

Home Credit Default Risk

Eric Huynh

Data Capstone 2

Problem & Overview:

- Countless numbers of loan/credit applications are refused each day by financial institutions
 - Due to monetary risk of credit defaults by clients
- Financial institutions are stringent on qualifications to request large loans.
 - Limits applicants to only clients that already have a strong credit record.
 - Forces clients with little to non-existent credit history to rely on unreliable lenders to meet their financial needs.

Solution and Objective

- Certain financial institutions have strived to expand the financial inclusion to those with insufficient or non-existing credit histories.
 - Utilize a variety of alternate data to gauge and measure the clients' risk profiles without relying on the conventional methods.
- Objective:
 - Draw reliable variables from the alternate data that can reliably gauge a clients' risk profile.
 - Develop a predictive model that can accurately classify clients as high-risk or low-risk for credit loans

Potential Clients

- Financial Institutions:
 - Expand their loans applicant pool by measuring new clients' risk profile with alternate data.
 - Higher yields both short-term and long-term as positive financial experience will bring in reoccurring clients.
- Loan Applicants:
 - Understand alternative attributes that can help them develop towards becoming a low-risk applicant despite insufficient credit histories.
 - Also applicable for anyone applying for a larger loan they may not currently qualify for under traditional requirements.

Data from Kaggle Competition

- Training Data:
 - Consists of 307,511 unique client IDs and 122 categories.
- 6 Supplemental Datasets (only 4 used):
 - “bureau.csv” – information regarding all previous credits
 - “credit_card_balance.csv” – previous credit information with Home Credit
 - “installments_payments.csv” – repayment history for approved credit by Home Credit
 - “previous_application.csv” – all previous applications made to Home Credit
- 300+ heterogenous mixture of both categorical and quantitative features

