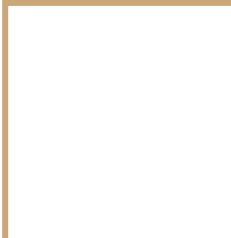# Availability of Home Mortgages

Russell Brown

Table of Contents

- Intro

- The Data

- Bayesian Inference

- Predictive / Inferential Modeling

- Conclusions

- Next Steps

Intro

- Home Mortgage Disclosure Act, enacted in 1975

- Grew out of concern about racial segregation

- Financial institutions keep a ledger of mortgage applications

- Data published to the public every year

- HMDA does not mandate any specific conditions be met

The Data

- 2015 dataset is VERY large (9.2 Gb of data)

- Maryland

- One-to-four family dwelling

- Home purchase

- Secured by first lien

The Data

Loan Acceptance Rates by Race (%)
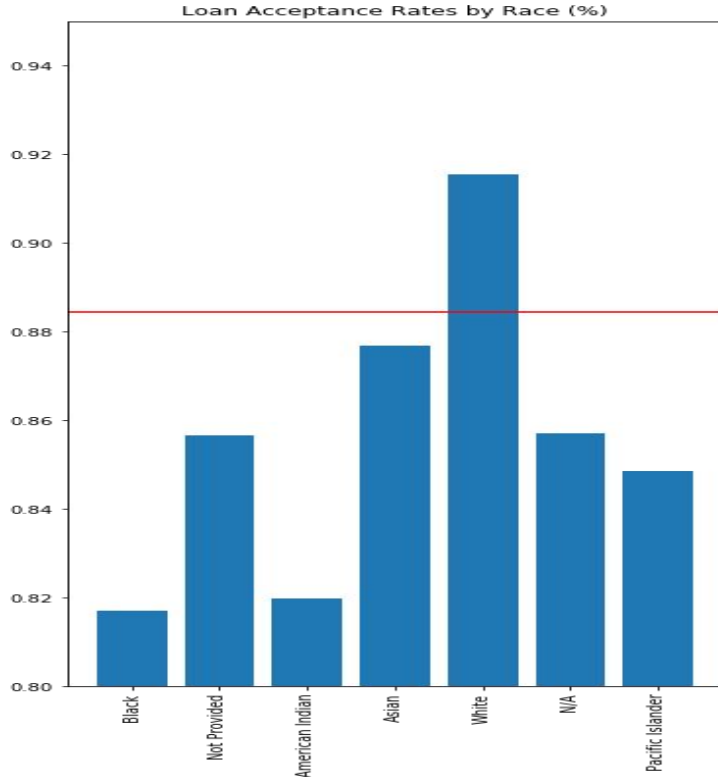
Sample Size

White	48177
Black or African American	17780
Information not provided….	10853
Asian	6707
American Indian or Alaska Native	345
Native Hawaiian or Other Pacific Islander	238
Not applicable	15

The Data

Loan Acceptance Rates by Co-Applicant Race (%)
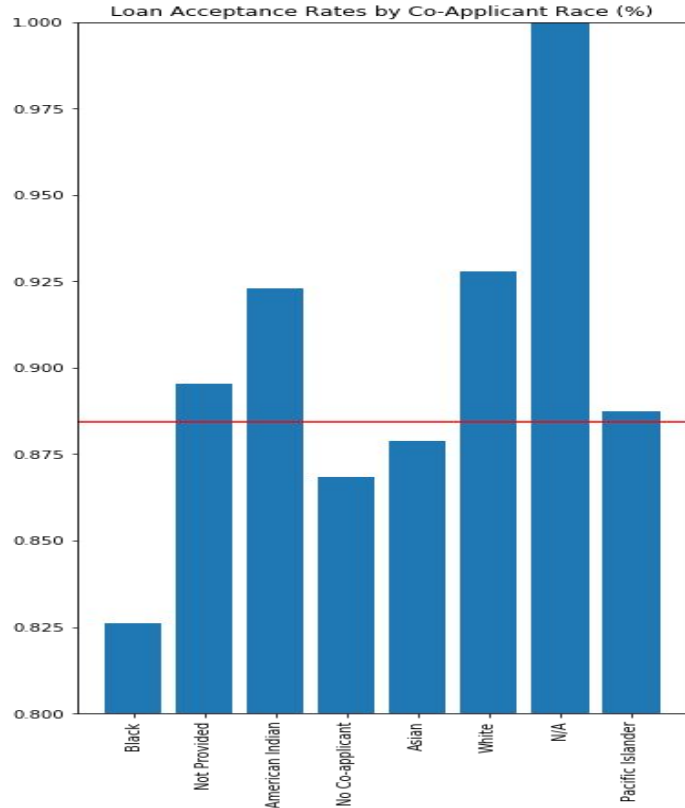
Sample Size

