Credit Risk Modelling Project

Credit Acquisition Scorecard

Authors: **Borislav Aymaliev, Stefan Antonov**

Our Dataset

Data set of over 21 000 credit applications

• Information related to the applicants
– age
– education
– income
– marital status
etc.
• Information about the performance of the applicants

Performance period selection

- Our data was for the time period from September 2015 until September 2016

	-18	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6
0	1201	852	840	674	1008	1066	1268	1156	1462	1424	1330	1275	1354
1	24	16	11	15	13	21	25	31	17	17	21	28	22

- Since september 2016 is just 6 months ago, we chose to have at least 9 months performance period, so our accepted customers from 15171 were cut down to 9700

Our key differentiators

We created a process flow that is:

- Linear - simple to comprehend and follow
- Reproducible - can lead to the very same results on different machines
- Highly automated - can easily be adapted to a different dataset with minimum manual intervention
- Implemented in the latest technology Data Science IDE - Jupyter Notebook on top of an R kernel

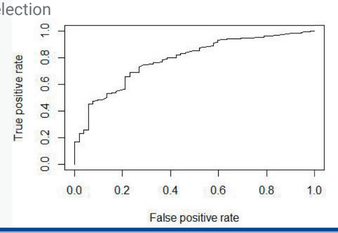
Procedures We Executed

- Correlation between continuous variables
- Optimal binning, based on IV
- Drop no-IV variables
- Correlation between all dummified variables
- Drop “year_bank”, “gr_months_from_OldestOpenDate” , “MARITAL_STATUSMarried” on the basis that other variables exist, which carry the same information

Modelling 1

Logistic regression with backwards selection

- 15min execution time
- Concordance of 80.37679%
- Somers D of 0.6124588
- Kendall's Tau A of 0.02139353
- Test set K-S = 0.5072132



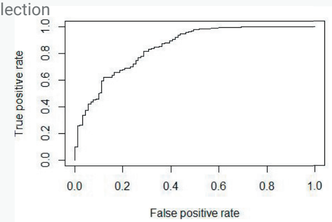
Reject Inference

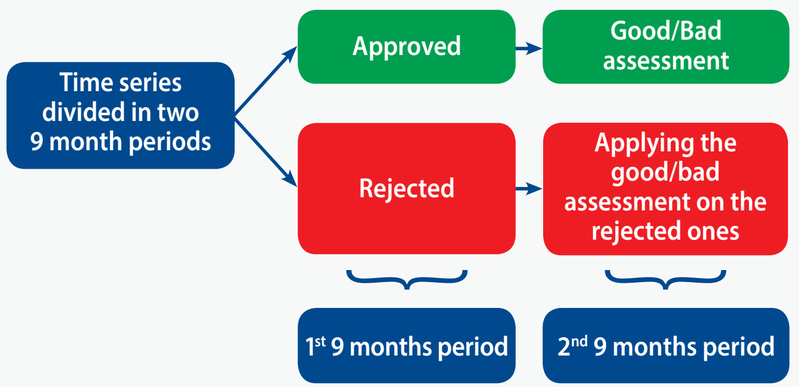
- Simple (Hard Cutoff) Augmentation
- Cutoff DR = 3,16%
  - 138 bad
  - out of 4367 total rejected
- Combined both datasets
- Split the resulting dataset into training_set_2 with 9819 observations and test_set_2 with 4248 observations by using stratified sampling in respect to the target

Modelling 2

Logistic regression with backwards selection

- 20min execution time
- Concordance of 85.89915%
- Somers D of 0.721613
- Kendall's Tau A of 0.03006119
- Test set K-S of 0.5576267





After applying the model over the rejected applicants we build a new model over the complete dataset. We divide on training set and test set both among the approved and on the rejected applicants. We do not use data fresher than 9 months.

Distribution of customers into score groups

GroupV	Data		
	Average - Score	Good Rate	Bad Rate
0	130.0066018	77.40%	22.60%
1	244.2741369	95.48%	4.52%
2	295.1880031	97.18%	2.82%
3	335.4708277	98.31%	1.69%
4	368.3888847	96.61%	3.39%
5	397.6556842	97.18%	2.82%
6	425.654851	99.44%	0.56%
7	443.6480693	98.87%	1.13%
8	466.413729	98.87%	1.13%
9	488.8760095	99.44%	0.56%
10	503.7162643	98.87%	1.13%
11	526.2738064	98.87%	1.13%
12	544.2351962	99.44%	0.56%
13	564.5734189	99.44%	0.56%
14	583.2689366	99.44%	0.56%
15	603.1085075	98.87%	1.13%
16	621.4094978	100.00%	0.00%
17	647.2055077	99.44%	0.56%
18	672.6002256	99.44%	0.56%
19	695.4665166	99.44%	0.56%
20	724.9304233	100.00%	0.00%
21	763.1279945	100.00%	0.00%
22	812.7609817	100.00%	0.00%
23	916.2911108	100.00%	0.00%
Total Result	**532.272716**	**98.00%**	2.00%