Prepayment Risk Analysis Using Mortgage Backed Securities (MBS)

TECHNOCOLABS SOFTWARES

INDORE

SUBMITTED BY

Mr. SHREYAS TARFE

M.Sc. (Applied Statistics)

PRN:23060641094

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS
STATISTICAL INSTITUTE

ACADEMIC YEAR 2023 - 24

Under the guidance of

YASEEN SHAH

Designation: DIRECTOR

Email Id: shahyaseen71@gmail.com

Mobile: 8319291391

Contents

1. Introduction:

This report provides a detailed analysis of a dataset, focusing on the preprocessing steps and initial data exploration performed in a Jupyter Notebook. The primary goal of the analysis appears to be the preparation and cleaning of data for further use, with a particular emphasis on filtering out rows containing specific unwanted data and exploring key features of the dataset. The dataset itself seems to be related to financial data, potentially focusing on mortgage or loan information, as suggested by the column names such as CreditScore, FirstPaymentDate, and EverDelinquent.

2. Data Preprocessing:

In the preprocessing stage, the dataset was filtered to remove rows that contained the character 'X' in any of its columns. This step is crucial for ensuring the quality and integrity of the data, especially if 'X' represents missing, erroneous, or placeholder data that could skew the results of subsequent analyses.

- Filtering Method: The filtering was done using a lambda function applied across all rows, checking if any column in a row contains 'X'. If such a condition was met, the row was excluded from the final dataset.
- Output: The cleaned dataset was saved as a new CSV file named filtered_dataset.csv. This file can be used for further analysis without the risk of including erroneous or placeholder data.

3. Data Exploration:

Following the preprocessing, the notebook explores the dataset by displaying the first 10 rows of a dataframe (df2). This step provides a snapshot of the data's structure, allowing for an initial assessment of the key variables.

- **Key Columns:**
  - CreditScore: Represents the credit score of the borrower.
  - FirstPaymentDate: The date of the first payment on the loan.
  - MaturityDate: The date by which the loan is expected to be fully repaid.
  - Occupancy: Indicates the occupancy status of the property (likely owner-occupied, investor, etc.).
  - OCLTV: The combined loan-to-value ratio.
  - DTI: Debt-to-Income ratio, an important indicator of borrower risk.
  - EverDelinquent: A binary variable indicating whether the borrower has ever been delinquent on the loan.

- Insights: The data presented shows a variety of factors that could be important for assessing the risk and performance of loans. For instance, the CreditScore and DTI ratios are critical for evaluating a borrower's creditworthiness, while the EverDelinquent variable is directly tied to past repayment behavior.

Exploratory Data Analysis :

1.Histogram

- **CreditScore**: The majority of the credit scores are clustered between 600 and 800, with a peak around 700-750. There are very few low or extremely high credit scores.

- **MSA**: The distribution of Metropolitan Statistical Area (MSA) codes shows a wide range with certain MSAs having a much higher frequency than others. There are distinct peaks, indicating that some MSAs are much more represented in the data.

- **MIP**: The Mortgage Insurance Premium (MIP) values are highly skewed, with most values clustered at the lower end, especially around 0. There are few entries with higher MIP values.

- **Units**: Almost all the entries have 1 unit, with very few cases having 2 or more units.

- **OCLTV**: The Combined Loan-to-Value (CLTV) ratio distribution is slightly skewed to the right, with a peak between 80 and 100. There are very few cases with CLTV below 60 or above 100.

- **DTI**: The Debt-to-Income (DTI) ratio is distributed more uniformly, with peaks around 30-40. There are fewer entries with very low or very high DTI ratios.

- **OrigUPB**: The Original Unpaid Principal Balance (UPB) has a right-skewed distribution, with most values between $100,000 and $300,000. There are very few large loan amounts.

- **LTV**: The Loan-to-Value (LTV) ratio shows a distribution with peaks around 80-100, similar to OCLTV, indicating a strong correlation between the two features.

- **OrigInterestRate**: The Original Interest Rate shows a normal distribution centered around 3-6%, with very few cases having extremely low or high rates.

