# Mortgage Classification using CatBoostclassifier

bokola

27 June, 2019

## Table of Contents

Executive Summary ..... 1
Methodology ..... 1
Description of Data ..... 2
Data Recoding and Visualization ..... 2
Data Recoding ..... 2
Data Visualization ..... 2
Feature Engineering ..... 11
Model Building ..... 11
Results and Conclusions ..... 12

## Executive Summary

The objective was Predicting Mortgage Approvals from Government Data, taking into consideration the **demographics, location, property type, lender**, and other factors to predict whether a mortgage application was accepted or denied. It is a classification problem - A classifier is a machine learning model that separates the **label** into categories or **classes**. In other words, classification models are **supervised** machine learning models which predict a categorical label. We apply relevant skills to classify the label **`accepted`** - whether a mortgage application was accepted or declined using data obtained across the United States.

We used a CatBoostclassifier model with 1000 iterations, 8 cross-validations returning an accuracy of 0.72.

## Methodology

The section outlines data manipulation procedures, visualizations and model selection.

## Description of Data

Data consisted of 23 variables (21 possible features and 1 binary label(`accepted`) and 1 arbitrary identifier(`row_id`). Both training and test sets spanned 500000 records

There were a number of variables in the train and test sets with missing values: applicant_income - 39948, population - 22465, minority_population_pct - 22466, ffiecmedian_family_income - 22440, tract_to_msa_md_income_pct - 22514, number_of_owner.occupied_units - 22565, number_of_1_to_4_family_units - 22530. The missing entries were recoded to -999 for numeric variables.

## Data Recoding and Visualization

Variable recoding and visualization are key steps that ensure every variable is of the desired class/type. Visualizaation is key to understand the distribution and separation of values by the two levels of a binary label.

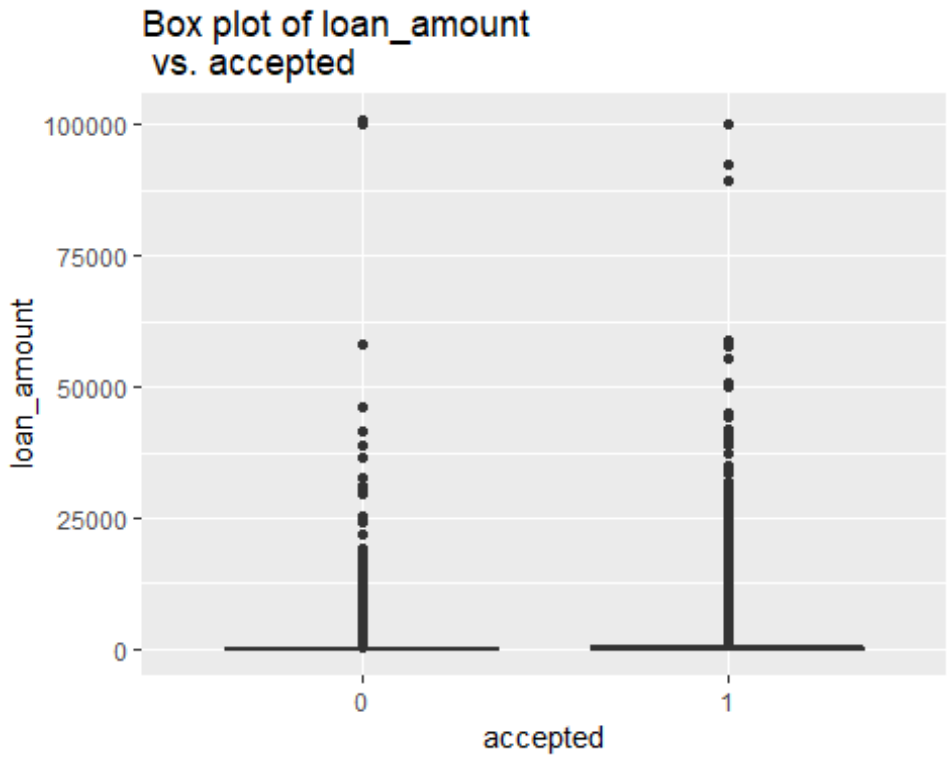
### Data Recoding

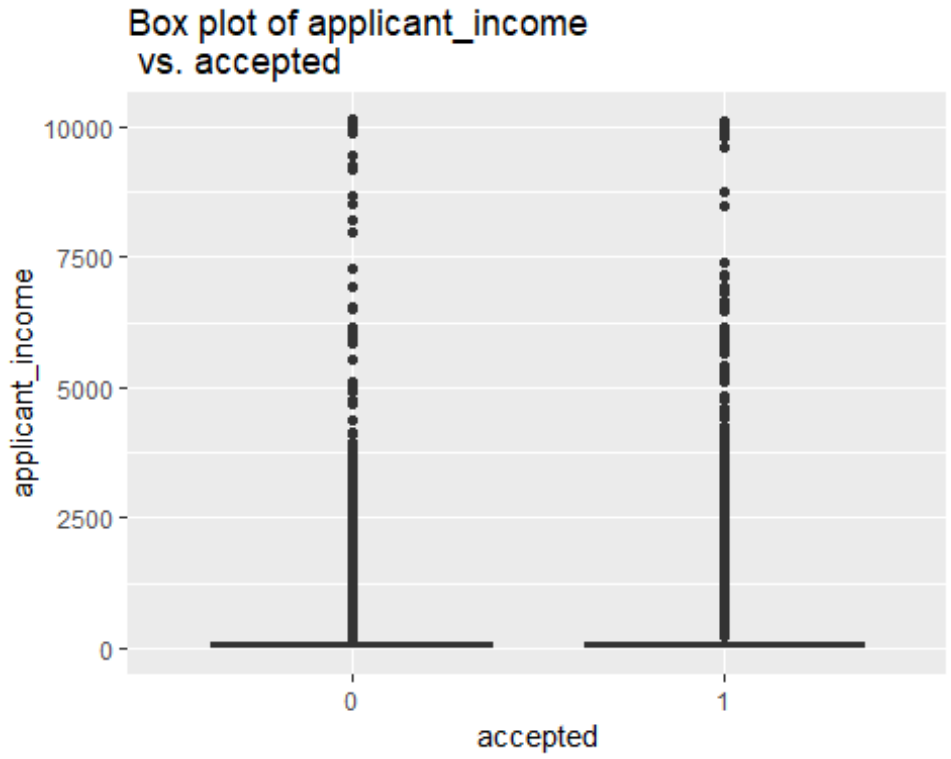
Our data had a number of categorical featuers (`msa_md`, `state_code`, `county_code`,`lender`,`loan_type`,`property_type`, `loan_purpose`, `occupancy`,`preapproval`,`applicant_ethnicity`,`applicant_race`, `applicant_sex`) but which are read in as numeric. We began by recoding such variables into the desired character class and explicitly supplying the categories in both train and test sets. We then visualize the separation of values between different levels of the label - `accepted(0,1)`.

