Mortage Loan Eligibility

Mark Athas

2022-10-21

Abstract

United States Federal law has defined a Consumer Financial Protection Bureau (CFPB). This Bureau has responsibility to track and report on lender mortgage loan activity. To fulfill that accountability, the CFPB collects loan activity and makes that data available to the public on an annual basis. This report utilizes that data to create a model to predict the granting of mortgage loans for single family, owner occupied mortgage applications.

Introduction

There is significant public interest in the mortgage loan activity of America's banks particularly as related to the sex, race and age of the applicant. The Consumer Financial Protection Bureau provides annual data sets of all mortgage loan activity from all 50 states. The latest year with complete data is 2021 which contains millions of loan applications. To manage the computational requirements of this analysis, it was limited to represent states with lowest, middle and highest median income.

To predict mortgages granted to the general public, applications for commercial and re-purchase loans are removed from the data set. Additionally, multi-unit mortgages were also removed, leaving only single-family loans made to occupying applicants. These mortgages are the focus of this analysis.

This report seeks to identify an optimal model to predict if a loan is granted to the applicant. The methodology to achieve this objective will involve a number of broad activities. First, the current condition of the data will be assessed. This will include describing the input data, evaluating its completeness and creating a tidy data set that can be further analyzed. Next, a number of machine learning models will be trained and their accuracy as determined by confusion matrix will be recorded. The final step will run the best-fit model(s) against a verification data set. The verification set accuracy achieved is 92.01 percent.

Methodology and Analysis

The 2021 data from the CFBP contains over 2.6 million mortgage loans in a more than 10 gigabyte file. (FFEIC 2021) This amount of data is too large to be processed in the available RStudio/R compute environment. To reduce the volume of data, mortgage loan applications were selected based on their representation among low, middle and high median incomes. The Federal Reserve median income by state data (FRED 2021)was used to make the state selection. This reduced the total number of mortgage loans to 528,151 which is manageable on the available compute platform.

Single Family Home Purchases by Individuals

The CFBP loans data contains both personal and commercial loans, new and re-finance loans, normal and reverse-mortgage loans–basically all loans provided by the banks in each state. For this analysis, loans included are only those from applications of single family, built on-site and owner occupied properties. The loans must also be for purchase and not refinance. Also, ony loans with applicant income are included.

Data wrangling was performed to filter the data as described previously, but also includes filling in missing variables with the median data of the corresponding Metropolitan Statistical Area (MSA).

Features and Outcome

The identified features are listed in Table 1 with a brief description. Feature selection was limited to these variables as they represent input from the application and publically available data about the property (e.g., property address, MSA).

Table 1: Available Features

Feature	Description
amount	principle amount of the loan
age	age of the primary applicant
debt_ratio	debt to income of the applicant
income	applicant income
lv_ratio	loan to value ratio
msa_incper	ratio of income of the tract to the contained-in MSA
msa_minc	median income of the MSA
race	1:Indigenous, 2:African American, 3:White, 4:Other
rate	interest rate of the mortgage
sex	1:Male, 2:Female, 3:Other, 4:Not Provided
term	term of the loan in months
value	value of the property mortgaged

The features include nominal discrete, ordinal discrete and continuous variables. Discrete values are re-coded to reduce the number of natural classes in the data as there were a significant number of class values with few observations. These were also converted to numeric values.

A histogram plot of the candidate features is provided in Figure 1. Most factors have a skewness. Additionally the box-plots in Figure 2 show that outliers exist in most features.

For continuous variables, a number of scaling functions were tried including logarithmic and exponentiation to reduce skewness, yet these proved to only have a negative impact on accuracy. Regularization and elimination of outliers also showed no improvement in prediction accuracy. Removal of these features, as a last-resort approach, would be arbitrary as they are more better described as legitimate extreme value rather than outliers due to measurement error.

