Analysis of Mortgage Approvals from US Government Data

Joeri Kiekens (Username: 30036766), April 2019

Executive Summary

This document presents an analysis of mortgage data performed on 500 000 observations. Each data point contained characteristics of the loan applicant, the lender, the property, and the census. The analysis will show meaningful relationships between the data and mortgage approval. This will be shown by using summary statistics and visualizations. The constructed model's accuracy on the test set was 72.53 %.

After performing the analysis, the author presents below the most significant features that affected mortgage approvals and recommends putting effort into these:

- Lender: The institution giving out the mortgage. Some lenders tended to have a higher approval rate than others. Out of all the features, this one was the most impacting.
- Co-applicant: Another person signing the mortgage with the applicant, in most cases the spouse. Having a co-applicant sign off too, increases chances of receiving the mortgage.
- Skipper: Skipping information on the application form greatly reduced chances of getting accepted.
- Company: Applying as a company (where ethnicity, race, sex are n/a) greatly increased chances of acceptance.

Contrary to popular belief, the following features were of less importance:

- Applicant income: Having a high income did not necessarily increase acceptance.
- Loan amount: A high/low loan amount was not significant for predicting loan acceptance.
- Loan type: Only for home improvement did a VA-guaranteed loan have a higher acceptance rate than the other types.

Disclaimer: Note that from the data, there is unfortunately a relationship between race, sex, ethnicity and mortgage approval. The constructed model is based on this data and to reach a high accuracy on new incoming data, the model had to be slightly biased. In the industry, ethically unjust data such as race, sex, ethnicity among others should be left out when deploying models. The purpose of this project however was not to deploy a model, rather 1) to create a model with optimal accuracy on the test data and 2) create a report of the project for the staff of the course/fellow students to show techniques used in constructing ML models.

Initial Data Exploration

Categorical Features

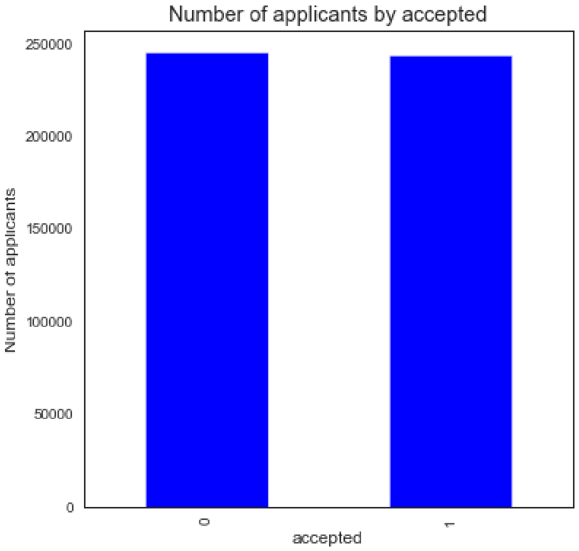
- **Loan Type**: Conventional, government-guaranteed, government-insured
- **Property Type**: One-to-four-family dwelling (other than manufactured housing), manufactured housing, multifamily dwelling

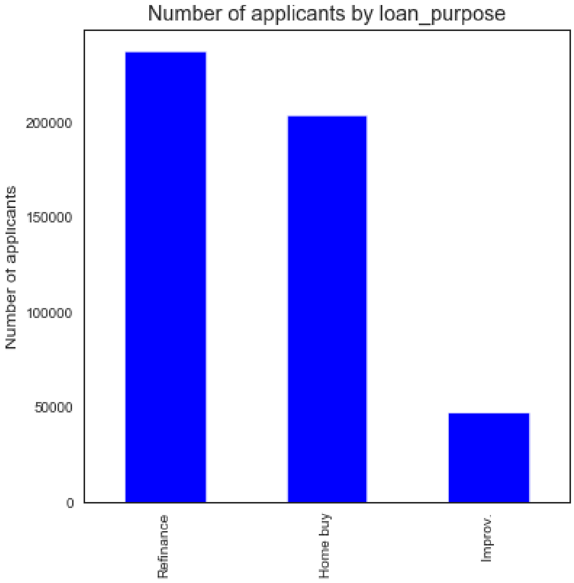
- **Loan Purpose**: Home purchase, home improvement, refinancing
- **Occupancy**: Owner's principal dwelling, not owner-occupied, not applicable
- **Preapproval**: Preapproval was requested, not requested, not applicable
- **MSA md**: Metropolitan Statistical Area/Metropolitan Division
- **State code**: US State
- **County code**: County
- **Applicant**: Ethnicity (1 ‘Hispanic or Latino’, 2 ‘Not Hispanic or Latino’, 3 ‘Not provided’, 4 ‘Not applicable’, race (1 ‘American Indian or Alaska Native’, 2 ‘Asian’, 3 ‘African American’, 4 ‘Native Hawaiian or Other Pacific Islander’, 5 ‘White’, 6 ‘Not provided’, 7 ‘Not Applicable’), sex (1 ‘Male’, 2 ‘Female’, 3 ‘Not provided’, 4 ‘Not applicable’)
- **Lender**: 6508 different lending institutions, each with a unique code
- **Accepted**: Our target/label (0 = application not accepted, 1 = application accepted)