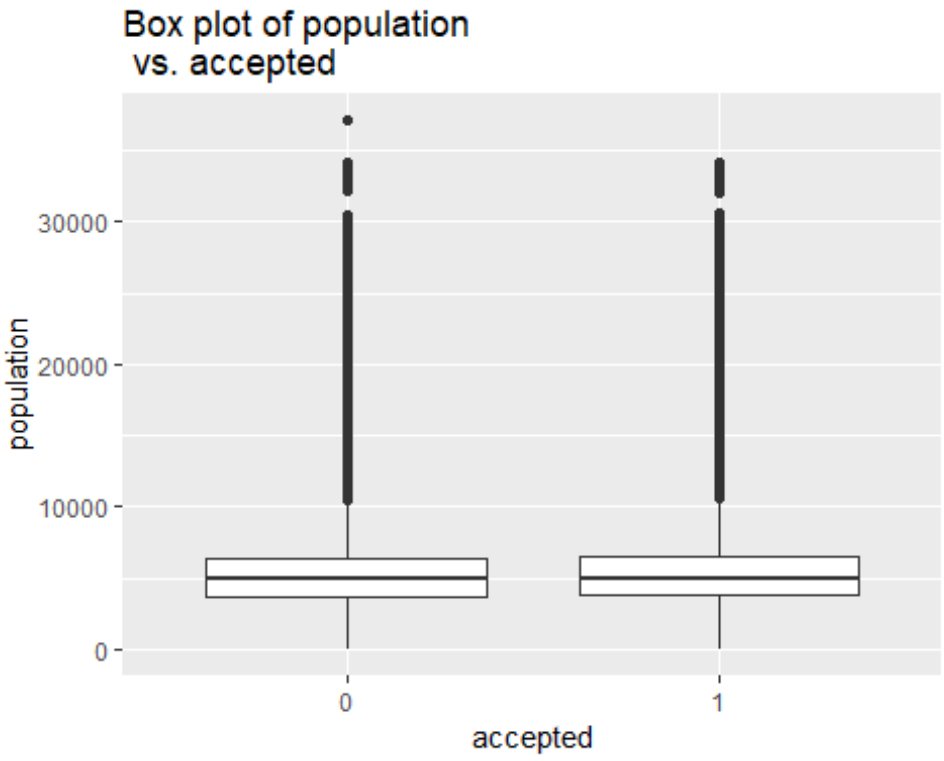
### Data Visualization

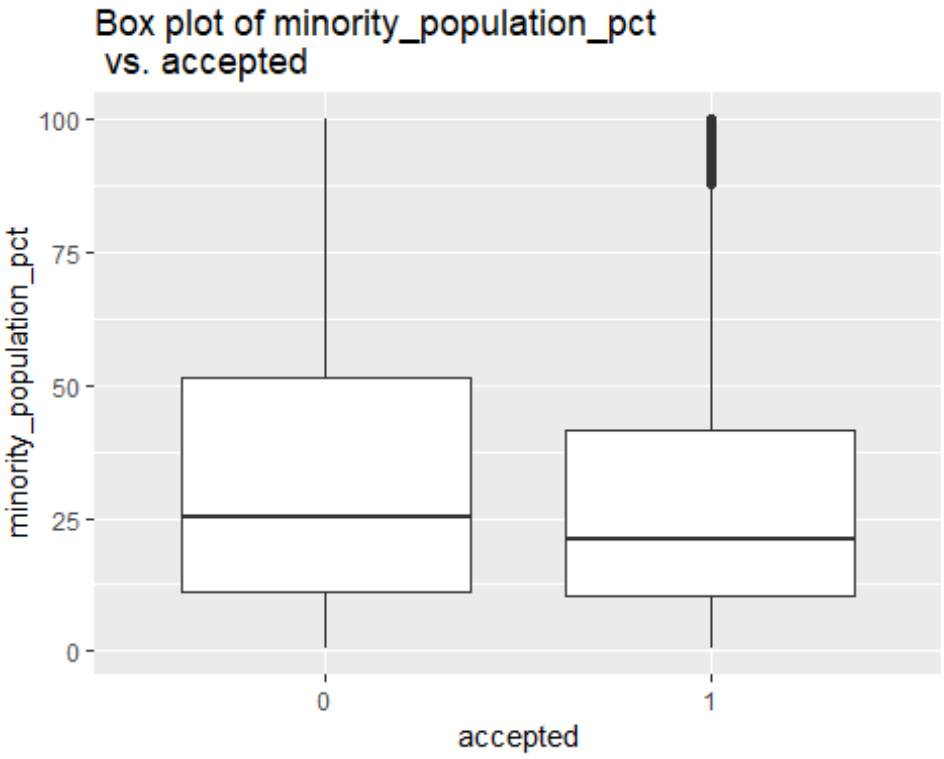
1. **Class separation by numeric variables**

