Customer behavior analysis

The main goals

The quality of service, the quality of products and, as a result, a satisfied customer are the key blocks in most business strategies. The business of lending is always a trade-off between risks and clients satisfaction. To be successful a company should balance these factors at the point that produces the highest value for the company and meets the customers' needs.

Certainly, clients are the most valuable asset for every business, however attracting new clients takes time and money and therefore it is so important to increase the number of returning clients. To achieve that, it is essential to understand current customer behavior (offer acceptance and reapplications rates) and determine key factors that influence credit acceptance.

Summary

The goal of the research is not only to study the ways to predict customer behavior, but also to understand what credit process is applied in the lender-company and how it can influence the 'future' or 'no future' relations with the clients.

What can cause clients' disappointment?

1) Slow service (decision) process. No answer from the company after application can cause 'new applications' (reapplication) before the answer is received.
2) Other loan conditions than those from application (term and/or amount reduction). There is a substantial demand for loans with terms over 6 months and amount up to 1 000.
3) Refusal to offer a loan for clients with "good history" (no data available - no declined applications and no offers with zero amount in dataset).
4) Loan repayment issues (no data available).

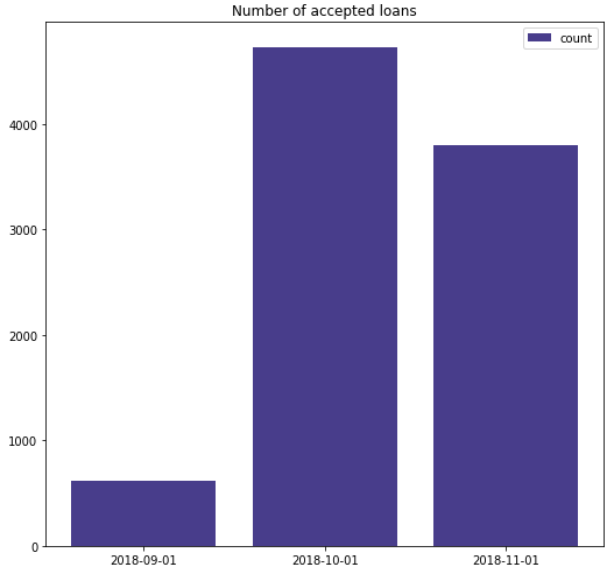
The obvious fact is that a strong negative correlation is observed between 'reapplied' and 'accepted_offer'. This represents the fact that most of the clients who disagree with the offer apply again. Correlation between 'accepted_offer' and 'reapplied' is -0.792.

Dataset

Dataset consists of 17 columns (including client's ID, and application ID) and 14 347 rows with the application information during the 3 months period from 2018-09-25 to 2018-11-21.

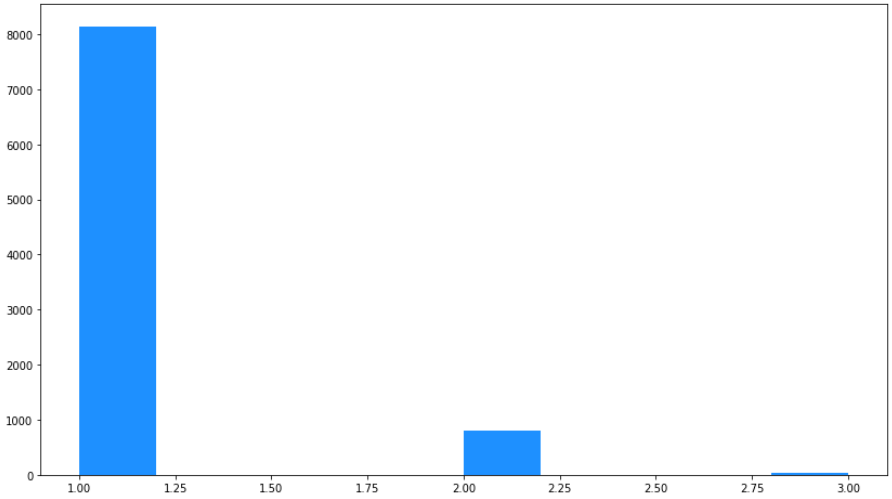
Exploratory Data Analysis

From September till November 2018 there was a huge rise in a number and total amount of loans issued. In September there were only 620 accepted offers with total amount of 177 290. In October there were impressive 4729 accepted offers with total amount of 1 184 790 (change 568.28%). In November there was a slight decrease in number of accepted offers (3802) with total amount resulted in 939 000 (change -20.75%).

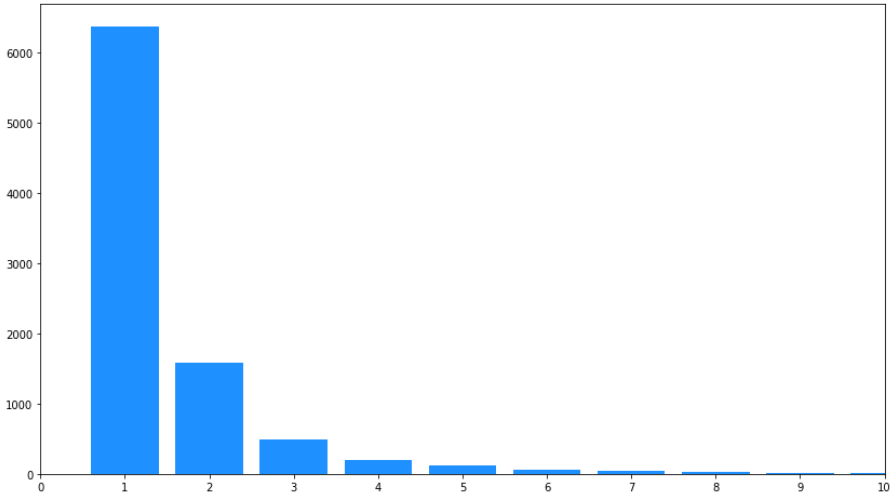


Clients

There are 8147 applicants out of 8981, whose applications are under only one scorecard lending program, 796 clients - under 2 and only 38 under all three programs. That shows that lender-company can take actions to offer other programs for clients.



