Springboard

Andrés García Molina, PhD
January 19, 2021

Mortgage Loan Applications in the USA (2018-2019): Will your application be approved?

I. Introduction

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.

Context

Starting in 2007, the US government's Consumer Financial Protection Bureau (CFPB) has provided yearly data pertaining to all mortgage applications across the country, responding to HMDA (Home Mortgage Disclosure Act) regulations which make mortgage-loan data public in order to assess discriminatory patterns and assist public officials in targeting public investment. Data from 2007 to 2019 is available, with millions of records for each year. Data is taken from the United States government's Consumer Financial Protection Bureau and the Federal Financial Institutions Examination Council (FFIEC), which provide access to mortgage loan data, following the 1975 Home Mortgage Disclosure Act (HMDA).

In an ideal world, securing a mortgage loan would be subject to objective metrics alone, such as loan-to-value ratio, personal income, or debt-to-income ratio. This notebook explores how complex the question can be: approval rates vary according to locality, and data on factors such as race and ethnicity suggests that access to mortgage loan is to a large extent reflective of existing systemic inequalities.

Stakeholders: Who might be interested in this report?

A wide range of audiences are invested in understanding how mortgage loan applications are evaluated, including:

- 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

Methods

Several classification machine learning models are trained and tested on a carefully selected subset of mortgage data in order to identify the most important factors that influence mortgage approval rates.

II. Data Wrangling

Given the large amount of data collected and recorded by the United States federal government, this project only considers data from 2018 and 2019[1], and filters the data according to the following main parameters:

- Purpose of the loan: Only loans for home purchases are examined, thus excluding loans for home improvement and refinancing.
- Action taken: Out of seven possible categories, only three are considered: loan originated, application approved but not accepted, and application denied. The four remaining outcomes include situations that are not necessarily relevant for the immediate problem.

As expected, there is great variance in the data obtained, which includes information about applicants and their income, debt-to-income-ratio, and their race, sex, and ethnicity; data about the property itself is also included, such as the value of the property and the characteristics of the census tract in which it is located. This is a challenge when handling over 8 million mortgage loan records, where the frequency of outliers reflects the heterogeneous nature of the population.

The data contains, altogether, 8,393,516 observations and a selection of 24 out of 99 features. In very general terms, the following factors influenced feature selection:

- Relevance: The data acquired by the US government includes minutiae about loan terms that describe contractual characteristics to be executed in the event that the loan is approved. For instance, the “intro_rate_period”, defined as the number of months until the first date the interest rate may change after closing or account opening. Other data describing loan terms belongs to majority classes: for instance, a feature that tells us whether an application is for a reverse mortgage (i.e. one that enables the borrower to access the unencumbered value

[1] *Note:* The reporting schema changed in 2018, expanding the features from 44 to 98.

of the property, often without having to issue monthly payments) reveals that over 99% of the applications are *not* for mortgage loans.
- Redundancy: Some of the features recorded are summaries or aggregates of other features, thus establishing linear dependencies. For instance, the "derived_sex" feature is a single aggregated sex categorization derived from applicant/borrower and co-applicant/co-borrower sex fields.

More details on the selection criteria for features can be found in the Data Wrangling notebook of this project. Additional documentation in the form of feature dictionaries is provided in the documentation section of the FFIEC's HMDA website; each year contains slight variations. This link provides access to the 2019 feature dictionary.

In addition to feature selection, the data required extensive cleaning and validation for common issues including missing values, out of range values (e.g. an applicant's age listed as 9999 or -1), and outliers. These steps are also described in the Data Wrangling notebook of this project.

III. Exploratory Data Analysis (EDA)

EDA was performed in order to get a general sense of distributions in the dataset. Emphasis was found on detecting anomalies in data pertaining to approval rates: that is, the relationship between features and approval rates was examined. The features explored can be classified along two general types:

- Numeric features: 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: Features that, in theory, should not be a factor in the decision making process, including ethnicity, race, gender, state of residence.

