

## Contents

Order of files:	1
<b>1. Executive Summary</b>	<b>3</b>
<b>2. Industry Analysis</b>	<b>3</b>
<b>3. Company Overview</b>	<b>4</b>
<b>4. Business Problem</b>	<b>4</b>
<b>5. Data Description</b>	<b>4</b>
<b>5.1 Data Source</b>	<b>4</b>
<b>5.2 Data Preparation</b>	<b>5</b>
<b>5.2.1 Data cleaning</b>	<b>6</b>
<b>5.2.2 Feature Engineering</b>	<b>7</b>
<b>5.2.3 Feature Selection</b>	<b>7</b>
<b>5.2.4 Up/down Sampling</b>	<b>7</b>
<b>6. Model Development</b>	<b>8</b>
<b>7. Model Evaluation</b>	<b>9</b>
<b>8. Conclusion &amp; Recommendations</b>	<b>12</b>
<b>Appendix</b>	<b>13</b>
Results from initial model development with only two data tables – application train and bureau.	13
Best hyperparameters used to tune the final models with for each dataset.	13
Figure 1: ERD of 7 tables used in the analysis	6
Figure 2: Distribution of target label	8
Table 1: Number of rows in each dataset after performing up sampling and downsampling	8
Table 2: ROC-AUC curve results from model development	10

## 1. Executive Summary

Small businesses and individuals with little to no credit history often struggle when attempting to receive financing. Alternative lending, where funds are provided by non-banks, are usually considered a good solution in these situations. Home Credit, an international consumer financial institution, primarily focuses on lending to individuals with little to no credit history. They determine the customer's ability to repay by using a non-traditional data-driven approach. In this report we analyze Home Credit's customer database and develop Gradient Boosting model, which can predict customer's repayment ability with an AUC score of 0.767. We chose ROC-AUC as a performance measure because it investigates the accuracy of the model's ability to separate positive from negative cases, that is repay versus difficulty to repay.

The following pages in this report provide background information on the lending industry, overview of the company – Home Credit, identification of the business problem, data preparation in order to use for development. We also include recommendations on how Home Credit and other alternative lending companies can implement this model to identify their customer's ability to repay a loan.

## 2. Industry Analysis

Credit comes in many forms including credit cards, automotive loans, mortgage, and personal loans. Each type of credit serves various purposes like buying a car or a house, paying for education or managing big purchases though affordable monthly installments. Regardless of the purpose, when being considered for a loan, financial institutions evaluate a customer's loan approval on several factors like credit score, current income, employment history, and assets. This is because lenders want to give low-risk loans and looking into a customer's credit and employment history can provide an indication of what kind of borrower they are.

When it comes to financial institutions lending money, credit score is one the main deterministic factors as it is influenced by many elements such as types of loans previously applied for, payment history, and balance on credit cards. These elements provide the lender a better understanding of the borrower's credibility. Therefore, a low credit score can automatically be a reason for loan disapproval.

Given the role credit score plays in loan approval, it is no surprise that individuals with little to no credit history often struggle to lend from traditional financial institutions. However, with integration of digital technology and use of data-driven solutions there has been a hike in use of non-bank and alternative lending institutions. Alternative lenders primarily focus on providing financial services to consumers with little to no credit history. According to a study conducted by Oracle's Digital Demand in Retail Banking with 5,200 consumers from 13 countries, more than 40% customers prefer non-banks as they provide better assistance with personal money management and investment needs<sup>1</sup>.

In the light of recent economic troubles across the world, consumers are looking for a lot more than receiving loans based on their credit history. They want personalized services catered to their financial needs. In which case non-banks and alternative financial intuitions seem to have more advantage.

