Loan Default Analysis: Are Unqualified Borrowers Being Targeted by Lenders?

Rohan Nitin Mahajan, Christy Sato, Chris Smith, Lennart Zeugner

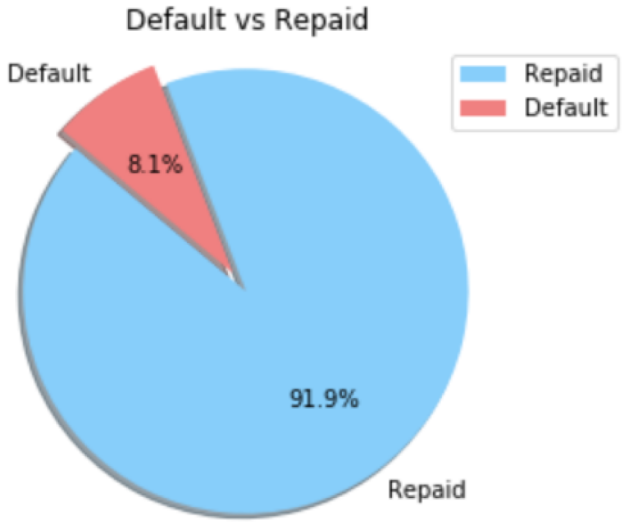
Problem Description

- Mortgage Crisis, 2008
- Loans lended to unqualified borrowers
  - What are the key indicators of unqualified borrowers?
- Prediction type
  - Classification

Dataset

- Home Credit Default Risk
- Kaggle
- Approximately 300,000 rows
- 120 variables

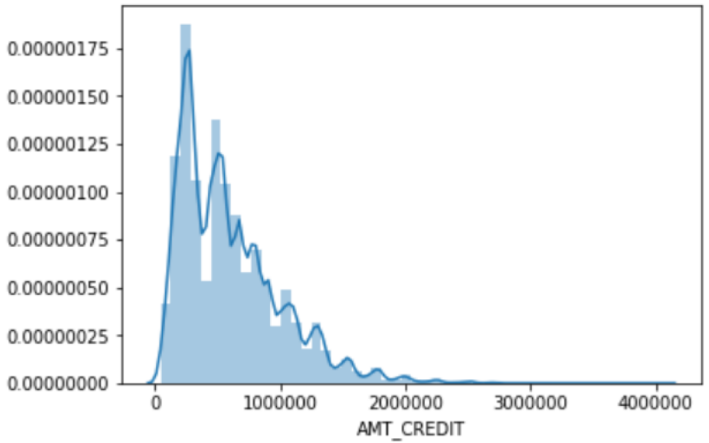
Descriptive Statistics - Target Variable



~ 92% of the loans in the dataset are repaid, while only ~8% default

The distribution of the target variable indicates an existing imbalance in our dataset

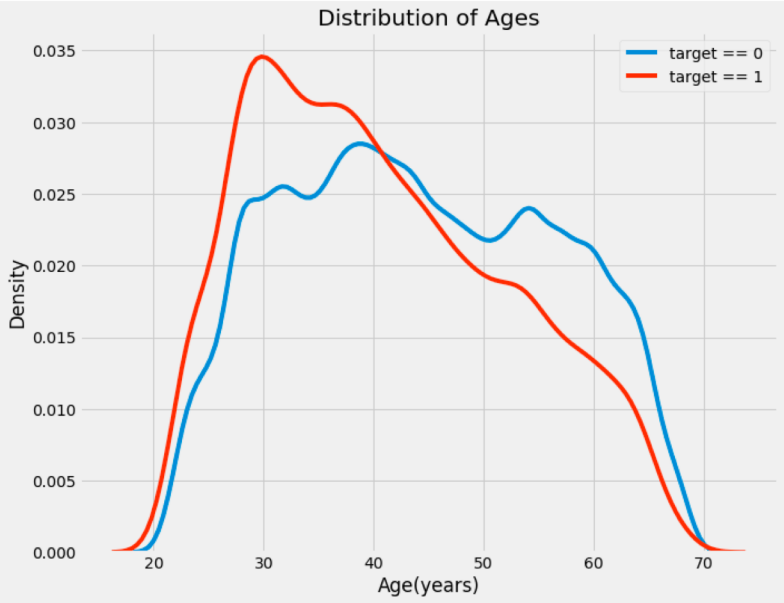
Descriptive Statistics - Total Credit Amount



