ASSESSING BIAS IN MORTGAGE LENDING AND PAYMENT TRENDS DURING THE COVID-19 PANDEMIC USING MACHINE LEARNING

Table of Contents

Introduction

The COVID-19 pandemic brought significant financial challenges to households across the United States, with unemployment, income reduction, and economic uncertainty affecting borrowers' ability to meet their mortgage obligations. However, these impacts were not evenly distributed. Racial disparities in mortgage lending and payment delinquencies, long an issue of concern before the pandemic, were further exacerbated by the economic fallout. Black and Hispanic borrowers, already vulnerable to higher interest rates, loan denials, and predatory lending practices, faced even greater risks during this period. The pandemic exposed systemic inequalities in the housing market that disproportionately affected minority homeowners. Understanding these disparities is crucial not only for addressing immediate financial insecurities but also for designing more equitable housing policies in the future.

This research aims to examine how racial disparities in mortgage lending and payments evolved during the COVID-19 pandemic and evaluate the effectiveness of mortgage relief programs in mitigating the impact on minority borrowers compared to White borrowers. Specifically, the study will investigate whether Black and Hispanic homeowners were more likely to experience mortgage payment difficulties compared to White homeowners, even when accounting for financial characteristics such as income, loan-to-value, and credit score. Using supervised machine learning, this study aims to provide significant insights into structural factors contributing to these disparities and inform policy recommendations for equitable housing solutions.

The data from the Federal Home Loan Bank System’s Public Use Database (2020-2022) is leveraged to employ advance multivariate analysis to identify patterns in mortgage delinquency risk among first-time homebuyers. The findings will expose the systemic nature of lending disparities and offer evidence-based guidance for addressing inequities in the housing market.

Question

What is the impact of the demographic characteristics of first-time homebuyers to predict mortgage delinquency rate?

Literature Review

The debate surrounding racial disparities in mortgage lending, payment delinquencies, and refinancing opportunities has persisted long before the onset of COVID-19. Research highlights that Black and Hispanic households are disproportionately more likely to rent than white households, and even among mortgage holders, these groups are more prone to falling behind on payments. However, banks like the Federal Reserve Bank of Minneapolis (2024) have contested the notion that mortgage lending itself causes racial disparities. Their analysis, using regression models to control for factors such as income, credit scores, and debt-to-income ratios, found that White applicants were more often denied loans due to collateral issues, whereas Black and Latino applicants faced higher denial rates due to their debt-to-income ratios. This highlight underlying structural financial inequities that disproportionately affect minority borrowers. Although these factors are well-researched, what remains less explored are the long-term implications of these disparities on generational wealth and homeownership sustainability.

The COVID-19 pandemic severely impacted the U.S. housing and mortgage markets, exacerbating pre-existing disparities. With rising unemployment and financial insecurity, mortgage defaults surged, particularly among minority borrowers. For instance, unemployment rates for Black and Hispanic workers continued to rise even after non-Hispanic White workers experienced a peak in April 2020. By early 2021, Black workers had the highest unemployment rates compared to other racial groups, with a slower recovery trajectory. These impacts underscore the vulnerability of minority populations during economic downturns, highlighting structural inequalities that predate the pandemic. However, while several studies document these disparities, there is limited research on the long-term effects of COVID-19 relief measures on minority homeowners and how effectively they mitigated mortgage payment challenges.

Current literature examining mortgage delinquency trends among minority groups during the pandemic consistently indicates that Black, Hispanic, and Asian households were at a higher risk of mortgage delinquency compared to their White counterparts. Studies have shown that African American homeowners, in particular, are more vulnerable to mortgage payment difficulties. Research by JungHo Park and Dongha Park (2023) found that Black homeowners were nearly twice as likely to be behind on their mortgage payments than White homeowners, supporting the hypothesis that minority groups were more susceptible to mortgage delinquency during the pandemic. However, while these studies effectively document the problem, they often fail to explore the specific factors driving these disparities beyond general economic vulnerability, leaving a gap in the understanding of how policy solutions may be better tailored for these communities.

Before addressing the effects of the pandemic, it is crucial to consider the longstanding housing inequities faced by minority groups in the U.S. Research consistently shows that Black, Hispanic, and Asian borrowers are more likely to face higher interest rates, loan denials, and predatory lending practices compared to White borrowers. For example, data from the Home Mortgage Disclosure Act (HMDA) revealed that Black applicants were denied conventional mortgage loans at more than twice the rate of White applicants. Additionally, Black homeowners seeking to refinance their loans were rejected at double the rate of White applicants, further exacerbating financial inequities. While this data underscores the severity of the issue, the literature lacks a comprehensive examination of how these long-standing disparities have shaped current lending practices and what actionable steps can be taken to address them.

Minority borrowers often have limited access to generational wealth, lower credit scores, and reduced financial flexibility, making them more susceptible to economic shocks like the pandemic. According to recent studies, Black and Hispanic borrowers are less likely to refinance their loans, leading to a loss of equity over time. Data from JPMorgan (2024) highlights that racial disparities in closing costs are particularly pronounced among nonbank lenders and brokers, where Black and Hispanic borrowers face significantly higher disparities compared to their White counterparts. These structural inequities in the lending process further compound the financial difficulties faced by minority homeowners, suggesting that targeted reforms in the nonbank lending sector are needed. However, much of the existing literature focuses on quantitative disparities, with less attention paid to the qualitative experiences of minority borrowers navigating these challenges.

These findings align with pre-pandemic trends but were amplified by the economic downturn. The heightened risk of delinquency for minority groups calls for an in-depth examination of the specific economic conditions that have placed them in this vulnerable position.

The CARES Act introduced mortgage forbearance programs, allowing borrowers to pause payments temporarily without facing immediate foreclosure. Although these programs provided much-needed relief, the extent of their effectiveness for minority borrowers remains questionable. Consumer Financial Protection Bureau (CFPB) (2021) indicates that Black and Hispanic borrowers accounted for a disproportionately high share of loans in forbearance and delinquency, despite making up a smaller portion of the overall mortgage market. This suggests that, while forbearance programs offered a temporary solution, they were insufficient in addressing the broader systemic issues that contribute to higher delinquency rates among minority homeowners.

Further exacerbating the problem, according to the Urban Institute Research Institute (2024), access to mortgage forbearance programs and foreclosure moratoriums varied across demographic lines. Homeowners in predominantly Black neighborhoods were more likely to be unprotected by forbearance programs, while those in predominantly White neighborhoods were less likely to face credit-related delinquencies. The gap in protection left many minority homeowners at risk of foreclosure, especially those with non-agency mortgages. Moreover, the lack of adequate protection caused a further decline in credit scores for these unprotected borrowers, limiting their ability to recover financially post-pandemic.

The article from the American Enterprise Institute (AEI) challenges the notion of pervasive racial bias in mortgage lending by analysing a robust dataset that includes conventional and FHA loans (2023). It found that minority borrowers have higher risk-adjusted default rates compared to non-Hispanic Whites. It suggests that the minority groups receive more lenient lending terms, contrary to claims of systemic discrimination. However, it acknowledges that some lenders might exhibit bias, though their impact in the overall mortgage market is minimal. This finding serves as an opposing viewpoint to the allegations of widespread racial bias in the mortgage industry, arguing instead the disparities in default rates is rooked based on financial risk assessment rather than racial discrimination.

The existing literature thoroughly documents the racial disparities in mortgage lending and delinquency trends, particularly during the COVID-19 pandemic. However, critical gaps remain, particularly in understanding the long-term effectiveness of pandemic relief programs for minority borrowers and how systemic changes in the mortgage industry might reduce racial inequities. While the available studies highlight the structural factors contributing to these disparities, future research must focus on the lived experiences of minority borrowers and explore policy reforms that can create more equitable outcomes in the housing market.