Each of the categorical features was explored with bar plots and cross tables to visualize their frequency. Some key observations from this exploration:

- Most applications were of the conventional loan type (74 %).
- Most common property type was the one-to-four family dwellings (96 %).
- The loan purpose was mainly refinancing (49 %). See below figure for the Loan Purpose distribution.
- 89 % of applicants would be the owner’s principal dwelling.
- Applicant ethnicity was primarily ‘non Hispanic/Latino’ (77 %), most common race was ‘White’ (72 %), and the most common sex was ‘male’ (63 %). In the ethnicity and race features, combinations of similar distributions were attempted, but resulted in a lower accuracy of the model, thus were eventually left as is. This unfortunately shows the dependence of race on mortgage acceptance rate.
- 40 % of applicants signed with a co-applicant.

Two bar plots are given below. The plot on the left hand side shows that there is no class bias in the provided dataset: half of the applications were accepted and half were denied. The plot on the right hand side shows the number of applicants per loan purpose.





Numerical Features

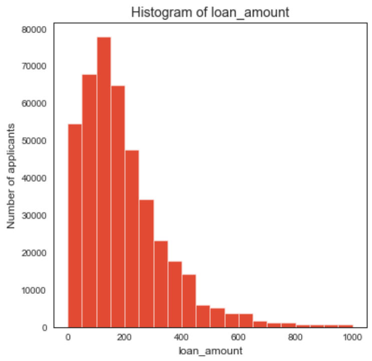
The following summary table contains the numerical features with their descriptive statistics.

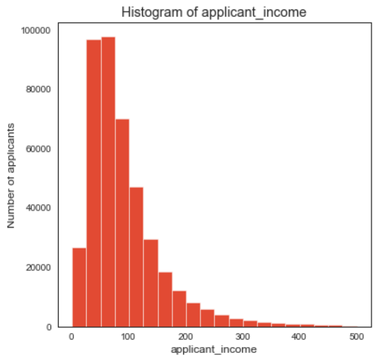
Columns	mean	stand. dev.	minimum	median	maximum	# entries
Loan Amount (k$)	222	591	1	162	100878	500000
Applicant Income (k$)	102	154	1	74	10139	460052
Population in tract	5417	2728	14	4975	37097	477535
MSA median family income ($)	69236	14810	17858	67526	125248	477560
Tract median/MSA median (%)	92	14	4	100	100	477486
Occupied living units in tract	1428	738	4	1327	8771	477435
1-4 family units in tract	1886	914	1	1753	13623	477470
Minority population in tract (%)	32	26	1	23	100	477534

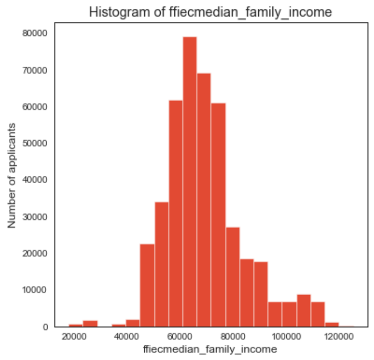
The numerical features were explored with histograms. Some key observations from this exploration:

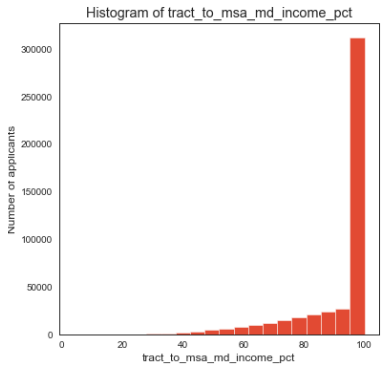
- Many right skew features were present: the mean was for most features higher than the median. Especially loan amount, applicant income, number of 1-4 family units in tract, number of occupied units, and population.
- The most obvious of these were applicants with high income and high loan amount. As outlier values greatly affect model accuracy, applicants requesting a very high loan amount (> 500 k$) or who have a very high income (> 400 k$) were dropped from the training data. Trial and error showed these cut-off values to be optimal for model accuracy.

