ANALYSIS OF MORTGAGE APPROVAL FROM GOVERNMENT DATA

Maureen Oamen, July 2019.

Executive Summary

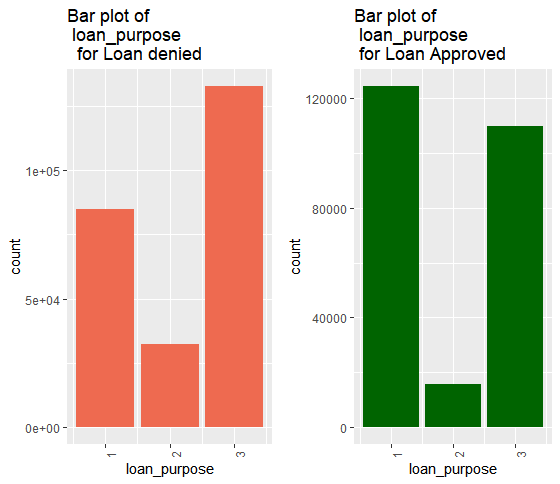
This document presents an analysis of data concerning Mortgage loans application and their approvals. The analysis is based on 500,000 observations of individual loan data, each containing applicant information, loan information, property information, census information, and loan approval status.

After exploring the data by calculating summary and descriptive statistics and by creating visualizations of the data, several potential relationships between loan approval and application were identified. A predictive model to classify loans approved or denied was created.

The following conclusions:

While many factors can help indicate if loans will be approved or denied, significant features found in this analysis in no particular order were

**Loan purpose**- A good number of individuals applying for the purpose of [1] Home purchase were approved followed by applicants who need loan for the purpose of [3] Refinancing, but just a very small number put in applications for [2] Home improvement were approved.

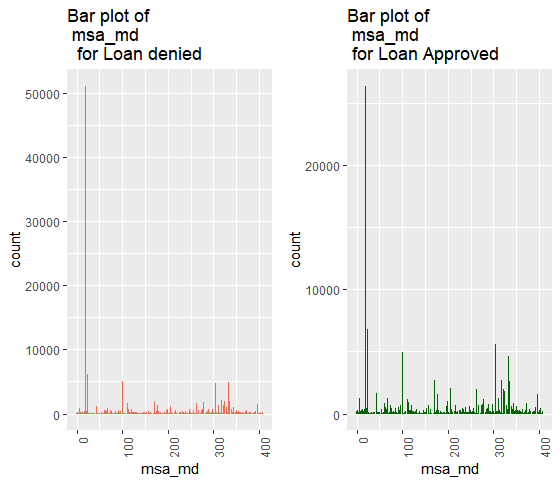
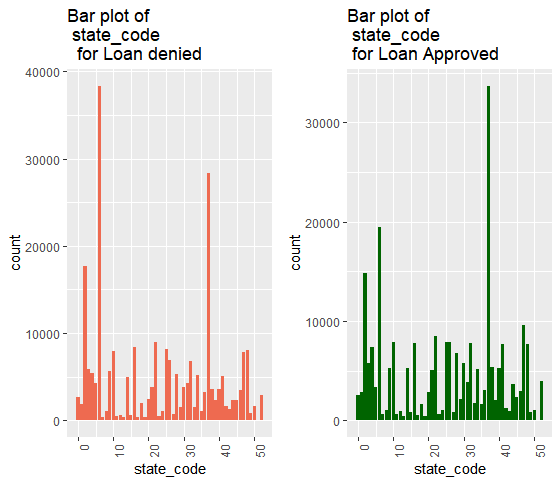
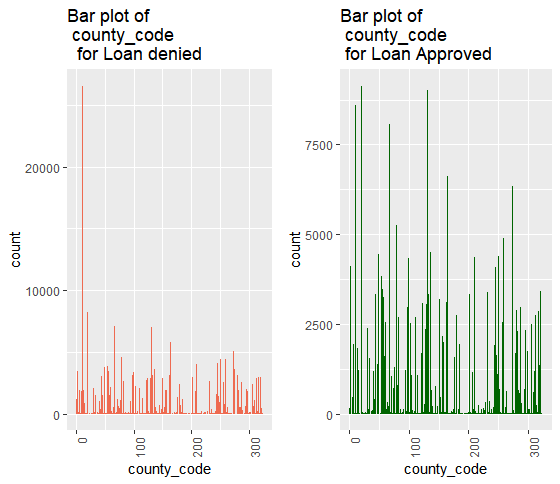


