Capstone project:
The biasness test and a self-checker on credit score for housing loan

Chalida Thangpetchr

01

Introduction

Motivation / Problem statements / Data source

Credit scores should be a result of individual's credential and not affected by their circumstances

Circumstance

- Age
- Race
- Ethnic origin
- Gender
- Location of birth/residence

Credential

- Payment history (35%)
- Amounts owed (30%)
- Length of credit history (15%)
- Credit mix (10%)
- New credit (10%)

Note: US FICO credit scoring system are calculated from this 5 categories in credential table

Problem statements

01

Main: Credit scores test for biasness

Does minority group get lower credit rating despite having the same credential when apply for the housing loan?

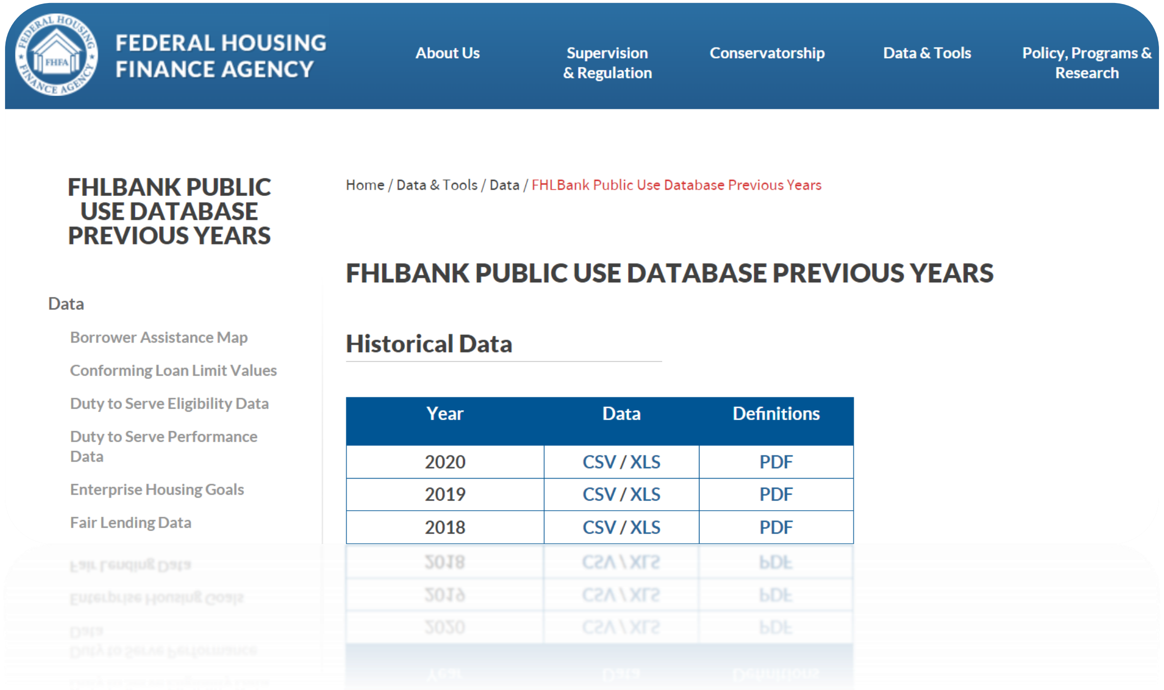
02

Secondary: Credit scores checker

Provide a tool for home buyer to personally check their credit rating based on their credential.

Source of data

- **The USA's federal housing finance agency (FHFA) loan-level Public Use Databases (PUDBs):** loan-level data on mortgages purchased by Fannie Mae and Freddie Mac, including borrower income, race, and gender.

