Client Behavior on Credit Repayments

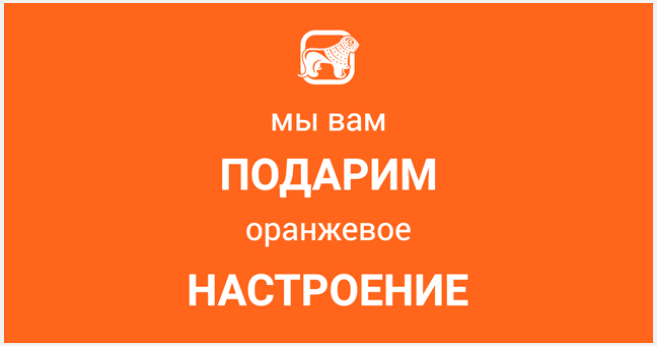
at a bank in the Eastern Europe

Aliaksandr Nekrashevich

Smith SCHOOL OF BUSINESS | Queen's University

Industrial Information

- Belarusky Narodny Bank (BNB-bank) – a small bank located in the Eastern Europe, Republic of Belarus.
- Key business directions: small and medium business, and personal finance.
- Financial capital of the Republic of Georgia in the Republic of Belarus.
- Main credit directions: car loans (9 credit lines) and mortgages (4 credit lines).
- URL: https://www.bnb.by



Problem Formulation

- We focus on the business line responsible for car credits.
- Although credit scoring is working, clients behave differently.
- The natural question is then how to use this variability in client behavior to make the bank more profitable?
- Financial atmosphere and economic ecology are very risky in Belarus, which adds initial challenges to the problem.

Some Context About It

- The problem and dataset were formulated at Datathon 2019. It is a competition in Data Science, typically held at Imaguru Startup Hub in Minsk, Belarus.
- Ideas studied in this project are in part related to what we were doing at the time of the competition. That time we did not manage to implement it.
- This time the ideas are finally implemented, and moreover, new directions and possibilities are studied.

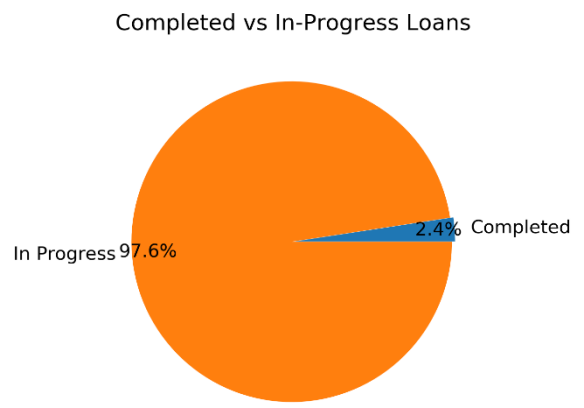
IMAGURU

IMAGURU
БІЗНЕС КЛУБ | STARTUP HUB

Dataset Description

DYNAMIC information contains the repayments information of the main loan body (without interest rate).

- **CONTRACT_ID** – contract identifier
- **PERIOD_ID** -- the month after the contract was issued. When PERIOD_ID = 1, it is the first month after the loan was provided
- **REPAYMENT_SCHEDULED** – amount of payment at the current period according to the contract
- **REPAYMENT_ACTUAL** -- factual payment by the client in the current period. When NULL, it means this period has not yet arrived



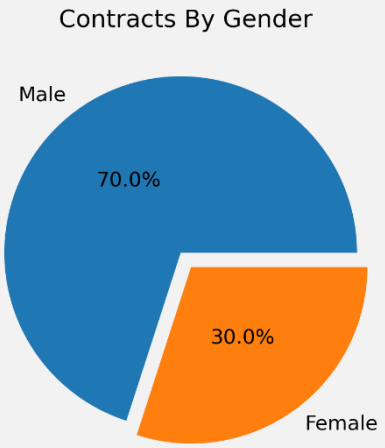
Amount of completed contracts: 91
Amount of in-progress contracts: 3701
Total amount of contracts: 3792
Maximal length of a completed contract: 18

	CONTRACT_ID	PERIOD_ID	REPAYMENT_SCHEDULED	REPAYMENT_ACTUAL
0	17228104	1	19.76	19.76
1	17228104	2	19.76	19.76
2	17228104	3	19.76	19.76
3	17228104	4	19.76	19.76
4	17228104	5	19.76	172.76
5	17228104	6	19.76	268.41

Dataset Description

STATIC information contains initial information about the client and the contract.

- **LOAN_TO_INCOME** – the ratio of the credit amount to the monthly client revenue
- **PAYMENT_TO_INCOME** – the ratio of monthly client payment to the monthly client revenue
- **DOWNPAYMENT** -- ratio of client self-participation in the car purchase. DOWNPAYMENT = 1 - (CONTRACT_SUM / cost of the automobile)
- **GRACE_PERIOD** – the length of grace period in months. At the beginning of the contract during this period, the interest rate is lower than the regular one afterwards. If GRACE_PERIOD = 0, there is no period with discounted interest rate
- **RATE_CHANGE_AFTER_GRACE** – how the interest rate changes after the grace period is over



