**Mortgage Payback Analytics**

Project 3

Sreeja Reddy Singidi

04-24-2024

**Table of Contents**

[1 Introduction 5](#_Toc164775638)

[2 Business Understanding 5](#_Toc164775639)

[3 Business Goal 5](#_Toc164775640)

[4 Analytical Approach 6](#_Toc164775641)

[5 Attribute Definition 6](#_Toc164775642)

[6 Data Understanding 9](#_Toc164775643)

[6.1 Number Of Records 9](#_Toc164775644)

[6.2 Checking For Missing Values 9](#_Toc164775645)

[6.3 Checking For 0 Values 10](#_Toc164775646)

[7 Data Preprocessing 12](#_Toc164775647)

[7.1 Optimizing Dataset For Model Efficiency 12](#_Toc164775648)

[7.2 Initial Transformation 12](#_Toc164775649)

[7.2.1 Time Attribute To Observation Count 12](#_Toc164775650)

[7.2.2 Origination Time 13](#_Toc164775651)

[7.2.3 First Time 13](#_Toc164775652)

[7.2.4 Handling Balance Time Column 13](#_Toc164775653)

[7.2.5 Ltv Ratio 13](#_Toc164775654)

[7.2.6 Interest Rate 13](#_Toc164775655)

[7.2.7 Fixed Interest Rate 13](#_Toc164775656)

[7.2.8 House Price Index, Unemployemnt, Gdp 13](#_Toc164775657)

[7.2.9 Real Estate Types 13](#_Toc164775658)

[7.2.10 Investor Type 14](#_Toc164775659)

[7.2.11 Origination Time Columns Handling 14](#_Toc164775660)

[7.2.12 Status Time 14](#_Toc164775661)

[8 Significance Of Analyzing Ongoing Loans 15](#_Toc164775662)

[9 Predictor Analysis 15](#_Toc164775663)

[9.1 Real Estate Types 15](#_Toc164775664)

[9.2 Analyzing Attribute Influence On Default And Payoff 15](#_Toc164775665)

[9.2.1 Investors Status 16](#_Toc164775666)

[9.3 Loan Origination Attributes 16](#_Toc164775667)

[9.3.1 Relation Between Origin Time (Default And Payoff) 16](#_Toc164775668)

[9.3.2 Fico Among Loan Status 17](#_Toc164775669)

[9.3.3 Distribution Of Ltv And Ltv Origin 17](#_Toc164775670)

[9.4 Relation Between Gdp And Unemployment Rate 18](#_Toc164775671)

[9.5 Fixed Interest Rate With Status Time: 19](#_Toc164775672)

[9.6 Economic Indicators For Selected Time Frame 19](#_Toc164775673)

[9.7 Relevancy 20](#_Toc164775674)

[10 Dimension Reduction 20](#_Toc164775675)

[11 Data Transformation 21](#_Toc164775676)

[12 Data Partitioning Methods 21](#_Toc164775677)

[12.1 Necessity 21](#_Toc164775678)

[12.2 Different Approaches 21](#_Toc164775679)

[12.3 Approach In This Project 22](#_Toc164775680)

[12.4 Importance 22](#_Toc164775681)

[13 Goal: Classifying Borrowers 22](#_Toc164775682)

[13.1 Model Selections 22](#_Toc164775683)

[13.1.1 Logistic Regression 22](#_Toc164775684)

[13.1.2 Decision Trees 23](#_Toc164775685)

[13.1.3 K-Nearest Neighbors (Knn) 23](#_Toc164775686)

[13.1.4 Naïve Bayes: 23](#_Toc164775687)

[13.2 Model Performance 23](#_Toc164775688)

[13.3 Best Model 23](#_Toc164775689)

[14 Predicting The Status On Ongoing Data 24](#_Toc164775690)

[15 Conclusion 24](#_Toc164775691)

[16 Future Work 24](#_Toc164775692)

[Executive Summary 25](#_Toc164775693)

**List Of Figures**

Figure 1: Missing values 10

Figure 2: Real estate distribution. 15

Figure 3: Comparison of Investor’s Distribution 16

Figure 4: Analysis of origination time. 17

Figure 5: Distribution of Credit Score 17

Figure 6: Ltv and Ltv origin 18

Figure 7: Relation between GDP and Uer 19

Figure 8: Fixed Interest with status time 19

Figure 9: Correlation Plot 20

**List of Tables**

Table 1: Attributes 9

Table 2: Missing values for each borrower 10

Table 3: Final Attribute Table 14

Table 5: Duplicate Records 21

Table 6: Data Transformation 22

Table 7: Model Performance 24

# **INTRODUCTION**

The dynamics of mortgage payback play a crucial role in the complex world of real estate financing, impacting loan companies' stability and profitability as well as the overall financial system. It's like solving a puzzle about how borrowers manage their mortgage payments. This project is all about figuring out patterns in how people repays their mortgages.

In the world of real estate finance, knowing how people pay back their mortgages is important. Mortgage payback analysis helps us understand how homeowners repay the money they borrowed to buy their houses. By looking at past data about when people got their mortgages and how they paid them back, we can learn a lot about how mortgages work. This analysis helps lenders and borrowers make better decisions about loans and housing. Lenders can adjust their risk management procedures and lending strategies by having a better understanding of the patterns observed.