Intuitively, it would make sense that total credit amount would be correlated with our target variable

The distribution of total credit amount is slightly right skewed with a **mean of ~ $599000**

Descriptive Statistics - Age and Target

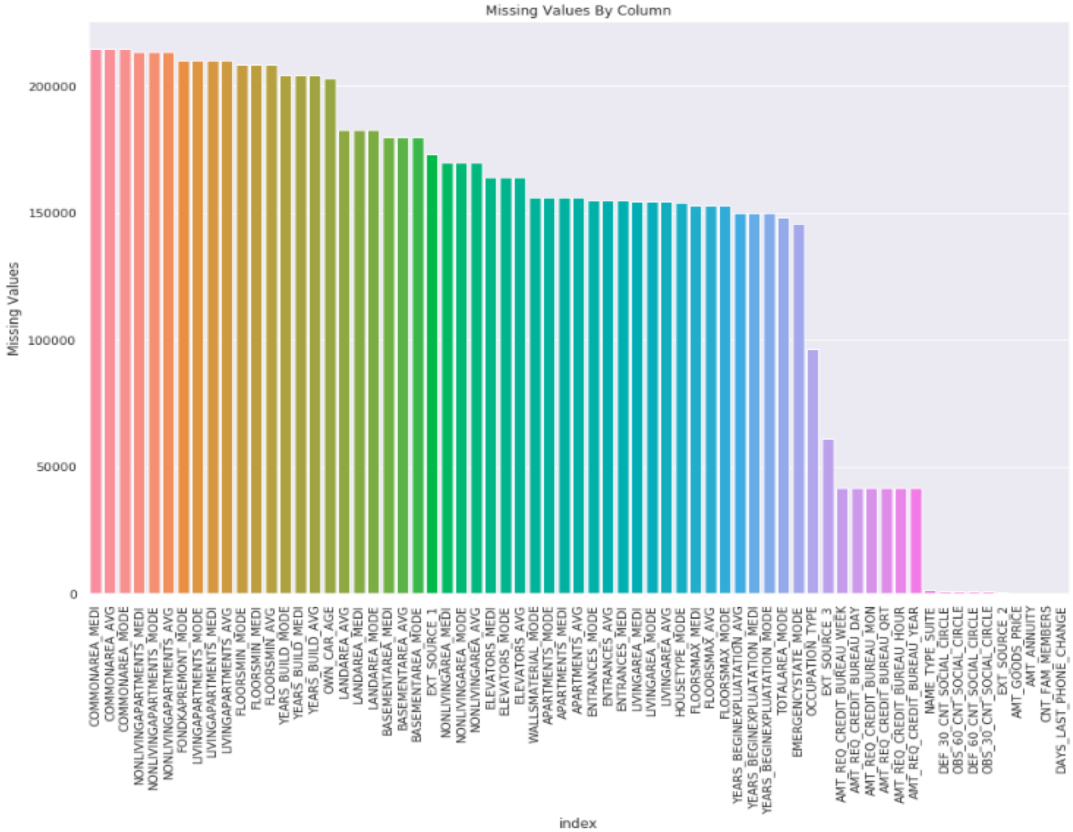


Those who defaulted, skewed toward younger ages

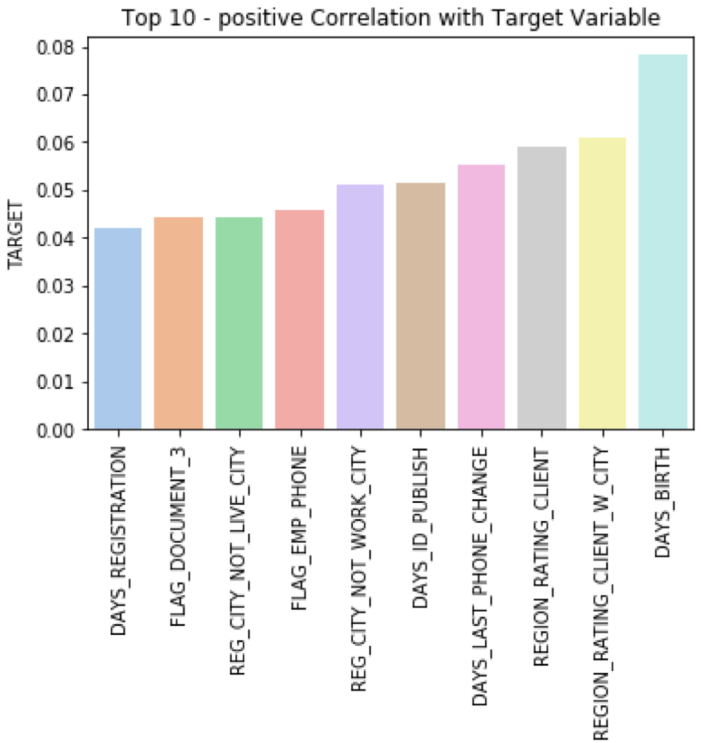
- Between the ages of 29 - 40

Descriptive Statistics - Missing Values

- Information about where client lives (common area, number of floors, number of entrances, etc)