Most of the clients have only one application.



Number of clients that have 1, 2 or 3 applications is 6380, 1586 and 495 respectively, which gives in total 94.21% of all clients. Out of the clients with only one application 6120 people have accepted the offer.

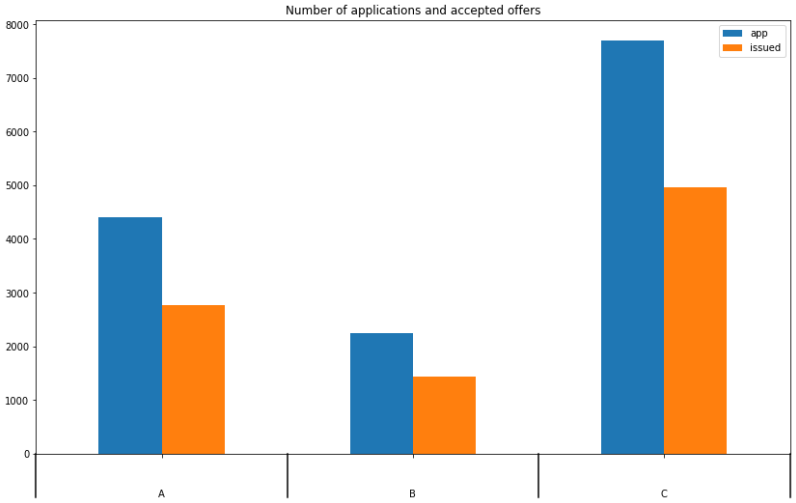
There are several clients that have outsize number of applications. The top 10 clients with the highest number of applications are:

customer_id	num_prev_applications
2002872	43
644473	40
521446	38
2010387	36
1230416	33
1548835	28
2003242	26
1484937	26
1714359	22
857505	19

There are 964 reapplications after accepting the offer.

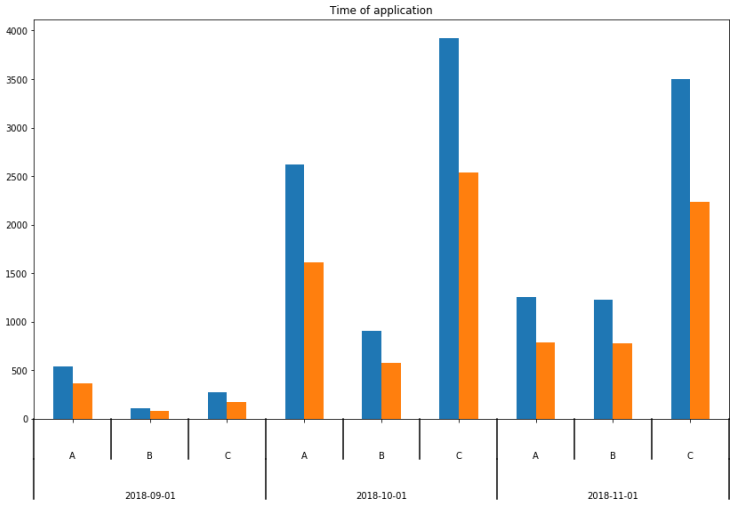
Scorecards

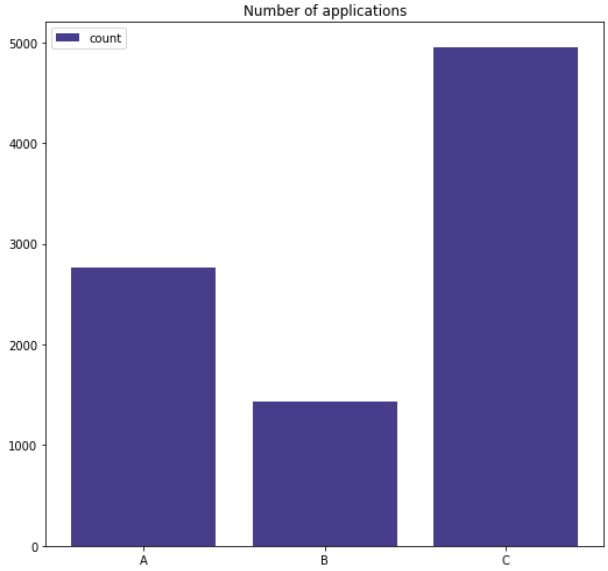
Clients can be scored under 3 loan programs (scorecard A, B and C).

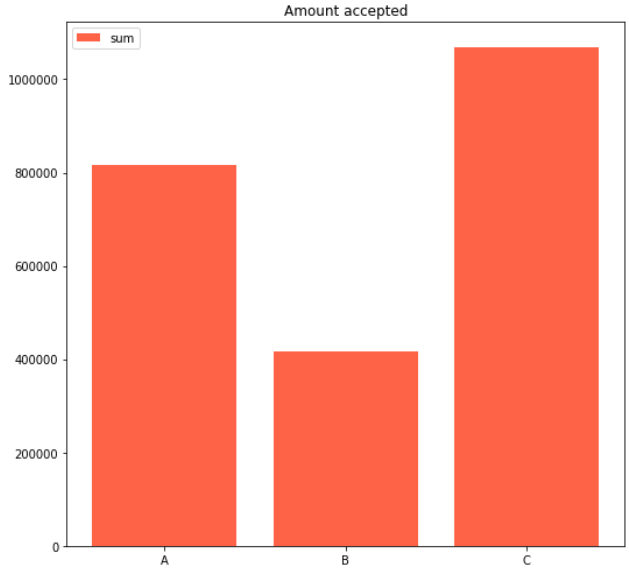


The most popular program is ‘Scorecard C’ and, as it is shown at the figure above, the greatest amount of applications has been received under this program.
Total amount of loans issued is 2 301 080 over the period from 2018-09-25 till 2018-11-21 . Most loans are issued under scorecard C program, however the mean amount of a loan is greater in scorecard A.
Although the number of applications and loans accepted by clients is different, the ratio between the offers accepted and the number of applications is almost the same (around 63-64%).