# **PROJECT OVERVIEW**

A screenshot of a diagram

Description automatically generated

Figure 1: Overview 1

*A diagram of a flowchart

Description automatically generated*

Figure 2: Overview 2

# **BUSINESS UNDERSTANDING**

Understanding the dynamics of mortgage payback is crucial for various stakeholders in the real estate finance industry, including lenders, borrowers, and policymakers. For loan companies, staying stable and making money mostly depends on how well borrowers pay back their mortgages. If they understand how borrowers usually repay their loans, they can predict when they'll get money back, handle their cash well, and use their resources wisely. By spotting trends in how people pay back mortgages, they can plan better to make as much profit as possible while keeping financially safe.

Agencies keep a close eye on the mortgage business to make sure people are treated fairly and the financial system stays strong. When lenders look at how mortgages are paid back, they can show they're following the rules set by the government. This includes being fair to borrowers, managing risks properly, and being transparent about the terms of the loans. By understanding repayment patterns, lenders can also get ahead of any new rules coming up and avoid getting into trouble with the law.

The status of the loan tells what's happening with each loan in the dataset- whether it's been paid back, not paid back, or still ongoing. When we look at this, we can see if a loan has had problems in the past, like not being paid back on time. If a loan has had problems before, it might mean it's risky to give out another loan and can be considered a **red flag**.

# **BUSINESS GOAL**

The main business goal is to look at how people paid back loans before to see if they might have trouble paying back a new loan. This helps the lending company avoid losing money if borrowers can't pay back their loans on time. To precisely determine **credit risk**, they make use of historical data and predictive analytics. Lenders can reduce risks and decrease losses by using the financial system to identify high-risk applicants and likely default scenarios early. If a system indicates a default status it shows that loans have not been paid. When the system spots these defaults early, it can act like rejecting risky loan applications or changing the loan terms to lower the chance of future defaults. This way, it helps the lender avoid losing money and keeps their loan portfolio in good shape.

It is more important to use information about how people pay back their loans to make more money. This could mean finding people who are likely to pay back their loans on time and offering them special deals. It might also involve changing how much interest is charged based on how reliable borrowers have been in the past, or making the rules for getting a loan stricter to attract people who are more likely to pay it back.

**Two Approaches:**

* Classifying a customer as whether they will pay off the loan or default.
* Classifying a new customer into payoff or default based on the historic data.

# **ANALYTICAL APPROACH**

Before diving into data preprocessing, thorough analysis of the dataset is very important, and Excel can be most helpful in this case.

**Excel analysis:**

After analysis, it appears that the "status\_time" variable contains data on the default and payment times and 0 denotes that the debt has not been paid off or is in default. Finding loans that are at risk of default or that have already defaulted is the business's top priority because these situations call for quick attention and possible action to reduce losses. We can determine whether a record falls into the category of payoff or default by analyzing these 0 records.

To ensure the output generated can be reliable, we should seperate customers whose last observation is 0, which means it is non-default and non-payoff into a new dataset.

Combining details about the type of properties linked with mortgage loans into one category makes the data clearer. It helps us to understand the various property types easily.

Typically, credit scores range from 300 to 850, with 300 being the lowest score possible. However, in this dataset, the lowest acceptable credit score is 400, which is different from the usual range.

Moreover, finding 0 values in the interest rate column needs more investigation to decide how to deal with them in the next stages because no loan is given without the interest rate.

**Classification Algorithm Development:**

Following thorough analysis and data preprocessing, the next step involves building classification algorithms. By assessing the probability of default or payoff, these algorithms are essential in evaluating the financial ability of loan applicants. Lenders can improve risk management and improve the loan process by using these forecasts to guide their decisions on whether to approve or deny loan applications.

