HMDA Mortgage Analysis and Prediction

ECE143 Group 12

Introduction

1. Importance
   a. Getting Loan is important;
   b. Fairness and Equality (Institutions);
   c. Application strategy (Individual)
2. Data Set
   a. HMDA Mortgage data set for California, 2017
   b. Representative Area & Recent Record; The 1.38 GB csv contains 1.7 million records in total with 78 columns of features.
   c. Well-organized compared with the most recent records (2018~ 2020) which only contains national records with more irrelevant columns .

Methodology

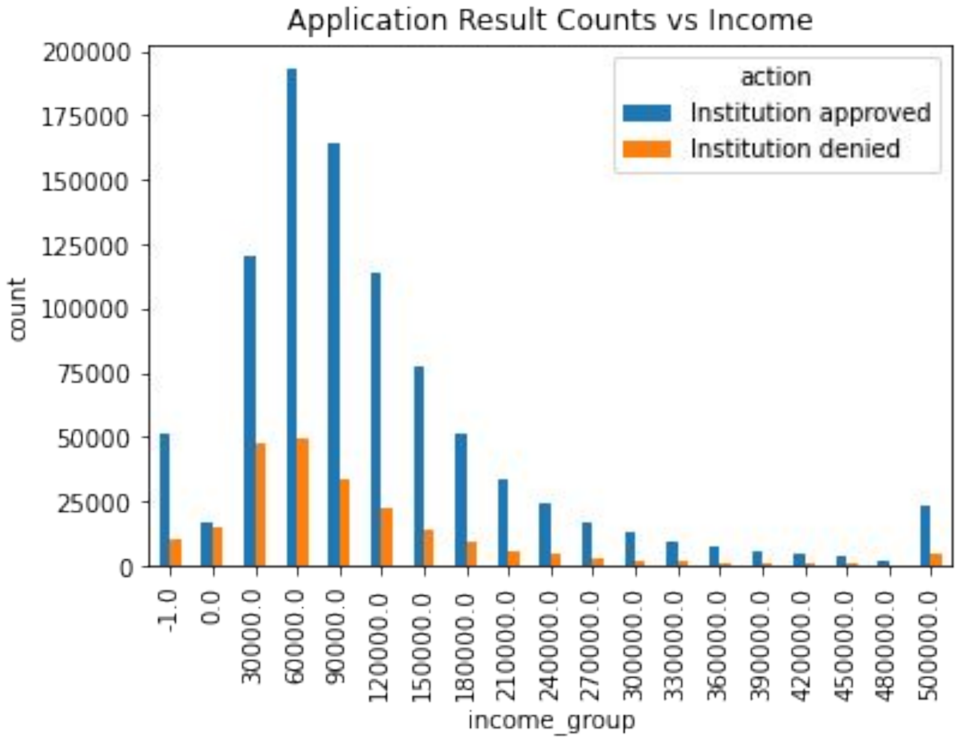
1. Down Scale Data Set/ Data Set Cleaning
    a. Mixed types, Null Values, Too many dimensions, Too many records: Only keep the records that clearly indicate institutions decisions and clean up messy data field.
    b. Result: 186MB new data set with 47 columns and 1.15million records.

2. EDA Important Features (Income, Gender, Race,Ethnicity)
    a. Evaluate each feature’s impact on the decision individually by assuming they are independent.