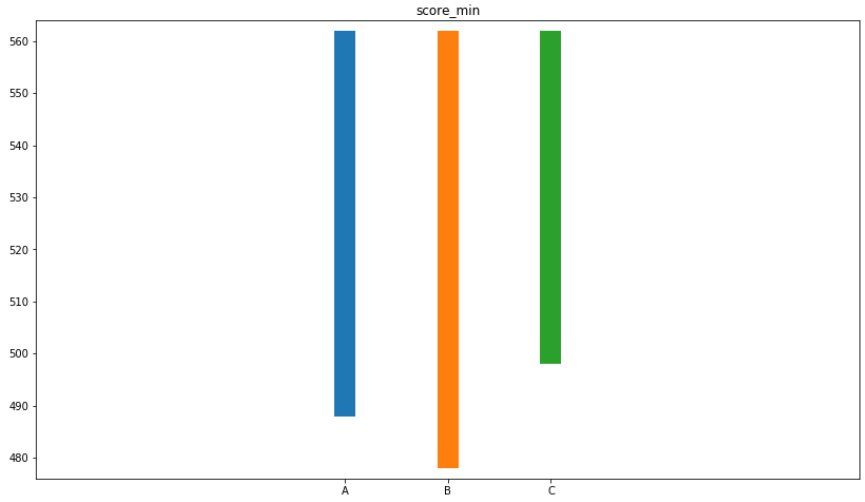
scorecard	Loans accepted	Applications	Percentage
A	2767	4411	62.73
B	1430	2243	63.75
C	4954	7693	64.4



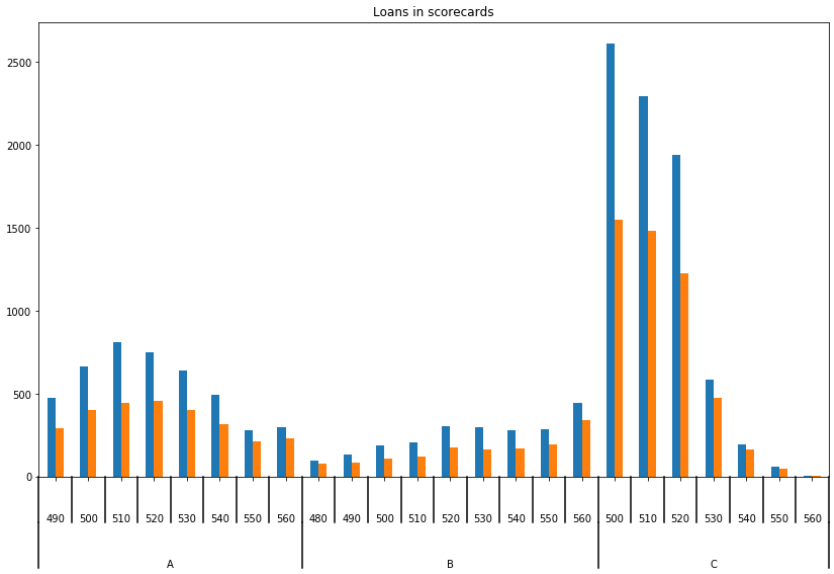




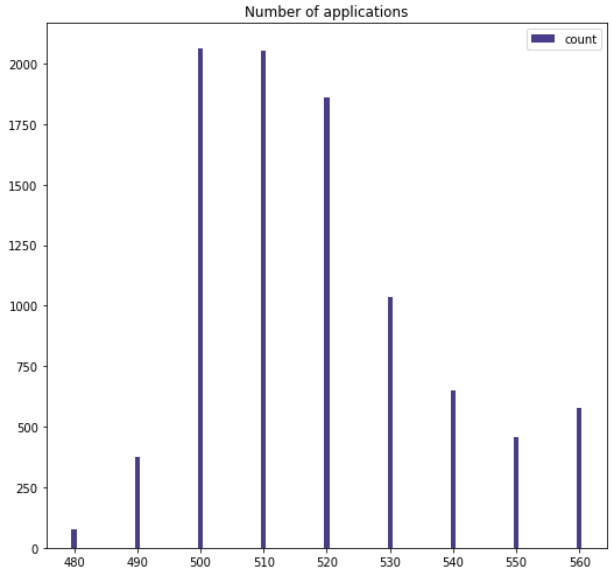
According to the data, the ‘scorecard B’ has the widest range of min_scores but the lowest amount and number of loans, whereas the ‘scorecard C’ has the narrowest range of scores but the greatest amount of loans issued (accepted offers).
For the purpose of risk diversification the ‘scorecard B’ program may need to be modified to attract more clients.

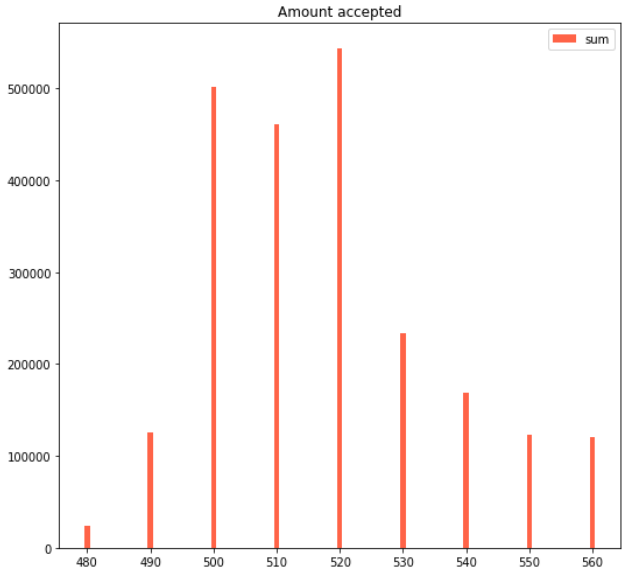


Most of the loans in ‘scorecard C’ tend to be issued to clients with lower scores, whereas ‘scorecard B’ has the greatest percentage of loans issued to clients with high scores.