# **ATTRIBUTE DEFINITION**

1. id: In the dataset, every mortgage borrower is individually identified by the "id" field. The dataset makes it possible to follow and analyze the performance and behavior of individual borrowers over time by giving each borrower a unique ID. This has 50,000 unique observations.
2. time: It represents the period at which the observation was recorded for each mortgage borrower. Time attribute has 60 time periods or observations, capturing data over some time for each loan.
3. orig\_time: The beginning of a loan's lifespan is shown by its origination time. It denotes the specific time when the borrower enters into a loan agreement with the lender, by signing the mortgage documents and officially initiating the loan process. Some loans originated before the start of the period are indicated by negative values.
4. first\_time: This indicates the initial observation time that has been recorded. It refers to the first time that information about the loan—including its balance, interest rate, and other details—was recorded or noticed. It shows starting from the earliest observed period (1) to the latest observed period (60).
5. mat\_time: The time at which the mortgage loans mature is the "mat\_time" attribute. It indicates the point in the loan term that the borrower has taken to pay back the entire loan balance plus any interest that has accumulated. The customer can also pay the entire amount before the maturity time.
6. balance\_time: This shows the remaining balance of the mortgage loan at a specific time point. It reflects the amount of money still owed by the borrower to the lender at each observation time.
7. LTV\_time: LTV stands for Loan-to-Value ratio. It is a financial term used by lenders to show the ratio of a loan to the appraised value of the property securing the loan which means- it measures the percentage of the property's appraised value that is being financed by the loan.
8. interest\_rate\_time: Shows the interest rate on the mortgage loan at the observation time. This represents the percentage of the loan amount that the borrower must pay to the lender as interest for borrowing the money.
9. hpi\_index: The House Price Index is a metric used to monitor changes in the cost of residential property over time. The base year is 10. It offers insights into the general patterns and movements in property values within a certain region. If the HPI for a specific area increases by 5% in a given period, it indicates that, on average, house prices in that area have increased by 5% relative to the base year. In the same way, a decrease in the HPI suggests a decline in property values.
10. gdp\_time: It is the Gross Domestic Product (GDP) at the observation time. This allows analysts to evaluate the state of the economy during the period. GDP fluctuations throughout time may be a sign of changes in economic activity or recession. These changes may affect the mortgages.
11. uer\_time: Represents the unemployment rate at the observed time. This is the percentage of individuals who are unemployed and are actively looking for work among the total labor force within the given timeframe.
12. REtype\_CO\_orig\_time: Real estate type- condominium- at loan origination. This categorical attribute specifies the type of real estate associated with the mortgage loan at the time of origination, indicating if it's a condominium. The condominium is nothing but a condo which is a type of housing (apartment).
13. REtype\_PU\_orig\_time: Real estate type - planned urban development - at loan origination. Similar to the previous attribute, this categorical attribute denotes the type of real estate associated with the loan at origination, indicating if it's a planned urban development.
14. REtype\_SF\_orig\_time: Real estate type - single family - at loan origination. This categorical attribute indicates the type of real estate associated with the loan at origination, indicating if it's a single-family property.
15. investor\_orig\_time: This indicates whether the borrower is an investor or not. If the borrower is an investor, the value would be 1, and if not, it would be 0.

The borrower is an individual who invests in loans rather than borrowing money for personal or business use. Investors provide funds to borrowers in exchange for earning interest or returns on their investments.

1. balance\_orig\_time: Original balance amount of the loan. Like the balance\_time, this attribute sh ows the initial balance of the mortgage loan at the time of origination.
2. FICO\_orig\_time: This refers to the FICO credit score assigned to an individual at the time when their mortgage loan originated. It is crucial to figure out how risky the borrower is and how likely it is that they will make timely loan repayments. To determine the borrower's credit risk and to make well-informed judgments on loan approval, terms, and interest rates, lenders utilize FICO scores. Higher FICO scores are often associated with lesser credit risk and might lead to better loan terms for the borrower.
3. LTV\_orig\_time: Loan-to-value (LTV) ratio at loan origination. This denotes the loan-to-value ratio at the time of loan origination, giving insights into the initial financing structure of the loan.
4. Interest\_Rate\_orig\_time: Interest rate at loan origination. Like the interest\_rate\_time, this shows the annual interest rate charged on the mortgage loan at the time of origination.
5. hpi\_orig\_time: House price index at loan origination. Like the hpi\_time, this attribute gives information about the house price index at the time of loan origination, providing information on the property market conditions at the loan's inception.
6. default\_time: This binary attribute indicates whether the mortgage loan has defaulted (1) or not (0) at the specific observation time. When a loan defaults, it means that the borrower has not made the required payments as scheduled, which is going against the contract with the lender.
7. payoff\_time: Payoff time refers to the point in time when a borrower completes the repayment of their mortgage loan in full. When the payoff time is 1, it shows that the mortgage has been completely paid off by the borrower. Same way, if the payoff time is 0, it means that the loan has not been fully repaid.
8. status\_time: Provides information about the loan's status at a specific observation time. This attribute helps in understanding the current condition of loans within the dataset. It is a key attribute for assessing the financial health of the loan portfolio and making decisions about risk management and servicing.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Type |
| id | Identifier for each mortgage borrower. | Numeric |
| time | Time period for each record. | Numeric |
| orig\_time | Original time period when the mortgage loan was originated | Numeric |
| first\_time | First observation time period for the mortgage borrower | Numeric |
| mat\_time | Maturity time period for the mortgage loan. | Numeric |
| balance\_time | Balance amount remaining on the mortgage loan at the observation time | Numeric |
| LTV\_time | Loan-to-Value (LTV) ratio at the observation time | Numeric |
| interest\_rate\_time | Interest rate on the mortgage loan at the observation time | Numeric |
| hpi\_time | House Price Index (HPI) at the observation time | Numeric |
| gdp\_time | Gross Domestic Product (GDP) at the observation time | Numeric |
| uer\_time | Unemployment rate at the observation time | Numeric |
| REtype\_CO\_orig\_time | Real estate type (condo) at the original loan origination time | Categorical |
| REtype\_PU\_orig\_time | Real estate type (planned urban development) at the original loan origination time | Categorical |
| REtype\_SF\_orig\_time | Real estate type (Single Family Home) at the original loan origination time | Categorical |
| investor\_orig\_time | Investor being a borrower is given by 1 otherwise 0. | Categorical |
| balance\_orig\_time | The original balance amount of the mortgage loan at origination time | Numeric |
| FICO\_orig\_time | FICO credit score at origination time | Numeric |
| LTV\_orig\_time | Loan-to-Value (LTV) ratio at origination time | Numeric |
| Interest\_Rate\_orig\_time | Interest rate on the mortgage loan at origination time | Numeric |
| hpi\_orig\_time | House Price Index (HPI) at origination time | Numeric |
| default\_time | Default status (0 for non-default, 1 for default) | Categorical |
| payoff\_time | Payoff status (0 for non-payoff, 1 for payoff) | Categorical |
| status\_time | Status indicator (0 for non-default/non-payoff, 1 for default, 2 for payoff). | Categorical |

