# Loan Sharks

**Illinois Institute of Technology**

- Anantanarayanan G Iyengar
- Bhuvnesh Tejwani
- Virat Joshi
- Xiaoman Shen
- Omkar Pawar

# Agenda

- Introduction and Motivation
- Problem definition
- Data description - Primary and Secondary sources
- Exploratory Data Analysis
- Data Cleaning and feature selection
- Data Modelling
- Model Deployment
- Takeaways
- Q&A

# Introduction

# Era of Big Data and ML

- Improving people's lives

  - Self driving cars

  - Smart meters

  - Healthcare, etc.

- Endless possibilities

- Is it all good?

  - 2007/2008 crisis

  - Smart policing?

  - Job applications screening, loans, etc.

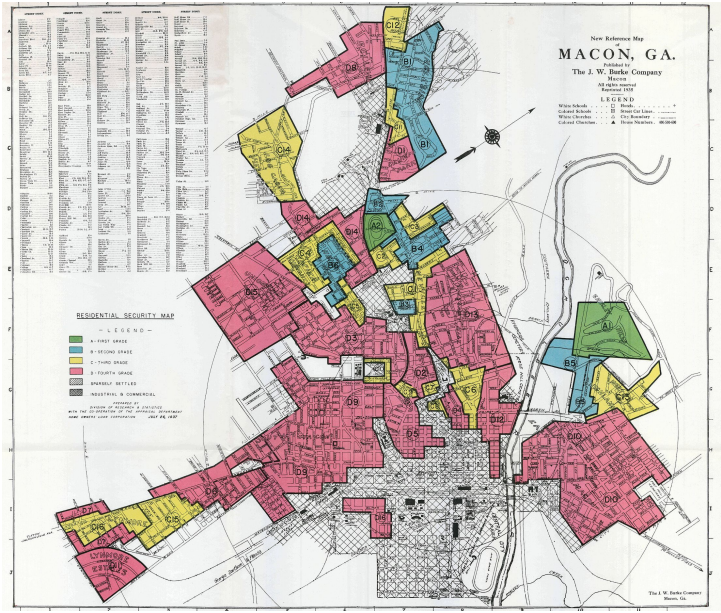
  - Winners and Losers -> long term effects?

# Racial bias?

- Traffic stops more for minorities (Stanford Open Policing)
- Higher car insurance rates
- Algorithms and models implicitly induce racial bias (Cathy ‘O’ Neil)
- Opacity, Scale, Feedback loop - More discrimination
- Effects on home ownership?. Is redlining back?

# Problem Statement

1. Racial disparity in home lending?
2. Analyze home loan data and provide insights.
3. Build predictive models to get the probability of applicants to get the loan



# Data Sources

# Data Sources

**Primary Data Source**

**Mortgage data (HMDA) (Required by law)**

- From Consumer Financial Protection Bureau.
- Choose a data set from years 2014 to 2017 for Illinois,  2007-2008, 2014-2017 for Pennsylvania.
- Dimensions (rows upto 310000, columns 78).
- Contains home loan application information per state and region.
- For e.g. Race, ethnicity, income, loan amount, loan type, loan decision, etc.

# Data Sources

**Secondary Data Source**

## Zillow Home Value Data

- This dataset includes Region id, State FIPS Code, Municipal FIPS Code and median monthly prices.
- Median home price in the county was added as a column to the hmda data used for models.

# Data Sources

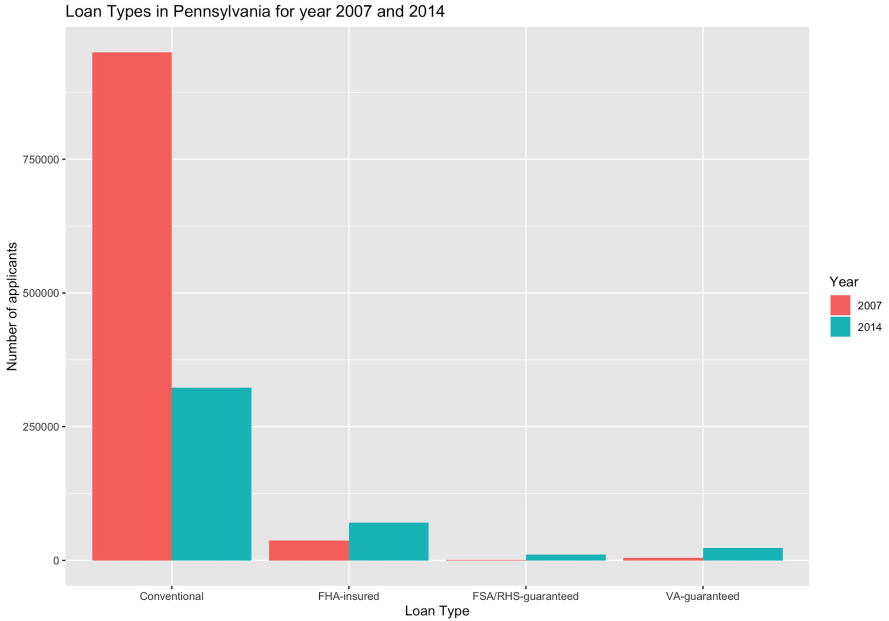
**Secondary Data Source**

**Census:** County demographic data

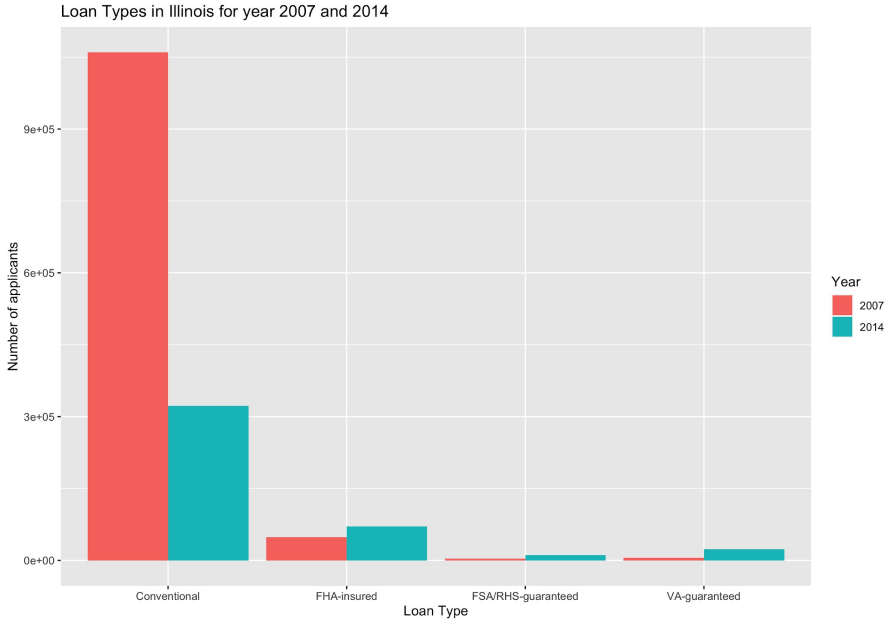
- From the United States Census Bureau
- This dataset includes State FIPS code, County FIPS code, County name, Total population and the numbers of men and women of different races.
- Percentage of races per county was used as an additional predictor in the modeling process.