Theory and Hypothesis

The theoretical framework should integrate aspects of structural inequality theory, alongside economic theory to fully encapsulate the systemic nature of disparities in mortgage lending. Structural inequality highlights how institutional practice, polices, and economic systems perpetuate unequal opportunities, particularly for minority borrowers.

H1 (Null Hypothesis): There is no significant relationship between a borrower's racial or ethnic background and the risk of mortgage delinquency when controlling for financial characteristics, such as loan-to-value ratio, housing expenses percentage, income, and credit score.

H2 (Alternative Hypothesis): There is significant relationship between a borrower's racial or ethnic background and the risk of mortgage delinquency when controlling for financial characteristics, such as loan-to-value ratio, housing expenses percentage, income, and credit score.

Data (Study Design)

Public Use Database - Federal Home Loan Bank System will be collected between 2020 and 2022 to investigate the impact of financial and socio-economic characteristics of first-time home buyers to predict the likelihood of mortgage delinquency. Most of the characteristics investigated contain no missing values. These are variables that will be selected in the research: '*TotalMonthlyIncomeAmount'*, '*NoteAmoun*t', '*LTVRatioPercent'*, '*HousingExpenseRatioPercent*', '*TotalDebtExpenseRatioPercent*', '*Borrower1CreditScoreValue*', '*PMICoveragePercent*', '*Borrower1AgeatApplicationYears*', *'PropertyType'*, '*MortgageType*', '*LoanAcquisitionActualUPBAmt*', '*Borrower1Race1Type*', '*Borrower1EthnicityType*', '*Borrower1GenderType*'.

Each of the variables will be measured by:

1. *TotalMonthlyIncomeAmount*: it is a continuous variable (in dollars) from borrower provided application forms
2. *NoteAmount*: a continuous variable representing the loan amount issued to the borrower.
3. *LTVRatioPercent*: a continuous variable that has been calculated in percentage.
4. *HousingExpenseRatioPercent*: a continuous variable that has been calculated in percentage.
5. *TotalDebtExpenseRatioPercen*t: a continuous variable that has been calculated in percentage.
6. *Borrower1CreditScoreValue*: it is ordinal categorical variable which are formed under ranges
7. *PMICoveragePercent*: it is a continuous variable that shows the percentage of mortgage balance at originated covered by low level PMI
8. *Borrower1AgeatApplicationYears*: this is a ratio variable that list the ages in years of a borrower.
9. *PropertyType*: it is a nominal categorical variable which looks at the property types.
10. *MortgageType*: it is a nominal categorical variable that looks at the mortgage types.

11. *LoanAcquisitionActualUPBAmt*: it is a continuous variable in dollars where the amount of unpaid principal balance is acquired by the FHLBank.
12. *Borrower1Race1Type*: it is nominal categorical demographic variable that showcase different races.
13. *Borrower1EthnicityType*: it is a nominal categorical demographic variable that showcase different ethnicities (Non-Hispanic and Hispanic)
14. *Borrower1GenderType*: It is a nominal categorical demographic variable that showcase different genders.

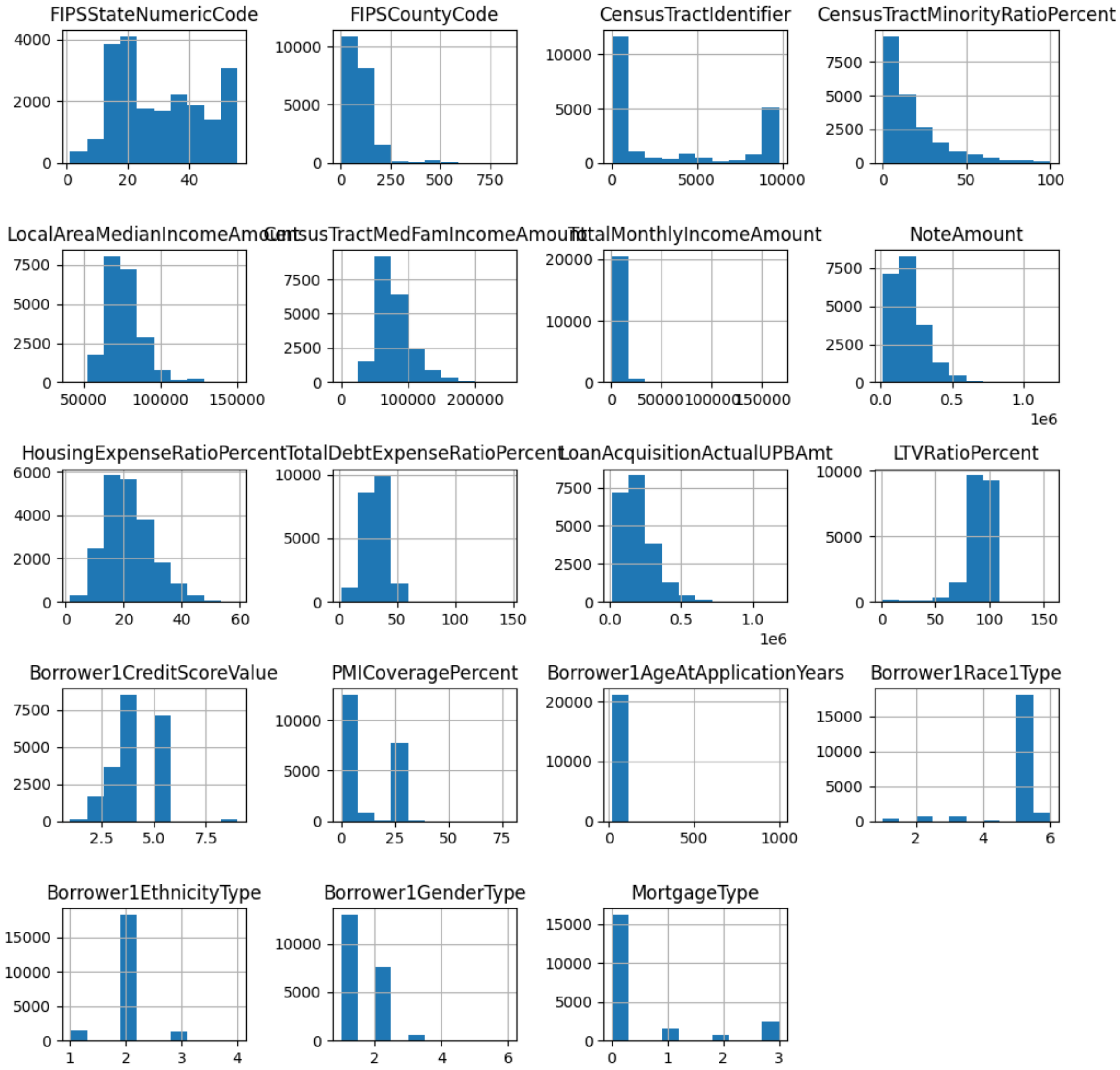


Figure 1: Histogram plots of the features

Method of Analysis (Study Design):

This study employs a quantitative approach using supervised machine learning to analyse racial disparities in mortgage delinquency among first-time homebuyer during the COVID-19 pandemic. The research draws on data from the Federal Home Loan Bank System from 2020 to 2022, focusing on borrowers who purchased primary residences by filtering '*BorrowerFirstTimeHomebuyer*' is equal to 1 and *'PropertyUsageType*' equal to 1.

The dataset includes 21 variables, comprising 13 predictors, one target variable and 7 census-related demographic variables. Key financial predictors include loan-to-value ratio, housing expense ratio, credit score, and total monthly income. Demographic predictors include race, ethnicity and gender. The target variable categorizes 21130 borrowers as either low-risk or high-risk for mortgage delinquency.