---

<sup>1</sup> <https://www.businessinsider.com/alternative-lending-non-bank-industry>

### 3. Company Overview

Home Credit is an international consumer financial institution founded in 1997 in the Czech Republic. It serves 9 different countries with over 1.16 million customers across Europe and Asia. Home Credit consumer finance products comprise of:

1. Point of Sales (POS) loans – offered to customers at the time of purchase at physical or online store
2. Cash loans – offered for consumer goods or services with no connection to point of sales
3. Revolving loans (such as credit cards) – offered to existing customer for their purchases of goods and services

In addition, they also offer products such as current accounts, insurance, and car loans. In this report we only analyze POS, cash and revolving loan applicants<sup>2</sup>.

Home Credit is unique compared to many financial institutions as its primary focus is lending to those with little to no credit history. It serves underserved borrowers such as blue collar and junior white-collar segments of the market who earn regular income and therefore are less likely to access financing from banks and other traditional lenders.

The business model of Home Credit is based on non-traditional credit scores as they make use of alternative data such as telecommunications and transactional to provide consumer loans. They provide customized cash loans and cards which increase their customer loyalty and deepens the relationship with retailers and manufacturers<sup>3</sup>.

### 4. Business Problem

As a non-bank financial institution, Home Credit aims to be an innovative, technology-driven lender that consistently manages risk, funding and customer satisfaction. For this reason, they rely on data-driven platforms that enable risk-based pricing of loans even for customers with no credit history.

In this report, we aim to use various data inputs from Home Credit to perform statistical and machine learning techniques to identify a customer ability to repay a loan they applied for. Our model will be helpful in analyzing Home Credit customers' credibility, approve their loan application accordingly and ideally avoid any repayment risks.

### 5. Data Description

#### 5.1 Data Source

The dataset used for this project is from one the competitions launched by Kaggle where the challenge is to predict applicant's loan repayment capability. The main data provider for this competition was Home Credit, an international consumer finance provider.

---

<sup>2</sup> <http://www.homecredit.net/about-us/our-products.aspx>

<sup>3</sup> <http://www.homecredit.net/about-us.aspx>

Seven different data sources/tables were used for this project. They contain various types of information related to applicants as described below.

1. **application\_train**: This is the main table with applicant's demographic information including their profession, education type, marital status etc. It has details on the properties the applicant owns as well as for the ones they are applying for. It also contains the binary TARGET label of where they repaid the loan or not. This table has a single record for each of the applicants.
2. **bureau**: This table consists of information related to applicants' previous credits from other financial institutions. It has multiple records per applicant, one for each previous credit.
3. **bureau\_balance**: This is one of the largest tables and contains monthly data for each previous credit for every applicant. This table is used to get historical credit balance and then make decisions based on the client's history.
4. **previous\_application**: As the name indicates, this table has a client's previous applications for loan at Home Credit. It is used to get the total number of previous loan applications an applicant has submitted and what is the current state of those applications. Table might have multiple instances for each applicant because the applicant might have applied to more than one loan previously.
5. **POS\_CASH\_BALANCE**: This contains data related to the applicant's previous point of sale or cash loans with Home Credit.
6. **credit\_card\_balance**: This table contains monthly level credit cards balance for each applicant. One applicant can have multiple credit cards and each credit card can have multiple monthly entries based on the card's activation period.
7. **installments\_payment**: This table has information on an applicant's previous Home Credit loans and their installment payment patterns.

## 5.2 Data Preparation

Exploratory data analysis was performed as part of the data preparation process. We created the entity relationship diagram (ERD), shown in Figure 1, to define the relationships between all 7 tables described above. This ERD helped us understand the relationship and granularity of the dataset.