No co-applicant	49171
White	22338
Information not provided....	4957
Black or African American	4320
Asian	3114
Pacific Islander	104
American Indian or Alaska Native	103
Not applicable	8

Bayesian Inference

- Bernoulli prior distribution with normally distributed p

- Approval rate for white people (most likely) vs. black people (least likely)

- Use pymc3 sampler to generate posterior distributions

- Determine if difference is statistically significant

Bayesian Inference
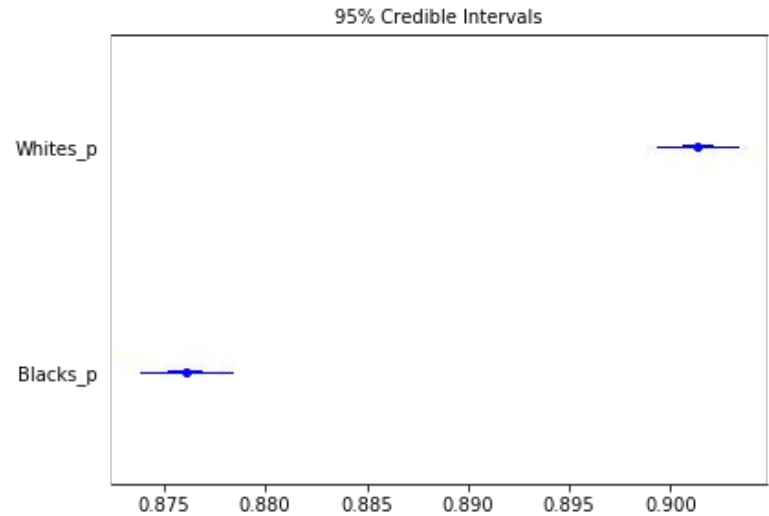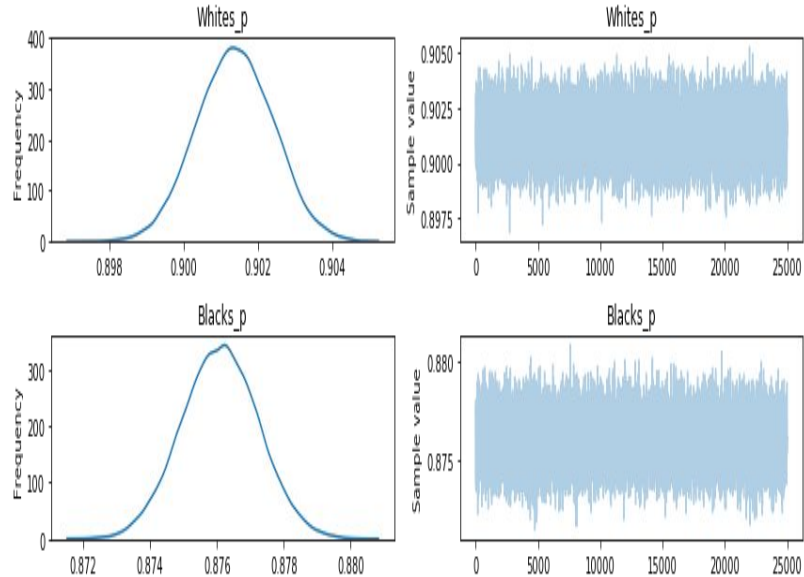
Predictive Modeling