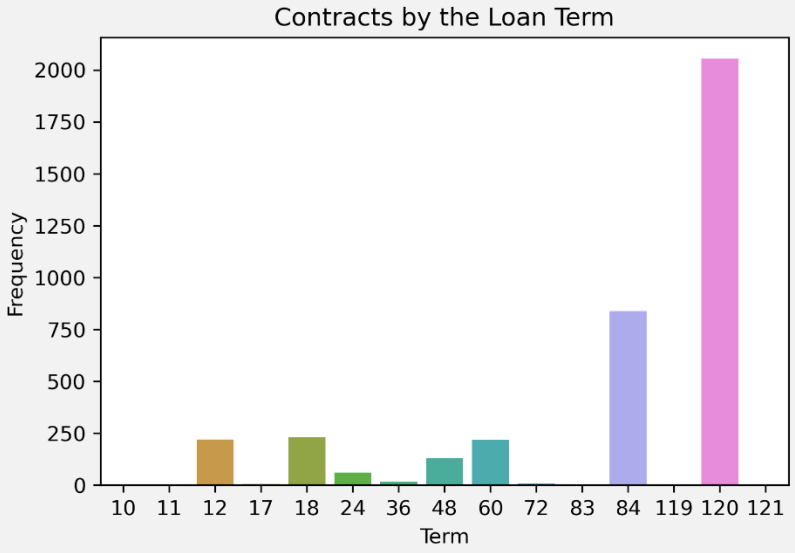
Male Gender: 2655
Female Gender: 1137

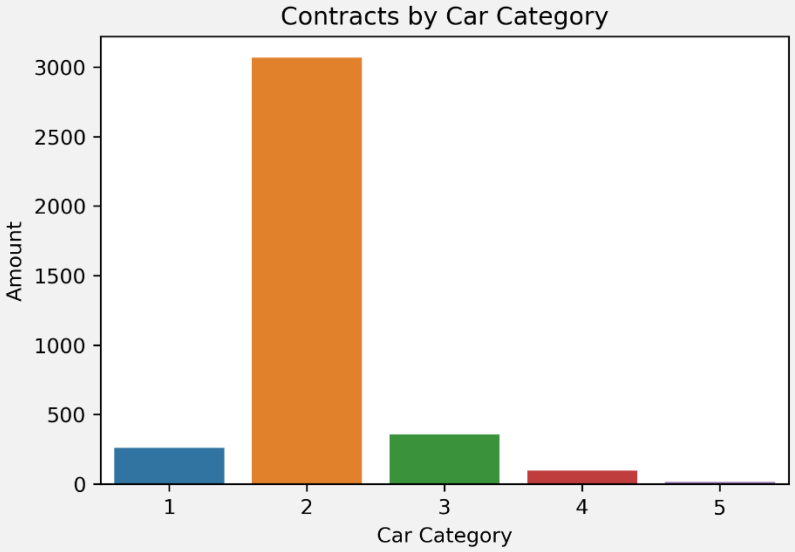
	CONTRACT_ID	CLIENT_ID	TERM	CONTRACT_SUM	GENDER	AGE	LOAN_TO_INCOME	PAYMENT_TO_INCOME	DOWNPAYMENT	CAR_CATEGORY	GRACE_PERIOD	RATE_CHANGE_AFTER_GRACE
0	17228104	251471	60	1185.75	M	32	10	0.22	0.4	2	6	13
1	17237409	251501	18	512.47	M	30	7	0.37	0.7	2	18	15
2	17276280	251669	60	1529.24	M	36	10	0.23	0.1	2	6	13
3	17282809	251684	60	906.53	M	27	6	0.15	0.3	1	6	13
4	17283247	251692	60	1593.5	F	50	18	0.42	0.1	2	6	13
5	17294333	251746	60	1442.03	M	26	16	0.36	0.2	2	6	13
6	17306398	251779	60	971.55	M	51	13	0.29	0.4	2	6	13
7	17320168	251852	60	1748.79	F	54	27	0.62	0.1	2	6	13

Dataset Description

Static information contains initial information about the client and the contract.

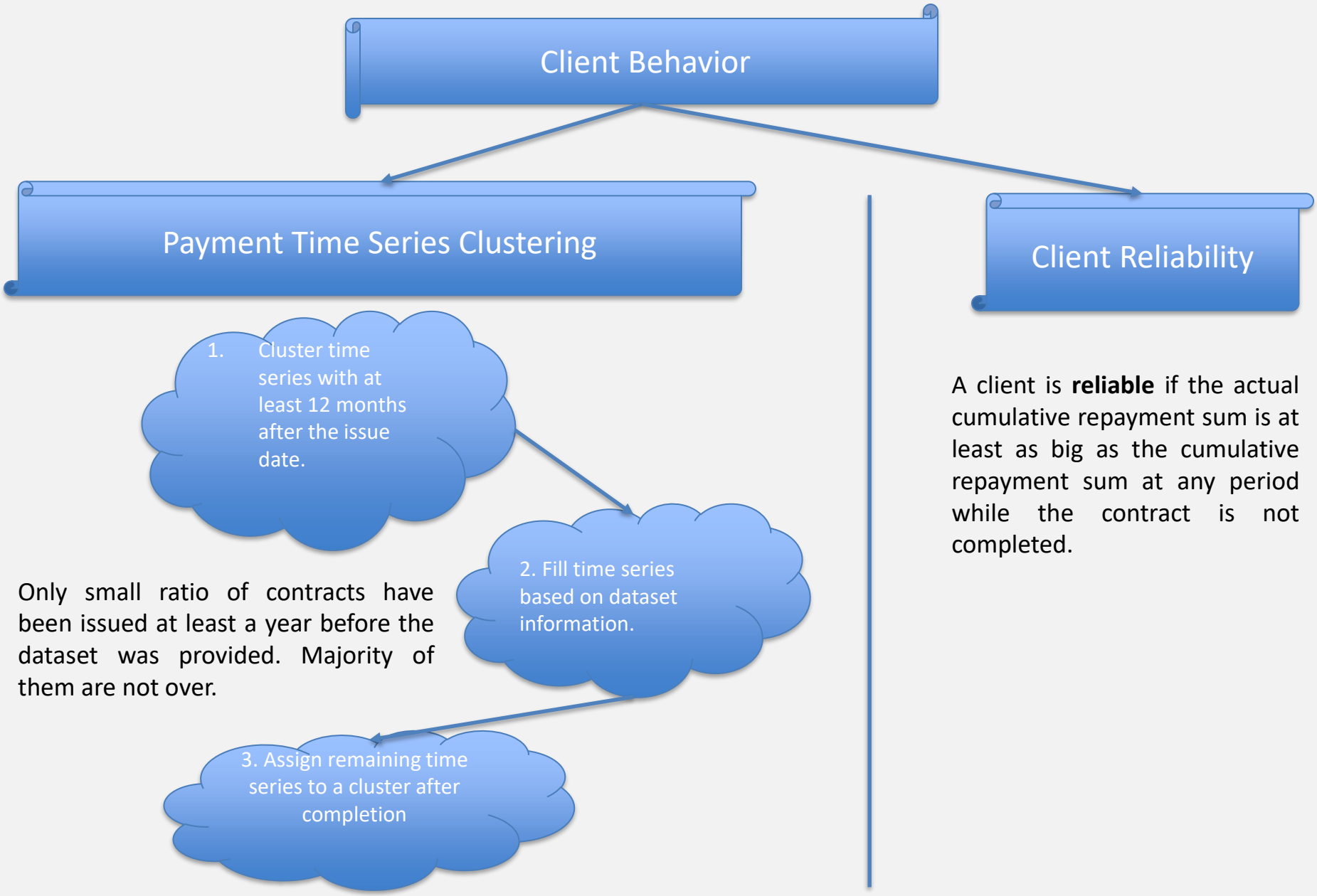
- **CONTRACT_ID** – contract identifier. The contract is a loan for a car purchase.
- **CLIENT_ID** -- client identifier
- **CONTRACT_SUM** – credit amount (in a hidden currency)
- **GENDER** – the gender of the client (M - Male, F - Female)
- **CAR_CATEGORY** – category of the purchased automobile. There are five different categories. The minimal budget category is 1, the maximal premium is 5.
- **TERM** -- amount of months for credit repayment
- **AGE** – age of the client at the time the loan was issued