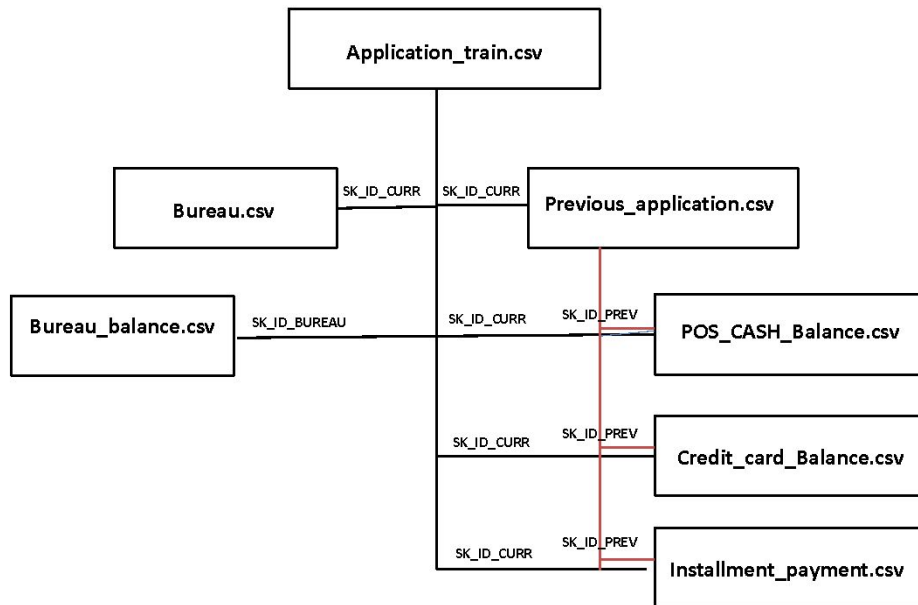


Figure 1: ERD of 7 tables used in the analysis

## 5.2.1 Data cleaning

Data cleaning was an ongoing process as we performed different methodologies to identify important features for our model. The steps below outline the data cleaning processes performed at different stages of our analysis.

1. Distribution of credit amount was highly left skewed, to achieve even distribution we performed log function transformation on *AMT\_CREDIT*.
2. Binning was done on different age groups to reduce the impact of outliers.
3. Features such as the number of *children* and *family members* where the value were high with low count, were replaced with more representable values.
4. All the features that were represented as days like *days after birth*, *days after first job*, and *days since last phone changed* were negative values. We converted these columns to positive values and converted them into years with two decimal places.
5. Some records had 365,243 in *days employed* column, which we believed was the default value for those who were never employed. This value was replaced by NaN and later imputed as well as with all other such (NaN) numerical values.
6. Missing values in numerical columns were imputed with different values (mean, most frequent, median) based on our understanding of the column.

7. Standard scalar was performed using `StandardScaler()` function to standardized numerical features.
8. One hot encoding was performed on all categorical features.
9. Columns that seemed to be less important, based on our experience and understanding of the business problem, were dropped.

### 5.2.2 Feature Engineering

As part of our initial modeling analysis we only used the *application\_train.csv* and *bureau.csv* tables for feature engineering.

1. Since *application\_train* table has demographical data only, we performed basic data cleaning like binning, log conversion, converting to categorical or Boolean data types, replacing unreasonable values with representable values as mentioned in Data Cleaning step above.
2. Count and Count\_Norm were performed for categorical features in *bureau* table.
3. Sum and Mean were calculated for numerical features in *bureau* table.

Later we included the features from the remaining 5 tables and performed these additional feature engineering techniques:

4. Count and mean operations on all the numerical values including the one hot encoded-categorical features. This helped us better understand applicant's credit history and past activity.

### 5.2.3 Feature Selection

To identify the most important features for our analysis, we performed multiple methodologies under feature selection.

1. Initially, when we only used the *application\_train* and *bureau* table, we checked for collinearity of the target label with all the features (including engineered ones) and identified top 30 features (top 15 positively correlated and top 15 negatively correlated) as part of our analysis.
2. However, after merging all the other tables and performing one hot encoding we had 157 features. Subsequently, we ran feature importance using `SelectKbest` and `ExtraTreesClassifier` and selected 126 features with the most importance. We also dropped any features that were highly correlated with each other.