*Table 1: Attributes*

# **DATA UNDERSTANDING**

Before proceeding with data preprocessing, it's crucial to thoroughly understand the dataset and make necessary adjustments, such as handling missing and zero values. Handling missing and zero values is particularly important as it ensures that analysis is based on complete and accurate information. This helps to better understand the behavior of each borrower by having a clear view of their data. These steps are **performed first** because based on all the records of each borrower we will know how they are behaving and **take appropriate measures** before removing the records.

## **NUMBER OF RECORDS**

This is a huge real-time dataset that consists of 622490 observations and 23 variables that are the characteristics of the mortgage.

## **CHECKING FOR MISSING VALUES**

We see from Figure 1, that there are 270 missing values and all of them are from LTV time.

A blue and black graph

Description automatically generated with medium confidence

*Figure 3: Missing values*

It has been noticed that these 270 records are from 18 different borrowers. Table 2, shows how many missing records are observed for each borrower ID.

|  |  |
| --- | --- |
| Borrower ID | Missing records |
| 39722 | 36 |
| 39723 | 36 |
| 39724 | 1 |
| 39725 | 36 |
| 39726 | 4 |
| 39727 | 10 |
| 39728 | 3 |
| 39729 | 2 |
| 39730 | 36 |
| 39731 | 19 |
| 39732 | 5 |
| 39733 | 36 |
| 39734 | 3 |
| 39735 | 17 |
| 39736 | 5 |
| 39737 | 8 |
| 39738 | 4 |
| 49658 | 9 |
| Total | **270** |

Table 2: Missing values for each borrower

Grouping by ID for imputation is impossible due to the pattern seen in the dataset, where most IDs had records with missing values in the "LTV\_time" column only. As of now, eliminating all IDs that have missing data.

## **CHECKING FOR 0 VALUES**

It was found that multiple columns in the dataset had 0 values when the dataset was examined for 0 values. It's crucial to remember that some of these characteristics- like real estate types, investor origin times, payoff and default times, and status times—are binary variables. When it comes to binary variables, one category (like non-default, not paid off, or a particular kind of real estate) is usually represented by a value of 0, and the other (like default, paid off, or a different kind of real estate) by a value of 1. As a result, the 0 values in these columns reflect the absence of a category rather than necessarily indicating missing or incorrect data. Because these 0 values contain important information, they shouldn't be regarded as true missing data.

A screen shot of a computer

Description automatically generated

Code Snippet 1: Zero values in each attribute.

**Origin time:**

The 0 values in "orig\_time" and the consistent value of 102.24 for "hpi\_orig\_time" from all these 0 values could mean that these loans originated during a time when the house price index remained relatively stable. It might be a default placeholder value applied to loans that were made during periods or under conditions.

These 0’s can also mean that the loans originated before the start of the observation period. This happens because some loans are already in existence when the data collection begins, and their exact origination time might not be available due to privacy or security concerns. So, instead of having specific origination timestamps, these loans are represented with 0 values in the dataset.

Negative values indicate that the loan was initiated sometime before the beginning of the data collection period.

**Balance time:**

When a borrower's balance on a mortgage loan is 0, it shows that they have paid back the entire loan amount. To confirm this, checked the status time which shows that that loan status is either paid off or still going on for all these records.

**LTV time:**

When the loan's outstanding balance is zero, it indicates that the borrower has paid back the entire loan amount. In these situations, as there is no longer a loan balance owed about the property's value, the LTV (Loan-to-Value) ratio would anyway be zero.

**Interest rate time and Interest rate origin time:**

Interest rate is shown in percentage terms that the lender is charging to borrowers. I’ve noticed that all these are default types of loans. These need to be handled because according to my knowledge, every loan will have at least a minimum interest rate.

Handling these 0’s can be done by imputation. Correctly handling these 0 values plays a major role in achieving the business goal.

In this case, it is handled by replacing 0 values with the median interest rate for each borrower ID but noticed that 2 IDs still have 0 interest rates. Those borrower ID records can be eliminated.

To handle 0 values in Interest rate origin time, the dataset and be grouped by borrower ID, and the median interest rate for each ID can be calculated based on the interest\_rate\_time column.

The median interest rate for each borrower ID can be then used to replace 0 values in the Interest\_Rate\_orig\_time column, ensuring that each borrower's interest rate is represented accurately.

# **DATA PREPROCESSING**

## **OPTIMIZING DATASET FOR MODEL EFFICIENCY**

In line with the business goal of accurately classifying customers as defaulters or payoffs, it's essential to simplify the dataset to enhance model efficiency. This involves examining the usefulness of records containing multiple timestamps leading up to deployment.