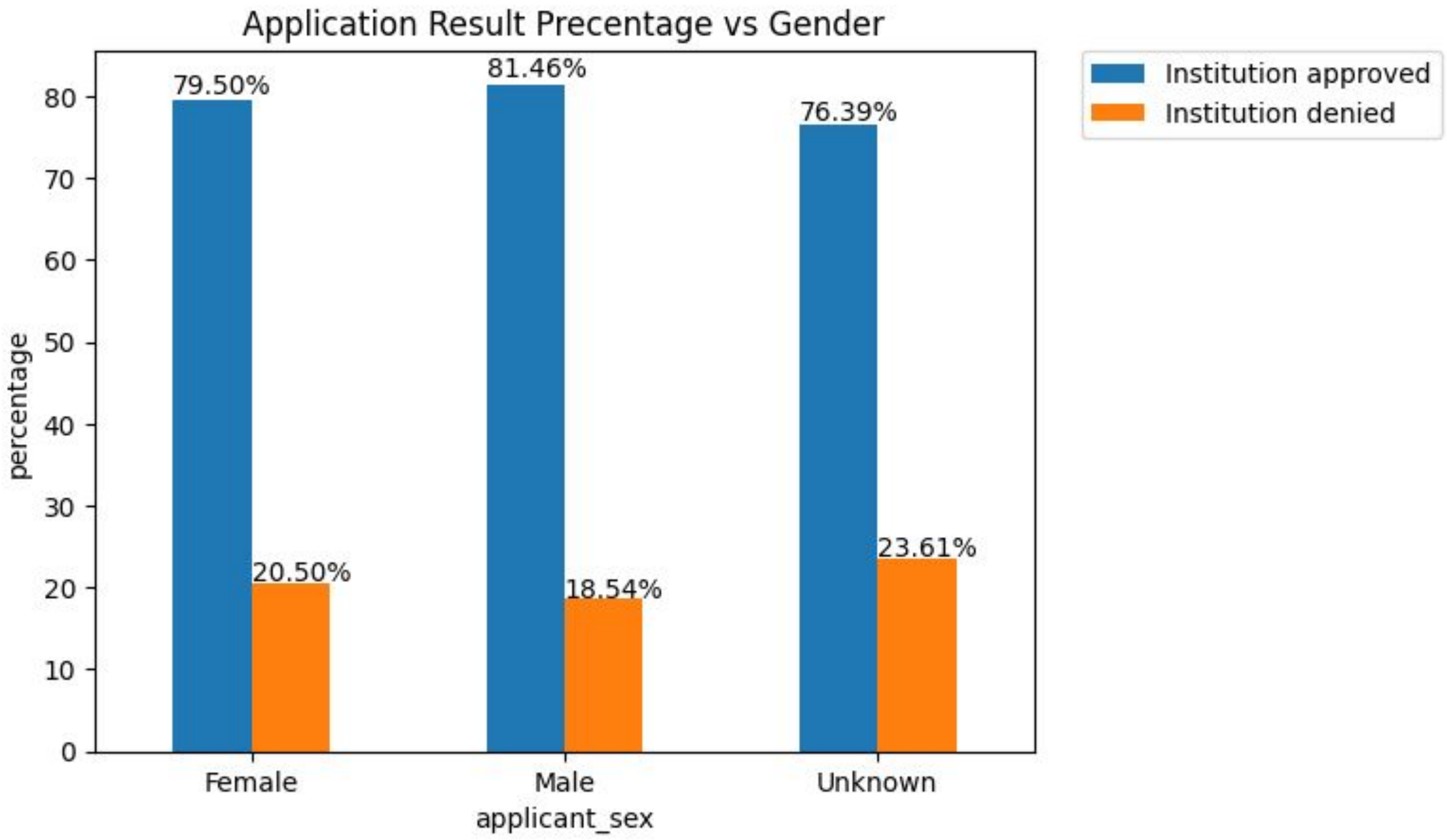
3. Predictive Model For Multiple Features and Better application advice
    a. Find a model to evaluate the influence of those features above combined as well as some other notable features such as loan amount.
    b. Apply the model to give a prediction result based on the certain features applicant have.

Income Analysis

- Huge Income Gap: 30k per bin (-1: Unknown,500k: 500k+ Income).
- Approved cases and denied cases shared the similar trend (right-skewed).
- Applicants whose income <=60k are less likely to be approved.