### 5.2.4 Up/down Sampling

Figure 2 shows the distribution of target labels in the dataset. It is a binary classification that labels if an application repaid the loan or had difficulty.

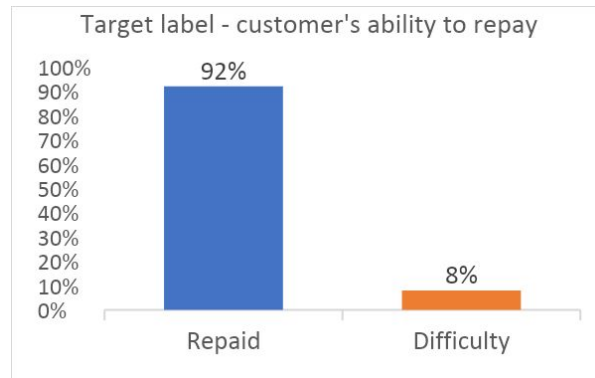


Figure 2: Distribution of target label

From the plot above we can see that our dataset was highly imbalanced with 92% of consumers who repaid the loan and 8% that had difficulty to do so. To avoid imbalance and selection bias while developing our model, we performed both up-sampling (using SMOTE technique) and down-sampling (using Near Miss) on the training set and ended up with three different datasets outlined in Table 1.

Table 1: Number of rows in each dataset after performing up sampling and down sampling

Dataset type	# of rows
Original	215, 257
Down sampling	34, 754
Up sampling	395, 760

## 6. Model Development

As part of model development, we initially ran three different models using the pre-processed data from *application train* table and *bureau* table. These models include:

1. XG Boost
2. Adaboost
3. Gradient Boosting Machine (GBM)

During our initial model development, we noticed that Gradient Boosting Machine performed best on all three datasets – up sampled, down sampled and original. [The results from our initial analysis are presented in the Appendix, Table A.](#) In addition, given the size of our dataset using XG Boost can be computationally expensive and time consuming for the purposes of this project<sup>4</sup>. Similarly, modelling Adaboost can make us vulnerable to overfitting and uniform noise<sup>5</sup>. For these reasons, when developing models with features from all 7 tables, we decided to continue with GBM alone.

As mentioned above, because our dataset was imbalanced, we developed GBM model using our final three datasets – original, down sampled and up sampled.

<sup>4</sup> <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/>

<sup>5</sup> <https://www.educba.com/adaboost-algorithm/>

Next, we performed hyperparameter tuning using Randomized Grid Search to identify the best parameters that fit our datasets. GBM has a total of 15 hyperparameters that can be tuned. In this report we only chose the top 9 parameters because it was computationally expensive to tune boosting methods. During the search process, we used 3-fold cross validation and evaluated using ROC\_AUC. Listed below are the top 9 hyperparameters we tuned, the values of the parameters for each dataset are outlined in the [Appendix, Table B](#).