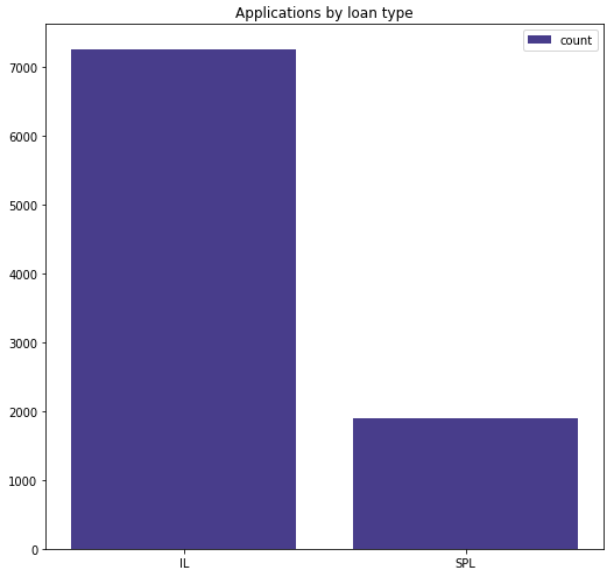


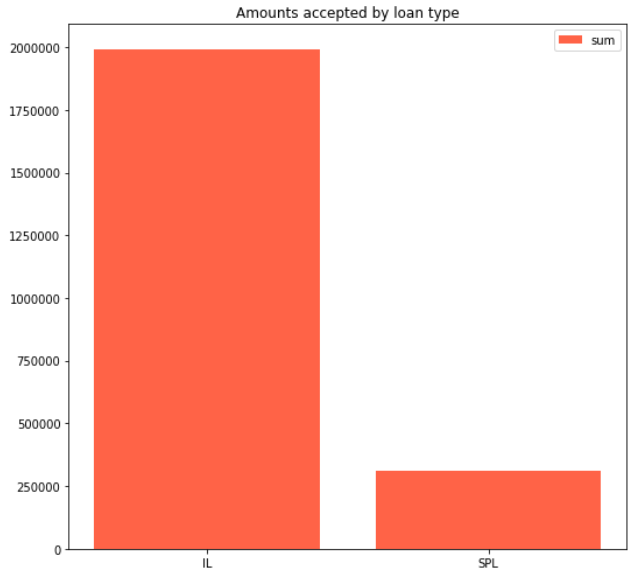
The study of clients’ distribution by score_min shows that most clients have the minimum score in the range between 500 and 520.



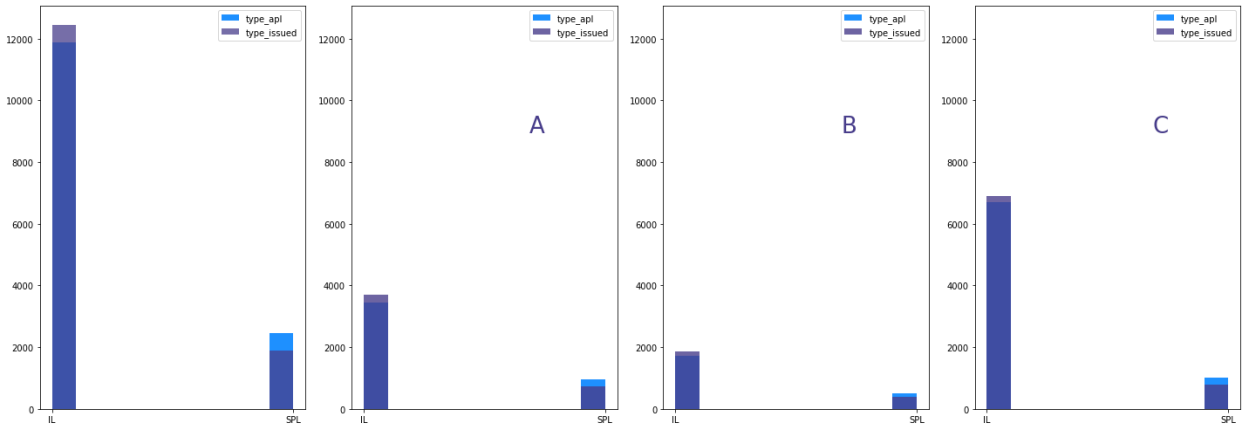


Major number and amount of loans are Installment Loans (type ‘IL’).



