

## Mortgage Loans in the USA (2018–2019): Will your application be approved?

Andrés García Molina



# The Problem

The ability to procure a mortgage loan conditions the quality of life of millions of people worldwide. What are the factors that contribute to a loan application being accepted or rejected?

This project implements machine learning models on millions of observations from mortgage applications in the United States in order to determine which are the most important predictors of loan application outcomes.

# Stakeholders: Who might be interested in understanding mortgage loan approvals and rejections?

- Applicants who want to purchase their primary residence or invest in real estate
- Mortgage loan providers, who want to identify trends in their decision-making and introduce changes deriving from this insight
- Regulatory bodies and policy-makers who want to audit the mortgage loan industry

# Data

https://www.consumerfinance.gov/data-research/hmda/

An official website of the United States government

Español 中文 Tiếng Việt 한국어 Tagalog Русский العربية Kreyòl Ayisyen (855) 411-2372


**cfpb** Consumer Financial Protection Bureau

Search Submit a Complaint

Rules & Policy ▾ Enforcement ▾ Compliance ▾ Consumer Education ▾ **Data & Research ▾** News ▾

## Mortgage data (HMDA)

HMDA data are the most comprehensive source of publicly available information on the U.S. mortgage market. Learn more about mortgage activity from these data or download the data for your own analysis.



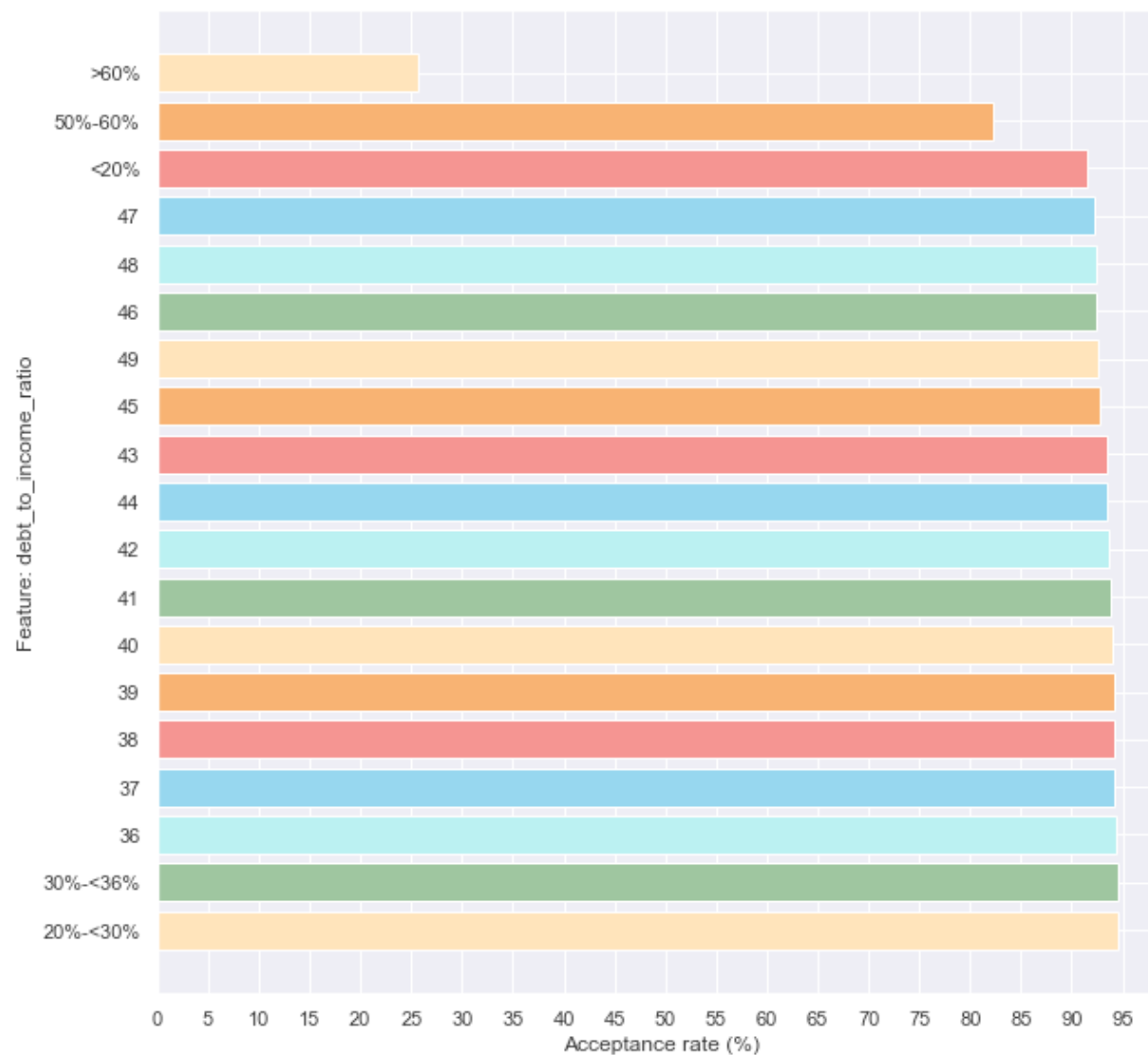
This project focuses on 2018–2019 data for *home purchases* in the United States. Over **8 million** observations were analyzed, selecting **24** out of **99** features according to *relevance* and *redundancy* criteria.

