Dataiku Data Scientist Technical Assessment

F. Ege Hosgungor

OUTLINE

Introduction

Exploratory Data Analysis

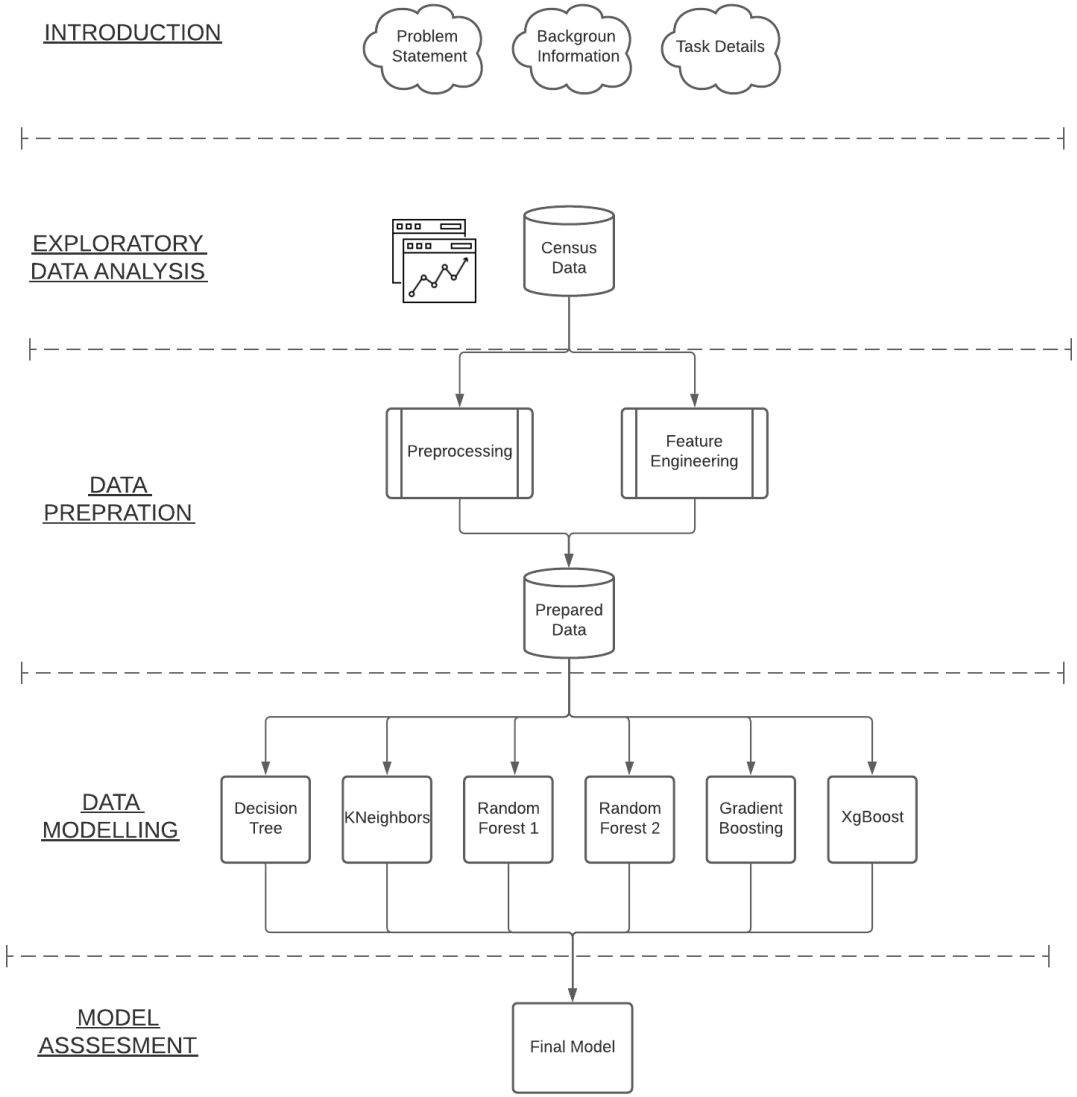
Data Preparation

Data Modelling

Model Assessment

Results

OUTLINE – Whole Process Diagram



INTRODUCTION

The Data:

US Census archive containing detailed, but anonymized, information for approximately 300,000 individuals.

