AI-Powered Lifecycle Financial Planning – Predictive Modeling in Liabilities

August 2023

Abstract

1 Introduction

1.1 Background

Our objective is to develop a comprehensive guide for managing people's liabilities throughout their lifecycle. Obtaining an accurate analysis of an individual's debts requires consideration of numerous factors, including job changes and economic policies. Accounting for all these possibilities can result in an overly complex and challenging prediction process. To address this challenge, we employ machine learning techniques to predict the typical liabilities that an average American may incur during various stages of their life.

The household debt report from the Federal Reserve Bank of New York[1] highlighted that housing debt constitutes a significant portion of the overall debt. Figure 1 shows that over the past decade, mortgages consistently represented nearly 70% of the total debt balance each quarter. Diverging into age demographics, Figure 2 reveals that housing debt remains a predominant component of the total debt balance across various age groups. Specifically, mortgages constitute approximately 50% of the total debt for individuals aged 18-29, 70% for those between 30-39 years, and a stable 75% for individuals aged 40 and above. Given the prominent role housing mortgages play throughout an individual's lifecycle, this paper will delve into predicting housing debt in subsequent sections.

[1] Federal Reserve Bank of New York. (2023). Quarterly Report on Household Debt and Credit: 2023 Q2. Retrieved from `https://www.newyorkfed.org/medialibrary/interactives/householdcredit/data/pdf/HHDC_2023Q2`. Released August 2023.

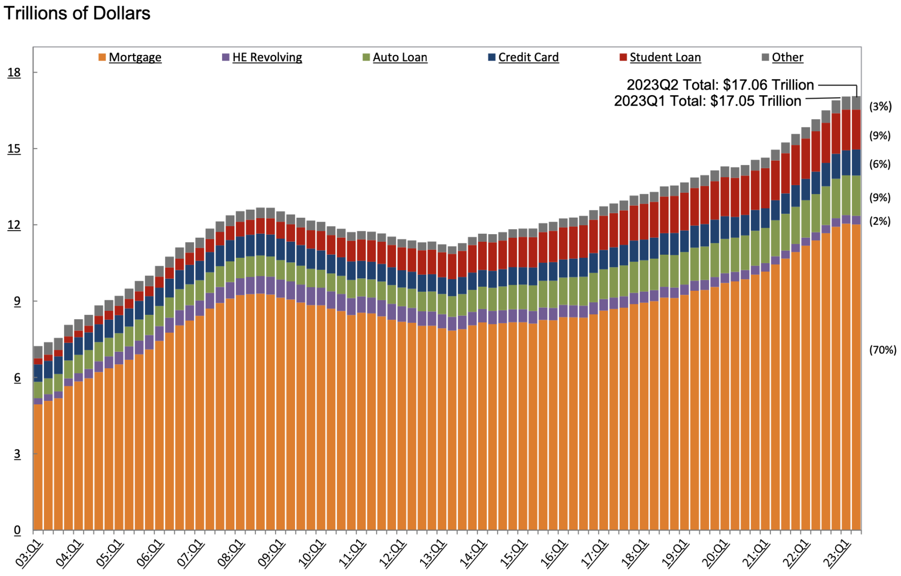


Figure 1: Total Debt Balance and its Composition 2003 Q1-2023 Q2)

