Home mortgage approval/denial in California counties

Henrique Martins

Useful information

Repository:

https://github.com/hpmartins/mlai-ucb-codes/tree/main/capstone_project

Data source:

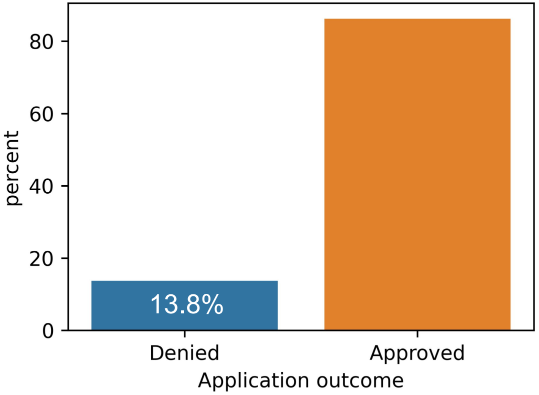
https://ffiec.cfpb.gov/

Data set documentation:

https://ffiec.cfpb.gov/documentation/publications/loan-level-datasets/lar-data-fields

Data set preprocessing

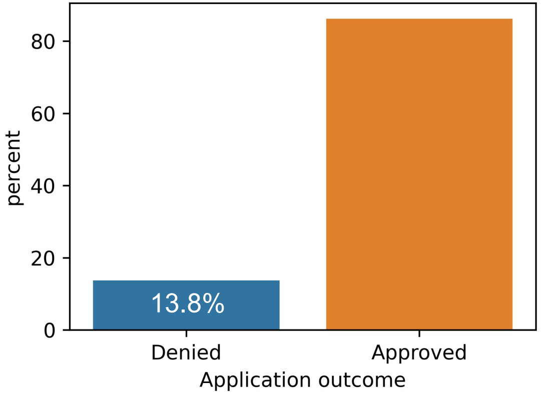
- The original data set contains 616388 entries and 99 features
- All relevant features are mapped into either numbers or categories, converting their values using the data documentation and grouping them into a new value whenever reasonable
- Some features are pre-filtered due to outliers (income, loan value, and others)
- All missing values are dealt with; categorical unknowns with very few samples are dropped

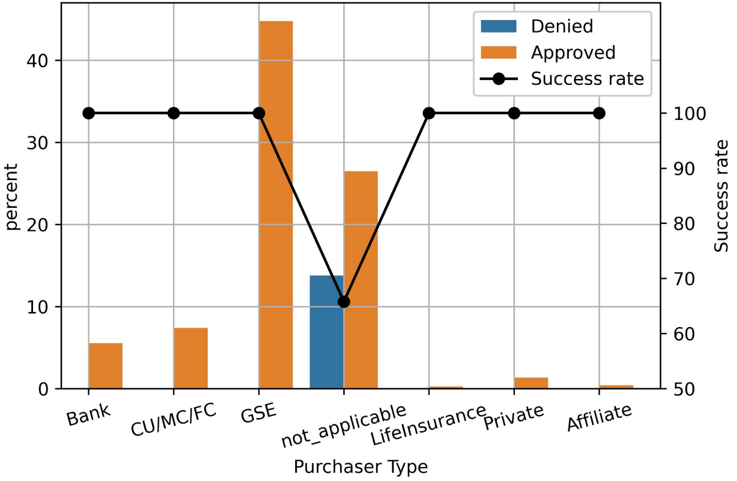


- Output of preprocessing step are 549263 entries and 38 features
- Application outcome presents heavily imbalance towards Approved applications

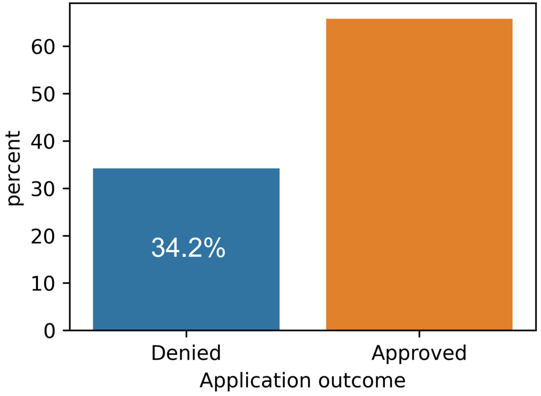
Data set preprocessing: Filtering “Purchaser type”

The *Purchaser Type* feature describes whether the applicant is an entity or not an entity. All named values are entities, while “not_applicable” are people (not entities). The data was filtered to only have people. That removed most of the target imbalance.