An analysis of each feature in relation to approval rate is included in the EDA notebook of the project. For this report, I focus on those features which showed the greater disparity across subjective features after presenting one example of a numeric feature's relation to approval rates.

As expected, an applicant's debt-to-income ratio is indeed a determining factor in the outcome of their loan applications, as exemplified below. Clearly, a debt-to-income ratio greater than 60% impacts acceptance rates negatively.

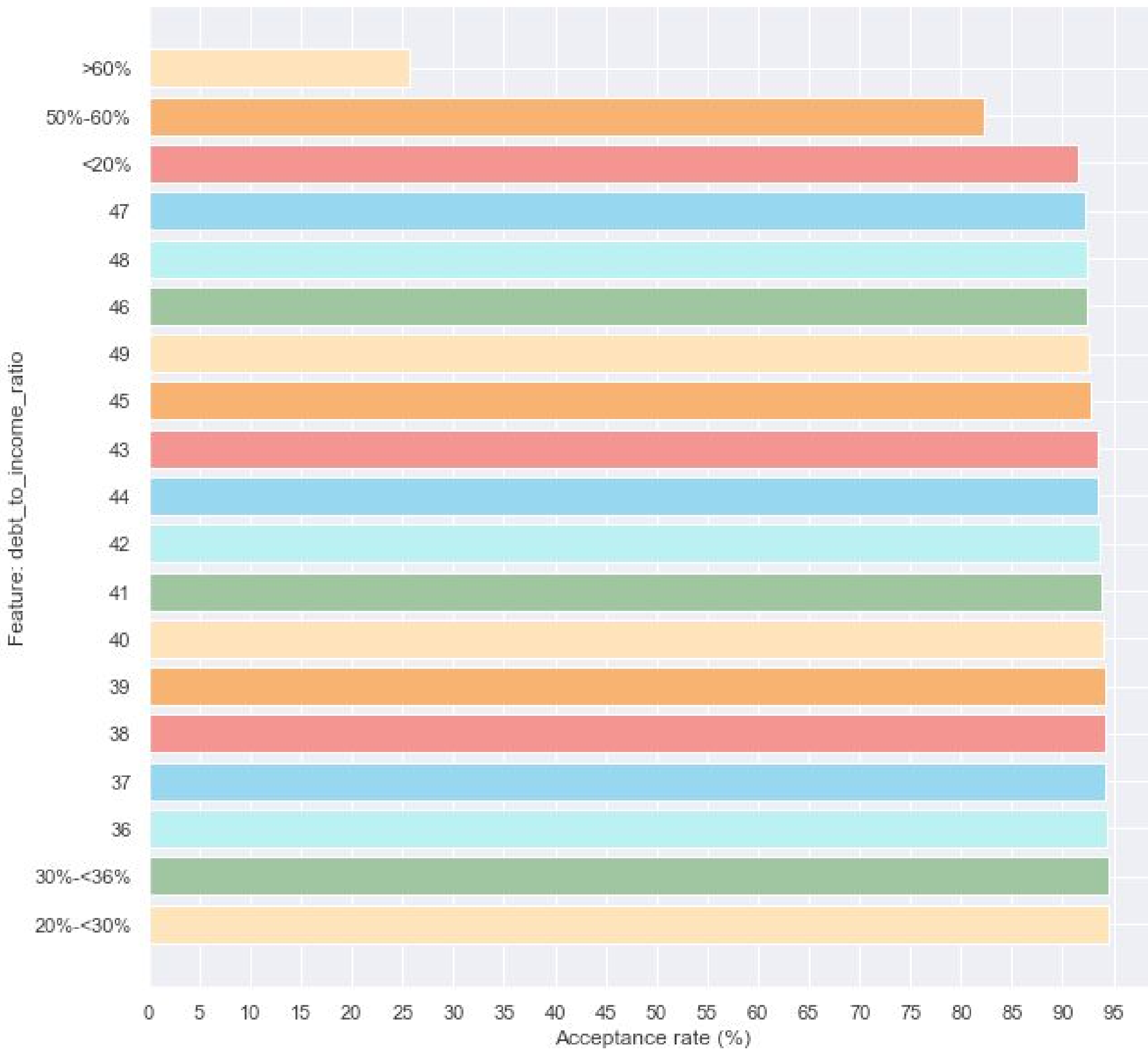

Figure 1: Acceptance rate in relation to debt-to-income ratio expressed in percentage brackets.