# Exploratory Data Analysis

# Loan Type Distribution for years 2007 and 2014



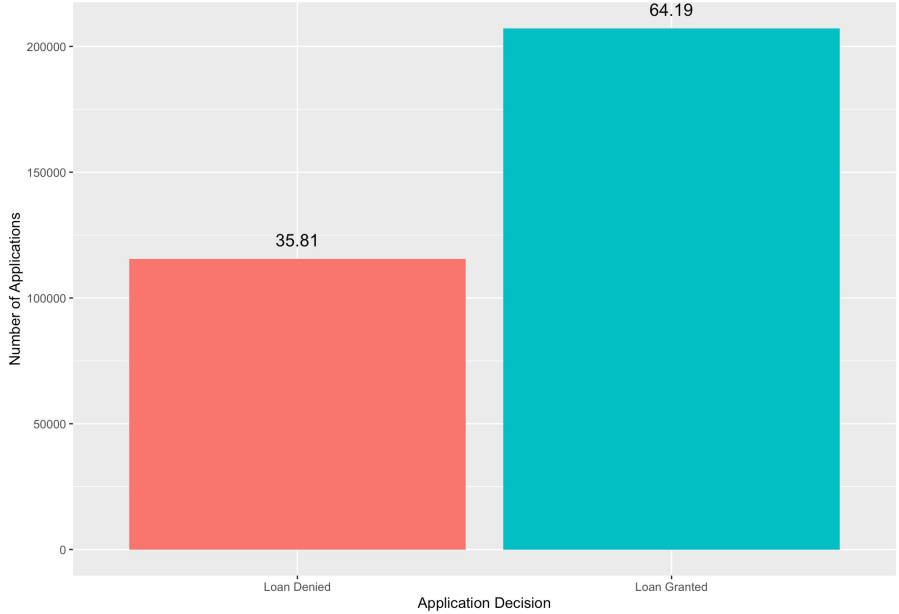

Pennsylvania



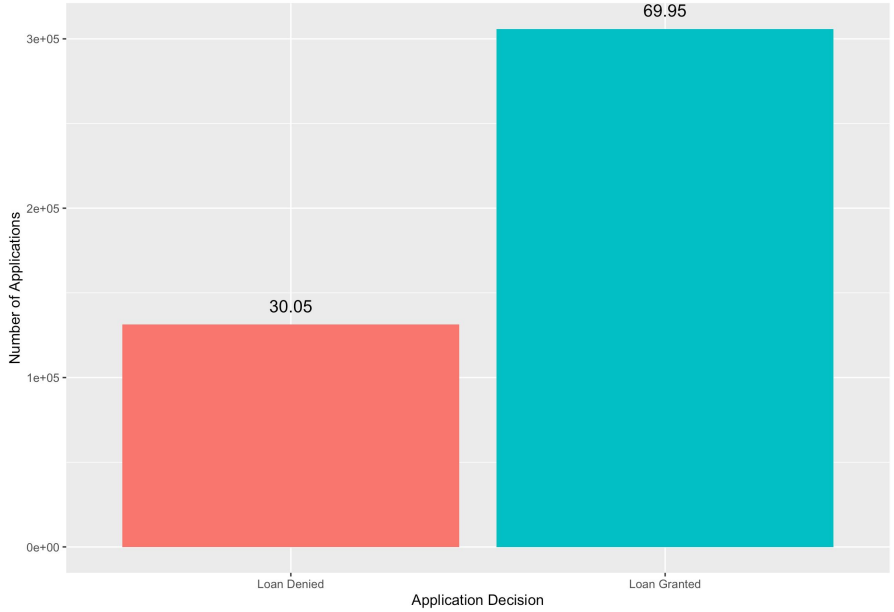

Illinois