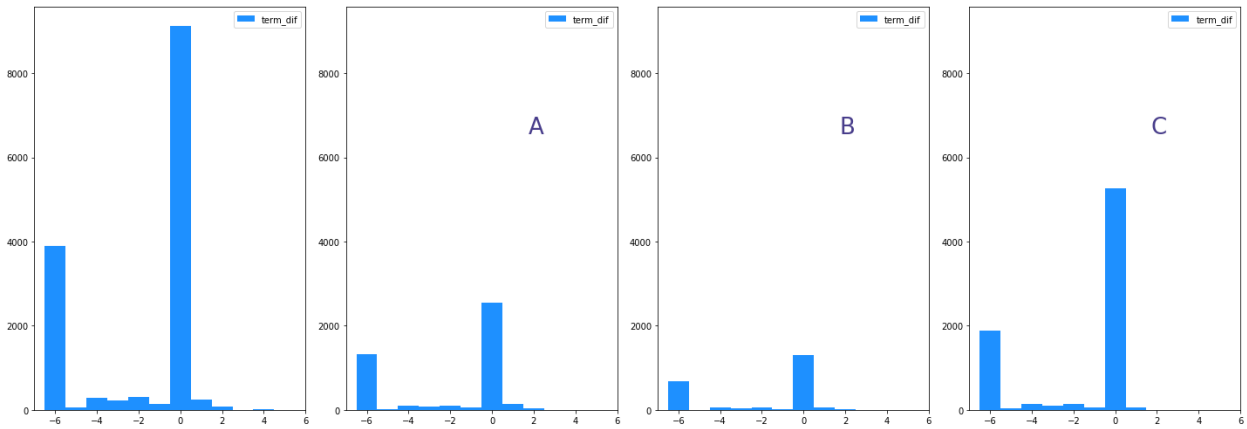


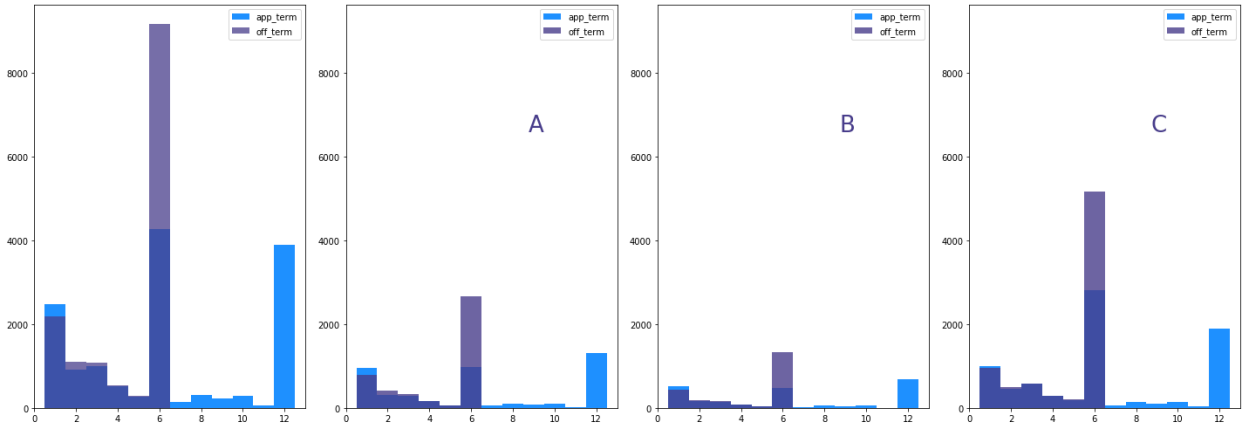
The company prefers to issue Installment Loans (‘IL’) rather than Sub Prime Loans (‘SPL’). This assumption comes from the fact that offered_loan_type was changes to ‘IL’ type despite the fact that ‘applied_loan_type’ ‘SPL’ had been chosen by client (loan type was upgraded for 282 applications).



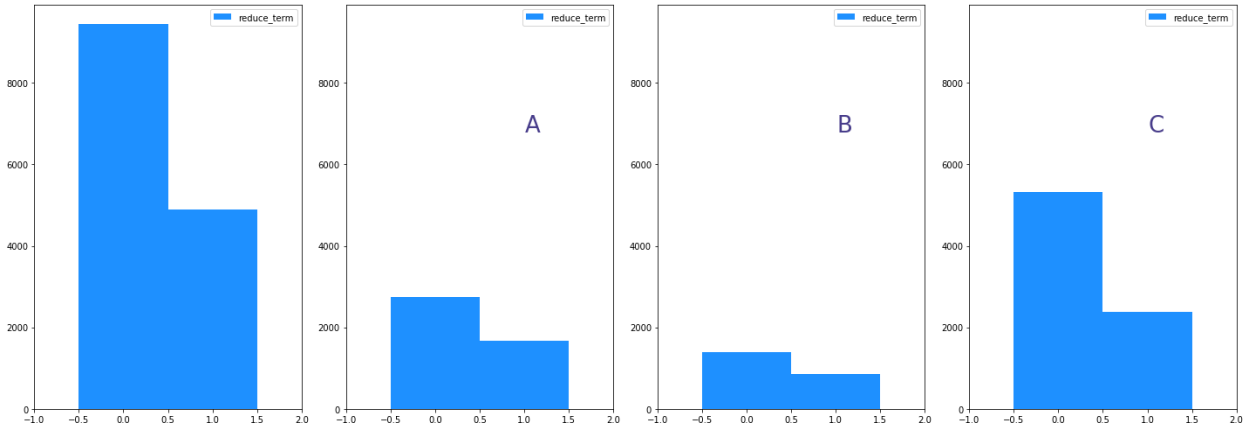
Terms

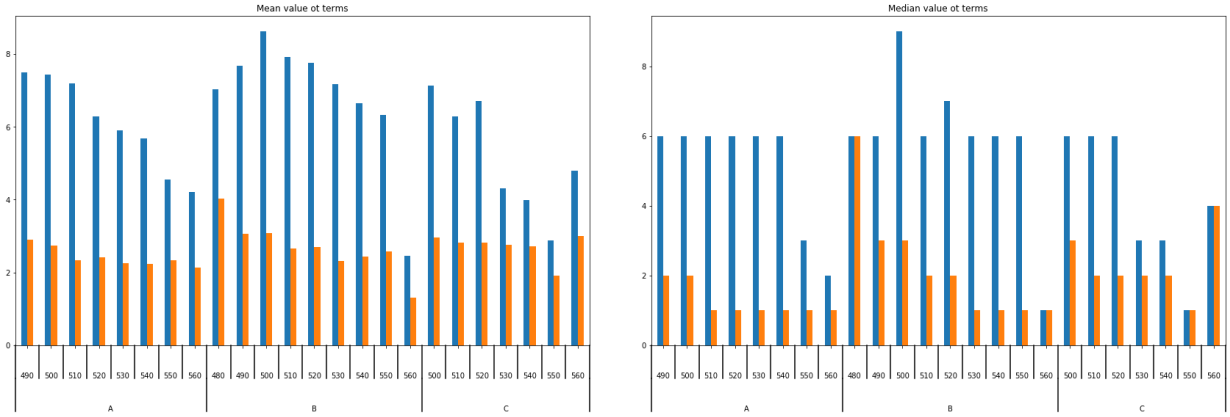
According to the data, a strong demand for longer loans (‘7-12’ group) can be determined, the ratio of applications for longer terms is 34.15% of all applications, but the company didn’t offer loans with offered term that exceeded 6 months though this option is offered to clients. In general, people are willing to have longer term to cut monthly payments. For some clients the loan term was increased up to 6 months in total, but all the loans with terms over 6 months were cut down to the duration ≤6 months. This term reduction can cause clients’ disappointment in cases when one applies for a year-long loan and gets only a half of that.