Problem Statement:

Identifying characteristics that are associated with a person making more or less than $50,000 per year.

INTRODUCTION – Background Information

The United States Census Bureau leads the country's Federal Statistical System

Responsible for collecting data every 10 years upon the American people and economy to help inform strategic initiatives

The data is also used for examining the demographic characteristics of population.

INTRODUCTION – Task Details

Binary Classification Task (Less Than $50,000 – More Than $50,000)

There are 200,000 training examples and 100,000 test examples (**2/3 ratio**)

Imbalance Data (Most of the data are example of majority class)

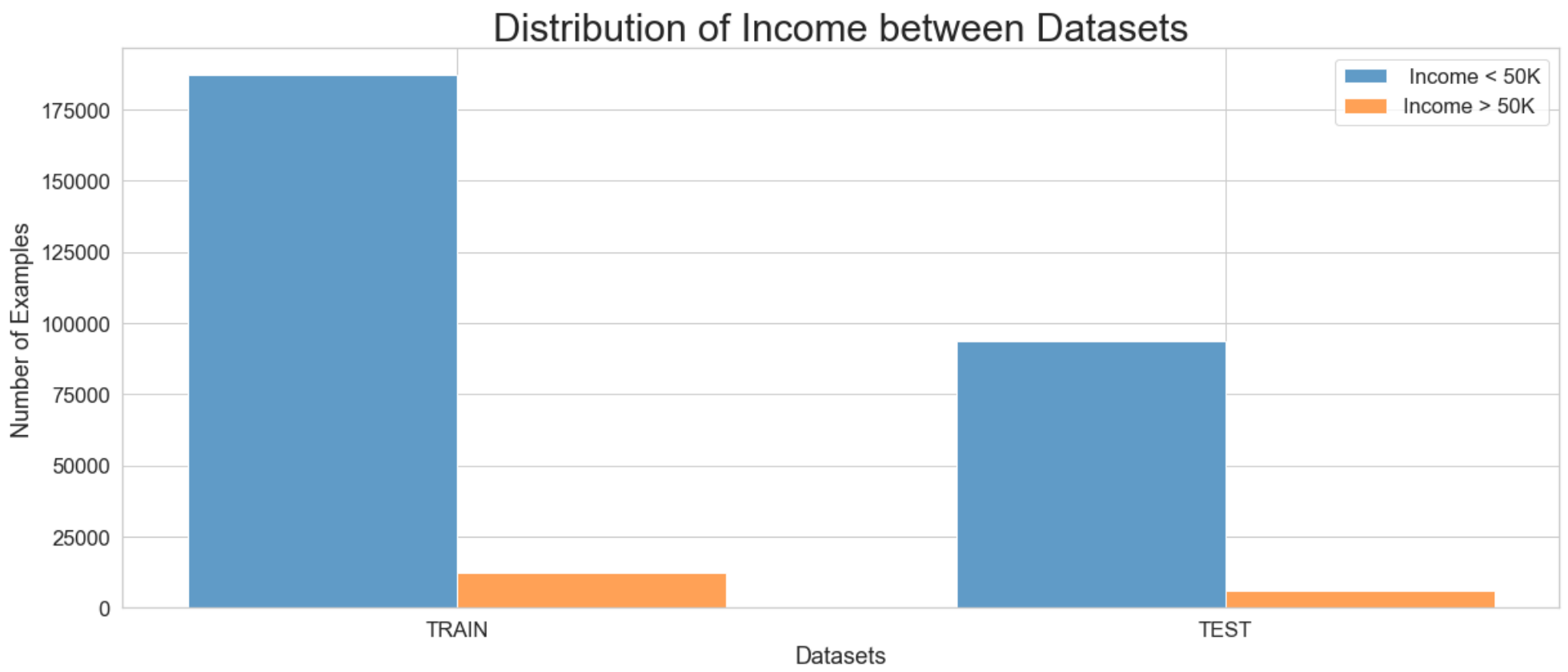
EXPLORATORY DATA ANALYSIS

There are 42 features in total, 35 nominal, 8 continuous including the label column.