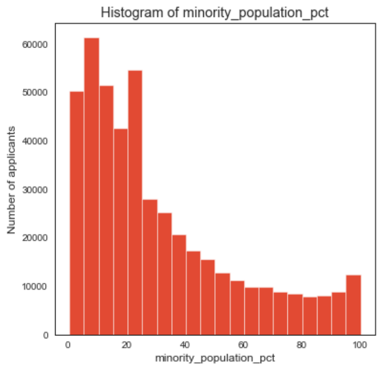
Below are some of the histograms of the raw data features, given with a conveniently chosen x range.









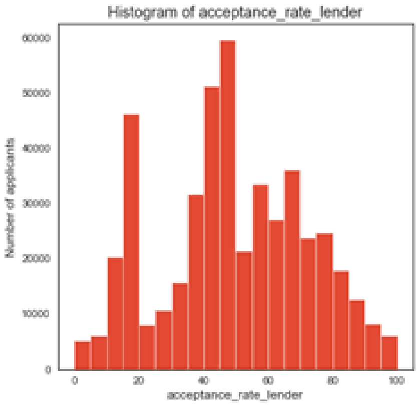


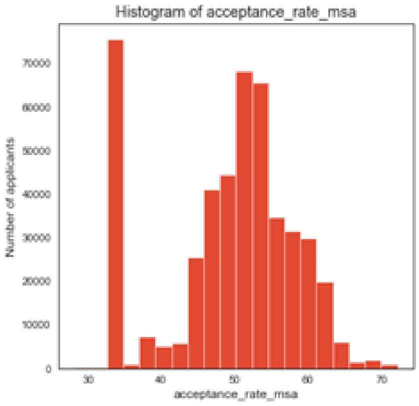
Additional features constructed through feature engineering

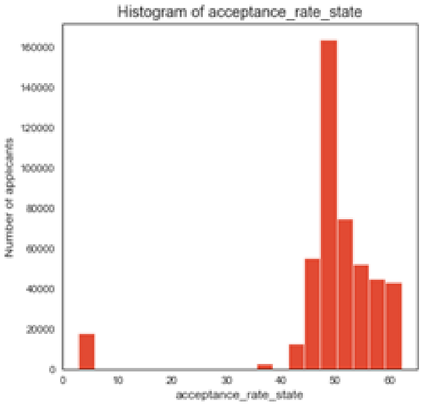
In addition to the numerical and categorical features given to us in the data, additional features were constructed. These are constructed below and some of their distributions are shown.

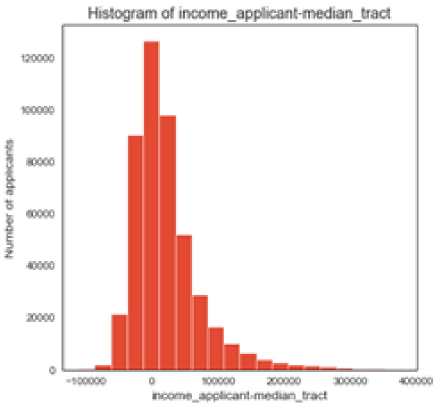
- Acceptance rate per lender: Some lenders might have a more rigorous mortgage lending system in place, making it harder to receive a mortgage at these lenders. This feature was become through an acceptance rate (%) per lender using a groupby.
- Acceptance rate per msa.
- Acceptance rate per state.
- Applicant income – median income tract (=f(applicant income, median family income for the MSA/MD in which the tract is located, tract to MSA/MD median family income): A relation between the applicant's income and the median income of the tract. Positive values mean you earn more than the median.
- Skippers: A 0/1 indication for applicants who skipped filling in the population of the tract, median income of tract and the other numerical columns had a very low acceptance rate (4.5 % of total applicants were skippers and only 4 % of the 'skippers' received a mortgage).
- Company: A 0/1 indication for applicants theorized to be companies. Their ethnicity, race, and sex were all filled in as 'not applicable'. 80 % of company applications were accepted. Note that these only accounted for 1.1 % of all applications.

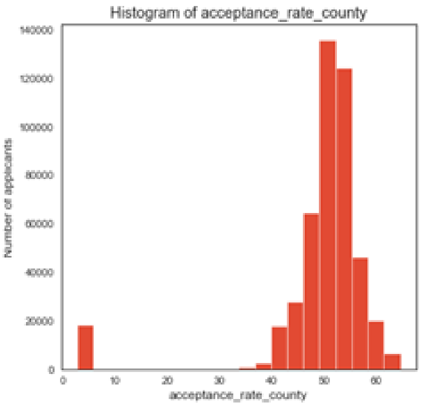
Below are histograms of all the engineered numerical features.











