.. but talking about the data we have creating a feature that represent the the average for the correlated feature can make the model faster to train and prediction..