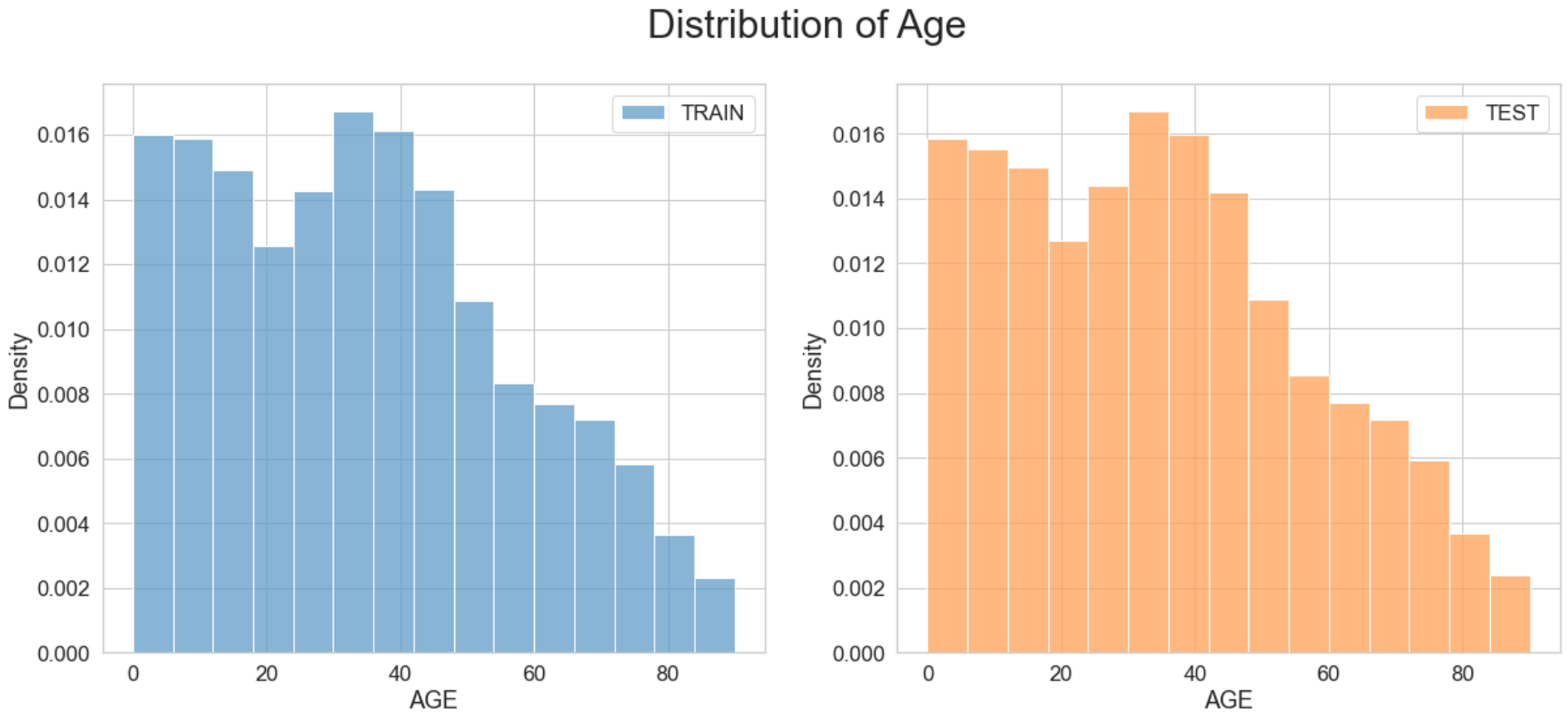
Here are the ones that I worked on most:

- Age
- Sex
- Education
- Occupation and Industry Codes
- **Income** (Target Feature)