Figure 1: Applicant Feature Histograms

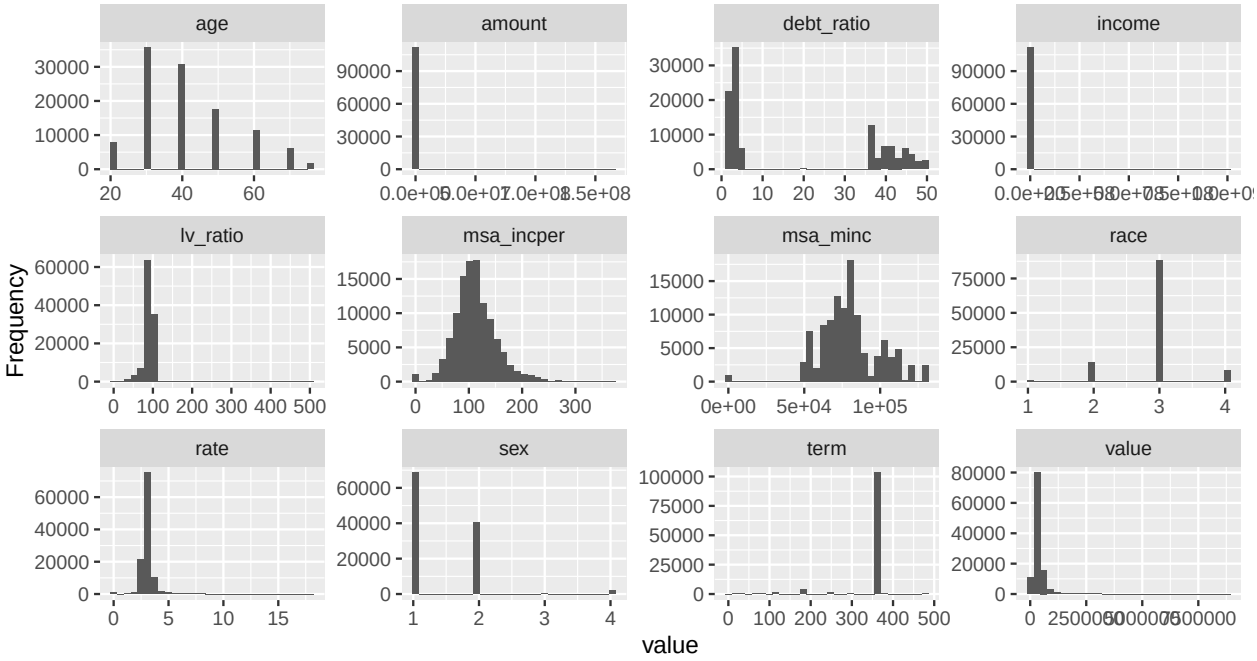
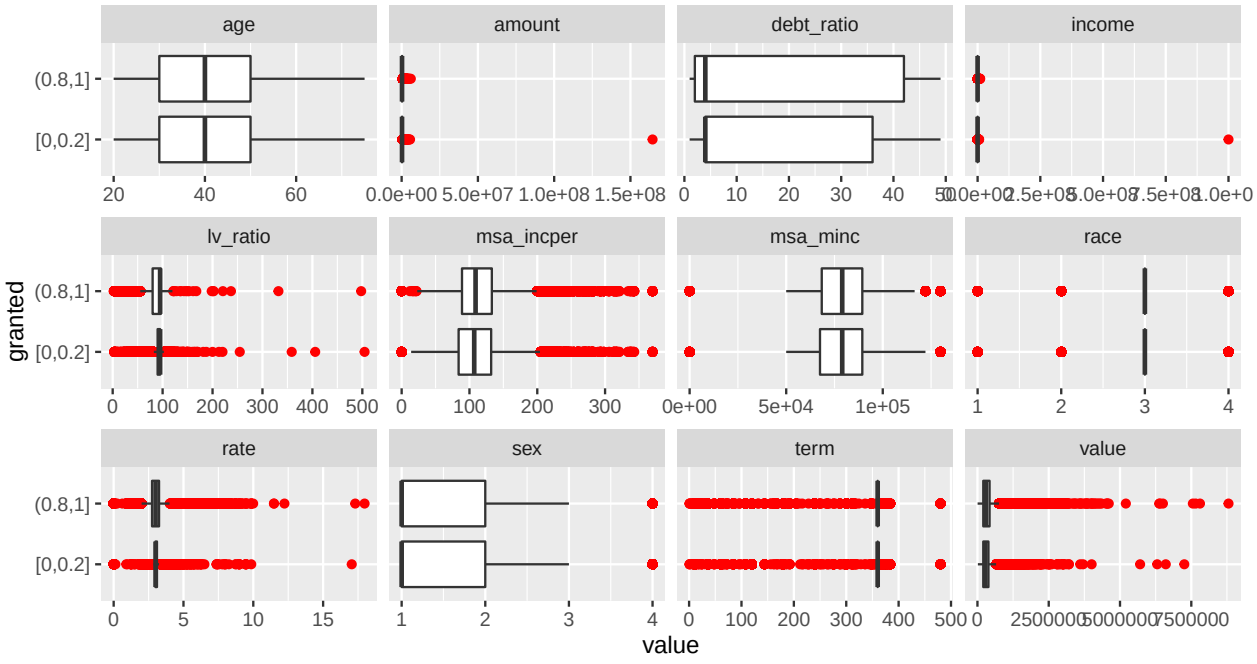