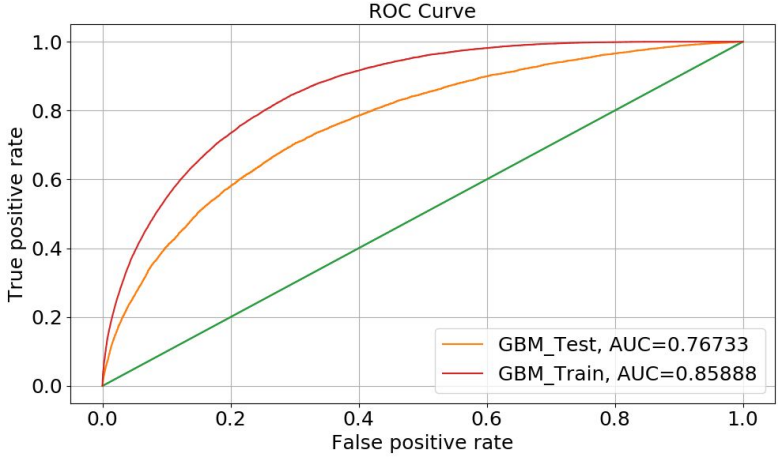
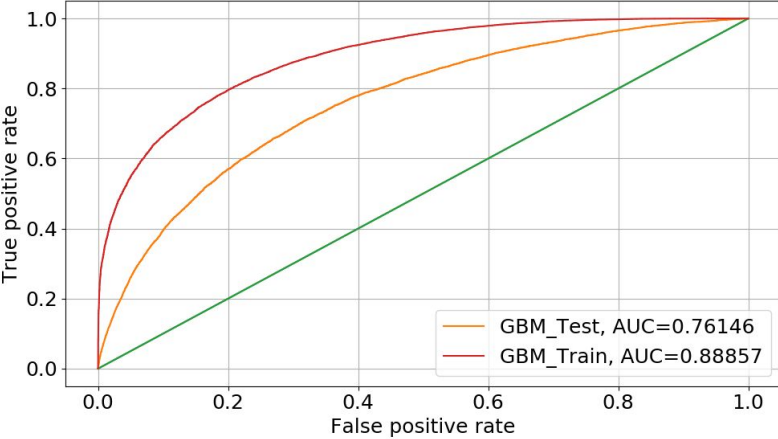
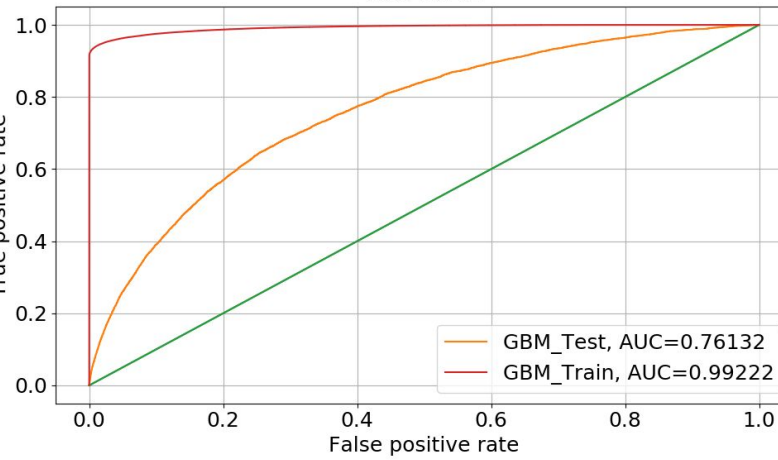
- loss
- learning\_rate
- n\_estimators
- max\_depth
- min\_samples\_split
- min\_samples\_leaf
- subsample
- random\_state

## 7. Model Evaluation

In order to evaluate the performance of our models and identify the best performing one, we chose ROC-AUC curve as our performance metric. This is because ROC-AUC investigates the accuracy of a model's ability to separate the two classes – repaid and difficult – independent of the frequency of the labels. That is, the higher the AUC the better the model is at distinguishing applicants that will repay the loan or not. Table 2 below shows the ROC-AUC plots for all three models.



Table 2: ROC-AUC curve results from model development

Dataset type	ROC-AUC Curve
Original	 <p>ROC Curve</p> <p>True positive rate</p> <p>False positive rate</p> <p>GBM_Test, AUC=0.76733</p> <p>GBM_Train, AUC=0.85888</p>
Down sampled	 <p>ROC Curve</p> <p>True positive rate</p> <p>False positive rate</p> <p>GBM_Test, AUC=0.76146</p> <p>GBM_Train, AUC=0.88857</p>
Up sampled	 <p>ROC Curve</p> <p>True positive rate</p> <p>False positive rate</p> <p>GBM_Test, AUC=0.76132</p> <p>GBM_Train, AUC=0.99222</p>

From the results above it is evident that the model with original dataset performed the best with an AUC score of 0.767, followed by down sampled dataset with a score of 0.7614 and finally the up sampled one with 0.7613. Overall, all models performed in a similar way with slight variations. They all performed much better than random guessing.

