Binary Classification of Mortgage Approval from Government Data

Pierre du Pont

April 22, 2019

Executive Summary

This document presents an analysis of data concerning mortgages and their approval. This analysis is based on data adapted from a datest ccreated by the Federal Financial Institutions Exemination Council (FFIEC). The data included one million total observations split into two sets—a test set with 500000 observations and a training set with 500000 observations. Each observation contains information around the applicant, the loan, and whether it was accepted. This report and analysis operates on the assumption that the train set is representative of the test set.

Initial analysis was completed by calculating summary and descriptive statistics along with visualizations of the data. After exploring the data, several models were used to predict mortgage acceptance in the test set. Throughout this process, additional features were created to more accurately represent important factors in mortgage acceptance.

After performing the analysis, the author presents the following conclusions:

Many factors are necessary to accurately predict mortgage approval. Significant features found in this analysis included

- **Loan Percent of Income** – the ratio of the loan amount to the applicant's annual income. Applicants with a higher ratio were less likely to be accepted.
- **MSA MD, State, and County** – certain localities had much higher acceptance rates than others, perhaps based on the average income of those localities. More analysis needs to be done to determine the cause.
- **Applicant Sex, Race, and Ethnicity** – despite laws against housing discrimination based on (among other things) race, gender, or ethnicity, there was a substantial difference between different races. Applicants who were white and/or male were more likely to be approved than other groups.
- **Loan Purpose** – home purchases were more likely to be accepted than re-financing or home improvement.
- **Co-Applicant** – people who applied with a co-applicant (for example, a spouse) were more likely to be accepted

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics around the train set. The columns in the data set are as follows (with descriptions taken from the HMDA Loan Application Register code sheet)

Feature Information

Row ID

A unique identifier with no intrinsic meaning

Loan Type

Indicates whether the loan granted, applied for, or purchased was conventional, government-guaranteed, or government-insured; available values are:

1. Conventional (any loan other than FHA, VA, FSA, or RHS loans)
2. FHA-insured (Federal Housing Administration)
3. VA-guaranteed (Veterans Administration)
4. FSA/RHS (Farm Service Agency or Rural Housing Service)

Property Type (categorical)

Indicates whether the loan or application was for a one-to-four-family dwelling (other than manufactured housing), manufactured housing, or multifamily dwelling; available values are:

1. One to four-family (other than manufactured housing)
2. Manufactured housing
3. Multifamily

Lender (categorical)

A categorical with no ordering indicating which of the lenders was the authority in approving or denying this loan

Loan Amount (int)

Size of the requested loan in thousands of dollars

Loan Purpose (categorical)

Indicates whether the purpose of the loan or application was for home purchase, home improvement, or refinancing; available values are:

1. Home purchase
2. Home improvement
3. Refinancing

Occupancy (categorical)

Indicates whether the property to which the loan application relates will be the owner's principal dwelling; available values are:

1. Owner-occupied as a principal dwelling
2. Not owner-occupied
3. Not applicable

Preapproval (categorical)

Indicate whether the application or loan involved a request for a pre-approval of a home purchase loan; available values are:

1. Preapproval was requested
2. Preapproval was not requested
3. Not applicable

MSA MD (categorical)

A categorical with no ordering indicating Metropolitan Statistical Area/Metropolitan Division where a value of -1 indicates a missing value.

State Code (categorical)

A categorical with no ordering indicating the U.S. state where a value of -1 indicates a missing value

County Code (categorical)

A categorical with no ordering indicating the county where a value of -1 indicates a missing value

Applicant Income (int)

Applicant income in thousands of dollars

Applicant Ethnicity (categorical)

Ethnicity of the applicant; available values are:

1. Hispanic or Latino
2. Not Hispanic or Latino
3. Information not provided by applicant in mail, Internet, or telephone pplication
4. Not applicable

Applicant Race (categorical)

Race of the applicant; available values are:

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 applicant in mail, Internet, or telephone application
7. Not applicable

Applicant Sex (categorical)

Sex of the applicant; available values are:

1. Male
2. Female
3. Information not provided by applicant in mail, Internet, or telephone application
4. Not applicable

Population

Total population in tract

Minority Population Pct

Percentage of minority population to total population for tract

FFIEC Median Family Income

FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)

Tract to MSA MD Income Pct

% of tract median family income compared to MSA/MD median family income

Number of owner occupied units

Number of dwellings, including individual condominiums, that are lived in by the owner

Number of 1 to 4 family units

Dwellings that are built to house fewer than 5 families

Accepted

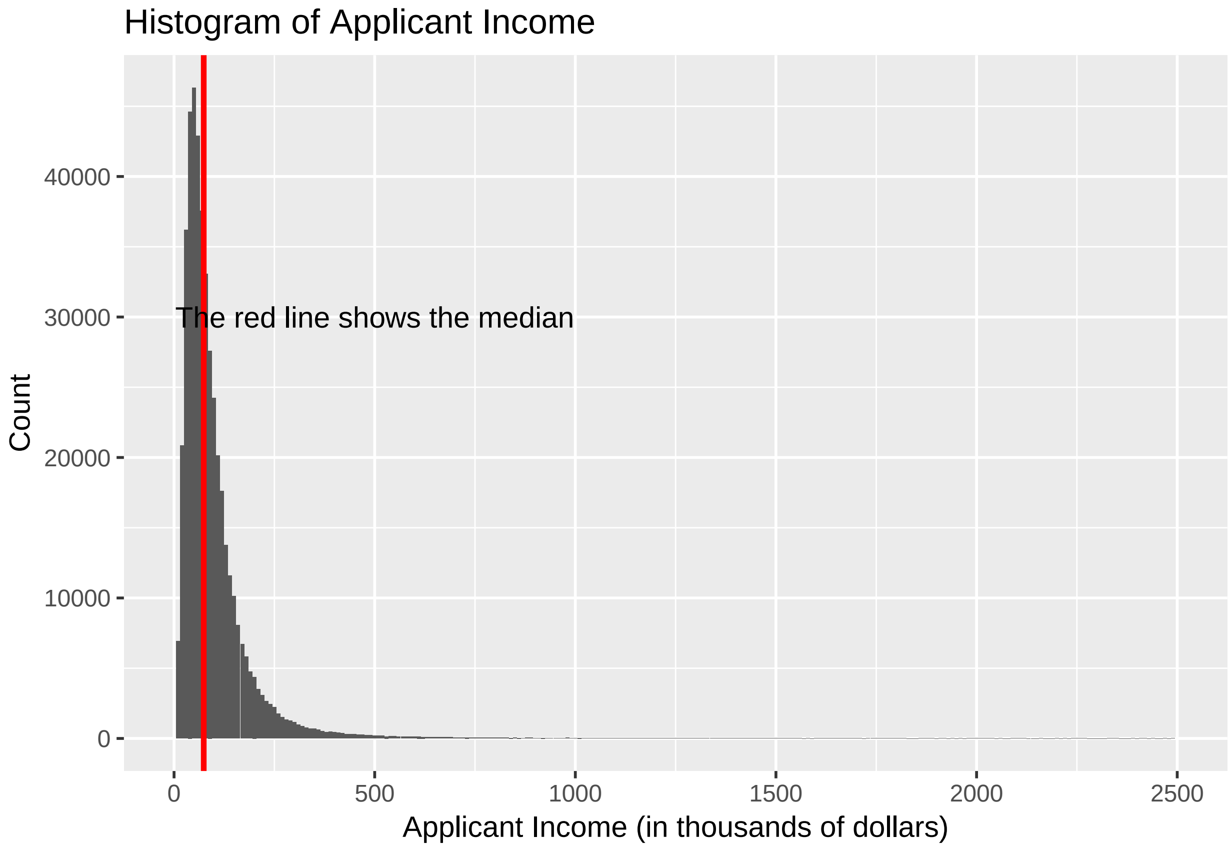
Indicates whether the mortgage application was accepted (successfully originated) with a value of 1 or denied with a value of 0. This feature is only present in the training set, and is the target variable for this analysis