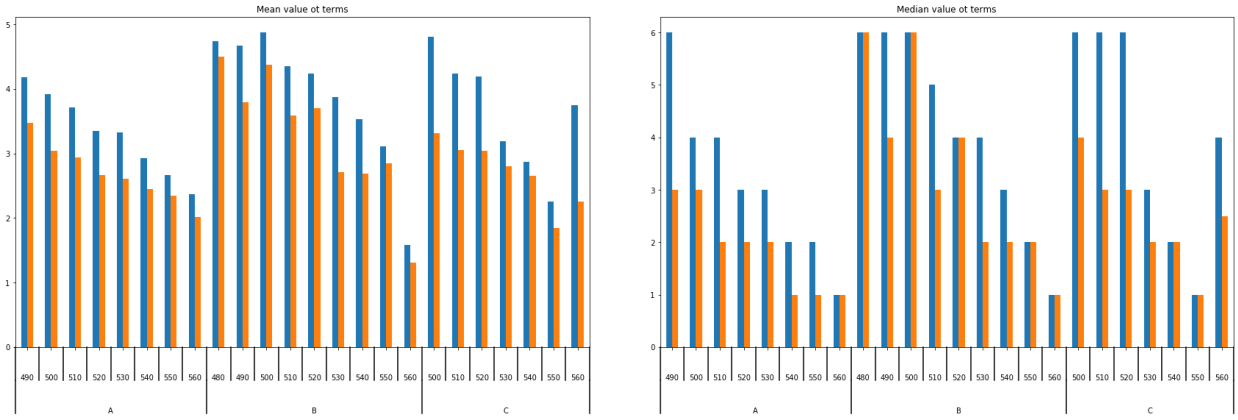


Almost one third of applications get reduced loan term (group 1 at the figure below represents applications that received an offer with reduced offered term).



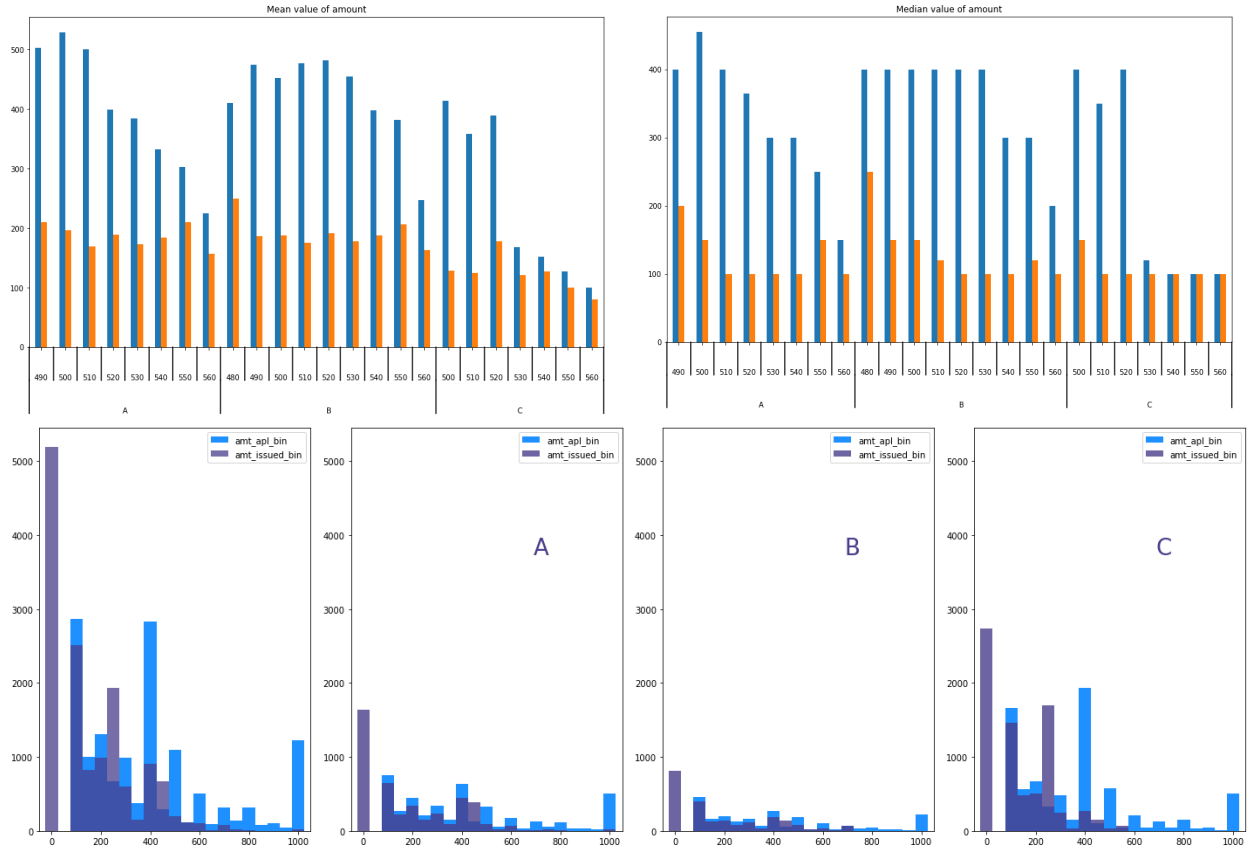


When we exclude applications with desired term over 6 months then we see a better match.