Pair-plot of Static Features

Framework Overview

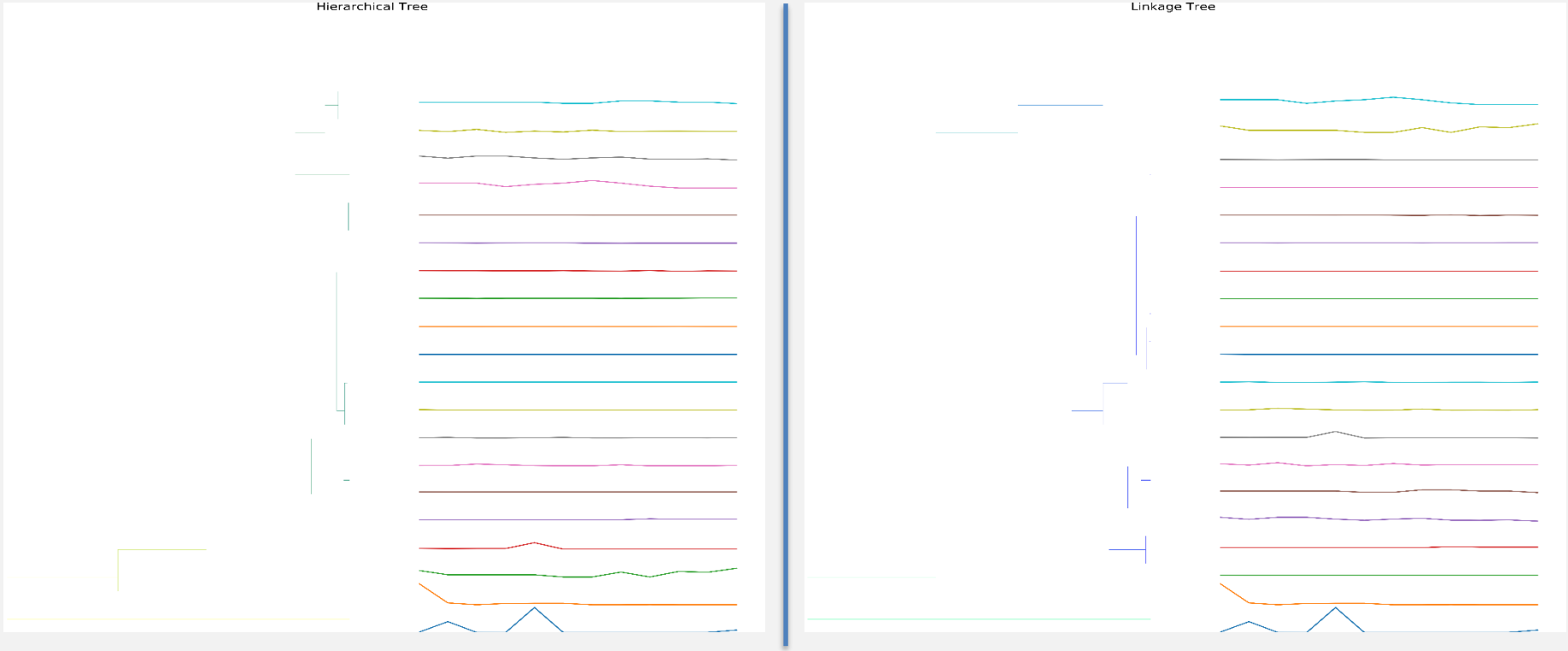


Only small ratio of contracts have been issued at least a year before the dataset was provided. Majority of them are not over.

A client is **reliable** if the actual cumulative repayment sum is at least as big as the cumulative repayment sum at any period while the contract is not completed.

Payment Time Series Clustering

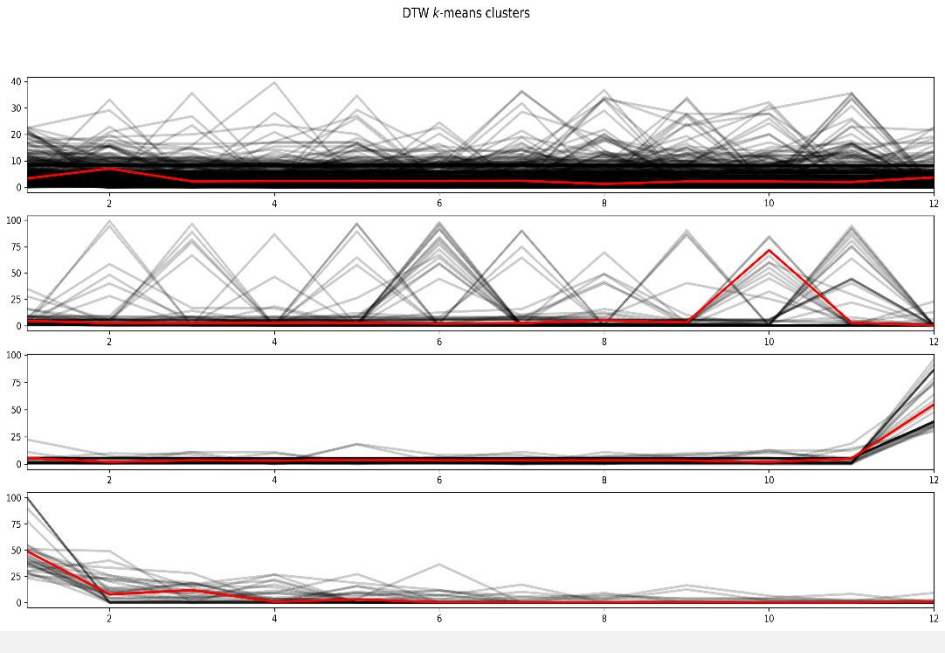
- **Step 1:** Guess the number of clusters. Hierarchical clustering on a small subset.
- **Assumption:** repayments are converted to relative (divided by contract sum and multiplied by 100%).

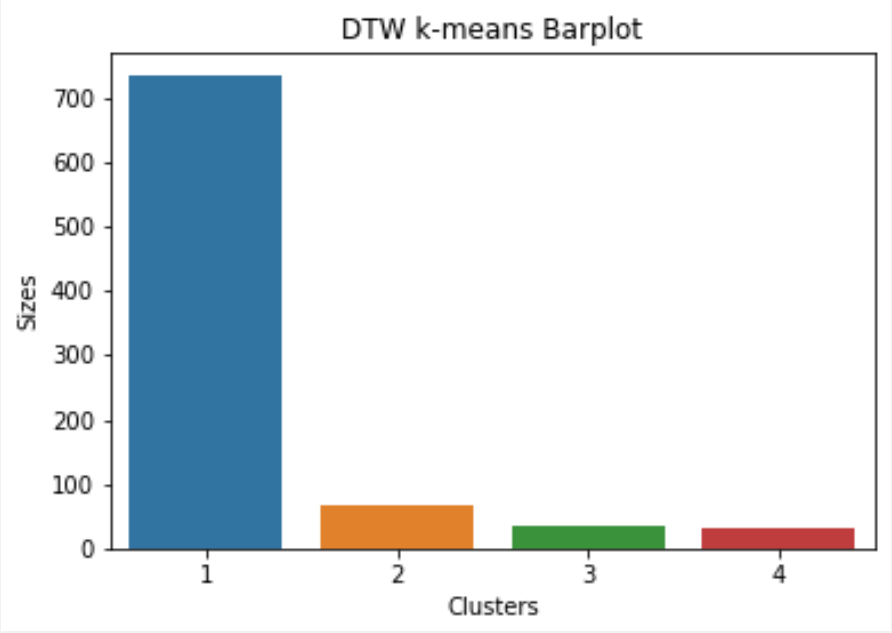


Visualization allows to make a guess a number of clusters as 3 to 5. The following behavior patters are visual: flat repayments, or single spike, or big initially then flat, or flat and big in the end.