# Proportion of Loans Granted and Denied

Pennsylvania

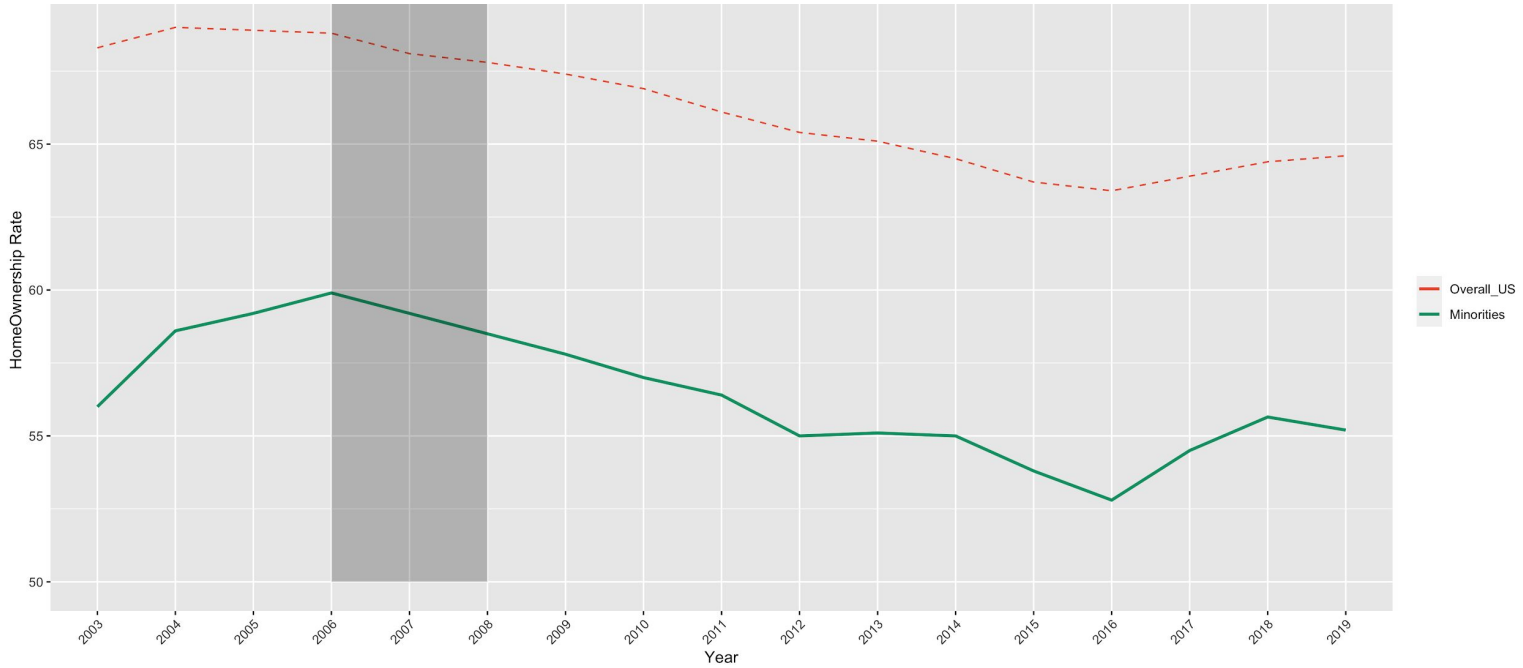



Illinois



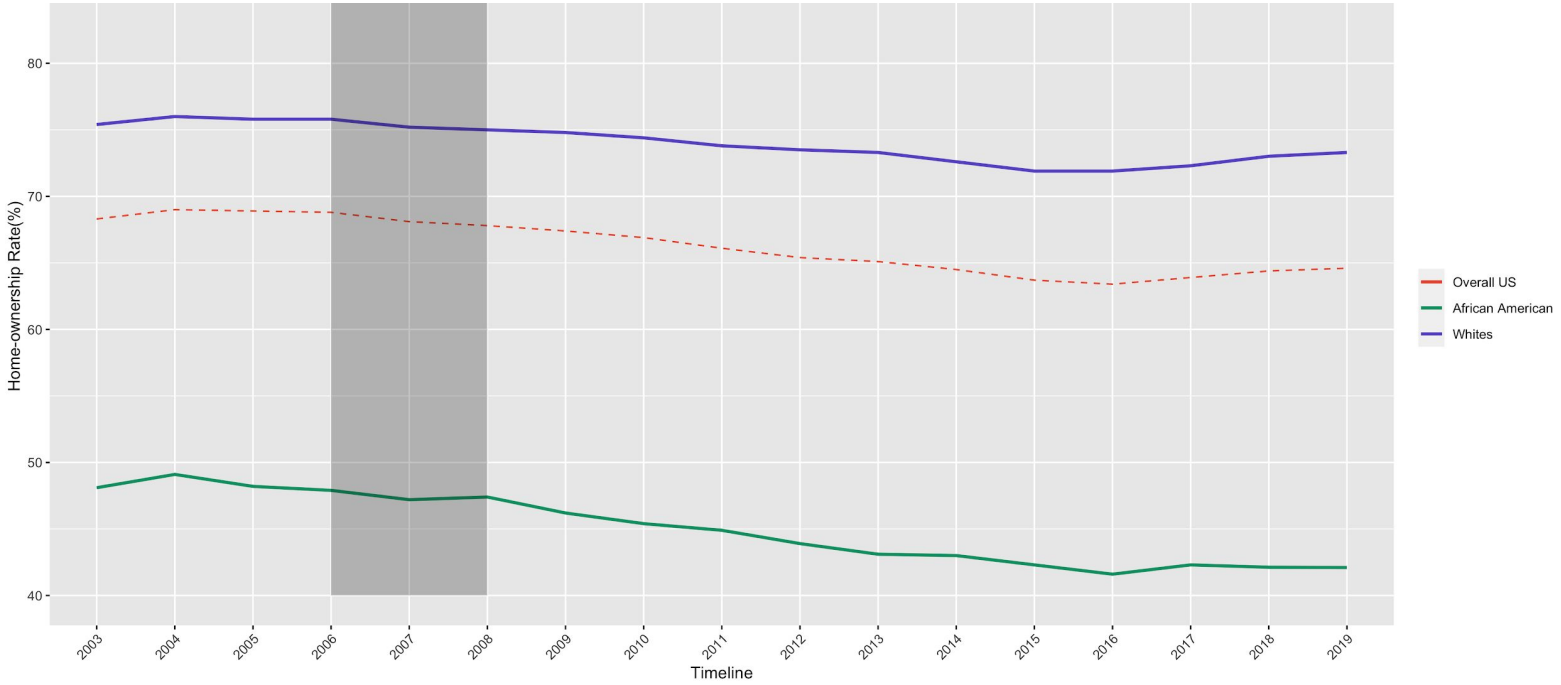
# Homeownership Share: 2003-2019

Homeownership Share for Overall US population and Minorities