In the dataset, each borrower has multiple records corresponding to different time stamps, capturing various attributes such as loan balance, LTV ratio, interest rate, and more. However, for analysis and modeling purposes, it's often more convenient to have consolidated data where all relevant information is captured into a single observation of a borrower.

The **goal** is to create a more **condensed** and structured dataset that retains all essential information while simplifying the analysis process.

This process works like:

The **first step** is to identify a unique identifier for each borrower, that is ID. This identifier will be used to group the data during the aggregation process.

In the **second step**, we group the dataset by the ID, ensuring that all records for each borrower are grouped together.

**Next**, within each group, we aggregate or summarize the information across different time stamps. This could involve statistics, or simply selecting the most recent value.

**Finally**, a new dataset will be created where each row represents a unique borrower, with all relevant information aggregated into columns. This condensed dataset is more suitable for analysis and modeling tasks.

All this will be performed by using **sqldf** in R so that the process does not take much time and is easy to understand.

## **INITIAL TRANSFORMATION**

### **TIME ATTRIBUTE TO OBSERVATION COUNT**

The time attribute consists of all the records of each borrower where he made a payment for each time stamp. This is not necessary, so we can consider how many time stamps or transactions are associated with each borrower. This calculation is helpful because it provides a quick overview of the activity of each borrower.

### **ORIGINATION TIME**

This is the time stamp for origination which will be same for borrower. So, considering the last time stamp’s corresponding value will be sufficient.

### **FIRST TIME**

This is the time stamp of the first observation. Like the origination time this can also consider the last time stamp’s value.

**Maturity attribute** is not considered in this new dataset because this dataset deals with 30 years of maturity only. As it is evident that all the borrowers have the same maturity time, we can exclude this column.

### **HANDLING BALANCE TIME COLUMN**

The difference between all balance times for each ID is calculated, identifying irregular payments. This column shows how many times the borrower did not make any payment between any 2-time stamps.

### **LTV RATIO**

The LTV ratio is a measure of the loan amount compared to the appraised value of the property securing the loan. It is considered to average the LTV for each borrower because watching how the average LTV ratio changes over time can tell us a lot about how risky a borrower's loans are becoming. For instance, if the average LTV ratio is going up, it could mean borrowers are borrowing more money compared to the value of their properties.

### **INTEREST RATE**

The average is calculated the same as LTV ratio.

### **FIXED INTEREST RATE**

A **new column** can be added based on Interest rate. The **fixed interest rate** indicates whether the interest rate remains constant or changes over time for each borrower's loan. The value 1 suggests that the interest rate remains fixed throughout the loan term, while a value of 0 indicates that the interest rate is variable and may change over time. If the interest rate remains fixed, borrowers may repay, leading to payoffs. If the interest rate is 0, fluctuations in the rate could impact borrower’s ability to make timely payments, resulting in defaults.

### **HOUSE PRICE INDEX, UNEMPLOYEMNT, GDP**

All these 3 are the economic indicators which will be same for each time stamp. Here the latest observation can be considered.

### **REAL ESTATE TYPES**

Based on the original type of real estate, each ID is categorized into four types: Single Family as these have majority records in the dataset as 1, Planned Urban Development as 2, Condo as 3 and 4 means other type of real estate. All these are merged into a single attribute called real estate type.

### **INVESTOR TYPE**

The investor type for each ID at the latest time stamp shows whether the borrower is an investor or not. In the future, we can analyze these investor records separately.

### **ORIGINATION TIME COLUMNS HANDLING**

Various origination time-related columns (such as balance time, FICO, LTV, interest rate and hpi index origination) for each ID at the latest time stamp are extracted.

### **STATUS TIME**

The status time for each ID at the latest time stamp is obtained.

Default time and pay off time are removed from the data because they contain the same information as in status time.

|  |  |  |
| --- | --- | --- |
| Attribute Name | Description | Type |
| id | Identifier for each mortgage borrower. | Numeric |
| observation\_count | How many time stamps or transactions are associated with each borrower. | Numeric |
| origin\_time | Last original time period when the mortgage loan was originated | Numeric |
| first\_time | First observation time for the mortgage borrower | Numeric |
| zero\_diff\_count | How many times the borrower did not make any payment between any 2-time stamps. | Numeric |
| ltv\_ratio | Average of the LTV time for each borrower | Numeric |
| interest\_rate | Average Interest rate on the mortgage loan at the observation time | Numeric |
| fixed\_interest | The interest rate remains constant or changes over time for each borrower's loan. 1= fixed, 0= not fixed | Categorical |
| hpi\_time | House Price Index (HPI) at the observation time | Numeric |
| gdp\_time | Gross Domestic Product (GDP) at the observation time | Numeric |
| uer\_time | Unemployment rate at the observation time | Numeric |
| real\_estate\_type | 1 = single family, 2 = planned urban development, 3 = condo, 4 = Others | Categorical |
| investor\_orig\_time | Investor being a borrower is given by 1 otherwise 0. | Categorical |
| balance\_orig\_time | The original balance amount of the mortgage loan at origination time | Numeric |
| FICO\_orig\_time | FICO credit score at origination time | Numeric |
| LTV\_orig\_time | Loan-to-Value (LTV) ratio at origination time | Numeric |
| Interest\_Rate\_orig\_time | Interest rate on the mortgage loan at origination time | Numeric |
| hpi\_orig\_time | House Price Index (HPI) at origination time | Numeric |
| status\_time | Status indicator (0 for non-default/non-payoff, 1 for default, 2 for payoff). | Categorical |

Table 3: Final Attribute Table

# **SIGNIFICANCE OF ANALYZING ONGOING LOANS**

Creating a separate dataset for loans with status equal to 0 (indicating that the loan process is still ongoing) plays a major role as this dataset will be used for predicting whether the profile will pay or default on the loans.

