Predicting credit card defaults

Artem Pashynskyi

October 9, 2020

Introduction

Background

- Client (retail bank) is trying to understand who from their existing credit card holders will default in the next period, it will allow the collection department to work in a more efficient way which will result in a lower amount of losses
- This project aims to predict whether the specific customer will default in the next period

Data acquisition and cleaning

Data source

- Data about the whole sample (30,000 clients) can be found in Kaggle dataset here. Overall data vas complete however it required quite a few changes and cleninng which will be described below.
- There are 25 variables:
  - **ID**: ID of each client
  - LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
  - SEX: Gender (1=male, 2=female)
  - EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
  - MARRIAGE: Marital status (1=married, 2=single, 3=others)
  - AGE: Age in years
  - PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
  - PAY_2: Repayment status in August, 2005 (scale same as above)
  - PAY_6: Repayment status in April, 2005 (scale same as above)
  - BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
  - BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
  - PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
  - PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
  - default.payment.next.month: Default payment (1=yes, 0=no)

Data acquisition and cleaning

Data cleaning

- First step was to analyze all categorical variables (Sex, Education, Marriage, Pay_0, Pay_2, Pay_3, Pay_4, Pay_5, Pay_6). Sex had 2 categories – logical; Education had 6 categories, based on data analysis and number of clients in category 5,4,6,0 were decided to combine into 1 category "Other"; Marriage had 4 categories – were decided to combine last 2 categories in "Others".
- One of the most important parameters is Pay, based on the dataset description it describe payment delay ("-1" means pay duly). Based on the data analysis categories "-2" and "0" are also in the dataset (as provided in the table on the right). It is not clear from the description what "-2" and "0" category means, thus, to make data more logical all clients in categories "-2", "-1", "0" are categorized as 0 – "pay duly".
- The next important category is BILL_AMT1 which provide information about amount of bill statement in a specific month. Based on the data analysis we identified c.2,500 clients with negative or 0 outstanding balance as of September, since bill amount should reflect outstanding credit balance it seems not logical to have negative BILL_AMT1 (c.2,000 customers) and in case BILL_AMT1 is 0 in September it is not possible to have default in the next period (because this month liability is zero), due to facts mentioned above I decided to drop customers with negative/0 outstanding balance as of September. Which resulted in train set reduction from 30,000 customers to 27,500 customers
- Next category is PAY_AMT, which describe payment in particular month. Based on my initial understanding, the following logic should be in the data: Bill_AMT_t = Bill_AMT_t+1 - PAY_AMT_t. However, based on further data analysis I realize that this equation is not holds in most cases, it seems that there are some additional elements for which information is not available, for example: expenses in a certain period, interest, etc.

Feature selection and Exploratory Data Analysis

Below is the list of key feature selection/modification:

- "normalization" of bill outstanding– to normalize bill outstanding values I have divided this columns by maximum credit limit column
- Client's willingness / discipline to pay the debt – I have estimated comparison between bill outstanding and what % of that bill was paid in the next period
- Categories based on age
- Categories based on sex and marital status

Predictive modelling

As a first step I have applied KNN, Decision Tree, Logistic regression , SVM models and I analyzed accuracy of these models:

Model	KNN	Decision Tree	Logistic regression	SVM
Accuracy	0.76	0.81	0.78	0.78

Which resulted in a quite high level of accuracy as for the first attempt, however, this is a wrong impression, the average default probability in the sample is 22% and based on the factor analysis the most critical element (with score of 0.72) is Pay_0, which means that our accuracy scores are high just because the average default is 22% (non-default - 78%), thus, I have tested accuracy F-1 score:

Model	KNN	Decision Tree	Logistic regression	SVM
F-1 score	0.72	0.79	0.68	0.68

Decision tree model provided the best results, thus, I decided to closely with this approach.

Based on further analysis of TP, FN, FP, TN I decided to use Receiver Operator Characteristic (ROC) curve and Area Under the Curve (AUC) as a primary classifier of the accuracy of the model. ROC_AUC score for initial Decision Tree was **0.66**

Predictive modelling

After 1) cleaning the data; 2) features creating 3) hyperparameters tuning I was able to get slightly better results of the decision tree:

Model	Initial Tree	Updated Tree
Accuracy	0.81	0.83
F-1 score	0.79	0.81
ROC_AUC	0.66	0.68

After analysis of different approaches I decide to try Light GBM model/approach. Light GBM is a gradient boosting framework that uses tree based learning algorithm. After analysis of Kaggle forums I used parameters for this model advise by another Kaggle participant. After applying Light GBM model I was able to achieve ROC_AUC score of 0.78:

Model	Tree-1	Tree-2	Light GBM
ROC_AUC score	0.66	0.68	0.78

Conclusions

In this study, I analyzed the relationship between default probability and client's payment information and social information. I identified the most important features and created features which maybe useful for this analysis. Initially I applied classification methods which were covered in IBM Machine Learning course, however, I was not able to improve accuracy of these methods by cleaning/preparing date and by creating new features, thus, I have used Light GBM model which provided quite a significant increase in ROC_AUC score.

Future steps

I was able to achieve ~12% improvement from the benchmark model in the regression problem, and ~78% ROC_AUC score in the classification problem. However, my understanding is that by analyzing further variables PAY_; BILL; PAY_Amount and creating some new features algorithm can be improved further.