Individual Feature Statistics

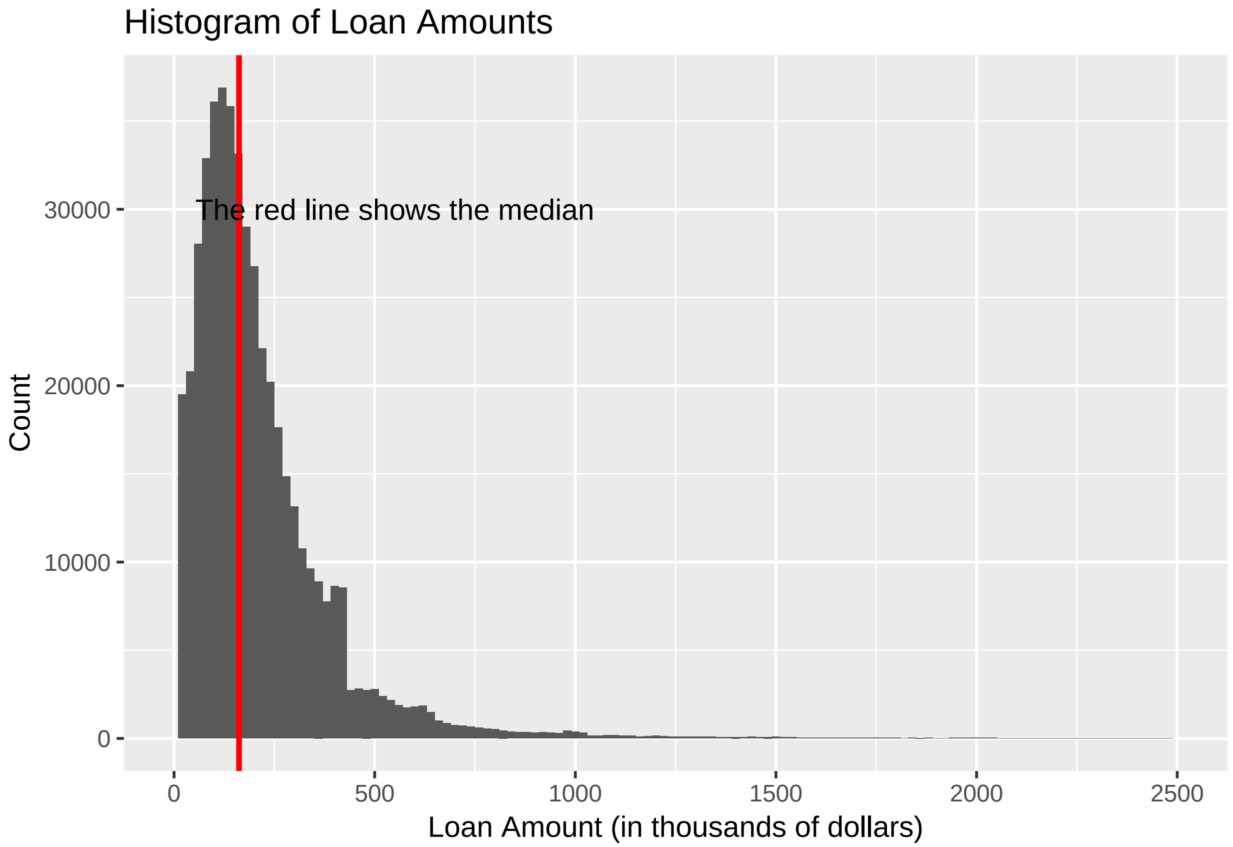
Summary statistics were computed for each numeric column in the training dataset. Results are below

```
##   loan_amount       applicant_income    population   
##  Min.   :     1.0   Min.   :    1.0   Min.   :   14  
##  1st Qu.:    93.0   1st Qu.:   47.0   1st Qu.: 3744  
##  Median :   162.0   Median :   74.0   Median : 4975  
##  Mean   :   221.8   Mean   :  102.4   Mean   : 5417  
##  3rd Qu.:   266.0   3rd Qu.:  117.0   3rd Qu.: 6467  
##  Max.   :100878.0   Max.   :10139.0   Max.   :37097  
##                     NA's   :39948     NA's   :22465  
##  minority_population_pct ffiecmedian_family_income
##  Min.   :  0.534         Min.   : 17858           
##  1st Qu.: 10.700         1st Qu.: 59731           
##  Median : 22.901         Median : 67526           
##  Mean   : 31.617         Mean   : 69236           
##  3rd Qu.: 46.020         3rd Qu.: 75351           
##  Max.   :100.000         Max.   :125248           
##  NA's   :22466           NA's   :22440            
##  tract_to_msa_md_income_pct number_of_owner-occupied_units
##  Min.   :  3.981            Min.   :   4                  
##  1st Qu.: 88.067            1st Qu.: 944                  
##  Median :100.000            Median :1327                  
##  Mean   : 91.833            Mean   :1428                  
##  3rd Qu.:100.000            3rd Qu.:1780                  
##  Max.   :100.000            Max.   :8771                  
##  NA's   :22514              NA's   :22565                 
##  number_of_1_to_4_family_units
##  Min.   :    1                
##  1st Qu.: 1301                
##  Median : 1753                
##  Mean   : 1886                
##  3rd Qu.: 2309                
##  Max.   :13623                
##  NA's   :22530
```

Most numeric columns show a slight right skew, since the median is less than the mean. This is paricularly notable in the columns Loan Amount and Applicant Income, where there are maxima over 1000 times larger than the median. A histogram of Applicant Income shows a peak near the median and a long right tail (note that this graph stops at an income of $2,500,000, even though income continues farther. This is to show the majority of the data set up close)



A histogram of loan amount shows a similar trend



Missing values in the test and train data set can cause problems with analysis. For numeric columns, missing values show up as NA, and the counts are visible in the data above. These missing values for numeric columns were replaced with the median of that column's values.

Factor statistics are below:

```
##      row_id       loan_type  property_type loan_purpose occupancy
##  0      :     1   1:370827   1:478217      1:209265     1:447394
##  1      :     1   2: 82430   2: 19741      2: 48065     2: 50417
##  2      :     1   3: 39521   3:  2042      3:242670     3:  2189
##  3      :     1   4:  7222
##  4      :     1
##  5      :     1
##  (Other):499994
##  preapproval     msa_md         state_code      county_code
##  1: 28748    -1     : 76982   37     : 61967   -1     : 20466
##  2: 60143    24     : 12993   6      : 38712   20     : 17368
##  3:411109    314    : 11014   2      : 32509   131    : 16043
##              305    : 10340   -1     : 19132   68     : 15146
##              101    : 10059   22     : 17476   11     : 14713
##              358    :  9590   47     : 17447   165    : 12448
##              (Other):369022   (Other):312757   (Other):403816
##  applicant_ethnicity applicant_race applicant_sex     lender
##  1: 50822           1:  4332       1:315806      6240   : 31685
##  2:386061           2: 25756       2:142876      5710   : 25125
##  3: 57298           3: 40495       3: 35643      3354   : 20450
##  4:  5819           4:  2339       4:  5675      5726   : 12281
##                         5:361538                     2458   : 11692
##                         6: 59862                     4701   : 10858
##                         7:  5678                     (Other):387909
##  co_applicant accepted
##  0:299974     0:249886
##  1:200026     1:250114
##
##
##
##
##
```

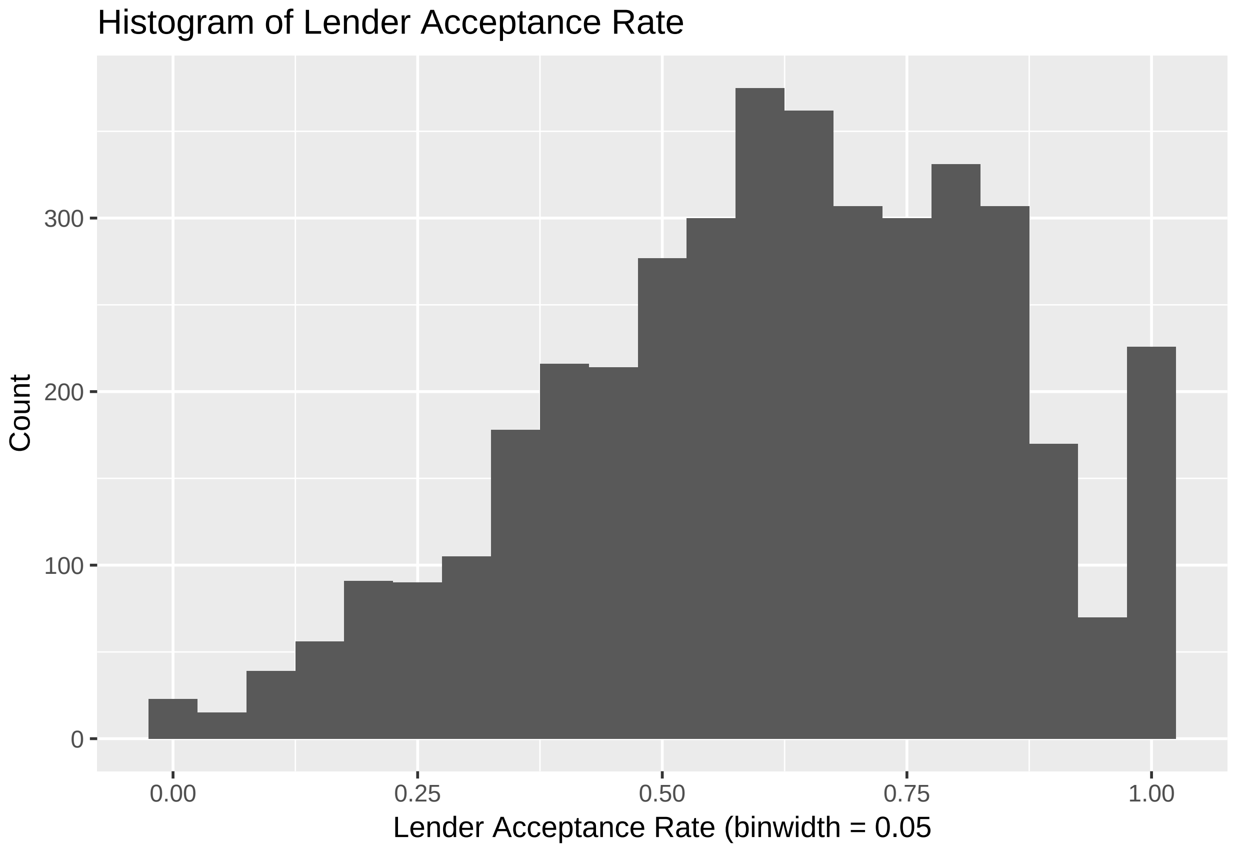
There is class imbalance in several factors, but the target variable shows no imbalance, which makes predictions easier. In particular, there is a large imbalance in property types and loan types, with very few manufactured houses or home improvement loans.

Relationships with Accepted

Several plots were created to determine the relationship between features and acceptance. As a classification problem, these plots were generally box plots or bar plots (for categorical features) or histograms (for numeric features).