-   Factorize categorical features
-   No need to preprocess numerical columns
-   Evaluate based on cross-validated ROC-AUC scores

Extra Random Trees -   Avg Score: 0.904
                          Std. Dev.: 0.01826

Random Forest -   Avg Score: 0.922
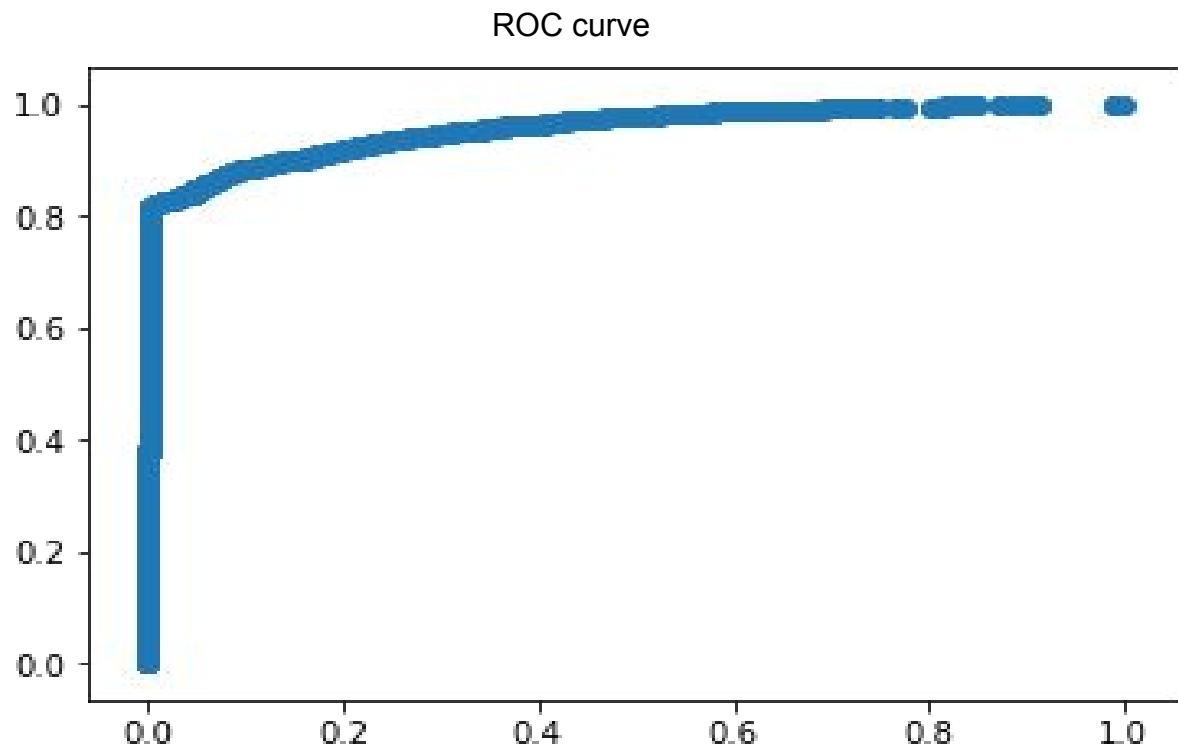   Std. Dev.: 0.01424

Extreme Gradient Boost - Avg Score: 0.956
                           Std. Dev.: 0.01605