By doing this, we can focus specifically on the loans that have not yet reached a conclusion, whether that be default or payoff. This allows us to analyze the characteristics and behaviors of these ongoing loans separately from those that have already defaulted or been paid off.

Once the models are trained and validated using historical data, they need to be evaluated on unseen data to assess their real-world performance. The ongoing loans dataset acts as this unseen data, allowing us to assess how well the models generalize to new instances and predict the outcomes of ongoing loans.

# **PREDICTOR ANALYSIS**

## **REAL ESTATE TYPES**

The graph from Figure 2, shows the distribution of real estate types and provides insights.

It is evident from Figure 2, that Single-family homes might dominate the local real estate market, leading to a higher proportion of loans secured by this property type. This might result from factors such as affordability or cultural preferences. Single-family homes might respond differently to economic fluctuations compared to other property types, showing their default and payoff rates.

Condos and urban developments have lower counts overall, indicating a smaller proportion of loans associated with these property types.

A graph of different colored squares

Description automatically generated

Figure 4: Real estate distribution.

## **ANALYZING ATTRIBUTE INFLUENCE ON DEFAULT AND PAYOFF**

The factors that influence mortgage default or payoff enable lenders to make more informed decisions about loan approval and interest rates. Lenders can adjust their lending criteria and pricing based on the risk profile associated with specific attributes, ultimately giving profits, and minimizing losses. This can also identify trends, patterns, and potential areas of concern.

### **INVESTORS STATUS**

Investor status allows lenders to make informed decisions about whether to approve loan applications from investors and non-investors. If loans initiated by investors have higher default rates, lenders may choose to apply stricter criteria or higher interest rates for such applicants.

From Figure 3, most borrowers, regardless of their loan outcome, are regular individuals rather than investors. The higher presence of investors among defaulters compared to payoffs could indicate potential differences in risk-taking behavior or investment strategies between investors and non-investors. Investors might be more willing to take on higher-risk loans, leading to a higher likelihood of default in **some** cases. We can provide **rewards to attract** borrowers who are investors.

A red and blue bars

Description automatically generated

Figure 5: Comparison of Investor’s Distribution

## **LOAN ORIGINATION ATTRIBUTES**

### **RELATION BETWEEN ORIGIN TIME (DEFAULT AND PAYOFF)**

The plot gives insights into how the status of loans has changed over time since origination. This understanding is valuable for assessing the performance of loans and identifying potential risk factors associated with origination time.

It is noticed that at the period of 23-25 months, there is a highest peak in the defaulter's status which shows that most loans are defaulted on in that period. Knowing when loans are most likely to go into default helps lenders use their money better, make specific plans, and change how loans work to lower risks and make more money overall.

A graph of a person with red squares

Description automatically generated with medium confidence

Figure 6: Analysis of origination time.

### **FICO AMONG LOAN STATUS**

Comparing FICO scores at the time of loan origination between defaulters and payoffs allows lenders to assess the success of their credit policies. In Figure 6, it's evident that loans given to borrowers with higher FICO scores tend to have lower default rates. Additionally, looking at the distribution of FICO scores for payoffs, we observe that borrowers with higher credit scores are more likely to successfully pay off their loans. This suggests that prioritizing borrowers with better credit scores can be an effective strategy for reducing default rates and improving loan repayments. Such insights help lenders make informed decisions about borrower eligibility criteria and risk management strategies.

|  |  |
| --- | --- |
| A graph of a graph  Description automatically generated with medium confidence | A graph of a graph  Description automatically generated |

Figure 7: Distribution of Credit Score

### **DISTRIBUTION OF LTV AND LTV ORIGIN**

The distribution of Loan-to-Value (LTV) ratios at a specific time and at the time of loan origination provides insights into the levels of borrowers. By analyzing these distributions, we can understand how the relationship between loan balances and property values changes over time.

A peak in the range of 50-100 suggests that a significant portion of borrowers have relatively high LTV ratios, indicating that they have borrowed a substantial amount compared to the value of their properties. This could imply higher risk for lenders, as borrowers with higher LTV ratios may be more vulnerable to default if property values decline or if they encounter financial difficulties.

A distribution concentrated below 100 suggests that many borrowers initially had lower LTV ratios, indicating that they had more equity in their properties at the time of loan origination. This could be less risky for lenders, as borrowers with lower initial LTV ratios may have a greater financial stake in their properties and may be less likely to default.

