Problem/Overview:

Financial institutions (e.g. banks) refuse a countless number of loan/credit applications to people in need each day, but they have a cause for doing so. These institutions put themselves at monetary risk by lending credit to others, especially with no guarantee of the money's return. As a result, financial institutions are meticulous and stringent on the qualifications required to request large loans. This means that they will only approve credit applications to clients that have a strong record for repayment, as well as proven financial means to repay the loans.

Although this strategy minimizes the risk of a loan default, it prevents the newer clients from receiving the loans they need. This is especially true since the only way to obtain good credit is to consistently spend smaller amounts of credit over a longer period. Those individuals uninformed about credit are unable to build their credit early, resulting in difficulty applying for larger loans at a later age. As a result, some of these individuals are forced to become patrons for unreliable lenders.

To address that, some financial institutions have taken the risk and strived to expand the financial inclusion to those with insufficient or non-existing credit histories. These institutions utilize a variety of alternate data to gauge and measure their clients' risk profile.

My objective for this project will be to experiment with a sample of credit data provided by Home Credit Group, an international financial institution focused on lending to the population with little to no credit history. I will analyze the data and generate a predictive model that can reliably classify each client as high-risk or low-risk for credit loans.

Potential Clients:

The potential clients for this project include two groups: financial institutions and loan applicants. All financial institutions can utilize this data to expand their loans applicant pool. If the model can reliably predict high-risk and low-risk loan applicants via alternate data, then the risk of defaults can be minimized. With a larger approved applicant pool, financial institutions will earn larger returns. Additionally, when a client has a positive loan experience, they are more likely to return as a reoccurring applicant. As a result, this would yield higher returns both short-term and long-term for the financial institutions.

Loan applicants, especially applicants with near non-existent credit history, can utilize this information to see and understand alternative attributes that can help them develop towards becoming a low-risk applicant despite short credit histories. However, this information is applicable across as larger spectrum, essentially to anyone who is applying for a loan they may not currently qualify for with a normal financial institution's requirements. This may help the applicant reliably acquire the money they need.

Data Wrangling:

This dataset was provided by the Home Credit Group to a currently ongoing Kaggle Competition. The data consists of one training dataset supplemented with six other datasets of varying information. No exact time-frame is provided for, but there are records for up to 8 years of prior credit history relative to the time of application data. Overall, the model will be trained with over three-hundred thousand unique applicants, and any associate information. This data contains over a total of three hundred attributes, consisting of a heterogeneous mixture of both categorical and quantitative information.

The Kaggle Dataset was already cleanly organized with specified merging columns between each of the seven datasets. However, a complete merge would result in over three hundred attributes in one table. As a result, the first step was to perform some feature selection.

Feature Selection:

With seven datasets and hundreds of features, I chose to go through each dataset one at a time, starting with the training data. I performed the same routine of feature selection, feature engineering, and data cleaning to each of the datasets. However, I chose to leave out two provided datasets, bureau_balance.csv and POS_CASH_balance.csv, because I felt they did not provide any additional details.

With my limited domain knowledge, I could not accurately select features with importance to finance data. However, with my limited computational power, I could not try every single variable to calculate feature importance. As a result, I performed some manual feature selection based on an informational .csv file containing information regarding every feature. My focus was to remove columns that provided redundant information, then select specific features that made practical sense in relaying information of the clients' characters.

Additionally, I added columns that I would use in feature engineering. From these select columns, I performed some feature engineering and added it to the columns of interest. For each of the supplementary datasets, I grouped the data by their unique client IDs so that each client could be represented by a single row. I also reduced the credit history to the last 12 months (if available) to reduce the weight of credit history when developing my model.

Data Cleaning:

Next, I began to address the null values in each of the datasets, filling them with values I felt were appropriate (e.g. 0, 1, -1, etc.) based on their numerical representations. Additionally, I sought out odd values that did not make contextual sense (e.g. amount of days since employment = 365243) and attempted to understand them and replace them with their appropriate representations. For the extreme outliers, I chose to scale them towards the second largest extreme value instead of outright removing them because, with the provided data, I could not

determine if the value was a mistake. I removed a couple of columns from the datasets as I found them to be redundant with other columns.

Once I cleaned every dataset individually, I began to merge the supplementary datasets to the training dataset with a left merge to retain all the unique client IDs in the training data (in this case, I don't need information about the clients not in the training data). With every merge, I needed to fill in the null values in each column for clients that had no prior history in that database. One surprising merge was with the credit card balance history, where there were more than 220,000 clients without history. The null values were fill accordingly to represent missing data. The completely merged data had a total of 35 heterogenous columns.

Exploratory Data Analysis and Data Story:

The data analysis started with the correlations of the numerical columns in the fully merged training data. To prevent multicollinearity, I used a heatmap to find variables that were highly correlated (Figure 1).