Figure 2: Applicant Feature Boxplots



Outcome Variable

The outcome variable is derived from the action taken by the lender. Such actions could include loan denied, loan not accepted, application rejected or loan originated. These values are consolidated into a loan granted and loan not granted value coded into a variable called *granted*. The granted variable is a binary categorical outcome with value of zero (loan not granted) or one (loan granted).

Creation of Train, Test and Verification Data

The loans from the selected states were split using `createDataPartition()`. A partitioning probability was selected after multiple runs of all models while altering the p value. Model accuracy and F1 scores for the best performing models resulted from a partitioning probability of 0.7. The partition split resulted in a train set of 77,996 loans and an intermediate set. Next, the intermediate set was split 50/50 into test set of 16,713 loans and verification set of 16,713 loans.

Model Analysis

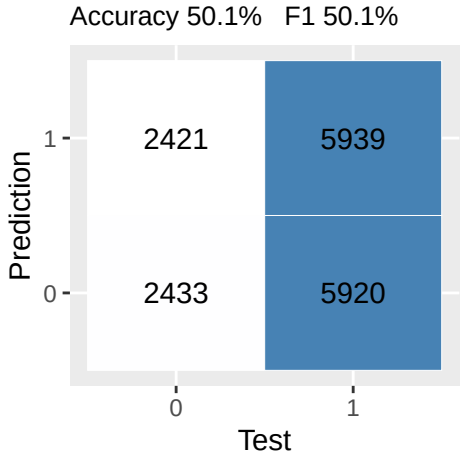
Five machine language models were applied in this analysis including; naive Bayes, knn, decision tree, random forest, and xgBoost. Each of these is compared to random guessing as a baseline.

Analysis of Random Guessing

To simulate a random selection mechanism which will act as a baseline, a binary random sample vector sized to matched the test set with a 50/50 probability is created using `sample()`. The resulting zero/one vector was passed along with the test set to a confusion matrix.

The random sample produced a result of 0.50093. Figure 3 shows the random sample confusion matrix which is approaching the 50% probability of a equalized random binary outcome resulting from 16,713 observations. However, guessing performs worse than the actual mean loans granted one could realize by inspection of the mean *granted* loans 0.70957 in the test data set.

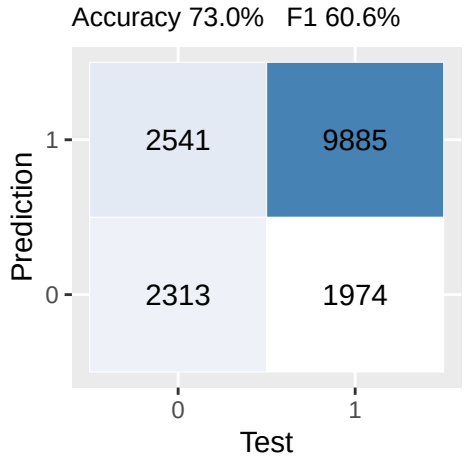
Figure 3: Random Guessing Confusion Matrix