A graph of a bar graph

Description automatically generated with medium confidence

Figure 8: Ltv and Ltv origin

## **RELATION BETWEEN GDP AND UNEMPLOYMENT RATE**

When the GDP of a country decreases, it often indicates a slowdown in economic activity. This slowdown may lead to reduced business investments, lower consumer spending, and decreased demand for goods and services. Consequently, businesses may lay off workers or reduce hiring, resulting in a higher unemployment rate.

When the GDP grows, it signifies an expansion in economic activity. This growth can stimulate business investments, boost consumer confidence, and increase demand for products and services. As businesses expand to meet this demand, they may hire more workers, leading to a decrease in the unemployment rate.

A graph with blue dots

Description automatically generated

Figure 9: Relation between GDP and Uer

## **FIXED INTEREST RATE WITH STATUS TIME:**

Loans with fixed interest rates tend to have a higher probability of being paid off compared to defaulting. This suggests that borrowers whose interest rate is fixed may have more stable financial circumstances contributing to successful loan repayment and a reduced no. of defaults.

A red and blue squares

Description automatically generated

Figure 10: Fixed Interest with status time

## **ECONOMIC INDICATORS FOR SELECTED TIME FRAME**

Economic indicators like GDP (the total value of goods and services in an economy), HPI (how housing prices change over time), and UER (the percentage of unemployed people), will be the **same** over each time stamp.

**Why does this matter for mortgages?**

When the economy is steady, people tend to feel more confident about borrowing money for things like house purchases. Also, lenders can better predict if borrowers will be able to pay back their loans on time from the unemployment rate. So, understanding when the economy is stable helps lenders make smarter decisions about who to give loans to and how much interest to charge. It's like having a clearer picture of how risky or safe it is to lend money at different times.

## **RELEVANCY**

Predictors like id and observation\_count have correlation close to 0 with most other predictors. This suggests that they may not be strongly correlated with other predictors and may not provide much additional information for modeling.

Also, investor\_orig\_time has low correlation with many other variables, indicating that it may not have strong linear relationships with other predictors, but low correlation may still be important predictors if their relationships with the status time are non-linear.

Now analyzing the status time, we notice that interest rate, investor origin time, balance origin, FICO and Interest rate origin, have weak correlations with status time.

For building the model all the predictors are considered.

A close-up of a graph

Description automatically generated

Figure 11: Correlation Plot

# **DIMENSION REDUCTION**

After all the preprocessing, again missing values and 0 values are checked. There are duplicate records of a few IDs which need to be handled.

When the same ID is listed more than once, it might mean there's a mistake in the dataset. Dealing with these duplicates shows we're making sure each piece of information is only counted once. This helps avoid having too much of the same data and keeps dataset accurate and reliable.

|  |  |
| --- | --- |
| ID | No. of duplicate records |
| 3014 | 63 |
| 36833 | 63 |
| 36910 | 63 |
| 37067 | 63 |
| 37130 | 63 |
| 37351 | 63 |
| 37528 | 63 |
| 37558 | 63 |
| 37580 | 63 |
| 37637 | 63 |
| 37792 | 63 |

Table 5: Duplicate Records

After removing these duplicate records, the no. of IDs in the data match with- after removing all the missing values from the original dataset.

# **DATA TRANSFORMATION**

The names of the attributes can be changed for better understanding. Changing to more descriptive and understandable terms enhances the clarity and interpretability of the data, making it easier for stakeholders to grasp the meaning of each variable.

|  |  |
| --- | --- |
| Original | Transformed |
| observation\_count | payment\_records\_count |
| first\_time | first\_payment\_time |
| zero\_diff\_count | stable\_balance\_periods |
| hpi\_time | hpi\_index |
| investor\_orig\_time | investor |
| balance\_orig\_time | balance\_at\_origin |
| FICO\_orig\_time | credit\_score |
| LTV\_orig\_time | ltv\_origin |
| Interest\_Rate\_orig\_time | interest\_rate\_origin |
| hpi\_orig\_time | hpi\_origin |
| status\_time | status\_type |

Table 6: Data Transformation

# **DATA PARTITIONING METHODS**

## **NECESSITY**

By dividing the dataset, we can train models on a portion of the data, assess their efficiency, and ensure the models operate well when applied to new, untested data. By preventing overfitting, optimizing model performance, and offering trustworthy smartphone pricing predictions, this method improves the effectiveness and reliability of the automated loan platform.

## **DIFFERENT APPROACHES**

Data partitioning can be done in different methods:

**Train-test division**: The process of train-test partitioning involves splitting the dataset into two subsets: a training set for model training, and a testing set for model evaluation.

**Cross Validation**: It is a resampling strategy that divides the dataset into several subsets of training and validation sets iteratively to assess the performance of the models. Training and testing a model on several subsets of the data lowers the likelihood of overfitting and offers a more reliable evaluation of the model's performance, which shows how well the model will generalize to new data.

## **APPROACH IN THIS PROJECT**

 The straightforward train-valid-test split approach will be used in this project, allocating **70%** of the dataset for training and **30%** for validation. This gives the models enough information to identify patterns and connections in the data and helps assess how well the trained models generalize to new data that is test data.

## **IMPORTANCE**

It ensures that the classification models are robust and reliable when deployed in real-world scenarios, contributing to better decision-making and risk assessment.