Payment Time Series Clustering

- **<u>Step 2</u>:** Trying methods with pre-defined cluster amount. They are k-Medoids, k-Shape, DTW k-Means, and Kernel k-Means.



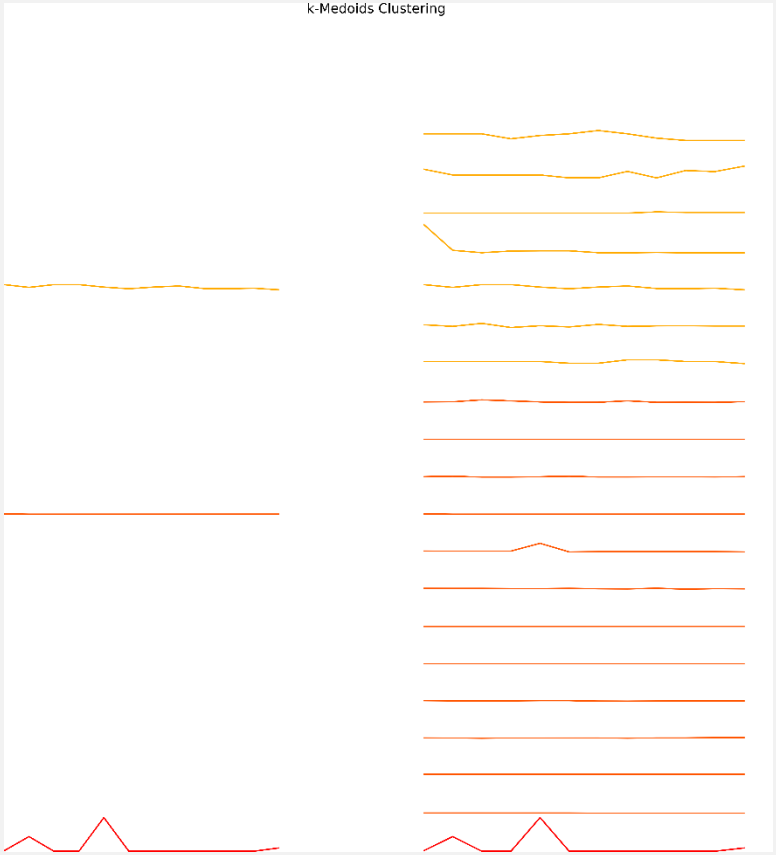


The most reasonable result is obtained by **DTW k-Means**. The following four patterns look the most realistic:

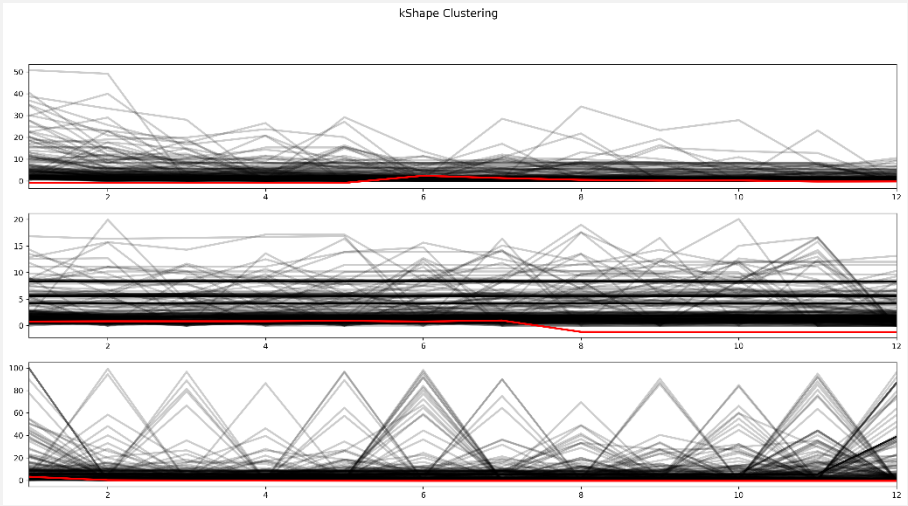
1. Flat repayment
2. Spiked repayment (flat, then fixing almost all debt, and then flatly concluding)
3. Flat and high payment at the end.
4. High payment initially and flat repayment until the end.

Payment Time Series Clustering

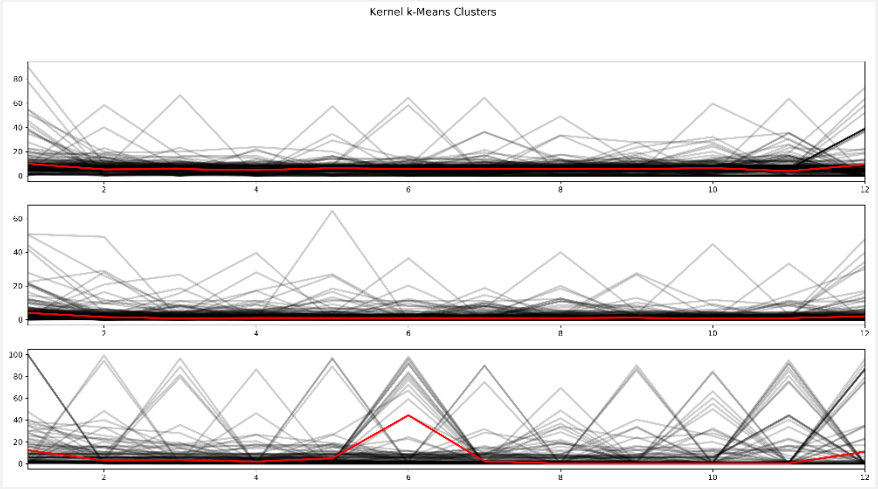
- Showing another options, with a purpose of comparison.



k-Medoids: flat and spike



K-Shape: flat and spike



Kernel k-means: flat and spike

Next Month Repayment Regression

- To cluster all time series, we first need to complete them. Standard time series techniques may not work, because of context, behavior structure and limited information.
- It can be reasonable to predict repayment, log-repayment, relative repayment, and log-relative repayment. But let's describe Feature Engineering first.