Data Exploration through comparison with the label

Categorical features vs. label (accepted).

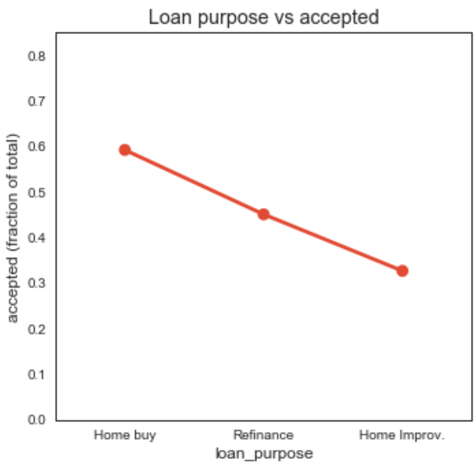
For each feature, it’s relation to the label was investigated and visualized. Visualizations of the categorical features were constructed using point plots. From these visualizations (see page 6) the following observations were made. (The y-axis shows the fraction of the total 500 000 applications. Having no influence on the acceptance rate would mean 0.5 on the y-axis. High on the y-axis means high acceptance rate of applications with the accompanying label and vice versa.)

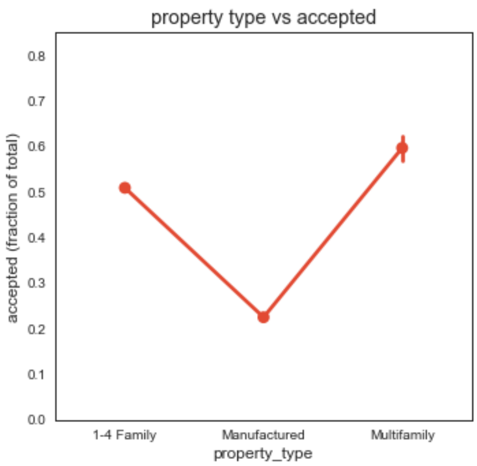
- The acceptance rate for a loan with the purpose of buying a home is much higher than for refinancing and home improvements. Furthermore, only 33 % of applications with the purpose of home improvement were accepted.
- The acceptance rate for manufactured housing in Property Type is significantly lower than 0.5 and for multifamily properties significantly higher than 0.5.
- A process without requesting preapproval was more often accepted. A loan application with a preapproval request was more often shot down.
- Applications where ethnicity, race, and sex were ‘not applicable’, had an 80 % acceptance rate. The author theorizes that these applications are made as a company, but cannot be derived from the given data.
- From the sex categories, the acceptance rate for ‘Male’ was a little higher than average (0.5).
- Having a co-applicant sign with the applicant had a higher rate of acceptance than not having one.
- Applicants skipping several columns had a significant lower acceptance rate than applicants filling in all the columns.
- Applicants applying as a company had an 80 % acceptance rate. Note that these only accounted for 1.1 % of all applications, but nearly all the ‘n/a’ from sex.

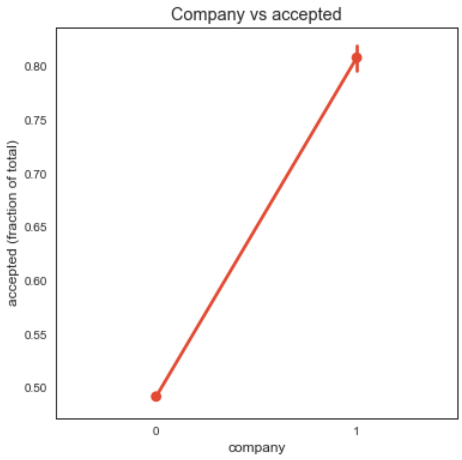
Numerical features vs. label (accepted).