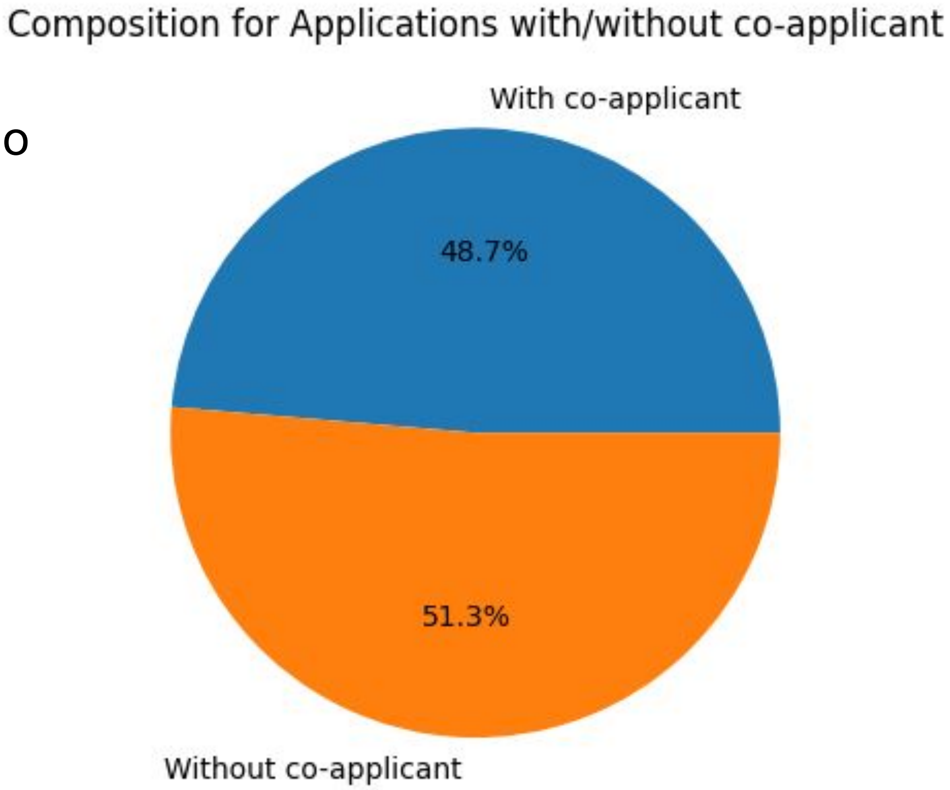


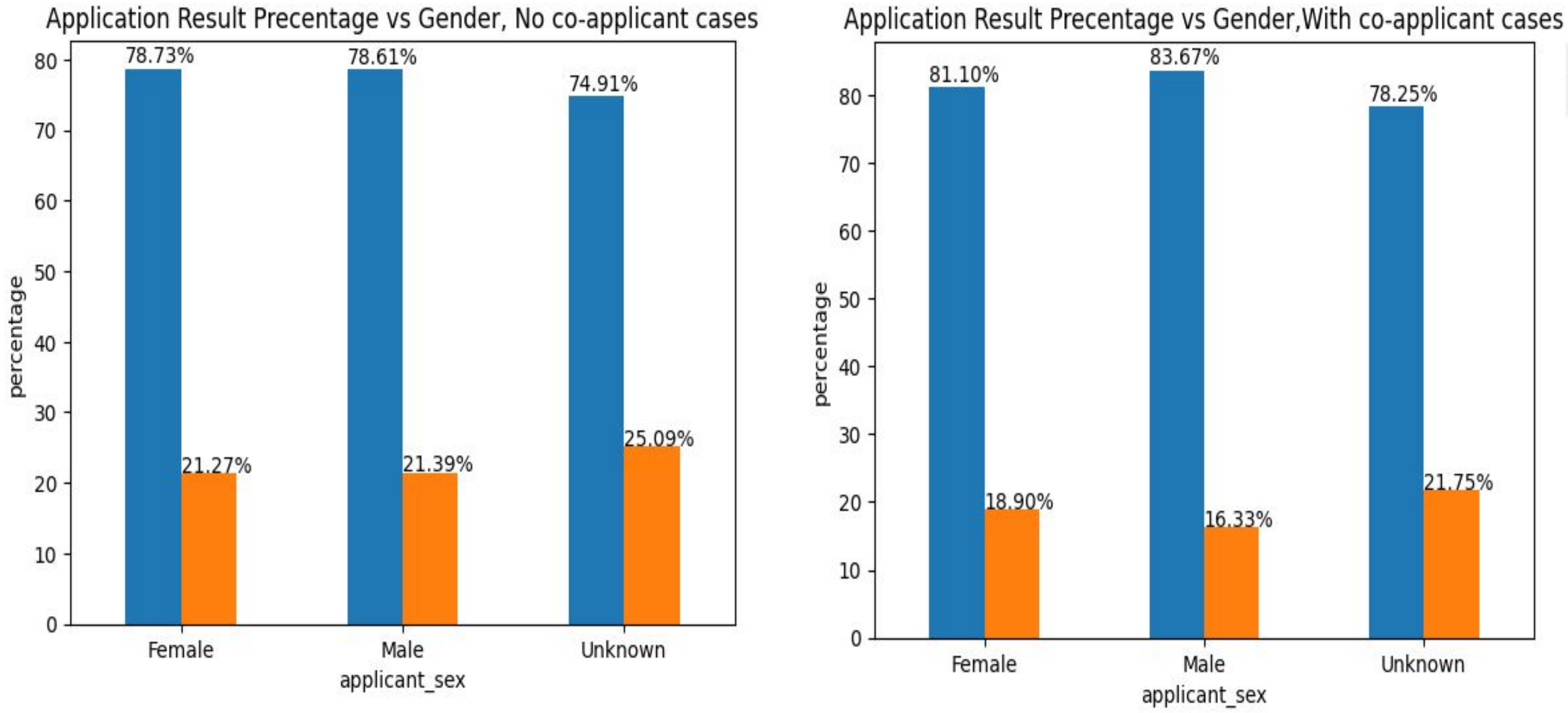
Gender Analysis



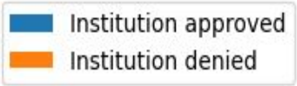
Without co-applicant vs With co-applicants