Features	Description
CUMSUM_*	Cumulative repayments (actual and scheduled).
AVERAGE_*	Average repayment until this period (actual and scheduled).
RELATIVE_*, RELATIVE_CUMSUM_*	Relative repayments, periodic, average and cumulative, actual and scheduled.
LOG_*, LOG_RELATIVE_*, LOG_RELATIVE_CUMSUM_*, etc.	Log-transformation of features (cumulative, average, and relative), np.log1p.
HISTORY_*, HISTORY_LOG_*, HISTORY_LOG_RELATIVE_*, etc.	Information from two last periods, replaced by scheduled amounts if it is one of the first two periods.
IS_GRACE_CONSTANT_*	If payments during grace were constant.
RATIO_*	If LOAN_TO_INCOME is above some threshold (5, 10, 20, 30, 40).
IS_CAR_*	Indicator dummy on car category.
IS_PERIOD_*	Dummy for first periods 1-2-3.
IS_GRACE_ON_*	Indicator if grace is on.

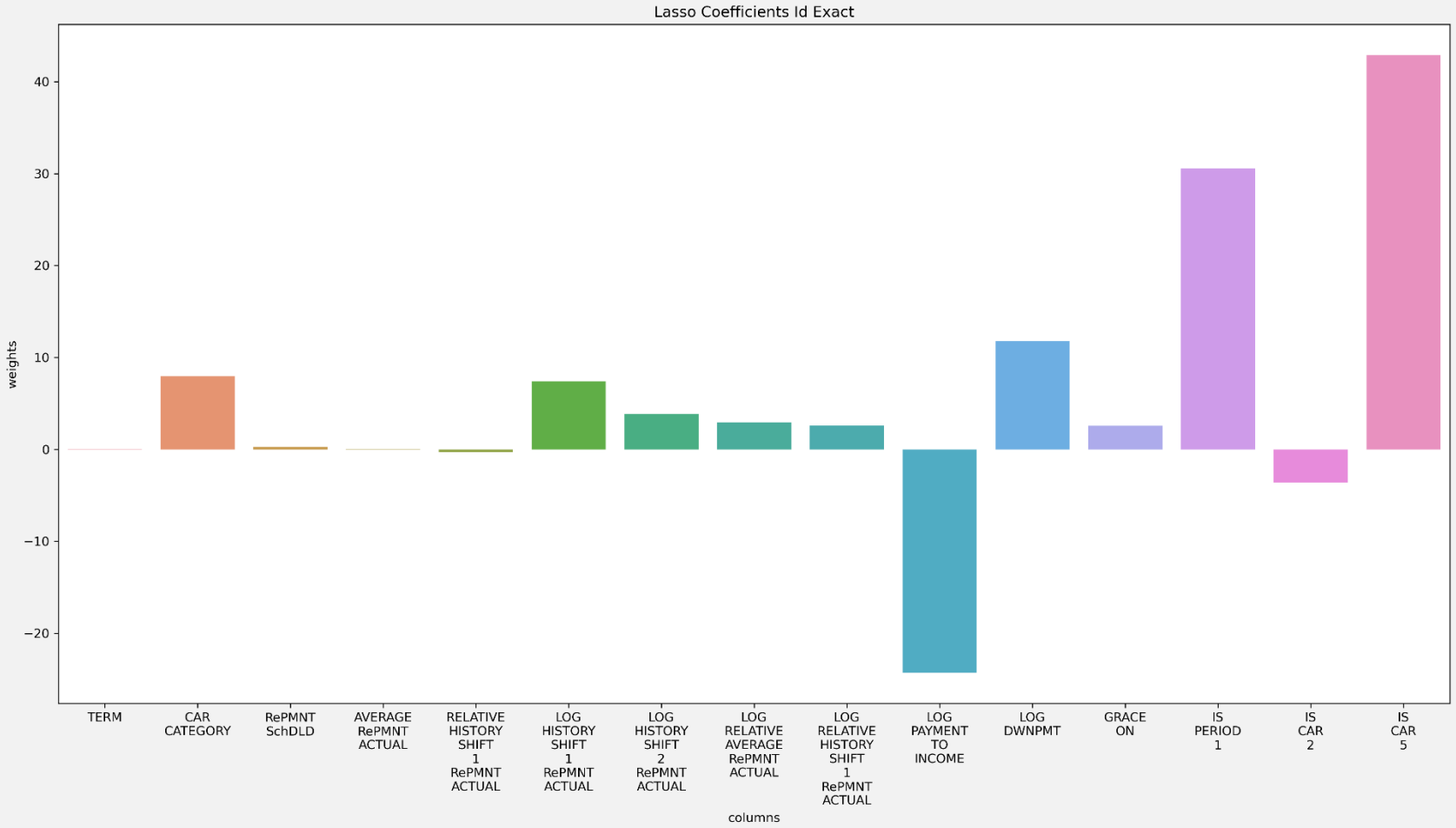
Next Month Repayment Regression

Algorithms	MAE	RMSE	MAXERR	EXPVAR
Baseline	2.314208	7.711478	99.1737	0.032126
ridge_regression_id_exact	2.453963	7.075056	99.29972	0.133426
ridge_regression_log_exact	1.9456	7.278206	98.96441	0.114408
ridge_regression_id_relative	2.480036	7.083142	99.38764	0.131396
ridge_regression_log_relative	1.90973	7.197327	99.02143	0.123661
lasso_regression_id_exact	2.666097	7.18833	98.30736	0.105932
lasso_regression_log_exact	2.811763	7.781552	99.04411	0.028616
lasso_regression_id_relative	3.428801	7.500937	96.54698	0.025795
lasso_regression_log_relative	3.015005	7.797025	98.13199	0
huber_regression_id_exact	1.957839	7.323161	98.92046	0.103697
huber_regression_log_exact	1.930997	7.247734	98.91808	0.115738
huber_regression_id_relative	1.994627	7.301279	99.10232	0.102372
huber_regression_log_relative	1.903285	7.309112	98.99213	0.097596
bayesian_regression_id_exact	2.455412	7.075383	99.2955	0.133342
bayesian_regression_log_exact	1.930281	7.27454	98.96367	0.114697
bayesian_regression_id_relative	2.480322	7.08103	99.41173	0.131897
bayesian_regression_log_relative	1.895273	7.187567	99.02326	0.125473
random_forest_id_exact	2.259798	7.082932	98.25778	0.131682
random_forest_log_exact	1.755068	7.193584	98.87177	0.130697
random_forest_id_relative	2.274521	7.061995	98.52509	0.136839
random_forest_log_relative	**1.727013**	**7.1157**	**98.80793**	**0.140065**
xgboost_id_exact	2.636633	8.1178	99.00763	-0.11386
xgboost_log_exact	2.942483	7.945959	99.45148	0.050481
xgboost_id_relative	2.282406	7.571633	98.55596	0.024771
xgboost_log_relative	2.345745	7.529081	98.75005	0.088634
sklearn_neural_net_id_exact	2.515001	7.041476	98.59635	0.143579
sklearn_neural_net_log_exact	2.211451	7.209658	98.56758	0.104569
sklearn_neural_net_id_relative	2.258668	7.140651	99.70385	0.128082
sklearn_neural_net_log_relative	3.507272	10.53171	176.5617	-0.92019
svr_id_exact	4.852726	9.094069	102.3217	-0.04404
svr_log_exact	3.269467	8.094878	100.0037	0.045698
svr_id_relative	5.29617	8.901884	104.3575	0.098913
svr_log_relative	3.831688	8.351416	100.528	0.034598
knn_regressor_id_exact	2.579049	7.408486	98.80412	0.049759
knn_regressor_log_exact	1.989511	7.253768	98.84488	0.114534
knn_regressor_id_relative	2.484756	7.32596	98.88255	0.070682
knn_regressor_log_relative	1.935938	7.200064	98.88734	0.121683