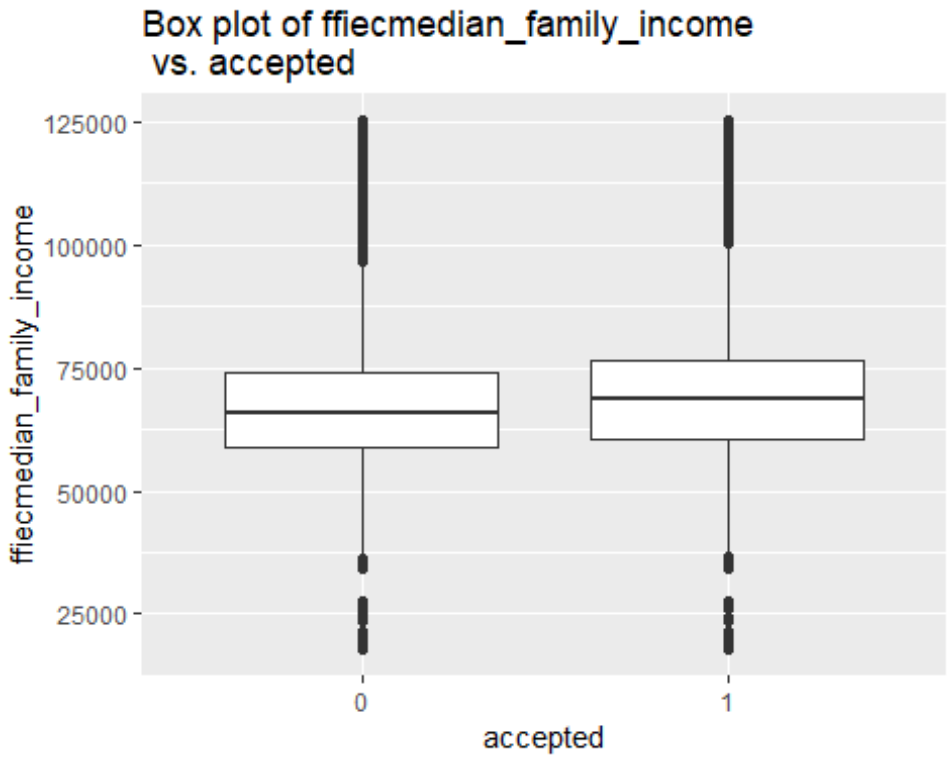
We explored the distinction of the distribution of values between the two levels of the label through boxplots.

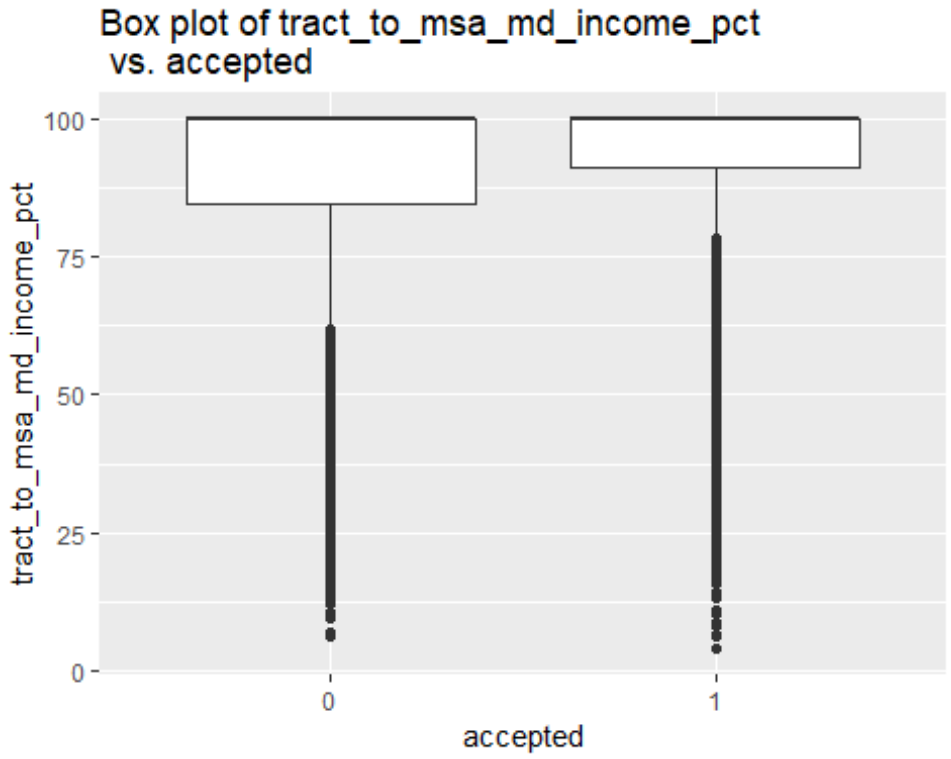

Box plot of loan_amount
vs. accepted
100000
75000
50000
25000
0
loan_amount
0
1
accepted