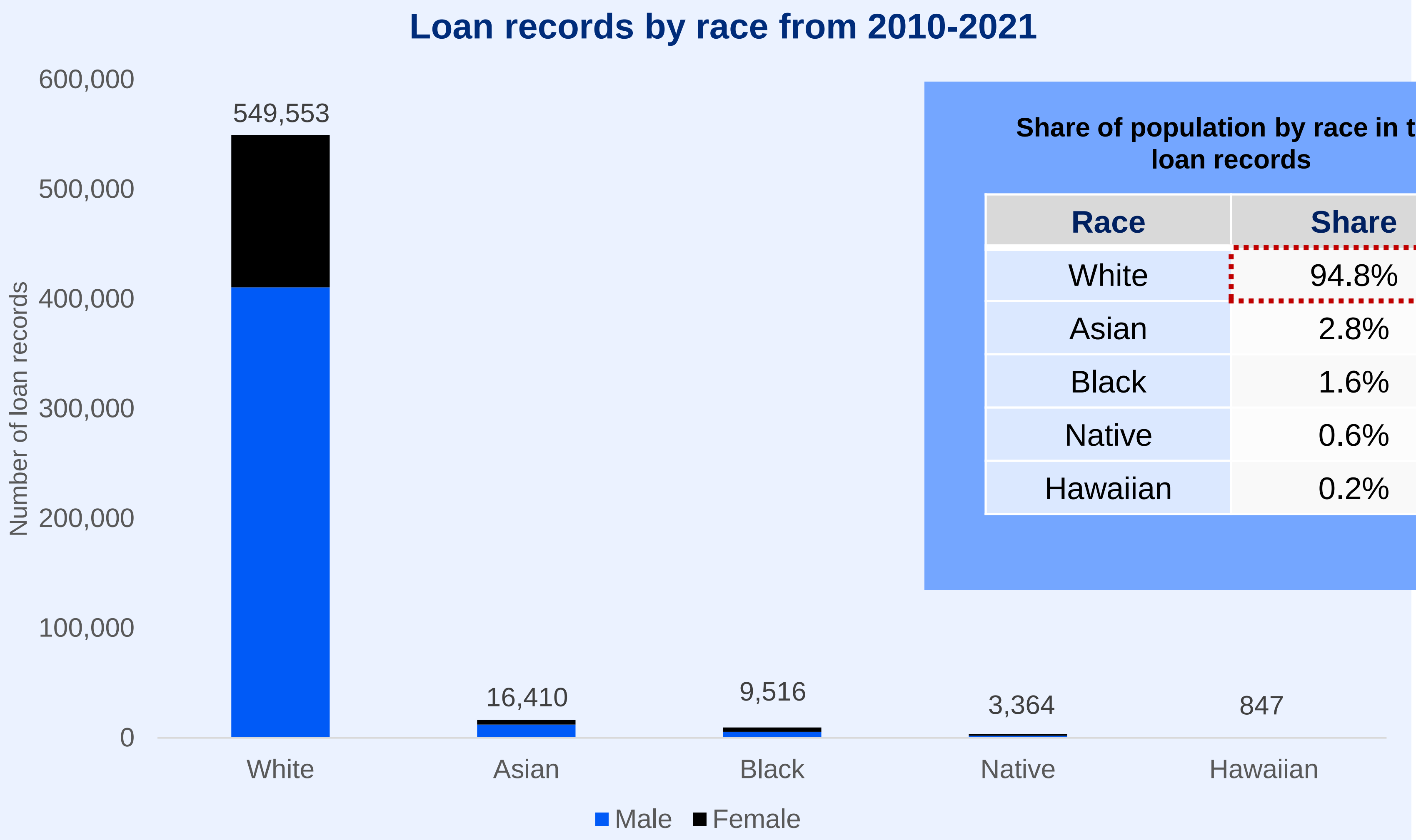


- **Dataset has 56 columns and 693,331 borrowers' records from 2010 – 2021**

02

Exploratory Data Analysis (EDA)

The proportion minorities in housing loan data set are much less than the minority ratio in the actual population