Subjective Features and Hypothesis Testing

The central question this project examines is the extent to which subjective factors might influence the ability to procure a loan. In this section I present some graphs that show the relationship between some subjective features and approval rates. The results of analysis of variance (ANOVA) hypothesis tests are also included.

We begin with applicant states: To what extent does living in a certain location affect approval rates? The plot below shows that living in US territories—PR (Puerto Rico) and Guam (GU)—as opposed to states presents a significant disadvantage:

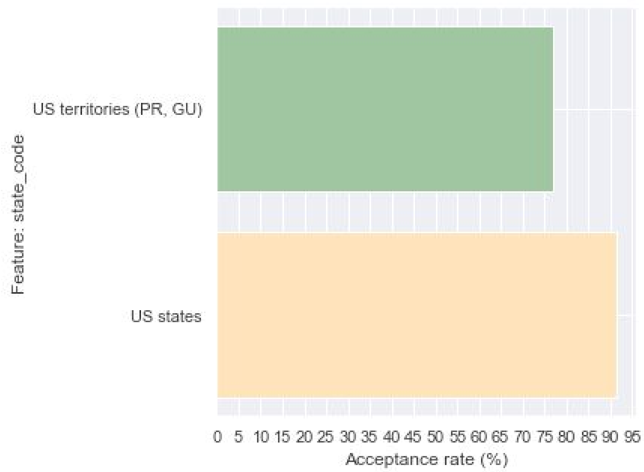

Figure 2: Acceptance rate by type of location (states and territories). A breakdown specifying approval rates per each individual state and territory is provided in the EDA notebook of the project.

An ANOVA test between territories and states yields an F value of 6147.30 and a p-value lower than 1.0 x 10-8. A Kruskal-Wallis H-test, which is the non-parametric version of ANOVA, confirms the results with an F value of 6142.80 and a p-value of lower than 1.0 x 10-8. These tests indicate that under the null hypothesis it would be impossible to obtain the current distribution of values.

Another way to frame the question is to ask whether we observe a similar disparity when we control by income. The following graph presents how each geographic group performs along binned income categories.