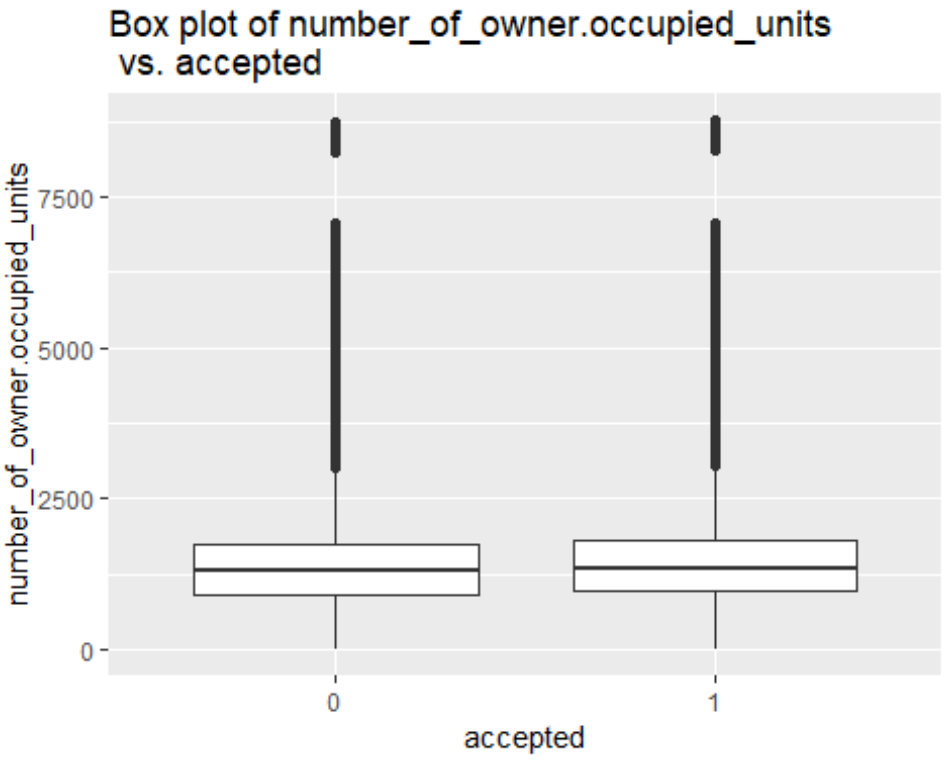

Box plot of applicant_income
vs. accepted
10000
7500
5000
2500
0
applicant_income
0
1
accepted

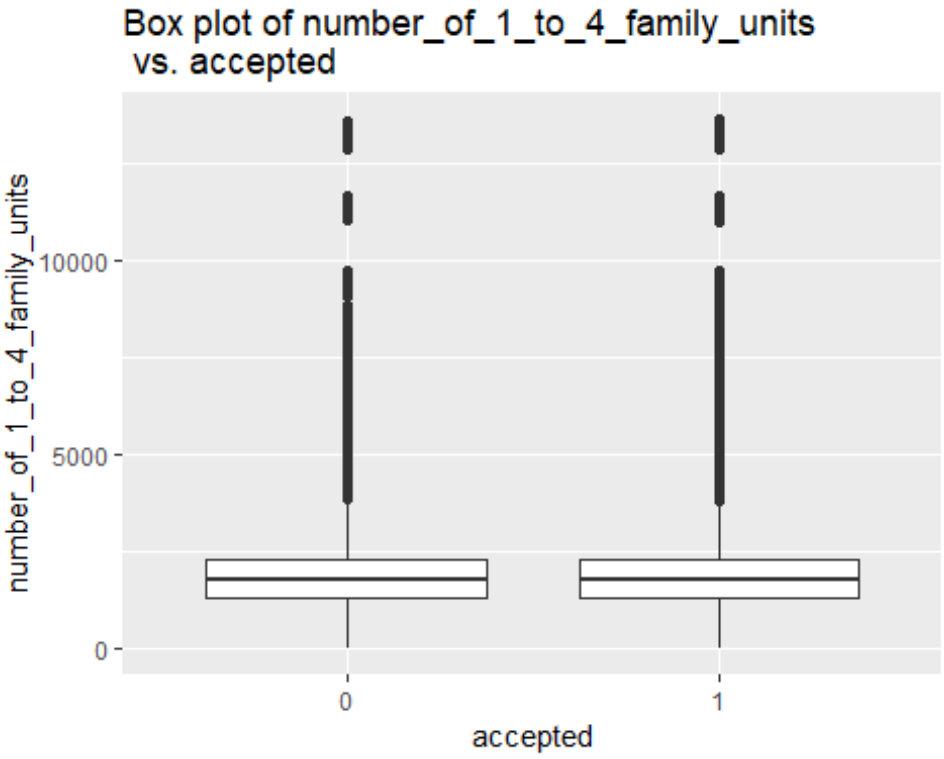

Box plot of population
vs. accepted
population
30000
20000
10000
0
0
1
accepted


Box plot of minority_population_pct
vs. accepted
minority_population_pct
100
75
50
25
0
0
1
accepted


Box plot of ffiecmedian_family_income
vs. accepted
ffiecmedian_family_income
125000
100000
75000
50000
25000
0
1
accepted


Box plot of tract_to_msa_md_income_pct
vs. accepted
tract_to_msa_md_income_pct
100
75
50
25
0
0
1
accepted