Note: The number on top of the column is the total Male and Female records

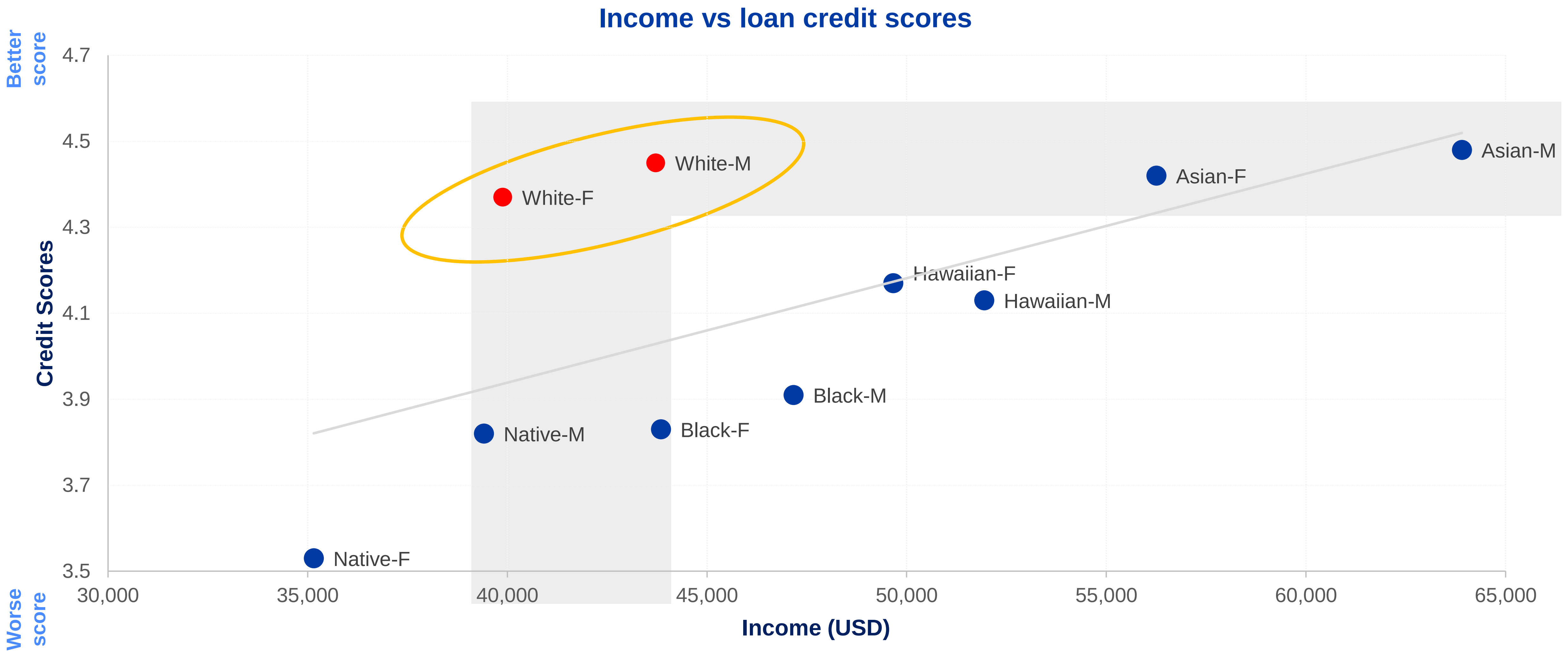
Share of population by race in the loan records