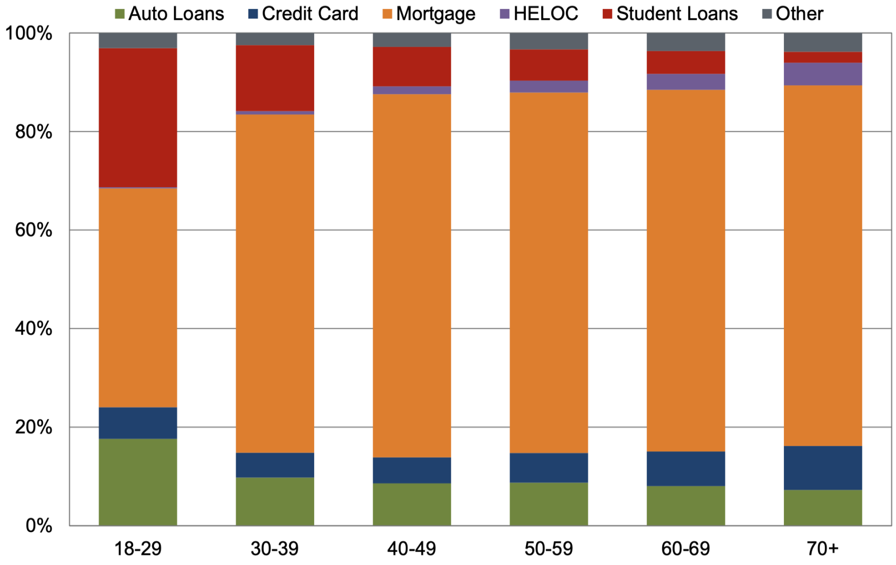


Figure 2: Debt Share by Product Type and Age (2023 Q2

1.2 Dataset Description

This is a public-use database that is about the federal home loan bank system. HERA Section 1212 requires the Director to make available to the public, in a form that is useful to the public (including forms accessible electronically), and to the extent practicable, census tract level data relating to mortgages purchased by each Federal Home Loan Bank. Besides, the contents of these files are unaudited. Specifically, this is for the 2021 Data Release. We use the updated data in 2021.

We found this dataset from Federal Housing 4 Finance Agency's website (At the bottom).[2] At the bottom of this link, we use the Dataset (FHLBank Public Use Database): 2021 CSV / 2021 Excel / 2021 Definitions, which has a more detailed description.

1.3 Data Preprocessing

We use the data "2021_PUDB_EXPORT_123121.xlsx", which contains 63,890 observations and 56 variables. The data was filtered to only consider rows where there is a single borrower. The ratio of single borrowers is approximately 40.7 percent of the total data. The shape of the filtered data is (25,986, 56). The percentage of missing values for specific columns (Borrower1Race1Type, Borrower1GenderType, Borrower1AgeAtApplicationYears, Borrower1CreditScoreValue, Borrower1EthnicityType) was calculated. All rows with these missing values were then removed, resulting in a data shape of (23,356, 56).

Table 1: The missing values of the dataset

BoRace	BoGender	BoAge	BoCreditScor	BoEth
0.0709613	0.0347110	0.0001539	0.0017702	0.0831602

2 EDA

2.1 Target Variable

For liability Data, our target value is **NoteAmount** (Mortgage balance at origination.)

Here is a distribution of the target variable.

Upon looking at the plot, we notice that the distribution is right-skewed. The tail on the right side of the distribution is longer and heavier, indicating positive kurtosis. This suggests that there is a higher concentration of data in the center of the distribution with a few extreme values on the right. This

[2] https://www.fhfa.gov/DataTools/Downloads/Pages/Public-Use-Databases.aspx

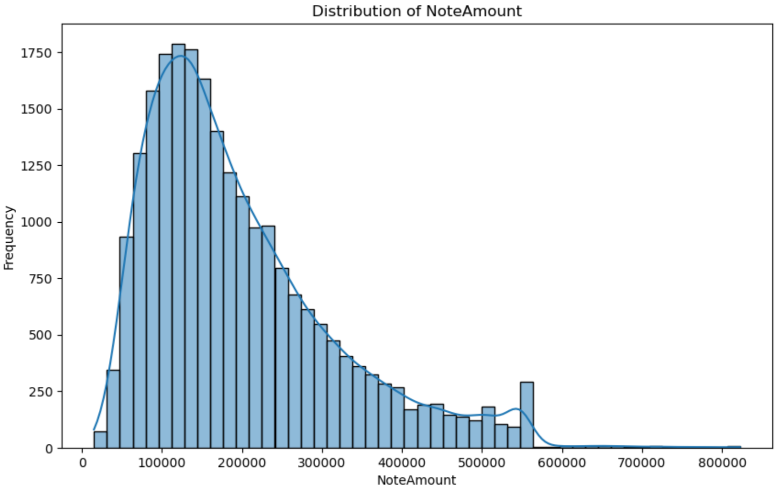


Figure 3: Distribution of NoteAmount

means that there are relatively few people with very high mortgage balance, while the majority of individuals have lower to moderate mortgage balance.