Applicants have approximately 50:50 ratio in with/without co-applicant

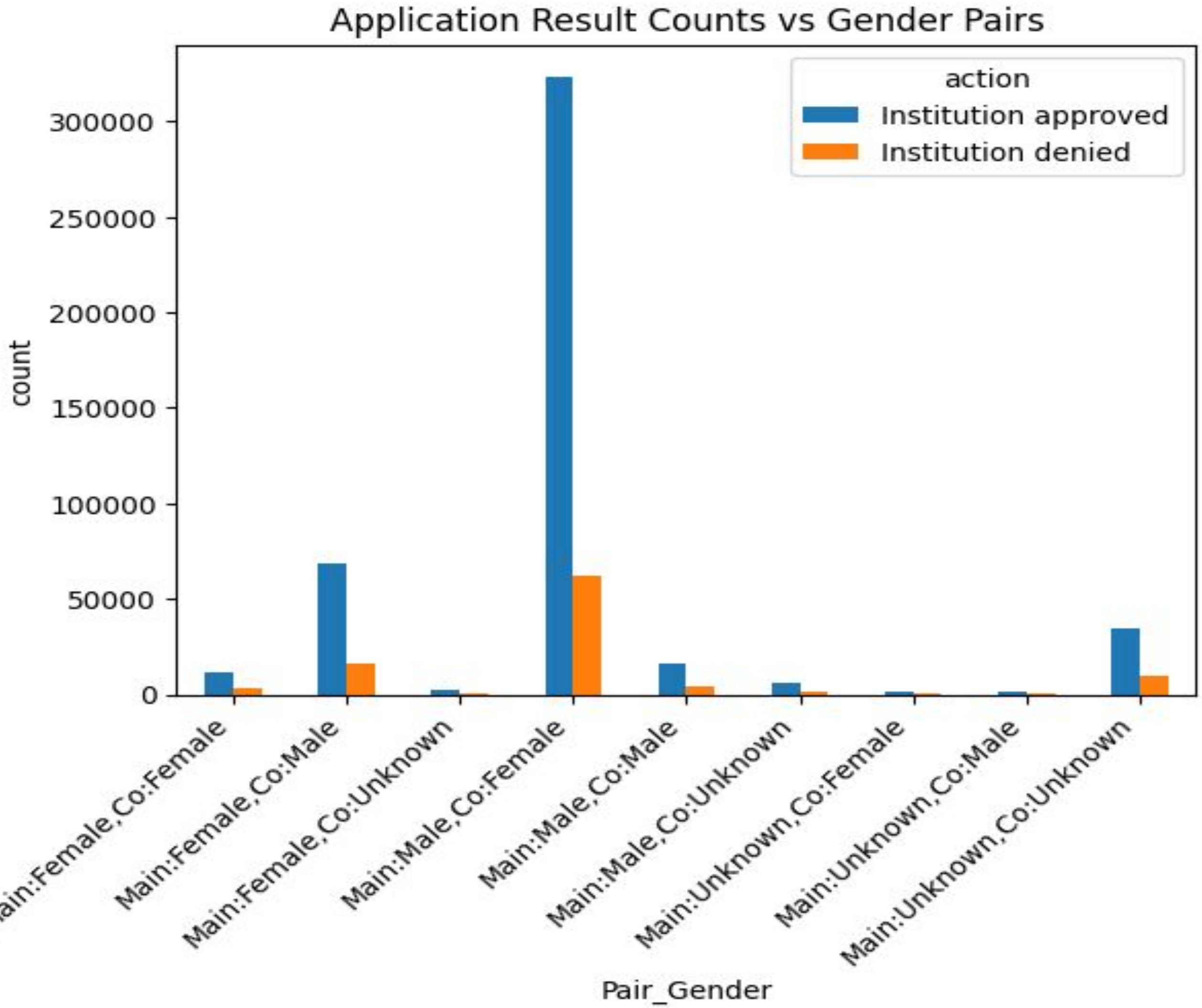




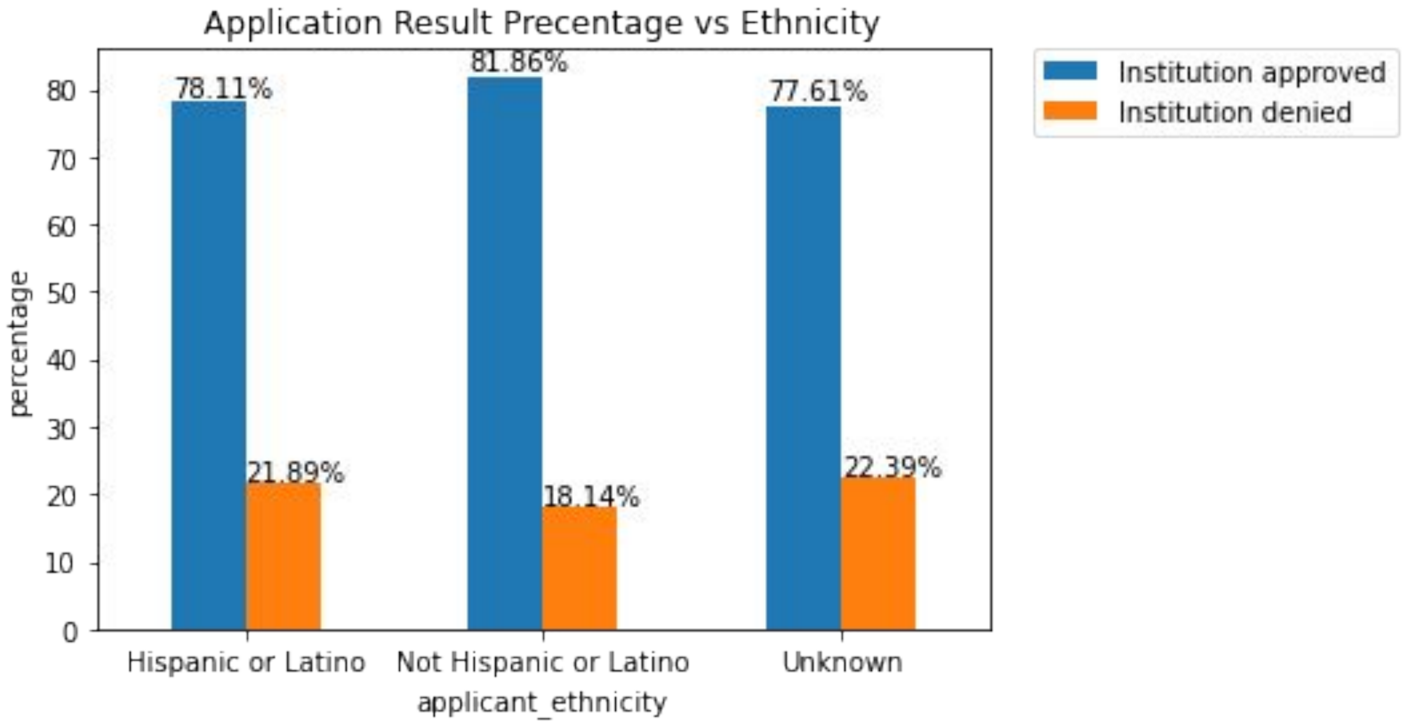
- Applicants tend to get approved if have a co-applicant
- Female applicants perform slightly better if have no co-applicants