Race	Share
White	94.8%
Asian	2.8%
Black	1.6%
Native	0.6%
Hawaiian	0.2%

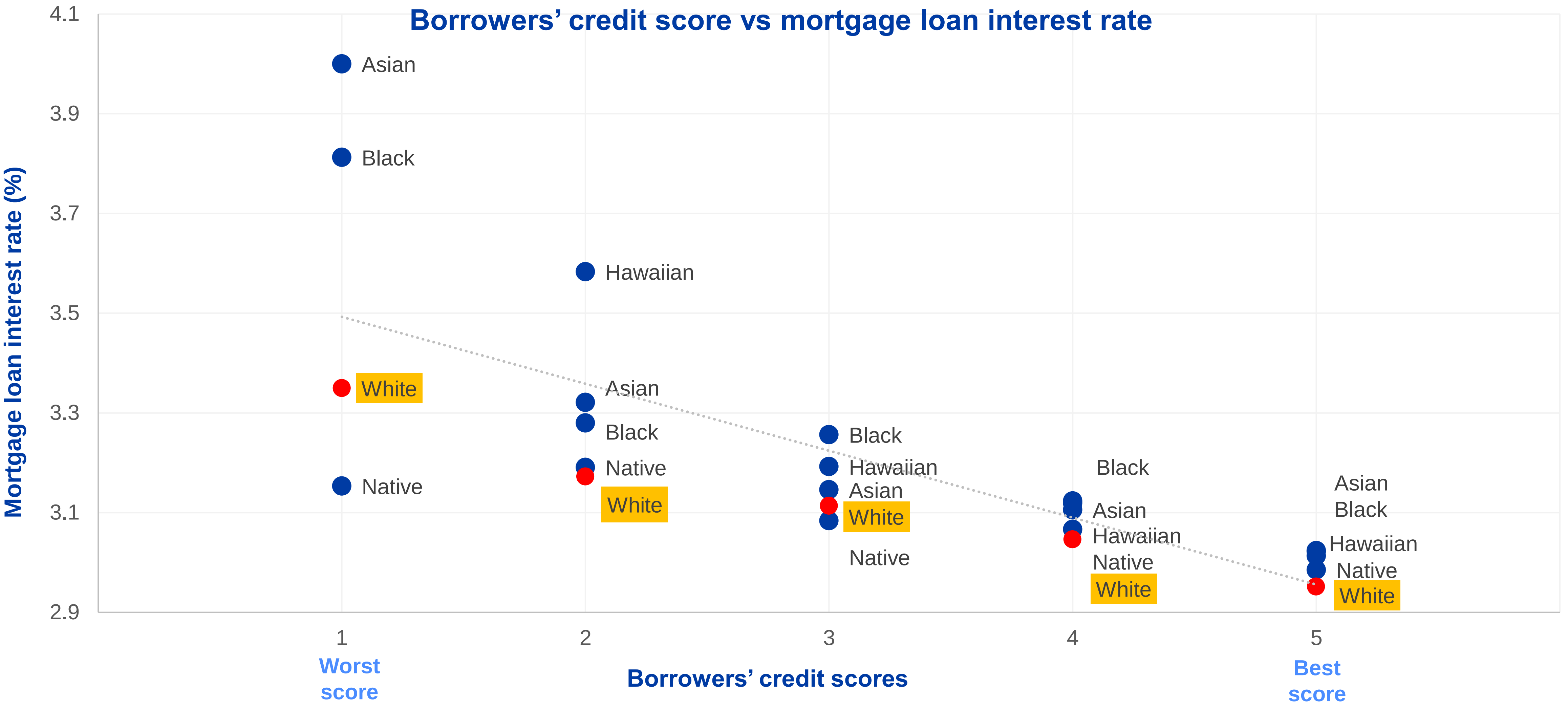
Share of the actual population by race in 2021

Race	Share
White	75.8%
Asian	6.1%
Black	13.6%
Native	1.3%
Hawaiian	0.3%
Mixed race	2.9%

White borrowers tend to have higher credit scores as compared to minorities despite having same range of income.



White borrowers tends to get more favorable mortgage loan interest rate despite having the same credit score



03

Methodology modeling and result

Problem statements

01

Main: Credit scores test for biasness

Does minority group get lower credit rating despite having the same credential when apply for the housing loan?

02

Secondary: Credit scores checker

Provide a tool for home buyer to personally check their credit rating based on their credential.