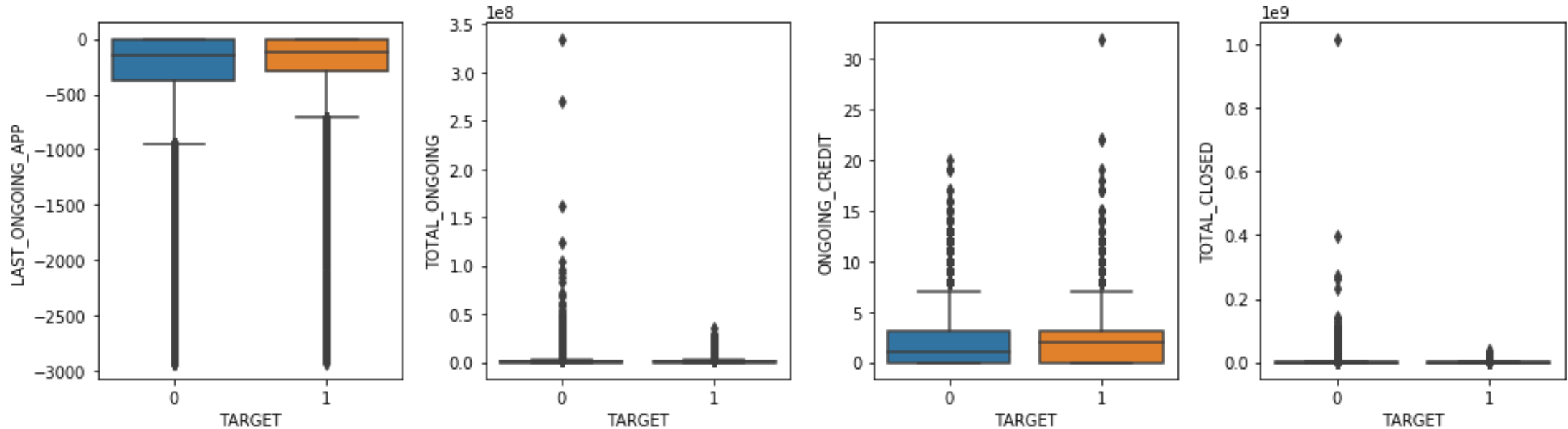
Numerical Correlation

AMT_INCOME_TOTAL	1	0.16	0.074	0.028	0.021	-0.015	-0.086	0.033	-0.015	0.13	0.061	0.077	0.037	-0.016	0.00073	-0.0043	0.067	0.0097	0.0087	0.02	0.0013	0.013	0.027	-0.037	0.0084	0.014
AMT_CREDIT	0.16	1	0.1	0.0096	0.028	-0.018	-0.1	0.0061	-0.052	0.13	0.05	0.096	0.059	-0.036	0.0077	0.037	0.098	-0.037	-0.039	0.06	-0.0018	0.0021	-0.055	-0.1	-0.0066	0.21
REGION_POPULATION_RELATIVE	0.074	0.1	1	-0.054	-0.011	0.0024	-0.54	0.025	-0.015	0.041	-0.019	0.029	-0.041	-0.025	0.003	0.018	0.051	-0.01	-0.013	0.00029	-0.0064	-0.026	-0.03	0.0019	-0.004	-0.0059
DAYS_REGISTRATION	0.028	0.0096	-0.054	1	0.045	0.0033	0.08	-0.023	0.017	0.092	0.027	0.032	0.00056	-0.0013	-0.0087	-0.031	-0.054	0.033	0.015	-0.0043	0.039	0.18	0.33	0.047	0.1	-0.0014
OWN_CAR_AGE	0.021	0.028	-0.011	0.045	1	-0.0059	0.026	-0.017	-0.0087	0.03	0.018	0.015	0.021	0.014	-0.00026	0.0055	-0.014	-0.012	-0.014	-0.012	0.0025	0.069	0.085	-0.011	0.013	-0.0065
DEF_30_CNT_SOCIAL_CIRCLE	-0.015	-0.018	0.0024	0.0033	-0.0059	1	0.018	0.0069	-0.0087	0.0081	-0.0089	-0.00089	-0.011	0.0023	0.0043	0.015	-0.0026	0.0017	-0.012	-0.014	0.00072	-0.0006	0.014	0.0021	0.023	
HOUSING	-0.086	-0.1	-0.54	0.08	0.026	0.018	1	-0.013	0.017	-0.042	0.031	-0.029	0.054	0.029	-0.0024	-0.011	-0.035	0.0059	0.012	-0.0066	-0.01	0.025	0.0093	0.0043	-0.0054	0.0014
ENQUIRIES	0.033	0.0061	0.025	-0.023	-0.017	0.014	-0.013	1	-0.099	0.07	0.2	0.05	0.21	-0.22	-0.016	-0.0056	0.3	0.14	0.19	0.079	-0.23	-0.03	0.071	-0.017	-0.065	0.021
LAST_ONGOING_APP	-0.015	-0.052	-0.015	0.017	-0.0087	0.0069	0.017	-0.099	1	-0.062	-0.067	-0.03	-0.065	0.015	-0.0041	-0.011	-0.041	0.027	0.018	-0.0056	0.016	-0.0046	0.044	0.047	0.065	-0.014
TOTAL_ONGOING	0.13	0.13	0.041	0.092	0.03	-0.0087	-0.042	0.07	-0.062	1	0.39	0.22	0.2	-0.014	3.3e-05	0.027	0.038	-0.043	-0.051	0.012	0.011	0.051	0.059	-0.093	-0.009	0.02
ONGOING_CREDIT	0.061	0.05	-0.019	0.027	0.018	0.0081	0.031	0.2	-0.067	0.39	1	0.14	0.46	-0.049	0.0068	0.036	0.052	-0.051	-0.045	-0.0013	-0.035	0.016	-0.0091	-0.098	-0.043	0.023
TOTAL_CLOSED	0.077	0.096	0.029	0.032	0.015	-0.0089	-0.029	0.05	-0.03	0.22	0.14	1	0.28	-0.014	0.0015	0.02	0.047	-0.014	-0.021	0.024	-0.0016	0.01	-0.014	-0.062	-0.026	0.015
CLOSED_CREDIT	0.037	0.059	-0.041	-0.00056	0.021	-0.00089	0.054	0.21	-0.065	0.2	0.46	0.28	1	-0.052	0.0054	0.041	0.085	-0.035	-0.041	0.037	-0.098	0.0074	-0.087	-0.13	-0.13	0.032
MONTHS_BALANCE	-0.016	-0.036	-0.025	-0.0013	0.014	-0.011	0.029	-0.22	0.015	-0.014	-0.049	-0.014	-0.052	1	-0.1	-0.051	-0.16	-0.066	-0.097	-0.038	0.097	0.0027	-0.0017	0.043	-0.0031	-0.093
SK_DPD	-0.00073	0.0077	0.003	-0.0087	-0.00026	0.0023	-0.0024	-0.016	-0.0041	3.3e-05	0.0068	0.0015	0.0054	-0.1	1	0.053	-0.018	-0.027	-0.029	-0.0022	0.018	-0.0028	-0.012	-0.011	-0.0037	0.0097
NET_PAID	-0.0043	0.037	0.018	-0.031	0.0055	0.0043	-0.011	-0.0056	-0.011	0.027	0.036	0.02	0.041	-0.051	0.053	1	-0.0054	-0.066	-0.066	0.0056	-0.018	-0.022	-0.064	-0.028	-0.031	-0.0056
PREV_CREDIT	0.067	0.098	0.051	-0.054	-0.014	0.015	-0.035	0.3	-0.041	0.038	0.052	0.047	0.085	-0.16	-0.018	-0.0054	1	0.063	0.22	0.1	-0.41	-0.055	-0.17	-0.061	-0.045	0.018