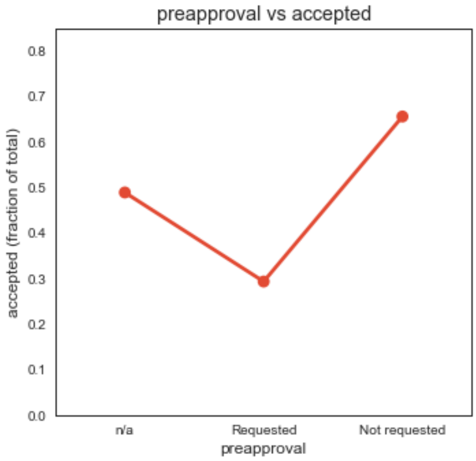
The violin plots on page 7 show the log of the skewed features in comparison with the label. Key observations are:

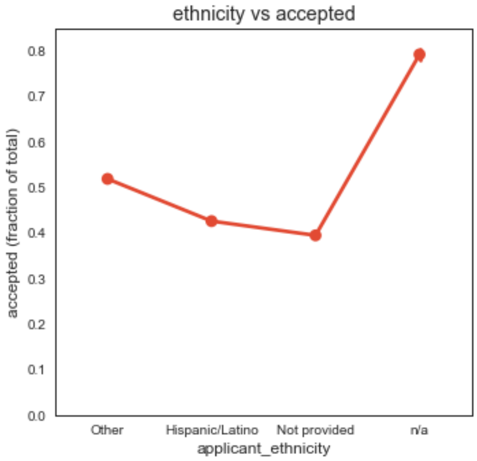
- The acceptance rate is highly dependent on the lender as presumed before. It is true that some lenders have indeed a more rigorous mortgage lending system in place, making it harder to receive a mortgage at these lenders. It is also the only feature that stands out as being very different between accepted and not accepted. This became apparent in model accuracy, which received a big boost by implementing this feature.
- Significance tests for comparing the means were necessary due to only slight visual differences. Their effect could also be measured by using model accuracy as a metric.
- Strangely enough, the effect of the applicant income and loan amount on acceptance rate is less pronounced than one would expect.

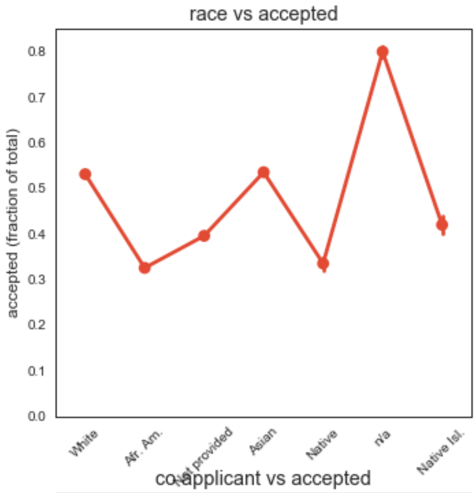

Loan purpose vs accepted
accepted (fraction of total)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Home buy
Refinance
Home Improv.
loan_purpose

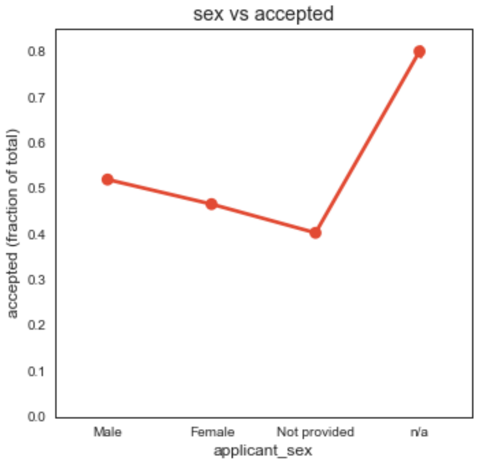

property type vs accepted
accepted (fraction of total)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1-4 Family
Manufactured
Multifamily
property_type

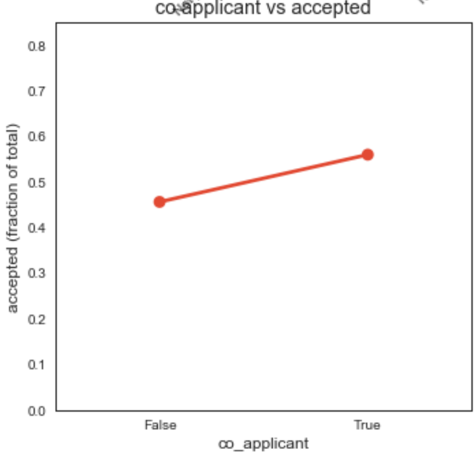

Company vs accepted
accepted (fraction of total)
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0
1
company