Race and gender are used to created different groups of circumstances

Circumstance

- Age
- Race
- Ethnic origin
- Gender
- Location of birth/residence

FHFA public data on borrowers

Borrower's **age**

Race: Native / Asian / Black / Hawaiian / White

Ethnicity: Hispanic or Latino / none

Gender: male / female

Location minority ratio

Local area **median income**

A collection of proxy variables indirectly related to the credit scores are used to test for biasness

Credential

- Payment history (35%)
- Amounts owed (30%)
- Length of credit history (15%)
- Credit mix (10%)
- New credit (10%)

Income per borrowers

- Indirectly effect payment history

Mortgage loan at origination

Unpaid loan balance (UPB)

- Indirectly linked to payment history
- Proxy for amounts owed

Housing payment to income ratio

Debt payment to income ratio

- Determine servicing ability and indirectly linked to payment history

Borrower First Time Home buyer

- Proxy for credit mix/amount owed

Credit score (S_{nm}) should be a function of buyer's credential to get a loan (L_n) but not circumstances (C_n)

$$L_n = \textit{loan credential cluster}$$
$$C_n = \textit{groups of individuals sharing the same circumstances}$$
$$S_{nm} = \textit{expected value of individual credit scoring}$$

For example

- C_1 White male
- C_2 Black female

	L_1	L_2	L_3	…	L_m
C_1	S_{11}	S_{12}	S_{13}	…	5
C_2	2	3	4	…	4
C_3	X_{31}	X_{32}	X_{33}	…	5
…	…	…	…	…	…
C_n	S_{n1}	S_{n2}	S_{n3}	…	S_{nm}

For example

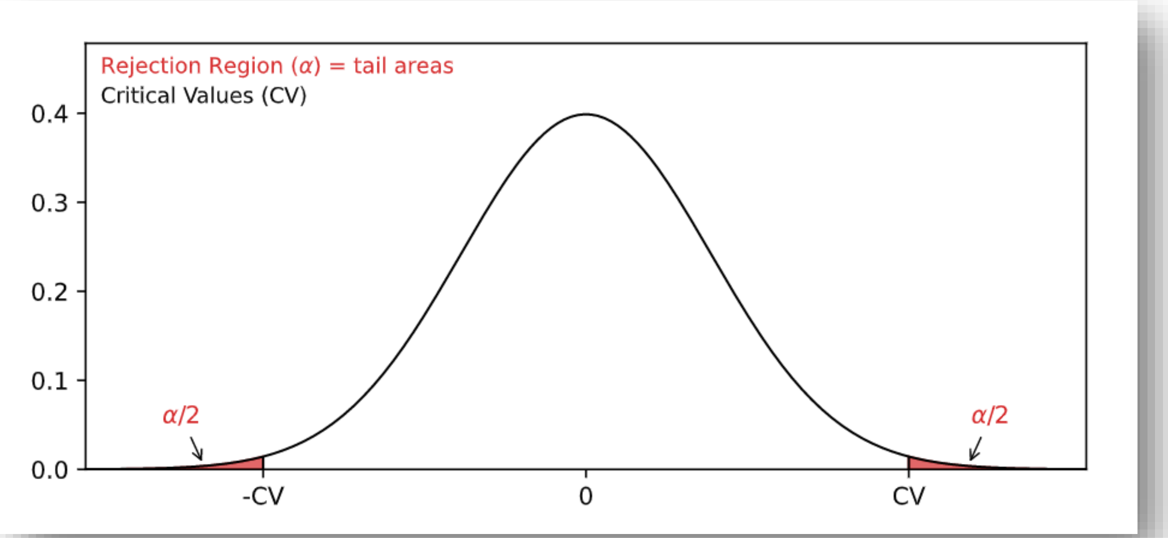
- L_1 Individual with no debt position
- L_2 Individual with 50% Debt to total income ratio

Clusters are created and tested for inequality within the same clusters

1) Use K-mean cluster machine learning to get 30 clusters of similar loan credentials
2) Inequality measure (Mean Log Deviation : MLD) is calculated for each cluster
3) If there is a credit score biasness, MLD should be different than zero (H_1: $\mu \neq 0$)

	L_1	L_2	…	L_m
C_1	S_{11}	S_{12}	…	5
C_2	2	3	…	4
C_3	X_{31}	X_{32}	…	5
…	…	…	…	…
C_n	S_{n1}	S_{n2}	…	S_{nm}

