CREDIT RISK ANALYSIS

Analysis of risk for Residential Mortgage and Equity

MGMT 715 – Group 2
Aditee Bhattarai
Nam Dang
Zihan Huang
Jingxin Yao

Table of Contents

Executive Summary

The world faced one of its worst economic crises in 2008. One of the leading factors of the downfall was bad lending practices and subprime mortgages. In the year 2022, as a lingering effect of the Covid-19 pandemic followed by other events around the world, such as wars, the world economy is unstable again. In the USA, unemployment rates are higher, GDP is declining, and the real estate market is volatile. Thus, it is essential that financial institutions pay attention to their mortgage loan and home equity loan lending practices.

The scope of this report is to explore and analyze past data on residential mortgage loans and residential equity loans to get a better understanding of customer creditworthiness. We utilized analytical tools such as Cluster Analysis, Decision Trees, K-Nearest Neighbor, Support Vector Machines, and Deep Neural Networks. Among all the prediction, models we found that Deep Neural Networks give us the best results for both residential mortgage loan data set and residential equity loan data set.

By working in conjunction with cluster analysis and descriptive data analysis to identify crucial variables or characteristics, we were able to concentrate the results on predicting customer creditworthiness with high accuracy. By limiting the number of incorrect predictions, the company will be able to minimize the potential losses incurred and make its residential mortgage loan or residence equity loan process more efficient.

Introduction

Managing credit risk has been a crucial aspect of all risk management frameworks of financial institutions. Loans are instruments to facilitate the economy and can benefit both the lender and the borrower given the right situation. However, for every loan, there is an inherent risk of default that, if not managed properly, could lead to a devastating effect on the economy. For example, in 2008, the housing bubble burst caused millions of dollars in damage, many companies went bankrupt, and the unemployment rate hit its peak. There were several factors contributing to the crisis, but the growth of the subprime mortgage market is considered the main driver. Since then, financial institutions have strived to enhance their credit risk management to prevent another housing mortgage economic crisis from happening again using analytics modeling to manage loan quality and customer creditworthiness.

The world has witnessed many cases of loan defaults which led to severe damage to the economy, but the housing bubble burst in 2008 was considered one of the worst economic crises ever happened in history. The reason is because of its unexpectedness. In 2006, the US economy was reaching its best performance in every category. For example, the unemployment rate was low at 4.4% and wages were rising by 4.2%, both considered particularly good by US economic standards. The real estate market became the core of the economy, which drove the Dow Jones index to reach its all-time high in October 2007. Still, the Economists and financial experts were optimistic about the prospect of the economy at that time (Magdoff & Yates, 2010). However, just two years later, in June 2009, the unemployment rate doubled from 4.4% to 9.5% and continued to increase in the months to come. The rate could even be higher at about 16.5% if considering all the people who had to work part-time involuntary or got layoff recently and still looking for a job (Magdoff & Yates, 2010). Housing prices fell by a record 9.5% in 2008, to $197,100, compared to $217,900 in 2007 (Christie, 2009). The Dow Jones Industrial Average fell by 777.68 points in intraday trading, which was the largest point drop in history at that time (Amadeo, 2022). All these figures demonstrated a fast and devastating change in the economy in 2008, a change so sudden that it became exceedingly difficult for economists and financial institutions to react and mitigate the damage.

In response to perceived failings of deregulation and to address the credit losses, the Basel Committee on Banking Supervision as known as BCBS devised the initial Basel Capital Accord, which defines the minimum capital requirements for financial institutions, with the primary objective of minimizing credit risk. With further adjustments, BCBS endorsed Basel II, III, and IV (incomplete implementation) to increase the quality and quantity of capital. The advanced Internal Ratings Based (IRB) approach outlined in the Basel II and Basel III Accords allows banks to calculate their regulatory capital requirements based on the internal credit risk model (Edward et al., 2010). Banks and bank holding companies need three key indicators for their loan portfolios: PD (probability of default in the next 12 months), LGD (loss given default), and EAD (exposure at default). LGD as a critical parameter for credit scores has been the main modeling objective. And for decades, mathematic methodologies are applied to predict LGD, typically including the Ordinary Least Squares (OLS) Regression, Two-Stage Approach, Tobit Regression, Censored Gamma Regression, and Zero-inflated Gamma Model.

The introduction of even stricter regulatory systems like Basel II and Basel III has required financial institutions to improve their capability, accuracy, and efficiency in credit scoring. The development of statistics and computer science has brought novel approaches to credit scoring methods. As an interdisciplinary of statistics, artificial intelligence, and machine learning, data mining has been applied by researchers for classification and prediction. Data mining tasks can be divided into unsupervised and supervised models. Unsupervised models attempt to extract patterns that represent and describe distinct features of the data, while supervised models focus more on using input variables to classify data or predict values for output variables. Data mining techniques are very powerful tools to distinguish data features or predict future values, but data mining usually focuses more on model performance rather than explanatory or economical intuition (Dong-sup Kim, Seungwoo Shin, 2021). Also, financial institutions may require quick response and decision-making, and data mining could have low computational efficiency (Pérez-Martín, Pérez-Torregrosa and Vaca 2018).

Different data mining algorithms have been applied for default risk and credit scoring research. Feldman and Gross (2005) used classification and regression trees for credit scoring research. Butaru, Chen et al. (2006) compared Decision Tree, Logistic Regression, and Random Forest models for risk management in the credit card industry. Pérez-Martín and Vaca (2017) used Quadratic Discriminant Analysis and Support Vector Machines with linear kernel and found LSVM is the better model in predicting default but has low efficiency. Neural network models have been used in credit scoring by Atiya (2001) and Bahrammirzaee (2010). Sun and Vasarhelyi (2017) concluded that a deep neural network had better performance over traditional neural networks, decision trees, naïve Bayes, and logistic regression. Addo, Guegan et al. (2018) predicted loan default probability with a binary classifiers deep learning model.

In this project, we will be exploring two separate datasets: (1) the mortgage loan dataset, and the (2) home equity dataset. The goal of our analysis is to find which analytics model is most effective in understanding loan customers and predicting customer creditworthiness according to the two datasets. Our prior research has suggested that to determine the best model, we need to consider both predicting performance and economic efficiency. Therefore, multiple algorithms such as decision trees, support vector machines, K-nearest neighbor, and deep learning will be applied and compared to find the optimal model.

Mortgage Loan Sample Dataset

The initial dataset used for analysis is a dataset taken from the mortgage loan company that contains the data of 50,000 customers (observations), including time stamps, real estate attributes, economy-relevant parameters, mortgage loan statutes, etc. After the process of data aggregation, 8,146 unique customers and 25 mortgage attributes are kept. There are 4 nominal variables and 21 numerical variables in the dataset.