Pairs Matter!



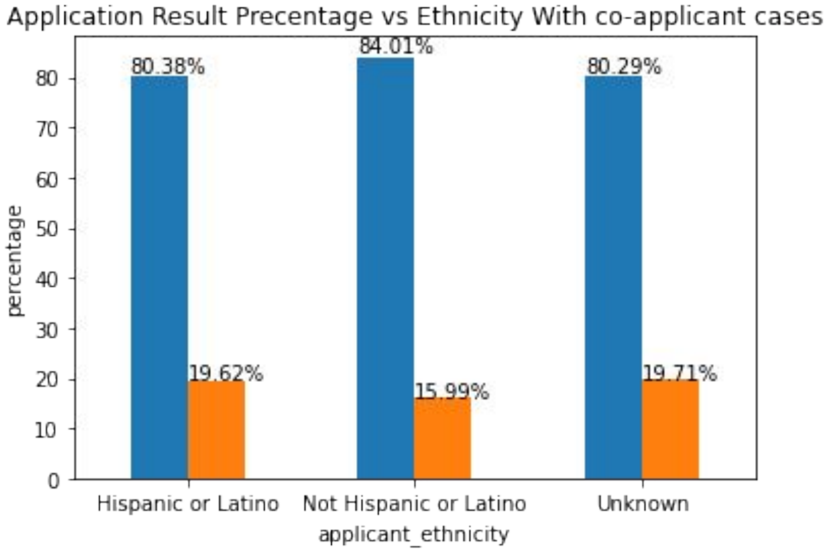
Ethnicity Analysis

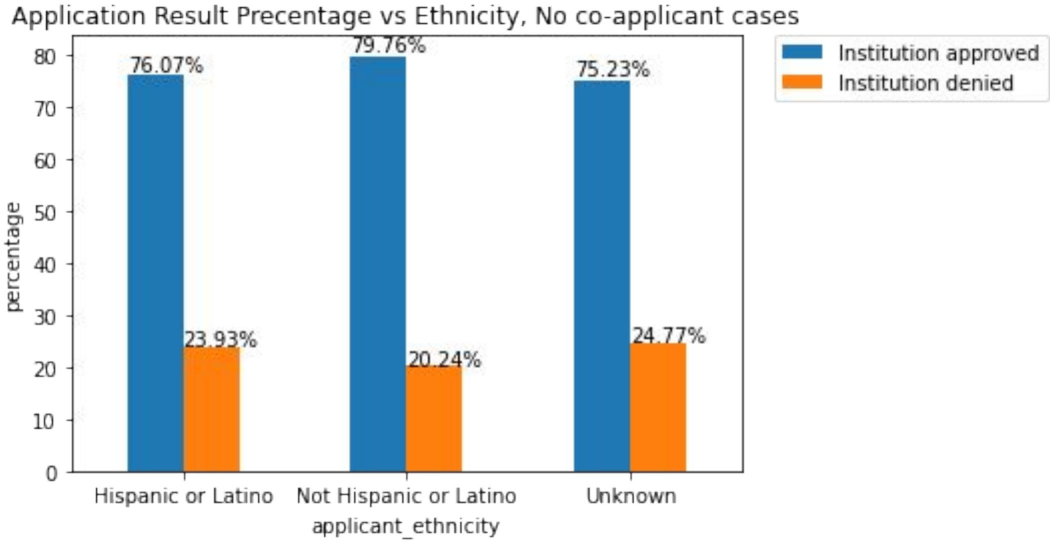