**Property location -**

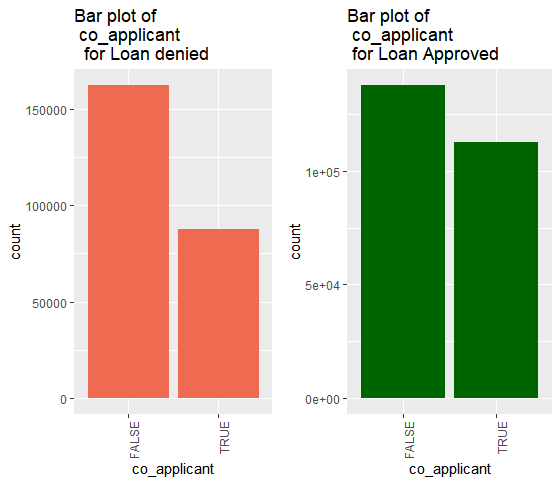
**Metropolitan Statistical Area/Metropolitan Division –** A lot of loans were approved for properties located within metropolitan division *100* – *380* with the highest number of properties coming from md-*375*

**For States**, many properties were approved in state with codes 26 – 52 with more thousands in state 37 and 41.

**County:** just a few hundred were approved within counties, with only county *131* having the greatest number of approvals



**Co applicant** – A lot of applicants who’s co applicant were their spouse who applied for loans were approved in the 10’s of thousand



Initial Data Exploration

**Individual Summary statistics**

Summary statistics for maximum, minimum, mean, median, standard deviation, and distinct

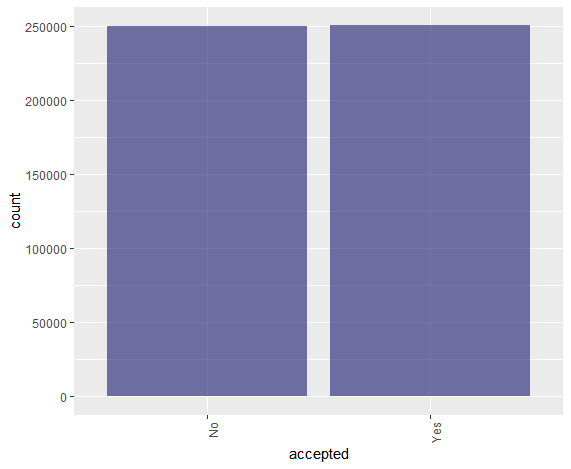
count was calculated for numeric columns and the results taken from 500,000 observations are shown here:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Features** | **Minimum** | **Maximum** | **Median** | **Mean** | **S. deviation** | **Count** |
| Applicant Income | 1 | 101390 | 74 | 100 | 144.47 | 1897 |
| Loan Amount | 1 | 100878 | 162 | 221 | 590.64 | 2997 |
| Population | 14 | 37097 | 4975 | 5397 | 2667.73 | 18202 |
| Minority population pct | 0.534 | 100.00 | 22.901 | 31.336 | 25.80 | 91923 |
| Ffiecmedian family income | 17858 | 125248 | 67526 | 69159 | 14478.23 | 68868 |
| Tract to msa-md income pct | 3.981 | 100 | 100 | 92.20 | 13.99 | 54535 |
| Number of owner-occupied\_ units | 4 | 8771 | 1327 | 1423 | 721.03 | 6088 |
| number\_of\_1\_to\_4\_family\_units | 1 | 13633 | 1753 | 1880 | 89372 | 7374 |

In addition, the mortgage approval data have 14 categories namely

* Loan type 4- Conventional, FHA, VA and FSA/RHS loan
* Property Type - One to four-family, Manufactured housing and Multifamily
* Loan purpose – Housing Purchase, Home improvement and Refinancing.
* Occupancy - principal dwelling Not owner, occupied, Not applicable
* Preapproval – Requested, No requested, not applicable
* Msa-md 408 distinct categories
* State Code 52 distinct categories
* County code 317 distinct categories
* Applicant ethnicity - Hispanic or Latino, Not Hispanic or Latino, information not provide, not applicable
* Applicant race - American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Information not provided by applicant in mail, Internet, or telephone application, Not applicable
* Applicant sex – Male, Female, not provided, not applicable
* Lender 6111 distinct loan providers
* Co applicant TRUE/FALSE (If spouse is a co applicant)
* Accepted loan approved or not

A Bar chart was created to show the frequency distribution of the loan approved or denied below, since it the category of interest



Where number of loans approved (250,114) are almost not significantly higher than the number of unapproved loans (249,886)