Analysis of Naive Bayes

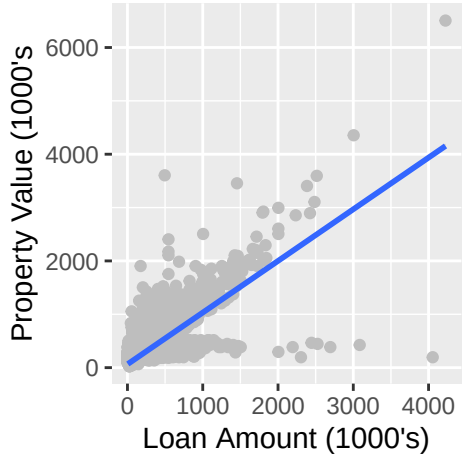
The Naive Bayes model applied to the test data resulted in a model with prediction accuracy of 0.72985. This is better than random and exceeds the means loans granted from the data. The confusion matrix of naive Bayes is provided in Figure 4.

Figure 4: Naive Bayes Confusion Matrix



Challenges with naive Bayes appear due to the assumptions of the model. First, it assumes the features act on the outcome with little or dependency on each other. However, with this data that assumption does not hold. Some features have dependency. For example, loan amount and property value correlate as seen in Figure 5. Other features also defy naive Bayes assumed independence.

Figure 5: Loan Amount to Property Value



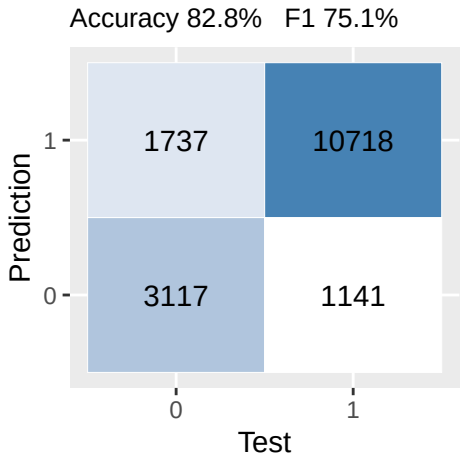
Additionally, as seen in Figure 1, features are not normal which also introduces difficulty for the naive Bayes model. Yet, naive Bayes is an improvement over random guessing.

Analysis of K-Nearest Neighbors

The knn model is the first to show potential. As a distance-based approach, the raw feature data will reduce accuracy as the features variety in domain, range, and distribution as visible in the Figure 2 boxplots. To remediate this, features are scaled to normalize their mean and standard deviation.

knn on the test data set produced an accuracy of 0.8278. Additionally, the F1 is also improved as noted in Figure 6.

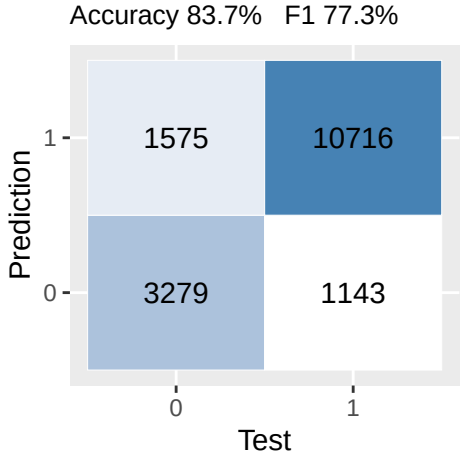
Figure 6: knn Confusion Matrix



Analysis of Decision Tree

A decision tree is a common model for learning as it is easy to explain, and particularly for this analysis, "can easily handle qualitative predictors." (Gareth et al. 2021) A number of features are categorical which benefit this model. Test data applied to the trained model produced an accuracy of 0.83737, an improvement of 67.2 percent over guessing.

