Will I Get to Buy My Dream House? - A Home Mortgage Loan Prediction Project

Authors: Ranjini (rr735), Murali (mt788), Subham (sb2564)

Abstract

The report outlines our approach to predict the loan approval process along with the denial reason. We start by identifying the key input features and look at their distributions. Following this, four models are built to predict the outputs action taken and denial reason like Support Vector Machine, XGBoost, Random Forest, Naive Bayes and Logistic Regression. **The best model identified in both cases after carefully cross validating and tuning shows logistic Regression performing well.** We finish the analysis by highlighting the key findings and show that our final model may not be biased against gender.

Section 1: Introduction

We have analyzed the HMDA (Home Mortgage disclosure act) dataset which contains the most comprehensive and publicly available information on mortgage market activity. The source of this data is the Consumer Financial Protection Bureau (CFPB). We have used the data from the year 2017 for our analysis. The loan origination process is a classification problem and we have built several models that were able to predict the binary response (Approved/Denied). Furthermore, we have also created a pipeline to predict the denial reasons such as Credit History and Debt to Income ratio for unapproved loans. In addition, we have applied Principal Component Analysis to study the variations in data and have built a logistic model with uncorrelated features. Finally, we studied gender bias in our model.

Section 2: Exploratory Data Analysis

As a first step in the data cleanup process, we imported the data as a csv file containing 100,000 data points and all those records that had missing values. We removed features like rate_spread which had too many missing values. On top of this, after carefully looking at categories involving NAs, we referred to guidance to decide allowed NAs as an option. The dataset finally had 19 variables and 84,157 records. *Table A.1* in the appendix shows a brief description of the features finally chosen. The stacked bar plots (*Figures 2.1, 2.2, A.2, A.3)* were constructed indicating the total count for the action taken on the loan (dependent variable) – Originated/ Denied for each category of the feature variables. Scatterplots of all the features are shown in *Figure A.1.*

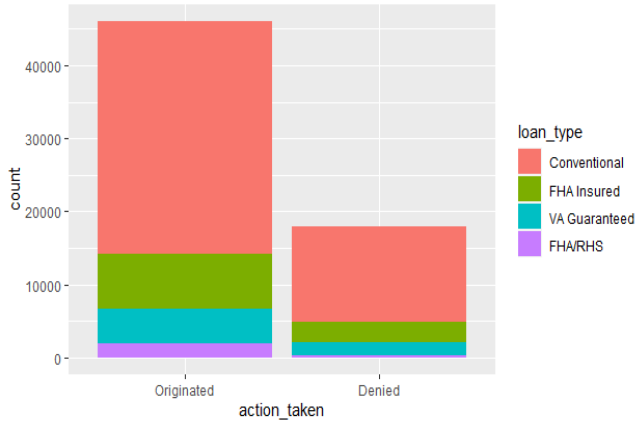


Figure 2.1

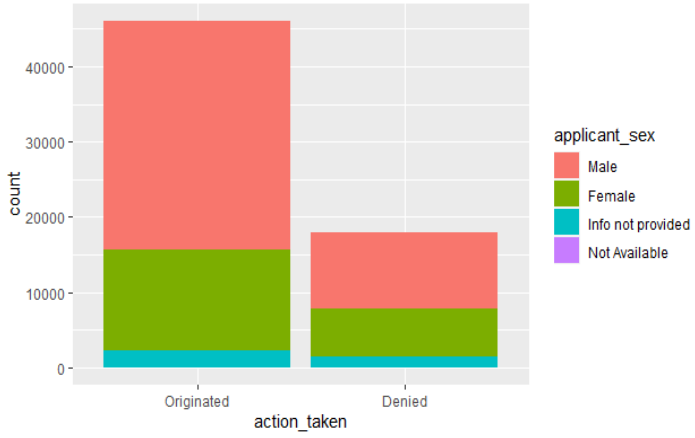


Figure 2.2

Section 3: Model Building and Metric

The first set of models built was for predicting the action taken. Table 3.1 captures the baseline model results obtained after sampling an equal number of samples in both classes of action_taken. **The metric we picked and therefore universally reported is Balanced Accuracy.** We realized the need to look at the differentiability power of the model as opposed to raw accuracy which may be biased on the number of examples per class and not show the full picture. *Figure 3.1* visualizes the model

building pipeline that we followed where we built two models one to predict action taken and the other is denial reason prediction in a sequential way.