Predictive Modeling

- XGBoost Classifier model

- GridSearchCV to choose best tuning parameters

- .9573 ROC-AUC on a 90/10 train-test split

- .9569 without race of applicant or co-applicant

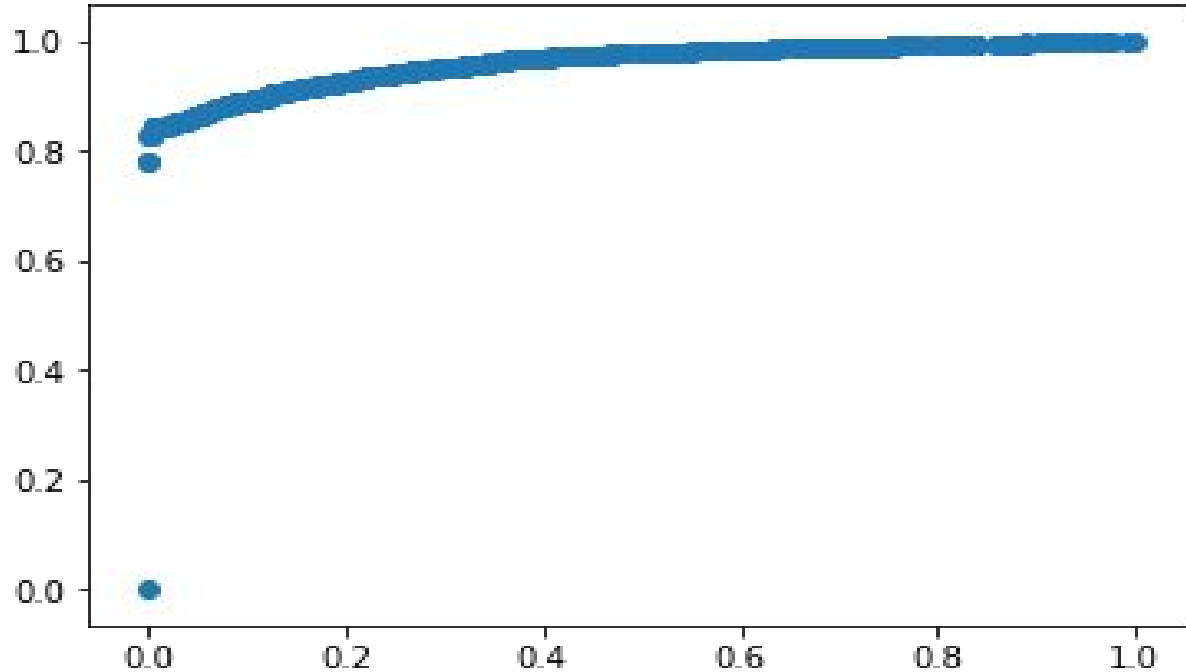# Predictive Modeling

ROC curve

AUC: 0.957

Inferential Modeling

- Logistic regression

- Predicts odds relative to baseline categories

- Performed just as well as Xgboost model

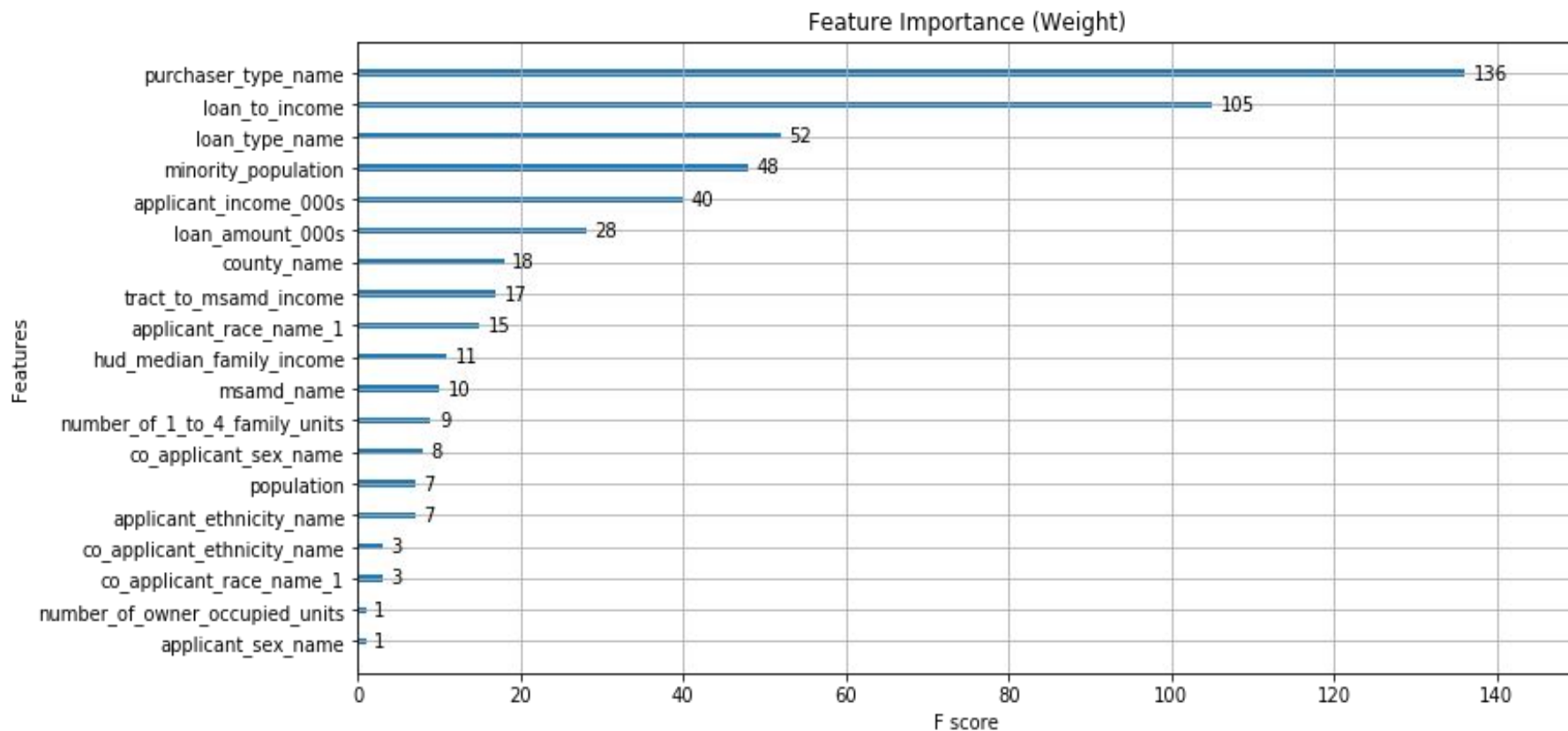- Tells us how large of an effect each feature has

Inferential Modeling

ROC Curve

AUC: 0.960

Conclusions

Feature Importance (Weight)

Features	F score
purchaser_type_name	136
loan_to_income	105
loan_type_name	52
minority_population	48
applicant_income_000s	40
loan_amount_000s	28
county_name	18
tract_to_msamd_income	17
applicant_race_name_1	15
hud_median_family_income	11
msamd_name	10
number_of_1_to_4_family_units	9
co_applicant_sex_name	8
population	7
applicant_ethnicity_name	7
co_applicant_ethnicity_name	3
co_applicant_race_name_1	3
number_of_owner_occupied_units	1
applicant_sex_name	1

Conclusions

Odds of Getting a Loan (Relative to Whites)

Not Provided:	0.914
Black:	0.723
Asian:	0.656
American Indian:	0.573
Pacific Islander:	0.550

Conclusions

- Effect of race dwarfed by other factors

- Removing race as a factor barely impacts model performance

- Race has low feature importance

- 3 different metrics (weight, gain, cover)

Next Steps

- Expand analysis to entire 2015 dataset

- Look for trends from previous years

- Freddie Mac Single-Family Loan-Level Dataset

- Include credit score