The spread of mortgage balance levels is relatively large, indicating substantial mortgage balance variability within the population. The mean mortgage balance is significantly higher than the median mortgage balance. This is a typical characteristic of right-skewed distributions, as the few high-expenditure individuals pull the mean to the right. The median mortgage balance is a more appropriate measure of central tendency in this case, as it is less affected by extreme values. It represents the mortgage balance level at which half of the population mortgages less, and half mortgages more.

One interesting thing is that there is a strange increase in distribution around 550000. Which might be an outlier.

2.2 Feature Variable

The clean data we used for the model has **19** features out of the target variable.

In order to facilitate the analysis, we list these variables by their meanings.

Variable Name	Description
TotalMonthlyIncomeAmount	The total monthly qualifying income used for underwriting in whole dollars for all borrowers on theloan.
LoanPurposeType	Purpose of Loan: 1 = Purchase, 2 = No-Cash Out Refinancing, 3 = Second Mortgage, 4 = New Construction, 5 = Rehabilitation or Home Improvement, 6 = Cash-out Refinancing, 7 = Other
MortgageType	Type of Mortgage and whether the mortgage is guaranteed: 0=Conventional, 1=FHA, 2=VA, 3=USDA Rural Housing-FSA Guaranteed, 4=HECMs, 5=Title1- FHA
LoanAmortizationMaxTermMonths	For Amortizing Mortgages, term of amortization in months; 998 if non-amortizing loan
MortgageLoanSellerInstType	Type of Institution from which the FHLBank acquired the mortgage. 01=Insured depository institution, 02=Housing Associate, 03=Insurance Company, 04=Non-Federally Insured CU, 05=Non-Depository CDFI, 06=Other FHLBank, 09=Other
BorrowerFirstTimeHomebuyer	Numeric code indicating whether borrower is a first time homebuyer. 0 = No, 1 = Yes
Borrower1Race1Type	Numeric code indicating the race of the Borrower. 1=American Indian or Alaska Native, 2=Asian, 3=Black or African American, 4=Native Hawaiian or other Pacific Islander, 5=White, 6=Information not provided by Borrower, 7=Not Applicable (First or primary borrower is an institution, corporation or partnership)
Borrower1GenderType	Numeric code indicating the sex of the first or primary borrower. 1=Male, 2=Female, 3=Information not provided by borrower, 4=Not Applicable (First or primary borrower is an institution, corporation or partnership), 6=Borrower selected both male and female
Borrower1AgeAtApplicationYears	Age in years of the borrower at time application submitted; 999=Age not provided, 998=Not Applicable (Borrower might be a legal entity like an LLC)
PropertyUsageType	Numeric code indicating whether property is owner occupied, second home or a rental investment property. 1=Principal Residence, 2=Second Home, 3=Investment Property
PropertyUnitCount	Total number of units in the property
NoteRatePercent	Interest rate on the mortgage at acquisition
NoteAmount	Mortgage balance at origination
Borrower1CreditScoreValue	Credit Scores are separated into a range: 1=¡620, 2=620 ¡ 660, 3=660 ¡ 700, 4=700 ¡ 760, 5=760 or greater, 9 = Missing or Not Applicable
PMICoveragePercent	Percent of mortgage balance at origination covered by loan level PMI
EmploymentBorrowerSelfEmployed	Numeric code indicating whether the borrower is selfemployed. 0=No, 1=Yes
PropertyType	PT01=Single family detached; PT02=Deminimus PUD; PT03=Single family attached; PT04=Two family; PT05=Townhouse; PT06=Low-rise condo; PT07=PUD; PT08=Duplex; PT09=Three family; PT10=Four family; PT11=Hi-res condo; PT12=Manufactured home not chattel; PT13=Manufactured home chattel; PT14=Five plus multifamily
MarginRatePercent	Margin added to the index used for the calculation of the interest on an ARM. 9999=Not Applicable
Borrower1EthnicityType	1=Hispanic or Latino; 2=Not Hispanic or Latino; 3=Information not provided; 4=Not applicable (First or primary borrower is an institution, corporation or partnership)
HOEPALoanStatusType	1=HOEPA: High-Cost Mortgage; 2=HOEPA: Not a HighCost Mortgage; 3=Not subject to HOEPA