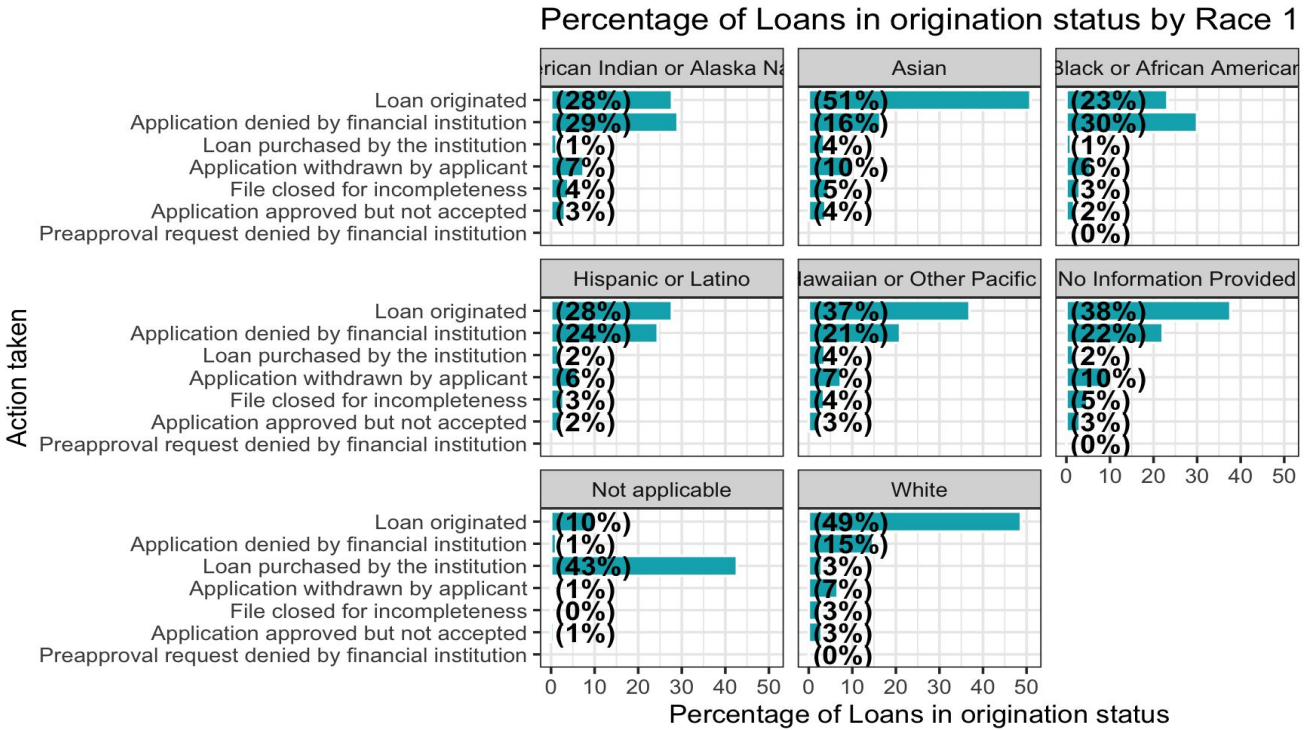


# How it compares between Whites and African Americans

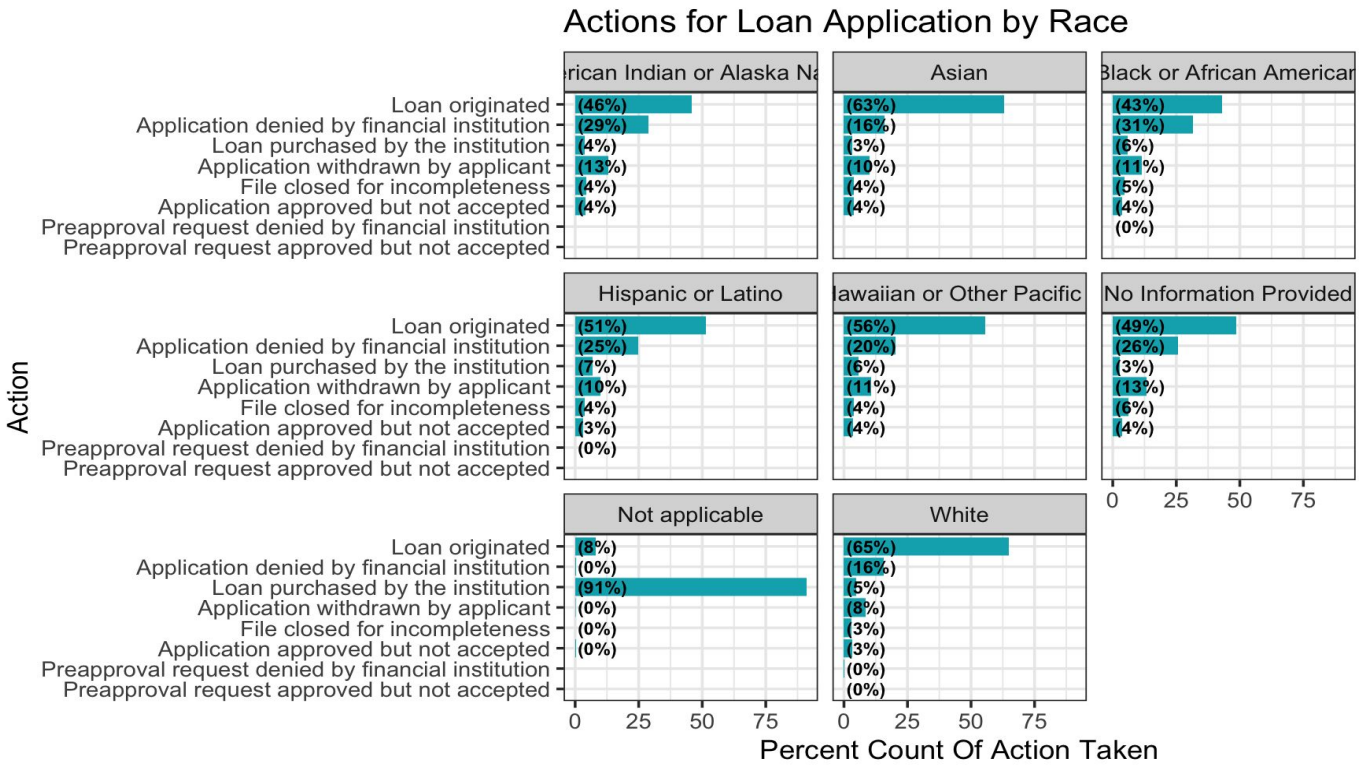
Homeownership Share Compared between Whites and African Americans