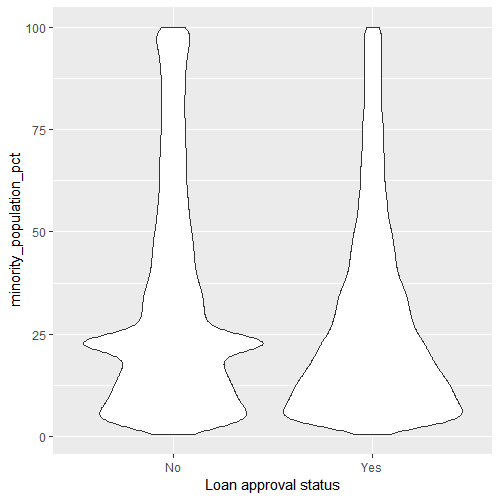
* Conventional loans are mostly applied for, about two time more than all the other three type of loans put together
* For property type, about 80% loans are for the One to four-family property dwelling compared to a manufactured housing or multifamily dwelling
* A relative high number of loans were directed to housing Purchase followed by Refinancing and just a small number of loans are for housing improvement
* Properties to which loan applies show that at least 88% of applicants occupies property as principal dwelling followed by not principal owner occupying property and the not applicable category is way less than 2%
* A good number of loans pre-approvals were not applicable to applicant followed by a small number of no Requested, then requested category for preapprovals
* Statistics for property location are distributed across metropolitan areas, states and counties
* Non-Hispano or Latino ethnicity loan applicant is far more than the combination of the Hispano or Latino, information not provided and not applicable ethnicity who apply for loan
* The White race apply for loan more than other race, followed by the information not provider, African American, Asian, not applicable then the Indians.
* Males apply the most followed by Female gender, information not provided then the not applicable category
* The lender with id’s 6240, 5710, 3354 have processed at least 20, 000 loan application while the remaining loan applications were processed by the remaining 6109 lending authority.
* For co applicants a good number of then were not spouse.

Apparent Relationships among feature

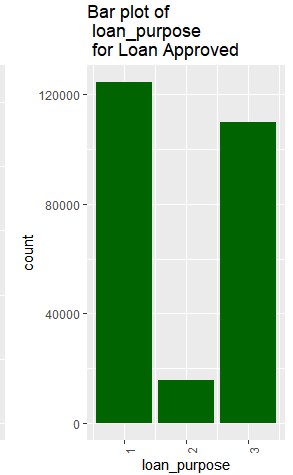
Relationships with Numeric Features

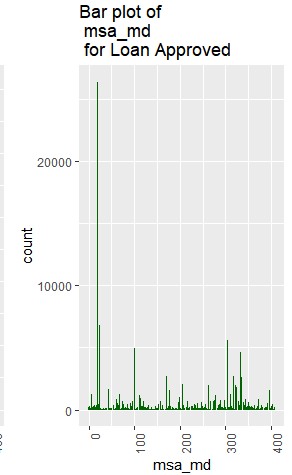
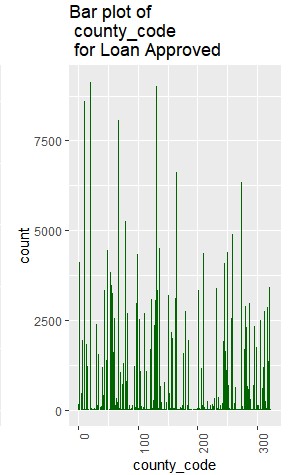
The only apparent relationship between loan application approved with numeric features are that with the percent of tract median family income compared to MSA/MD median family income and Percentage of minority population to total population for tract,

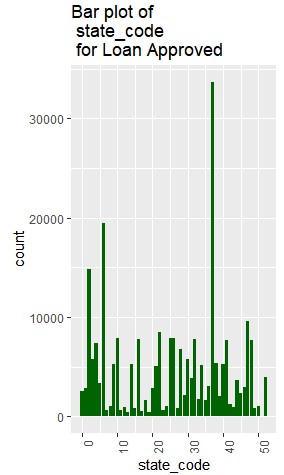
Where a good number of loan applicant whose percent of tract median family income compared to MSA/MD median family income fall between *90-10* tract were approved, also a good number of applicants who were part of the percentage of minority population to total population for tract *1-13* were approved

