Machine Learning Engineer Nanodegree

Capstone Proposal

Hyungoo Shim

July 22, 2018

Home Credit Default Risk

Domain Background

This project is be based the Kaggle competition Home Credit Default Risk (https://www.kaggle.com/c/home-credit-default-risk). It's competition's goal is that can you predict how capable each applicant is of repaying a loan? Many people struggled to get loans due to bad credit history. It can aggravate your financial life. There are lots people are in a dangerous financial situation with who is earning a minimum wage. Every time someone borrows money from the bank but If you have a non-credit history or bad credit report. what happens? There are two important effects of having no credit history.

1. It will be difficult to borrow money like mortgages, car loans, student loans, etc.
2. There is no way to gauge your creditworthiness. It is an important consideration to gauge if you are high-risk borrow or not.

Like this, If you don't have any credit record, unfortunately, you will suffer from financial problems. Home Credit is a non-financial institution, founded in 1997 in the Czech Republic. The company operates in 14 countries and focuses on lending primarily to people with no credit record and will become victims of untrustworthy lenders.
Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. However, One thing's for certain, when you don't have a credit report, you don't have any dept. In the other word, If you don't have any debt, then you do not have any credit history.
In order to make sure this underserved population, Home Credit company uses a variety of data like telco and transactional information to predict clients' repayment abilities.

Now this problem's solution is very important in our society. At least, It can help increase the client's potential stable financial situation and help achieve their dream of starting a family or sharing a home.

Problem Statement

The purpose of Home Credit Default Risk is to predict using historical loan application data. We have to know outcome that whether or not the customer will be able to repay a loan. It's definitely supervised and categorical problem like Email labeled as spam / not spam.
The labeled datas are included in the training data and our model is to learn to predict the label and This label is a zero or 1. Zero is will repay the loan on time and 1 is will have difficulty repaying the money.

Datasets and Inputs

For undestand our task, check the data structure. The datasets by Home Credit are 7 different sources of data. These Dataframes contain 24 attributes, 30000 instances, and more 220 columns across seven files.

Dataframe	Description
application_train / test	It is the main table. It contains the loan and loan application. Every loan has row and this row is identified by **SK_ID_CURR** what is connecting with the *bureau* data, *previous_application.*
bureau	Application data from previous loans that client got from other institution.
bureau_balance	Monthly balance of credits in Credit Bureau
previous_application	Application data of client's previous loans in Home Credit
POS_CASH_BALANCR	Monthly balance of client's previous loans in Home Credit
credit_card_balance	Monthly balance of client's previous credit card loans in Home Credit
installments_payment	Post payment data for each installments of previous credits in Home Credit related to loans

With application_csv datasets, The dependent variable is the label we want to predict. 0 for the loan was repaid on time and 1 indicating the client had payment difficulties. However, 1 is about 8% and 0 is approximately 90%. from this information, we know that it is a highly unbalanced pattern.

And the independent variable case, We can categorize these variables into six type dataset. First is feature of population statistics like gender, age, family_status, and education type. Second is a feature of Occupation and income. Third is assets like car, house, property. and fourth is contact information. Lastly, we can define Records of Credit Bureau and type of loans as independent variable.

Solution Statement

Above all, We need preprocessing step to handle such as missing or duplicate value in datasets. another option would be to drop columns with a high percentage of missing values. but it is a wrong decision to know that if these values will be helpful to our prediction. Thus we need to deal with categorical variables using one-hot-encoding or Label encoder for Machine learning model. After this we will aligning training and test data. There are some features in both the training set and test set. One-hot-encoding could create more columns in two datasets because there are some categorical variables with categories not represented in the testing data.

Next step is that we will treat correlation and outliers and looking for the correlation between the features and target then we can calculate the Pearson correlation coefficient between all variable and target. Another important step is Feature Engineering. It represents one of the patterns in machine learning and refers to adding new features from the existing data and choose only the most significant features or other methods of dimensionality reduction. Then We will perform a comprehensive comparison of different classification models. Some suitable supervised learning algorithms to predict the client who capable of repaying a loan are:

- Gradient Boosting,
- Adaboosting
- Random Forest
- Logistic Regression
- Decision tree

Finally, An optimal model would be chosen with the high scoring of ROC AUC’s probability.

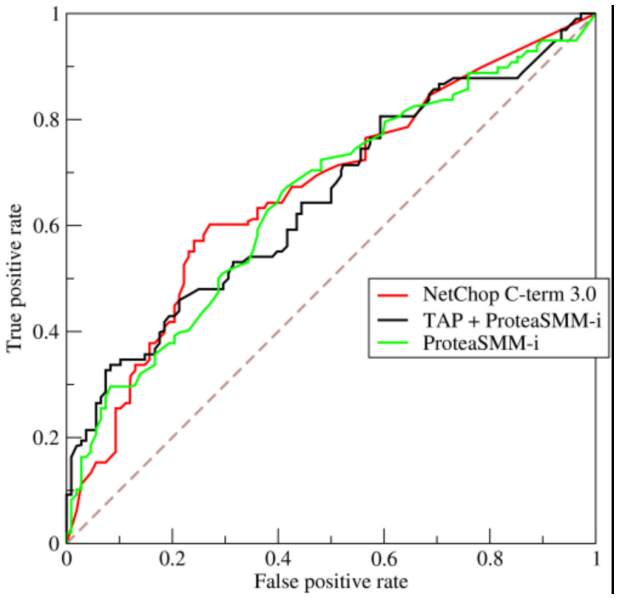
Benchmark Model

This competition evaluated on area under the ROC Area Under Curve [1] between the prediction probability and the observed target. This metrics is between 0 and 1 with a better model scoring higher. when we measure a classifier according to the ROC, we don't generation 0 or 1 predictions but rather a probability between 0 and 1, not accuracy because this dataset is unbalanced classes and we have to understand that Confusion Matrix when we make a binary prediction.