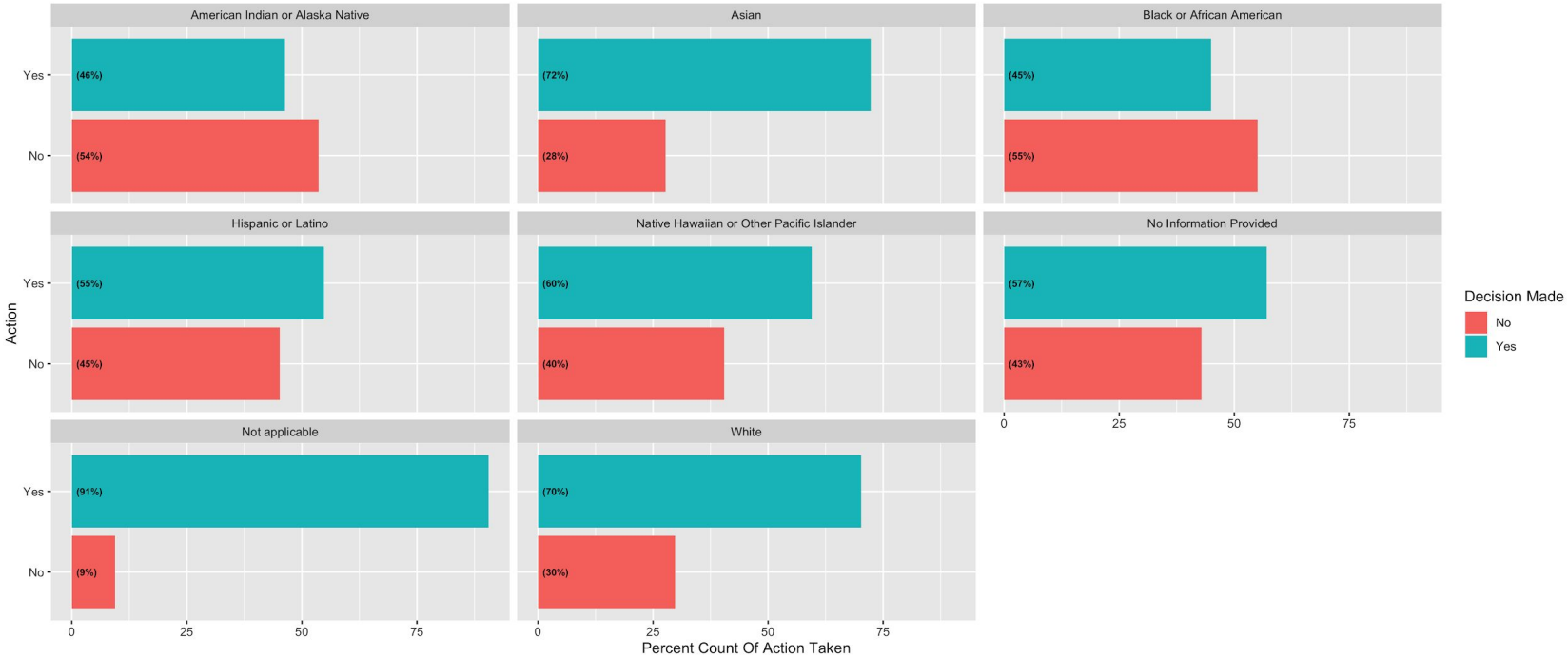
# Loan Action Type distribution PA 2014



# Loan Action Type distribution IL 2017



# Decision Distribution according to Race and Ethnicity



# Data Cleaning and Feature Engineering

# Data Cleaning

**Drop**

- Irrelevant features
- Features with minimal information
- Redundant Features
- Rows with most of the values missing.