Algorithms	MAE	RMSE	MAXERR	EXPVAR
Baseline	26.38009	98.94643	1947.34	0.012502
ridge_regression_id_exact	29.39517	93.34552	1867.763	0.074884
ridge_regression_log_exact	22.93969	95.13858	1897.17	0.066014
ridge_regression_id_relative	29.98678	93.18889	1840.458	0.077915
ridge_regression_log_relative	22.57219	94.11738	1880.579	0.076144
lasso_regression_id_exact	31.33117	94.09595	1884.1	0.060028
lasso_regression_log_exact	31.85741	100.0341	1948.598	-4.44E-16
lasso_regression_id_relative	44.18033	98.76968	1865.038	-0.0288
lasso_regression_log_relative	34.79259	98.83935	1916.248	-0.01258
huber_regression_id_exact	23.0813	95.75555	1915.885	0.054024
huber_regression_log_exact	22.67185	94.80628	1878.079	0.067304
huber_regression_id_relative	23.63774	95.04246	1885.542	0.061504
huber_regression_log_relative	22.55198	94.70119	1880.344	0.066677
bayesian_regression_id_exact	29.40727	93.34155	1867.654	0.07496
bayesian_regression_log_exact	22.8654	95.08873	1896.491	0.066656
bayesian_regression_id_relative	30.05022	93.19813	1839.585	0.077713
bayesian_regression_log_relative	22.49696	94.06656	1879.973	0.076907
random_forest_id_exact	26.88662	93.30122	1836.525	0.075736
random_forest_log_exact	21.06228	94.68814	1882.059	0.071564
random_forest_id_relative	27.923	93.59628	1842.353	0.070433
random_forest_log_relative	**20.73473**	**93.7506**	**1874.312**	**0.080352**
xgboost_id_exact	31.05579	103.3138	1906.516	-0.11151
xgboost_log_exact	33.40411	101.287	1945.639	0.023602
xgboost_id_relative	26.37833	96.50098	1894.067	0.023366
xgboost_log_relative	26.43769	96.89034	1898.132	0.053083
sklearn_neural_net_id_exact	30.68476	93.14774	1869.026	0.081145
sklearn_neural_net_log_exact	26.60569	94.63519	1878.667	0.052348
sklearn_neural_net_id_relative	28.83009	96.25677	1909.721	0.034581
sklearn_neural_net_log_relative	43.56092	142.6136	2384.555	-1.1583
svr_id_exact	51.88139	107.2871	1952.762	0.047778
svr_log_exact	37.48597	102.8881	1946.624	0.019756
svr_id_relative	70.68696	121.1039	2003.991	-0.04067
svr_log_relative	43.23752	103.8542	1905.198	0.035913
knn_regressor_id_exact	30.28266	96.35495	1878.207	0.014076
knn_regressor_log_exact	23.61601	95.36429	1903.723	0.059622
knn_regressor_id_relative	29.77508	96.17761	1862.24	0.01769
knn_regressor_log_relative	23.06363	94.42481	1878.843	0.069547