When comparing the AUC scores across the three models, they performed much better on all the training datasets. This is because the training datasets we imputed with missing values gave the model much more valuable information in order to predict accurately. Whereas, with test datasets we did not adjust for these changes.

## 8. Conclusion & Recommendations

Reduce the length of the application as we have noticed that not all questions were answered by the applications nor were they important for the analysis of the customer's ability to repay the loan.

- By reducing the length of the application, we believe customers will fill in the most important entries and we will get much better results – similar to trained set, where we imputed values for missing entries.
- Additionally, we can inform the application approval agents on what forms are important and for them to focus on receiving higher completion rates for items that matter.

Given that the Gradient Boost Machine is a type of decision tree model, further work can be achieved to graph the trees produced. This allows for Home Credit to retain the ability to explain the model and its reproducibility. This feature of decision trees gives model interpretation and shows where an applicant needs to improve their score, in order to be considered a lower risk to the lending institution.

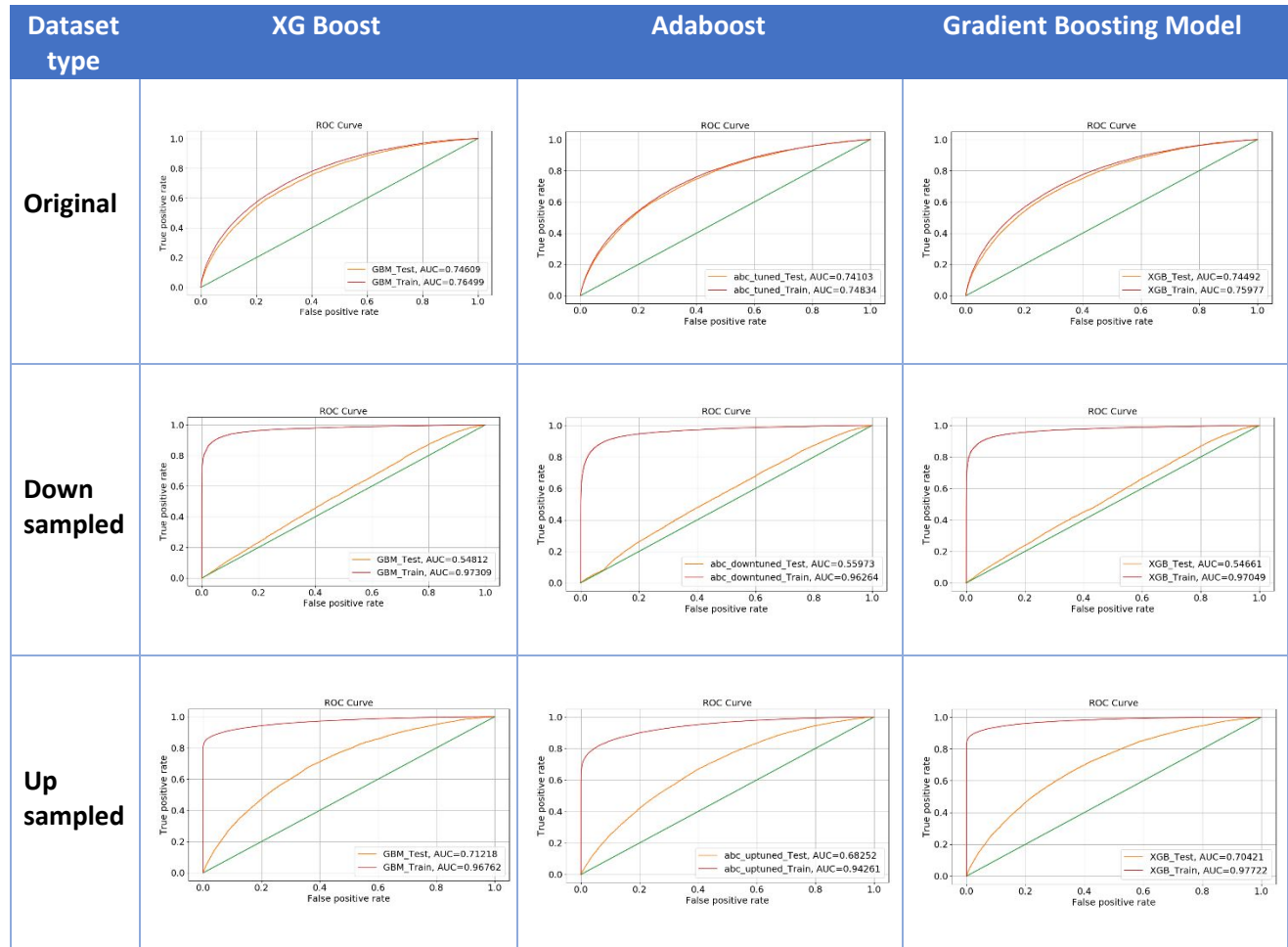
We would advise the lending institution to use this model as a way to price the cost of borrowing, rather than outright deny someone's application. With a 76% probability for the test case, our GBM model can meaningfully inform the lender whether they are dealing with a higher risk applicant. A function of the ratio of probability towards the Target (having difficulty to repay) could be implemented to offer a more appropriate cost of borrowing associated with the loan. This is currently done by the industry in a linear capacity, where they will offer a higher rate of borrowing for a higher risk client. The approach to using the model prediction probability would allow for a more confident decision towards how much higher the risk of borrowing should be presented to the applicant, and lower the risk of default to the lender in return.