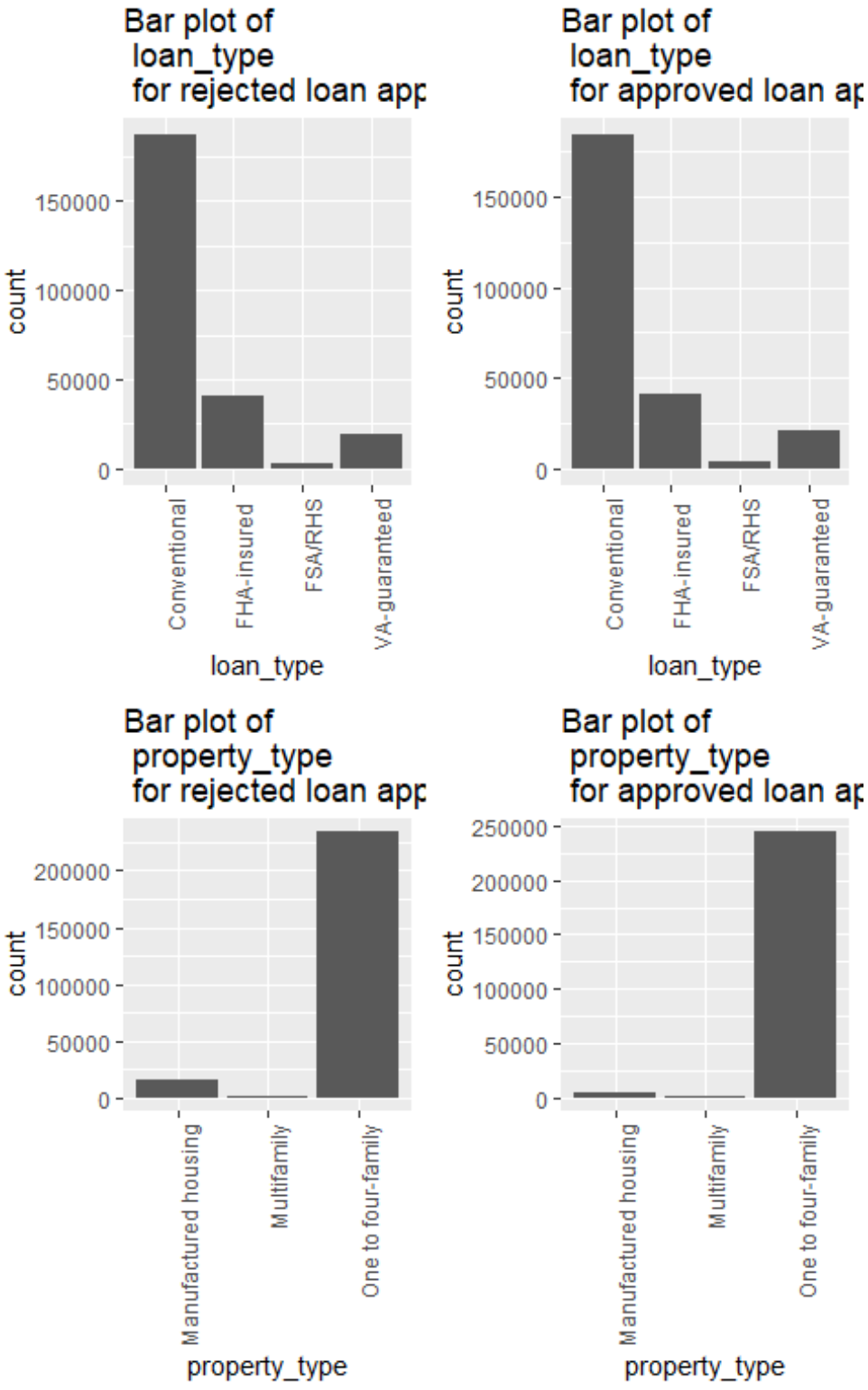


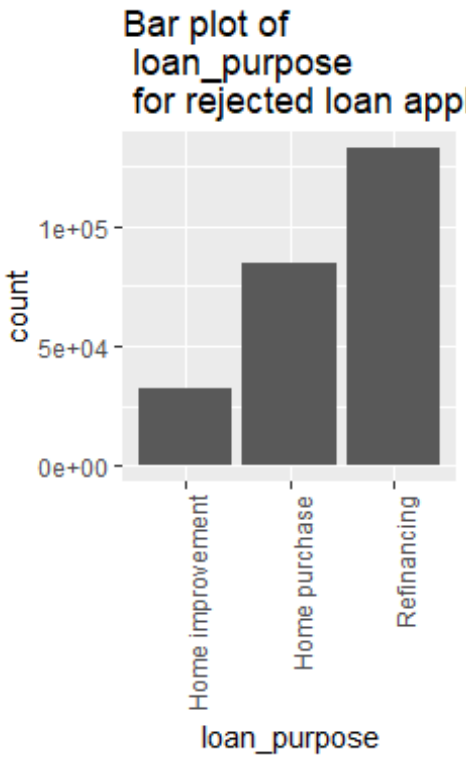



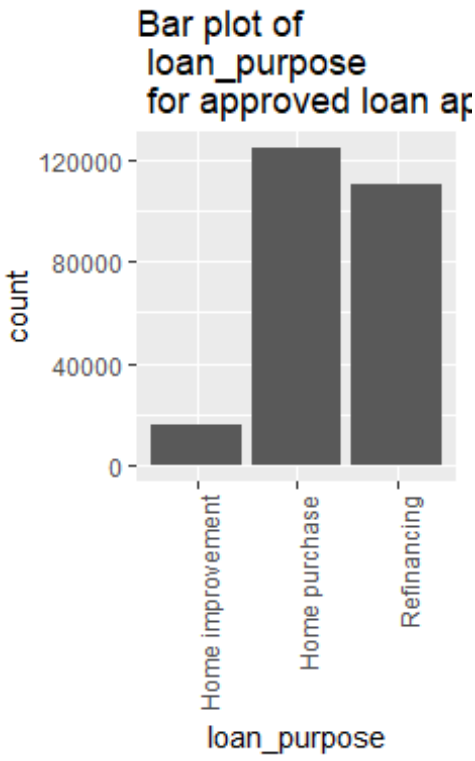

From the plots, three variables - loan_amount, minority_population_pct and tract_to_msa_md_income_pct show clear separation of values between the two levels. Next we explore this separation by categorical variables.