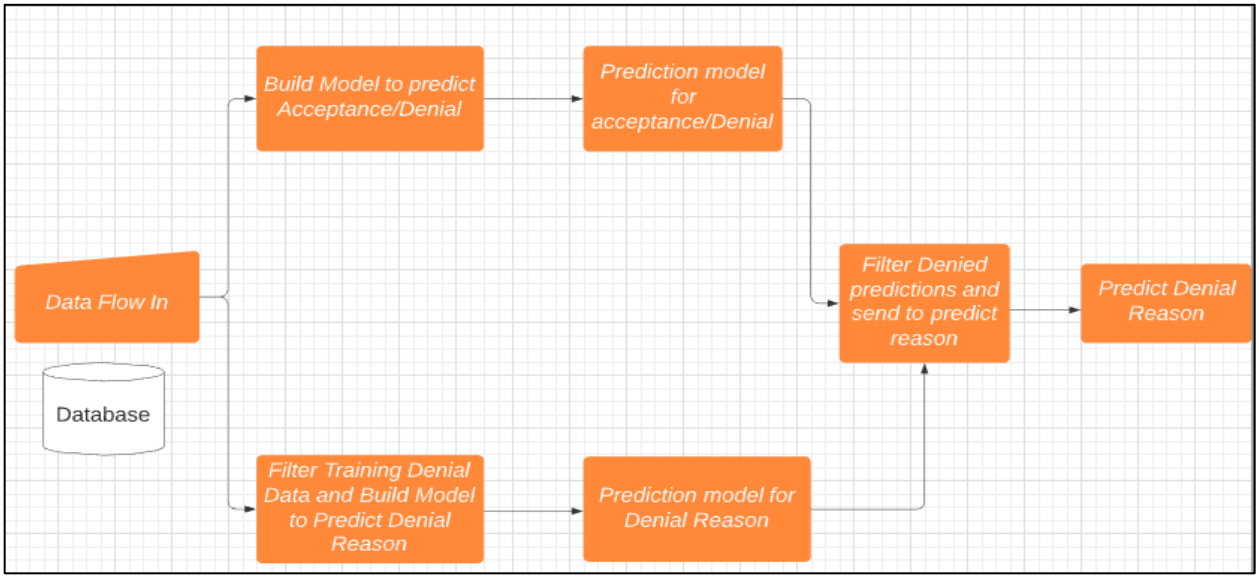


Figure 3.1: Model Building and inference Pipeline Visualization

Table 3.1: Results for Action Taken Prediction

Model	Training Balanced Accuracy	Testing Balanced Accuracy	Cross Val Accuracy (Std dev)	Tuned Training Balanced Accuracy	Tuned Testing Balanced Accuracy
Naive Bayes (Gaussian)	66.98	65.75	66.90% (0.53%)	-	
Random Forest	98.18	67.10	67.71% (0.64%)	87.90	71.52
XGBoost Classifier	71.98	70.80	71.30% (0.57%)	70.85	70.95
Logistic Classifier	69.26	69.20	69.11% (0.54%)	69.58	69.36
SVM with rbf kernel	99.77	55.02	55.29% (0.90%)	-	
Polynomial function (GAM)	70.95	70.53	-	-	
Natural Spline	70.64	70.83	-	-	

Table 3.2: Results for Denial Reason Prediction

Model	Training Balanced Accuracy	Testing Balanced Accuracy	Cross Val Accuracy (Std dev)	Tuned Training Balanced Accuracy	Tuned Testing Balanced Accuracy
Naive Bayes (Gaussian)	61.29	60.36	61.18% (2.05%)	-	
Random Forest	98.31	58.86	60.94% (1.77%)	88.47	64.43
XGBoost Classifier	98.75	61.28	62.60% (1.48%)	68.05	64.80
Logistic Classifier	63.76	63.15	63.33% (2.23%)	63.87	63.07
SVM with rbf kernel	99.96	51.71	51.72% (0.90%)	-	

The "best" models seem to be doing equally well within 65-71% Balanced Accuracy and so we pick GAM and logistic regression because of the interpretability advantage. It is also interesting to see that both Random Forest and SVM badly overfit the training data. We also chose to balance the data because the results obtained were better than the unbalanced case.