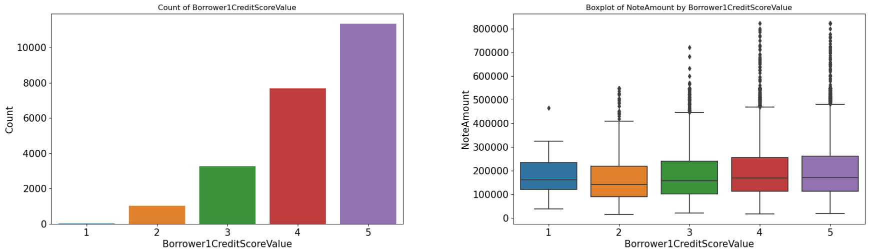
In order to have a clear look about the relationship between target variable and feature variables, we pick several examples from each group and split them between categorical and numerical variables.

2.2.1 Categorical Variable

Borrower1CreditScoreValue

We get the distribution of “Borrower1CreditScoreValue” variable and the boxplot of “NoteAmount” by “Borrower1CreditScoreValue”. In this data, we have most people in range of 3 to 5.

The median of each group is approximately closer, implies that, on average, they have a similar mortgage balance. The height of box for group 1 is narrower than the box for others, showing that the mortgage balances in group 1 are more tightly clustered around the median. The whiskers for group 1 are shorter than those for others, it indicates that group 1 has fewer extreme mortgage balances. Group 4 and 5 have more outliers than male, it suggests that they may have more people with exceptionally high mortgage balances.



(a) Count of Borrower1CreditScoreValue

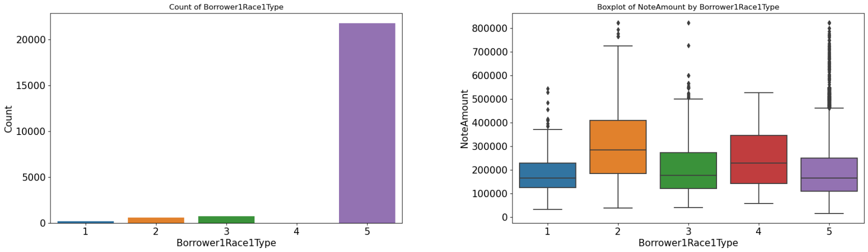
(b) Boxplot of NoteAmount by Borrower1CreditScoreValue

Figure 4: Borrower1CreditScoreValue plots

Borrower1Race1Type

We get the distribution of “Borrower1Race1Type” variable and the boxplot of “NoteAmount” by “Borrower1Race1Type”. In this data, most people’s race type is in group 5.

Group 2 has a higher median, implies that, on average, it have a higher mortgage balance. The height of box for group 1 is narrower than boxes for other groups, showing that the mortgage balances in group 1 are more tightly clustered around the median. The whiskers for group 1 are shorter than those for other groups, it indicates that group 1 has fewer extreme mortgage balance. Group 5 has more outliers than other groups, it suggests that group 5 may have more people with exceptionally high mortgage balances.



(a) Count of Borrower1Race1Type

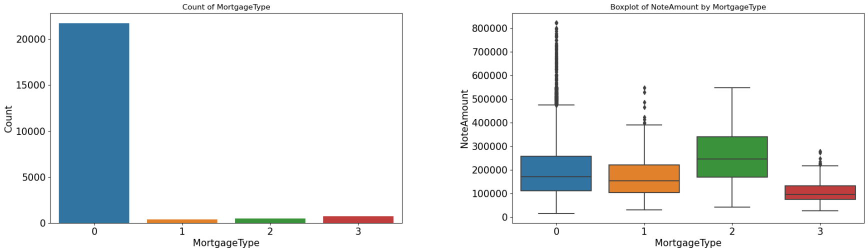
(b) Boxplot of NoteAmount by Borrower1Race1Type

Figure 5: Borrower1Race1Type plots

MortgageType

We get the distribution of "MortgageType" variable and the boxplot of "NoteAmount" by "MortgageType". In this data, most people's mortgage type is in the group 0.