only not_applicable

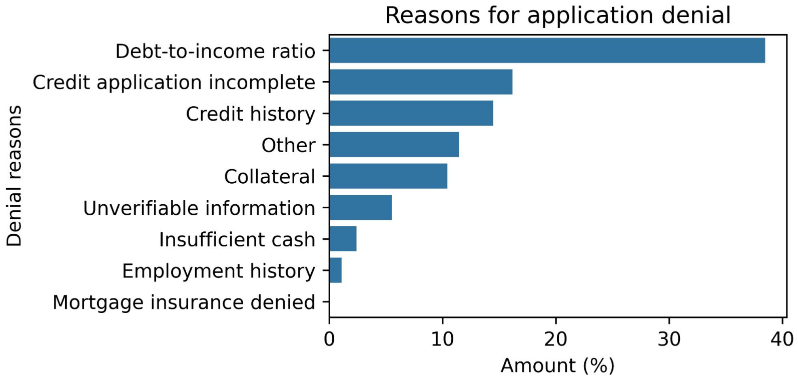


Findings: Summary

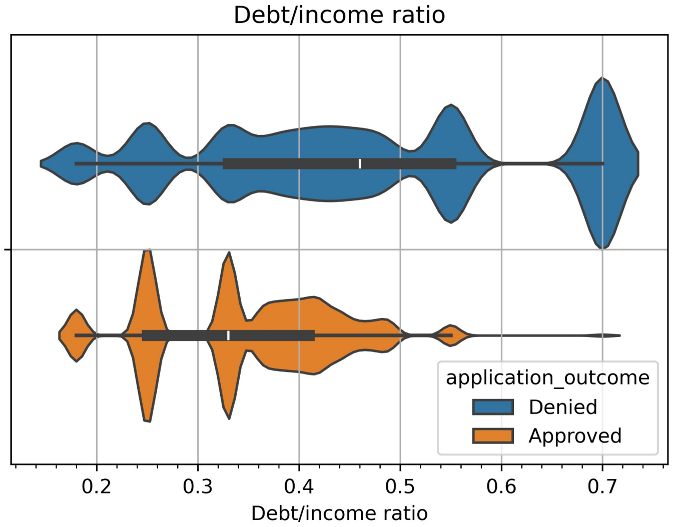
- The main reason for denial of application is low debt-to-income ratio
- Higher approval rates at younger ages
- Higher approval rates for males and joint applicants
- Different rates of approval depending on the applicant’s race and ethnicity
- Lower approval rates for lower income applicants
- Higher approval rates for applicants that use more expensive properties to secure the loan

Findings: Reasons for application denial

The main reason for denial of application is debt-to-income ratio, followed by credit application incomplete and credit history.



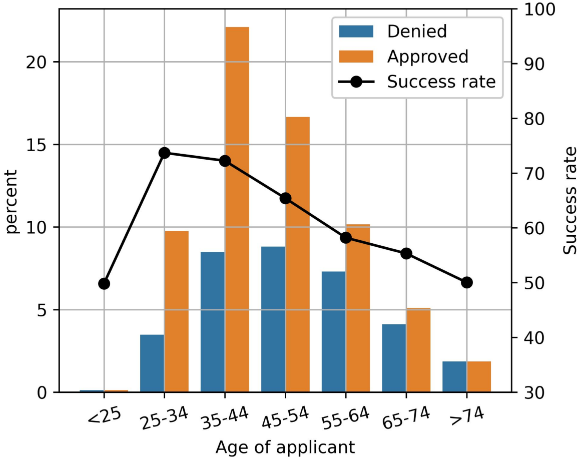
Most approved applicants have a debt/income ratio of around 0.3, while above 0.5 the applications are mostly all denied



Findings: Applicant’s age matter

Higher approval rates at younger ages

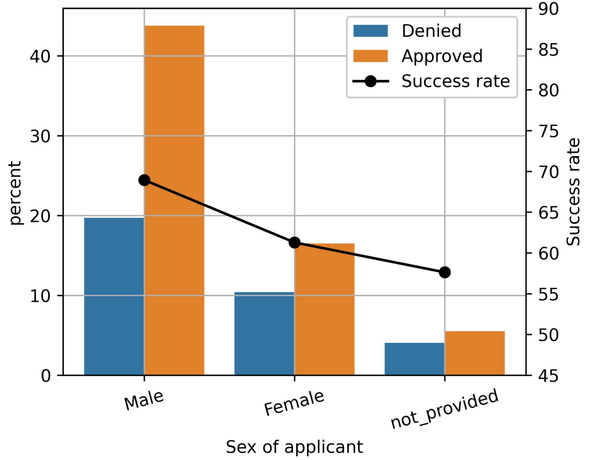
The application success rate depends on the age of the applicant: for ages below 25 the approval rate is around 50%, and it increases to around 75% between ages of 25-34. Beyond that it decreases as the age increases, going back to 50% for applicants above 74 years old.