The dataset underwent several processing steps to ensure quality. Continuous variables, such as '*TotalMonthlyIncomeAmount'* and '*LoanAcquistionActualUPBAmt*', were log-transformed to normalise their distribution, while standardization ensured consistent scaling across all features. To address potential multicollinearity and reduce dimensionality, the Principal Component Analysis (PCA) was applied, retaining 95% of the variance. Furthermore, the dataset was split into training (80%) and testing (20%) subsets using a random state of 42 to ensure the reproducibility of the results. These preprocessing steps facilitated the effective of machine learning models.

Decision trees is also called a classification tree as the target variable is categorical. The purpose of the decision tree is to produce predictive algorithms for the target variable. The classification trees comprise of root nodes, internal nodes and leaf nodes. The model is formed by top-to-bottom branches where each branch consists of the root node through international nodes to the leaf node. The model algorithm develops a splitting criterion that will partition each internal nodes across the hierarchy of trees and attempts to reduce each node's impurity. This binary partitioning procedure becomes recursive and continues till the stopping rule is fully applied in the tree (Song, 2015). My research used the random state of 42 for my decision tree model.

Furthermore, pruning the decision tree will work in this situation to ensure the model is more interpretable by using the cost complexity pruning. Cost complexity pruning involves finding the optimal parameter for alpha values and checking the accuracy with the pruned trees using the confusion matrix, ROC curve (Receiver Operating Characteristic) and AUC curve (Area Under Curve).

Random Forests model avoids being prone to overfitting compared to Decision Tree model, as the ensembled trees are built under bootstrap samples rather than the principal component sample. The variable for each split in the classification tree and the response variable is predicted as an average vote of the predictions of all trees. The measure that I will use in random forests is the feature importance measure to help identify which features contribute the most predictive power to the target variable (Au, 2018). However, it tends to require more time and data power to make predictions which is not cost effective (GeeksforGeeks, 2024).

Logistic regression model is a predictive model which will be used as the target variable is binary '*Low Risk*' and '*High Risk*'. This model is only effective when using the principal component analysis as it may be sensitive to outliers when since some of the variables were still slightly skewed.

The LightGBM model is an improved version of the boosting model, developed by Microsoft. It uses the leaf-wise decision tree-based gradient boosting method to improve model efficiency. It uses two techniques, Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

I adopted the common performance metrics in the model evaluation that balance, precision, recall, and overall accuracy. Confusion matrices were generated to examine classification performance, while the Receiver Operating Characteristics (ROC) curve and Area Under Curve (AUC) scores assessed model effectiveness in distinguishing between low-risk and high-risk borrowers. F1-scores were also calculated to evaluate the balance between precision and recall, particularly for minority class. These evaluation metrics ensured a robust assessment of the models' predictive capabilities.

Expected Results

The study is anticipated to show the following evidence to accept the null hypothesis according to existing literature and prior analysis of mortgage lending disparities. African and Hispanic homeowner applicants are expected to be more susceptible to being mortgage delinquent compared to non-Hispanic White homeowner applicants. Lending disparities may be more pronounced in metropolitan regions compared to rural areas due to differences in housing markets and lender policies.

While the study leverages nationwide data, lending practices and borrower experiences may vary significantly across counties, states, and metropolitan areas. Aggregating data at the national level might obscure local trends and nuances. Despite its comprehensiveness, the dataset is large and complex, potentially exceeding the processing capacity of internal systems (e.g., IDEs). To address this, a smaller, representative sample may need to be extracted for efficient analysis by cross-validation. Although anonymized, some fields in the dataset may lack granularity or contain missing information, potentially limiting the ability to draw precise conclusions. There may be uncontrolled variable influencing lending outcomes, such as lender-specific policies, credit scores, or local economic conditions, may not be fully captured in the dataset.

Findings (Analysis)

The study is anticipated to show the following evidence to reject the null hypothesis according to existing literature and prior analysis of mortgage lending disparities. African and Hispanic homeowner applicants are expected to experience significantly higher risk of mortgage delinquency compared to non-Hispanic White homeowner applicants. Lending disparities may be more pronounced in metropolitan regions compared to rural areas due to differences in housing markets and lender policies.

While the study leverages nationwide data, lending practices and borrower experiences may vary significantly across counties, states, and metropolitan areas. Aggregating data at the national level might obscure local trends and nuances. Despite its comprehensiveness, the dataset is large and complex, potentially exceeding the processing capacity of internal systems (e.g., IDEs). To address this, a smaller, representative sample may need to be extracted for efficient analysis. Although anonymized, some fields in the dataset may lack granularity or contain missing information, potentially limiting the ability to draw precise conclusions. There may be uncontrolled variable influencing lending outcomes, such as lender-specific policies, credit scores, or local economic conditions, may not be fully captured in the dataset.

As shown in figure 2, the non-Hispanic White first-time home buyers (group 5) are overrepresented in the dataset and considered largely low risk of mortgage delinquency. Whereas, African American (group 3) first-time home buyers show higher delinquency risk, potentially pointing to systemic disparities in income, lending policies and financial stability.

As shown in figure 3, the non-Hispanic white borrowers dominate in being categorised as low-risk individuals at 84.5%, suggesting they have more favourable financial characteristics to being offered a loan. The Hispanic white borrowers have a much smaller share of low-risk and high-risk categories, which demonstrates that they are less likely to be unfavoured in mortgage terms due to socioeconomic differences, and creditworthiness.

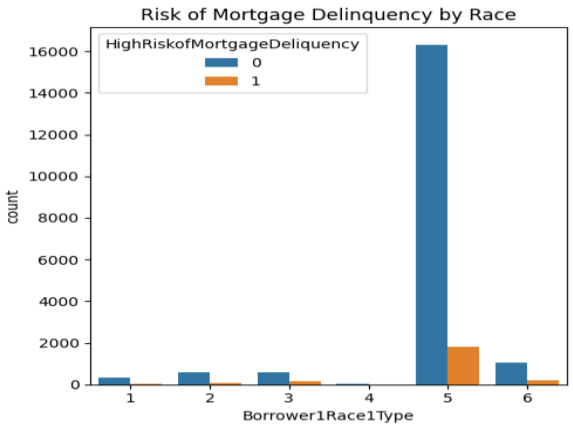


Figure 2: Histogram plot of Low- and High-Risk mortgage delinquency by the race of the first-time homebuyer borrowers

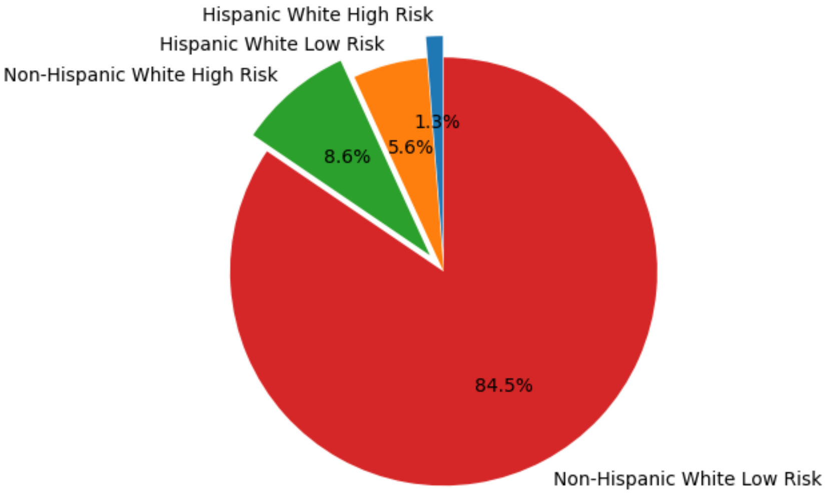


Figure 3: Pie chart of the Low-and High-Risk mortgage delinquency by the Non-Hispanic and Hispanic first-time homebuyer borrowers

In this section, I describe the experimental results. The results are extracted via cross validation for both datasets.