The fact that logistic regression model does almost as well as the other tuned models shows that the data must be *inherently linearly separable* and that this generalizes well to the testing set. The most important features identified include loan amount, owner occupancy, and property type. We also fitted a GAM model with quadratic function for the quantitative predictor variables and a step function for the categorical variables. Visual inspection of the *Figures 3.2*, *3.3* indicate that there may not be a gender bias in the data. An interesting insight observed is that increased population in a tract

tends to be associated with higher probability of loan approvals. This can be attributed to higher economic activity causing people to stay large numbers (Eg. NYC/SFC). Some more plots (*Figures A.4, A.5*) can be found in the appendix.

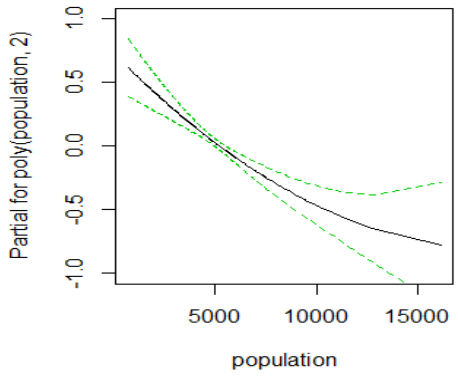


Figure 3.2

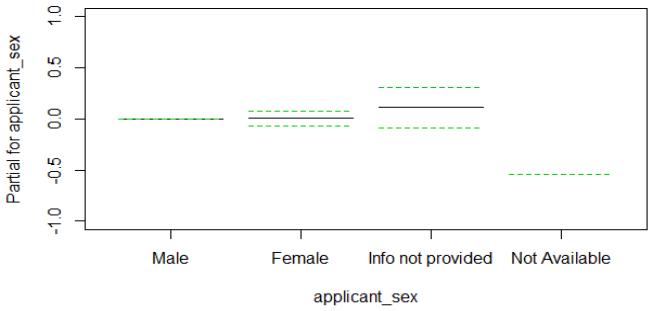


Figure 3.3

Section 4: Principal Component Analysis and Regression

In this section, we have explored the role of continuous predictors in the loan approval process. The idea behind this analysis is to build a model that is agnostic to categorical variables such as ethnicity, gender, etc.

The correlation plot (*Figure 4.1)* clearly shows that several variables are correlated with each other suggesting multicollinearity (high dependence) within predictors. To tackle this issue, we implement Principal Component Analysis. PCA helped us visualize the variations in data. The Variables - PCA plot *(Figure 4.2)* shows the correlations between the variables and their contribution in the first two Principal Components. Positively correlated variables point to the same side of the plot as opposed to negatively correlated variables that point to opposite sides. We observe that applicant income and loan amount are positively correlated indicating that high earners typically apply for large size loans.

This seems quite natural since high earners have high purchasing power and hence might buy more expensive houses. Additionally, it is interesting to note that loan amounts and minority population are negatively correlated which indicates that large size loans are requested at tracts with less minority population. One could ponder whether equal funding opportunities are available to all tracts in the US.