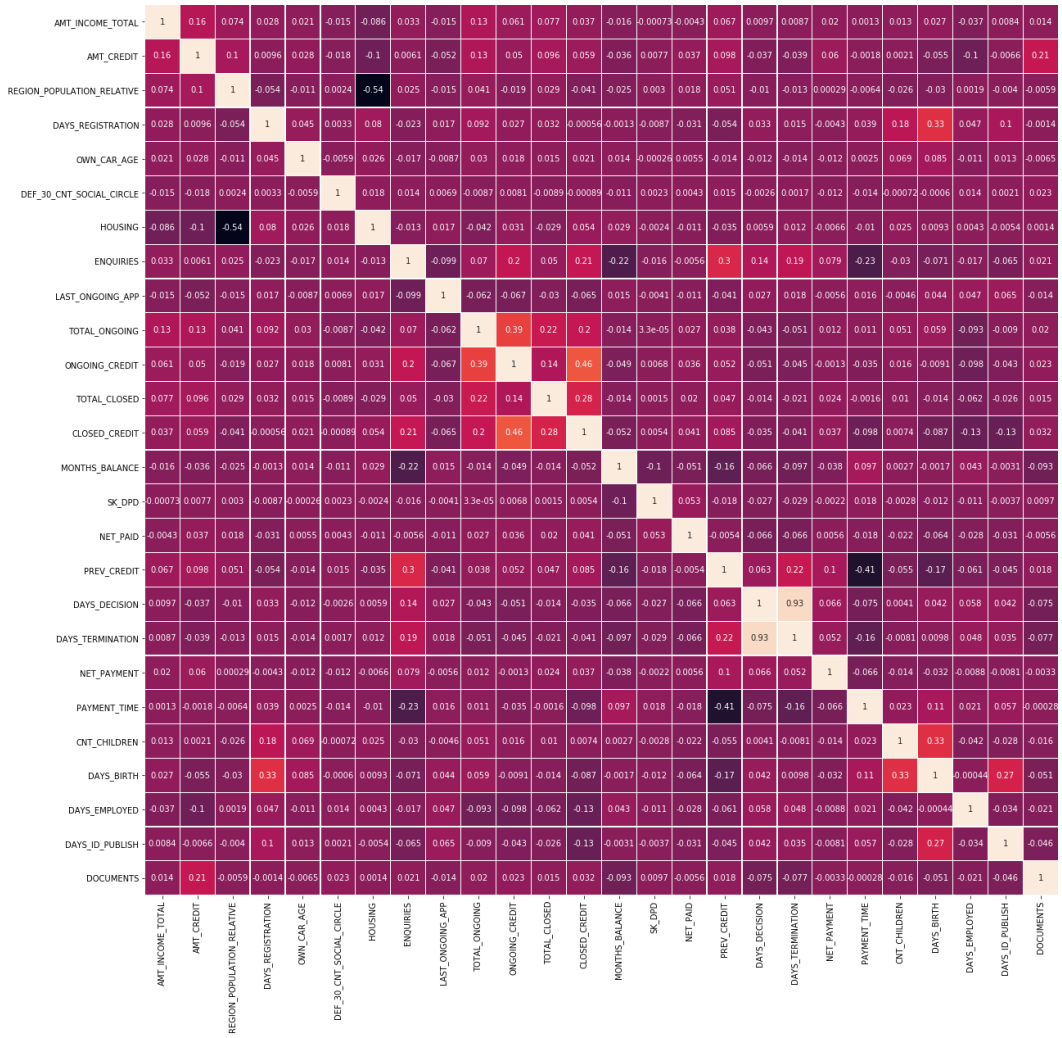

Figure 1. Seaborn heatmap of the correlations of the numerical columns in training data.

Of six notable correlations, I only acted on the correlation between DAYS_DECISION (when the last application was approved) and DAYS_TERMINATION (when the last approved application will end), removing DAYS_DECISION. For the remainder of the correlations, I plotted them on Seaborn's regression plot, and noticed that the distribution was random (Figure 2).