MLD_1 MLD_2 MLD_3 MLD_m



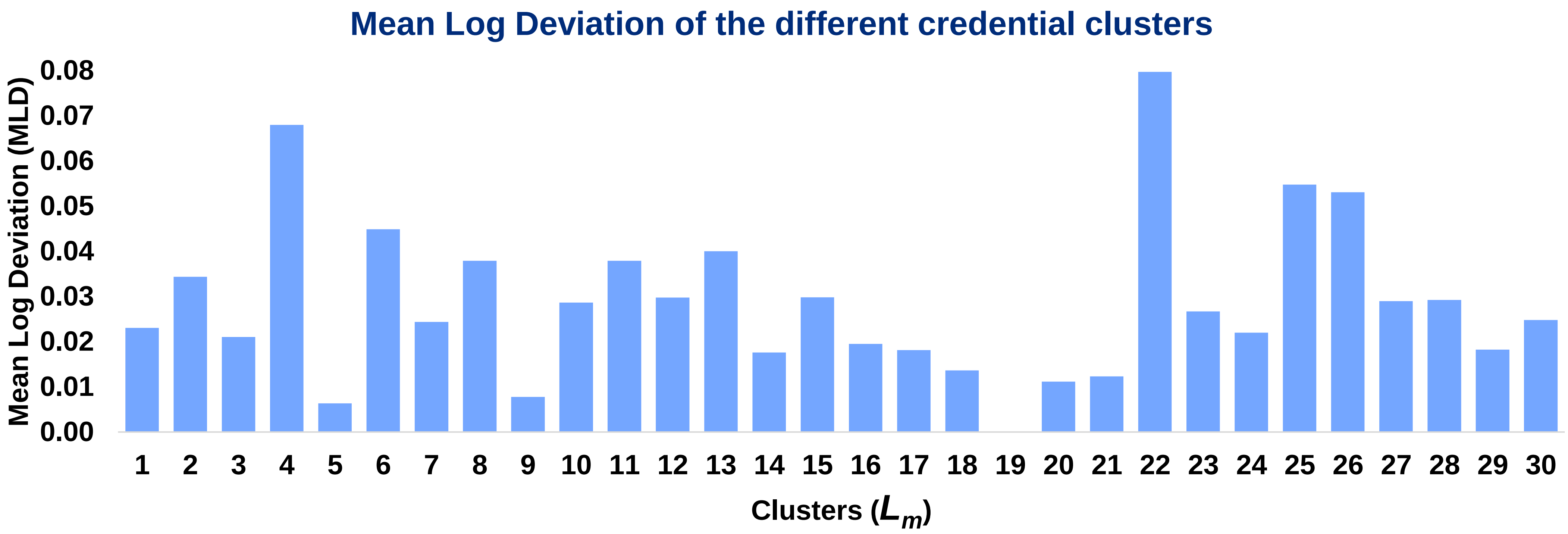
Set up hypothesis testing for MLD distribution