**Replace**

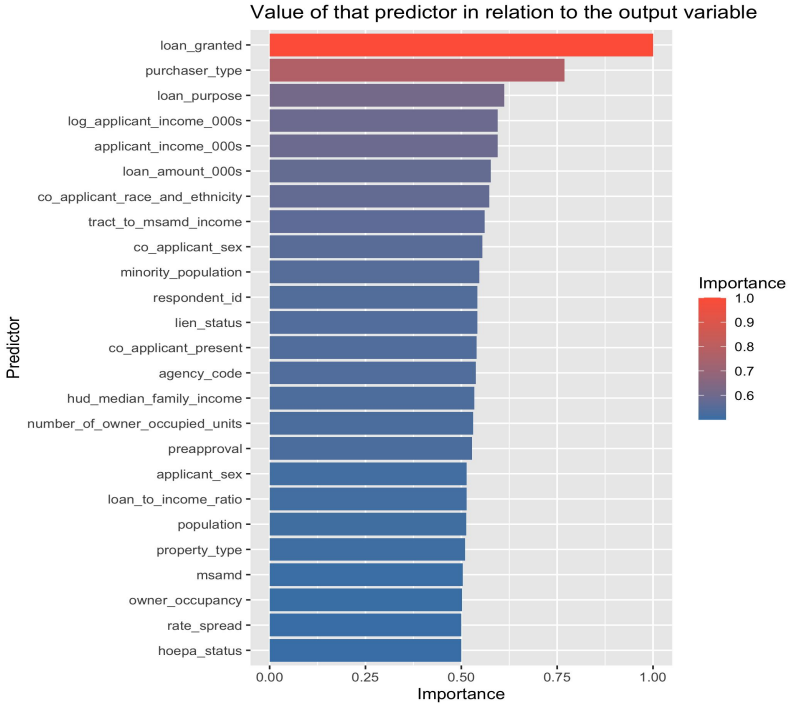
- By Median
- By Mean

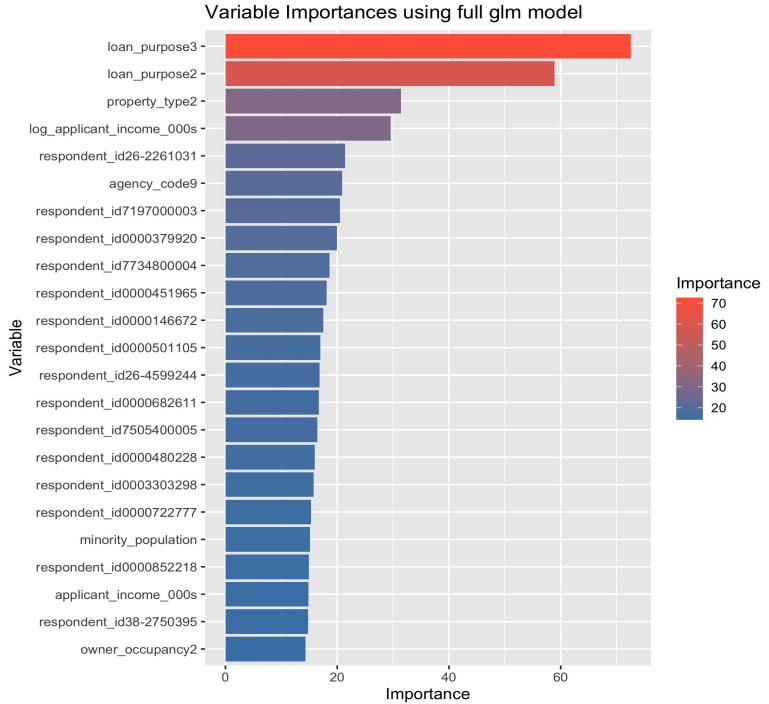
**Impute**

- Using Mice

# Feature Engineering

## Identify most relevant features.



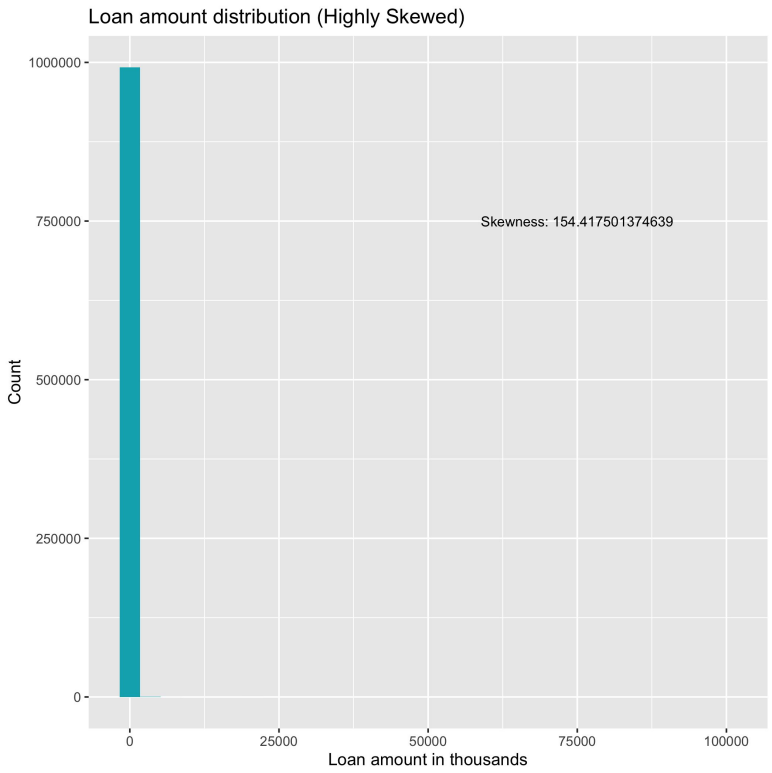



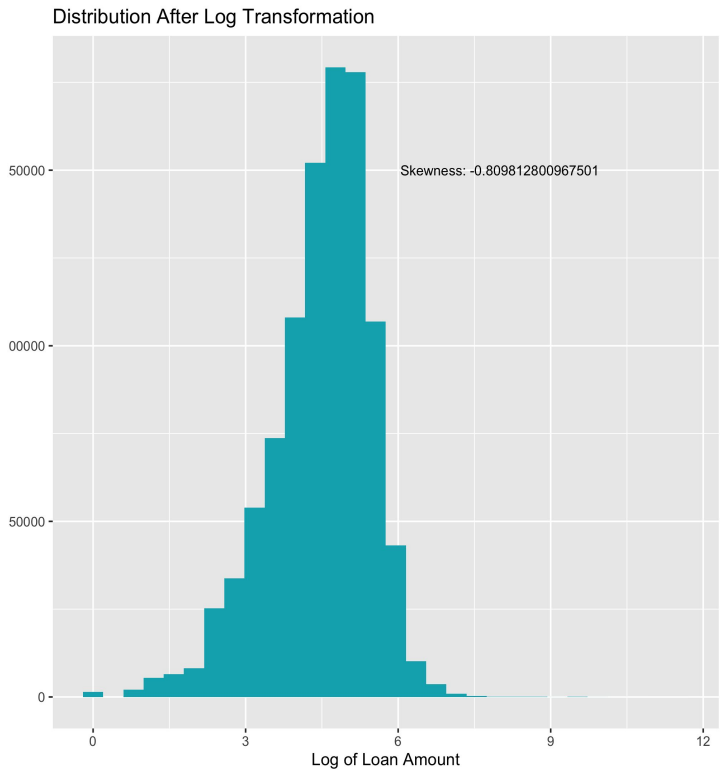
# Feature Engineering

**New features**

- Add new features based upon secondary data (Census, Zillow Dataset)
  - Median House value per county
- Create new features from existing predictors.
  - Loan to income Ratio
- Transform highly skewed predictors
  - Applicant income
  - Loan Amount

# Log Transformation



# Data Modelling

# Approaches

- Naive Bayes
- Decision Trees
- Random Forest
- Logistic Regression

# Performance Metrics

- **Precision**
  - Minimize the risk

- **Recall**
  - Maximize profit

- **F1 Score**
  - Weighted average of Precision and Recall.

# Naive Bayes

- **Why?**

- Easy to train and implement
- Less model complexity
- Can be used as a benchmark model.

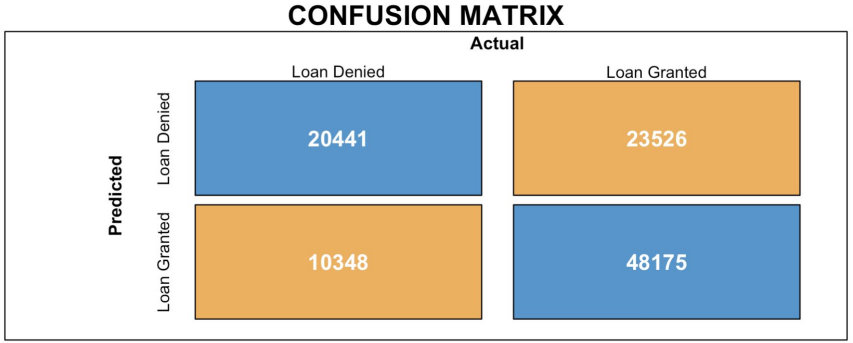
# Naive Bayes: Model Performance

**PA 2014-15**

- **How does it perform?**

- Less Accuracy

- Decent Precision but very low Recall




**DETAILS**

| Sensitivity | Specificity | Precision | Recall | F1 |
|---|---|---|---|---|
| 0.672 | 0.664 | 0.823 | 0.672 | 0.74 |