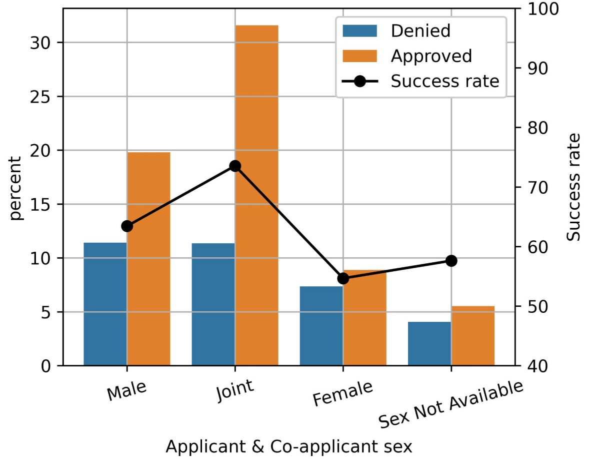
Findings: Applicant's sex matters

Higher approval rates for males and joint applicants

The main applicant's sex seems to play a role in the application outcome. Male applicants have around 70% approval rate, whereas female applicants are around 57%.

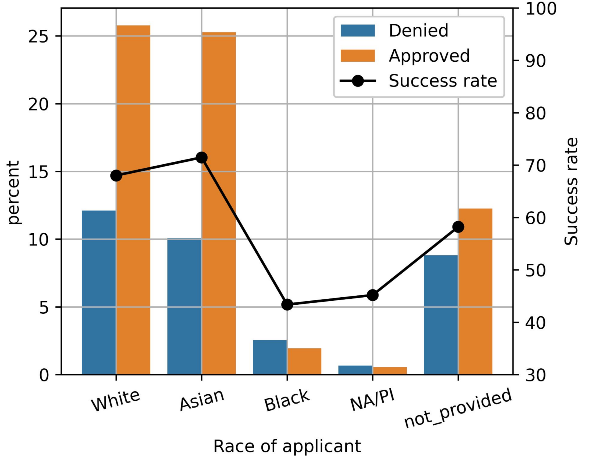


If we group both the applicant and co-applicant's sex together, the approval rate for male-male goes down to 62%, while filing having both sexes has an approval rate of around 73%. Female-female remains around 55%.

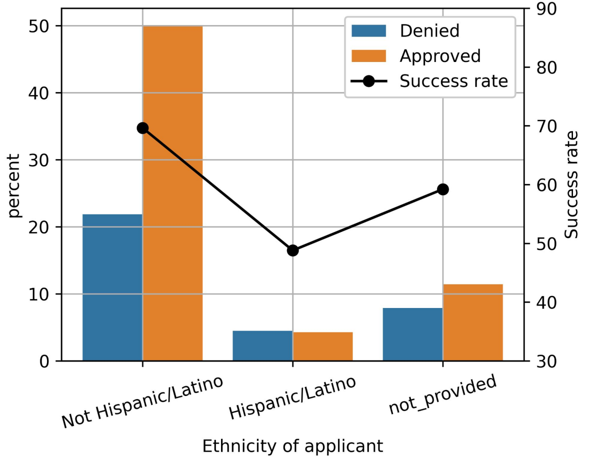


Findings: Applicant’s race and ethnicity

The approval rate for Asian and White races is around 70% and higher than for Native Americans, Pacific Islanders, and Blacks/African Americans which sit at around 42%.

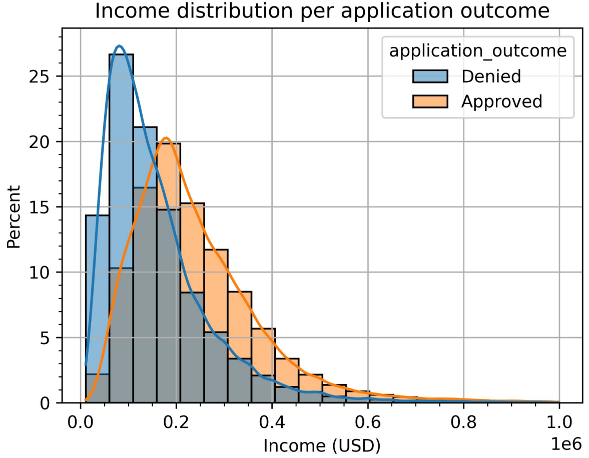


The approval rate is around 50% for Hispanic/Latino ethnicity and 70% for non-Hispanic/Latino ethnicity.