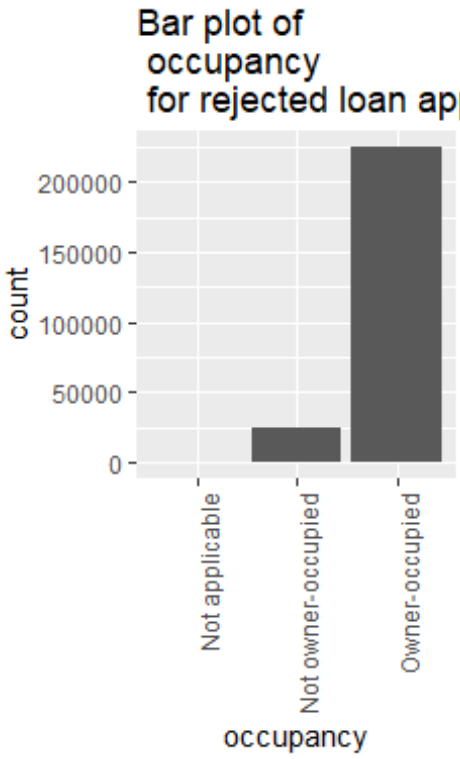
1. **Class separation by categorical variables**

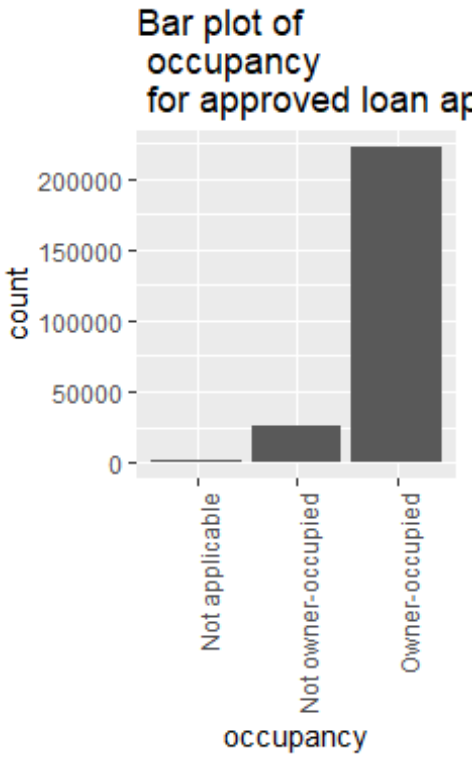
We explore the distinction of the distribution of values between the two levels of the label through bar plots.

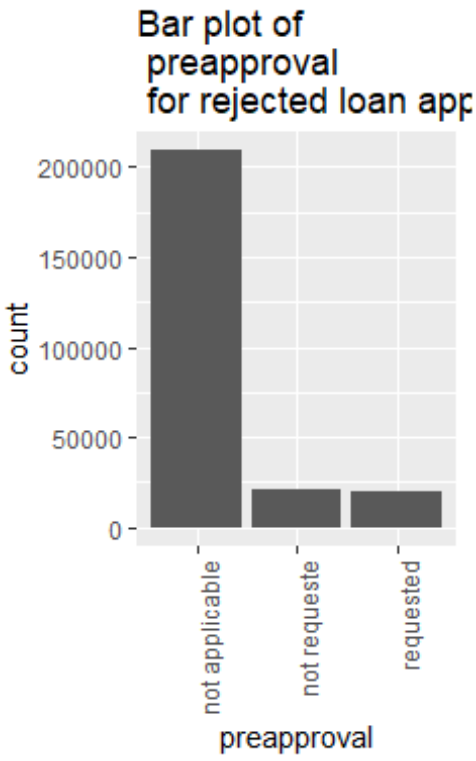


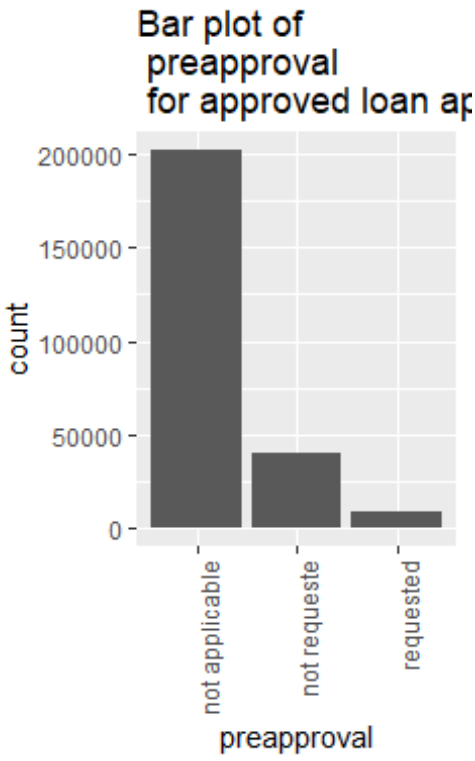

Bar plot of
loan_purpose
for rejected loan appl
count
1e+05
5e+04
0e+00
Home improvement
Home purchase
Refinancing
loan_purpose

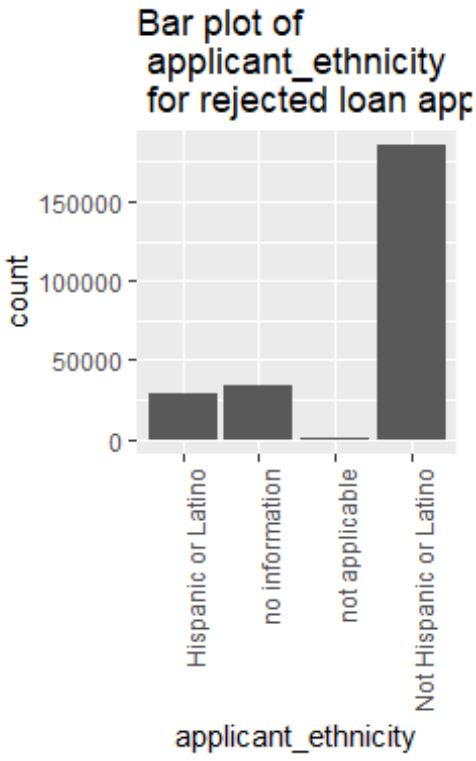

Bar plot of
loan_purpose
for approved loan ap
count
120000
80000
40000
0
Home improvement
Home purchase
Refinancing
loan_purpose