| Accuracy | Kappa |
|---|---|
| 0.669 | 0.299 |

# Decision Tree

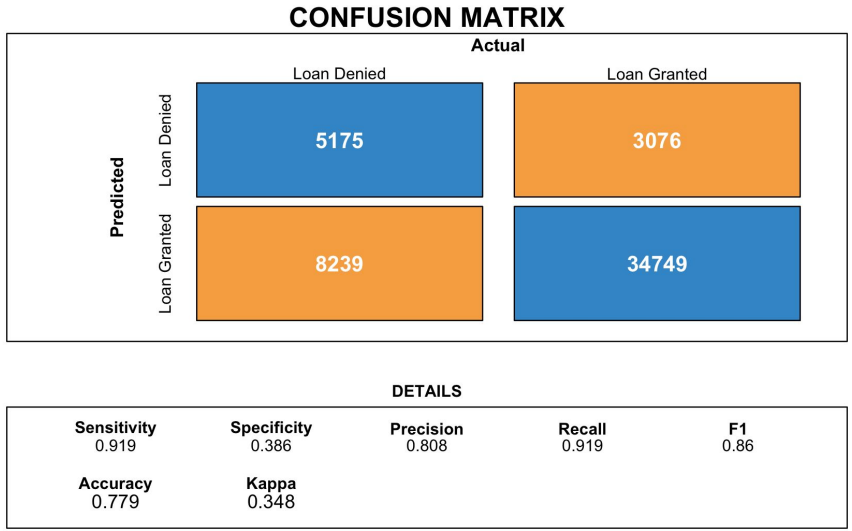
- **Why?**

- Non- Parametric Model
- Can get the probabilities for each class
- Feature selection happens based on information gain
- Closely mimics human decision making process.

# Decision Tree: Model Performance

**PA 2014**

- **Better than Naive Bayes?**
  - Accuracy improved by 11%
  - Precision and Recall values are also
- **Hyperparameter Tuning?**
  - Used grid search to find the best hyperparameters

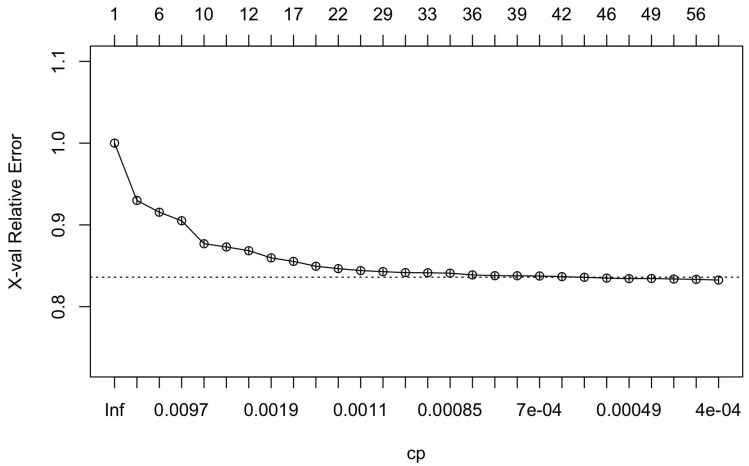


# Decision Tree: Hyperparameters

Method Used: **Grid Search**

Values that optimized model architecture

- Minimum Split: 14
- CP: 0.0007
- Maximum Depth: 20



# Random Forest

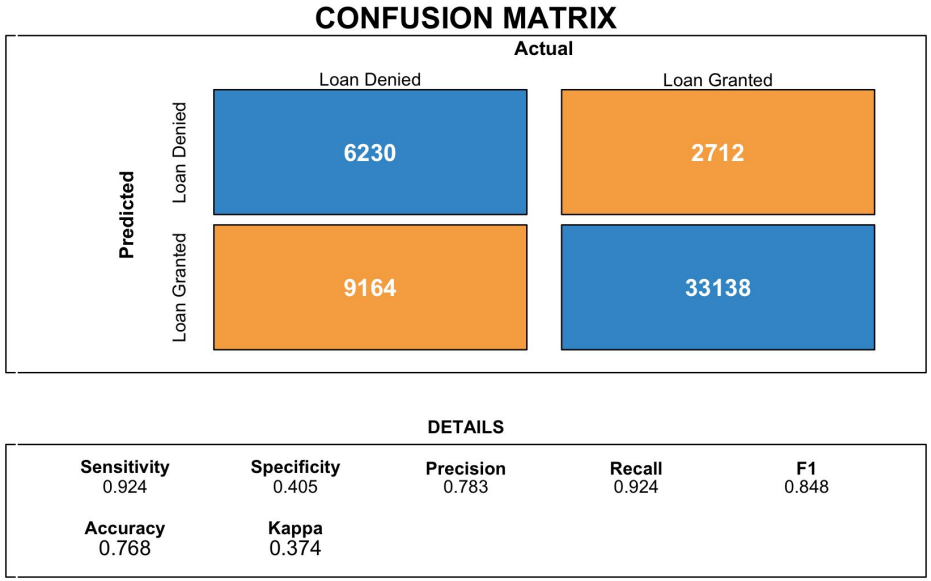
- **Why?**

- Built by growing multiple trees in the model.
- Boasts of Improved Accuracy
- Doesn’t allow overfitting
- Can handle highly-dimensional data

# Random Forest: Model Performance

**PA 2014**

- **Better than Decision Tree?**
  - Does not perform any better
  - Recall value goes up
  - Precision is very low.



# Logistic Regression

- **Why?**

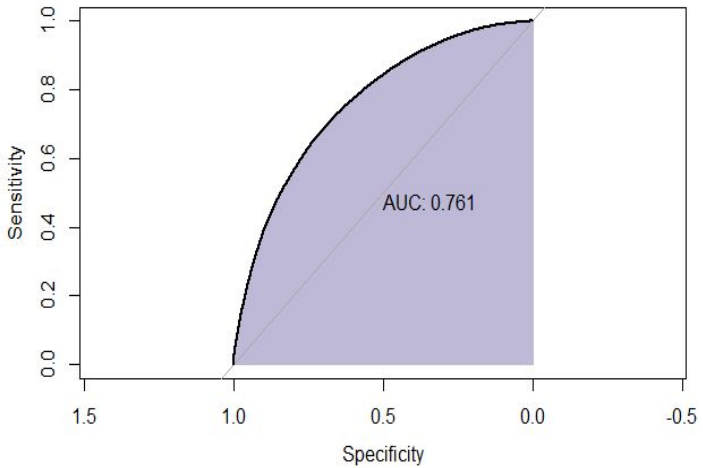
- Can interpret each variable independently
- Easy to implement
- Get odds ratio per coefficient (predictor).
- Can get probability of each observation. So each applicant can be interpreted individually.