Table 1: Data Attributes and Type

#	Variable Name	Type	Description
1	MID	ID	Unique mortgage customer identification number
2	time	Numerical	Time stamp (deidentified) of observation
3	origin_time	Numerical	Time stamp (deidentified) for mortgage origination
4	first_ob_time	Numerical	Time stamp (deidentified) for first mortgage customer observation
5	maturity_time	Numerical	Time stamp (deidentified) for mortgage loan maturity
6	RE_Type	Nominal	Real estate type. Condominium = “CO”, planned urban development = “PU”, single-family home = “SF”, Other = “OTH”
7	investor	Nominal	Customer is investor borrower = 1, otherwise = 0
8	balance_orig_time	Numerical	Outstanding balance amount (USD) at mortgage origination time
9	FICO_orig_time	Numerical	Middle FICO score at mortgage origination time
10	LTV_orig_time	Numerical	Loan-to-value ratio at mortgage origination time, in %
11	Interest_Rate_orig_time	Numerical	Interest rate at mortgage origination time, in %
12	hpi_orig_time	Numerical	House price index at mortgage origination time, base year = 100
13	status_time	Nominal	Default (1), and Non_Default (0) observation at observation time
14	LTV_time_mean	Numerical	The average of LTV per customer
15	LTV_time_sd	Numerical	The standard deviation of LTV per customer
16	balance_mean	Numerical	The average of outstanding balance amount per customer
17	balance_sd	Numerical	The standard deviation of outstanding balance amount per customer
18	interest_rate_time_mean	Numerical	The average of interest rate per customer
19	interest_rate_time_sd	Numerical	The standard deviation of interest rate per customer
20	hpi_time_mean	Numerical	The average of house price index per customer
21	hpi_time_sd	Numerical	The standard deviation of house price index per customer

22	gdp_time_mean	Numerical	The average of Gross domestic product (GDP) growth per customer
23	gdp_time_sd	Numerical	The standard deviation of Gross domestic product (GDP) growth per customer
24	uer_time_mean	Numerical	The average of unemployment rate per customer
25	uer_time_sd	Numerical	The standard deviation of unemployment rate per customer

Descriptive Statistics

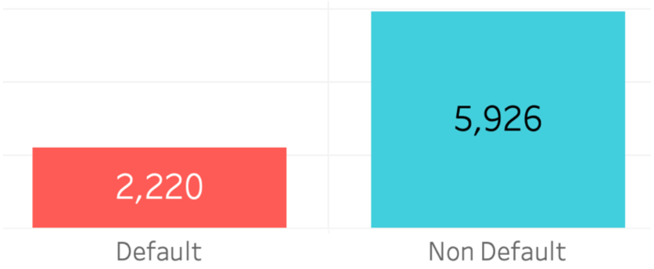
The descriptive analysis summarizes the data points in the dataset. The summary of the customer data is captured in the form of tables (see *Table A1* and *A2* in Appendix A).

Data Exploration and Data Visualization

To observe relations among variables and understand the customer situation of the mortgage loan, we conduct data exploration with supporting visualization before conducting data analysis.

Class imbalance of the target variable suggests that the number of observations for each level in the target variable is not equal, which could influence the model's capability of classifying correctly. In the mortgage dataset, the target variable 'status_time' had 3 levels: 'non-default/non-payoff', 'default', and 'payoff'. Although class imbalance exists, 'Default' is not the level with the least observations.

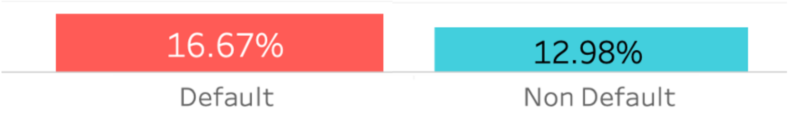
Figure 1: Target Variable Scale Before Conversion



Therefore, we decided to combine 'no action' and 'payoff' to 'Non_Default', the final levels are shown in Figure 1. After conversion, the number in 'Non-Default' category is 5,926. Handling class imbalance will be discussed in further details at the analysis section.

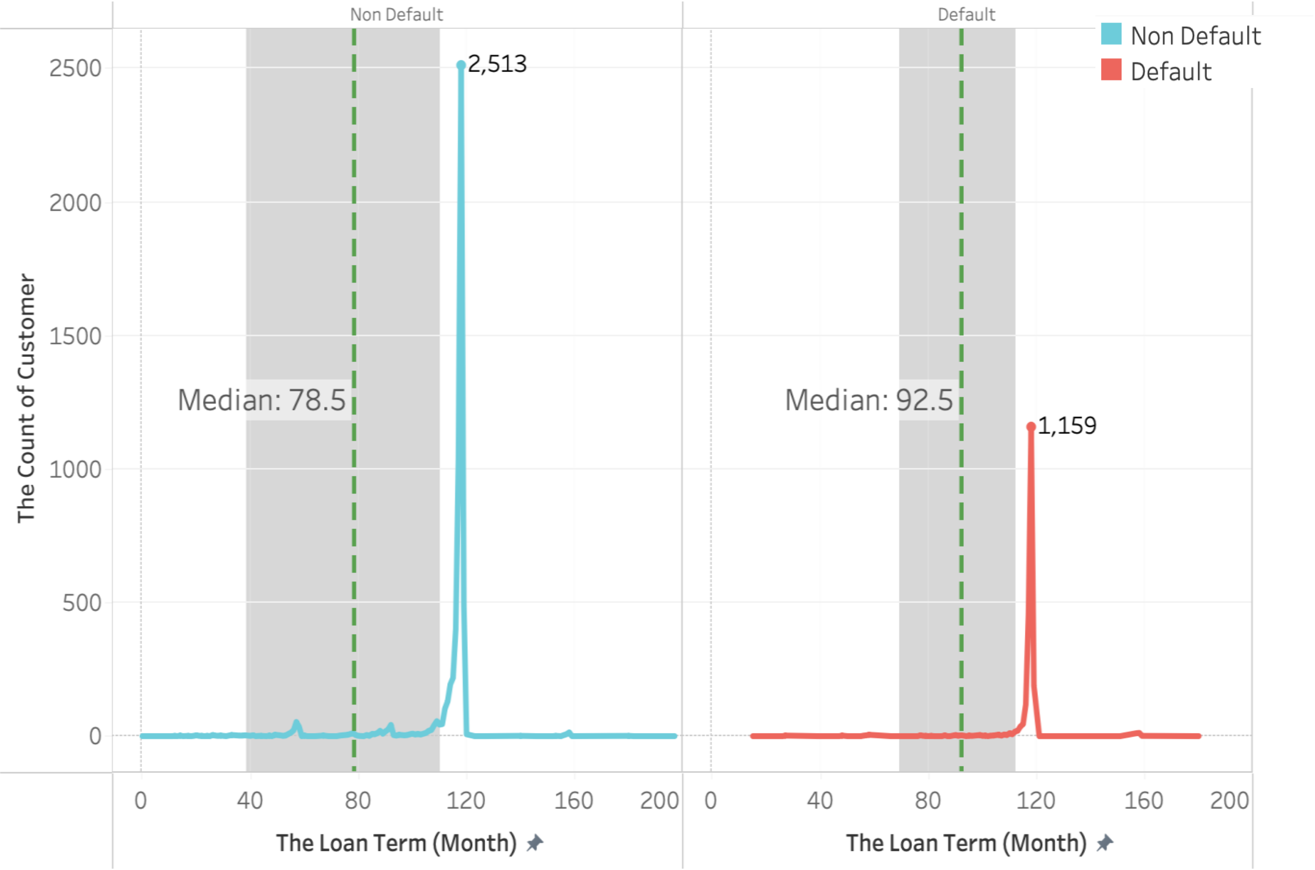
Customers who are investors are more populated in the Default group. Figure 2 displays that 16.67% of customers who defaulted are investor borrowers. This indicates that investor borrower has a higher risk of loan default.

Figure 2: The Bar Chart of Investor Distribution by Groups



For all customers, we observed that the median loan term in the non-default group is 78.5 months (6.5 years). On the other hand, the median loan term in the default group is 92.5 months (7.7 years). One thing in common is that customers in both groups paid off the loan before 120 months. It suggests that most customers would like to pay off their loan before 10 years. Figure 3 below visualizes our findings in detail.