- **OrigLoanTerm**: The loan term is almost entirely concentrated at 360 months (30 years), with very few loans of shorter terms.

- **EverDelinquent**: This binary feature shows that most loans have never been delinquent, with a small proportion that has.

- **MonthsDelinquent**: The number of months delinquent is heavily skewed towards zero, indicating that most loans have not been delinquent for many months, if at all.

- **MonthsInRepayment**: This feature has a long right tail, with most loans being in repayment for less than 100 months.

2.Scatter Plots :

The scatter plots show various variables plotted against an index (likely a data sample or observation number):

- **CreditScore vs. Index**: Shows a positive trend, where the CreditScore increases with the index, possibly indicating that higher index values correspond to higher credit scores.
- **MSA vs. Index**: This plot is highly scattered with no clear pattern, suggesting that the Metropolitan Statistical Area (MSA) data is fairly evenly distributed across the index.
- **MP (Mortgage Payment) vs. Index**: This shows a few distinct lines with scattered points, indicating some grouping in the mortgage payment data.
- **Units vs. Index**: This plot seems to have a lot of zero or low values, indicating many samples have the same or no units.
- **OCLTV (Original Combined Loan-to-Value Ratio) vs. Index**: Scattered plot with no clear trend, indicating a random distribution across the index.
- **DTI (Debt-to-Income Ratio) vs. Index**: Similar to OCLTV, shows a random distribution without any clear trend.
- **OriginalUPB (Unpaid Principal Balance) vs. Index**: Shows a wide scatter, indicating variability in unpaid balances across the index.
- **InterestRate vs. Index**: There seems to be a flat line pattern, suggesting that the interest rate might be relatively consistent across the samples.
- **EverDelinquent vs. Index**: Shows a lot of zero values, indicating many samples have not experienced delinquency.
- **MonthsDelinquent vs. Index**: Scattered points at low values, suggesting most delinquencies are recent or limited in duration.
- **MonthsInRepayment vs. Index**: Scattered with some noticeable groupings, indicating variability in repayment durations.

3. Pie Charts

- **Pie Chart of Occupancy**:

The chart represents the distribution of different types of occupancy.The largest portion (94.8%) is labeled as "O," indicating the majority category.Other categories are "I" (3.0%), "S" (2.2%), and "Other" (0.0%), which occupy much smaller portions of the distribution.

- **Pie Chart of FirstTimeHomebuyer**:

This chart shows the distribution of first-time homebuyers.The majority category (64.5%) is labeled as "N," indicating those who are not first-time homebuyers.The next largest group

(25.8%) is labeled as "X," followed by "Y" (9.6%).There is a small or negligible category labeled "Other" (0.0%).

- **Pie Chart of PPM :**

This chart shows the distribution of first-time PPM.The majority category (96.6%) is labeled as "N," indicating majority category.The next largest group (2.0%) is labeled as "X," followed by "Y" (1.3%).There is a small or negligible category labeled "Other" (0.0%).

- **Pie Chart of Product type :**

This chart shows the distribution of first-time Product type.The majority category (100%) is labeled as "FRM," indicating majority category. There is a small or negligible category labeled "Other" (0.0%).

Pie Chart of Property State: This pie chart indicates the proportion of different sellers in the dataset. "Other" makes up the largest portion (46.4%), followed by "CA" (14.7%), "FL" (6.3%), "MI" (6.3%), and others with smaller percentages. This suggests that a few Property state dominate the dataset, with a long tail of smaller property state.

- **Pie Chart of Property type :**

This chart shows the distribution of Property type.The majority category (84.4%) is labeled as "SF," indicating majority category .The next largest group (9.2%) is labeled as "PU" followed by "CO" (6.1%).. There is a small or negligible category labeled "Other" (0.0%).

- **Pie Chart of Loan Purpose :**

This pie chart indicates the proportion of different loan purpose in the dataset. "P" makes up the largest portion (41.4%), followed by "N" (37.6%), "FL" (6.3%), "C" (20.5%), and others with smaller percentages.