I. Log Transformation

According to the histograms after log transformation, the ‘*TotalMonthlyIncomeAmount*’ and ‘*LoanAcquisitionActualUPBAmt*’ variables in the dataset slightly follows a normal distribution. However, the ‘*LTVRatioPercent*’, ‘*HousingExpenseRatioPercent*’ and ‘*TotalDebtExpenseRatioPercent*' in the dataset was left-skewed. However, there were still outliers for these selected columns: '*TotalMonthlyIncomeAmount*', '*LTVRatioPercent*', '*LoanAcquisitionActualUPBAmt*', '*HousingExpenseRatioPercent*', '*TotalDebtExpenseRatioPercent*'. Therefore, PCA model was used to help reduce the dimensionality for machine learning methods to do multivariate analysis.

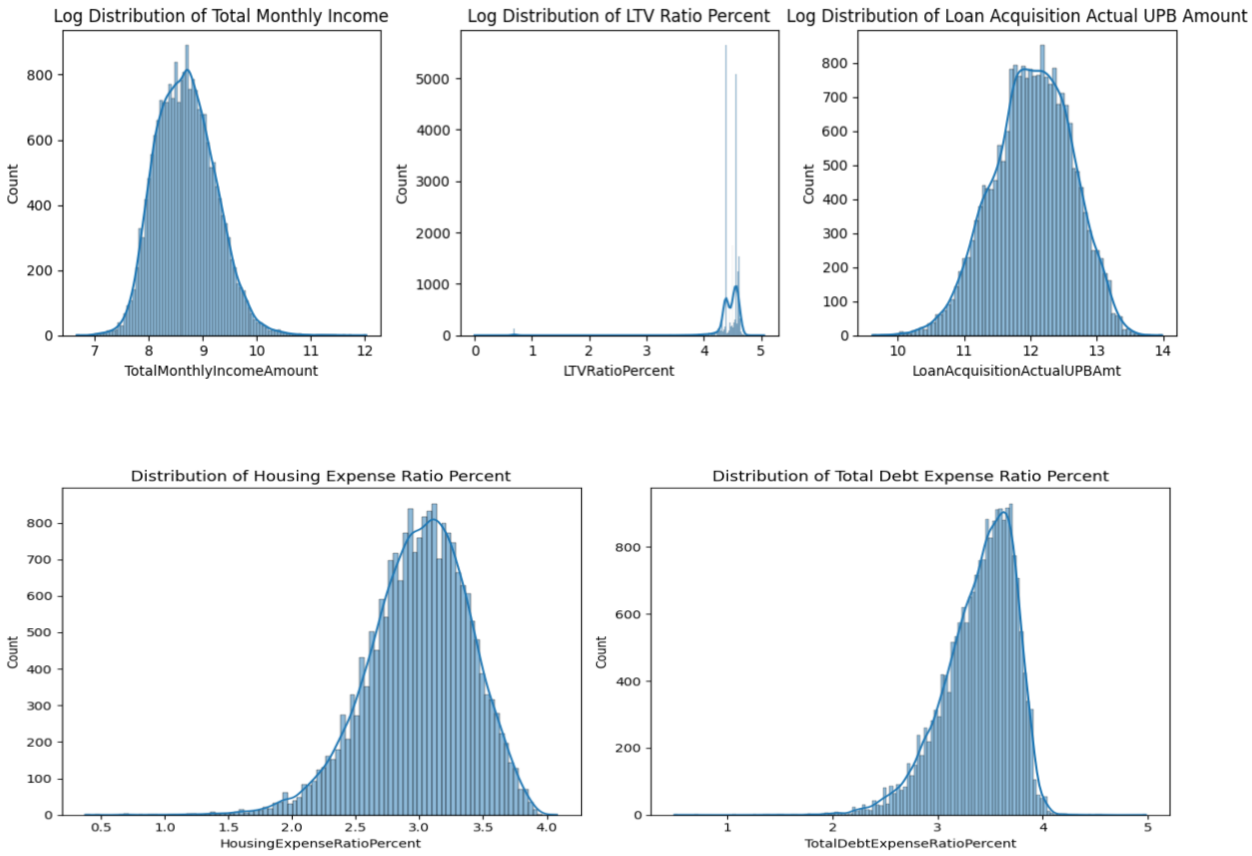
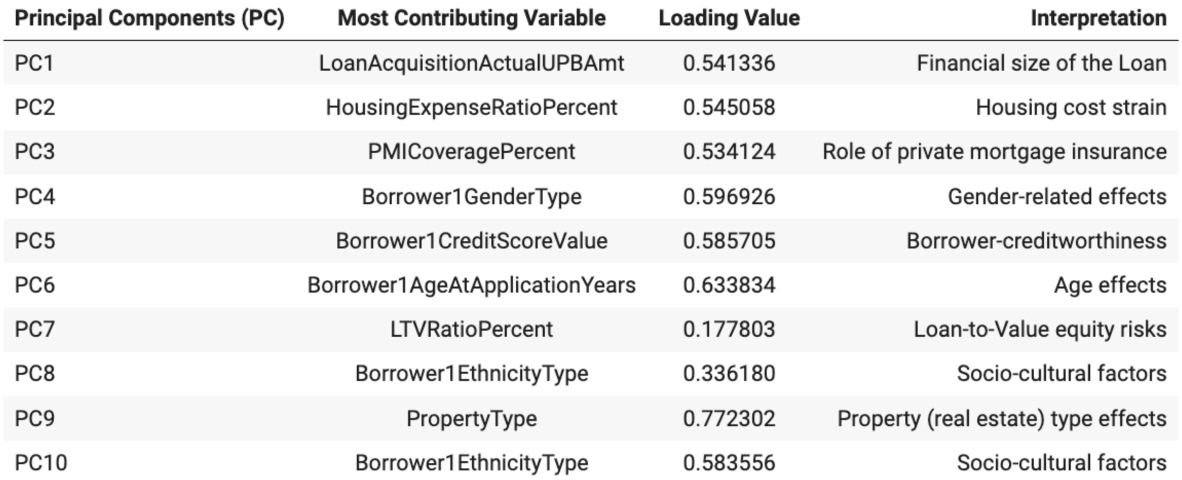


Figure 4: Histogram Plots of the selected features when transformed into logs

II. Principal Component Analysis (PCA)

I partitioned the dataset into training and test with a random state of 42, then scaled the features to visualize the number of features that explains 95% variance. To explain 95%

variance in our normalized *fin_char_data_without_census_v3* dataset, I reduced the dimensionality from 13 to 10 principal components. Then, I calculated the PCA loadings to observe the most contributing variable to each principal component. This table shows the following results:



Principal Components (PC)	Most Contributing Variable	Loading Value	Interpretation
PC1	LoanAcquisitionActualUPBAmt	0.541336	Financial size of the Loan
PC2	HousingExpenseRatioPercent	0.545058	Housing cost strain
PC3	PMICoveragePercent	0.534124	Role of private mortgage insurance
PC4	Borrower1GenderType	0.596926	Gender-related effects
PC5	Borrower1CreditScoreValue	0.585705	Borrower-creditworthiness
PC6	Borrower1AgeAtApplicationYears	0.633834	Age effects
PC7	LTVRatioPercent	0.177803	Loan-to-Value equity risks
PC8	Borrower1EthnicityType	0.336180	Socio-cultural factors
PC9	PropertyType	0.772302	Property (real estate) type effects
PC10	Borrower1EthnicityType	0.583556	Socio-cultural factors

Figure 5: Principipal Components and contributing variables analysis

III. **Decision Tree**

After pruning the tree, the model was more interpretable. The decision tree implies that the creditworthiness, loan size, housing expense ratio and loan-t-value ratio are key factors in determining whether a borrower is at low or high risk for mortgage delinquency. Borrowers with lower loan amounts, better credit scores and lower housing expense ratios are consistently predicted as low risk. Conversely, those with higher financial burdens, including higher loan amounts, housing expense ratios and poor credit scores are flagged as high risk.