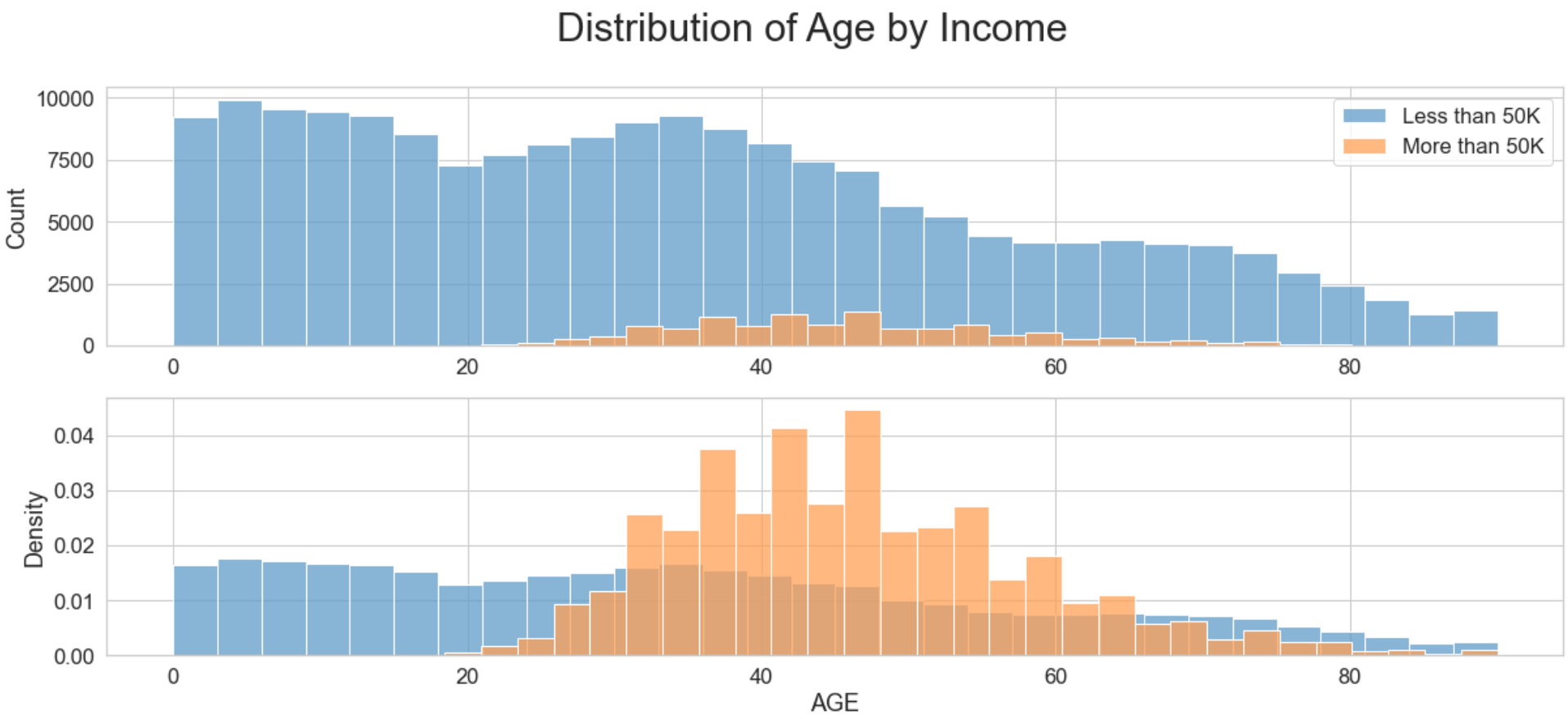
EXPLORATORY DATA ANALYSIS



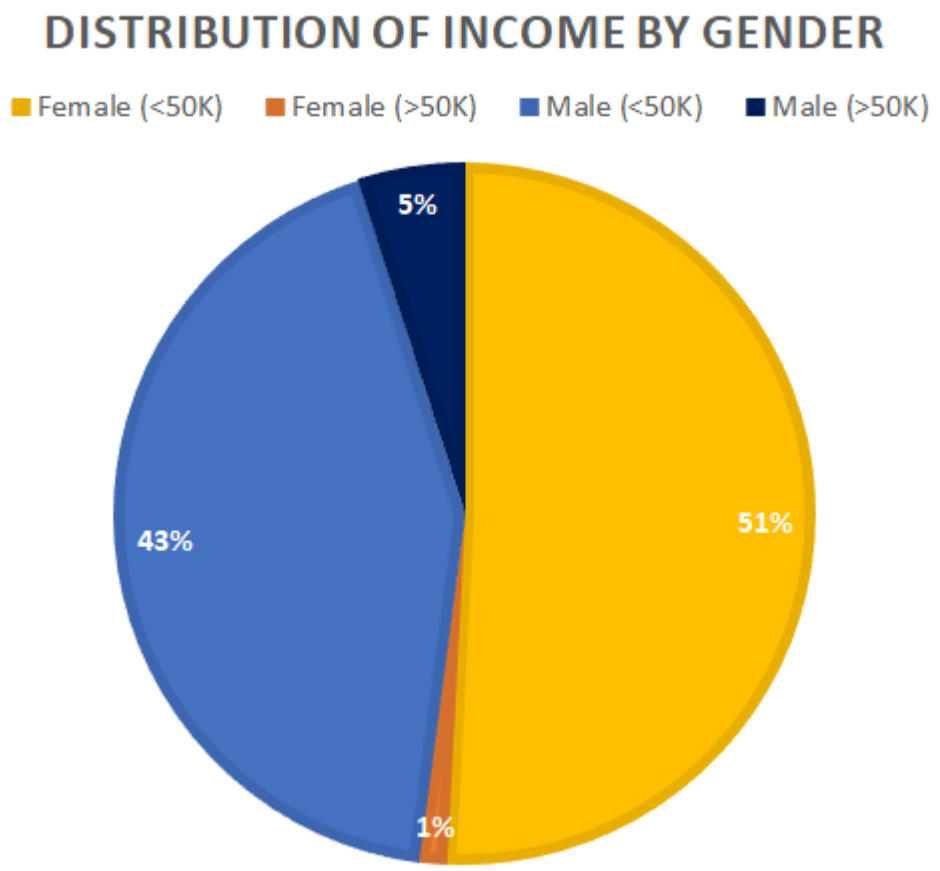