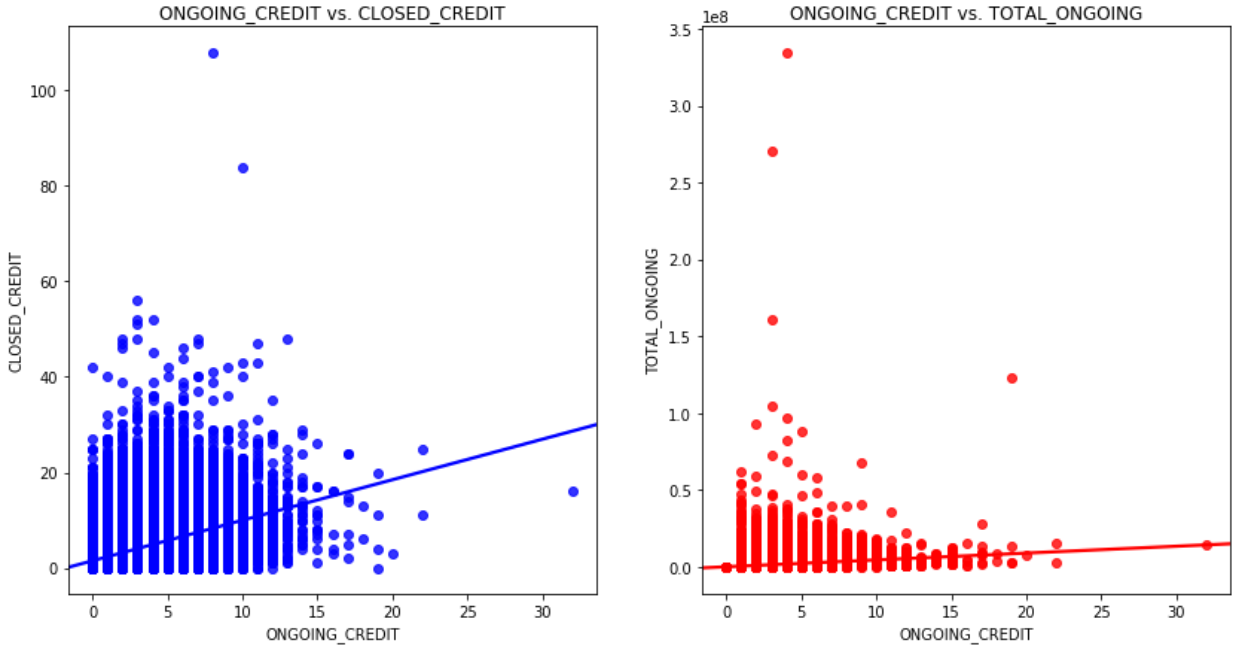

Figure 2. Regplot of 2 positive correlations between numerical columns.

Next, I attempted to utilize a series of boxplots to visually search for statistical differences between the target (difficulty repaying loans) and non-target groups (Figure 3). For the most part, most of the boxplots were obscure and difficult to draw any conclusive insights from. From the visualizations, I drew a couple of hypotheses for variables that could be strong predictors for the target group. To validate this, I used some inferential statistics.

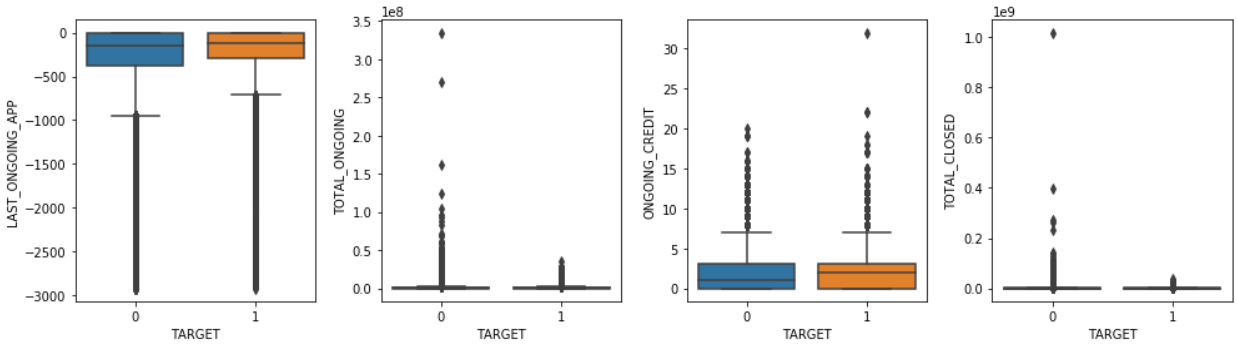

Figure 3. Boxplots of numerical columns split according to target and non-target groups.

I did not perform any visualizations for the categorical data as I felt it was difficult to accurately present without generating misrepresented plots.

Inferential Statistics:

Using a p-value of 0.05, I attempted to prove that there were statistical differences in each variable between the target and non-target groups. I applied the inferential statistics to both the numerical and binary categorical data.

For the numerical columns, there was strong support from both the two-sided t-test and frequentist bootstrap approach that there were no statistical differences between the two groups for OWN_CAR_AGE (0.205), ENQUIRIES (0.1), SK_DPD (0.62), and DAYS_TERMINATION (0.69). Unexpectedly, SK_DPD was a variable I flagged as potential correlation in my observations of the boxplots. Additionally, AMT_INCOME_TOTAL was also shown to have no statistical differences by the frequentist approach. However, a closer observation of the 95% confidence intervals showed that there was a small range for the total income for the non-target data, whereas the target data had a large spread that encompassed the smaller range.

For the binary categorical columns, I first used LabelEncoder from sklearn to convert the text to integers. From there, I applied the two-sided t-test (p-value of 0.05) and the frequentist approach to the five binary categories. Both tests support that there are not statistical differences between the two groups for the VALID_MOBILE category (if a valid mobile phone was provided).

Since I'm working with a heavily unbalanced dataset (# non-targets >> # targets), these tests can be biased. For all the columns that failed to show statistical differences, I chose to temporarily keep them and train my model with and without them.