If the '*NoteAmount*' ≤ 1.41, they are more likely to be at low risk of mortgage delinquency, because it is easier to manage repayments with smaller loan sizes. Furthermore, if the '*HousingExpenseRatioPercen*t' ≤ 1.90, borrowers are less likely to default as they have smaller housing costs. Borrowers with a '*CreditScoreValue*' > 0.752 have a stronger financial responsibility, thus associated with lower risk of delinquency. Lastly, if the normalized '*LoanAcquisitionActualUPBAmt*' ≤ 0.31, the outstanding amount of unpaid principal balance is lower, reducing the risk of being a delinquent.

To examine the performance of the model, I calculated the ROC curve and AUC. We can see the decision model has an AUC score of 70%.

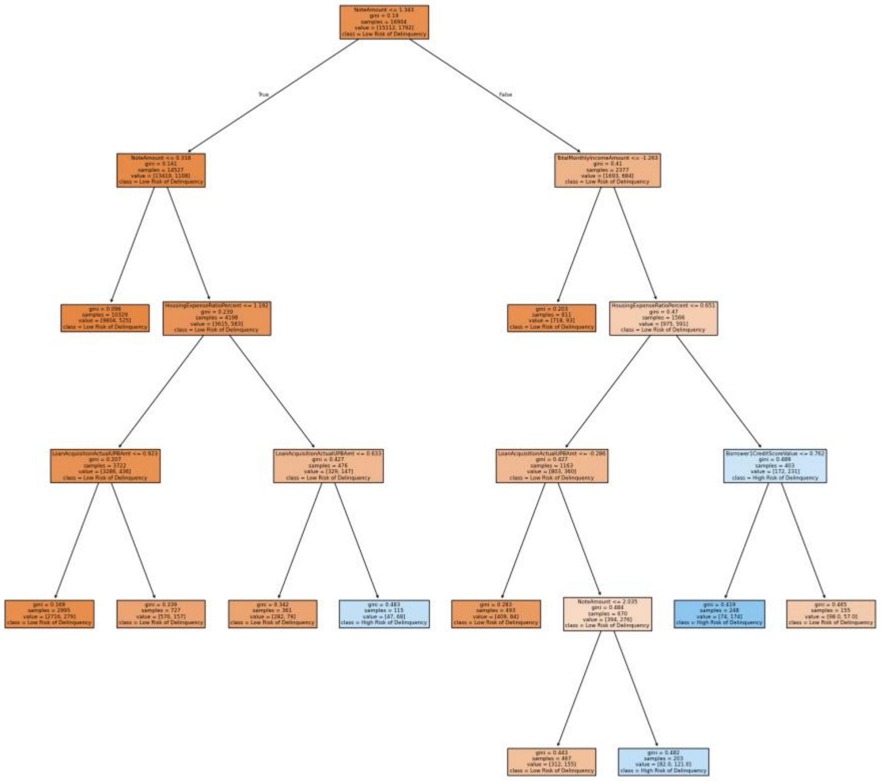


Figure 6: Decision Tree (classification) model without PCA

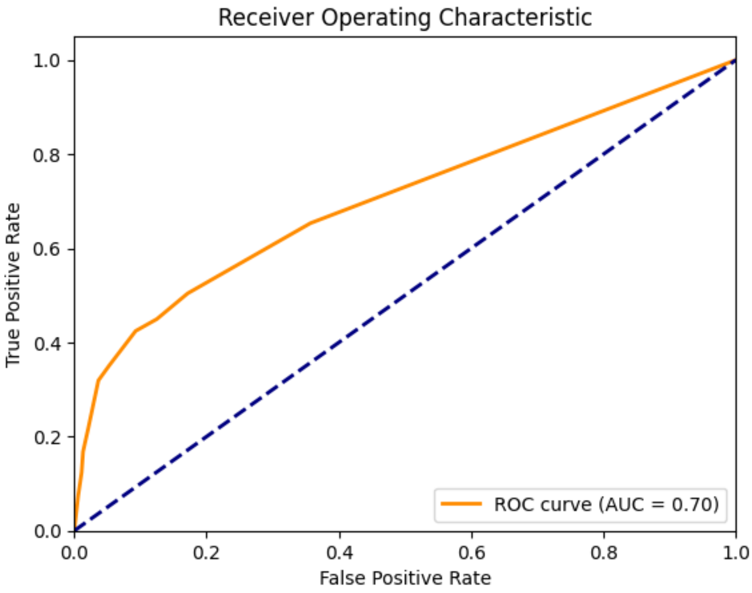


Figure 7: ROC Curve with AUC score using the training set for Decision Tree model

IV. **Random Forest**

We can observe that ‘*HousingExpenseRatioPercent*’ is the most important predictor as housing cost strain is critical in assessing the risk of mortgage delinquency as first-time homebuyer. Furthermore, ‘*TotalMonthlyIncomeAmount*’, ‘*LoanAcquisitionActualUPBAmt*’ and ‘*NoteAmount*’ show a significant importance in predicting the risk. Features, like ‘*Borrower1CreditScoreValue*’, ‘*LTVRatioPercent*’, and ‘*Borrower1AgeAtApplicationYears*’ are moderately relevant in predicting the risk of mortgage delinquency. However, demographic variables including ‘*Borrower1Race1Type*’, ‘*Borrower1EthnicityType*’, and ‘*Borrower1GenderType*’ have minimal importance to assessing in the risk of delinquency rate, which means that race, ethnicity and gender are not highly contributing factors in showcasing risk of mortgage delinquency, as suggested by the existing literatures.

The graph with PCA also expresses the same results in which that ‘*HousingExpenseRatioPercent*’ and ‘*LoanAcquisitionActualUPBAmt*’ have notable importance in predicting the risk of mortgage delinquency.

I examined the performance metric for the random forest model, and it shows that without using the principal components, the AUC score of 81% is slightly higher than using the principal components in the random forest model.

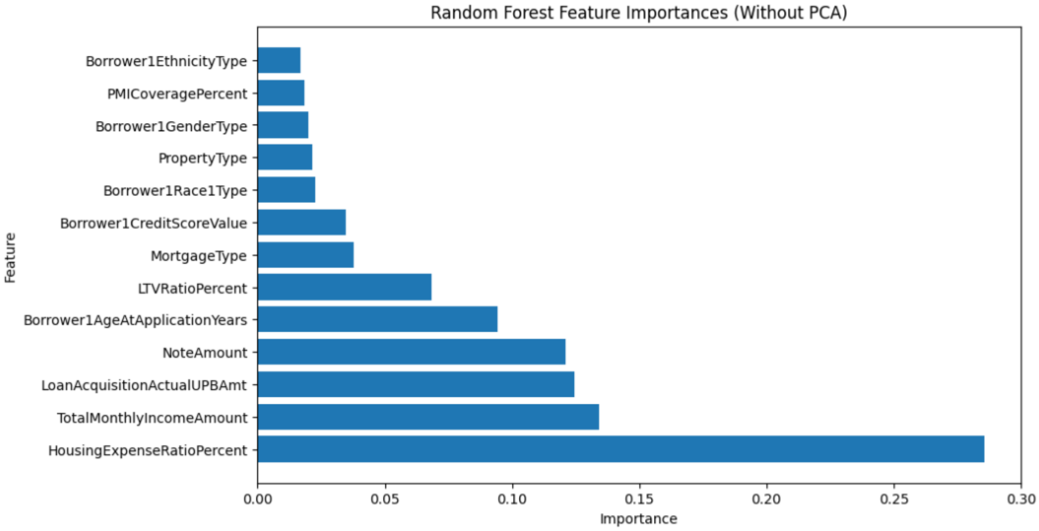


Figure 8: Random Forest Feature Importance (Without PCA)