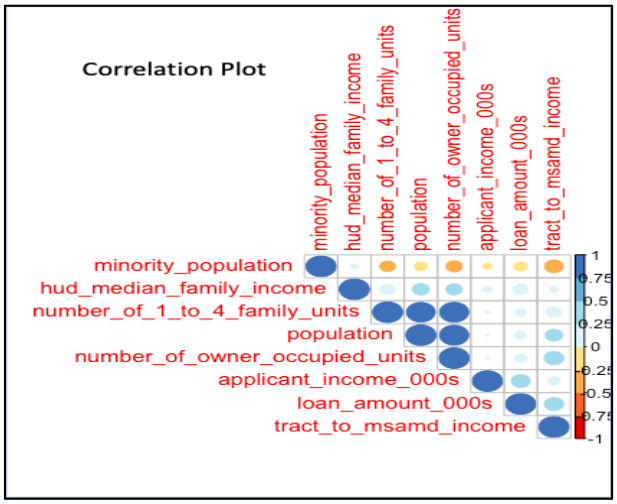


Figure 4.1

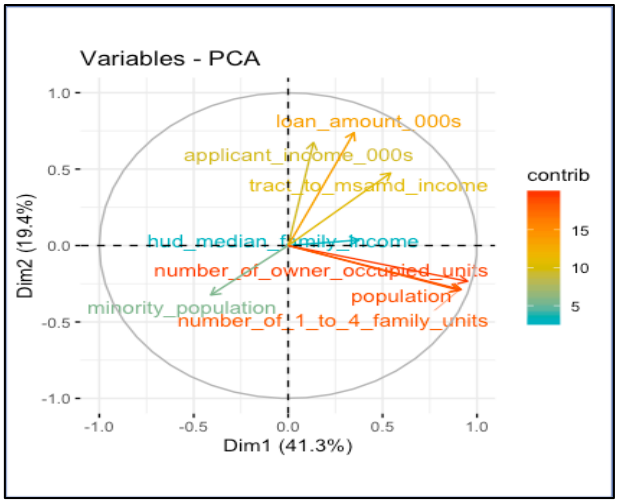


Figure 4.2

We have used these principal components as predictors in a logistic regression model to predict approval/denial on the test dataset. The Scree plot *(Figure 4.3)* shows that the first 6 PCs explain the majority (98%) of variance in the data. Furthermore, we have used 10-fold Cross Validation to find out the optimal no. of principal components to be used in our model. The AUC curve *(Figure 4.4)* for

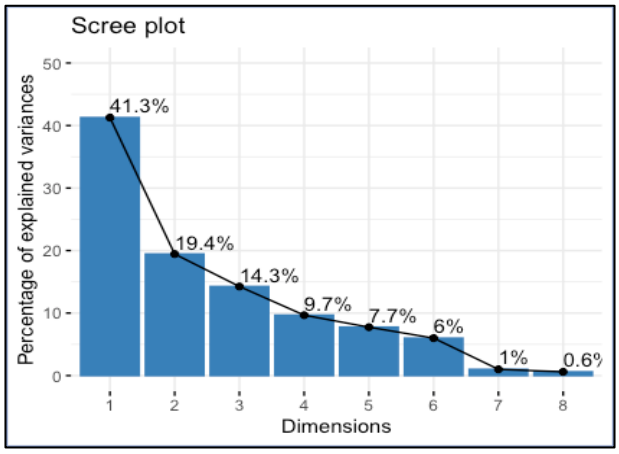


Figure 4.3

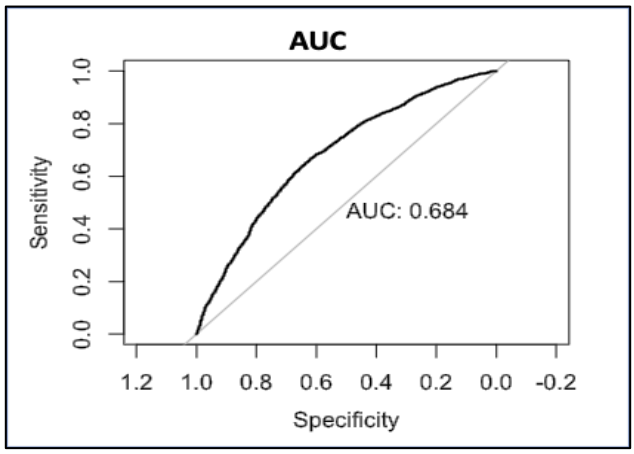


Figure 4.4

the best model using 5 PCs is shown. The Logistic classifier was able to achieve a balanced accuracy of 63.60%. PCA helped reduce the set of correlated predictor variables to a smaller, uncorrelated set that still contained most of the information in the dataset.