- **Pie Chart of Num Borrowers :**

This pie chart indicates the proportion of different Num Borrowers in the dataset. "2" makes up the largest portion (64.6%), followed by "N" (37.6%), "1" (35.3%).

Pie Chart of SellerName: This pie chart indicates the proportion of different sellers in the dataset. "Ot" (likely short for Other) makes up the largest portion (28.9%), followed by "CO" (12.9%), "FL" (9.6%), "FI" (9.2%), and others with smaller percentages. This suggests that a few sellers dominate the dataset, with a long tail of smaller sellers.

4. Boxplots

The boxplots depict the distribution of various variables grouped by whether or not someone is a first-time homebuyer:

- **CreditScore by FirstTimeHomebuyer**: The distribution of credit scores is shown for first-time and non-first-time homebuyers. The credit scores seem to be slightly lower for first-time homebuyers.
- **MSA by FirstTimeHomebuyer**: Shows the distribution of MSAs, with no clear distinction between first-time and non-first-time homebuyers.
- **MP by FirstTimeHomebuyer**: Mortgage payments have a similar distribution across both groups, with some outliers.
- **Units by FirstTimeHomebuyer**: Both groups have a significant number of zero values, with few higher values scattered.
- **OCLTV by FirstTimeHomebuyer**: The loan-to-value ratios appear slightly higher for first-time homebuyers.
- **DTI by FirstTimeHomebuyer**: Debt-to-income ratios are similar across groups with outliers.
- **OriginalUPB by FirstTimeHomebuyer**: Higher unpaid principal balances are slightly more common among non-first-time homebuyers.
- **InterestRate by FirstTimeHomebuyer**: Interest rates seem consistent across both groups.
- **EverDelinquent by FirstTimeHomebuyer**: This plot has many zeros, indicating most have never been delinquent, with a similar pattern in both groups.
- **MonthsDelinquent by FirstTimeHomebuyer**: Shows outliers for delinquency months, indicating some individuals have experienced longer periods of delinquency.
- **MonthsInRepayment by FirstTimeHomebuyer**: Similar distribution between both groups with some variability.

Heatmap of Pearson Correlation Coefficients:

- This heatmap visualizes the correlation between different features.
- Correlation values range from -1 to 1, where:
    - 1 indicates a perfect positive correlation.
    - -1 indicates a perfect negative correlation.
    - 0 indicates no correlation.

Principal Component Analysis (PCA):

PCA is applied as a dimensionality reduction technique after data preprocessing. PCA is used to transform the high-dimensional dataset into a set of uncorrelated variables, known as principal components, which capture the most variance in the data.

Steps Involved in PCA:

1. Data Standardization:

The dataset is first standardized to ensure that all variables contribute equally to the PCA. This is done using the `StandardScaler` from `sklearn.preprocessing`, which scales the data to have a mean of 0 and a standard deviation of 1.

2. Applying PCA:

PCA is then applied to the standardized data. The `PCA` class from `sklearn.decomposition` is used to compute the principal components.

3. Explained Variance Ratio:

The explained variance ratio for each principal component is calculated. This ratio indicates the proportion of the dataset's variance that each principal component accounts for. The scree plot is generated to visualize the explained variance ratio for each component, helping to determine how many principal components should be retained.

4. Principal Components:

The principal components themselves are analyzed and stored in a DataFrame. This allows for the inspection of how each original feature contributes to the principal components.

5. 3D Scatter Plot:

A 3D scatter plot is created for three selected features (Credit Score, DTI, and OrigUPB) to visualize the data in three dimensions before applying PCA. This visualization helps to understand the distribution of data points across these features.

Purpose of PCA :

The primary purpose of applying PCA in this code is to reduce the dimensionality of the dataset while retaining as much variance as possible. This reduction is crucial for simplifying the dataset, improving the performance of machine learning algorithms, and mitigating multicollinearity issues. By analyzing the explained variance ratio and the principal components, one can decide the number of components to retain for further analysis or modeling.