- Number of inquiries to Credit Bureau
- # of observations of client's social circle



Exploratory Analysis - Positive Correlations



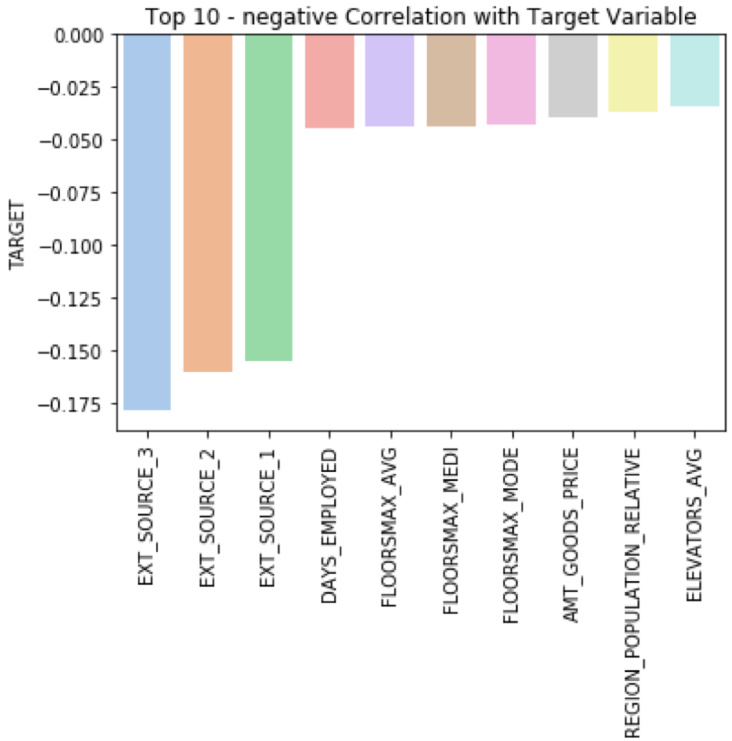
Top 3 Correlations:

1) Client Age

1) Lender rating of the region where client lives with taking city into account (1,2,3)

1) Lender rating of the region where client lives (1,2,3)

Exploratory Analysis - Negative Correlations



Top 5 Correlations:

1) Top 3 are normalized scores from external data sources. Metadata does not provide more information.