# Exploratory Data Analysis

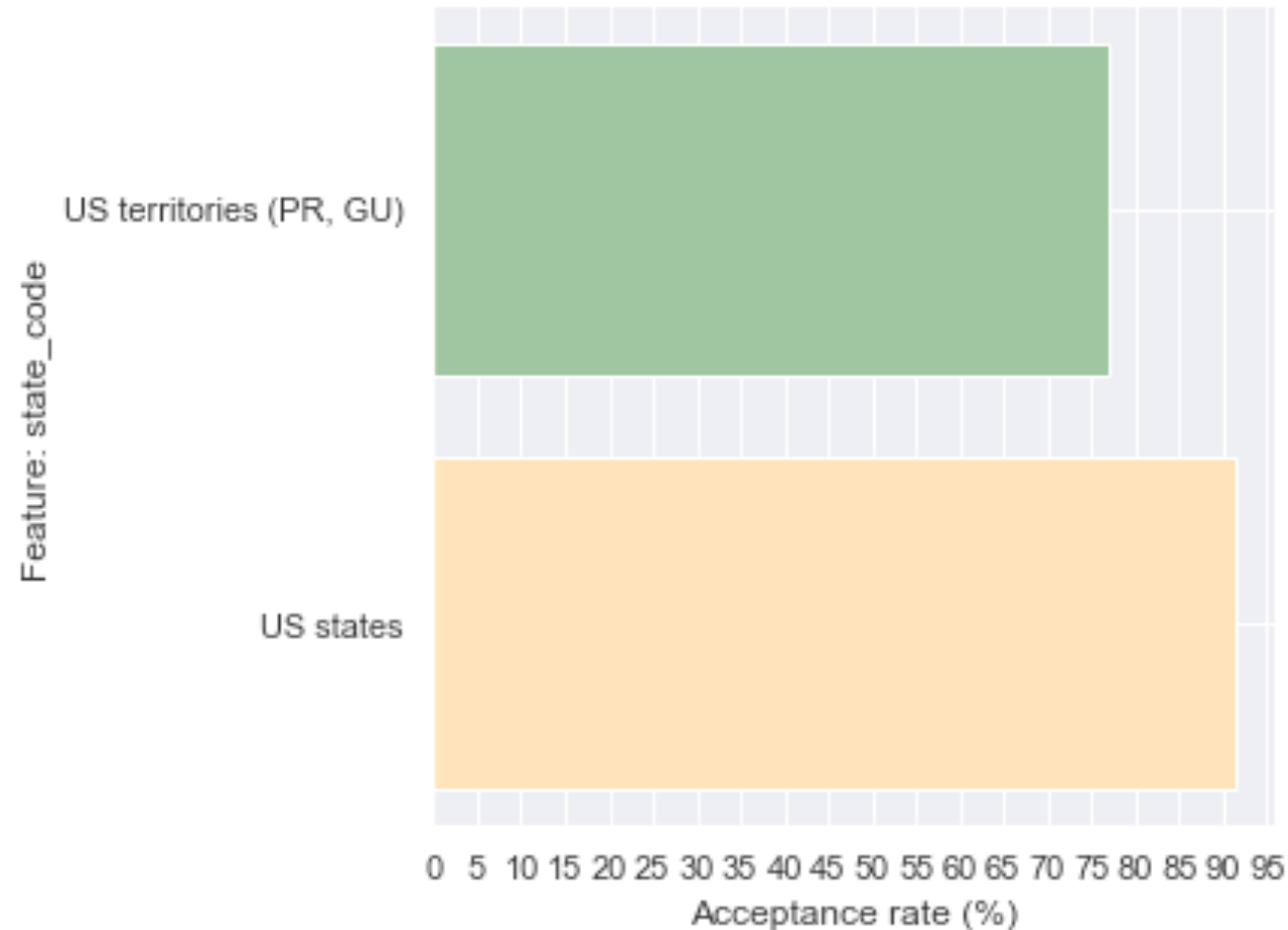
The features explored can be classified along two general types:

- **Numeric:** Features that are about economic and objective characteristics of the property, the applicant, and the loan. These include income, loan-to-value ratio, loan term, property value, property units.
- **Subjective:** Features that, in theory, should not be a factor in the decision-making process, including ethnicity, race, gender, state of residence.

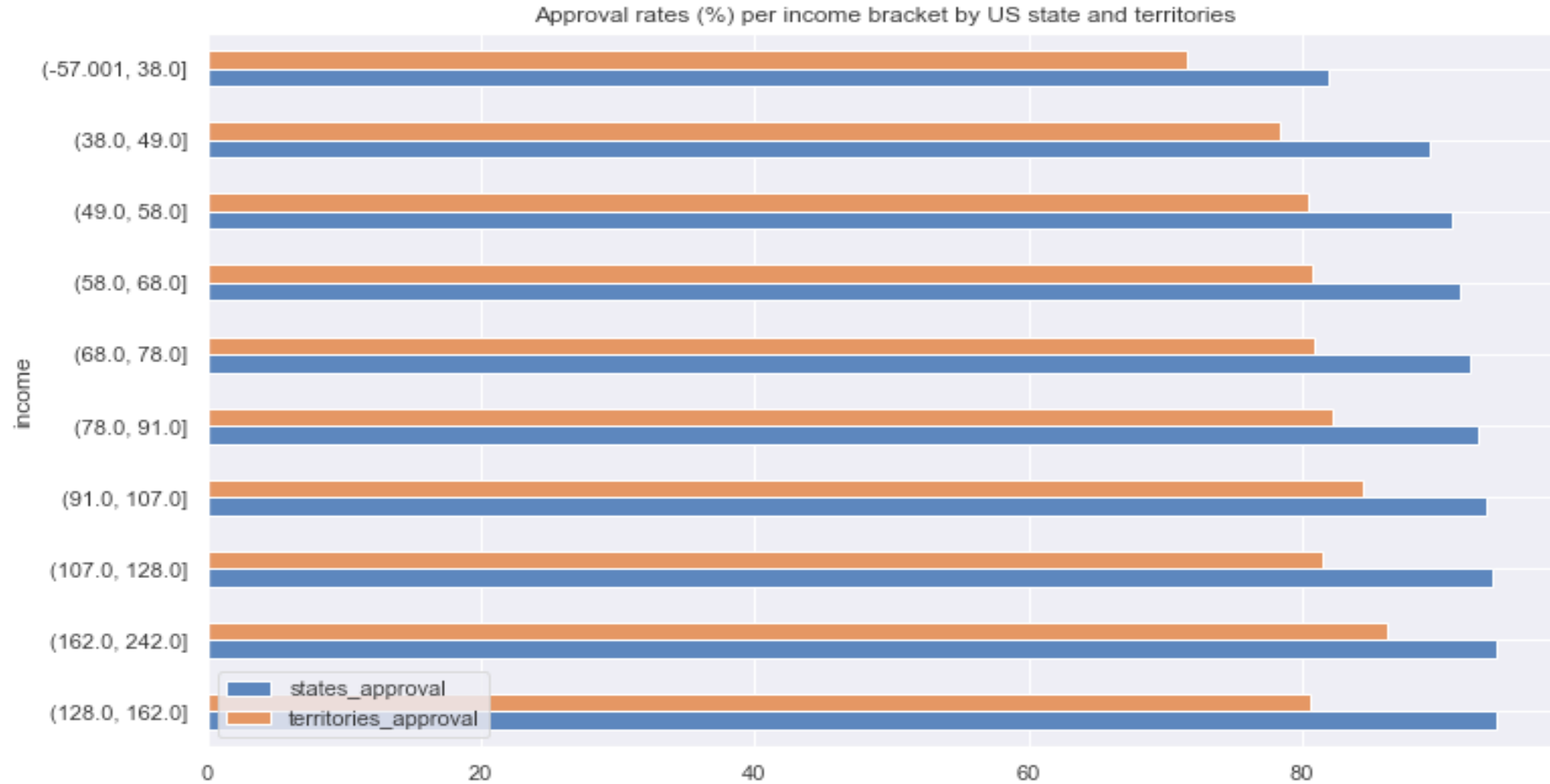
# Debt-to-income ratio in relation to acceptance rate percentage