DAYS_DECISION	-0.0097	-0.037	-0.01	0.033	-0.012	-0.0026	0.0059	0.14	0.027	-0.043	-0.051	-0.014	-0.035	-0.066	-0.027	-0.066	0.063	1	0.93	0.066	-0.075	0.0041	0.042	0.058	0.042	-0.075
DAYS_TERMINATION	-0.0087	-0.039	-0.013	0.015	-0.014	0.0017	0.012	0.19	0.018	-0.051	-0.045	-0.021	-0.041	-0.097	-0.029	-0.066	0.22	0.93	1	0.052	-0.16	-0.0081	0.0098	0.048	0.035	-0.077
NET_PAYMENT	0.02	0.06	-0.00029	-0.0043	-0.012	-0.012	-0.0066	0.079	-0.0056	0.012	-0.0013	0.024	0.037	-0.038	-0.0022	0.0056	0.1	0.066	0.052	1	-0.066	-0.014	-0.032	-0.0088	-0.0081	-0.0033
PAYMENT_TIME	-0.0013	-0.0018	-0.0064	0.039	0.0025	-0.014	-0.01	-0.23	0.016	0.011	-0.035	-0.0016	-0.098	0.097	0.018	-0.018	-0.41	-0.075	-0.16	-0.066	1	0.023	0.11	0.021	-0.057	-0.00028
CNT_CHILDREN	0.013	0.0021	-0.026	0.18	0.069	-0.00072	0.025	-0.03	-0.0046	0.051	0.016	0.01	0.0074	0.0027	-0.0028	-0.022	-0.055	0.0041	-0.0081	-0.014	0.023	1	0.33	-0.042	-0.028	-0.016
DAYS_BIRTH	0.027	-0.055	-0.03	0.33	0.085	-0.0006	0.0093	-0.071	0.044	0.059	-0.0091	-0.014	-0.087	-0.0017	-0.012	-0.064	-0.17	0.042	0.0098	-0.032	0.11	0.33	1	-0.0044	0.27	-0.051
DAYS_EMPLOYED	-0.037	-0.1	0.0019	0.047	-0.011	0.014	0.0043	-0.017	0.047	-0.093	-0.098	-0.062	-0.13	0.043	-0.011	-0.028	-0.061	0.058	0.048	-0.0088	0.021	-0.042	0.00044	1	-0.034	-0.021
DAYS_ID_PUBLISH	-0.0084	-0.0066	-0.004	0.1	0.013	0.0021	-0.0054	-0.065	0.065	-0.009	-0.043	-0.026	-0.13	-0.031	-0.0037	-0.031	-0.045	0.042	0.035	-0.0081	0.057	-0.028	0.27	-0.034	1	-0.046
DOCUMENTS	-0.014	0.21	-0.0059	-0.0014	-0.0065	0.023	0.0014	0.021	-0.014	0.02	0.023	0.015	0.032	-0.093	0.0097	-0.0056	0.018	-0.075	-0.077	-0.0033	-0.00028	-0.016	-0.051	-0.021	-0.046	1
AMT_INCOME_TOTAL																										
AMT_CREDIT																										
REGION_POPULATION_RELATIVE																										
DAYS_REGISTRATION																										
OWN_CAR_AGE																										
DEF_30_CNT_SOCIAL_CIRCLE																										
HOUSING																										
ENQUIRIES																										
LAST_ONGOING_APP																										
TOTAL_ONGOING																										
ONGOING_CREDIT																										
TOTAL_CLOSED																										
CLOSED_CREDIT																										
MONTHS_BALANCE																										
SK_DPD																										
NET_PAID																										
PREV_CREDIT																										
DAYS_DECISION																										
DAYS_TERMINATION																										
NET_PAYMENT																										
PAYMENT_TIME																										
CNT_CHILDREN																										
DAYS_BIRTH																										
DAYS_EMPLOYED																										
DAYS_ID_PUBLISH																										
DOCUMENTS																										

Key Points:

- Strong correlation between DAYS_TERMINATION and DAYS_DECISION which indicates multicollinearity.
 - DAYS_DECISION removed
- 5 other additional correlations that show correlation, but provide completely different information.
 - None removed

Numerical Variables as Predictors



Key Points:

- Visual representations of the data were obscure and difficult to draw insights from.
- Instead, I chose to utilize inferential statistics to find statistical differences between target and non-target groups.