Amounts

According to the data, the amount offered is never increased compared to application, but mostly is decreased. Also there are no declined applications in the data (offers with zero amount).



The major amount issued falls into the range from 100 to 400. A strong demand for loans with amount of 400 can be observed, but the offered amount for these loans was downgraded to 200 and less.

Target Modeling (Binary Classification)

Environment

- Anaconda Spyder (Python 3.6)

Target variable:

1. ‘accepted_offer’ (binary classification)
2. ‘reapplied’ (binary classification)

In order to build dataset for these classifications, different ways to design features were applied:

1. Raw features
2. Raw features grouped to bins according to value
3. Features based on the difference between application and offer
4. Features based on difference between current and previous applications (if available)
5. Counters (number of applications per client, number of previous accepted offers etc.)
6. Latest previous application data (history)

Based on feature type several approaches were applied:

- Categorical:
  - 2 values: set 1 for major class, 0 for minor
  - More than 2: One Hot Encoding with dropping less frequent class
- Numerical:
  - Raw features
  - Log-transformed features (to reduce the influence of large numbers)

Training and Validation design was the same for both targets:

1. Trainset/Testset split 70/30 Stratified by target
2. Train models on balanced 50/50 subsamples of Trainset
3. Validation:
   - StratifiedKFold (10 folds with same balance of classes)
   - EarlyStopping (to prevent overfitting)
   - Out of sample validation Testset

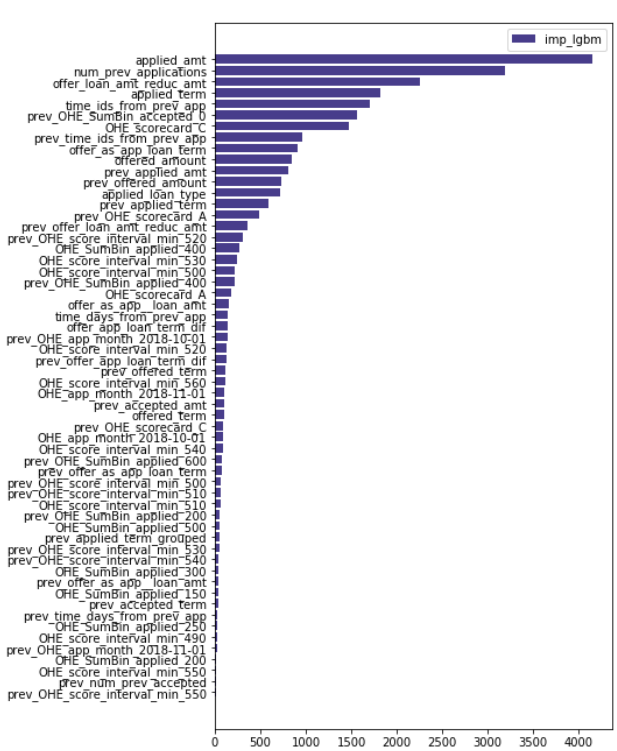
4. For the purpose of classification two classifiers were used: XGBClassifier and LGBMClassifier.
5. Metric for training “ROC AUC”.
6. Several parameters of classifiers were chosen by applying the random grid search.
7. To eliminate bias in classification the ensemble of models was designed based on different random seeds (4 seeds, 10 folds, 2 classifiers) and training subsamples.
8. Quality assessment metrics for binary Classification:
    - ROC AUC
    - Accuracy (exact match)
    - Precision (Positive predictive value - predicted class 1 compared to data)
    - Recall (True positive rate – predicted classes in class 1)
    - F1 score

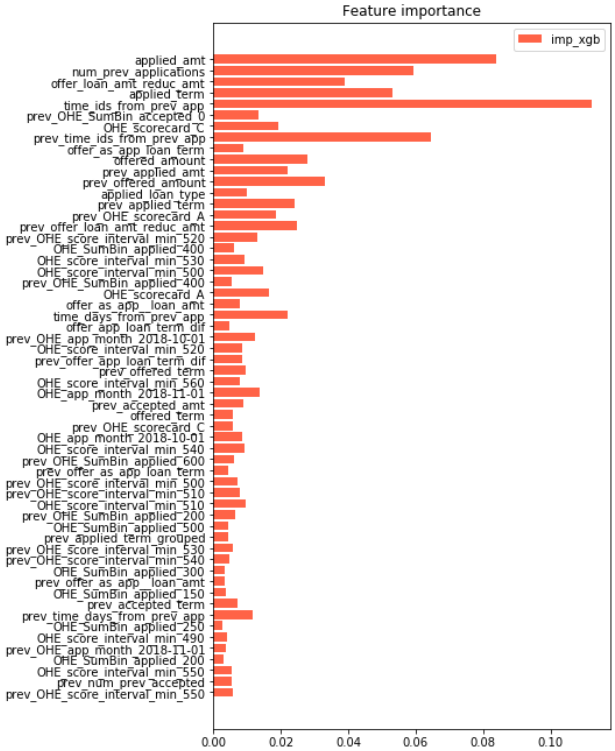
    The quality of prediction could be improved by adding information about the exact time of application, time of offer and current loan balance.
    The other significant information missing is the repayment schedule and client’s behavior during that period.