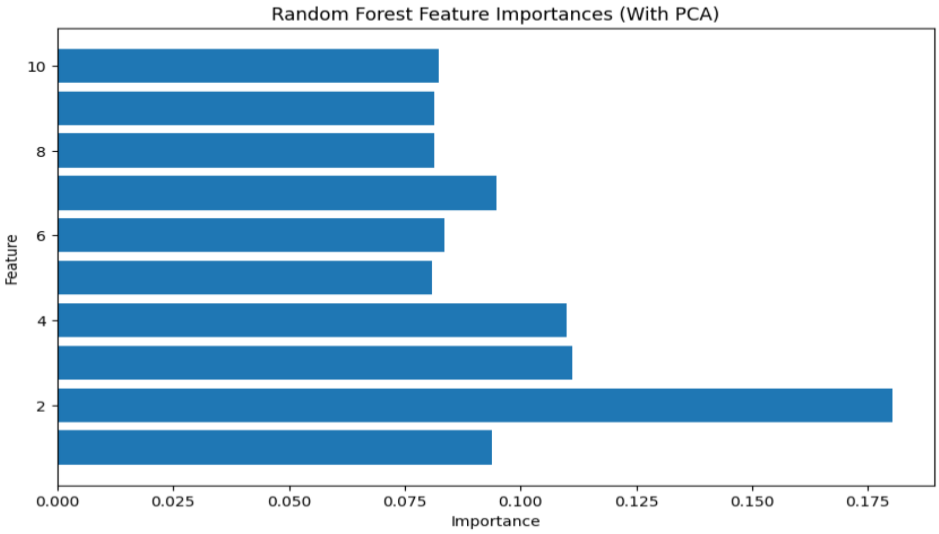


Figure 9: Random Forest Feature Importance (With PCA)

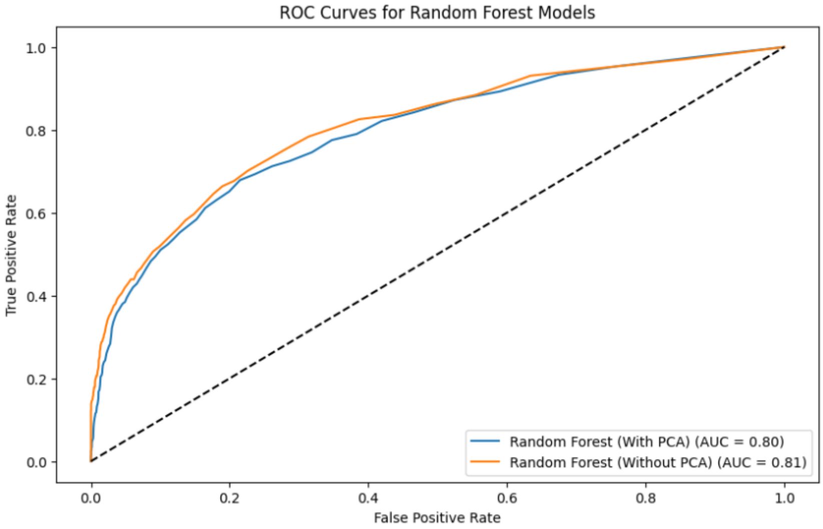
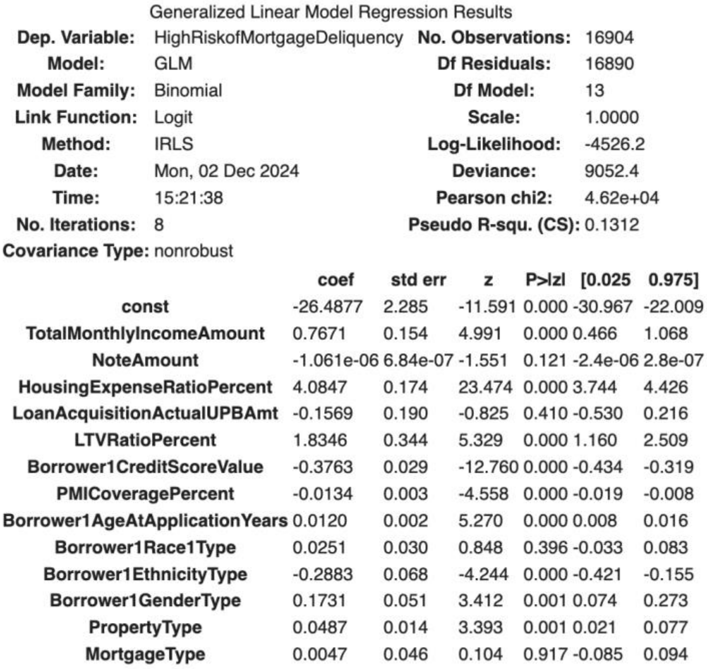


Figure 10: ROC Curve with AUC score using the training set for Random Forest model (With and Without PCA)

V. Logistic Regression

Under the generalized linear model regression analysis, I observed the logistic regression without the PCA, but with some scaled features, for more interpretability. A sample of 16904 observations were used, and the R-squared is 13.125% of the variability in the dependent variable. The model identifies the key financial factors that are significant contributors to mortgage delinquency risk, which are '*TotalMonthlyIncomeAmount*' (income), '*HousingExpenseRatioPercent*' (housing expense ratio), '*Borrower1CreditScoreValue*' (credit score). Furthermore, the '*LTVRatioPercent*' (loan-to-value ratio) and '*PMICoveragePercent*' (PMI coverage) are relevant to assess the risk of mortgage delinquency when managing loans of first-time home buyers. Demographic variables (age, and ethnicity) express as significance predictors as their p-value is less than 1% significance level. Thus, this shows that they may play a role in determining mortgage delinquency, despite other machine learning models saying otherwise. However, a borrower's race type is not statistically significant, therefore, cannot conclude that racial category of the borrower has a meaningful impact on the likelihood of mortgage delinquency.



```
                     Generalized Linear Model Regression Results
  Dep. Variable:    HighRiskofMortgageDeliquency  No. Observations:   16904
      Model:        GLM                            Df Residuals:      16890
  Model Family:     Binomial                        Df Model:         13
  Link Function:    Logit                             Scale:          1.0000
     Method:        IRLS                           Log-Likelihood:    -4526.2
      Date:         Mon, 02 Dec 2024                 Deviance:        9052.4
      Time:         15:21:38                       Pearson chi2:      4.62e+04
 No. Iterations:    8                          Pseudo R-squ. (CS): 0.1312
Covariance Type: nonrobust
                                   coef       std err   z       P>|z| [0.025    0.975]
              const                -26.4877   2.285     -11.591 0.000 -30.967   -22.009
     TotalMonthlyIncomeAmount      0.7671     0.154     4.991   0.000 0.466     1.068
            NoteAmount             -1.061e-06 6.84e-07  -1.551  0.121 -2.4e-06  2.8e-07
   HousingExpenseRatioPercent      4.0847     0.174     23.474  0.000 3.744     4.426
  LoanAcquisitionActualUPBAmt      -0.1569    0.190     -0.825  0.410 -0.530    0.216
          LTVRatioPercent          1.8346     0.344     5.329   0.000 1.160     2.509
   Borrower1CreditScoreValue       -0.3763    0.029     -12.760 0.000 -0.434    -0.319
       PMICoveragePercent          -0.0134    0.003     -4.558  0.000 -0.019    -0.008
Borrower1AgeAtApplicationYears     0.0120     0.002     5.270   0.000 0.008     0.016
        Borrower1Race1Type         0.0251     0.030     0.848   0.396 -0.033    0.083
      Borrower1EthnicityType       -0.2883    0.068     -4.244  0.000 -0.421    -0.155
       Borrower1GenderType         0.1731     0.051     3.412   0.001 0.074     0.273
           PropertyType            0.0487     0.014     3.393   0.001 0.021     0.077
           MortgageType            0.0047     0.046     0.104   0.917 -0.085    0.094
```