To get the Confusion matrix, we go over all the predictions made by the model and count how many times each of those 4 types of outcome occurs:

		Predicted class	
		Class 1	Class 0
Actual class	Class 1	True positive (TP)	False negative (FN)
	Class 2	False positive (FP)	True negative (TN)

The ROC curve is created by plotting the true positive rate (TPR)as known as sensitivity, recall in machine learning. Against the false positive rate (FPR) is also known as the fall-out with many different thresholds. Then plot them on a single graph, with the FPR values on the abscissa and the TPR values on the ordinate. The resulting curve is called ROC curve and the metric we consider is the AUC of this curve. It graphically showed below:



https://en.wikipedia.org/wiki/Receiver_operating_characteristic

We know the background of the data we are using and the metric to maximize. And as you know we are asked to predict the probability, so we are entirely unsure, we would guess 0.5 for all observation on the test set.
For the benchmark model, I'll use a slightly more sophisticated Logistic regression model for the actual baseline. Baseline model should have an Area Under the Precision Recall Curve (AUPRC) score of 75% or greater. Therefore, Logistic regression will be good choice to compare to other prediction models because it is generally produced high accuracy rates between 70 to 75%.

Evaluation Metrics

Given that this is binary classification problem, I will measure the AUPRC score. this curve computes two metrics from the Confusion matrix. To combine the FPR and the TPR into one

single metric, then we compute the two former metrics with threshold like 0.00, 0.01, 0.02 and so on for Logistic regression.

$$\text{TPR} = \frac{TP}{TP + FP} \qquad \text{FPR} = \frac{FP}{FP + TN}$$

- True positive rate (TPR), aka, sensitivity or recall. This metric correspond to the proportion of positive data. The higher TPR, the fewer positive data points we will miss.

- False positive rate, aka, fall-out. This metric correspond to the proportion of negative data. The higher FPR, the more negative data points will be misclassified.

The ROC curse is created by plotting the TPR against FPR at various threshold setting. As well as, we can describe that mathematically. Assume the following:

1. A is the distribution of scores the model produces for data points that are actually in the positive class.
2. B is the distribution of scores the model produces for data points that are actually in the negative class.
3. τ is the cutoff threshold. If a data point get's a score greater than this, it's predicted as belonging to the positive class.

So, we can show that TPR is given by: $P(A > \mathcal{T})$ and the FPR is given be: $P(B > \mathcal{T})$. We plot The TPR on the y-axis and FPR on the x-axis, draw the curve for various τ and calculate the area Under this curve. We get:

$$AUC = \int_0^1 TPR(x)dx = \int_0^1 P(A > \mathcal{T}(x))dx)$$

Where x is the FPR. One way to calculate this integral is to consider x as belonging to a uniform distribution. It simply becomes the expectation of the TPR.

$$AUC = E_x[P(A > \mathcal{T}(x))] \quad --- (1)$$

$$\text{x} = FPR = P(B > \mathcal{T})$$

x here was FPR, We considered x to be from a uniform distribution.

$$P(B > \mathcal{T}(x)) \sim U$$
$$\Rightarrow P(B < \mathcal{T}(x)) \sim (1 - U) \sim U$$
$$\Rightarrow F_B(\mathcal{T}(x)) \sim U \qquad --- (2)$$

We know from the inverse transform law that for any random variable X, if $F_X(Y) \sim U$ then $Y \sim X$. this follows since taking any random variable and applying its own CDF to it leads to the uniform. Using the fact in equation (2) gives us:

$$\mathscr{T}(x) \sim B$$

Finally, we can get AUC equation: $AUC = E_x(P(A > B)) = P(A > B)$
There are also other methods than ROC curves but are also related to the TPR and FPR like precision-recall, F1-Score, and Lorenz curves.

Project Design

step 1: Exploring the Data

We will explore each feature one at a time. Investigation of the dataset will determine how individuals fit into either group.

step 2: Preparing the Data

Before data can be used as input for machine learning algorithms, it often must be cleaned, formatted, and restructured. There are missing entries we must deal with, however, there are some qualities about certain features that must be adjusted. This preprocessing can help tremendously with the outcome nearly all learning algorithms.

step 3: Evaluating Model performance

we will investigate different supervised algorithms, and determine which is best at modeling the data. I will build the Logistic regression model as a base model and will train other more powerful ensemble methods.

step 4: Improving Results

We will choose from the different supervised learning models the best model to predict who will repay the loan on time. then we calculate the Area Under the ROC curve for the model over the entire training set by tuning at least one parameter to improve it.

step 5: Final model evaluation

After training and validation of different algorithms, the best performing model is finally selected and will optimize using hyperparameter tuning.

Citations

- Home Credit Group, http://www.homecredit.net/
- Home Credit, Financial report, http://www.homecredit.net/~/media/Files/H/Home-Credit-Group/documents/reports/2016/hcbv-ar-2015.pdf
- Some mathematical properties of the ROC curve and their applications (Università degli Studi di Napoli “Federico II” 2015)
- A gentle introduction for Home Credit Default Risk, http://medium.com/@williamkoehsen/
- Competition for Home Credit, https://medium.com/mighty-data-science-bootcamp/
- Function to disregard outliers when plotting, using matplotlib, https://stackoverflow.com/questions/11882393/matplotlib-disregard-outliers-when-plotting