Classification models:

Logistic Regression
A linear model used for binary classification is called logistic regression. By fitting the data to a logistic function, it forecasts the likelihood of a binary outcome. When there is roughly a linear relationship between the features and the target variable, the model performs especially well.

Decision Trees
A non-linear model called the Decision Tree Classifier creates a tree structure by dividing the data into subsets according to feature values. A decision rule is represented by each node, and an outcome is represented by each leaf. Although this model is adaptable and capable of capturing intricate relationships between features, overfitting is a risk.

Naive Bayes Classifier

Naive Bayes is a probabilistic model based on Bayes' theorem. It assumes that the features are independent of each other given the target variable, which is often not true in practice, hence the term "naive." Despite this assumption, Naive Bayes performs well on various tasks, particularly when the independence assumption is approximately valid.

Measures of Performance
The models were assessed using the following performance metrics:
The proportion of correctly predicted cases to all instances is known as accuracy.
Precision is defined as the ratio of the model's total number of successful predictions to its true positive predictions.
The ratio of true positive predictions to all actual positives in the dataset is known as recall (sensitivity).
The F1-Score is a balanced indicator of model performance that is calculated as the harmonic mean of precision and recall.
Region Higher values indicate better performance. Under the Receiver Operating Characteristic Curve (AUC-ROC): A measure of the model's ability to distinguish between the classes.

Results and discussion:

Logistic Regression achieved the highest AUC-ROC score of 0.87, indicating a strong ability to distinguish between borrowers who will be delinquent and those who will not. The model also performed well in terms of precision and recall, with a balanced F1-Score.

The Decision Tree Classifier had the lowest performance metrics among the three models, with an AUC-ROC score of 0.82. The model is prone to overfitting, which may explain its lower generalization ability compared to Logistic Regression. However, it is still a useful model when interpretability is crucial, as decision trees provide clear decision rules.

Naive Bayes performed moderately well, with an AUC-ROC of 0.86. Despite its simplicity and the independence assumption, it provided competitive performance, particularly in precision. The model is also computationally efficient, making it suitable for large datasets.

The model that performed the best in this task was logistic regression, which offered a good balance between accuracy, precision, recall, and AUC-ROC. Because features and the target variable have an approximately linear relationship, it is especially well suited for this problem. Even though it is less effective, the Decision Tree Classifier can still be helpful in situations where interpretability is crucial. Even though it was straightforward, the Naive Bayes Classifier performed competitively, particularly in situations where computational efficiency is crucial.This study focuses on utilizing Ridge Regression and Linear Regression, two regression models, to predict mortgage prepayment. The continuous variable "prepayment," the target variable, indicates the amount or probability of paying off a mortgage before the end of the agreed-upon term. Using suitable evaluation metrics, we will compare these models' performances in terms of their capacity to forecast this continuous outcome.

Regression Models:

Linear Regression

Linear Regression is a fundamental and widely used model in regression analysis. It assumes a linear relationship between the independent variables (features) and the dependent variable (target). The model tries to minimize the sum of squared differences between the observed and predicted values, yielding the best-fitting linear line through the data.

Ridge Regression

Ridge Regression is an extension of Linear Regression that includes a regularization term (L2 penalty) in the loss function. This regularization helps to prevent overfitting by penalizing large coefficients, thus improving the model's generalization to unseen data. Ridge Regression is particularly useful when dealing with multicollinearity, where independent variables are highly correlated.

Measures of Performance
The models were assessed using the following performance metrics:
A measure of prediction accuracy called Mean Absolute Error (MAE) is calculated by averaging the absolute differences between the values that were predicted and the actual values.
The average of the squared differences between the expected and actual values, with a higher weight assigned to larger errors, is called the mean squared error, or MSE.
Root The square root of mean squared error (MSE) gives an error measurement in the same units

as the target variable.
Model fit is indicated by R-squared (R²), which is the percentage of the target variable's variance that can be predicted from the independent variables.