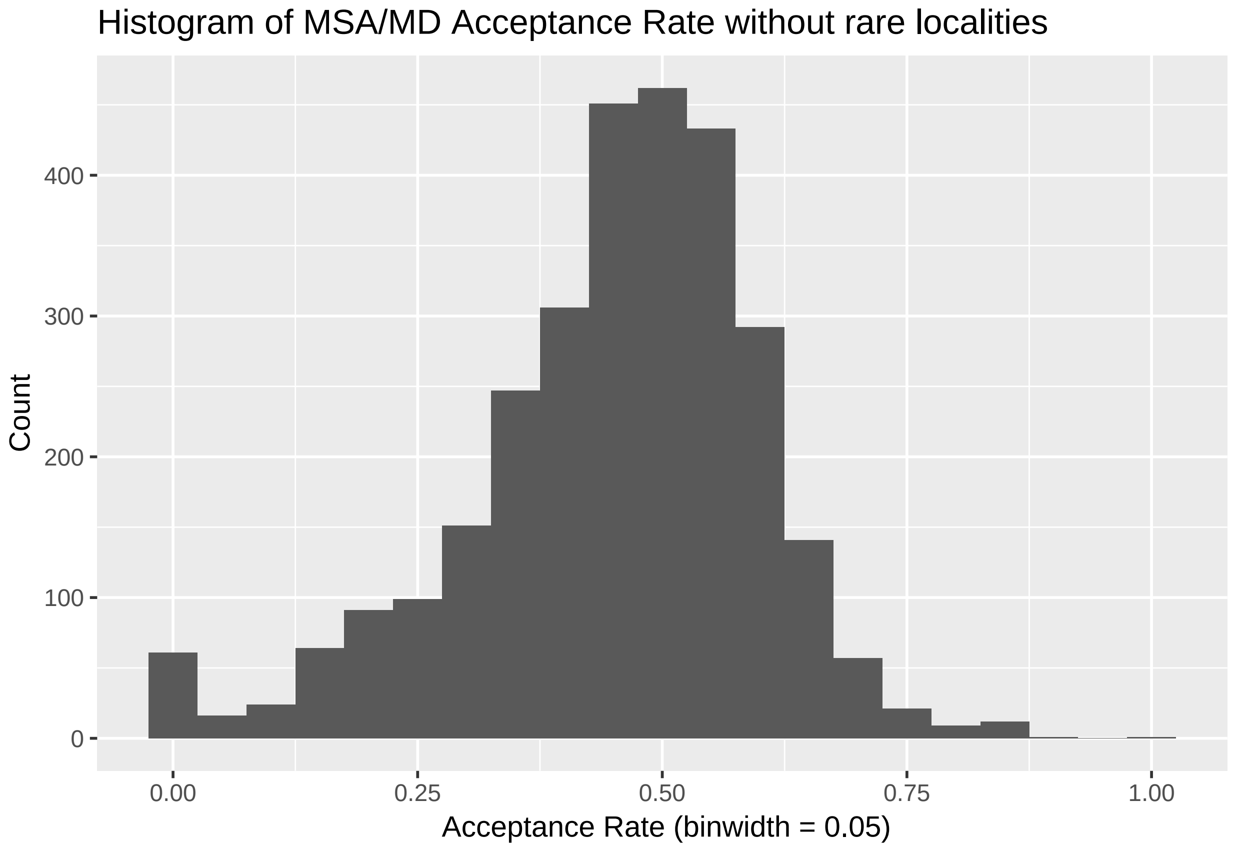
Numerical Relationships

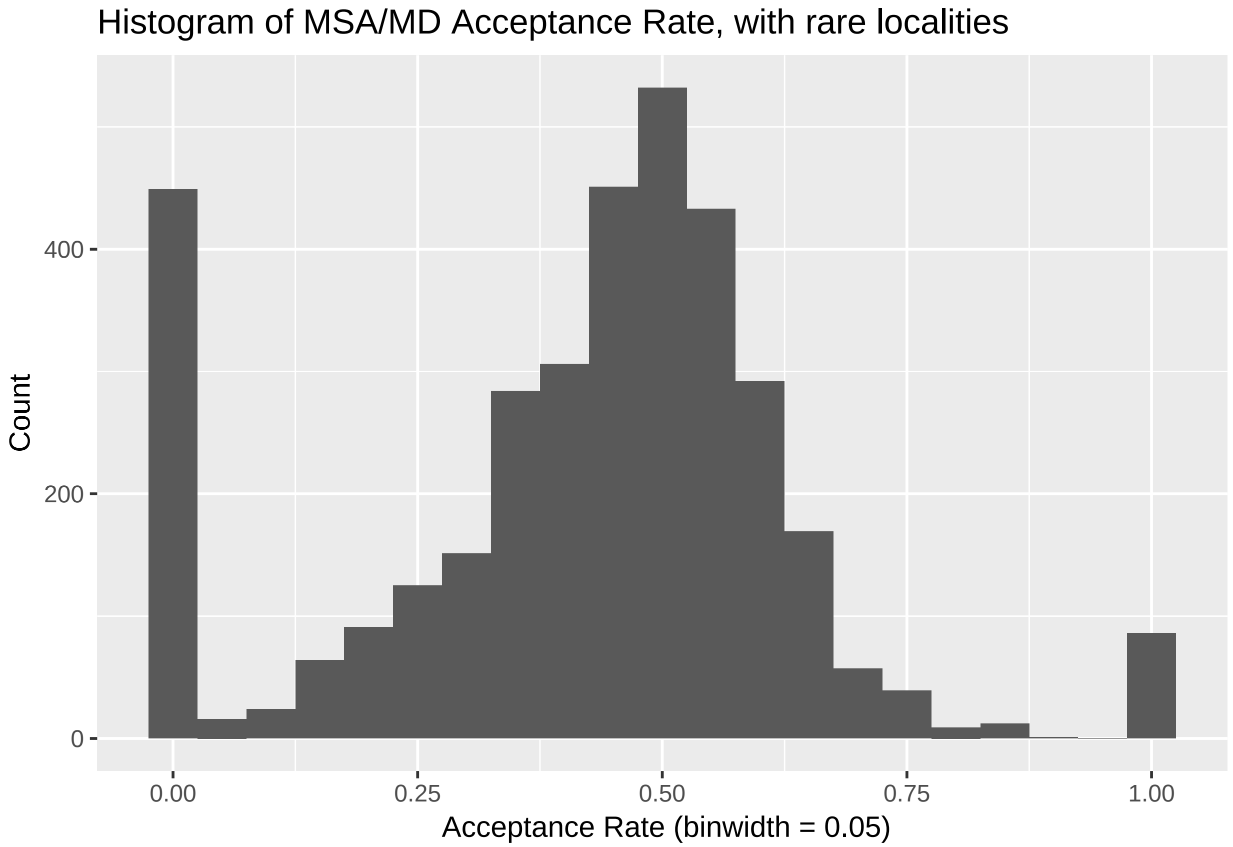
Certain lenders may be more likely to accept or decline applicants. A histogram shows the acceptance rate for lenders between 0% and 100% (calculated as accepted applications divided by total applications for each lender). Lenders who had reviewed fewer than five applications were ignored for the case of this visualization.



From this plot, it would appear that the distribution of lender acceptance is slightly left skewed, with a median near 0.65.

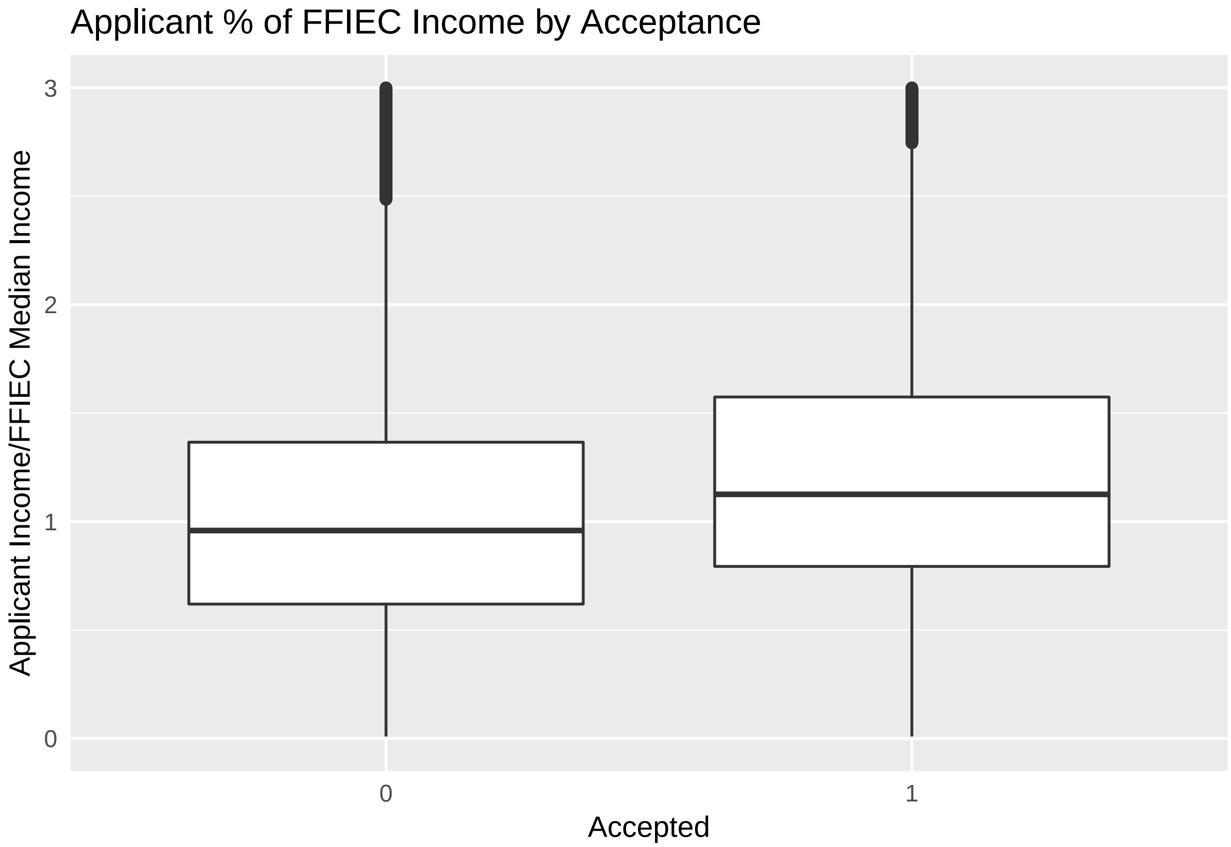
A similar plot was created for the MSA/MD acceptance rates, which shows an approximately normal distribution. This was plotted twice–once with areas with fewer than five observations, and once without. Notice the peaks at 0% and 100% that disappear without small counts.