Figure 3: The Distribution of Loan Terms by Groups



Moreover, data visualization plays an important role to help us understand how the finance-related factors are during the observing period and whether they have associations with customers who would default or not. We analyzed all relevant factors about the mortgage loan, and Figure 4 shows that the average GDP growth during the observation time is lower for default customers compared with non-default customers. However, in Figure 5, default customers have higher values of loan-to-value (LTV) ratio, interest rate, and house price index (HPI), with 87.53%, 7.76%, and 209.21, respectively.

Figure 4: The Average Variation of GDP Growth

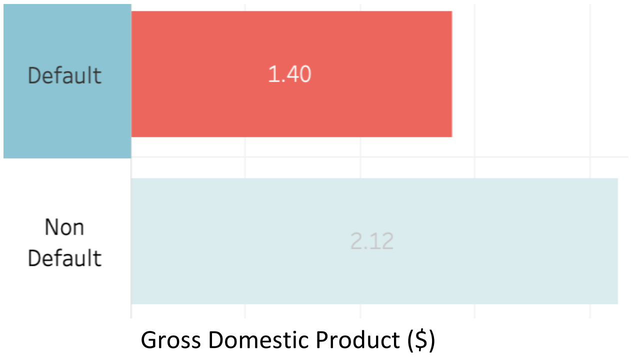
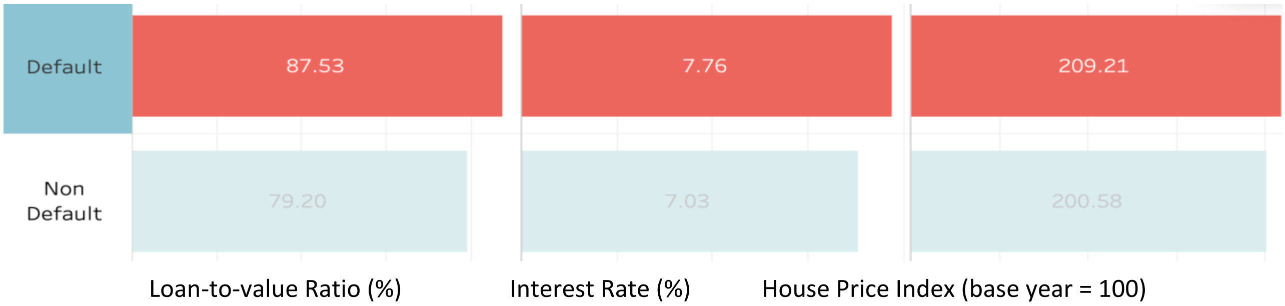


Figure 5: Average Variations among Other Variables



Data Preparation

In the mortgage dataset, each variable calculating standard deviation had 763 missing values. These missing values are generated because there are customers with only one observation and their standard deviations were recorded as missing. We imputed those missing values with 0, which is the standard deviation for a single number.

Outliers are data that are significantly different from other observations. We used the z-score method and found 12 numerical variables with outliers. Outliers could cause higher mean values and influence statistical analysis. However, we understand that for our business case, presence of outliers in certain variables is reasonable as banks may occasionally issue larger loan for priority customers. In our cases, some customers are considered outlier, but their presence could have economical meaning. Therefore, we decided not to remove the outliers.

We also identified variables ‘interest_rate_time_std’, ‘balance_std’ and ‘Interest_Rate_orig_time' were severely sparse that over 20% of the observations were 0. Therefore, we redefined these variables so that every observation not equal to 0 was recorded as 1.

Data Analysis

1. Decision Tree

In our literature review, it is common among prior research to choose decision tree model for default prediction and credit scoring. The advantage of using a decision tree is that it is self-explanatory and easy to interpret. It is robust to irrelevant, redundant features, outliers, and missing values. Decision tree models are also easy and fast to train and test.

For the mortgage dataset, we performed hyperparameter tuning to help control the learning process to optimize performance. To handle class imbalance, we chose class weighting which will not change the size of the dataset. The target variable ‘status_time’ had two levels, and we assigned a weightage of 0.689 to ‘payoff/no_action’ and a weightage of 1.824 to ‘default.’

First, we wanted to conduct a grid search over the complexity parameter that is associated with the smallest cross-validation error. We tuned the model with the gird search, over a control object. The metric we chose to find the optimal model is accuracy. The tuned model gave us the optimal complexity parameter of 0.00066, a small complexity parameter value suggesting that the penalty of having many splits was low, and the decision tree was large. The performance measure values are in Table 2:

Table 2: Weighted Decision Tree Performance

	Training	**Testing**
Accuracy	0.80	0.77
Kappa	0.55	0.50
Sensitivity	0.82	0.79
Specificity	0.79	0.77
Precision	0.60	0.56
Recall	0.82	0.79
F1	0.69	0.66

Our decision tree model had good accuracy on both the train and test datasets. The kappa value for the test dataset showed moderate agreement. After class weighting, sensitivity and specificity were improved, and the model could classify both types of customers well. However, precision which is the proportion of predicted default customers that are actual default is just moderate.

2. **Support Vector Machines**

Support Vector Machines are applied often in prior research and are often considered one of the most powerful machine learning algorisms. (Pérez-Martín, 2018) Support Vector Machines are robust to irrelevant and redundant variables and noises, and not very prone to overfitting. The SVM is effective with high-dimensional data. The weakness is that the model is complex and can take a lot of time to train. The model does not give predicted probabilities and it can be difficult to interpret. We also built SVM models to compare with other models.

For the mortgage dataset, we performed hyperparameter tuning. A center and scale normalization was applied in the model. We conducted a random search over the cost and sigma value. In Support Vector Machines, cost C acts as a regulation parameter associated with training error that with larger cost, training error is minimized. Sigma controls the level of non-linearity introduced in the model. Next, we set up a control object. The class imbalance was taken into consideration and handled with class weighting. We also chose accuracy as the metric. The method we chose was radial SVM. The optimal model had a cost value of 8 and a sigma value of 0.0307118. This suggested that our model focused on maximizing the margin while the decision boundary is highly nonlinear. The model performance is shown in Table 3**:**

Table 3: Support Vector Machines Model Performance

	Training	**Testing**
Accuracy	0.86	0.84
Kappa	0.63	0.58
Sensitivity	0.67	0.65
Specificity	0.93	0.91
Precision	0.79	0.73
Recall	0.67	0.65
F1	0.72	0.69

The SVM model had high accuracy over both the train and test datasets. Again, the kappa value showed moderate agreement. The model had high specificity in that it can classify customers that do not default correctly. The precision was over on the test dataset, which meant that the model could predict customers that have defaulted moderately. However, the recall was lower so that the model had a lower capability of classifying default customers. The model had similar performance over the train and test datasets and was balanced.

3. **k-Nearest Neighbors**

kNN（k-Nearest Neighbors）is a supervised learning method that can class the target observation according to the distance between it and other nearest points which are defined by the Euclidean Distance: $dist_{ed}(X,Y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$.

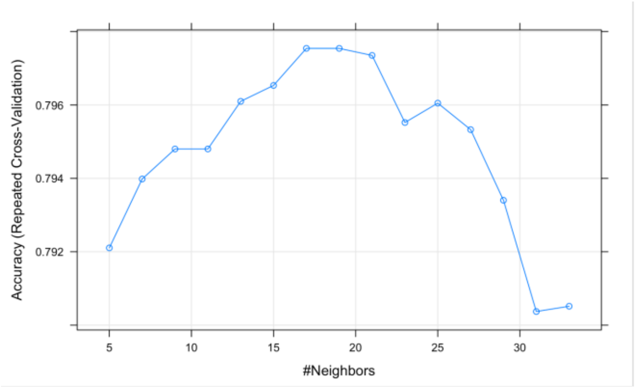
Considering the kNN sensitivity to missing values, redundant variables, irrelevant variables, outliers, and non-normalized values, all data has been transformed.