Group 2 has a higher median, implies that, on average, people in this group have a higher mortgage balances. The height of box for group 3 is narrower than boxes for other groups, showing that the mortgage balances in group 3 are more tightly clustered around the median. The whiskers for group 3 is shorter than those for other groups, it indicates that people in this group have fewer extreme mortgage balances. Group 0 has more outliers than other groups, it suggests that group 0 may have more people with exceptionally high mortgage balances.



(a) Count of MortgageType

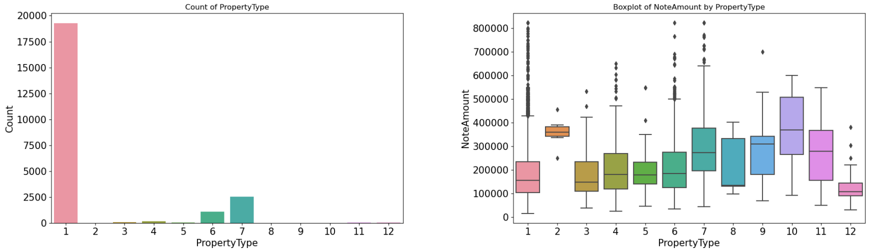
(b) Boxplot of NoteAmount by MortgageType

Figure 6: MortgageType plots

PropertyType

We get the distribution of “PropertyType” variable and the boxplot of “NoteAmount” by “PropertyType”. In this data, most people’s propertyType is in group 1.

Group 2 has a higher median, implies that, on average, people in group 2 have a higher mortgage balances. The height of box for group 2 and 12 are narrower than boxes for other groups, showing that the mortgage balances in group 21 and 12 are more tightly clustered around the median. The whiskers for group 2 is shorter than those for other groups, it indicates that people in this groups have fewer extreme mortgage balances. Group 1 has more outliers than other groups, it suggests that group 1 may have more people with exceptionally high mortgage balances.



(a) Count of PropertyType

(b) Boxplot of NoteAmount by PropertyType

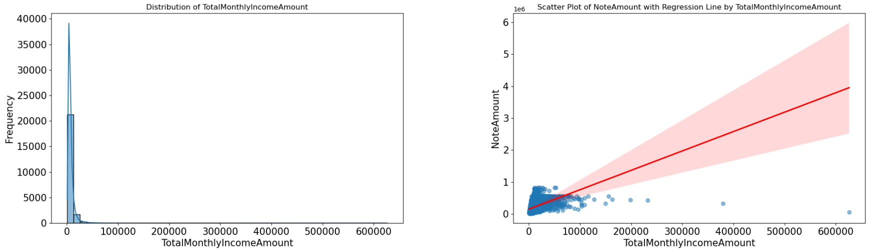
Figure 7: PropertyType plots

2.2.2 Numerical Variable

TotalMonthlyIncomeAmount

We get the distribution of “TotalMonthlyIncomeAmount” variable and the scatter Plot of “NoteAmount” with Regression Line by “TotalMonthlyIncomeAmount”. From the plot we find that there are most people have low TotalMonthlyIncomeAmount in this data.

From the scatter plot we could see a positive slope trend indicating a positive correlation between “NoteAmount” and “TotalMonthlyIncomeAmount”. Also the points are cluster together to the regression line, showing that this relationship is very strong. However, we find some points spread extremely far away from other points, indicates that they are outliers and explains why we see very strange increase in 550000 in the distribution of “NoteAmount”.