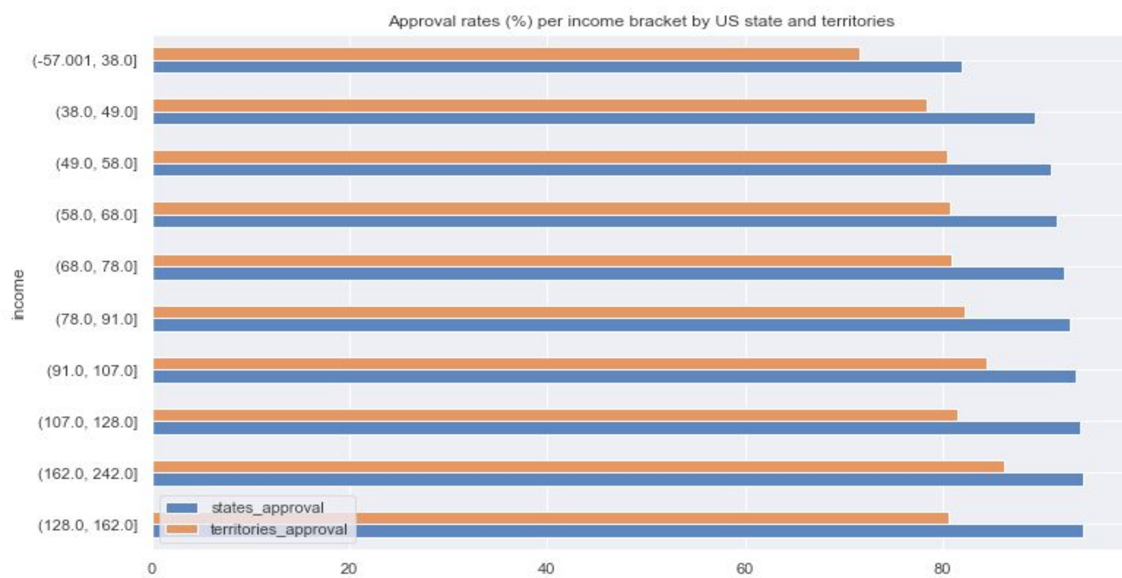

Figure 3: The income categories are prescribed by the data, and are expressed in thousands of dollars. Territories consistently underperform across binned income categories.

The full notebook outlines how other subjective categories—such as applicant ethnicity and race—are related to approval rates unequally. While these discoveries do not suggest a clear-cut relation of causality, they warrant more investigation. One last example is included here that underlines disparity in approval rates across identity categories:

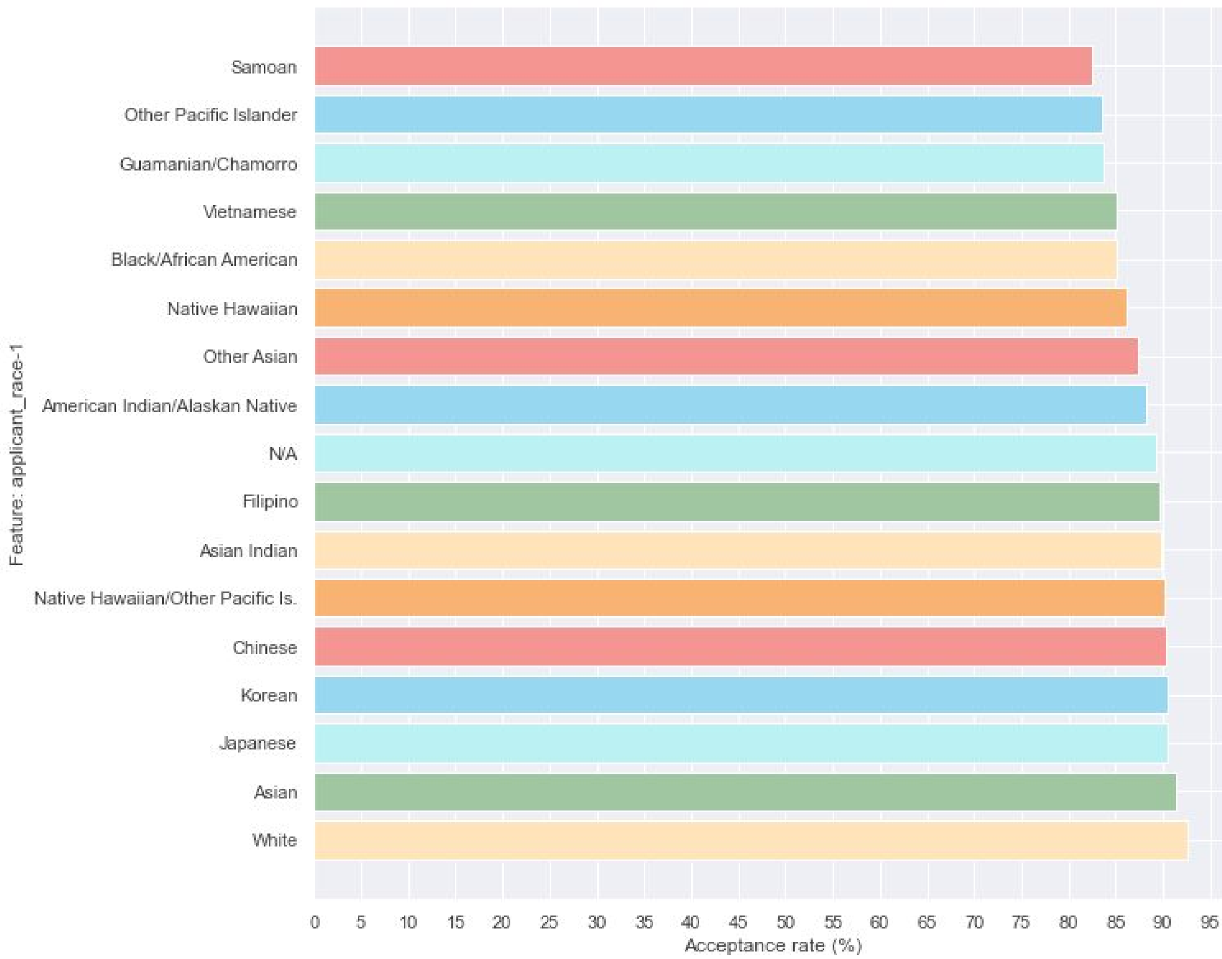

Figure 4: Acceptance rate by race categories.

As in the case with states and territories, ANOVA and Kruskal-Wallis tests indicate that the null hypothesis with regards to applicant race has to be rejected.

IV. Modeling

Prior to training and testing models, the following steps were taken, outlined in greater detail in the Feature Engineering notebook of the project:

1. 'Dummy' features for categorical variables were created in order to facilitate machine learning implementation.

2. The data was split into training and testing sets.
3. Numerical features were normalized.

Another important step that was taken involved addressing class imbalance, given that the ratio between the majority class (approvals) and the minority class (rejections) is roughly 9 to 1. The following two strategies were adopted and compared, with results presented at length in the Modeling notebook of the project:

1. Addressing class imbalance through built-in model parameters.
2. Addressing class imbalance in the dataset prior to modeling, and training models on balanced data.

Various models were tested, including training without hyperparameter tuning but addressing :

- 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

Because the most valuable data is the one for the minority class (rejections), and since the dataset contains millions of observations, downsampling the majority class was the approach selected, yielding an overall better performance.

Because of the relatively high processing times required when testing models, a first pass was performed without hyperparameter tuning to get a sense of time and metrics per model. Hyperparametric tuning with RandomizedSearchCV was performed on the top three performing models: Random Forest, Logistic Regression with SGD Training, and Extreme Gradient Boosting.

Metrics

The most relevant metric in this project is **specificity**, or recall of the negative class, since in this project emphasis is placed on the correct identification of mortgage denials; after all, recall penalizes for false negatives. While undersampling means losing a significant amount of data in the majority class, the purpose of this project is to identify the factors that affect the minority class. Therefore, the decision to downsample is justified.

Scikit-learn offers no scoring optimizer for hyperparameter tuning that is based on specificity alone; in any case, optimizing for specificity alone might compromise predictions for the positive class. The metric most suitable to use in this case, and which takes specificity into account is the Area Under the Receiver Operating Characteristic Curve (ROC AUC).

There are other cases in which other metrics might be more suitable. For instance, placing emphasis on **precision**, a metric that penalizes false positives, could be of interest if loan institutions are concerned about approving loans that should have been withheld. Conversely, an emphasis on **recall**, a metric that penalizes false negatives, would prioritize cases in which loans that should have been approved were not. A rationale could be to say that this is business lost for lenders and reduced opportunities for borrowers.

	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

Figure 5: Model metrics comparison.

<u>Feature importance</u>

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.

The following graphs describe the top fifteen most important features according to each of the top three models. All coincide that by far the most weight is given to debt to income ratio, especially if its value is greater than 50%; the most weight given to debt to income ratios above 60%. Also important is data related to income and loan amount. In the three top models selected, race and ethnicity are within the top 15 important features, although their overall importance is less than 5%.