First of all, we performed min-max normalization on the numerical to make the value fall between 0 and 1. The formula for min-max normalization is:

$$x' = \frac{x - \min{(x)}}{\max(\quad) - \min{(x)}}$$

To tune the model, we performed hyperparameter tuning to help us get the best possible k value. In this process, the training control object is set. The final k value used for the training model was 17. This plot gives a general view of the training accuracy changing with different hyperparameter values, k (number of neighbors).

Figure 3: K Value and Accuracy



Next, we practiced the tuned training model with an accuracy is 0.7975. And then we predicted the target variables based on the tuned model, resulting in an Accuracy of 0.792 and the Kappa of 0.4189. However, the potential problem in the mortgage loan data is the default customers compose a minority which can result in an unfair prediction. In terms of the imbalanced classification, we resampled the targeted variables by setting the class weight value to prepare a balance training dataset. Finally, the class-weighted model was used to predict default customers on both the training and testing datasets. The model performance is shown in Table 4:

Table 4:k-NN Model Performance

	Training	Testing
Accuracy	0.82	0.79
Kappa	0.49	0.42
Sensitivity	0.52	0.46
Specificity	0.92	0.92
Precision	0.72	0.67
Recall	0.52	0.46
F1	0.61	0.55

Overall, the kNN model had similar performance and had high accuracy both on training and testing datasets. High specificity can classify Nondefault customers correctly whereas the low sensitivity and recall suggested that it is difficult to classify Default customers correctly. Besides, the kappa of the kNN model indicates a moderate model agreement. Considering the scenario of predicting default customers, the sensitivity should be highlighted as the evaluation value of the kNN model.

4. **Deep Learning**

Deep learning is a set of machine methods based on artificial neural networks which were inspired by biological systems. The difference between deep learning and normal artificial neural networks is that deep learning involves multiple hidden layers in the model. Deep learning was applied in default prediction and was proved to be a strong classification model over the other machine learning algorithms.

Deep learning model can be influenced by redundant variables, which means the information of the variable is contained in the dataset by other variables or variables. Redundant variables could increase the complexity of the model, and lower the accuracy and computational efficiency. We calculated correlation between predictor variables and found that 6 variables ('time', 'uer_time_std', 'gdp_time_std', hpi_time_std', 'hpi_orig_time' and 'balance_orig_time') had high correlations above 75% and removed them. Also, we applied min-max normalization before the model was tuned

To build a deep learning model, we used hyperparameter parameter tuning which searched over 10 hidden layers to find the optimal number of nodes. We also used the early stop to avoid overfitting and improve effectiveness. For the class imbalance problem, we used the SMOTE algorithm for oversampling. The model was used to predict default on both the train and test datasets. The model performance is shown in Table 5:

Table 5: Deep Learning Model Performance

	Training	**Testing**
Accuracy	0.82	0.81
Kappa	0.64	0.61
Sensitivity	0.85	0.83
Specificity	0.79	0.79
Precision	0.80	0.79
Recall	0.85	0.83
F1	0.82	0.81

The deep learning model had high accuracy and good agreement on both the train and test datasets. The SMOTE algorithm successfully handled class balance as both sensitivity and specificity were high, suggesting that the model was classifying default and non-default customers correctly. The model also had high precision and F1 value, which meant that the model could predict default customers that actually defaulted properly. The model had overall similar performance over the train and test dataset, and we believe it was balanced.

Home Equity Loan Dataset

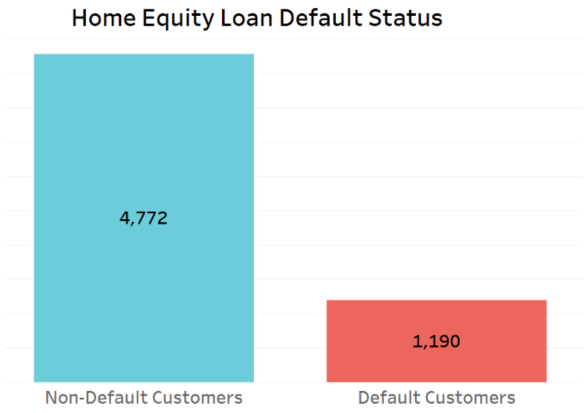
The Home Equity Loan Dataset has a total of 5964 rows and 14 columns. The types of variables in the data are shown in Table 6.

Table 6: Data Attributes and Types

#	Variable Name	Type	Description
1	HEID	ID	Home equity customer identification number
2	Loan_Amount	Numerical	Amount of loan requested (USD)
3	Mort_Bal	Numerical	Amount due on existing mortgage
4	Home_Val	Numerical	Value of current property
5	Reason_HE	Nominal	Stated reason for loan
6	Occupation	Nominal	Occupation of the customer
7	YOJ	Numerical	Number of years at current job
8	Num_Derog	Numerical	Number of major derogatory reports
9	Num_Delinq	Numerical	Number of delinquent credit lines
10	CL_Age	Numerical	Age of oldest credit line (months)
11	Num_Inq	Numerical	Number of recent credit inquiries
12	Num_CL	Numerical	Number of credit lines
13	Debt_Inc	Numerical	Debt-to-income ratio
14	Default	Nominal	Status of loan default/payoff

The variable HEID is a unique identifier and has been dropped from the dataset as it is not relevant to the scope of our analysis. Our target variable is Default. Default consists of 2 class levels 0 and 1, where 1 indicates default and 0 indicates no default or payoff.

Figure 4: Class Imbalance in Target Variable



We have also observed that there is a class imbalance in our target variable, Default. Only 20% of the observation have “1” under “Default”, which means only 20% of the customers in our dataset have defaulted on their loans. We can see the representation in Figure 6. We will need correct class imbalance during data pre-processing in order to get accurate results.

Descriptive Statistics

The descriptive analysis summarizes the data points in the dataset. The summary of the customer data is captured in the form of tables (*see Table A3 and A4 in Appendix A*).

Data Exploration and Visualization

Through initial data exploration, we found that most of the variables have means higher than the median. This shows that our data is positively skewed, which will need to be corrected through standardization and/or normalization during data preparation.

An interesting pattern that we found during our initial exploration is that the Debt-to-Income ratio, as we see in Figure 7, is slightly higher in defaulters. The number of delinquent credit lines is proportionately higher in defaulters, which can be seen in Figure 8. Default is also higher among people who have a higher-than-average number of derogatory reports.

Figure 5:Average Debt to Income Ratio

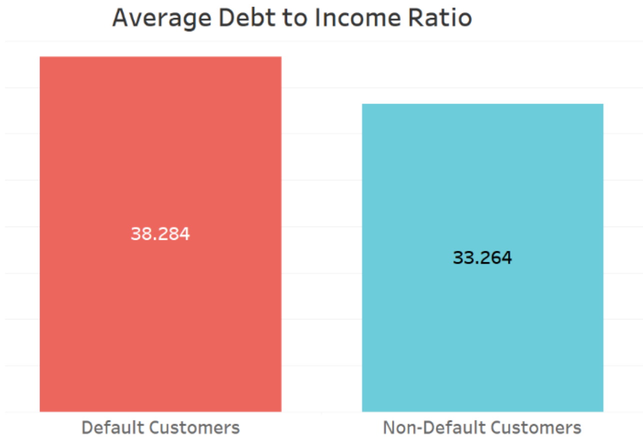
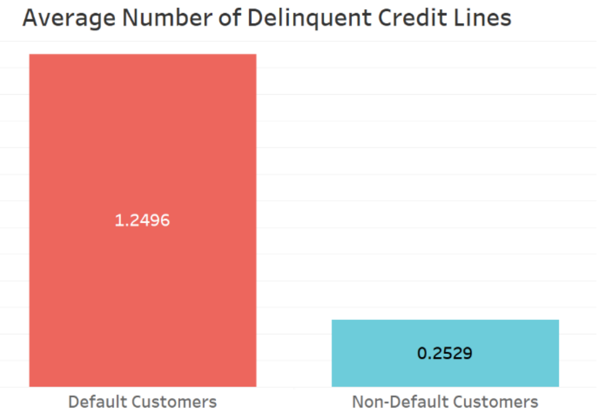


Figure 6: Average Number of Delinquent Credit Lines



We also observed that Credit Line Age along with Years on Job is lower among defaulters from Figure 9 and Figure 10, respectively.