Goal 1: predict “accepted_offer”

“Reapplied” variable has high negative correlation to the “accepted_offer” and by itself is a good predictor for clients behavior. As there was no information about time of reapplication in the dataset and how this time compares to time of offer the “Reapplied” was eliminated from the training data to avoid possible feature from the future error.
The designed ensemble of models shows relatively high predicting power with quite low deviation. The result is consistent if applied to Out of sample data (Testset).

	median	mean	std	min	max
In Sample (Validation set)					
ROC_AUC_score	0.8238	0.8227	0.015	0.7808	0.8627
OUT of Sample (Testset)					
ROC_AUC_score	0.8342	0.834	0.002	0.8267	0.839
Accuracy	0.748	0.7481	0.0037	0.7361	0.7614
F1_score	0.7778	0.7785	0.0063	0.7609	0.8025
Precision_score	0.8884	0.8864	0.0109	0.8423	0.9088
Recall_score	0.6917	0.6944	0.0165	0.6548	0.7615





Among the most important features were several raw features (applied amount, applied term, offered amount), but new engineered features such as number of previous applications, reduction of amount offered, time between current and previous applications played important role in predicting clients behavior.
The resulting model has relatively high predicting power and allows to predict clients new applications with high accuracy (AUC≈0.83).

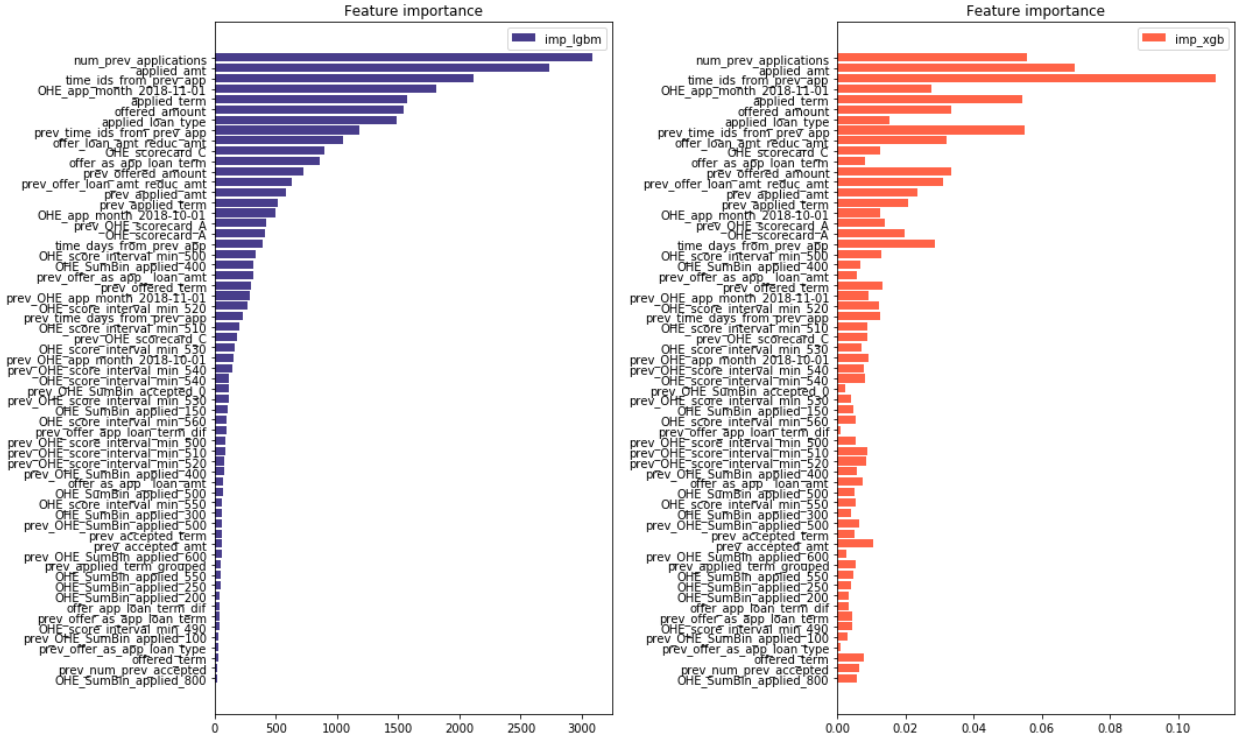
Goal 2: predict “reapplied” (Binary Classification)

The ‘reapplied’ variable can be treated at least in two ways:

- negative scenario, reapplication can be a result of client’s disagreement with the conditions of an offer (amount and term reduction)
- positive scenario, returning client

As there was no information about the time of acceptance of the offer in the dataset and how it compares to time of reapplication the “offer_accepted” was eliminated from the training data to avoid possible feature from the future error.
The designed ensemble of models shows relatively high predicting power with quite low deviation. The result is consistent if applied to Out of sample data (Testset).

	median	mean	std	min	max
In Sample (Validation set)					
ROC_AUC_score	0.7652	0.7638	0.016	0.7054	0.8098
OUT of Sample (Testset)					
ROC_AUC_score	0.7688	0.7681	0.0031	0.7519	0.7763
Accuracy	0.6846	0.6834	0.0074	0.6532	0.7006
F1_score	0.6538	0.6535	0.0067	0.6255	0.6696
Precision_score	0.5793	0.5787	0.0115	0.5399	0.6165
Recall_score	0.7535	0.7522	0.0327	0.6515	0.8585



As well as in previous case (offer acceptance), among the most important features were several raw features (applied amount, applied term, offered amount, applied loan type). New features such as a number of previous applications, a reduction of amount offered, a time between current and previous applications provide additional significant information and play important role in predicting clients’ behavior.

The resulting model has relatively high predicting power and allows to predict clients new applications with relatively high accuracy (AUC≈0.76).

Conclusion

The research showed that, based on data provided, clients' offer acceptance behavior is more straightforward and, thus, easier to predict than reapplication. In general, offer acceptance proves the customer demand for money at the time of the application, while reapplication is a prediction of client's future needs. Moreover, the period of analysis is quite short and most of the loans do not mature by the end date. Models could be improved not only by adding extra information like repayment schedule and repayment history, but also by extending time interval in the dataset.