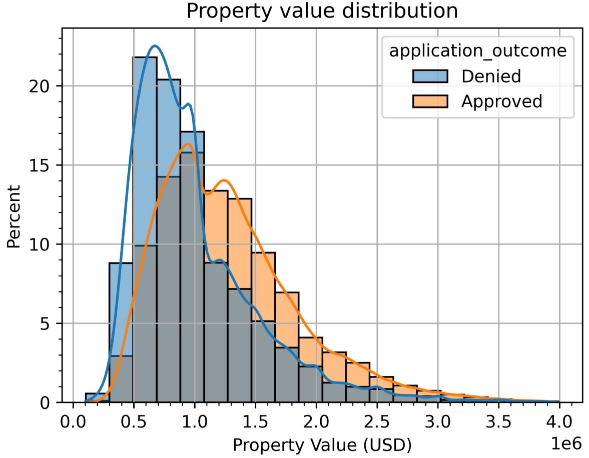


Findings: Applicant’s income and property value

The approval rate goes above 50% for incomes above US$ 150 thousand per year. Most of the denied applicants have incomes lower than that.



The value of the property proposed to secure the loan has a similar behavior, with most denied applicants having used a property valued at less than US$ 1 million to secure the loan.



Models summary

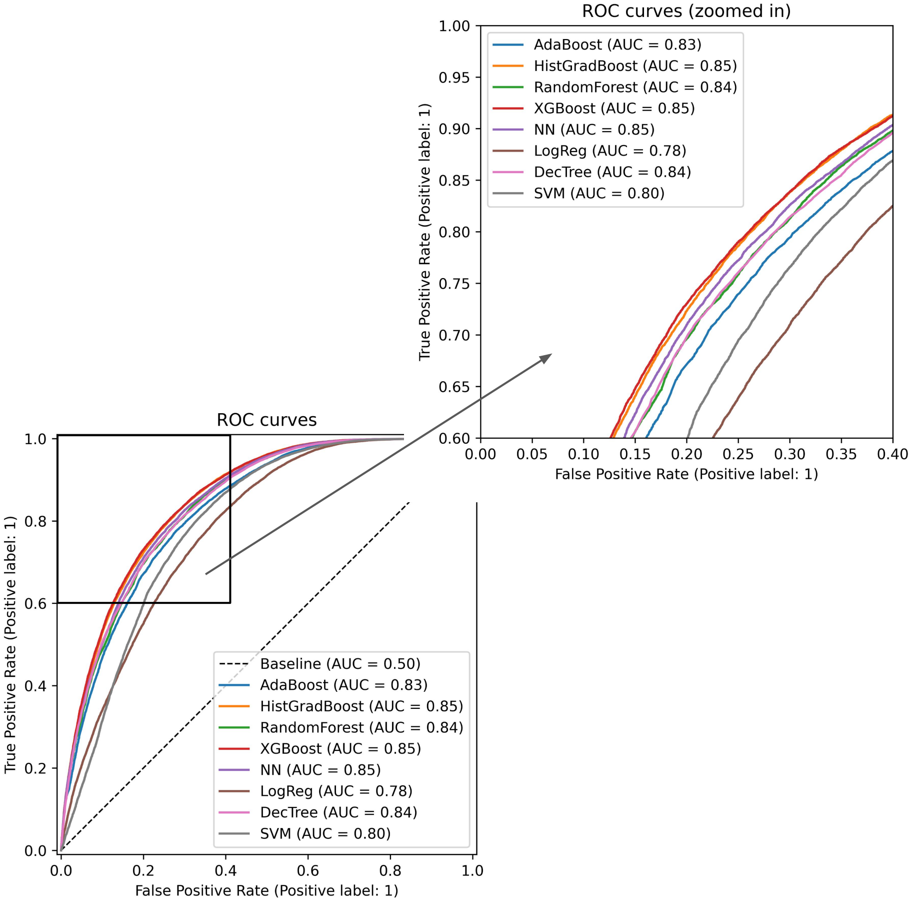
- DummyClassifier (baseline)
- LogisticRegression, Decision Tree, SVM
  - Relatively good and very fast
  - GridSearchCV on same models: hyperparameters do not improve much
- Ensemble:
  - HistGradientBoosting
  - AdaBoost
  - RandomForest
- XGBoost
  - Booster: DART; dropout rate at 0.1
  - Objective function: binary:logistic
  - Tracked classification error at multiple thresholds and AUC-PR for early stop
- Neural Networks:
  - Loss function: BinaryFocalCrossentropy

