Predicting credit card defaults

Artem Pashynskyi
October 9, 2020

1. Introduction

1.1 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

1.2 Problem

This project aims to predict whether the specific customer will default in the next period

1.3 Interest

Obviously, the main stakeholder is the bank, however, results of this work maybe used by other market players (banks, credit organizations)

2. Data acquisition and cleaning

2.1 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.

2.2 Dataset description

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_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)
- **PAY_5**: Repayment status in May, 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_AMT2**: Amount of bill statement in August, 2005 (NT dollar)

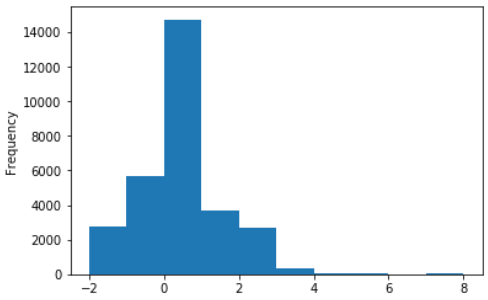
- **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
- **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
- **default.payment.next.month**: Default payment (1=yes, 0=no)

2.3 Data cleaning

As mentioned before, data was quite complete, however, further investigation resulted in a data anomalies and non-logical behavior.

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.

3. Feature selection and Exploratory Data Analysis

3.1 Bill outstanding normalization

After data cleaning we ended up with 27,402 samples and 25 features in the data. First of all it is clear that each client different amount of balance which depends on their income, thus, categories 'BILL_AMT1', 'BILL_AMT2','BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6' are not really helpful until we will normalize them. As the proxy for normalization I used 'LIMIT_BAL' (total limit available for each customer), my assumption that customers with higher limit will have higher payments and higher bills. Thus, I divided categories 'BILL_AMT1', 'BILL_AMT2','BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6' by column 'LIMIT_BAL' to get "normalized" payments and balances. As a result, of this step I have dropped 'BILL_AMT1', 'BILL_AMT2','BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6' and instead of that added normalized columns

3.2 Ability/willingness to pay (financial discipline)

In the next step I tried to determine Client's ability/willingness to pay debt, for this I have estimated ration between BILL_AMT_T+1 and PAY_AMT_T (comparison between bill outstanding and what % of that bill was paid in the next period). For example, bill in Arp = USD100 in May client is paying USD90, Client paid 90% of the required amount. So, I have estimated this for each month and then estimated average payment, as a result, of this step I have removed 'PAY_AMT1','PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6', instead of this I have added BILL_AMT_T+1/PAY_AMT_T for each month and then average for all 5monhts 'PAY_AMT_Average_N'.

3.3 Age categories

Next step is to check whether default (ability to meet obligations) are depends on age. Initially I have categorized age by the following categories: 20-30; 30-40; 40-50; 50-60; 60-70. Categories 40-50; 50-60; 60-70 had much less amount of people and they had comparable default probabilities, thus, I have combined Clients from this age bin and then made pore detailed breakdown for 20-40 years:

	Non-Default	Default	Probability
20-25	1861	708	28%
25-30	5026	1343	21%
30-35	4445	1042	19%
35-40	3679	971	21%
40-45	2730	782	22%
45-50	1828	564	24%
>50	1840	583	24%

As presented in the table above, the highest probability of default is in category 20-25 (28%) then it is going down and rise starting from 35-40 years. Not sure what is the explanation here, but my assumption is that people from category 20-25 have lower income / lower financial discipline. So instead of AGE category we will use the following age bins.

3.4 Sex and marital status

Probability of default based on the sex:

	Non-Default	Default	Probability
Male	8398	2636	24%
Female	13011	3357	21%

Probability of default based on the marital status:

	Non-Default	Default	Probability
Married	9527	2879	23%
Not married	11612	3030	21%
Other	270	84	24%

Based on the tables above, it seems that male clients have higher probability of default and married clients have higher probability of default. Therefore I decided to combine these categories into one column which may be beneficial for our analysis:

	Non-Default	Default	Probability
married man	3564	1218	25%
single man	4733	1380	23%
"other" man	101	38	27%
married woman	5963	1661	22%
single woman	6879	1650	19%
"other" woman	169	46	21%

Based on the table above it seems that single woman have the lowest probability to default

4. Predictive modelling

This is classification problem, we should determine whether Client will default (default = 1) or will not default (default = 0) in the next month based on the data provided for each client.

4.1 Initial 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**

4.2 Further 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

As you can see, margin effect of 2 p.p. were achieved.

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:

```
'boosting_type': 'gbdt',
'objective': 'binary',
'metric':'auc',
'learning_rate': 0.05,
'num_leaves': 7,  # we should let it be smaller than 2^(max_depth)
'max_depth': 4,  # -1 means no limit
'min_child_samples': 100,  # Minimum number of data need in a child(min_data_in_leaf)
'max_bin': 100,  # Number of bucketed bin for feature values
'subsample': 0.9,  # Subsample ratio of the training instance.
'subsample_freq': 1,  # frequence of subsample, <=0 means no enable
'colsample_bytree': 0.7,  # Subsample ratio of columns when constructing each tree.
'min_child_weight': 0,  # Minimum sum of instance weight(hessian) needed in a child(leaf)
'min_split_gain': 0,  # lambda_l1, lambda_l2 and min_gain_to_split to regularization
'nthread': 8,
'verbose': 0,
'scale_pos_weight':50, # because training data is sightly unbalanced
```

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

5. 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.

6. 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.