EXPLORATORY DATA ANALYSIS



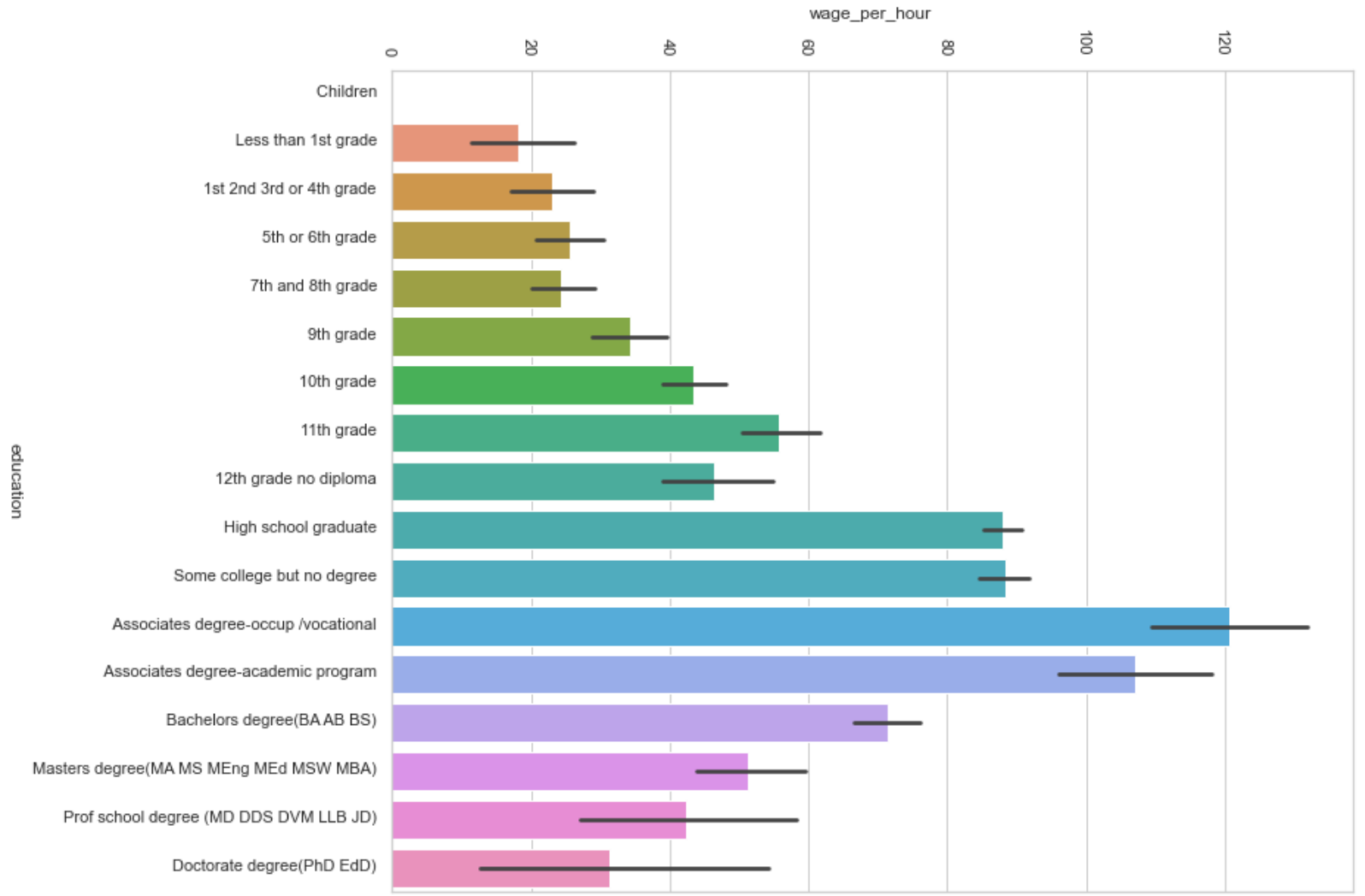