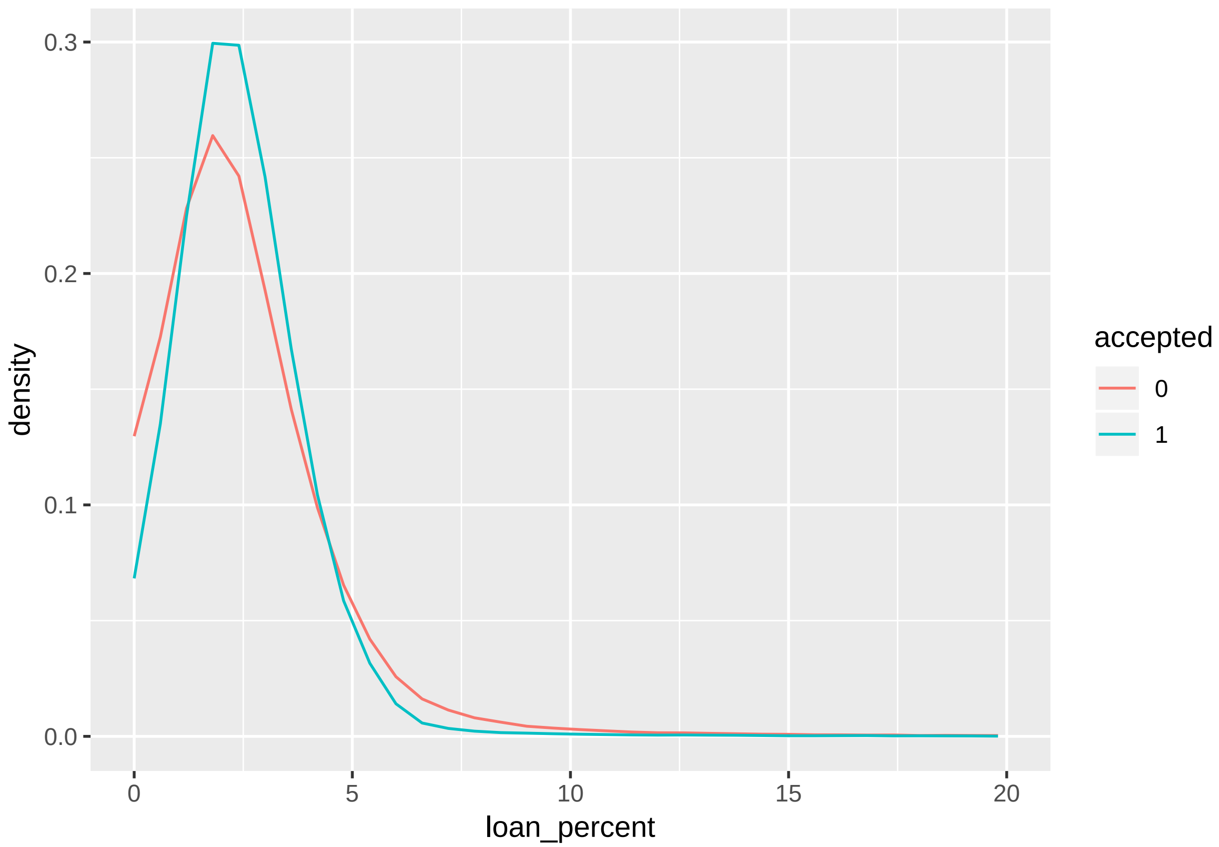


The plots above show that lender and locality acceptance rates can differ dramatically, although this analysis makes no claim as to the cause

An additional feature that appears to have a correlation with acceptance is the applicant's income compared to the FFIEC Median Income in the area. A boxplot shows a difference (note that approximately 33,000 data points were cut off the top as outliers)



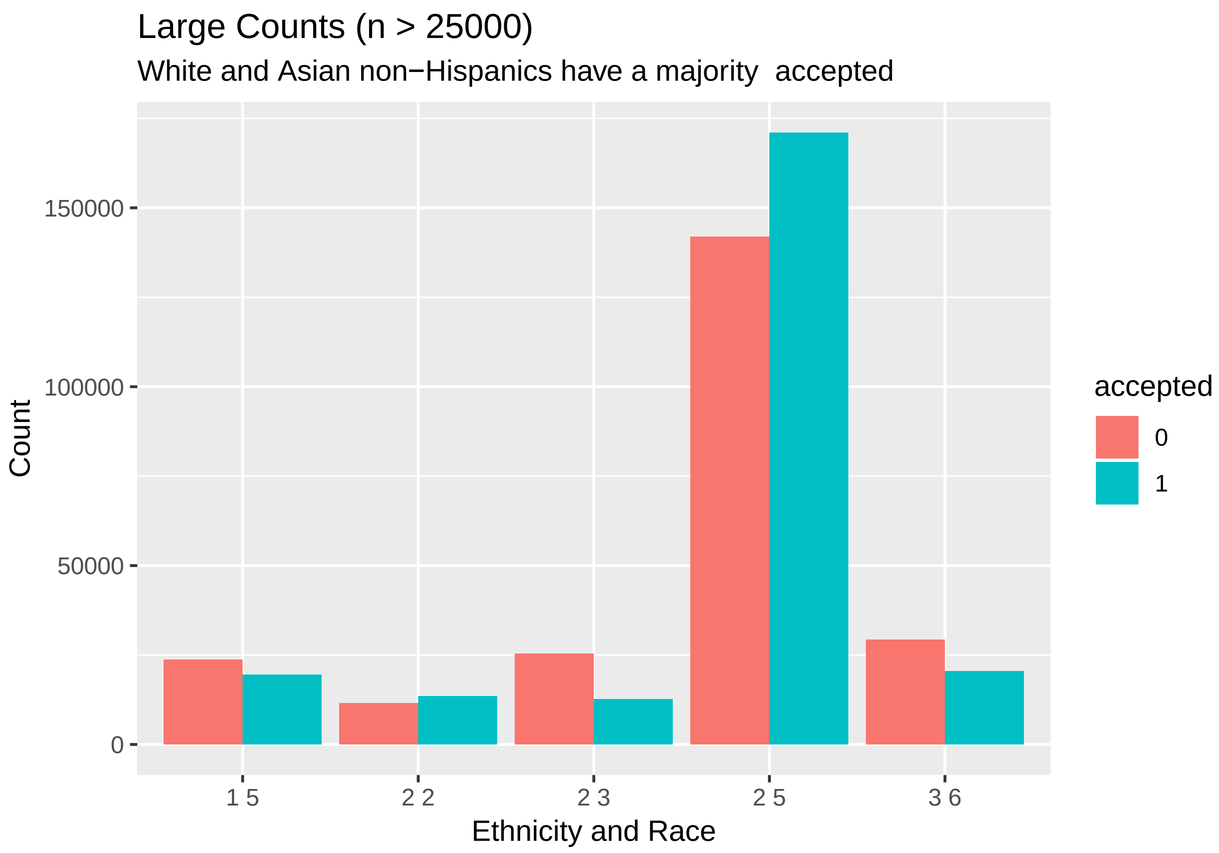
Loan percent (the ratio of loan amount to the applicant's income), also shows a small difference between accepted and denied, with a thicker tail for denied and a higher value of accepted for the lower ranges.



Categorical Relationships

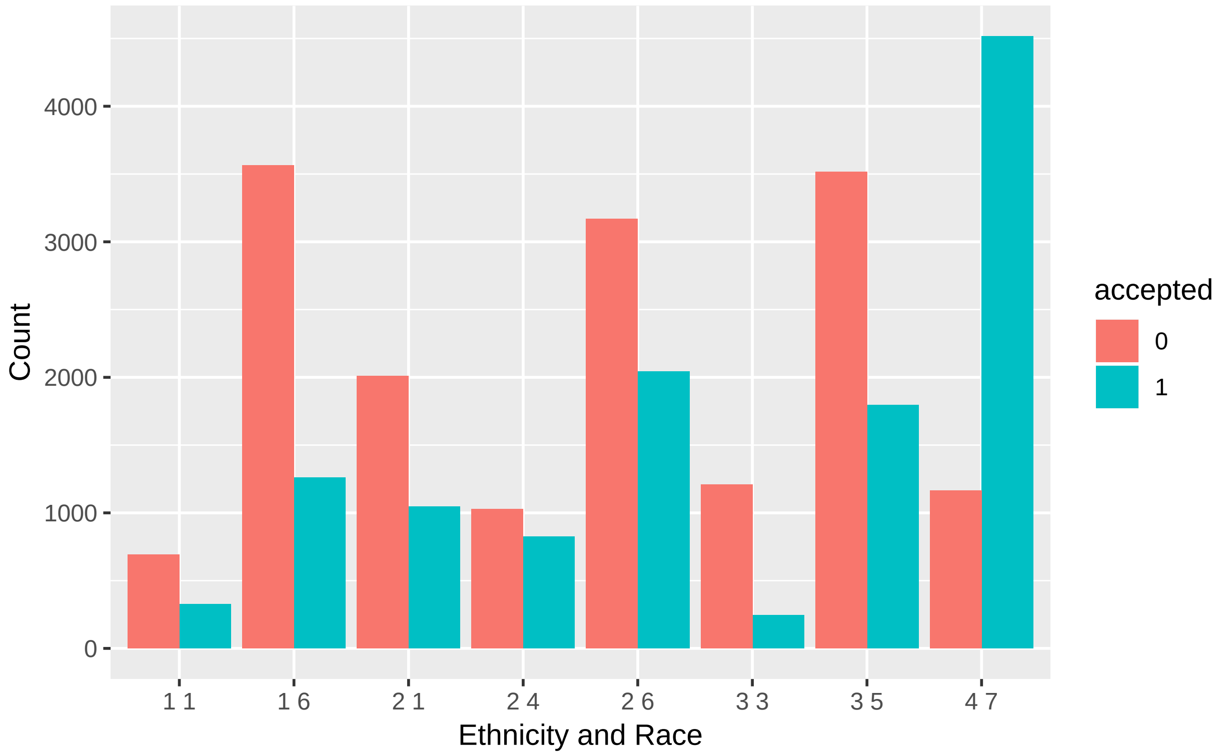
Most of the features used in the model are categorical. To visualize these, boxplots or barplots were used.