Results and discussion:

With an R-squared value of 0.82, which indicates that 82% of the variance in the "prepayment" variable can be explained by the model, Linear Regression produced a good baseline model. But if there is multicollinearity in the data or if there isn't a strictly linear relationship between the features and the target variable, the model might be prone to overfitting.
With a lower MAE, MSE, and RMSE value and a higher R-squared value of 0.85, Ridge Regression performed better than Linear Regression in every metric. The regularization term's inclusion lessened overfitting and improved the model's ability to generalize to new data. When the independent variables have a high degree of correlation, this improvement is especially apparent.In comparison to Linear Regression, Ridge Regression proved to be a more accurate and more general model for predicting mortgage prepayments. Ridge Regression's regularization term successfully reduces the possibility of overfitting, particularly when multicollinearity is present. Even though it is more straightforward, Linear Regression can still be a helpful baseline model; however, if the assumptions of linearity and independence between features are broken, it may not perform as well.

Other regularization methods, such as Elastic Net or Lasso Regression, might be investigated in future research to improve model performance even more. Furthermore, feature engineering and the incorporation of domain-specific knowledge may enhance the models' capacity for prediction.

Key Points:

- **Pipeline Benefits:** Pipelines streamline the machine learning workflow by automating the sequence of preprocessing and modeling steps. This ensures consistent data handling during training and prediction.
- **Handling Missing Data:** The inclusion of SimpleImputer helps address missing values, which are common in real-world datasets.
- **Feature Scaling:** StandardScaler ensures that features are on a similar scale, which is often necessary for optimal model performance.
- **Categorical Encoding:** OneHotEncoder converts categorical features into a format suitable for machine learning algorithms.

- **Model Combination:** The pipeline cleverly combines the outputs of two different models, demonstrating how to integrate different types of predictions into a unified workflow.

Data Preparation and Feature Engineering

- **Feature and Target Separation:** The code begins by separating the dataset into features (predictors) stored in 'X' and target variables for logistic regression ('EverDelinquent') and linear regression ('Prepayment').
- **Feature Type Identification:** It then identifies which columns in the dataset contain categorical data (e.g., text categories) and which contain numerical data (numbers).
- **Preprocessing Pipelines:** Two preprocessing pipelines are established:
  - **Numerical Pipeline:** This pipeline standardizes the numerical features (scales them to have a mean of 0 and a standard deviation of 1) using StandardScaler.
  - **Categorical Pipeline:** This pipeline one-hot encodes the categorical features, transforming them into a format suitable for machine learning algorithms. The handle_unknown='ignore' parameter ensures that if any new categories are encountered during prediction, they will be gracefully handled without causing errors.
- **ColumnTransformer:** The ColumnTransformer combines these two preprocessing pipelines, applying the appropriate transformations to the corresponding columns in the dataset.

Model Building and Training

- **Pipelines for Logistic and Linear Regression:** Two machine learning pipelines are created:
  - **Logistic Regression Pipeline:** This pipeline first applies the preprocessing steps defined in the ColumnTransformer, then trains a logistic regression model to predict the binary outcome 'EverDelinquent' (whether or not a borrower will become delinquent).
  - **Linear Regression Pipeline:** This pipeline also applies the same preprocessing steps, then trains a linear regression model to predict the continuous numerical outcome 'Prepayment' (the amount a borrower is likely to prepay on their loan).

Acknowledgement: -

I would like to express my sincere gratitude to Mr. Yaseen Shah, Director of Technocolabs Softwares, for his invaluable guidance and support throughout the course of this project. His expertise and insights were instrumental in shaping the direction and execution of this analysis.

I would also like to thank the entire team at Technocolabs Softwares for providing the resources necessary for the successful completion of this work. Their dedication to fostering innovation and excellence has been a significant driving force behind this project.

Finally, I extend my appreciation to all my colleagues and mentors who offered their assistance and encouragement during the various stages of this project. Your contributions have been greatly appreciated.

References: -

https://github.com/Technocolabs100/Attrition-Analysis-and-Prediction

Kaggle

YouTube Channels such as: Great learning, Nptel, etc.

Edura

Forage