(a) Distribution of TotalMonthlyIncomeAmount

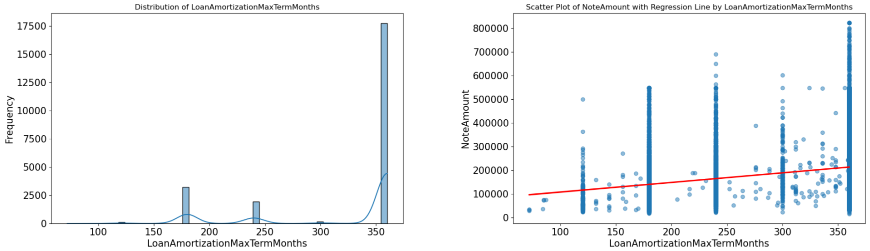
(b) Scatter Plot of NoteAmount with Regression Line by TotalMonthlyIncomeAmount

Figure 8: TotalMonthlyIncomeAmount plots

LoanAmortizationMaxTermMonths

We get the distribution of “LoanAmortizationMaxTermMonths” variable and the scatter Plot of “NoteAmount” with Regression Line by “LoanAmortizationMaxTermMonths”. From the plot we find that most people have there loan term months as 350 months.

From the scatter plot we could see a positive slope trend indicating a positive correlation between “NoteAmount” and “LoanAmortizationMaxTermMonths”. Also the points are cluster together to the regression line, showing that this relationship is quite strong. As there are more data in 350 than other ranges, there might be outliers that have influence to this correlation.



(a) Distribution of LoanAmortizationMaxTermMonths

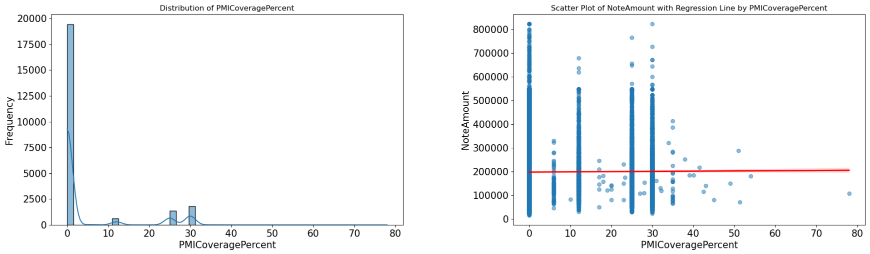
(b) Scatter Plot of NoteAmount with Regression Line by LoanAmortizationMaxTermMonths

Figure 9: LoanAmortizationMaxTermMonths plots

PMICoveragePercent

We get the distribution of “PMICoveragePercent” variable and the scatter Plot of “NoteAmount” with Regression Line by “PMICoveragePercent”. From the plot we find that most people 's PMI Coverage percent is 0.

From the scatter plot we could see a very flat slope trend indicating a small correlation between “NoteAmount” and “PMICoveragePercent”. Also the points are spead on two sides of line relationship is not strong enough. As there are more data in range 0, there might be outliers that have influence to this correlation.



(a) Distribution of PMICoveragePercent

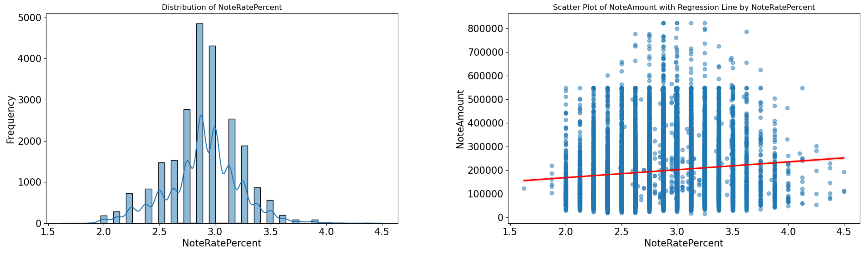
(b) Scatter Plot of NoteAmount with Regression Line by PMICoveragePercent

Figure 10: PMICoveragePercent plots

NoteRatePercent

We get the distribution of “NoteRatePercent” variable and the scatter Plot of “NoteAmount” with Regression Line by “NoteRatePercent”. From the plot we find that this distribution is quite normal, and most people have there note rate percent in group 3.0.

From the scatter plot we could see a positive slope trend indicating a postive correlation between “NoteAmount” and “NoteRatePercent”. Also the points are spread evenly on two sides of the regression line, showing that this relationship is quite strong.



(a) Distribution of NoteRatePercent

(b) Scatter Plot of NoteAmount with Regression Line by NoteRatePercent

Figure 11: NoteRatePercent plots

2.3 Heat Map

We also create a heat map to check the correlation between each numerical variables.