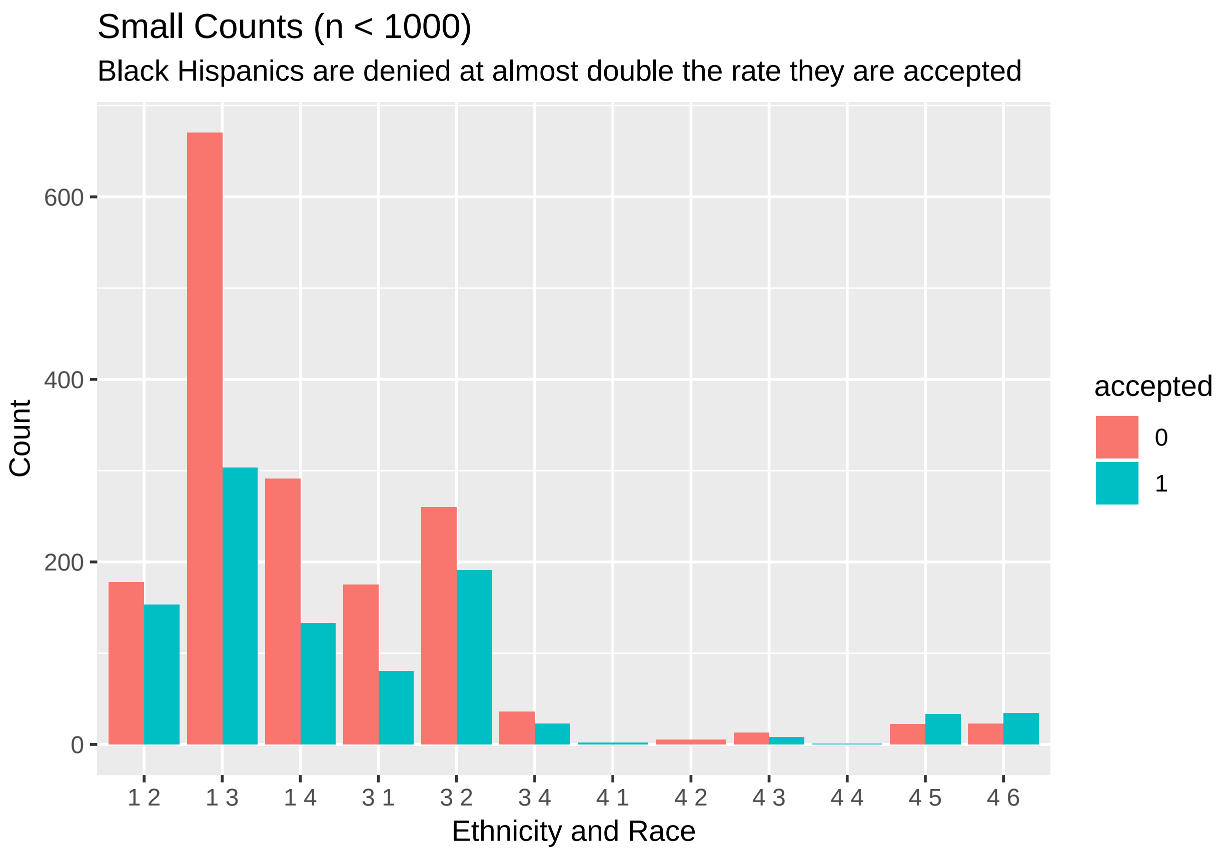
As ethnicity and race are often linked, they were combined together into one feature, which resulted in up to 28 different factors; however, some combinations were not seen. Because the counts of these combinations were very different, three plots were created so that groups with similar counts could be compared.


Large Counts (n > 25000)
White and Asian non−Hispanics have a majority accepted
Count
150000
100000
50000
0
1 5
2 2
2 3
2 5
3 6
Ethnicity and Race
accepted
0
1

Medium Counts (n > 1000)

Corporations are the only medium count where the majority are accepted



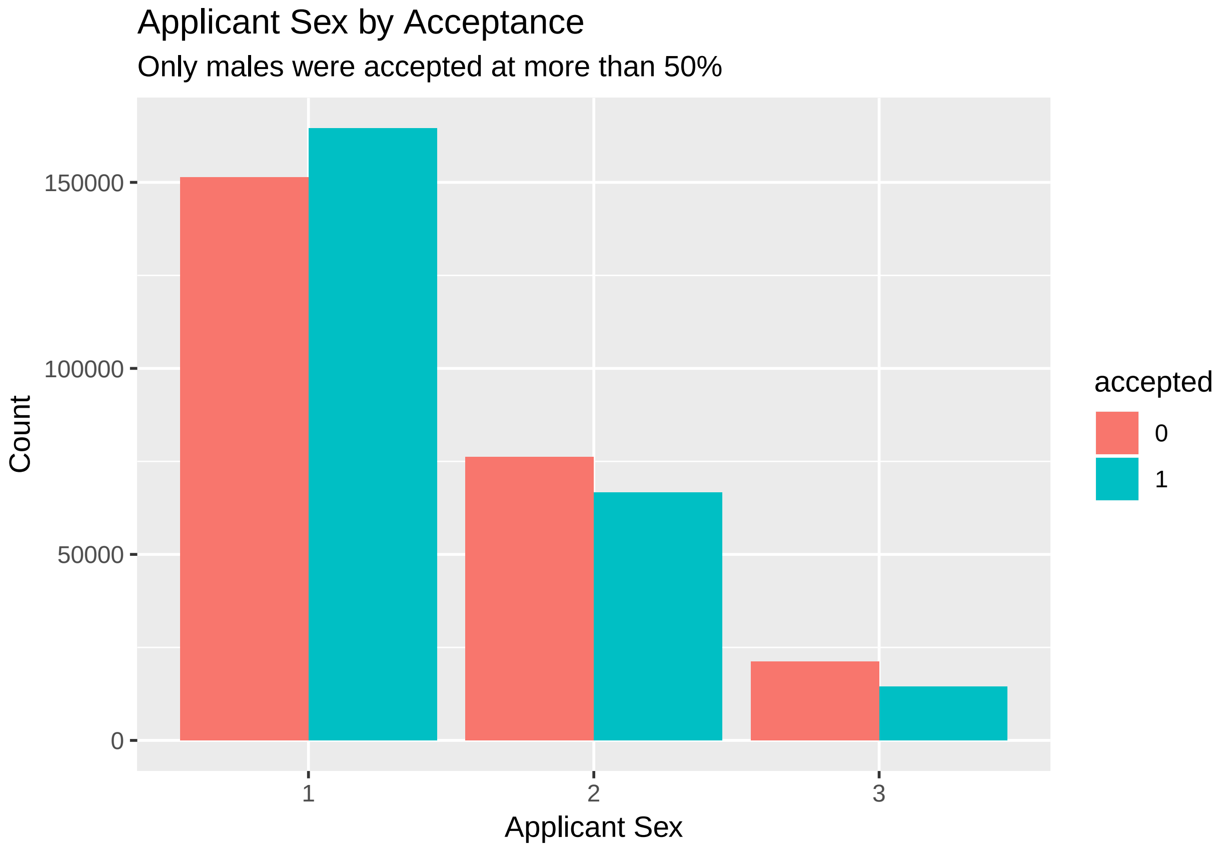


applicant_race	rate
1	0.3358726
2	0.5368070
3	0.3253982
4	0.4202651
5	0.5319164
6	0.3984164
7	0.7951744