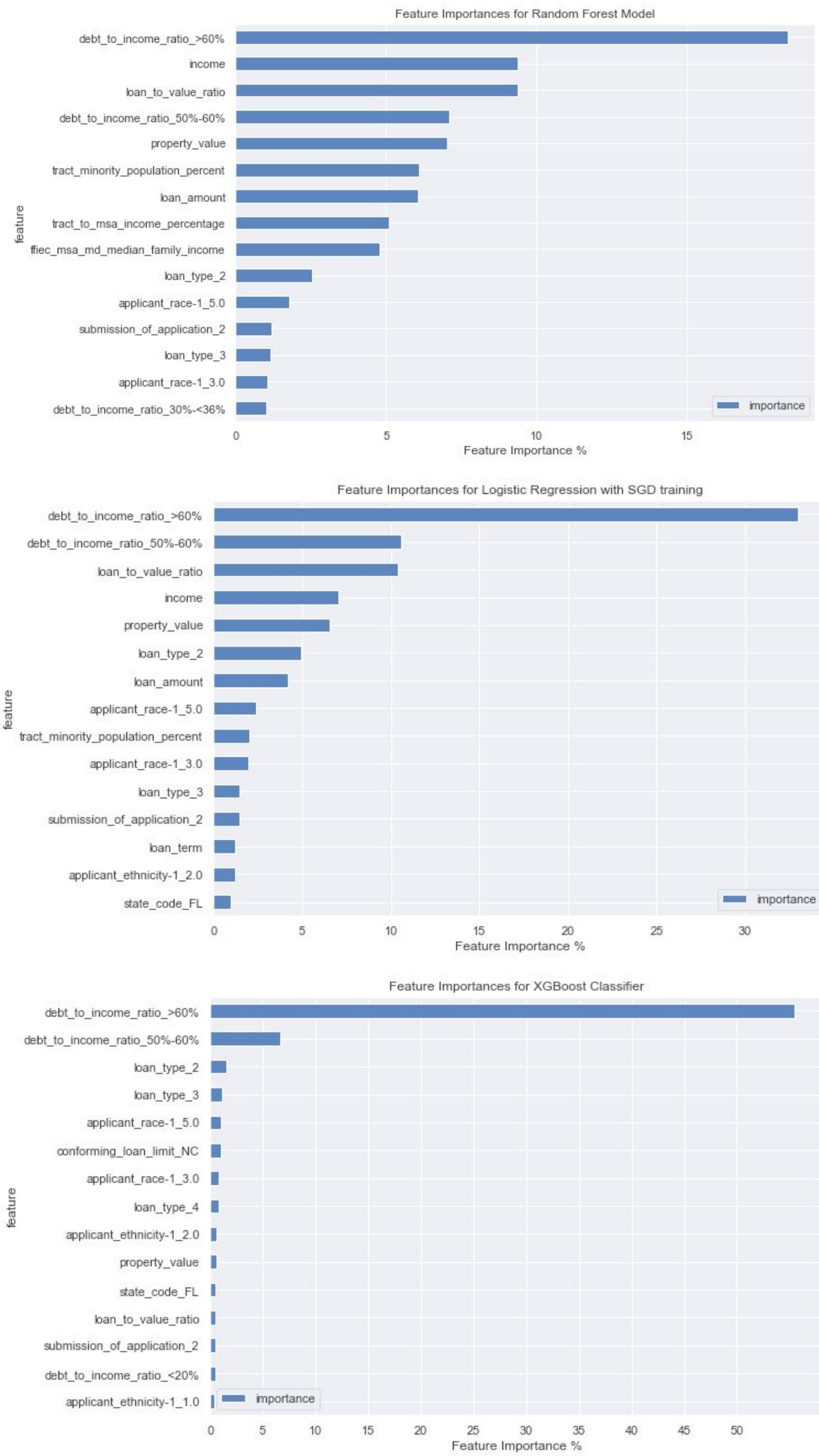

Figure 6: Feature Importances for each of the three top models.

V. 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.