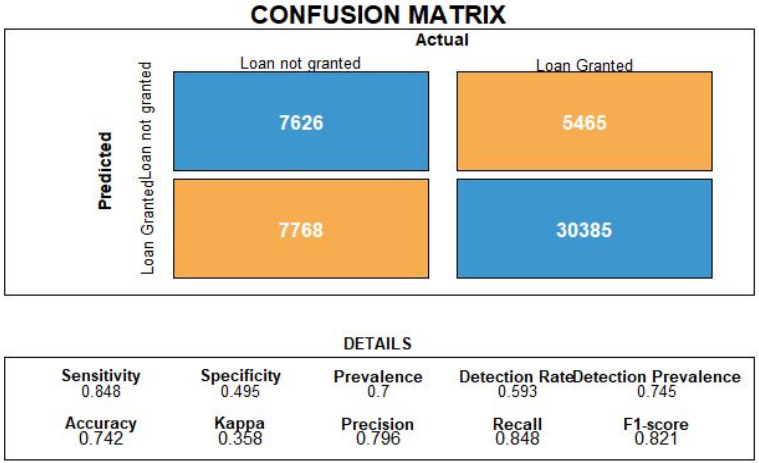
# Logistic regression best models

- Log transformed loan amount and applicant income
- Log transformed loan amount and loan to income ratio
- HMDA data merged with census
- HMDA data with median home values.
- Optimizing for precision along with other metrics.

# Model performance for PA for 2014

A



# Coefficients for race and ethnicity PA 2014

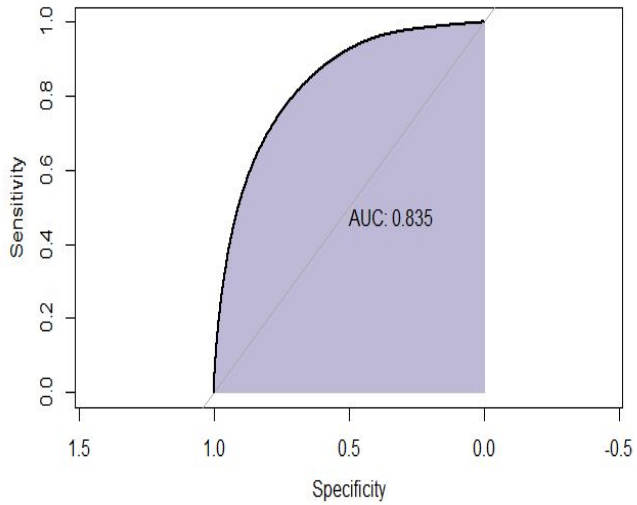
| Coefficients | |
|---|---|
| Asian | 0.4509005 |
| Black or African American | 0.1727634 |
| Hispanic or Latino | 0.1865617 |
| White | 0.7282939 |
| Native Hawaiian | 0.2847779 |
| No Information Provided | 0.3617244 |
| Not applicable | 0.5592027 |

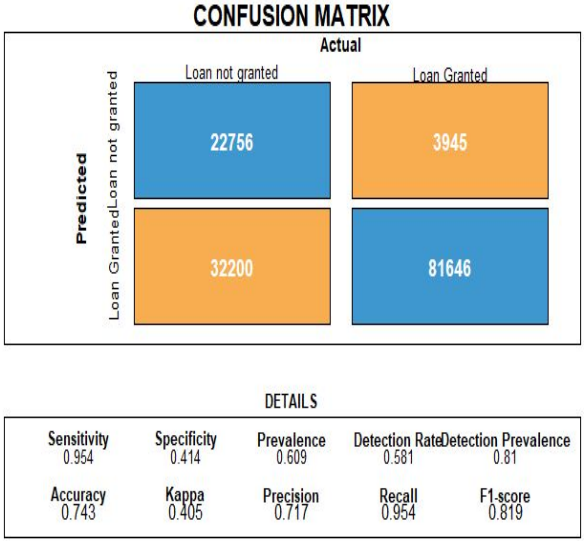
**P values for predictors**

| | |
|---|---|
| applicant_race_and_ethnicity | 0.00000000000000022 *** |
| owner_occupancy | 0.00000002049 *** |
| preapproval | 0.00000000000000022 *** |
| log_loan_amount_000s | 0.00000000000000022 *** |
| applicant_sex | 0.00000000000000022 *** |
| county_code | 0.00000000000000022 *** |
| tract_to_msamd_income | 0.00000000000000022 *** |
| co_applicant_present | 0.00000000000000022 *** |
| agency_code | 0.00000000000000022 *** |
| minority_population | 0.00000000000000022 *** |
| loan_purpose | 0.00000000000000022 *** |
| rate_spread | 0.00000000000000022 *** |
| property_type | 0.00000000000000022 *** |
| loan_to_income_ratio | 0.00000000000000022 *** |
| respondent_id | 0.00000000000000022 *** |

# Comparison with 2007/2008

AUC curve



# Logistic regression key findings for PA

- Odds of getting a loan for African Americans range from 18.9% to 29%
- For White's it ranges from 107% to 150%
- For Asian's it ranges from 56% to 101%
- For Hispanics it ranges from 20% to 46%

# Logistic regression key findings for IL

- Odds of getting a loan for minorities including Hispanics, African Americans, etc is very low (30% to 50%)
- For Whites it is 102%

# Model Deployment

# Takeaways

- The Equal Credit Opportunity Act (ECOA 1974)  prohibits race based discrimination.
- Data is telling a different story
- Models show some evidence of bias against minorities
- Factors like debt to income ratio, credit scores, education level, bank balance, loan to value ratio, etc are missing in HMDA data
- Models can be improved with the above data

# Recommendations

- Loan decisions should be based on financial standing alone
- Loan decision models have societal impact. Objective functions cannot be black and white (profit and loss, for e.g.)
- Applicants need visibility into the decision process
- HMDA reporting requirements should be strengthened
- Freedom of information act should apply to value-added models.

# Questions