Figure 7: Decision Tree Confusion Matrix

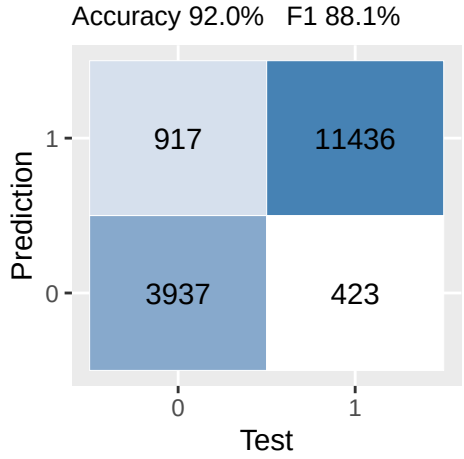


Analysis of Random Forest

The Random Forest, due to the bootstrapping (splitting the test data into groups) and randomness provides an improvement over the decision tree. As explained by Irizarry, "These two factors combined explain the name: the bootstrap makes the individual trees *randomly* different, and the combination of trees is the *forest*." (Irizarry 2020) Additionally, these qualities help reduce overfitting. IBM states, "when there's a robust number of decision trees in a random forest, the classifier won't overfit." (IBM 2020) Also, random forest reduces the impact of outliers in the model as noted by the Corporate Finance Institute, "cases of missing values and outliers have less significance on the decision tree's data." (Corporate Finance Institute 2022). In this analysis, random forest is among the better performing models.

The random forest confusion matrix is presented in Figure 8. Random forest accuracy was 0.91982. Which is an improvement of 83.6 percent over guessing.

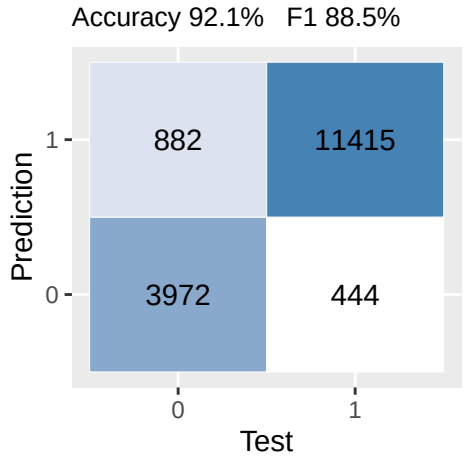
Figure 8: Random Forest Confusion Matrix



Analysis of xgBoost

XgBoost is a well known model for its performance. It replaces the bootstrapping of random forest with boosting where, "the trees are grown sequentially: each tree is grown using information from previously grown trees." (Gareth et al. 2021) Although the improvement over random forest is slight, it also retains a favorable level of F1. The xgBoost produced an accuracy 0.92066. Which is an improvement of 83.8 percent over guessing.

Figure 9: xgBoost Confusion Matrix



Conclusion

Of the models applied in this analysis, naive Bayes produced the weakest performance of 72.99. The xgBoost model produced the best performance at 92.07. In addition to an improvement in accuracy, the F1 value of xgBoost is also improved as shown in Table 2 below.

Table 2: Summary Test Results

Method	Accuracy	F1	Improvement
random guess	0.50093	0.50102	0.00000
naive bayes	0.72985	0.60638	0.45700
knn	0.82780	0.75083	0.65253
rpart	0.83737	0.77310	0.67164
rf	0.91982	0.88109	0.83624
xgboost	0.92066	0.88458	0.83791

Verification Results

As a cross-check, a verification run of the top two models was performed. Verification on random forest produced an accuracy of 92.13. XgBoost accuracy on verification data is 92.01. Details of the verification listed in Table 3 compare favorably for random forest and xgBoost test runs detailed in Table 2. Both models limited the effect of outliers and overfitting.