Figure 11: Logistic Regression results

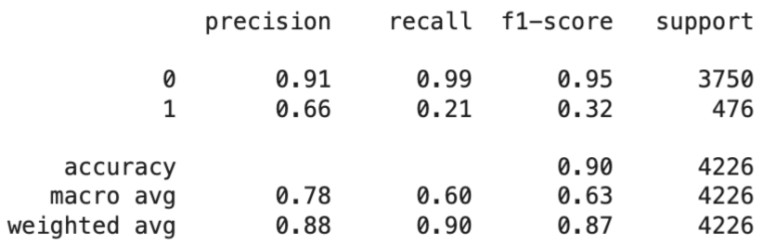
VI. LightGBM

Using the principal components, the low-risk mortgage delinquency class has high precision (91%) and recall (99%), indicating the model is effective in identifying low-risk borrowers with minimal false positive. Furthermore, it has an excellent F1-Score of 0.95 reflecting the overall performance of the class. For the class of high-risk mortgage delinquency, it has low to moderate precision (66%) and recall (21%), which means that high-risk borrowers are misclassified as low risk. It has a poor F1-score of 32% which means there is a low overall performance for this class.

Without the principal components, the low-risk mortgage delinquency class has high precision (91%) and recall (99%), also indicating the model performs well for the majority class. Whereas the high-risk class have slightly higher precision (72%) and recall (26%) compared to the principal components case. The F1-score improve marginally to 39% compared to the principal components case, indicating a slightly better detection of high-risk cases.

The accuracy without the PCA is slightly higher at 91% compared to without PCA, which is still dominated by the low-risk class. Furthermore, the AUC score of 83% without using the principal components dominates the AUC score of 81% using the principal components.

Without the PCA, the '*HousingExpenseRatioPercent*' leads as the most important feature to predict the risk of mortgage delinquency, followed by '*TotalMonthlyIncomeAmount*' and '*NoteAmount*'. While demographic features, '*Borrower1Ethnicitytype*' and '*Borrower1Race1Type*' have the least impact on mortgage delinquency risk.

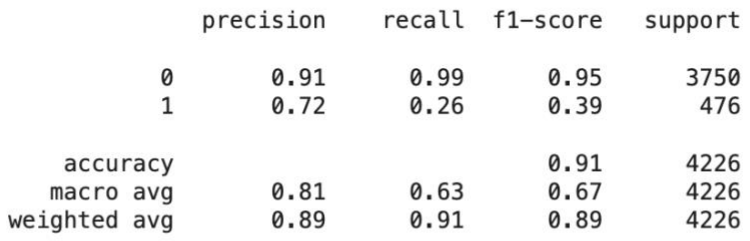


```
              precision    recall  f1-score   support

           0       0.91      0.99      0.95      3750
           1       0.66      0.21      0.32       476

    accuracy                           0.90      4226
   macro avg       0.78      0.60      0.63      4226
weighted avg       0.88      0.90      0.87      4226
```

Figure 12: LightGBM metrics with PCA



```
              precision    recall  f1-score   support

           0       0.91      0.99      0.95      3750
           1       0.72      0.26      0.39       476

    accuracy                           0.91      4226
   macro avg       0.81      0.63      0.67      4226
weighted avg       0.89      0.91      0.89      4226
```

Figure 13: LightGBM without PCA

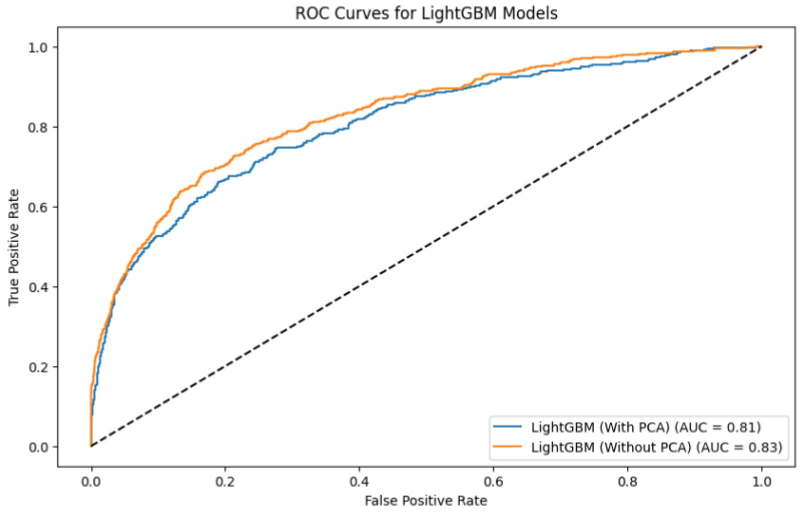


Figure 14: ROC Curve with AUC score using the training set for LightGBM model (With and Without PCA)

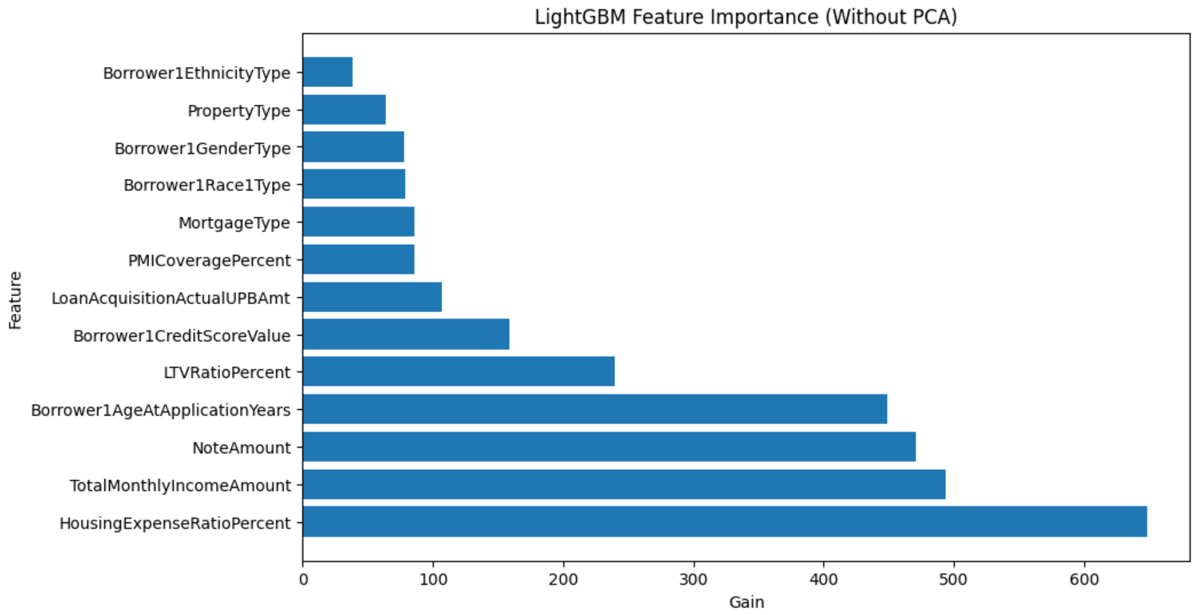


Figure 15: LightGBM Feature Importance (Without PCA)

Conclusion

In conclusion, the LightGBM, logistic regression, and random forests models are the most effective machine learning model in predicting the risk of mortgage delinquency. By observing my hypothesis, we can accept the null hypothesis as there is no significant relationship between a borrower's racial or ethnic background and the risk of mortgage delinquency when controlling for financial characteristics, including housing expenses ratio, income, credit score, loan-to-value ratio.

The findings emphasize the crucial role of the financial predictors, such as housing expense ratio, credit score, and loan-to-value ratio in assessing delinquency risk. However, the demographic disparities revealed through machine learning models suggest that these financial metrics along cannot totally explain the inequities faced by minority borrowers.