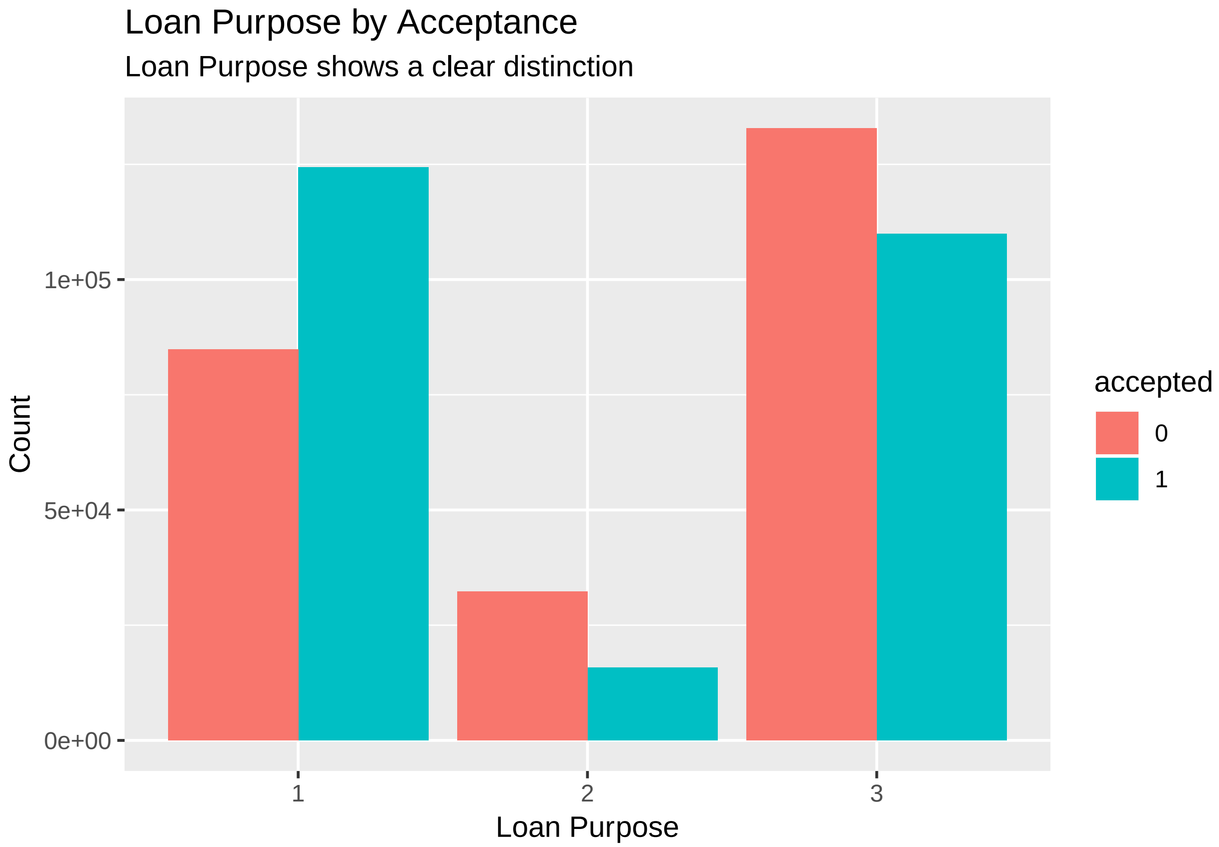
applicant_ethnicity	rate
1	0.4274133
2	0.5205188
3	0.3987399
4	0.7893109

The graphs and table show that Whites and Asians (race codes 2 and 5), as well as non-hispanics (ethnicity code 2), are accepted at a greater than 50% rate, while non-human entities (typically corporations, ethnicity code 4 and race code 7) were accepted at almost 80% rates. Minorities other than Asian were typically denied.

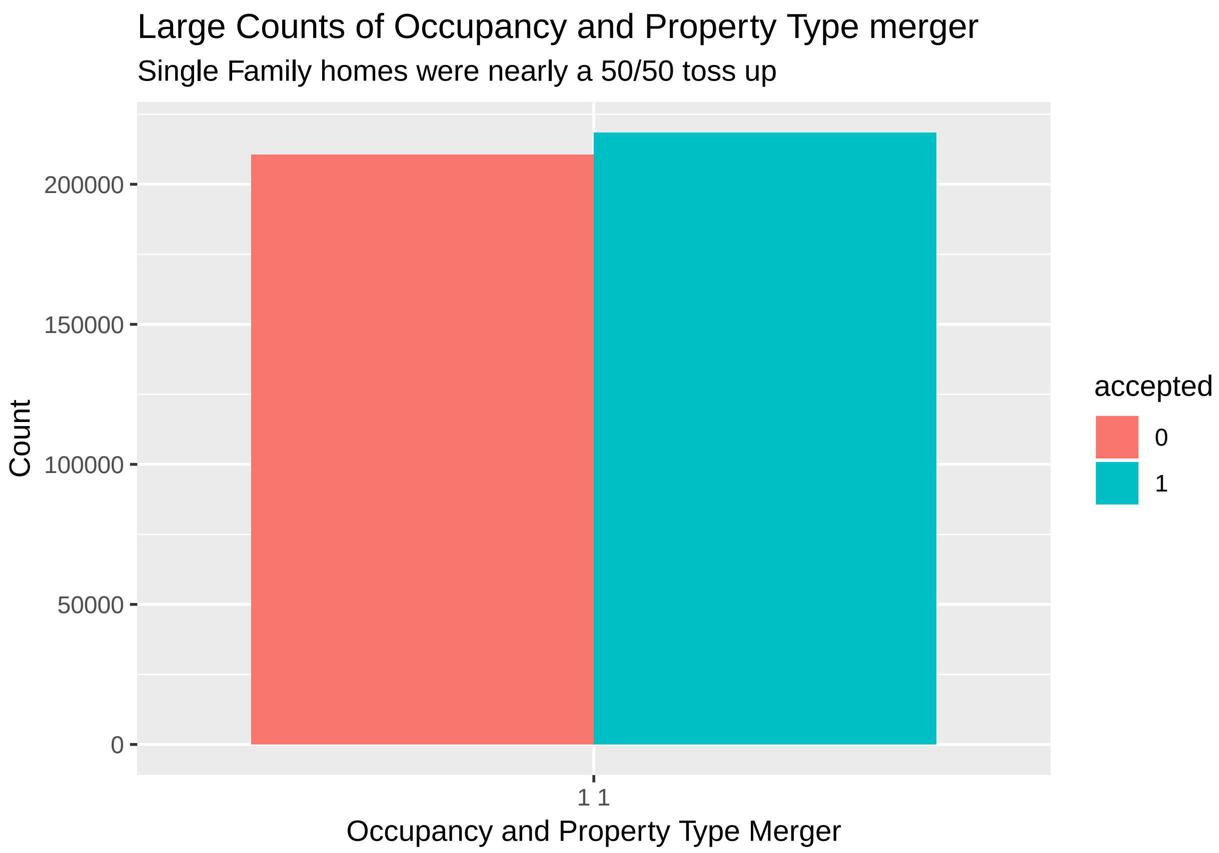
A similar trend shows up with sex. Note that in the following graph, code 4 (not applicable, typically corporations) was excluded.

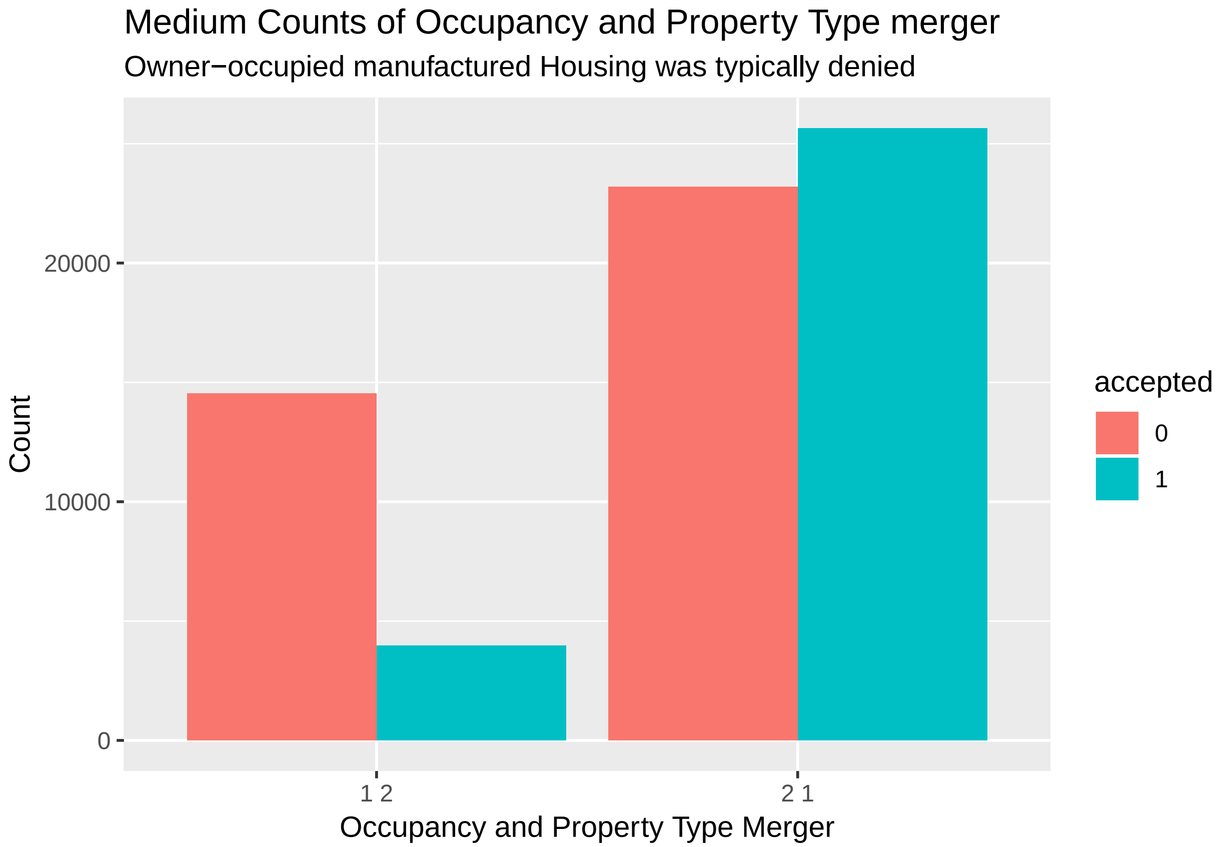


Finally, loan purpose showed a clear distiction–loans to buy a house were approved more often than not. However, loans for refinancing or home improvement were typically not approved.