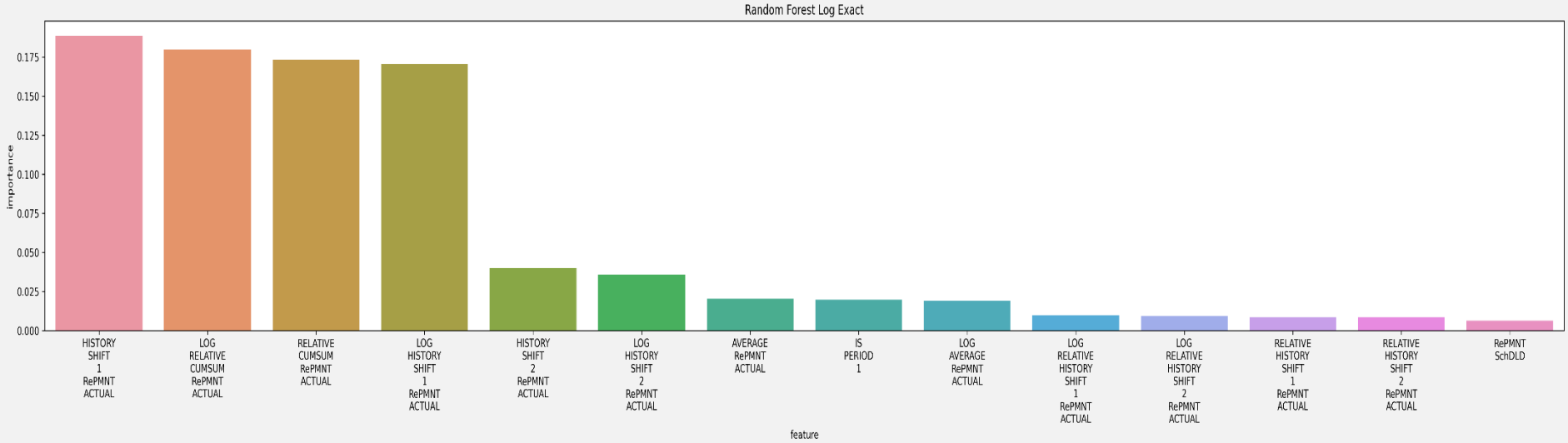
Next Month Repayment Regression

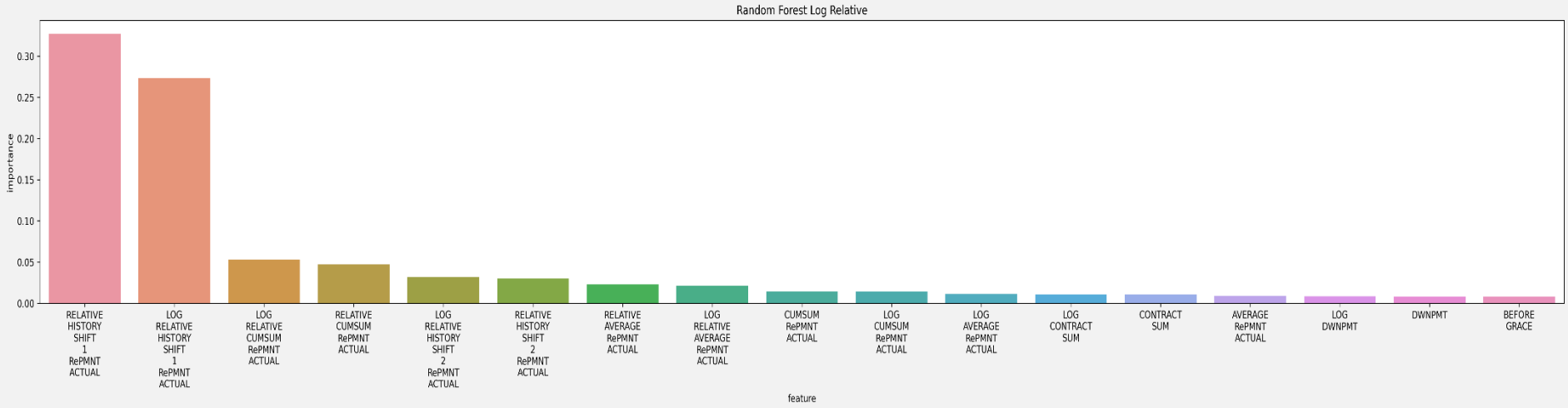
- As a side effect, we can estimate how features impact next month repayments.



Next Month Repayment Regression

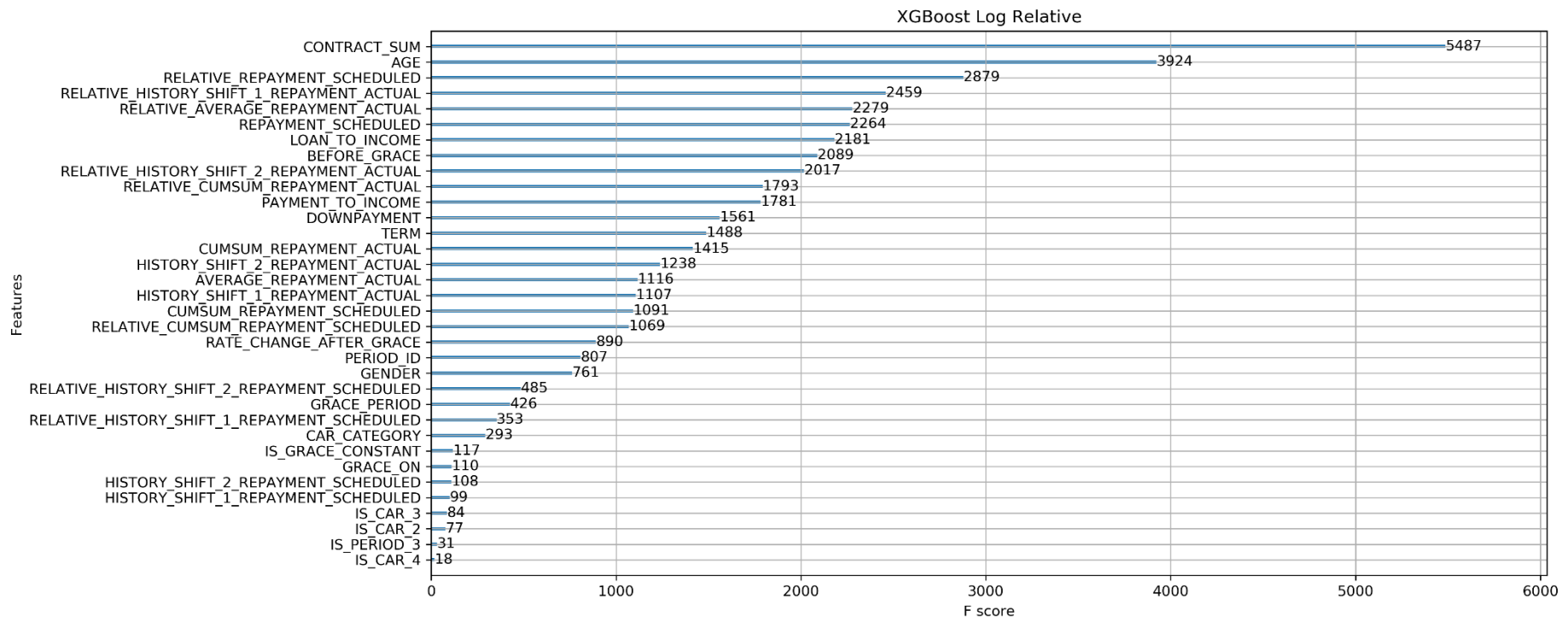
- As a side effect, we can estimate how features impact next month repayments.