While the study provides valuable insights, it is limited by the scope of available data. Therefore, future research should incorporate quantitative analysis to better understand borrowers' experiences and explore how local housing policies contribute to systemic disparities.

Bibliography (APA format)

American Enterprise Institute. (2023). Racial bias in mortgage biz? New data says no, researchers find. https://www.aei.org/research-products/one-pager/racial-bias-in-mortgage-biz-new-data-says-no-researchers-find/

Au, Timothy C. (2018). Random Forests, Decision Trees, and Categorical Predictors: The "Absent Levels" Problem. Journal of Machine Learning Research 19 (2018) 1-30. https://jmlr.csail.mit.edu/papers/volume19/16-474/16-474.pdf

Ben Horowitz, K.-E. K. (2024). *Lender-reported reasons for mortgage denials don't explain racial disparities*. Retrieved from: https://www.minneapolisfed.org/article/2024/lender-reported-reasons-for-mortgage-denials-dont-explain-racial-disparities

Because the CARES Act's foreclosure moratorium does not cover borrowers with non-agency mortgages, unprotected homeowners with non-agency mortgages face a higher likelihood of losing their home if they are not in a privately agreed-upon forbearance plan. Delinquency without the protection of forbearance, for either agency or non-agency mortgage borrowers, will also lower that borrower's credit score. These unprotected borrowers have credit scores that, at the median, are almost 200 points lower than protected and current borrowers. Continuing to miss mortgage payments without the protection of forbearance will further weigh on their scores. The combination of low credit scores and tight lending standards makes it impossible for these delinquent borrowers to refinance to lower their payment or extract home equity and makes it more difficult to get a personal loan at a reasonable rate to weather this crisis. Historically, households of color, and Black households in particular, had lower credit scores than white households, partly because of structural barriers in employment, income, and access to credit. Consequently, the greater proportion of unprotected mortgaged homeowners in neighborhoods of color could further exacerbate pre-pandemic disparities in credit scoring and access to wealth-building tools, while slowing the recovery of homeowners of color and leaving them further behind.

GeeksforGeeks. (2024). What are the Advantages and Disadvantages of Random Forest? GeeksforGeeks. https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/

JPMorganChase. (2024). *Hidden costs of homeownership: Race, income, and lender differences in loan closing costs*. Retrieved from JPMorganChase: https://www.jpmorganchase.com/institute/all-topics/financial-health-wealth-creation/hidden-costs-of-homeownership-race-income-and-lender-differences-in-loan-closing-costs

The analysis highlights significant racial disparities in closing costs for home loans, particularly among nonbank lenders and brokers. Black and Hispanic borrowers face higher costs compared to White borrowers, even with banks. Specifically, Black borrowers pay an average of $256 more, and Hispanic borrowers $270 more. Nonbanks and brokers exacerbate these disparities, with Black borrowers incurring an average of $812 in additional costs through brokers. The findings suggest that socioeconomic factors intersect with racial identity, warranting further investigation into how these dynamics impact homeownership costs.

Michael Neal, C. Y. (2020). *Delinquent Homeowners in Neighborhoods of Color Are Less Likely to Be Protected by Forbearance*. Retrieved from Urban Institute: https://www.urban.org/urban-wire/delinquent-homeowners-neighborhoods-color-are-less-likely-be-protected-forbearance

The CARES Act provided forbearance to homeowners with federally backed mortgages, protecting them from default and credit score damage if they missed payments due to the pandemic. However, many homeowners are still delinquent without forbearance, particularly those with non-agency mortgages or those who have not taken advantage of forbearance. This leaves a group of "unprotected homeowners," disproportionately found in predominantly Black and Hispanic neighborhoods. These homeowners are more likely to face foreclosure and experience lower credit scores, exacerbating pre-existing racial disparities in credit access and wealth-building opportunities. Public policy efforts are necessary to target these communities with outreach and assistance to mitigate long-term economic harm and ensure a more equitable recovery.

Michelle Singletary (2021). Covid forced more borrowers to be late on their mortgages than at any time since the Great Recession, CFPB reports. Retrieved from Washington Post: https://www.washingtonpost.com/business/2021/05/07/cfpb-reports-mortgage-borrowers-covid-challenges/

The Consumer Financial Protection Bureau (CFPB) has reported that more borrowers are behind on their mortgages now than at any time since the Great Recession, with Black and Hispanic homeowners disproportionately affected. These borrowers are twice as likely as White homeowners to be delinquent or in forbearance, despite making up a smaller share of total mortgage holders. Federal forbearance programs introduced during the pandemic helped many avoid foreclosure, but as these programs wind down, Black and Hispanic homeowners remain at higher risk of losing their homes, potentially widening the racial wealth gap. The CFPB urges both borrowers and lenders to explore available options to prevent foreclosure.

Park, J., & Park, D. (2023). Racial disparities in unemployment benefits among U.S. mortgage borrowers during COVID-19. Journal of housing and the built environment : HBE, 1–31. Advance online publication. https://doi.org/10.1007/s10901-022-10006-w

The article highlights that during the COVID-19 pandemic, Black unemployment insurance (UI) recipients were more likely to delay mortgage payments and express concerns about future payments compared to White recipients. Despite receiving UI benefits, Black borrowers faced significant mortgage payment challenges, compounded by income loss, food insecurity, and mental health issues. The findings align with previous research indicating that while UI benefits can assist borrowers, they disproportionately benefit White individuals over people of color. The study suggests that a more race-conscious approach to UI management could address the unique challenges faced by minority borrowers, who experienced higher rates of mortgage difficulties even after receiving benefits. Issues like application surges and delays in benefit distribution during the UI program's expansion further affected racial and ethnic groups differently. Instead of considering mortgage deferrals for specific groups—which could lead to perceptions of unfairness—governments should enhance education and financial counselling programs for minority borrowers. Additionally, while national forbearance and foreclosure moratoriums, along with expanded UI programs, alleviated some financial distress for many borrowers, minority groups continued to face greater difficulties in managing payments and securing refinancing options. The pandemic exacerbated existing inequalities, as racial minorities frequently experienced job losses and health challenges. As the pandemic's effects linger, housing policymakers are urged to prioritize support for minority mortgage borrowers, recognizing that their housing challenges are intertwined with broader socioeconomic hardships exacerbated by the pandemic.

Research, O. o. (2021). ***Consumer Financial Protection Bureau*****. Retrieved from Characteristics of Mortgage Borrowers During the COVID-19 Pandemic. Retrieved from Consumer Financial Protection Bureau: https://www.consumerfinance.gov/data-research/research-reports/characteristics-mortgage-borrowers-during-covid-19-pandemic/**

By early 2021, a year into the pandemic, a significant share of mortgage borrowers remained in forbearance programs or had delinquent loans. Many of these borrowers likely entered into forbearance at the start of the pandemic, between March and May of 2020, when the largest wave of forbearances took place. Among delinquent mortgage loans, a significant share was delinquent prior to the pandemic with the delinquency continuing throughout the economic crisis. As programs designed to aid mortgage borrowers during the pandemic wind down, borrowers in forbearance programs or who are delinquent may be disproportionately at risk of losing their home. The Bureau’s analysis of mortgage data shows that a significant share of these borrowers were minorities, lived in majority-minority tracts, and lived in relatively lower-income areas. Many of these borrowers also may be single-income households, making it more difficult for them to recover from income shocks. A significant share of borrowers also showed distress in terms of non-mortgage products. In addition, the fact that many delinquent borrowers were delinquent prior to the pandemic may mean it is even more difficult for those borrowers to recover after-the-fact. Given the borrower characteristics associated with forbearance and delinquency, these borrowers also may have difficulty recovering from the extended period of forbearance or delinquency.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044