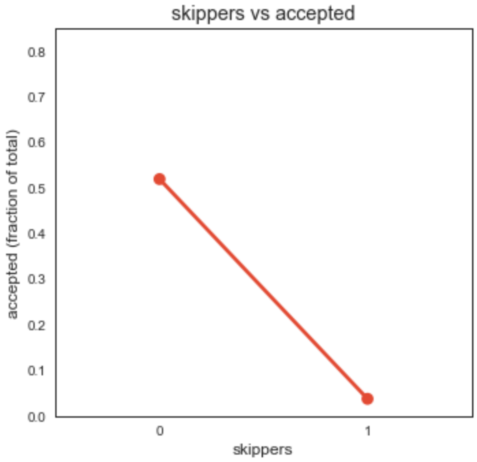

preapproval vs accepted
accepted (fraction of total)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
n/a
Requested
Not requested
preapproval

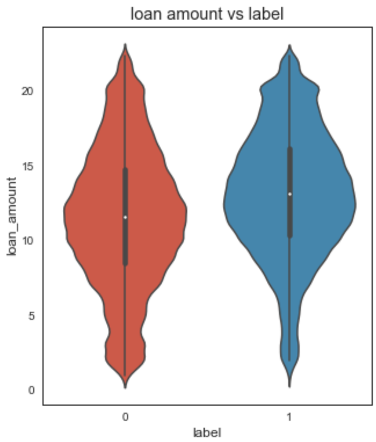

ethnicity vs accepted
accepted (fraction of total)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Other
Hispanic/Latino
Not provided
n/a
applicant_ethnicity

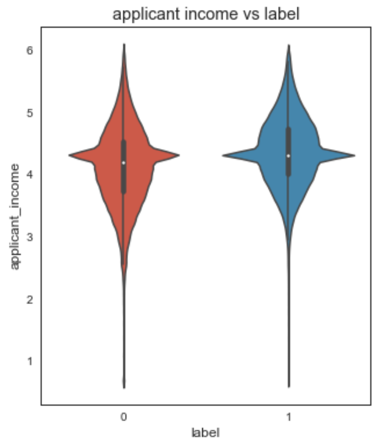

race vs accepted
accepted (fraction of total)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
White
Afr. Am.
Not provided
Asian
Native
n/a
Native Isl.

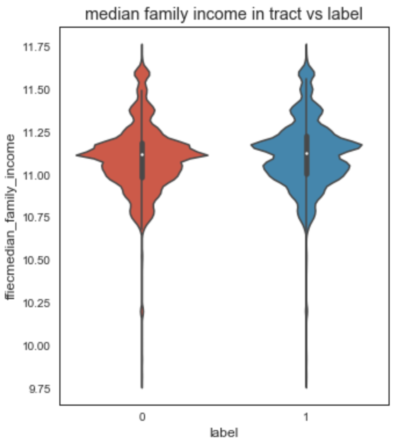

sex vs accepted
accepted (fraction of total)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Male
Female
Not provided
n/a
applicant_sex

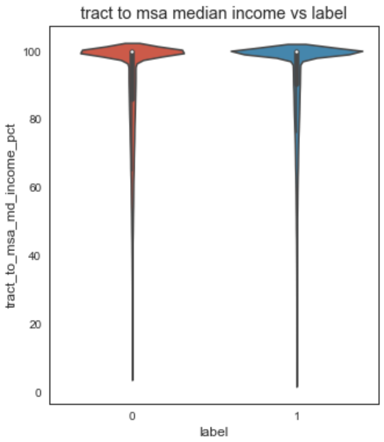

co_applicant vs accepted
accepted (fraction of total)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
False
True
co_applicant

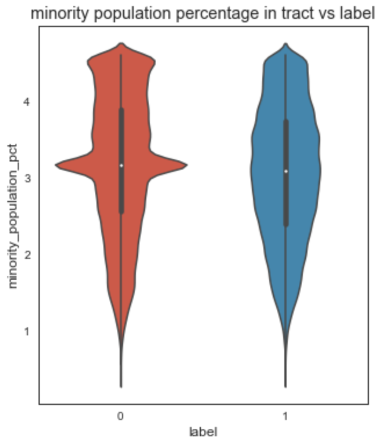

skippers vs accepted
accepted (fraction of total)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
skippers