Figure 7: Credit Line Age

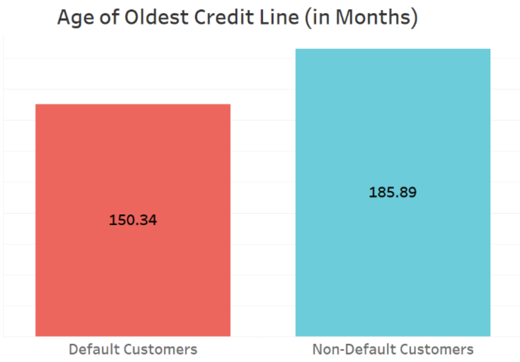
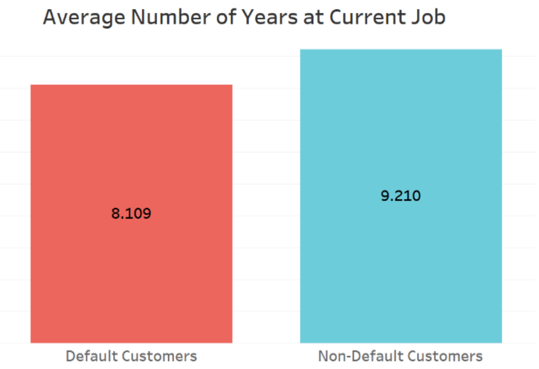


Figure 8: Average Number of Years on Job



Default seems proportionately higher in the Occupation: Sales. Around 35% of the people in Sales have defaulted on their Home Equity Loans. Default seems to be lowest among Professional/Expert, which is around 16%.

Data Preparation:

The data frame was also checked for duplicate values. There were 2 duplicate values, which have been removed from the data.

One of the most challenging problems with the Home Equity dataset is that it is plagued with missing values. After checking for missing values using the complete cases method, we observed that there are 2620 entries with at least 1 missing value, which is equivalent to 43% of total observations in the data set. Table 7 shows the number and proportion of missing data in each variable.

Table 7: Missing Values in Home Equity

Variable	Missing	Percent of Total
Loan Amount	0	0.00%
Mortgage Balance	518	8.69%
Home Value	112	1.88%
Reason for Home Equity	252	4.23%
Occupation	280	4.69%
Years on Job	518	8.69%
Number of Derogatory Reports	710	11.90%
Number of Delinquent Credit Lines	582	9.76%
Credit Line Age	310	5.20%
Number of Inquiries	511	8.57%
Number of Credit Lines	223	3.74%
Debt to Income Ratio	1267	21.24%
Default	2	0.03%

Due to high missing values across all variables, we proceed cautiously to handle these missing values. Firstly, we investigated the variables case-by-case to determine if there is any pattern in missing data. For example, we suspected that customers that have high-value loans might be more hesitant to provide information about their loans. Fortunately, we did not find any significant difference between the frequency of missing value in high-value loans compared to lower-value loans. We concluded that there is no clear pattern behind the distribution of missing values in our dataset.

Secondly, we want to determine if we have a case of missing at random (MAR) or missing completely at random (MCAR). MAR is defined as missing values of a Y variable depending on the X variable, but not on Y. MCAR, on the other hand, means observed values of Y are truly a random sample of all Y values (Hair et al., 2019). Categorizing the missing values as MAR or MCAR is important because, for each case, the approach to handling missing values would be different to minimize bias in our dataset. To test for the level of randomness, we performed a t-test of missingness, which is testing the differences between cases with missing data versus those not missing data. In our case, the mean of Loan Amount will be compared between the two testing cases, since it is the only variable that contains 0 missing values. The result of the test is shown in Table 8 below:

Table 8: Result of t-test of missingness for Home Equity Dataset

Mean	17,908.54	19,154.17
Variance	134,271,897.75	118,239,736.57
Observations	2,599.00	3,365.00
Hypothesized Mean Difference	-	
df	5,403.00	
t Stat	(4.22)	
P(T<=t) one-tail	0.00	
t Critical one-tail	1.64	
P(T<=t) two-tail	**0.00**	
t Critical two-tail	1.96	

The result of the t-test of missingness has a very low P-value (~0.00002). Therefore, we reject the null hypothesis, or in other words, the means of Loan Amount is significantly different between cases with missing data versus not missing data. We concluded that the level of missingness in our variables is MAR.

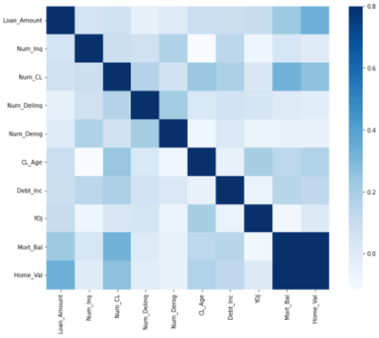
The best remedy to handle MAR is using model-based imputation. For this dataset, we decided to use the MICE package, which is very helpful in our case because MICE would automatically infer the type of data and apply the appropriate model for imputation. Table 9 below shows the imputation methods determined by MICE for each variable in our dataset:

Table 9: Imputation methods for missing values in each variable in Home Equity Dataset

Variable	Imputation Method
Loan Amount	
Mortgage Balance	PNN
Home Value	PNN
Reason for Home Equity	Logistic Regression
Occupation	Polytomous Regression
Years on Job	PNN
Number of Derogatory Reports	PNN
Number of Delinquent Credit Lines	PNN
Credit Line Age	PNN
Number of Inquiries	PNN
Number of Credit Lines	PNN
Debt to Income Ratio	PNN
Default	Logistic Regression

Redundant variables increase the complexity of prediction models without adding a lot of value. They lower the model’s efficiency as well as the accuracy. Upon conducting a correlation test, we found that Mortgage Balance had a high positive correlation of 0.87 with the variable Home Value. When we refer to Table 7, we see that Mortgage Balance has 8.69% missing values whereas Home Value only has 1.88% missing values. Thus, for models that cannot handle redundancies, we will drop Mortgage Balance from our data.

Figure 9: Correlation among Variables



We used the z-score method to detect outliers. We observed outliers in every numerical variable. The outliers have not been treated as we do not want to compromise the direction of our analysis, which can be biased if we remove or impute the outliers. For example, outliers in home values exist because there is a huge variation in home prices, depending on things such as location and type of property. We can see a visual representation of this in *Appendix B – Figure B1*. From our data summary, we know the standard deviation is around 57, 471. We will opt for prediction models that are more robust to outliers.

