Estimation of default probability using bank account data (Tinkoff challenge)

Ilya Melnikov

1 Introduction

For this task we decided to use data from Tinkoff machine learning challenge 2013. As they suggested, we try to predict the probability of a loan payment being overdue within the first year of loan given information on customers' credit bank accounts. Such tasks appear when banks want to decide whether they should give a loan to a customer, or they should not. We use logistic regression for this purpose, because it is better suited than linear regression and simple enough to implement.

A more thorough problem statement and details of the original challenge can be found in the next section. A brief overview of data provided is in section 3. Sections 4 and 5 are entirely built on our scripts "data_filtering.R" and "analysis.R" respectively and are aimed to give interpretation to what are we doing in these scripts. Section 6 concludes.

2 Problem statement

Every credit institution faces certain risks when giving loans to people or firms. No bank wants to lose a trustworthy loanee or give loans to untrustworthy ones. In order to minimize losses due to such risks, banks put their clients through careful vetting procedures before approving a loan. However, manual checks may be costly in terms of time and human resources, but a total lack of such checks poses a threat to bank's stability. Thus, creating a reliable algorithm that can reveal the type ("good" or "bad") of a loanee may substantially increase expected gains of a credit institution while reducing risks.

This task was originally included in Tinkoff machine learning tournament in 2013 (available here). We aim to predict the probability of default (credit payment being overdue for more than 90 days during the first year, as stated in the task) using data on bank accounts provided by the challenge authors. Ideally, one should also take into account economic information (for example, currency exchange rates when estimating risks of loans taken in foreign currency) and information about individual performance of entities (data on salary and employment history for people and financial statements for firms), but we chose to stick to the provided dataset since it has abundant information about loanee's credit history, but doesn't make analysis overly complex. A more elaborate discussion of restrictions of such approach can be found in the next section.

3 Data understanding

Details regarding particular variables (brief description and types) are included in the attached archive. This section justifies our choice of variables in the model. Challenge authors provided 2 datasets: one with 50000 customer IDs and respective "default" dummies (separated into test and train subsets, the former being only meaningful for the challenge) and the other with detailed information regarding credit history of borrowers (a total 280942 entries). The first dataset is used solely as a source for the dependent variable, so we can focus on the second one.

The correspondence between IDs (and the dependent variable) and account data is not 1-to-1, because for every entity there can be several account histories. There are a few ways to handle this: either use some statistic to use as a proxy for all account data of a particular customer (for example, mean for numeric values and mode for categorical ones), assign respective values of dependent variables to every ID's account entry, or analyse data as is extracting information from "sequences" of accounts throughout time. Obviously, the last approach is the most preferable, but inadequately complex. The first one could be justified in simpler cases, but is not easily implemented here. Thus, we decided to stick to the second approach.

Some variables can be excluded right away because they are either within the chosen approach or for our purpose in general. Examples of these variables are: *bureau_cd* (the bureau that provided account data), *bki_request_date* (date when the bureau request was made), *open_date* (the data of account creation, useless if we don't consider "account sequences"), etc. A summary of variables we use may be found in the next section.

4 Data preparation

First of all, the part of the sample reserved for the challenge submission was excluded. Some variables were deleted right away. Except for the variables mentioned above, we also deleted *final_pmt_date*, *inf_confirm_date*, *fact_close_date* (same reason as *open_date*), *pmt_string_84m* (it is included in *ttl_delq_xx* dummies), *status* (it is unclear how to include this in the model), *outstanding* and *next_pmt* (not informative without the full history of payments).

After that, we kept only entries with *"relationship==1"*, because all codes but 1 and 9 are reserved for indirect payees, and firms (code 9) require a separate model, but lack in numbers (only 205 entries including train sample). Thus, this column contains only 1 value and thus was deleted after sampling.

Most of variables didn't have NAs, but some had them in abundance, e.g., *max_delq_balance*. This variable has 0's, other numeric values and NAs. Since we cannot interpret NA in this case (those who had no allowed debt have 0), we deleted this variable. Since *current_delq*, *curr_balance_amt* and *delq_balance* have similar problems, we excluded them as well.