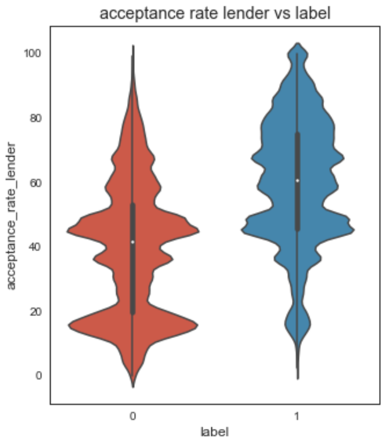

loan amount vs label
loan_amount
0
5
10
15
20
0
1
label

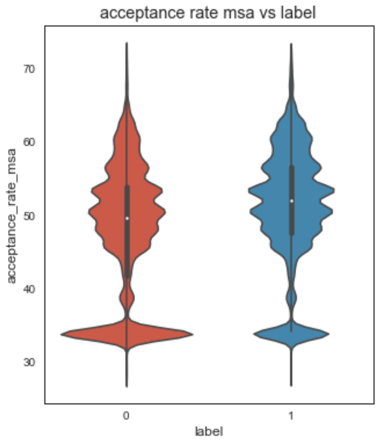

applicant income vs label
applicant_income
1
2
3
4
5
6
0
1
label

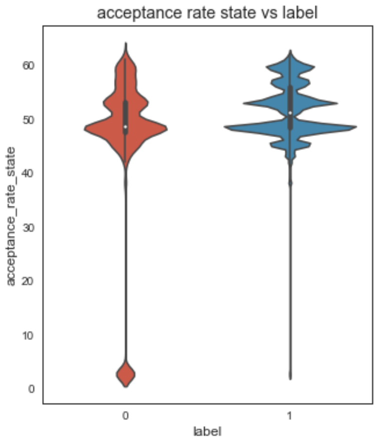

median family income in tract vs label
ffiecmedian_family_income
9.75
10.00
10.25
10.50
10.75
11.00
11.25
11.50
11.75
0
1
label

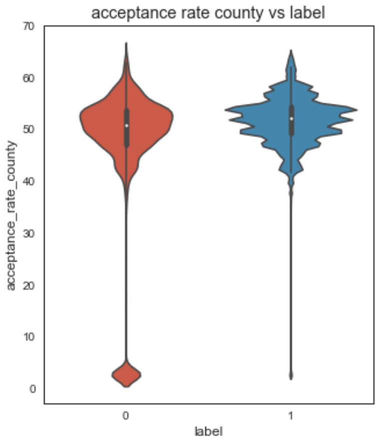

tract to msa median income vs label
tract_to_msa_md_income_pct
0
20
40
60
80
100
0
1
label

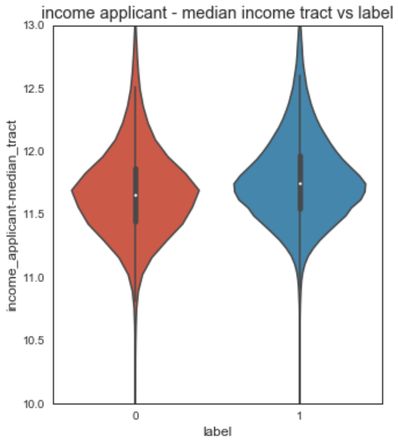

minority population percentage in tract vs label
minority_population_pct
1
2
3
4
0
1
label


acceptance rate lender vs label
acceptance_rate_lender
0
20
40
60
80
100
0
1
label


acceptance rate msa vs label
acceptance_rate_msa
30
40
50
60
70
0
1
label


acceptance rate state vs label
acceptance_rate_state
0
10
20
30
40
50
60
0
1
label

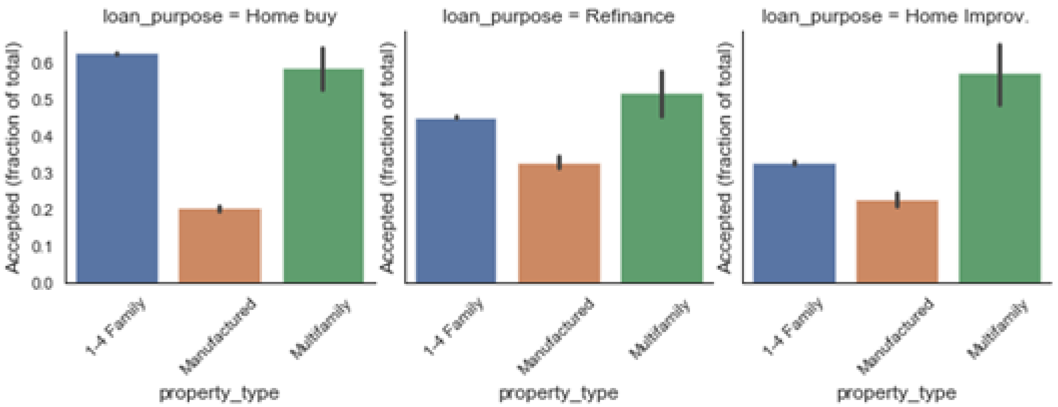

acceptance rate county vs label
acceptance_rate_county
0
10
20
30
40
50
60
70
0
1
label