As discussed before, our target variable, Default, has a class imbalance which could influence the model's capability of classifying correctly. There are 4772 values in the negative class (no-default) 0 and 1190 values in the positive class (default) 1. Only 19.95% of observations in the dataset have defaulted on their loans. We have chosen 2 methods: Random Oversampling and Class weighting, which will be discussed in the respective models that they have been applied.

Customer Understanding Through Cluster Analysis

In order to get a better understanding of the loan customers, we conducted Cluster Analysis. We are interested in finding patterns that naturally exist in the data to help us understand customer behavior. We also believe that cluster analysis will serve as a basis for further analysis of the data.

We have a lot of outliers in our data that we do not want to remove. For this reason, we have opted for K-Medoids Cluster Analysis (PAM) as it is more robust to outliers and noise. Three of the variables, namely Number of Derogatory Reports, Number of Delinquent Credit Lines, and Number of Inquiries, have a lot of 0 values and a limited number of other values. We have converted these 3 numerical variables into categorical variables to assist our cluster analysis. To standardize our data, we used a combination of the Center Scale and Yeo-Johnson transformation. Due to the mixed nature of our data, we decided to use Gower Distance to calculate the distance.

Table 10: External Validation of cluster analysis using Target Variable

Clusters	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
No Default	794	536	574	217	664	904	685	398
Default	228	93	83	287	153	162	97	87
% Default	22.31%	14.79%	12.63%	56.94%	18.73%	15.20%	12.40%	17.94%

From table 10, we can distinguish the clusters into two broad zones. Clusters 2, 3, 6, and 7 have a lower rate of Default customers, with cluster 7 having the lowest rate. Clusters 1, 4, 5, and 8 have higher rates of Default, with cluster 4 with a significantly higher rate of 56.94%.

After examining the variables based on the clusters, we found that Cluster 4 also has the most customers with a high average debt-to-income ratio of 22.85. Cluster 4 also has the highest number of people with a lower Credit Line Age. People in cluster 4 also have lower average Years on Job. This cluster has people with more numbers of credit lines, and a greater number of people with 1 or more derogatory reports. All the customers in this cluster have 1 or more delinquent credit lines. The occupation category “Others” seems to be highly concentrated in Cluster 4.

Cluster 1 and cluster 5 also follow a pattern similar to cluster 4. The debt-to-income ratio is higher, and the Years on Job are lower. Cluster 1 also has the lowest loan amount, mortgage balance, and home value among all the clusters.

On the other hand, Cluster 7, with the lowest default rate of 12.40%, has a higher loan amount, mortgage balance, and home value. This cluster also has people with the lowest average debt-to-income ratio of 20.23. The occupation category “Professionals/ Executives” is concentrated in this cluster. Similarly, cluster 2 also has a high average home value and mortgage balance. This cluster also has people with the oldest credit lines.

To validate our cluster analysis, we used External Validation as seen in Table number. We also checked the RAND index, which is 0.01288. The cluster statistics for this clustering solution are shown in Table 11.

Table 11: Cluster Statistics

Average Between	Average Within	Dunn
0.280915	0.162105	0.004402

Higher "Average Between" values, lower "Average Within" values, and a higher "Dunn" index is preferred in a cluster solution. Our solution does not satisfy those conditions adequately. All these validation measures suggest that our cluster analysis does not have a good recovery.

Although we cannot rely solely on the insights of this analysis, we believe that the cluster information found from this analysis will prove useful to the company to get a better understanding of their loan customers. The important variables, based on this analysis, that can help to determine the default risk a customer poses are Debt to Income Ratio, Number of Derogatory Reports, Number of Delinquent Credit Lines, Years on Job, and Credit Line Age.

Data Analysis:

1. Decision Tree

As discussed in the Mortgage Loan Dataset, decision trees are self-explanatory and easy to interpret if small. They are also robust to irrelevant, redundant features, outliers, and missing values. Decision tree models are easy and fast to train and test.

We want to assess the factors that contribute the most to a default. We have used a Decision Tree for this data. Even though a decision tree can handle missing values, we have imputed all the missing values.

Our decision tree model is tuned to the cost complexity parameter by performing a grid search for the optimal complexity parameter value. We treated the class imbalance by using the class weighting method. The majority class 0 (No-Default) was given a weight of 0.625 and minority class 1 (Default) was given a weight of 2.505. We have split our dataset in the ratio of 80/20, which means 80% of the data is used in the training set and 20% of our data is used in the testing set. 4770 observations are used in the training set and 1192 observations have been used in the testing set.

Based on our weighted model, the top five variables of importance are Debt to Income Ratio, Number of Delinquent Credit Lines, Credit Line Age, Home Value, and Mortgage Balance.

Even though we can get an idea about the important predictor variables, the overall training and testing performance of the model is not optimal.

Table 12: Weighted Decision Tree Performance

	Training	Testing
Accuracy	0.87	0.78
Kappa	0.66	0.42
Sensitivity	0.94	0.71
Specificity	0.85	0.79
Precision	0.61	0.46
Recall	0.94	0.71
F1	0.74	0.56

The training set accuracy and kappa value are at 87.3% and 66.7% respectively. However, the testing set accuracy is at 78% and the kappa value is at 42.5%. The low kappa value indicates that this model is comparable to random guessing. The model is also overfitting the training dataset. Other measurements such as Precision, Recall, and F1 do not compensate for the low performance either. Therefore, we would like to explore more prediction models to get better performance while predicting default.

2. Support Vector Machines

A problem with our data is that most numerical variables in our dataset are plagued with outliers. SVM can handle this problem because the model itself is robust to redundant variables and outliers. Also, SVM could potentially provide a more balanced model since it is not very prone to overfitting. For the above reason, we believe that SVM could be a good candidate to be our prediction model for this data set.

We have split our dataset in the ratio of 85/15, which means 85% of the data is used in the training set and 15% of our data is used in the testing set. To correct the class imbalance, the majority class 0 (No-Default) has been down-sampled to meet the minority class 1 (Default). The data has been transformed using center-scale normalization. Categorical variables, Reason for Home Equity and Occupation, have been binarized for the analysis.

The optimal parameters used for the model were sigma = 0.2302467 and C = 62.52626 . We train our SVM model and achieve the results below. The SVM model is quite balanced, with accuracy metrics of Training and Testing are 93% and 89% respectively. Kappa value is at a moderately high level of 67%. Notably, the model has high Sensitivity and Recall, which means the false negative rate is low, or the model is less likely to wrongly predict a customer that will not default.

Table 13: Support Vector Machines Performance

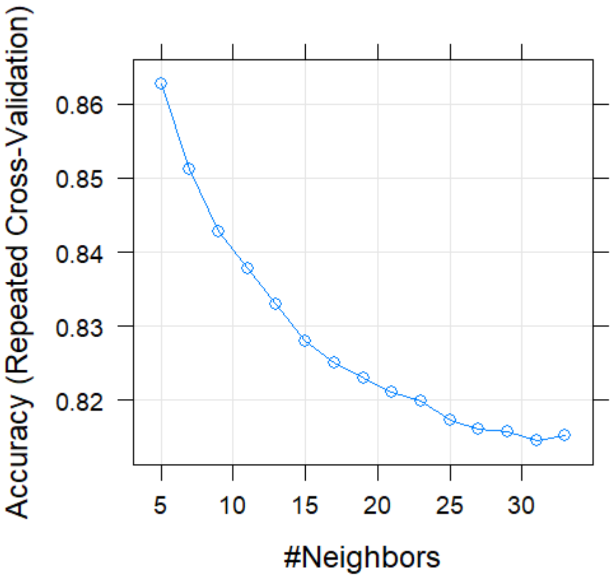
	Training	Testing
Accuracy	0.93	0.89
Kappa	0.79	0.67
Sensitivity	0.99	0.81
Specificity	0.91	0.91
Precision	0.73	0.68
Recall	0.99	0.81
F1	0.84	0.74

3. K-Nearest Neighbor

We implemented the K-Nearest Neighbor algorithm to our dataset because it is proven to be effective when the dataset is large and contains noisy data. However, the algorithm does not produce a model, which could limit the ability to understand how the features are related to the class.