Note that most correlations between numerical variables are as small as 0.18, -0.02, indicating that they have small cor relationship.

We do find the highest correlation between “TotalMonthlyIncomeAmount” and “NoteRatePercent” as 0.58. This can be explained as people with more income have more ability to apply loan with higher interest rate.

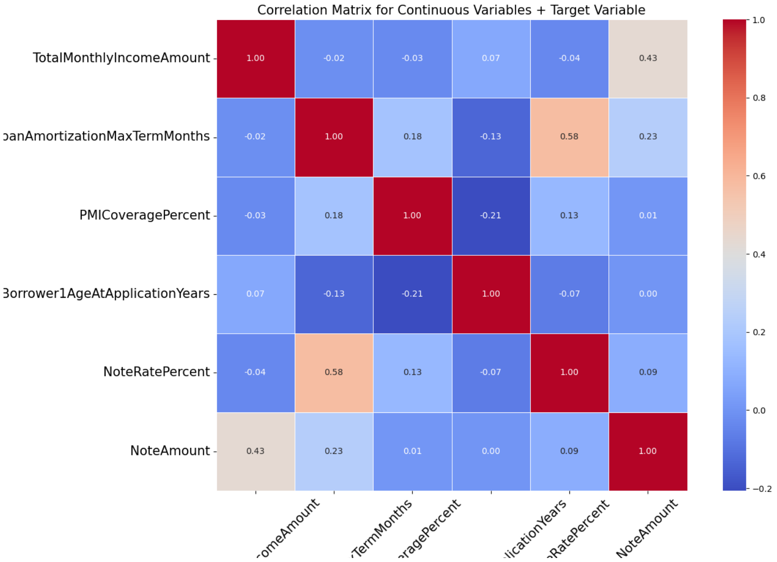


Figure 12: Heat map for Numerical Variables

3 Modeling

3.1 Models

3.1.1 Linear Regression

The first model that we built is the linear regression model. To process the data for linear regression, we used one-hot encoding to turn all the categorical variables into dummy variables. Consequently, we have 5 numeric features and 49 dummy features. We built a function called standardize in the code file to standardize the predictors which are calculated as:

$$\text{standardized\_value} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

3.1.2 Lasso regression

We used *LassoCV* from *sklearn.linear_model* package to do the lasso model. Lasso regression is a technique to reduce model complexity and prevent overfitting, which may result from simple linear regression. By adding penalty to the weights of the variables, it provides greater prediction accuracy as compared to simple linear regression models.

3.1.3 Recursive Feature Elimination

We used $RFECV$ and RFE from $sklearn.feature\_selection$ package, and *LinearRegression* to do recursive feature elimination fit. Recursive feature elimination is a feature selection method that fits a model and removes the weakest features until the specified number of features is reached. To find the optimal number of features cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features. $RFECV$ implements recursive feature elimination with built in cross-validated selection of the best number of features. In our model, We set the minimum features to select to be 20, and cross-validation to be 3, and step to be 1 (default).

3.1.4 Random Forest

We used $RandomForestRegressor$ from the *sklearn.ensemble* package to fit the random forest regression model. We use 100 trees in the forest with a default value of none max depth, 2 splits, "auto" max features. This method significantly reduce the error compared to the linear regression model.

3.1.5 LightGBM

We used $LGBMRegressor$ from the *lightgbm* package to fit the lightGBM model. LightGBM is a gradient boosting framework that employs tree-based learning algorithms. The performance for this model is even better than random forest.