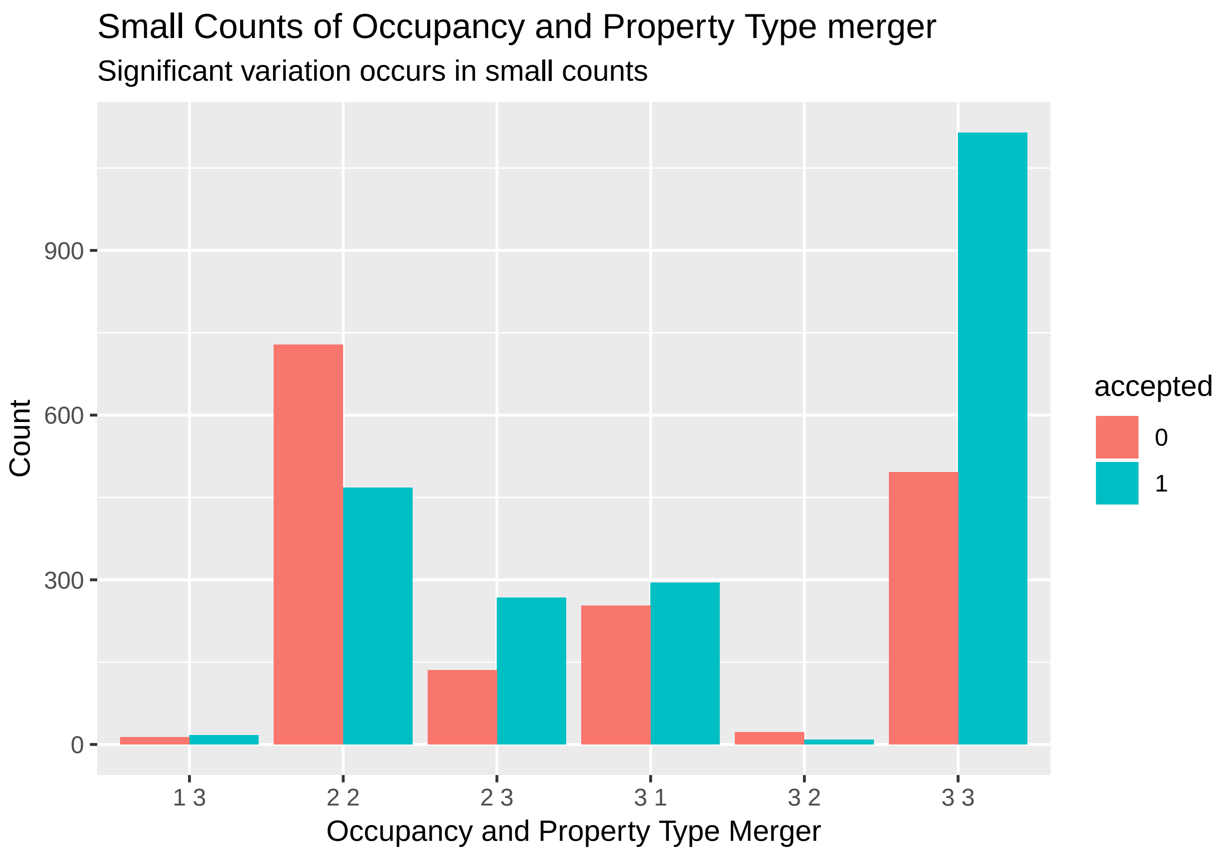
Certain combinations of other features led to distinctions. By combining property type and occupancy, a feature was created which helps clarify combinations of what the mortgage is for and who lives there. Examples of combinations include owner-occupied one family housing (i.e., single family homes), or non-owner-occupied multifamily (i.e., apartment complexes). There was significant variation both in acceptance rates and counts, so three different plots were used, as with race and ethnicity above.


Large Counts of Occupancy and Property Type merger
Single Family homes were nearly a 50/50 toss up
Count
200000
150000
100000
50000
0
1 1
Occupancy and Property Type Merger
accepted
0
1

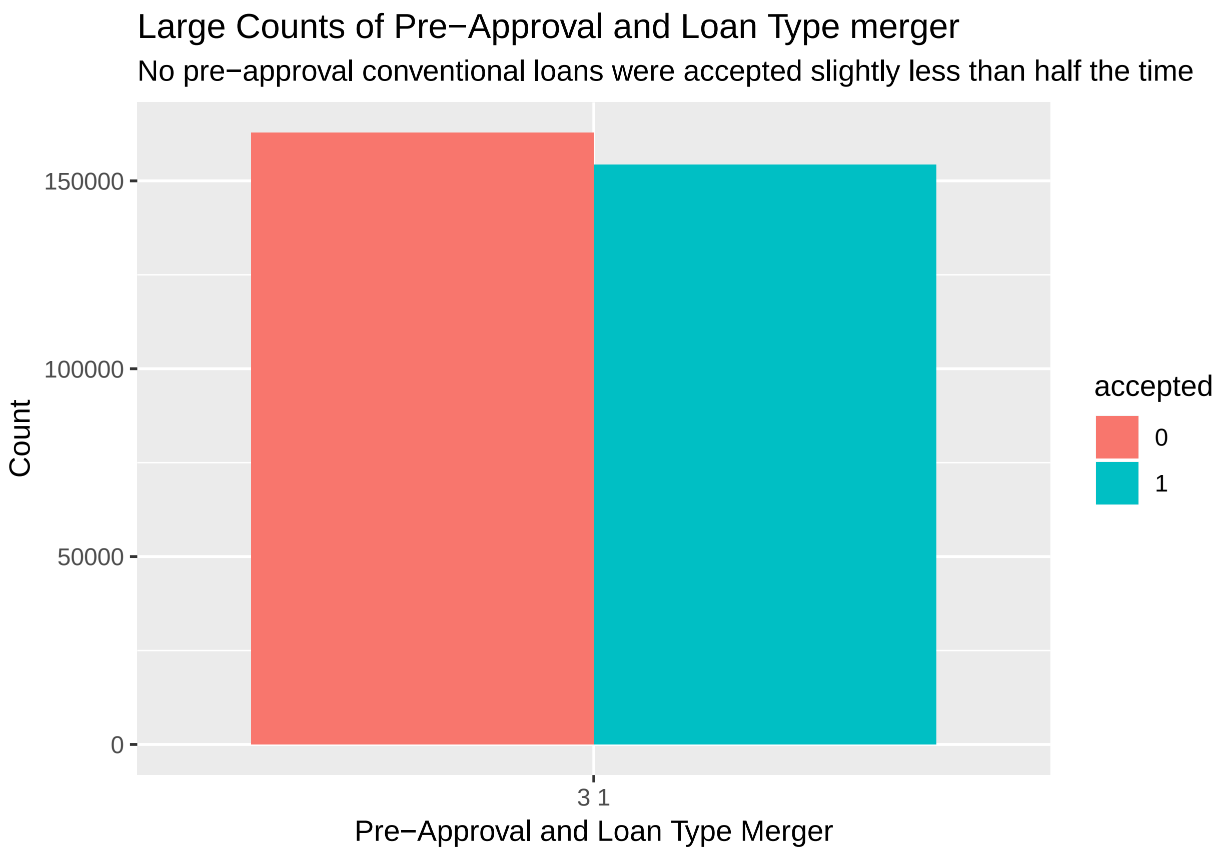


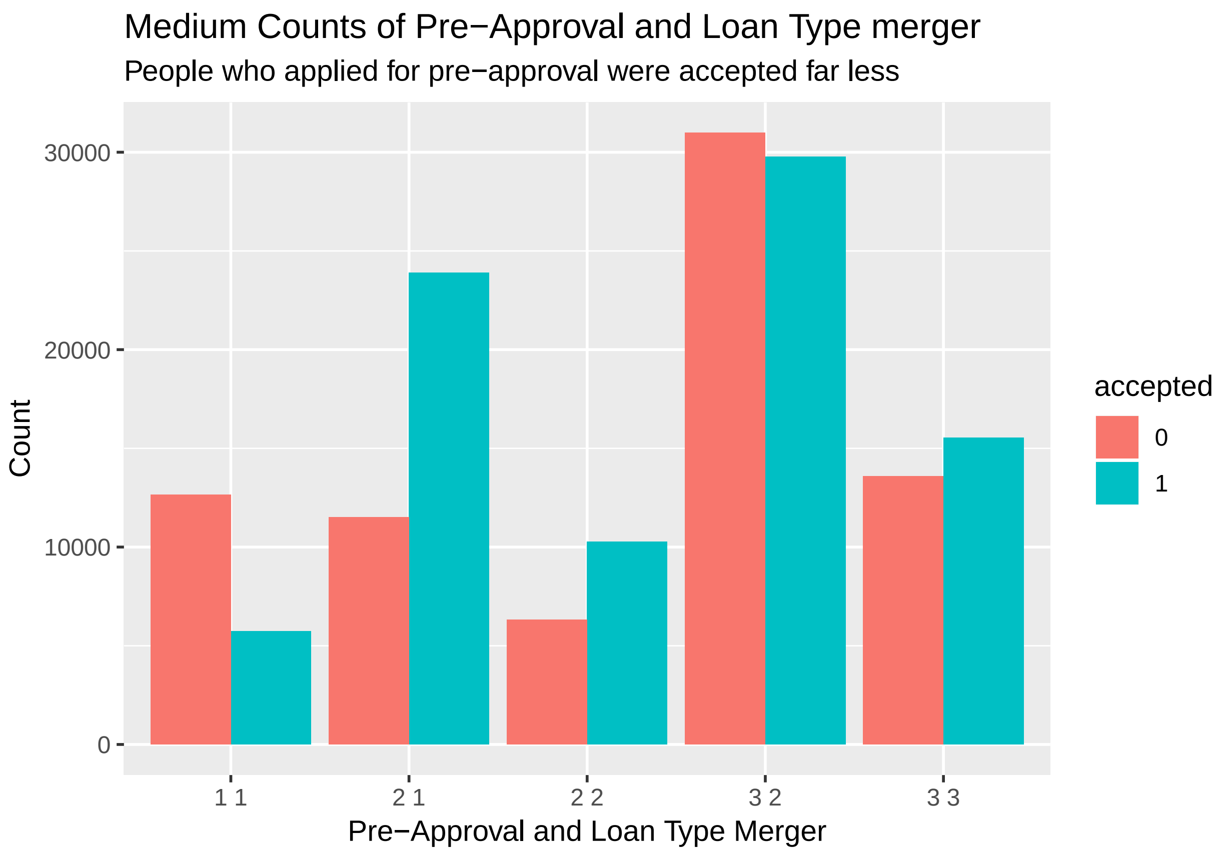
Medium Counts of Occupancy and Property Type merger