For K-Nearest Neighbour, the dataset is split in the ratio of 85/15, which means 85% of the data is used in the training set and 15% of our data is used in the testing set. The data has been transformed using min-max range normalization. Categorical variables, Reason for Home Equity and Occupation, have been binarized for the analysis.

Figure 10: Cross Validation of KNN



The kNN algorithm produced the results below. The accuracy of the classification is relatively high at 87%, which means kNN performs well with our dataset. However, sensitivity and recall are low compared to the SVM model. This also indicates that kNN has a higher rate of false negatives, which makes it less optimal when compared with other models we used.

Table 14: KNN Model Performance

	Tuned
Accuracy	0.87
Kappa	0.47
Sensitivity	0.39
Specificity	0.98
Precision	0.86
Recall	0.39
F1	0.53

5. **Deep Learning**

   As mentioned, the Deep learning model can be influenced by redundant variables, so we proceed to determine and remove redundant variables if necessary. We calculated the correlation between predictor variables and found that only 2 variables (Mortgage Balance and Home Value) had high correlations above 75%. As mentioned in the data exploration, we decided to drop Mortgage Balance over Home Value due to data quality. Also, we applied min-max normalization to the variables.

   For the hyperparameter parameter tuning, we searched over 6 hidden layers to find the optimal number of nodes. For the class imbalance problem, we used the SMOTE algorithm for oversampling. The model performance is shown in Table 15:

Table 15: Deep Learning Model Performance

	Training	**Testing**
Accuracy	0.90	0.89
Kappa	0.80	0.78
Sensitivity	0.84	0.84
Specificity	0.84	0.84
Precision	0.95	0.93
Recall	0.84	0.84
F1	0.89	0.89

The model was balanced as training and testing are approximately equal across measurements, suggesting that SMOTE algorithm handled class imbalance well. The deep learning model had a high accuracy of 89%, which is on par with the performance of SVM. However, the model performed better than all other models for this dataset with high precision, F1 value, and especially recall. This was a deciding factor for us to choose the Deep learning model as the best model for the home equity dataset.

Conclusion

The world economy right now has become unstable more than ever because of the effect of the Covid-19 pandemic. With a high unemployment rate, stagnated GDP, and volatile real estate market, the current situation reminds us of the economy before the crisis in '08. Even though the causes of economic crises usually vary from one to another, we cannot say for certain that a housing bubble crash will not happen again in the future. At these times, it is best to be prudent and learn from our mistakes of letting the real estate bubble burst from happening. Therefore, we believe that the findings from our research are relevant and would add value to people working in the real estate industry. Companies can use our insights to set up and enhance the risk management indicators in their mortgage loan portfolio.

For the mortgage dataset, our goal is to identify important factors when accessing mortgage risk based on macroeconomic factors and the nature of the loan. Based on the dataset, the factors that significantly impacted the default rate are GDP, unemployment rate, housing price index, LTV, and interest rate. Among these, GDP and the unemployment rate stood out as strong indicators. If we take a closer look, on average, the default rate is more sensitive when GDP decreases compared to when the unemployment rate increases. Therefore, we would suggest the company give a higher weight to GDP among the macroeconomic factors when accessing mortgage loan risk.

For the home equity dataset, our analysis helped us to determine the important factors of the customer profile for companies to consider when issuing a mortgage loan. Although the clustering solution that we applied in this dataset was not very sufficient due to the low RAND index, it still provided some notable insights in terms of customer understanding. Our results showed that people with higher years on jobs usually have a lesser rate of default. Also, in the cluster where the rate of default is lowest, most customers are categorized as working in the Professionals/ Executives occupation. Credit line age is also an interesting factor, as people with a credit line age greater than two are much more likely to default compared to people with a credit line age of one or smaller. We believe these factors are relevant to our research goal and companies should consider incorporating them as indicators when accessing customer profiles for a loan.

In terms of predicting default, in this research, we applied four machine learning models: Decision Tree, SVM, K-NN, and Deep Learning. To evaluate the performance of our model, we considered many metrics such as accuracy, kappa, precision, recall, F1 value, etc. However, we gave the highest weight to recall, as we wanted our models to have minimized false negative rate. This is important in the business context, as we believe that the proposed model should support the company to detect high-risk customers before issuing the loan. We want to avoid the situation where the model predicts that a specific customer will not default, but that customer ends up defaulting after receiving his/her loan. This could incur an unexpected loss for the company, so we aim to minimize the false negative rate as much as we can. Based on these criteria, our results indicated that Deep Learning has the best performance. Not only does the Deep Learning model produce high accuracy when predicting default, but it also performed better in terms of recall rate when compared with other models. Thus, we would recommend the company apply our Deep Learning model to their risk detection framework to predict customer creditworthiness.

For future improvements, we believe that more information about the datasets would enhance the quality of our analysis. For example, in the home equity dataset, over 20% of observations for variable Occupation are categorized as "Other". We think that a more granular breakdown of the variable could help us to derive better insight and improve our clustering solution. Another variable that we would prefer to include in the home equity dataset in the future is the FICO score. FICO score is a strong indicator of a customer's credit strength, so it could potentially be relevant in building customer profiles for risk management.

References

- Amadeo, K. (2022). When and why did the stock market crash in 2008? The Balance. Retrieved July 8, 2022, from https://www.thebalance.com/stock-market-crash-of-2008-3305535#:~:text=The%20stock%20market%20crash%20of%202008%20occurred%20on%20September%2029,largest%20point%20drop%20in%20history.
- Chad Cowden, Frank J. Fabozzi, and Abdolreza Nazemi. (2019), Default Prediction of Commercial Real Estate Properties Using Machine Learning Techniques. Journal of portfolio management, 2019, Vol.45 (7), p.55-67. Retrieved from Default Prediction of Commercial Real Estate Properties Using Machine Learning Techniques - ProQuest
- Christie, L. (2009, February 12). Home prices in Record Plunge. CNNMoney. Retrieved July 8, 2022, from https://money.cnn.com/2009/02/12/real_estate/Latest_median_prices/#:~:text=The%20median%20price%20for%20a,1.6%25%20between%202006%20and%202007
- Edward N.C. Tong∗, Christophe Mues & Lyn Thomas (2013). A zero-adjusted gamma model for mortgage loan loss given default, *International Journal of Forecasting*, 29 (4), 548-562. Retrieved from https://doi.org/10.1016/j.ijforecast.2013.03.003
- Emad Azhar Ali, Syed; Sajjad Hussain Rizvi, Syed; Lai, Fong-Woon; Faizan Ali, Rao; Ali Jan, Ahmad (2021), Predicting Delinquency on Mortgage Loans: An Exhaustive Parametric Comparison of Machine Learning Techniques. International Journal of Industrial Engineering and Management, 2021, Vol.12 (Issue 1), p.1-13. Retrieved from Predicting Delinquency on Mortgage Loans: An Exhaustive Parametric Comparison of Machine Learning Techniques - ProQuest (drexel.edu)
- Feibelman, A. (2022). BANKRUPTCY AND THE STATE. *Emory Bankruptcy Developments Journal, 38*(1), 1-50. Retrieved from http://ezproxy2.library.drexel.edu/login?url=https://www.proquest.com/scholarly-journals/bankruptcy-state/docview/2637689114/se-2?accountid=10559
- Goolsbee, Austan D., and Alan B. Krueger. 2015. "A Retrospective Look at Rescuing and Restructuring General Motors and Chrysler." *Journal of Economic Perspectives*, 29 (2): 3-24.DOI: 10.1257/jep.29.2.3
- Jackson, J. R., & Kaserman, D. L. (1980). Default risk on home mortgage loans: A test of competing hypotheses: ABSTRACT. *Journal of Risk and Insurance (Pre-1986), 47*(4), 678. Retrieved from http://ezproxy2.library.drexel.edu/login?url=https://www.proquest.com/scholarly-journals/default-risk-on-home-mortgage-loans-test/docview/235110404/se-2?accountid=10559
- Kim, Dong-sup; Shin, Seungwoo, THE ECONOMIC EXPLAINABILITY OF MACHINE LEARNING AND STANDARD ECONOMETRIC MODELS-AN APPLICATION TO THE U.S. MORTGAGE DEFAULT RISK. International journal of strategic property management, 2021, Vol.25 (5), p.396-412. Retrieved from THE ECONOMIC EXPLAINABILITY OF MACHINE LEARNING AND STANDARD ECONOMETRIC MODELS-AN APPLICATION TO THE U.S. MORTGAGE DEFAULT RISK. - Document - Gale Academic OneFile (drexel.edu)