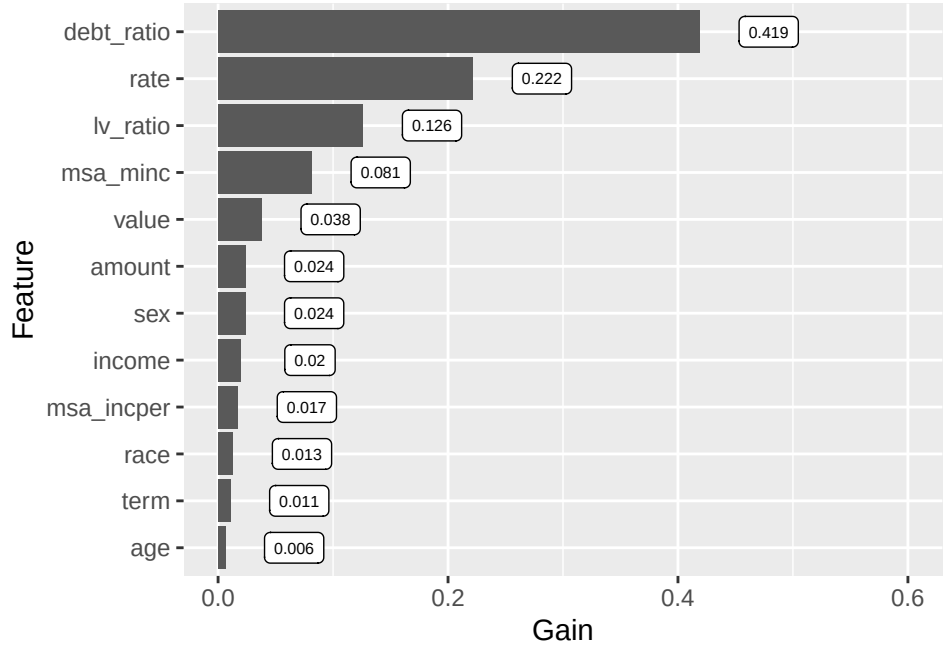
Table 3: Summary Verification Results

Method	Accuracy	F1
Ver rf	0.92132	0.88399
Ver xgb	0.92006	0.88765

Returning to the pubic concerns with regard to mortgage loans, the feature importance provides some insight into the expected drivers of underwrite decision making. Figure 10 below shows the importance variables of the best performing model, xgBoost. Variables with low importance in underwriting include: age, term and race. Sex is relatively low as well. One could conclude that bank underwriting is achieving the social goals of these features because their importance is low, inferring they are not impactful in loan underwriting.

The top three variables in terms of importance are: debt ratio, interest rate and loan-to-value ratio. This could be expected as applicant debt ratio is a strong indicator of confidence in the applicants ability to re-pay. Importance of loan-to-value would signify the level of risk protection the bank could hold in the property. Interest rate splits the top and third importance variables, which could infer the criticality of loan profitability for the lender.

Figure 10: xgBoost Variable Importance



Future Study

Due to the compute environment for this analysis, loan data was significantly restricted from over 2.6M loans to 528,151. A processing environment that could handle the entire 2021 data set would certainly impact the outcome of this analysis.

Tuning ML models is a laborious process in the RStudio environment. Cloud platform solutions such as *DataRobot* (Data Robot Corporation 2022) could not only handle the many giga-bytes in the original loan data set, but its built-in tuning capabilities applied automatically across many ML models, could produce better outcomes.

Computational Environment

This analysis was performed using RStudio 2022.07.1, build 554 with R 4.2.1, running on a MacBook Pro (2018) with an Intel Core I7 and 16GB Memory.

References

Corporate Finance Institute. 2022. "Random Forest." https://corporatefinanceinstitute.com/resources/knowledge/other/decision-tree/.

Data Robot Corporation. 2022. "Data Robot." https://www.datarobot.com/.

FFEIC. 2021. "Dynamic National Loan-Level Dataset." https://ffiec.cfpb.gov/data-publication/dynamic-national-loan-level-dataset/2021.

FRED. 2021. "FRED Economic Data, Real Median Household Income by State." https://fred.stlouisfed.org/release/tables?eid=259515&rid=249#.

Gareth, James, Danella Witten, Trevor Hastie, and Robert Tibsharani. 2021. *An Introduction to Statistical Learning*. New Yory, NY: Springer.

IBM. 2020. "Random Forest." https://www.ibm.com/cloud/learn/random-forest.

Irizarry, Rafael A. 2020. *Introduction to Data Science*. Boca Raton, FL: CRC Press.