Section 5: Bias Studies

One interesting study we attempted was to check the bias in the XGBoost model built. For this we calculate the following rates.

$$Gender\ Acceptance\ Rate = \frac{Number\ of\ approved\ loans\ for\ that\ gender}{Total\ Number\ of\ loans\ applied\ by\ that\ gender}$$

Table 5.1: Gender acceptance rate

Metric	Calculating with the actual data	Sign	Calculating with the predicted data
Male acceptance rate	0.58	<	0.66
Female acceptance rate	0.53	<	0.62

Table 5.2: Inverting gender and studying acceptance/rejection probabilities

Change/Probability	Acceptance Probability	Rejection Probability
Male to Female	0.46	0.53
Female to Male	0.55	0.44

From *Table 5.1 and 5.2*, it is clear that the acceptance rates have gone up in the prediction, although the gap in the acceptance rates between the two genders have dropped. Taking a look at the probabilities, it is clear that the differences are very close to 0.5 which may show that the model and the inherent data is unbiased. We conclude from our studies that our model is unbiased with some confidence and *figure 3.3* supports this. However, the model has higher acceptance rates for both genders.

Section 6: Conclusion

From the discussion above, it is clear that logistic regression performs well for both the models suggesting linear separability is inherent to the data. The model and the inherent data does not seem to be biased against gender which is a finding we report. It is important to note that we used only 100,000 data points because of computational capabilities and that additional data or even sampling in a different way may lead to a different result depending on the input distribution. The overall pipeline accuracy is 44.10% Balanced Accuracy which is greater than random guess (Baseline is 33% Balanced Accuracy). As an additional step, the bias in the model was checked from a gender perspective and we show that our model may not be biased. There may be other factors such as co-applicant sex which we did not venture into but leave as potential future work.

Section 7: Bibliography

1. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
2. https://www.consumerfinance.gov/data-research/hmda/

Section 8: Appendix

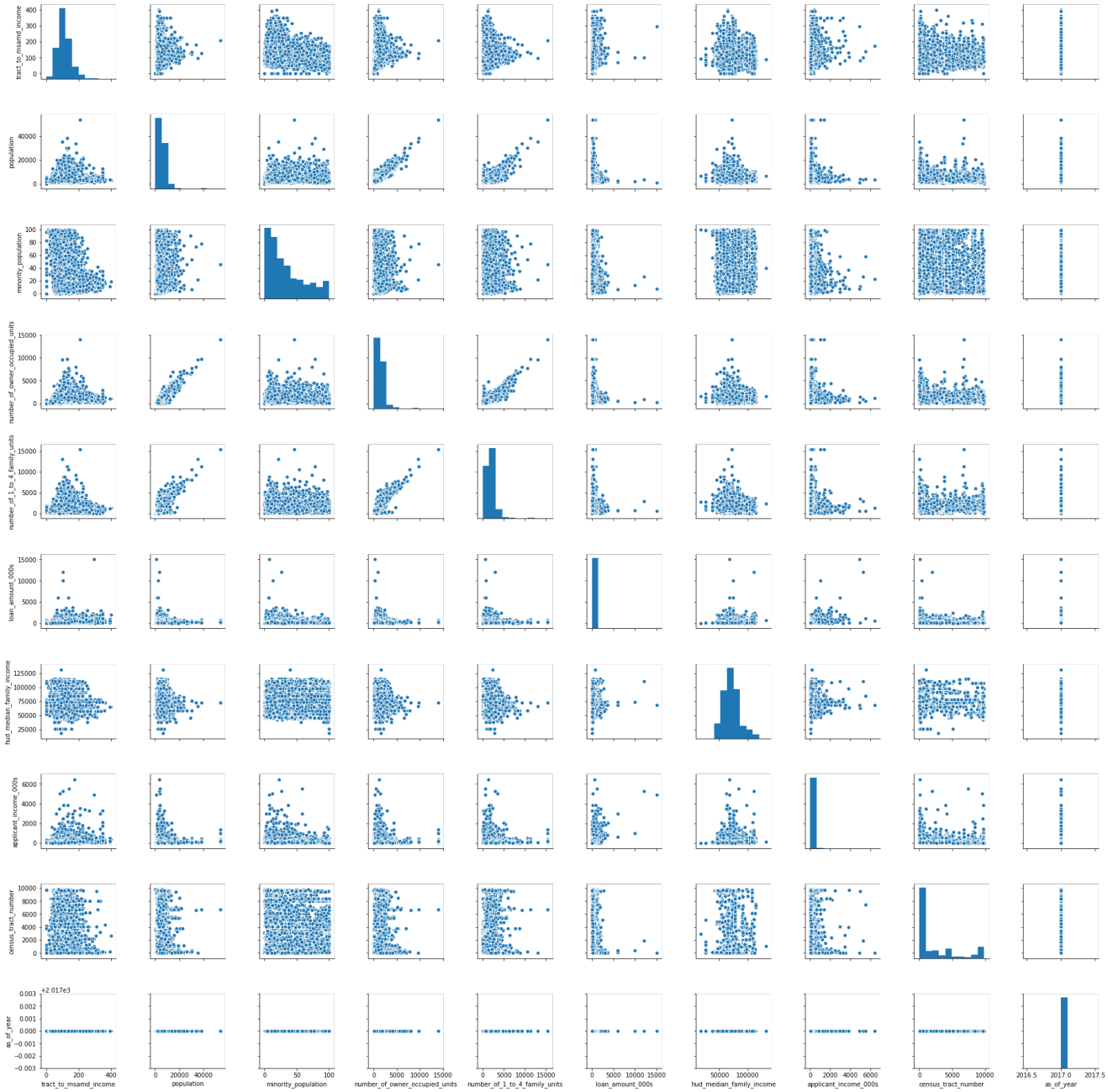
Table A.1: Variable Names and Associated Types