As the final step, we filtered out values of *pmt_freq* that have either too few occurrences (less than 10 as a rule of thumb) or code 7 (we cannot say anything about them). We also deleted values of *type* with codes 4, 14, and 99 (the latter is not informative, the first 2 have less than 10 occurrences).

We transformed 3 variables. According to the intuition we used, for people who have to pay every week it is less probable to allow for a debt to exist for 90+ days. Thus, the higher the frequency of payment, the lesser the probability of default. We assigned numeric values to *pmt_freq* by dividing a calendar year by the base period of payments, which yields the supposed number of payments per year. Surprisingly, there were 21504 observations with code 0, which is not included in variable description. However arguable this step is, we decided to exclude those (along with 145 empty entries). Probably it would have been better to assume monthly payments (mode of this variable), but thus we can distort the results. Some other levels were excluded as too rare (<10 entries).

We also transformed factor variables *currency* and *type*. We assigned labels for types according to its description and expanded these 2 variables as dummies. The final dataframe was merged by tcs_customer_id (excluded during merging) from 2 datasets provided. It can be obtained by using the data_filtering script from data available online.

5 Model building and evaluation

For this model we must use models with binary dependent variables. The most obvious choice is to use logistic regression. As the first step, we tried to include all variables from the final dataset. Here is the summary of the output regression:

```
Call:
glm(formula = bad ~ ., family = binomial(link = "logit"), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9498  -0.4615  -0.4511  -0.4229   3.7780

Coefficients: (2 not defined because of singularities)
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.370e+00  8.045e-01  -4.189 2.80e-05 ***
credit_limit     -8.400e-07  9.715e-08  -8.646  < 2e-16 ***
ttl_delq_5       -1.023e+01  3.246e+02  -0.032 0.974865
ttl_delq_5_29    -1.017e+01  3.246e+02  -0.031 0.975014
ttl_delq_30       1.017e+01  3.246e+02   0.031 0.974994
```

```
ttl_delq_30_59    -8.730e-03  1.399e-02  -0.624 0.532638
ttl_delq_60_89     5.883e-02  2.182e-02   2.696 0.007025 **
ttl_delq_90_plus  -3.461e-03  3.154e-03  -1.097 0.272567
interest_rate     -4.660e-03  1.640e-03  -2.842 0.004482 **
pmt_freq           1.080e-02  2.599e-03   4.155 3.25e-05 ***
typeauto          -5.714e-03  7.500e-01  -0.008 0.993921
typemort           3.213e-01  7.625e-01   0.421 0.673475
typecard          -1.616e-01  7.463e-01  -0.216 0.828602
typecons           1.128e-02  7.460e-01   0.015 0.987935
typebus           -5.493e-01  8.412e-01  -0.653 0.513751
typeca                    NA         NA      NA       NA
currencyEUR       -8.841e+00  5.372e+01  -0.165 0.869284
currencyRUB        1.054e+00  2.957e-01   3.564 0.000365 ***
currencyUSD               NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 84507  on 135274  degrees of freedom
Residual deviance: 84237  on 135258  degrees of freedom
AIC: 84271

Number of Fisher Scoring iterations: 11
```

We can see a few interesting results. First of all, there are 2 coefficients that seem to be NA. This happened because we did not choose and exclude the "base" dummies when trying to expand *type* and *currency* factors. Thus, one dummy per group becomes 100% dependent on other dummies within its group. We might exclude *currencyRUB* dummy from the regression, because the vast majority of loans is taken in this currency in the sample. As for the *type* variable group, it seems irrelevant to our model.

Another interesting fact is that payment frequency positively affects the probability of default. We might assume a non-linear relation between these variables and include variable *pmt_freq_sq*, that is pmt_freq squared. For obvious reasons, *ttl_delq_60_89* is significant. But we can see that variables within this group are strongly correlated if the periods are close. We might use only 60-89 and 30 days since they are weakly correlated (ρ=0.15) and can tell us something about the client's discipline.

Coefficients for variables *interest_rate* and *credit_limit* are also significant. We assume that *credit_limit* may be negatively correlated with the probability of default due to the fact that this variable may be a proxy for individual characteristics or other unobservables of the customer (e.g., discipline or employability), since the better his or her reputation is the larger loan will be approved. It is not as simple to find a good interpretation for the *interest_rate* coefficient. We speculate that this may be due to the fact that the loanee does not want to pay larger interest on missed payments. However we believe that this interpretation is quite weak.