Comparison

With Co-applicant

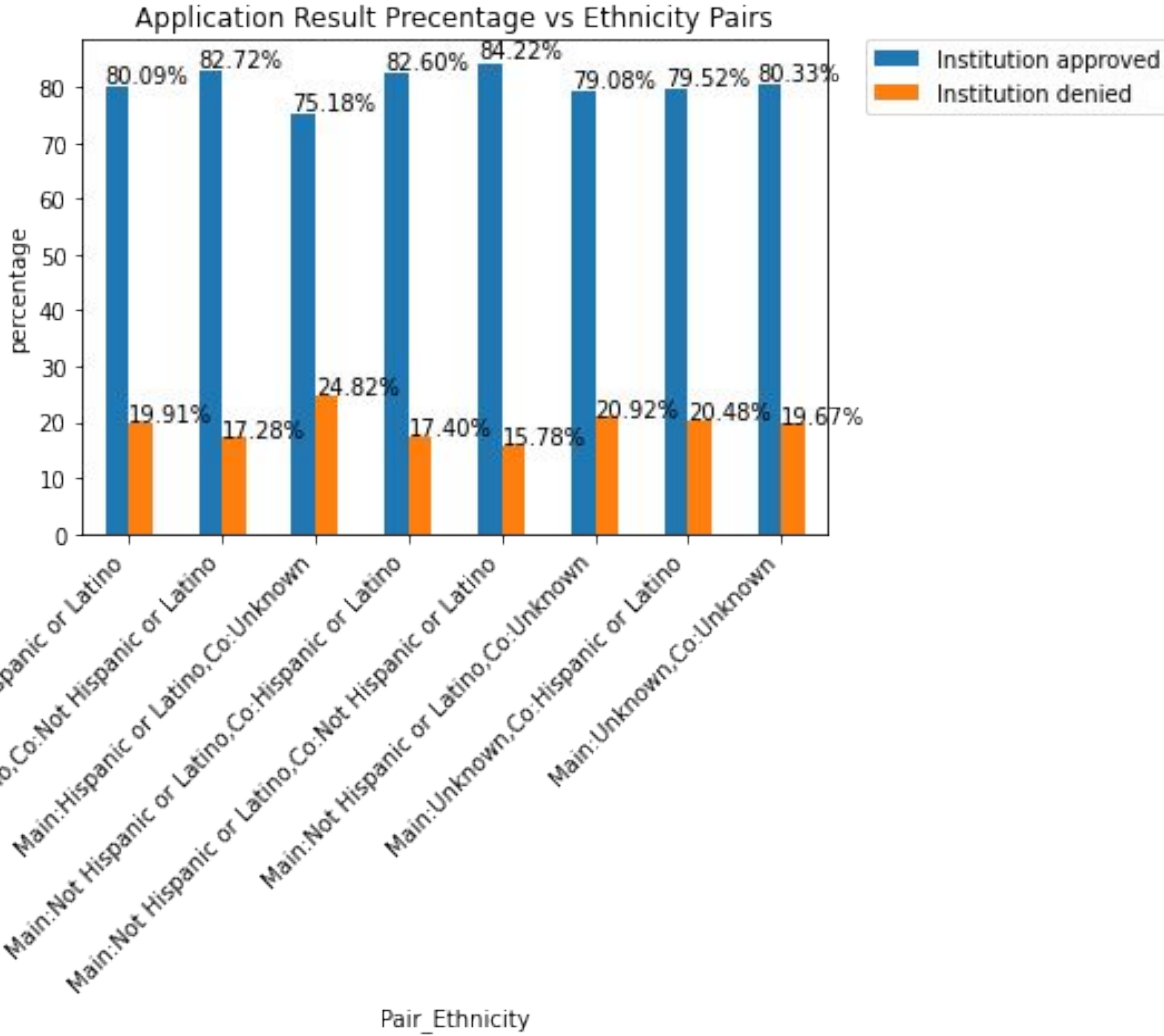
No Co-applicant



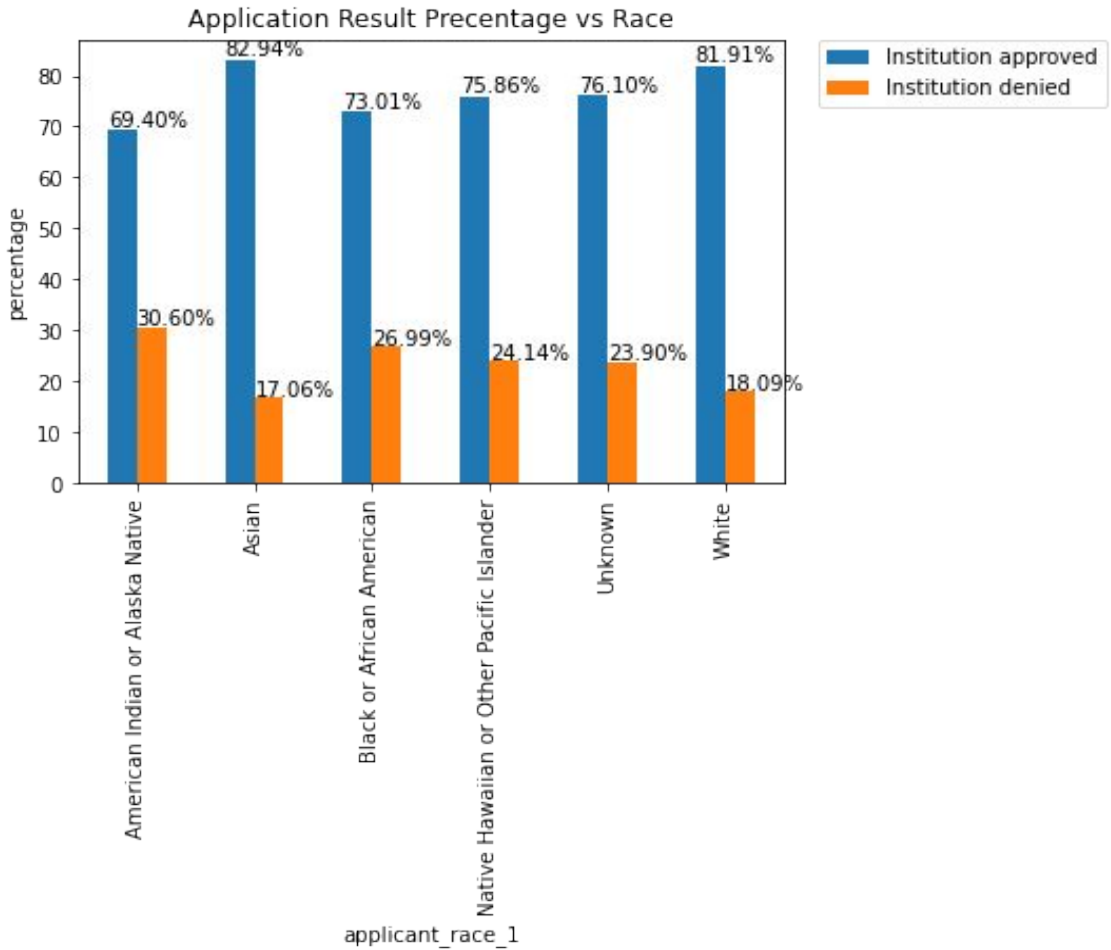