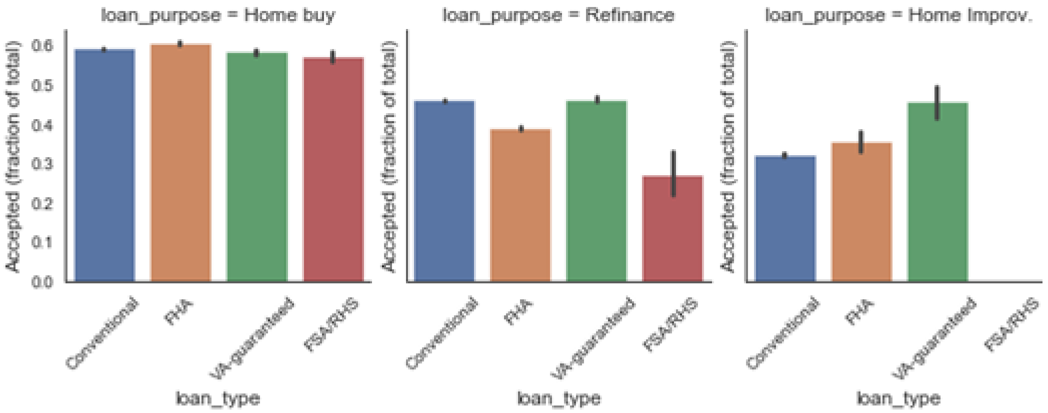

income applicant - median income tract vs label
income_applicant-median_tract
10.0
10.5
11.0
11.5
12.0
12.5
13.0
0
1
label

Multidimensional Analysis

For greater understanding of the data, a multidimensional analysis was performed using the categorical features label (accepted (0/1)) and the loan purpose as a grid. The acceptance rate is given for each combination of other categorical features. Observations for this data:

- Acceptance rate of applications for buying a 1-4 family dwelling were higher than 60 %. However applicants wanting to improve their 1-4 family dwelling only had an acceptance rate of half that.
- For buying a home, the type of loan does not seem to matter, but for home improvement, VA-guaranteed loans have the highest acceptance rate.



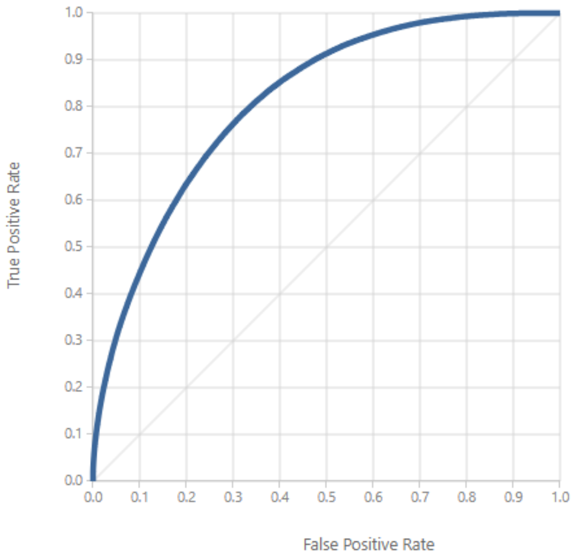


Classification of applicants according to accepted or rejected

The model was created using the Two-Class Boosted Decision Trees algorithm in Azure Machine Learning Studio with tuned hyperparameters and cross-validated using 5 folds.

• True positives:	186321
• True negatives:	171195
• False positives:	73930
• False negatives:	57267

The Received Operator Characteristic (ROC) curve for the model is shown below with the blue line indicating the model's performance at varying classification threshold values. The diagonal line represents the results of a random guess.



The model's metrics on the training data with cross validation 10 folds are given as follows:

Accuracy:	0.732
Precision:	0.716
Recall:	0.767
Area Under Curve (AUC):	0.813
F1 Score:	0.741

Conclusion

The analysis has shown that mortgage acceptance can be predicted from given characteristics. The biggest factors deciding higher acceptance rate were: which lender the applicant applied to, not skipping information on the application, applying as a company, having your spouse or another co-applicant sign with you.

The model predicts reasonably well, however the number of false positives and negatives has to be taken into consideration, concluding that the model is not a perfect predictor. The author believes the reason a certain amount of variance/noise in the data is due to the fact that mortgages are partly determined by the lender (human interaction – differing from person to person) and by the lender's local system (automized).