EXPLORATORY DATA ANALYSIS



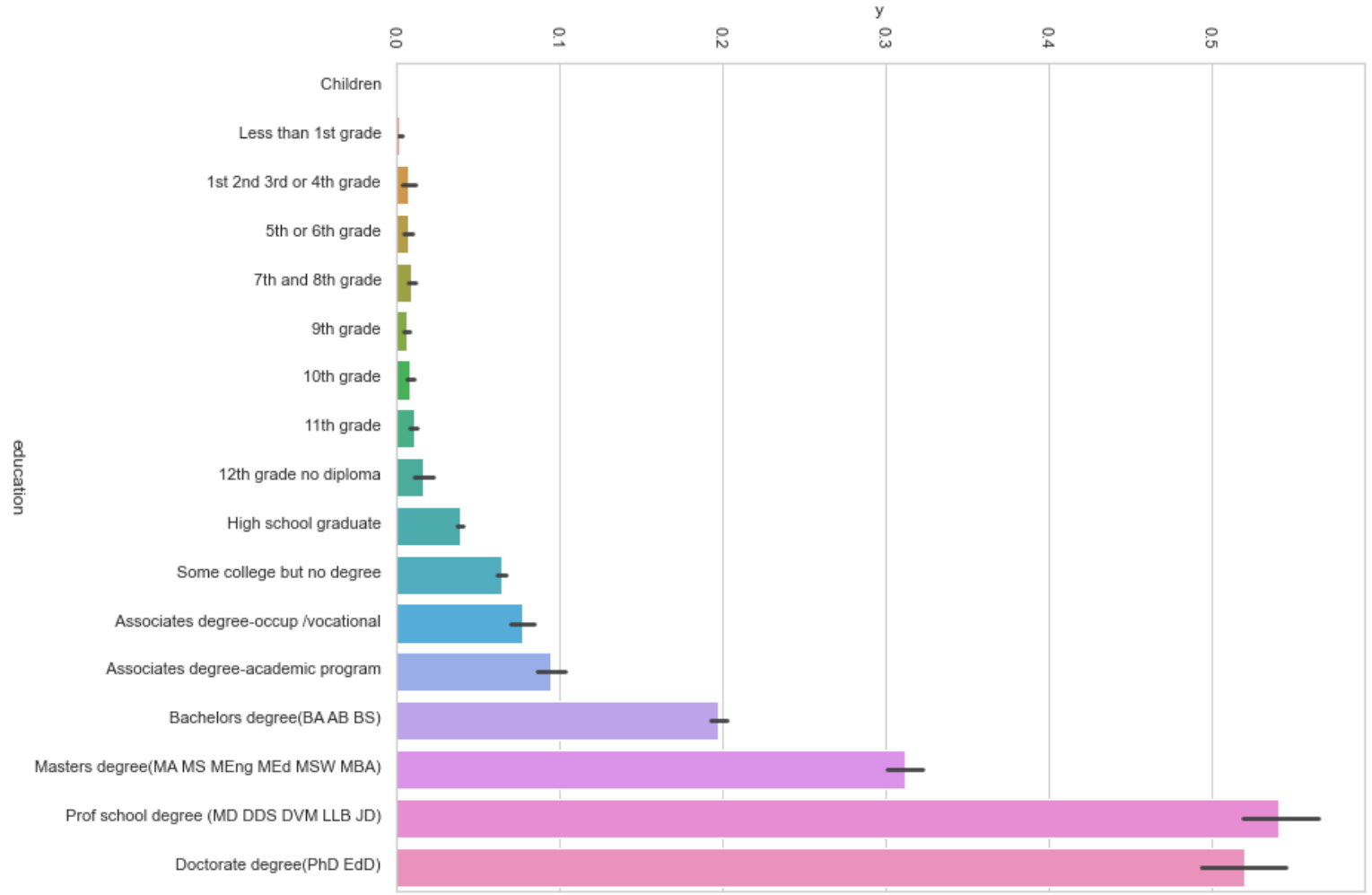
EXPLORATORY DATA ANALYSIS