- Applicants tend to get approved if have a co-applicant

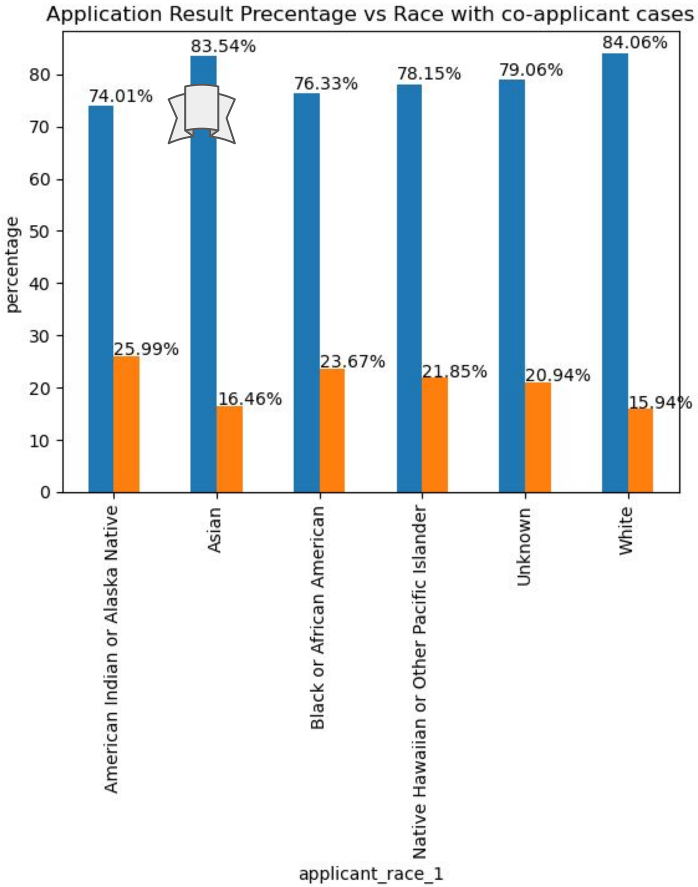
Pairs Matter!