From the results above we decided to run another model. We decided to include *typecard* and *typemort* here just in case. Summary:

```
Call:
glm(formula = bad ~ credit_limit + ttl_delq_30 + ttl_delq_60_89 +
    interest_rate + typecard + typemort + pmt_freq + pmt_freq_sq +
    currencyEUR + currencyUSD, family = binomial(link = "logit"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9883  -0.4612  -0.4501  -0.4236   3.7847

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.731e+00  1.038e-01 -16.669  < 2e-16 ***
credit_limit   -8.558e-07  9.036e-08  -9.470  < 2e-16 ***
ttl_delq_30    -9.923e-03  4.372e-03  -2.270 0.023218 *
ttl_delq_60_89  5.130e-02  1.762e-02   2.911 0.003600 **
interest_rate  -4.557e-03  1.636e-03  -2.785 0.005348 **
typecard       -1.680e-01  2.645e-02  -6.351 2.13e-10 ***
typemort        3.217e-01  1.580e-01   2.036 0.041705 *
pmt_freq       -5.074e-02  1.069e-02  -4.747 2.07e-06 ***
pmt_freq_sq     1.161e-03  1.905e-04   6.092 1.11e-09 ***
currencyEUR    -9.885e+00  5.364e+01  -0.184 0.853787
currencyUSD    -1.014e+00  2.950e-01  -3.438 0.000586 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 84507  on 135274  degrees of freedom
Residual deviance: 84222  on 135264  degrees of freedom
AIC: 84244
```

```
Number of Fisher Scoring iterations: 11
```

We can see that this model is better by looking at AIC. The second model has better results, but we can still cut out *currencyEUR* since it is insignificant.

```
Call:
glm(formula = bad ~ credit_limit + ttl_delq_30 + ttl_delq_60_89 +
    interest_rate + typecard + typemort + pmt_freq + pmt_freq_sq +
    currencyUSD, family = binomial(link = "logit"), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9889  -0.4612  -0.4501  -0.4234   3.7870

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.730e+00  1.038e-01 -16.667  < 2e-16 ***
credit_limit   -8.580e-07  9.036e-08  -9.496  < 2e-16 ***
ttl_delq_30    -9.935e-03  4.372e-03  -2.273 0.023047 *
ttl_delq_60_89  5.136e-02  1.762e-02   2.915 0.003560 **
interest_rate  -4.555e-03  1.636e-03  -2.784 0.005368 **
typecard       -1.689e-01  2.645e-02  -6.386 1.71e-10 ***
typemort        3.233e-01  1.580e-01   2.046 0.040732 *
pmt_freq       -5.076e-02  1.069e-02  -4.749 2.05e-06 ***
pmt_freq_sq     1.161e-03  1.905e-04   6.094 1.10e-09 ***
currencyUSD    -1.013e+00  2.950e-01  -3.435 0.000593 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 84507  on 135274  degrees of freedom
Residual deviance: 84227  on 135265  degrees of freedom
AIC: 84247

Number of Fisher Scoring iterations: 6
```

While this model is marginally worse than the second one, all its coefficients are significant, so we will not change it. Positive and negative correlation between the dependent variable and *typemort* and *typecard* respectively may be explained as the effect of payment sizes for each of these types (assuming that larger payment sizes increase the probability of default). The negative effect of *currencyUSD* for a dataset from 2013 is adequate and may be either due to lower rates on such loans (it is not confirmed by correlation) or due to higher income of loanees. As for the effect of *pmt_freq*, we can see that it is a parabola with minimum at 26 payments per months and maximum (in our sample) in 52 payments (the effect here is actually positive), which means that weekly payments lead to higher risks.

6 Discussion

This model can be used in order to determine whether a certain loanee is trustworthy. Some results must be re-evaluated (the effect of currency, e.g.) to fit in today's realia, it is likely that the model can be improved by certain data (loanee's salary).

As it was already mentioned, there might be better approaches to fit these data like considering sequences of account history. Moreover, this model works only for people, but not firms. However, we believe that our approach can be a used to determine a rough estimate of client's credibility.