# **GOAL: CLASSIFYING BORROWERS**

The primary objective is to identify individuals who are at risk of defaulting on their loans and may require additional assistance to ensure timely repayment. Those IDs might be targeted for loan repayment outreach. By accurately predicting which borrowers are more likely to default, this system can allocate resources and engage in targeted outreach efforts to reduce the risk of loan defaults. These IDs require additional support. Here we will be using **active records to test**, by which we can assess the effectiveness of the classification models in predicting the future outcomes of these borrowers.

## **MODEL SELECTIONS**

### **LOGISTIC REGRESSION**

The glm() model is well suited for classification problems that give probabilities of belonging to a particular class. The implementation and interpretation of logistic regression are made easier by its computing efficiency and simplicity. This is a simple yet powerful model for binary classification.

**Stepwise** can be performed to check the important predictors.

### **DECISION TREES**

Decision trees partition the feature space into regions and make predictions based on the majority class in each region. They are useful in this project for their simplicity and interpretability, making them easy to understand and explain to stakeholders.

### **K-NEAREST NEIGHBORS (KNN)**

KNN is a non-parametric algorithm that makes predictions based on the majority class of the k nearest neighbors. It might not be an appropriate approach for this project as KNN's performance decreases as the dimensionality of the data increases. This is known as the **"curse of dimensionality"**.

### **NAÏVE BAYES:**

Naive Bayes classification is a probabilistic algorithm that is based on Bayes' Theorem, with the "naive" assumption that features are conditionally independent given the class label. This model works better with categorical features which means it might not be suitable in this case.

## **MODEL PERFORMANCE**

After building the models, their performance is evaluated. Sensitivity measures how well the models identify payoffs, while specificity shows their ability to recognize default loans correctly. It is important to assess sensitivity and specificity because they provide insights into how well the models perform in identifying payoffs and default loans, respectively. **KNN and Naïve Bayes** need **not be considered** for model selection as they might not be suitable for this dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| Models | Accuracy | Sensitivity | Specificity |
| Logistic Regression | 77.11% | 86.50% | 61.05% |
| Stepwise | 77.16% | 86.45% | 61.27% |
| Classification Tree | 76.37% | 84.35% | 62.72% |
| KNN | 75.48% | 81.25% | 65.39% |
| Naïve Bayes | 76.08% | 80.68% | 68.03% |

Table 7: Model Performance

According to the business goal it is **important** to identify the **default loans** so that the system can initiate communication with the defaulter to understand their circumstances and explore options for repayment or restructuring of the loan or may pursue legal action against the defaulter for repayment. The system can also engage **third-party debt collection agencies** to recover the outstanding debt on their behalf. These agencies specialize in contacting and negotiating with defaulters to secure repayment.

## **BEST MODEL**

Among the logistic regression, stepwise, and classification tree models, the **classification tree model** has the highest specificity at 62.72%. Specificity is crucial in this context because it represents the ability of the model to correctly identify default loans, which is essential for the business goal of taking appropriate actions.

Now considering this model, we need to predict on the ongoing dataset that is separated so that we will know who will pay and who will not pay in the future. By this we can take appropriate measures on defaulters.

# **PREDICTING THE STATUS ON ONGOING DATA**

Using this model, we will predict the loan repayment status of borrowers in an ongoing dataset. By doing so, it aims to distinguish between borrowers who are likely to fulfill their loan obligations and those who may default. This predictive insight will enable us to implement appropriate measures for managing default risks effectively.

The ongoing data consists of 8256 records where no.of **defaulters predicted is 1579** and rest all borrowers 6677, have paid off according to the predictions. So, these defaulters need extra measures. Ultimately, by identifying potential defaulters beforehand, we can take proactive steps. This proactive approach helps to reduce potential losses and optimize loan portfolio management.

# **CONCLUSION**

This project highlights the critical importance of understanding mortgage repayment dynamics for lenders and borrowers in the real estate finance industry.

By analyzing historical repayment patterns and building predictive models, lenders can optimize their loan portfolio management strategies to minimize default risks and maximize profitability. By identifying individuals who are at risk of defaulting on their loans, lenders can allocate resources and handle the risk of loan defaults. This helps minimize financial losses, improve overall portfolio performance, and enhance customer satisfaction.

# **FUTURE WORK**

Refine loan approval criteria based on insights from repayment patterns to attract reliable customers with lower default probabilities. Adjust interest rates, loan terms, and eligibility requirements to align with the risk profile of potential borrowers and optimize profitability.

# **EXECUTIVE SUMMARY**

**NAME:** Sreeja Reddy Singidi

**DATE:** 04-24-2024

**OPPORTUNITIES:**

The automated loan approval system gives big chances for banks to make things easier, lower risks, and make customers happier. Also, these systems help banks spot risks early by looking at borrower information right away, so they can make smarter decisions and prevent defaults. Plus, they save money by doing things automatically and following rules better, making banks better prepared for the future and more competitive in the loan world that's always changing.

**SOLUTIONS:**

To achieve the goal, the system should focus on predicting if someone might not be able to pay back a loan so that special assistance will be provided to those borrowers. These classification models will look at information about borrowers and help the system make smart decisions about who to lend money to. Also, the system needs to keep an eye on how well it works and make improvements when needed to keep up with changes in the loan market.