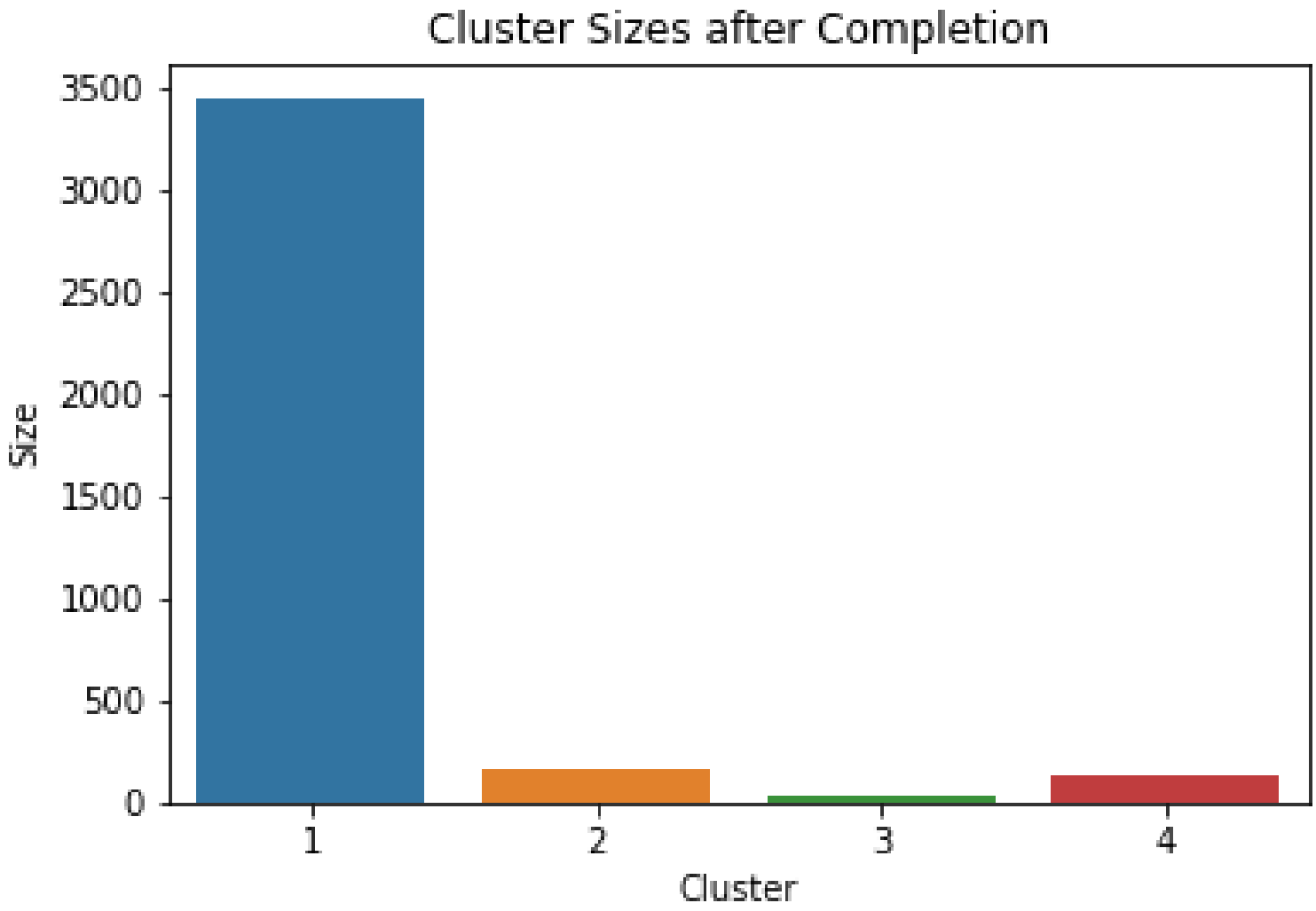
Next Month Repayment Regression

- As a side effect, we can estimate how features impact next month repayments.



XGBoost Feature Importances

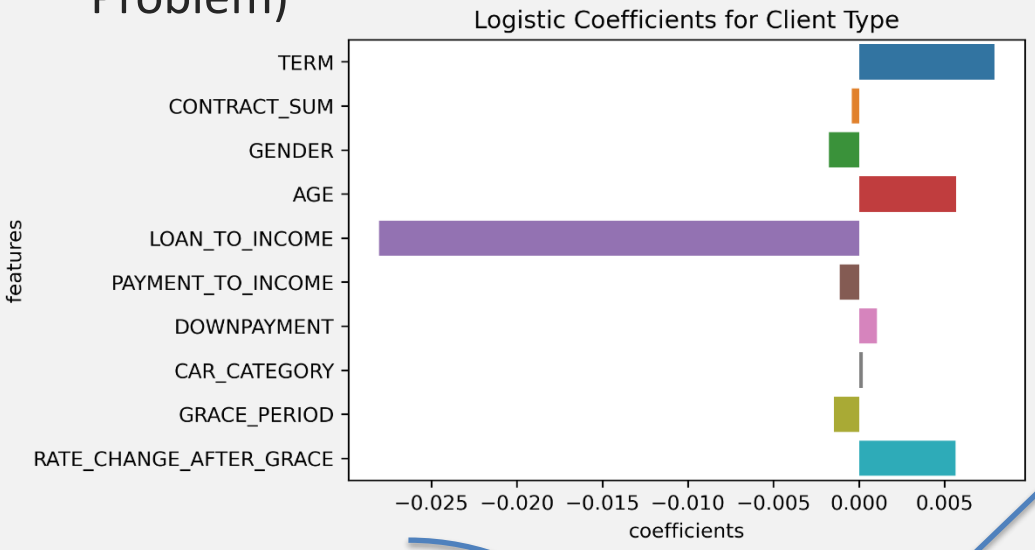
Clustering After Completion



Majority of clients are still very much linked to the schedule. However, we have found more clients with less standard behavior patterns.

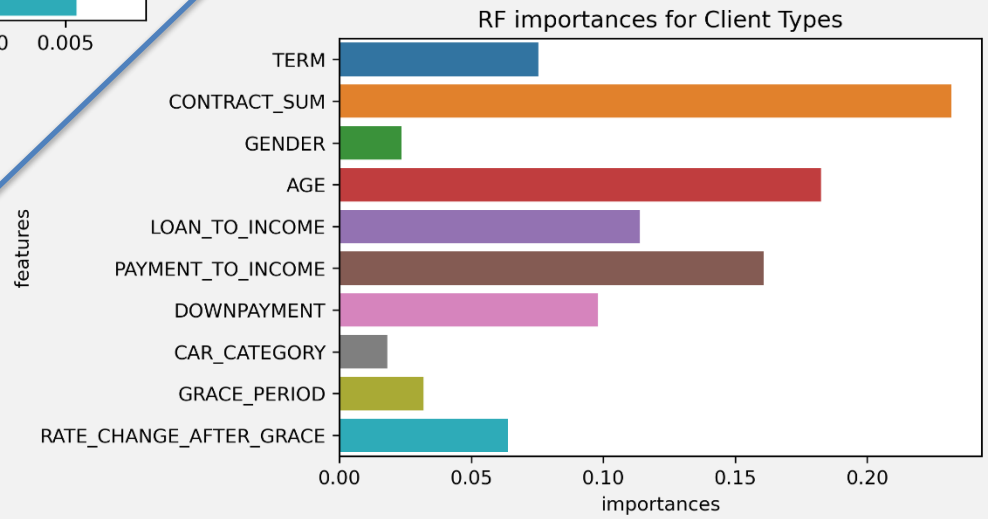
Client Reliability

- 3348 reliable, 444 unreliable in the training set (Imbalanced Classification Problem)



A client is **reliable** if the actual cumulative repayment sum is at least as big as the cumulative repayment sum at any period while the contract is not completed.

Reliable money can-be reinvested and improve Funds Transfer Pricing.



Algorithm	accuracy	precision	recall	f1-score	reliable precision	reliable recall	unreliable precision	unreliable recall
Logistic Regression	0.583388	0.81586	0.583388	0.655304	0.906976744	0.585585586	0.159817352	0.159817352
Random Forest	0.8029	0.808723	0.8029	0.805762	0.891584534	0.882882883	0.212121212	0.212121212

Next Steps

Promising Directions:

- Time Series estimation via Recurrent Neural Nets, e.g. LSTM.
- Cluster prediction using only static features.
- Cluster prediction using static features and dynamic features generated by Tsfresh.
- Imbalanced Classification for Reliability Prediction (ImbLearn: SMOTE, ROSE, etc.)
- Your Suggestions?

Summary

Results:

- Client repayment clustering into four behavior patterns.
- Interpretation for repayment drivers at the next period.
- Client reliability forecasting and interpretation.

Thank you for your attention!
Questions? Remarks, Suggestions, Comments?