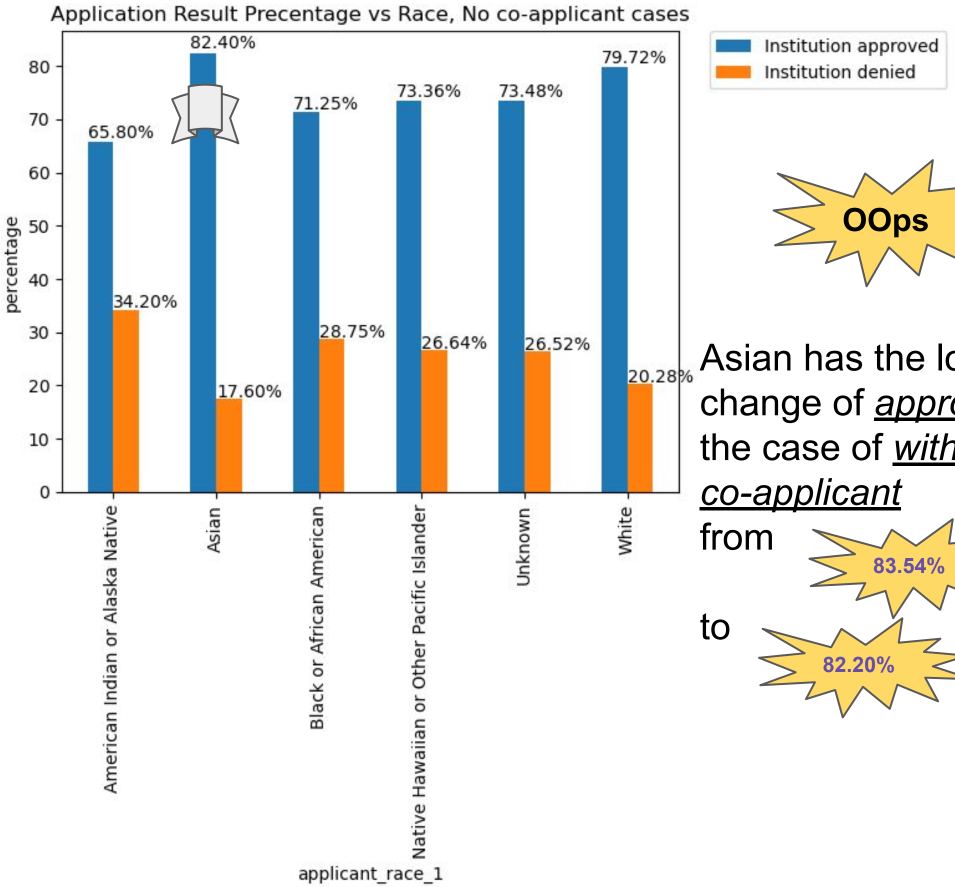


Race Analysis



Co-applicant and No co-applicant comparison





OOps

Asian has the lowest change of <u>*approved* rate</u> in the case of <u>*with or without co-applicant*</u>

from **83.54%**

to **82.20%**

Predictive Model

- Logistic Regression: linear space
- Containing different kinds of features:
  - Numerical: Income, loan amount
  - Categorical: Property type, Sex, *has_coapplicant*

One-hot Encoding

E.g. 1: [1,0,0]; 2: [0,1,0]; 3:[0,0,1]

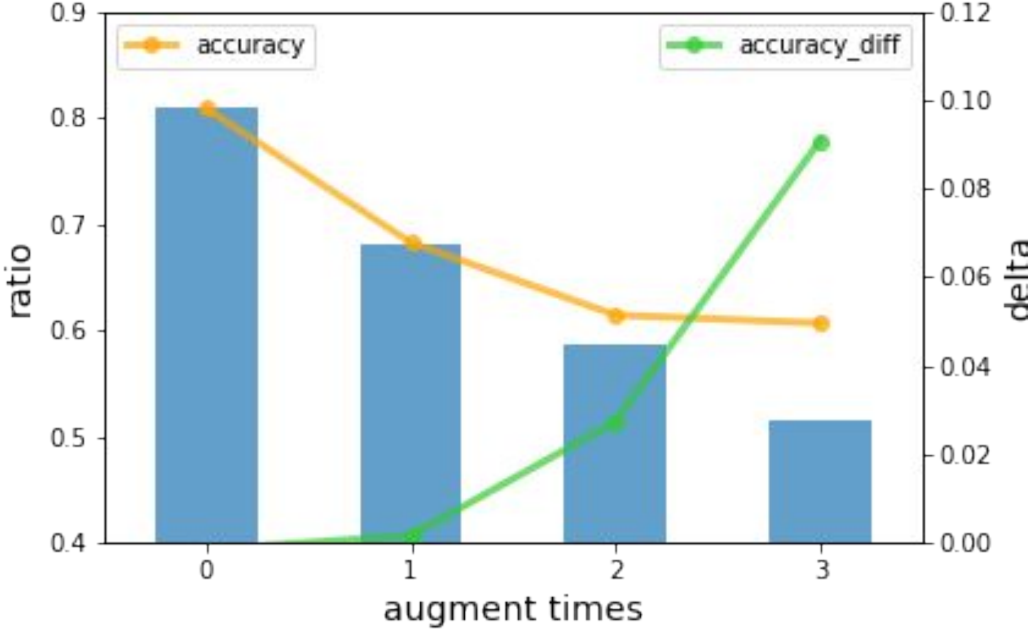
- What about Multi-Valued Categorical Features?
  - Race: you can select several to reflect your family

“Multi-hot” Dummy Encoding

E.g. 1&2: [1,1,0]; 1&2&3: [1,1,1]

Training

- First sight: 80% Accuracy
- How?
  - The model mostly fits 80% of approved samples
  - No improvement compared to all-approve
- Data Augmentation
  - Accuracy seems to drop
  - But the model excels itself from all-approve



Conclusion

- Institution decisions:
  - Being fair on Gender (though seemingly not that fair on surface) and reasonable on income.
  - Show preference on applicant’s race/ethnicity, potential unfair treatment on specific groups like Hispanic people or American Indian. Prefer Asian even without co-applicant provided.
- Applicant advice:
  - Do not leave anything blank/Not applicable.
  - Find a co-applicant.
  - Try to find a stable income source that give one 60k+ annually.
  - Apply our predictive model to preemptively check if application can be approved or not
- Social Insights
  - Huge income gap.
  - Gender inequality conventions/ideas.