- Magdoff, F., & Yates, M. D. (2010). ABCs of the economic crisis: What Working People Need To Know. Aakar Books. Retrieved from https://books.google.com/books?hl=en&lr=&id=LzIVCgAAQBAJ&oi=fnd&pg=PA9&dq=what+happens+after+economic+crisis&ots=Pd0VyZNGJW&sig=hnlfe1t32HVBmMj1hIpiN7UXPVY#v=onepage&q=what%20happens%20after%20economic%20crisis&f=false.
- Mitrašević, M., & Bardarova, S. (2020). CAUSES OF NON-PAYMENT OF MORTGAGE LOANS: THEORETICAL AND PRACTICAL ASPECTS. *UTMS Journal of Economics, 11*(2), 138-150. Retrieved from http://ezproxy2.library.drexel.edu/login?url=https://www.proquest.com/scholarly-journals/causes-non-payment-mortgage-loans-theoretical/docview/2571154652/se-2
- Pérez-Martín*, Pérez-Torregrosa, M. Vaca. (2018), Big Data techniques to measure credit banking risk in home equity loans. Journal of business research, 2018, Vol.89, p.448-454 Retrieved from Big Data techniques to measure credit banking risk in home equity loans - ScienceDirect (drexel.edu)
- Song, G. (2022). Large US bank takeovers in 2008: Performance and implications. *Journal of Capital Markets Studies*, *6*(1), 33-47. https://doi.org/10.1108/JCMS-06-2021-0021
- Teng, Huei-Wen; Lee, Michael. (2019), Estimation Procedures of Using Five Alternative Machine Learning Methods for Predicting Credit Card Default. Review of Pacific basin financial markets and policies, 2019, Vol.22 (3), p.1950021. Retrieved from Estimation Procedures of Using Five Alternative Machine Learning Methods for...: EBSCOhost (drexel.edu)
- THE FINANCIAL CRISIS INQUIRY COMMISSION (2011), Final Report of the Causes of the Financial and Economic Crisis in the United States. Retrieved from
- https://cybercemetery.unt.edu/archive/fcic/20110310173545/http://c0182732.cdn1.cloudfiles.rackspacecloud.com/fcic_final_report_full.pdf

Appendix A

Tables A1 and A2 reflect the descriptive statistics information for Mortgage Loan Dataset.

Table A1: Descriptive Statistics for Numerical Predictors for Mortgage Loan Dataset

Variable	Mean	Median	Standard Deviation
time	33.88	30.00	15.02
origin_time	22.52	24.00	8.93
first_ob_time	22.52	24.00	8.93
maturity_time	136.40	141.00	16.89
balance_orig_time	252153.00	188000.00	234385.80
FICO_orig_time	667.50	672.00	74.05
LTV_orig_time	78.91	80.00	9.49
Interest_Rate_orig_time	5.52	6.25	3.47
hpi_orig_time	202.90	221.90	33.78
LTV_time_mean	81.47	80.24	15.50
LTV_time_sd	6.42	5.08	5.72
blance_time_mean	245917.00	184001.00	231038.30
balance_time_sd	4271.50	383.80	16682.93
interest_rate_time_mean	7.23	6.93	2.15
interest_rate_time_sd	0.36	0.00	1.13
hpi_time_mean	193.00	196.60	28.71
hpi_time_sd	13.51	12.52	10.40
gdp_time_mean	1.93	1.91	0.98
gdp_time_sd	1.06	0.74	0.82
uer_time_mean	5.53	5.11	1.04
uer_time_sd	0.77	0.32	0.81

Table A2: Count for Nominal Predictors for Mortgage Loan Dataset

Variable	Class	Count
RE_Type	SF	4867
	PU	1150
	CO	615
	OTH	1514
investor	0	5926
	1	2220

Tables A3 and A4 reflect the descriptive statistics of Home Equity Loan Dataset.

Table A3:Descriptive Statistics of Numerical Predictors for Home Equity Loan Dataset

Variables	Median	Mean	Standard Deviation*
Loan Amount	16350	18611	11208.25
Home Value	89236	101778	57471.92
Mortgage Balance	65019	73767	45079.55
Years on Job	7	8.919	7.63
Number of Derogatory Reports	0.00	0.2545	0.85
Number of Delinquent Credit Lines	0.00	0.4493	1.12
Credit Line Age	173.5	179.8	86.54
Number of Inquiries	1.00	1.187	1.73
Credit Line Age	20.00	21.3	10.13
Debt to Income Ratio	34.8183	33.7825	8.84

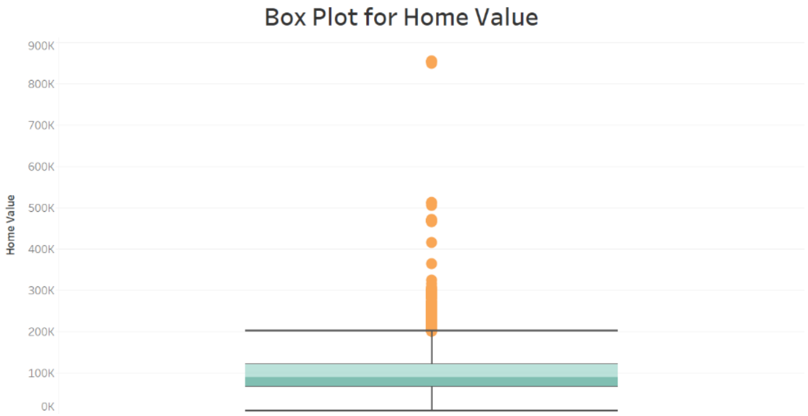
Table A4:Count of Nominal Predictors for Home Equity Loan Dataset

Variable	Class	Count
Occupation	Sales	109
	Office	948
	Manager	767
	Professional Executive	1276
	Self Employed	193
	Other	2390
Reason for Home Equity	Debt Consolidation	3930
	Home Improvement	1780

Appendix B:

The box plot detecting outliers for the variable home value is shown in Figure B1.

Figure B1: Outliers in Home Value for Home Equity Dataset


Box Plot for Home Value
Home Value
900K
800K
700K
600K
500K
400K
300K
200K
100K
0K