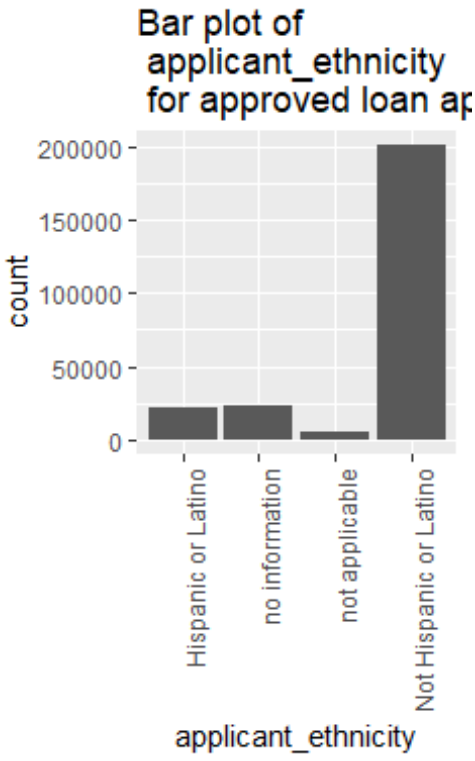

Bar plot of
occupancy
for rejected loan app
count
200000
150000
100000
50000
0
Not applicable
Not owner-occupied
Owner-occupied
occupancy

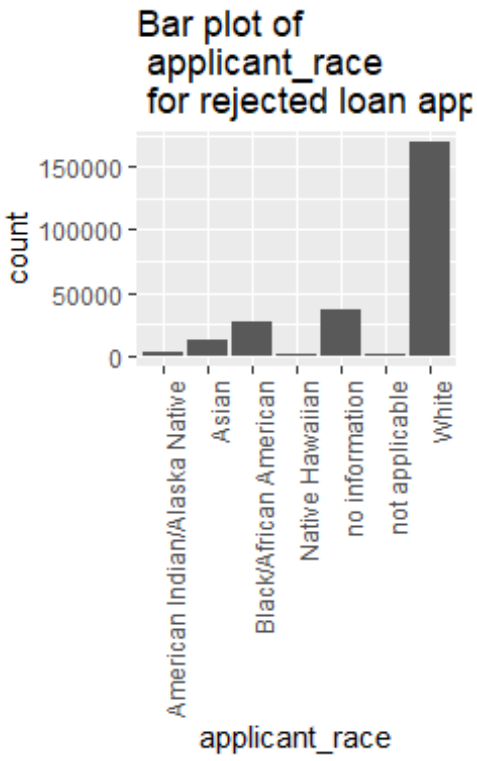

Bar plot of
occupancy
for approved loan ap
count
200000
150000
100000
50000
0
Not applicable
Not owner-occupied
Owner-occupied
occupancy

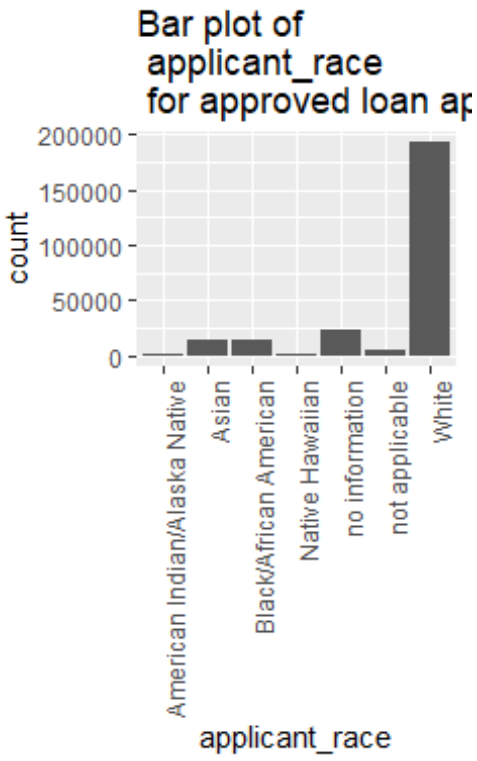

Bar plot of
preapproval
for rejected loan app
count
200000
150000
100000
50000
0
not applicable
not requeste
requested
preapproval

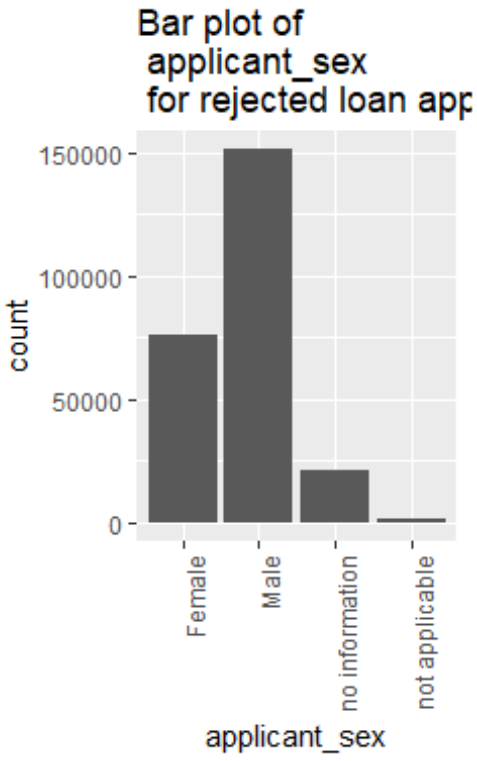

Bar plot of
preapproval
for approved loan ap
count
200000
150000
100000
50000
0
not applicable
not requeste
requested
preapproval