S.No	Variable Name	Type	Categories/Description
1	Loan Type	Categorical	1-- Conventional 2 -- FHA-insured 3--VA-guaranteed 4 -- FSA/RHS
2	Property Type	Categorical	1 -- One to four-family 2 -- Manufactured housing
3	Loan Purpose	Categorical	1 -- Home purchase 2 -- Home improvement 3 -- Refinancing
4	Owner Occupancy	Categorical	1.Owner-occupied as a principal dwelling 2 -- Not owner-occupied 3 -- Not applicable
5	Action Taken	Categorical	1 -- Loan originated 3 -- Application denied
6	Applicant Ethnicity	Categorical	1 -- Hispanic or Latino 2 -- Not Hispanic or Latino 3 --Info not provided 4 -- Not applicable 5 -- No co-applicant

7	Applicant Sex	Categorical	1 -- Male 2 -- Female 3 -- Info not provided 4 -- Not applicable 5 -- No co-applicant
8	Loan Amount	Continuous	The amount of the covered loan/the amount applied for.
9	tract_to_msamd income	Continuous	Percentage of tract median family income to MSA/MD family income
10	Applicant Income	Continuous	If a credit decision is made, gross annual income relied on in making the credit decision/ if a credit decision was not made, the gross annual income relied on in processing the application.
11	HOEPA Status	Categorical	1 -- HOEPA loan 2 -- Not a HOEPA loan
12	Lien Status	Categorical	1 -- Secured by a first lien 2 -- Secured by a subordinate lien 3 -- Not secured by a lien 4 --Not applicable
13	Number of owner-occupied units	Continuous	Number of dwellings, including individual condominiums, that are lived in by the owner

14	number_of_1_to_4_family_units	Continuous	Dwellings that are built to house fewer than 5 families.
15	Population	Continuous	Total population in tract. Loans were only filed for tracts that had population over 30,000
16	minority_population	Continuous	Percentage of minority population to total population for tract.
17	hud_median_family_income	Continuous	Housing and urban development median family income
18	co_applicant_ethnicity	Categorical	1 -- Hispanic or Latino 2 -- Not Hispanic or Latino 3 -- Info not provided 4 -- Not applicable 5 -- No co-applicant
19	co_applicant_sex	Categorical	1 -- Male 2 -- Female 3 -- Info not provided 4 -- Not applicable 5 -- No co-applicant