Summary table of performances and
metrics on next slide

Model performances

Model	Train score	Test score	Precision		Recall	
			Approved	Denied	Approved	Denied
DummyClassifier	0.658	-	-	-	-	-
LogisticRegression	0.770	0.773	0.77	0.78	0.93	0.46
Decision Tree	0.796	0.798	0.79	0.81	0.93	0.54
Support Vector	0.786	0.788	0.79	0.79	0.93	0.51
Random Forest	0.798	0.797	0.78	0.86	0.96	0.49
HistGradientBoosting	0.740	0.743	0.80	0.82	0.94	0.55
AdaBoost	0.785	0.788	0.77	0.85	0.96	0.46
XGBoost	**0.808**	**0.804**	**0.80**	**0.83**	**0.94**	**0.53**
Neural Network	0.797	0.802	0.80	0.81	0.93	0.55

Model performances



The XGBoost is the model that seemed to perform better overall at classifying the outcome of the applications.

The Denied outcome has a low recall value for all models. HistGradientBoosting and Neural Network models had better outcomes.

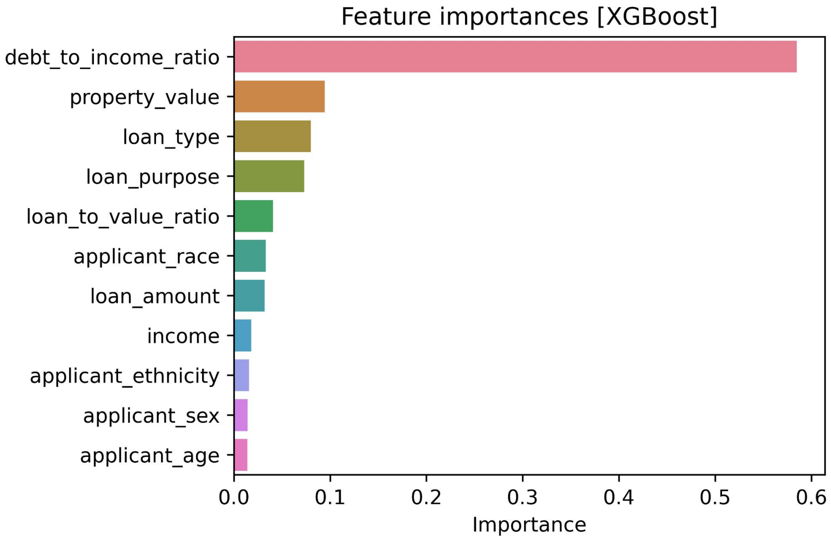
Model	Train score	Test score	Precision		Recall	
			Approved	Denied	Approved	Denied
DummyClassifier	0.658	-	-	-	-	-
LogisticRegression	0.770	0.773	0.77	0.78	0.93	0.46
Decision Tree	0.800	0.799	0.80	0.80	0.93	0.55
Support Vector	0.786	0.788	0.79	0.79	0.93	0.51
Random Forest	0.798	0.797	0.78	0.86	0.96	0.49
HistGradientBoosting	0.751	0.744	0.80	0.82	0.94	**0.56**
AdaBoost	0.785	0.788	0.77	0.85	0.96	0.46
XGBoost	**0.807**	**0.803**	**0.79**	**0.87**	**0.96**	**0.50**
Neural Network	0.797	0.802	0.80	0.79	0.92	**0.57**

Feature importances

- *Debt/income ratio* is by far the most influential factor, suggesting that the model heavily relies on this metric to assess the applicant's financial risk.

- *Property value*: given a good debt/income ratio, the value of the property proposed to secure the loan has a strong influence on the outcome. The loan type and the purpose of the loan also affect in the same order of magnitude.

- *Fairness and Bias*: The low importance of demographic features might appear positive from a fairness perspective. However, as noticed during the EDA phase, there is influence (direct or indirect) on the application outcome.


Feature importances [XGBoost]
debt_to_income_ratio
property_value
loan_type
loan_purpose
loan_to_value_ratio
applicant_race
loan_amount
income
applicant_ethnicity
applicant_sex
applicant_age
Importance