1) How long applicant has been employed at current job

1) Average number of floors of the building applicant lives in

Data Preprocessing & Feature Engineering

- All categorical variables were converted to dummy variables
- Created by udf called quick_dummies() to create several separate data frames
  - Used SK_ID_CURR as key for joins
- Dropped columns with more than 60 percent of data missing
  - This will be something we work on moving forward
- 224 features -> 163 features after cleaning

Model Overview

- Two different classification models
  - Logistic Regression
  - Using varying elastic net parameters and PCA k values
  - 60% training, 30% validation, 10% testing
- Infer what variables indicate a high probability of default
- Moving towards more complex model in the future
  - Random Forest
  - Clustering - K-Means

Models

- AUC Score - It can be thought as how good is the model at ranking the probabilities of the real labels

Models	Parameters	AUC Score (Val)
Model 1 Logistic Regress	maxIter = 10	.62
Model 2 Logistic Regress	maxIter = 20 StandardScaler PCA(k=10) elasticNet=.2	.70

	feature	pca_0	pca_1	pca_2	pca_3	pca_4	pca_5	pca_6	pca_7	pca_8	pca_9
0	CNT_CHILDREN	1.541711e-02	-5.000117e-02	3.218654e-01	-0.213280	5.172046e-03	-0.237248	1.851318e-02	-8.885754e-02	-1.545809e-01	4.191563e-02
1	AMT_INCOME_TOTAL	-3.056546e-01	-1.276710e-02	4.507604e-02	0.022044	4.027882e-02	0.144756	6.881739e-02	-1.537706e-02	1.852703e-02	-6.181973e-02
2	AMT_CREDIT	-2.707864e-01	-1.831456e-01	9.213557e-02	0.098138	-1.124033e-01	0.210548	3.010201e-01	-1.029287e-01	5.295026e-02	-3.647660e-02
3	AMT_GOODS_PRICE	-2.773363e-01	-1.812467e-01	9.073546e-02	0.090590	-1.066503e-01	0.209727	2.913812e-01	-9.717152e-02	5.060236e-02	-3.387634e-02
4	REGION_POPULATION_RELATIVE	-2.538078e-01	2.390151e-02	-7.348504e-02	0.146886	7.268901e-02	-0.245938	-6.329163e-02	-8.642057e-02	-4.687827e-02	6.678368e-02
5	DAYS_BIRTH	3.238538e-02	2.329705e-01	5.462648e-02	-0.314234	4.091830e-02	-0.123406	7.814457e-02	-6.043386e-02	1.988677e-02	3.914418e-02
6	DAYS_EMPLOYED	1.793977e-02	2.508883e-01	4.425831e-02	-0.109493	1.047082e-01	-0.079601	1.659372e-01	6.325123e-02	-8.635969e-03	1.309106e-02
7	DAYS_REGISTRATION	8.877898e-03	7.362227e-02	1.236905e-01	-0.211704	1.490340e-02	-0.008643	6.777354e-02	3.293175e-02	-1.158271e-02	-1.402577e-02
8	DAYS_ID_PUBLISH	2.044916e-02	8.243999e-02	-7.600381e-02	-0.049643	7.159724e-03	-0.026227	5.280377e-02	5.545623e-02	4.779687e-02	-4.152942e-02
9	FLAG_MOBIL	5.421011e-20	-1.734723e-18	3.469447e-18	0.000000	9.714451e-17	0.000000	-2.081668e-16	-2.775558e-16	-5.551115e-17	-1.804112e-16

Evaluation Metrics

- Three of the most popular classification metrics are AUC, Recall, and Precision

Best Performing Model - Model 2:

- ROC Area Under Curve (Test)
  - .72
- Recall (Test)
  - 0 = .91
  - 1 = .41
- Precision (Test)
  - 0 = .91
  - 1 = .03

Problems Encountered & Solutions

Problems:

- Imbalanced Dataset
- Missing Values
- Low recall score

Solutions:

- Random sample size of default values to balance dataset
- More advanced models to improve performance metrics