Owner−occupied manufactured Housing was typically denied

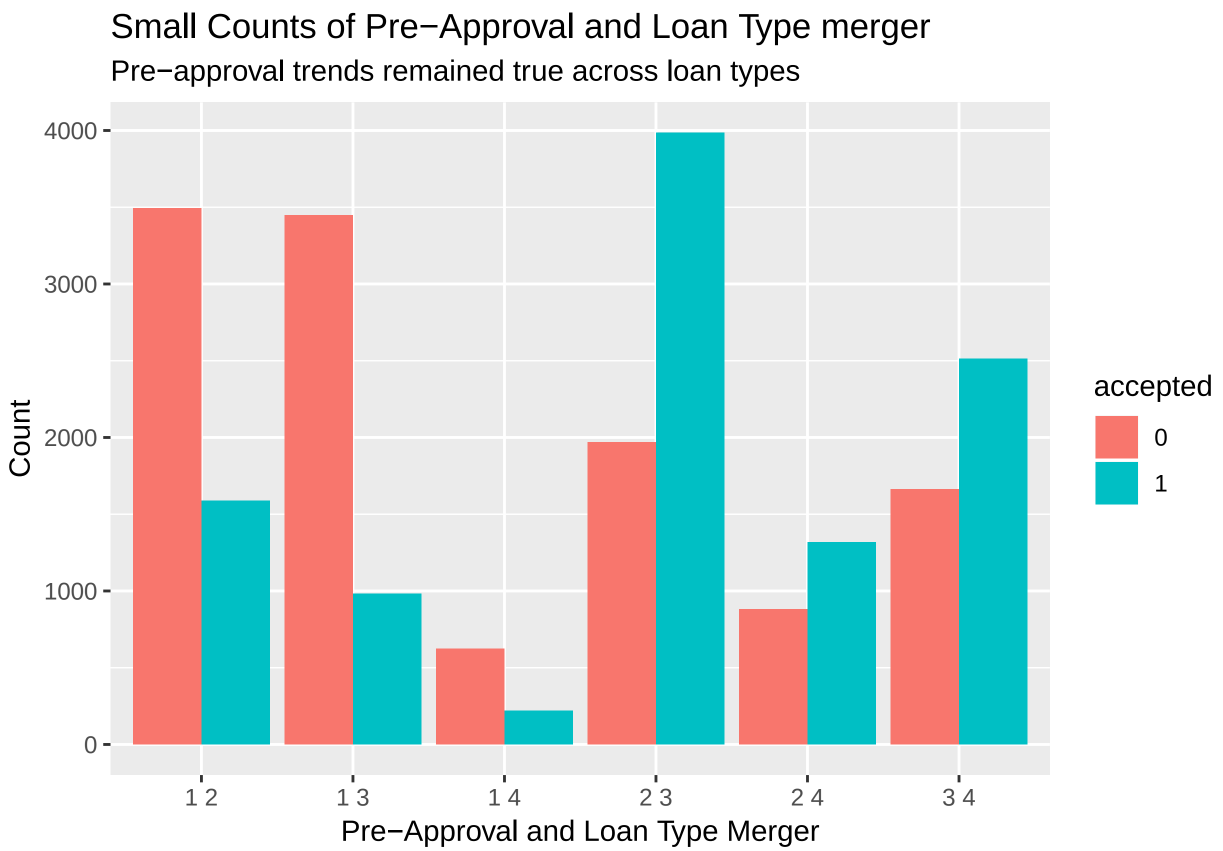


Finally, whether applicants requested pre-approval was combined with the loan type to create another feature which showed large differences.


Large Counts of Pre−Approval and Loan Type merger
No pre−approval conventional loans were accepted slightly less than half the time
Count
150000
100000
50000
0
3 1
Pre−Approval and Loan Type Merger
accepted
0
1



Medium Counts of Pre−Approval and Loan Type merger
People who applied for pre−approval were accepted far less
Count
30000
20000
10000
0
1 1
2 1
2 2
3 2
3 3
Pre−Approval and Loan Type Merger
accepted
0
1



People who applied for pre-approval (codes starting with 1) were accepted far less often in all cases than those who did not. This result is surprising, but perhaps means that those who were initiall less likely to be accepted would apply for pre-approval to help their chances.

Classification Results

Based on the analysis of the mortgage data, a series of predictive models were used to classify applications into accepted or denied. Four models were used 1. GLM - A generalized linear model 2. LDA - Linear Discriminant Analysis 3. GBM - Generalized Boosted Regression Models 4. RF - The ranger implementation of the random forest mmodel

The first two models were typically used to see if changes would be likely to increase or decrease accuracy. Both had runtimes in the range of 10-45 seconds on the training set. GBM had runtimes around 10-20 minutes, while the ranger algorithm could take up to five hours to run. Ranger was typically the most accurate, but because it was so time-intensive it was used sparingly after changes had produced positve results in the other models.

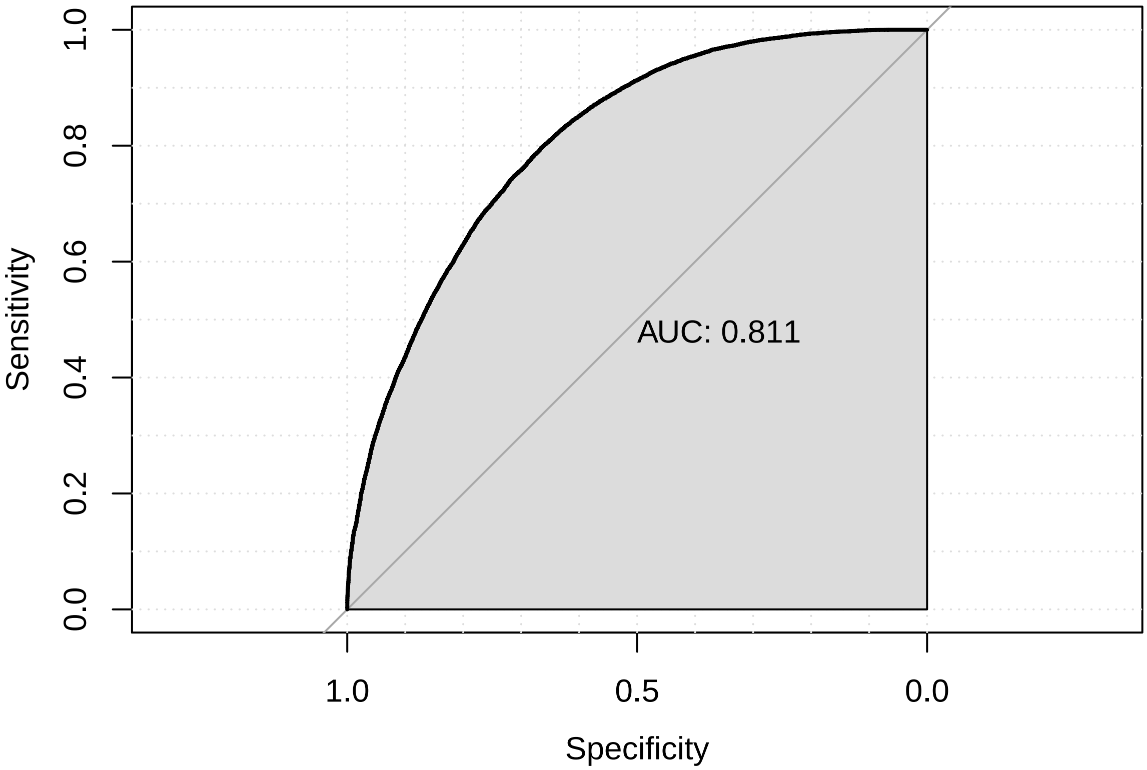
For performance reasons, some of the larger factors (MSA/MD and Lender, among others) were turned into numerical features by calculating the acceptance rate for each level. This made it possible to run the analysis on a personal laptop while still including important features.

Each model was trained using 5-fold cross validation on 90% of the training set. The model was then tested on the remaining 10%. Results above .72 were used on the test data and submitted. The most accurate model is detailed below with a confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 16464  5021
##          1  8525 19991
##
##                Accuracy : 0.7291
##                  95% CI : (0.7252, 0.733)
##     No Information Rate : 0.5002
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4581
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6588
##             Specificity : 0.7993
##          Pos Pred Value : 0.7663
##          Neg Pred Value : 0.7010
##              Prevalence : 0.4998
##          Detection Rate : 0.3293
##    Detection Prevalence : 0.4297
##       Balanced Accuracy : 0.7291
##
##        'Positive' Class : 0
##
```

This model used the ranger algorithm. The accuracy was 0.7291 (which translated to 0.7155 on the official test set), which indicates some overfitting. Attempts were made to reduce overfitting by using cross validation. The sensitivity was low, indicating a low true positive rate, while the specificity is high, indicating a low false positive rate. This suggests that the model is underestimating the chances for each applicant.

A ROC curve of the model is plotted below. The AUC is 0.811, and the F1 Score is 0.71.



```
##
## Call:
## roc.default(response = split_test$accepted, predictor = x$Y,     plot = T, auc.polygon = T, grid = T
##
## Data: x$Y in 24989 controls (split_test$accepted 0) < 25012 cases (split_test$accepted 1).
## Area under the curve: 0.8106
```

Conclusion

This analysis has shown that whether a mortgage application is accepted can be predicted from features of the application. In a troubling result, some of the features which lead to the most accuracy are features which should not be considered—race, ethnicity, and sex. Additional research needs to be done to determine if these features are indicative of something else (income, location, etc.) or whether they are used in applications (which is against the law).

Other characteristics of the application which affect approval dramatically are less troublesome. Corporations are far more likely to have their mortgage applications accepted, while the lower the income is compared to the loan amount requested, the less likely the application is to be approved.