H_0: $\mu = 0$

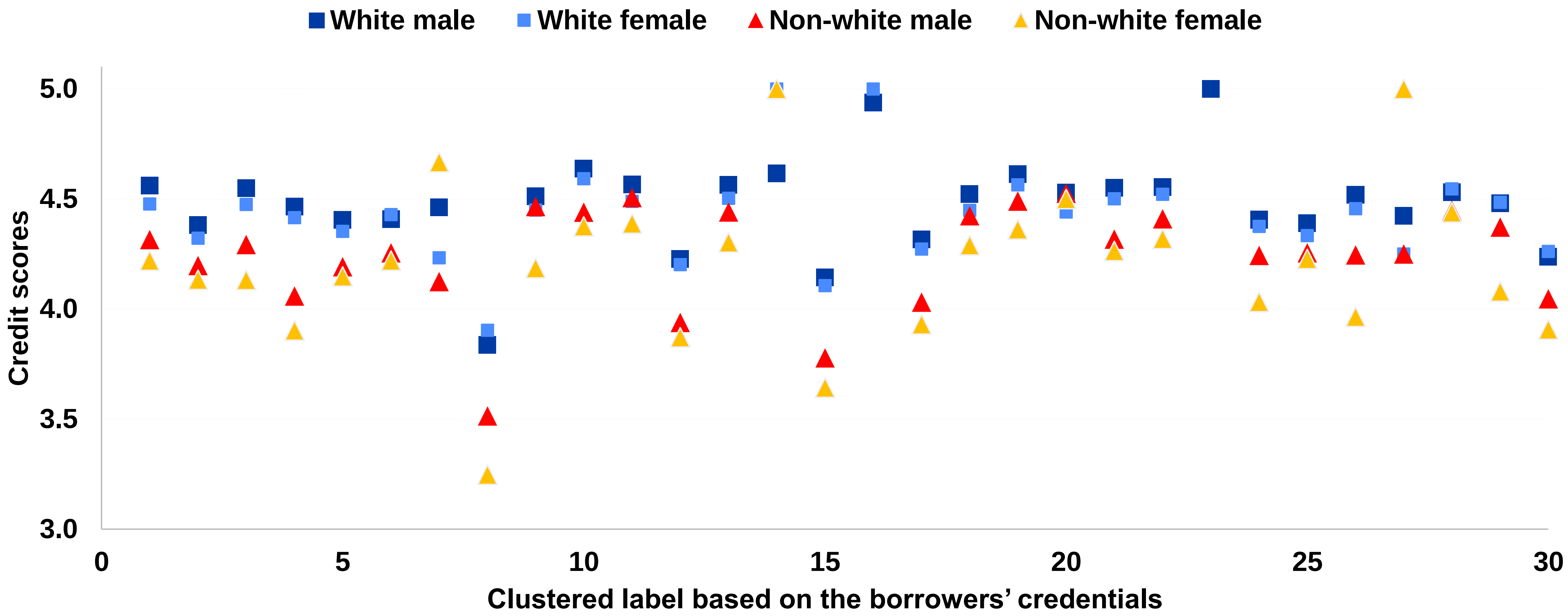
H_1: $\mu \neq 0$

Result: There is a biasness for borrowers within the same cluster of credential measured by Mean Log Deviation (MLD)

- K-mean model presented silhouette score of 0.5, this score measures how separate and cohesive our clusters are with the score range of -1 to 1
- The Mean Log Deviations are significantly different than zero with the p-value of 0.00
- However, the MLD values are considered as very low inequality



Racial minority borrowers in the same credential cluster as the others generally have the lower credit score



Problem statements

01 Main: Credit scores test for biasness

Does minority group get lower credit rating despite having the same credential when apply for the housing loan?

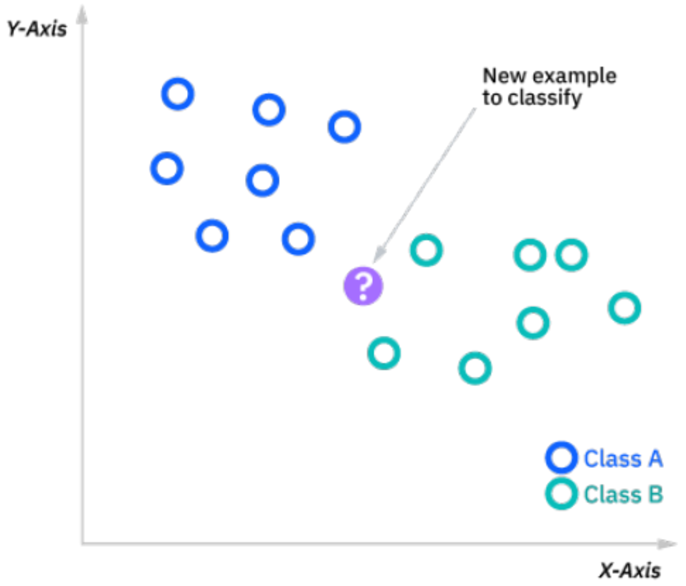
02 Secondary: Credit scores checker

Provide a tool for home buyer to personally check their credit rating based on their credential.