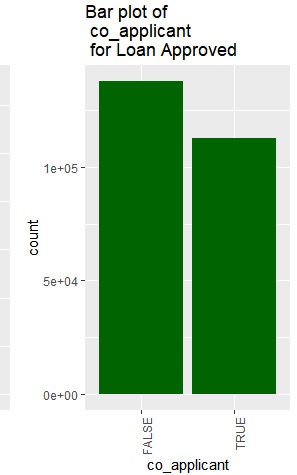


Relationships with categorical Features



Property Location





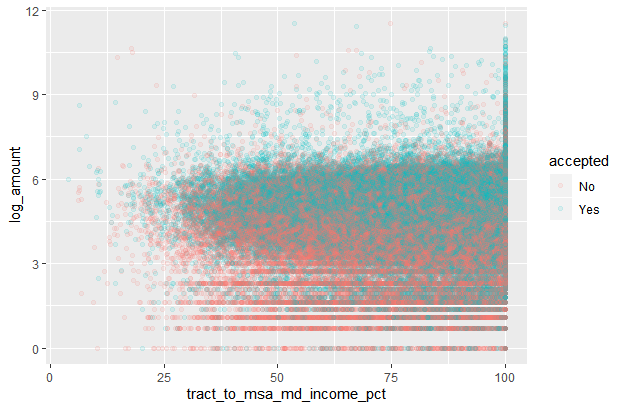
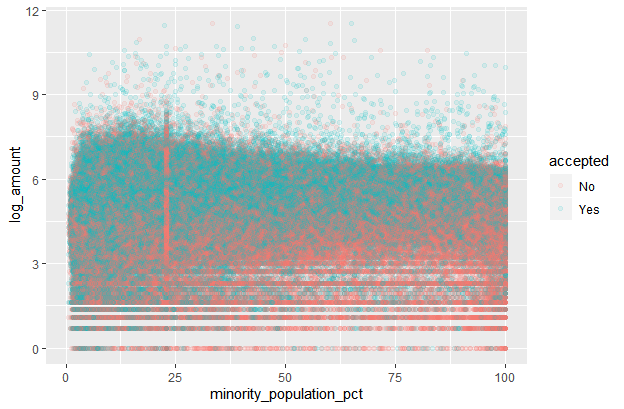
Multi-faceted Relationships

Apparent relationships between loan application accepted and individual features are helpful in determining predictive information of a model. However, relationships are often more complex, and may only become apparent when multiple features are considered in combination with one another. To help identify these more complex relationships, some faceted plots were created.

Logically one would think loan amount, applicant income would have some significant or predictive power in determining if loan application will be accepted or rejected but that wasn’t the case, It is shown that when loan amount is considered with other features the following plots revealed some interesting aspect that a large number of loan approved were for applicants who earn above the average income, a logical explanation would be these applicants had financial capacity of some sort to pay back the loan

Also considering the loan amount with the percent of tract median family income compared to MSA/MD median family income and percent of minority population to total population for tract, the greatest number of loans approved were for little above the average loan amount to the highest loan amount.



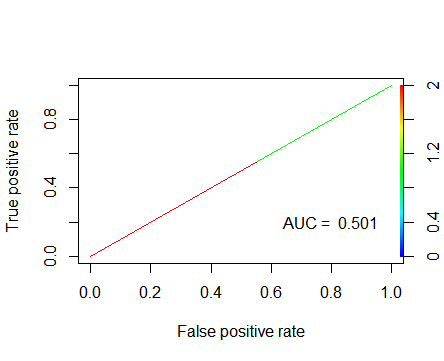
Classification of mortgage loan Based on their approval status

Based on the analysis of the mortgage loans approval data, a predictive model to classify loan application into two categories: Approved(loan application accepted) and Denied (loan denied).

The model was created using the Two-Class Boosted Decision Trees algorithm and trained with 18% of the data. Testing the model with the remaining % of the data yielded the following results:

* True Positives: 139145
* True Negatives: 110969
* False Positives: 138770
* False Negatives: 111116

The Received Operator Characteristic (ROC) curve for the model is shown here, with the diagonal line showing the expected results of a random guess:



The above plot translates in to the following standard performance metrics for classification:

* Accuracy: 73.1%
* Precision: 50.1%
* Recall: 55.6%
* F1 Score: 52.7%

Conclusion

This analysis has shown that an applicant loan approval status can be confidently predicted from applicant information, loan information, property information, census information. In particular, the important features would be loan purpose, co applicant application, where the property is located; metropolitan areas, state and county have a significant effect on the if a loan will be approved. The only secondary features important are applicant income and loan amount further revealed that loan approval will depend on how much an applicant earns compare with how much the applicant wants to borrow.