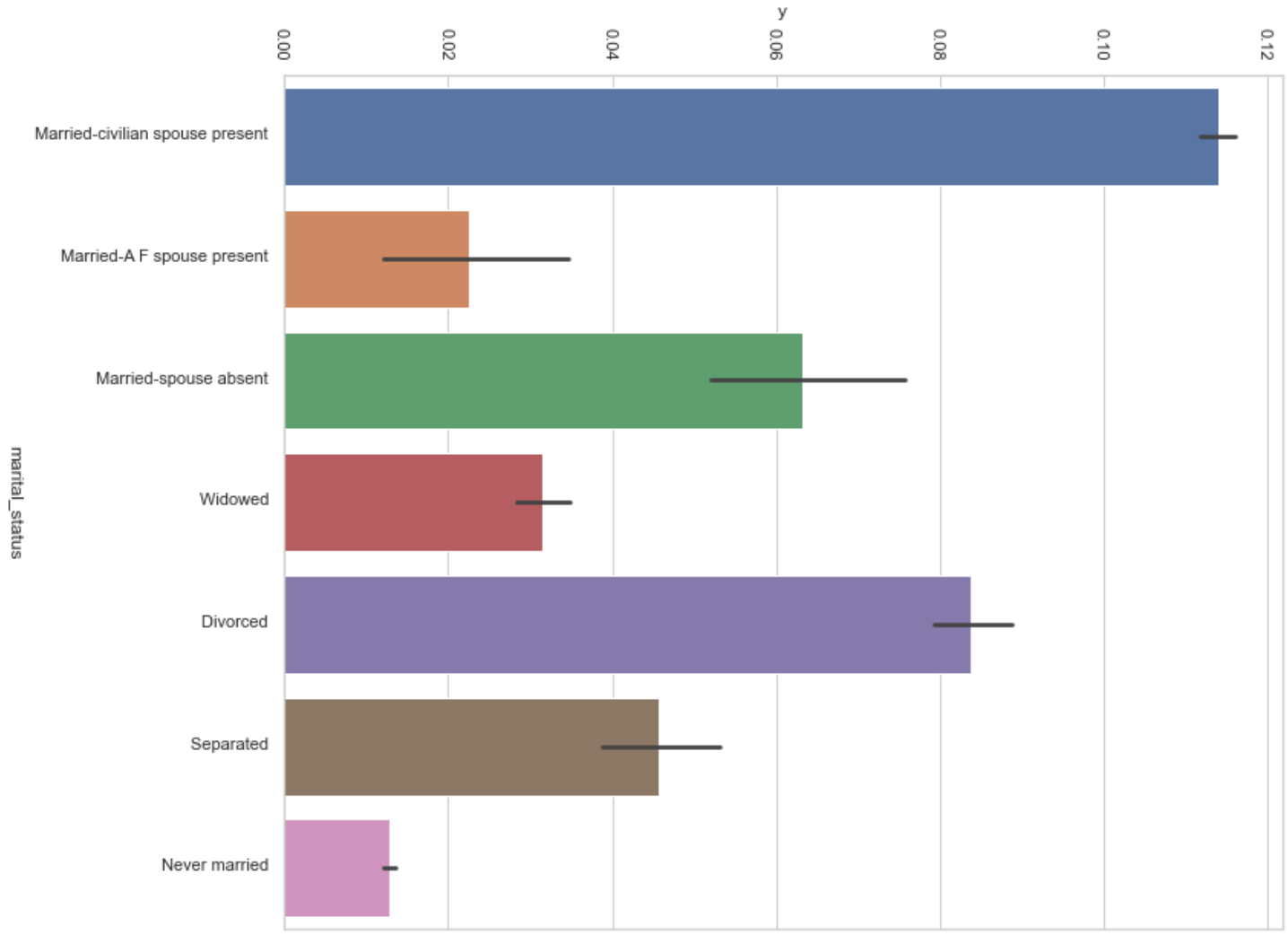
EXPLORATORY DATA ANALYSIS – Education VS Wage p. Hour