# Should where you live influence your likelihood of securing a loan?



# Should where you live influence your likelihood of securing a loan?





# Modeling

Prior to training and testing models, the following steps were taken:

- 'Dummy' encoding for categorical variables.
- Feature normalization.

Another important step that was taken involved addressing class imbalance, given that the ratio between the majority class (approvals) and the minority class (rejections) was roughly 9 to 1. The following two strategies were adopted and compared:

- Using built-in model parameters.
- Random under-sampling of majority class.

# Models

Various models were tested, including:

Bagged Decision Trees with Random Undersampling, Standard Random Forest, Random Forest With Class Weighting, Random Forest With Bootstrap Class Weighting, Random Forest With Random Undersampling, Easy Ensemble Classifier, Weighted Logistic Regression, Linear Support Vector Classifier, Extreme Gradient Boosting

# Metrics and Top Three Models

Which metric should be used to optimize and assess models?

- Precision, recall, accuracy, specificity, ROC, etc...

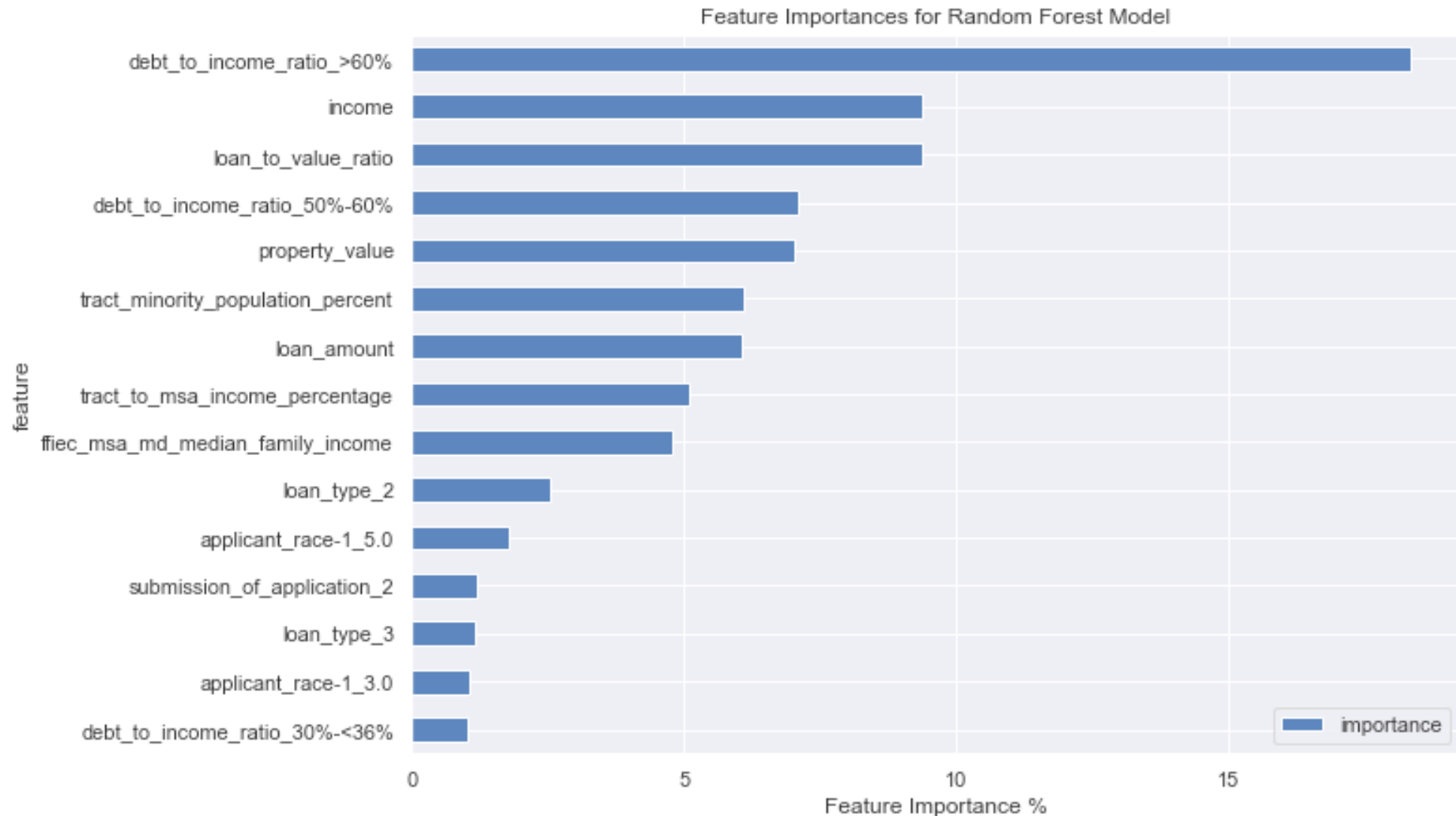
	Model	ROC_AUC Score	Accuracy
0	Random Forest	0.78	0.71
1	Logistic Regression with SGD Training	0.73	0.67
2	XG Boost	0.71	0.71

# Feature importances

As stated throughout this report, the main purpose of this project is not necessarily, or not exclusively, to be able to predict the outcome of mortgage loan applications; rather, understanding the importance of parameters that influence the actions taken by lenders is important for understanding the extent to which subjective characteristics of applicants (e.g. gender, race, ethnicity) appear to have an impact in loan application outcomes.

In sum, **explainability is crucial** in this particular case study.

# Feature importances for top model



# Further Research and Recommendations

- Further work concerning the influence of subjective factors in the mortgage loan application could examine trends along different granularities: by state, by metropolitan area, by tract. More models can be trained accordingly, with the ability to dedicate more resources to hyperparameter tuning as the sample sizes grow smaller with increased geographical specificity.
- Due to limited resources, this project only examined data from two years. A more longitudinal study could increase model accuracy as well as assist in identifying longer trends in access to mortgage loans.
- What could lenders do to improve approval rates within minorities and in underperforming locales? The first step is recognizing this is indeed an issue. This notebook contributes towards that insight.