3.2 Evaluation Metrics

For evaluating the performance of our regression models, since the outcome is continuous, we choose these four numeric metrics to calculate the error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$PE = \frac{|\hat{y}_i - y_i|}{y_i} * 100$$

3.3 Result

For comparison we have the results for 5 models:

Validation Error of Different Models

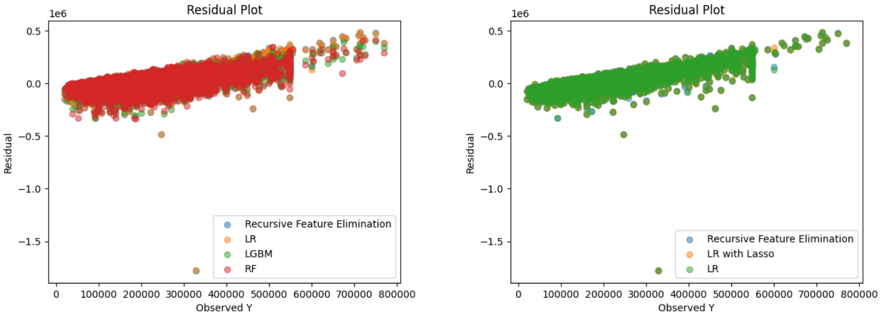
Model	RMSE	MAE	ME	PE
Standard Linear Regression	96657.1255	70405.0837	-1123.4371	25.6200
Lasso Regression	96605.7569	70419.3795	-1224.6346	25.6766
Recursive Feature Elimination	97005.3396	70682.7340	-1099.4237	25.8815
Random Forest	79887.9503	58525.9467	257.2903	16.1573
LightGBM	76513.8960	55887.7718	-1731.6140	14.7951

Note that the LightGBM model has the overall best performance with the lowest RMSE 76513.8960 and the lowest PE 14.7951.

3.4 Residual Plots

The residual is defined as the difference between the observed value and the predicted value. We plot the residual of each observation in the test data, and the smaller residual indicates that the predictions are closer to the ground truth and the model less prediction error.

Here We create two separate residual plots for more direct comparison. One is the comparison between linear models and tree-based models, and the other one is the comparison of different kinds of linear models.



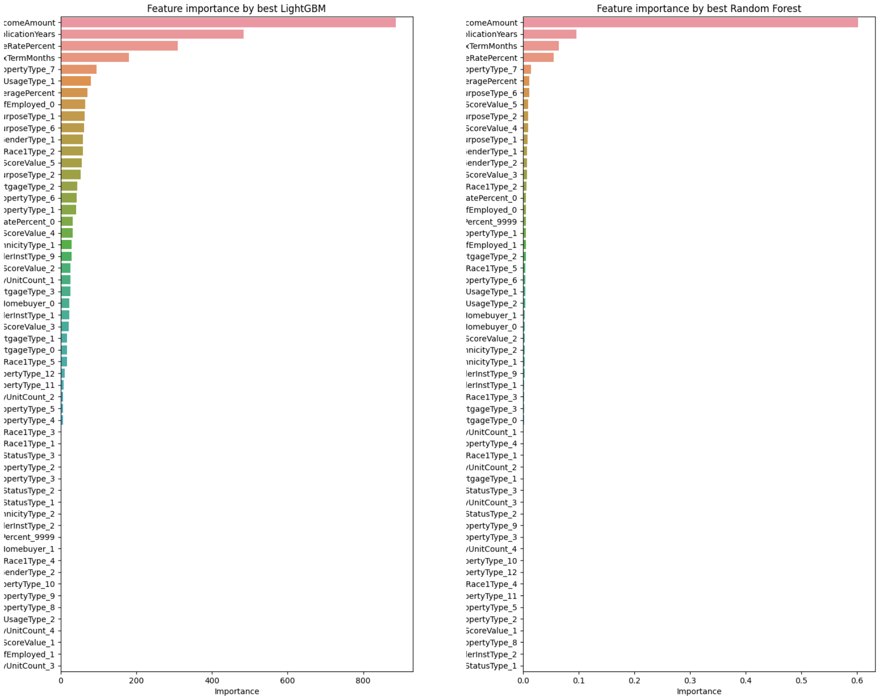
(a) Residual Plot

(b) Residual Plot LR

Figure 13: Residual plots

3.5 Feature Importance

To get an intuition of which features are most related to the total expenditure, we plotted the feature importance of tree-based models. The feature importance plots generated from our random forest and lightgbm model.



(a) Feature Importance by LightGBM (b) Feature Importance by Random Forest

Figure 14: Feature Importance

Note that "TotalMonthlyIncomeAmount" is the most important feature. This might because people with more income will have more access to mortgage.