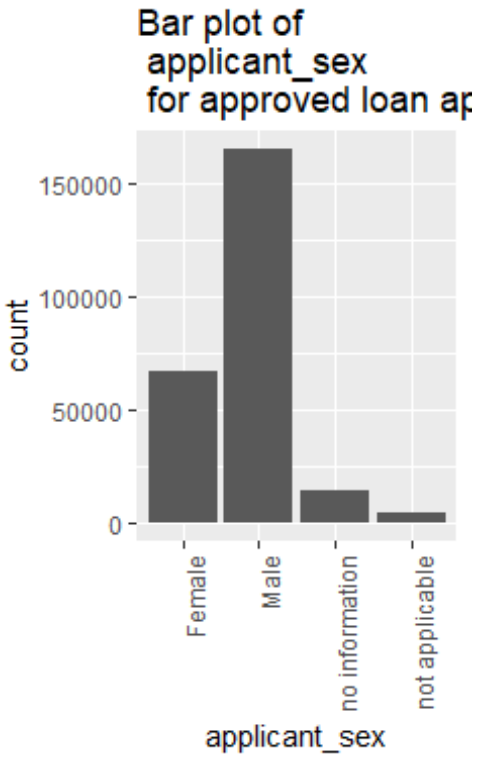

Bar plot of
applicant_ethnicity
for rejected loan app
count
150000
100000
50000
0
Hispanic or Latino
no information
not applicable
Not Hispanic or Latino
applicant_ethnicity

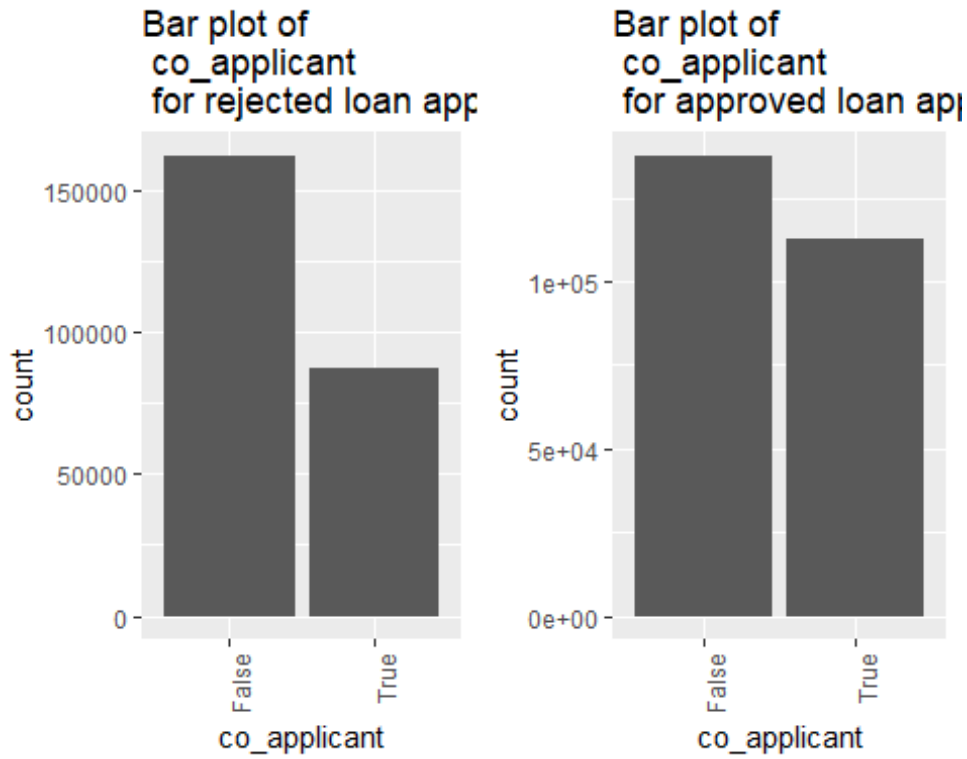

Bar plot of
applicant_ethnicity
for approved loan ap
count
200000
150000
100000
50000
0
Hispanic or Latino
no information
not applicable
Not Hispanic or Latino
applicant_ethnicity


Bar plot of
applicant_race
for rejected loan app
count
150000
100000
50000
0
American Indian/Alaska Native
Asian
Black/African American
Native Hawaiian
no information
not applicable
White
applicant_race


Bar plot of
applicant_race
for approved loan ap
count
200000
150000
100000
50000
0
American Indian/Alaska Native
Asian
Black/African American
Native Hawaiian
no information
not applicable
White
applicant_race


Bar plot of
applicant_sex
for rejected loan app
count
150000
100000
50000
0
Female
Male
no information
not applicable
applicant_sex


Bar plot of
applicant_sex
for approved loan ap
count
150000
100000
50000
0
Female
Male
no information
not applicable
applicant_sex




We observe the following:

1. Some features such as property_type, loan_purpose, applicant_ethnicity, applicant_race and co_applicant have an almost significant different distribution of categories between the label categories.

2. Others features such as preapproval, applicant_sex, occupancy show small differences, but these differences are unlikely to be significant.

Next we do feature engineering in readiness for modelling.

## Feature Engineering

Data preparation is an important step. We recoded loan_amount and applicant_income to categorical strings which helped in distinguishing levels within the dataset and imputed missing numeric variables with means.

Feature engineering and modelling were done in python

## Model Building

Modelling was done with python, selecting the CatBoostclassifier library for the classification. As earlier indicated we settled on 8 cross-validations with 1000 iterations to minimize bias and variance of the model.

## Results and Conclusions

The model was able to correctly classify 72% of the accepted values  with AUC of 0.8. There were discrepancies with the supplied test values indicating that our model was overfitting. However, it offers a good starting point to build on and improve.