Three popular classification model are used to classify credit scores with the same independent variables in clustering model

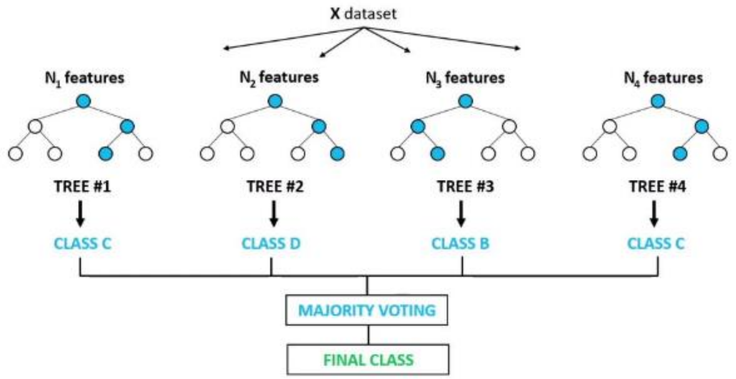
K-nearest neighbors

Estimating the likelihood that a data point will be its member based on the nearest points



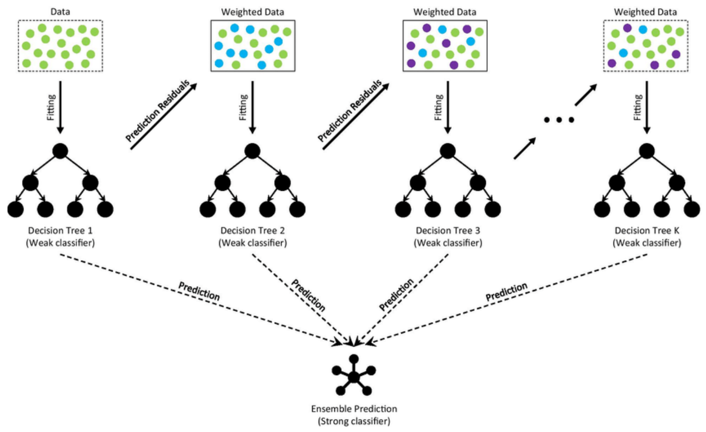
Random forest

Building decision trees on different samples and takes their majority vote for classification



Gradient Boosting

Combining several weak learning models to produce a powerful predicting model



Result: Classification model result slightly improved after using Random Forest/Gradient Boosting

Model used	Variable used	Train accuracy	Test accuracy	Differences
Baseline model KNN	**Income / loan amount/ unpaid loan / mortgage to income / debt payment to income / first owner**	**60.4%**	**54.7%**	**-5.7%**
Model 1 Random Forest	**Income / loan amount / first time buyer**	**59.6%**	**59.1%**	**-0.5%**
Model 2 Gradient Boosting	**Income / loan amount/ unpaid loan / mortgage to income / debt payment to income / first owner**	**59.8%**	**59.7%**	**-0.1%**

- Based on one of the classification model, the demo app had been created so home buyer can try keying in their credential and check their credit scores through this link.

04

Conclustion

Conclusion

- **Main problem statement: There is a biasness for borrowers within the same cluster** of credential measured by Mean Log Deviation (MLD)
- Racial minority borrowers in the same credential cluster as the others generally have the lower credit score
- **Secondary problem statement: Gradient Boosting classifier has improved accuracy from the baseline model in test data,** as well as giving more stable result between test and train data set. And from this model, we're able to create a credit scores self checker demo.

Caveat and future development

- Borrower records in data set are those who had already got housing loan approved. **But biasness can start even before the process** and cause the minorities to be excluded from getting the loan.

- The featured **variables used in this study are just proxies** since FICO credit scores attributes are not publicly available.

- The credit score categories currently used are on the **crude scale of 1-5**, the actual scoring is much wider from 300 - 850.

- **Classification model accuracy can still be improved** through feature engineering, exploring wider range of machine learning models or expansion of data set.

THANK YOU