Figure A.1. Distribution Plots of Input Features



Additional plots and visualization

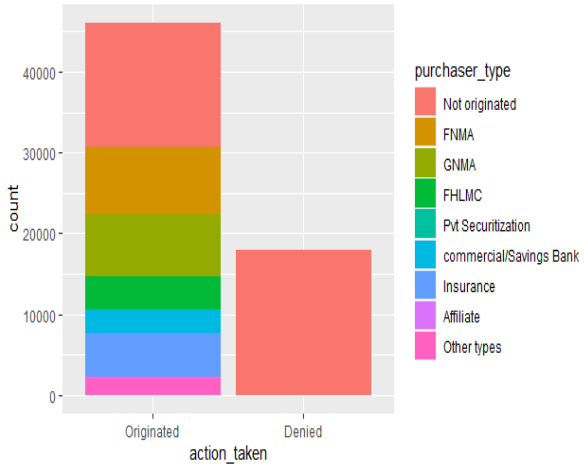


Figure A.2

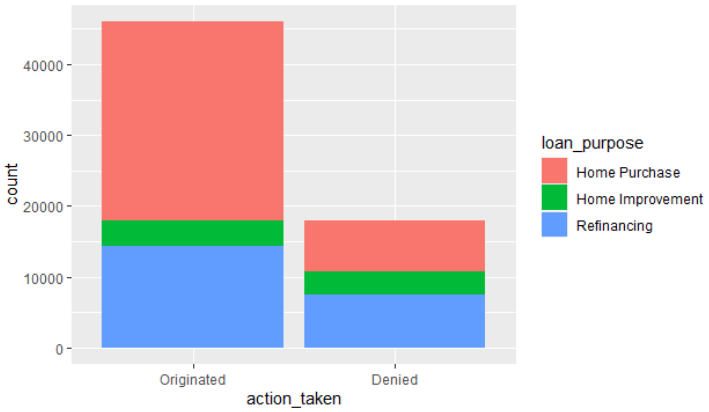


Figure A.3

The following plots indicate that in tracts where there is a higher minority population, there is a higher probability of loans being denied which implies some bias. A mild parabolic curve associated with HUD median family income implies there is a very small bias against the higher and lower median family income groups compared to the mid- income level groups for loan approvals.

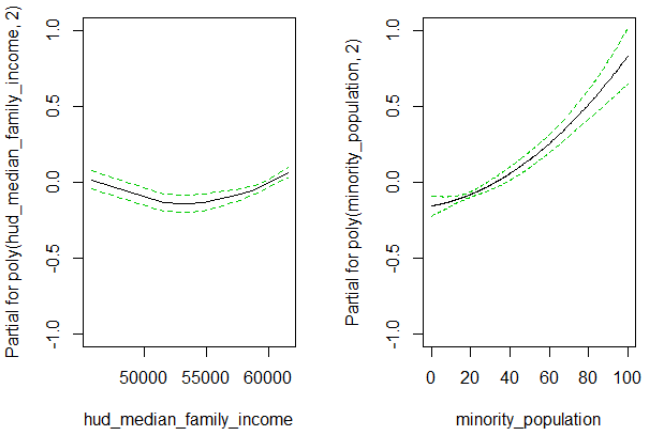


Figure A.4

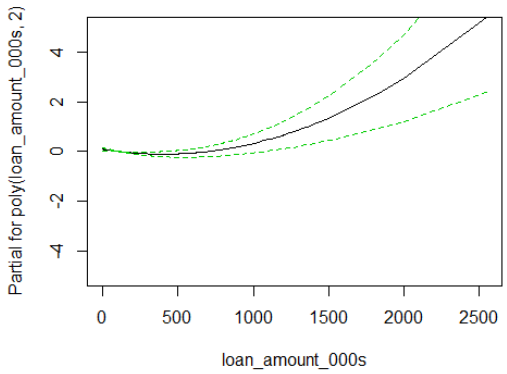


Figure A.5