Machine learning methods are powerful in capturing a non-linear relationship in the data. If we look at a traditional relationship for lending, a lender cares mostly about a ratio of total carrying debt vs the ability to service this debt from the income stream. With a GBM model, if we graph the Feature Importance, just using our initial Application + Credit Bureau data sets, we can see that the most important features influencing our target were [EXT\_Source\_2, EXT\_Source\_3, and EXT\_Source\_1]. These features represent External Data Sources linked to by the lender in their data set. They could either be credit score ratings, or a behavior rating on the propensity of an applicant to service their debt. Including these 3 data sources. builds a nonlinear relationship and allows for the credit issuer to have more stability in their confidence to lend.

This predictive model on the Home Credit's current portfolio of debt can be used to improve the prediction of default risk, and stress test for liquidity risk of the whole institution. Our GBM model can improve analysis and confidence of the whole portfolio. By identifying the notes at risk, Home Credit can set aside the needed capital to shore up their internal operations, and implement interventions for the individual notes at their local branches. This can positively influence their consumers to make their payments and improve the relationship for future repeat business.

## Appendix

### A) Results from initial model development with only two data tables – *application* train and *bureau*.



### B) Best hyperparameters used to tune the final models with for each dataset.

Dataset type	Hyperparameter values
<b>Original</b>	'subsample': 0.6, 'random_state': 21, 'n_estimators': 1875, 'min_samples_split': 2750, 'min_samples_leaf': 100, 'max_features': 17, 'max_depth': 11, 'loss': 'exponential', 'learning_rate': 0.01
<b>Down sampled</b>	'subsample': 0.7, 'random_state': 21, 'n_estimators': 1275, 'min_samples_split': 500, 'min_samples_leaf': 30, 'max_features': 17, 'max_depth': 11, 'loss': 'exponential', 'learning_rate': 0.01
<b>Up sampled</b>	'subsample': 0.9, 'random_state': 21, 'n_estimators': 925, 'min_samples_split': 500, 'min_samples_leaf': 90, 'max_features': 19, 'max_depth': 13, 'loss': 'deviance', 'learning_rate': 0.01