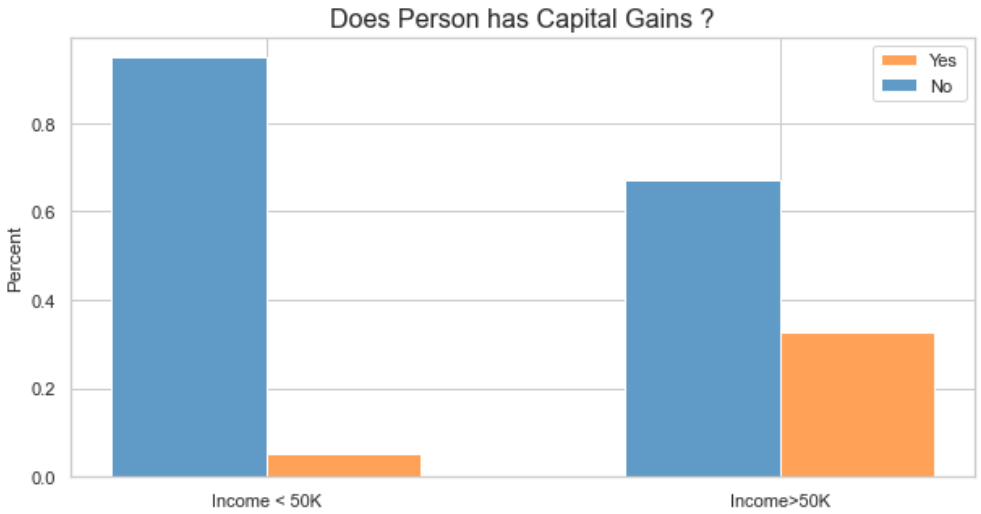
EXPLORATORY DATA ANALYSIS – Education VS Income

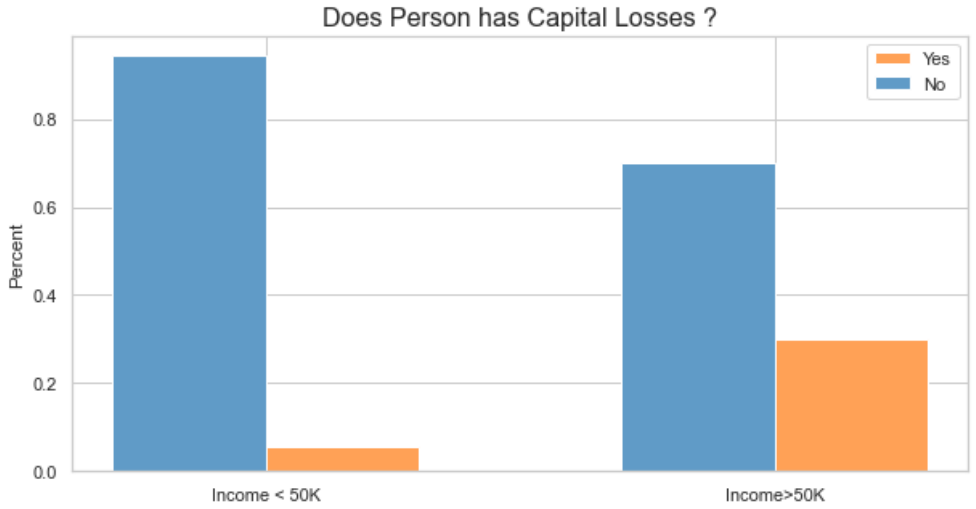


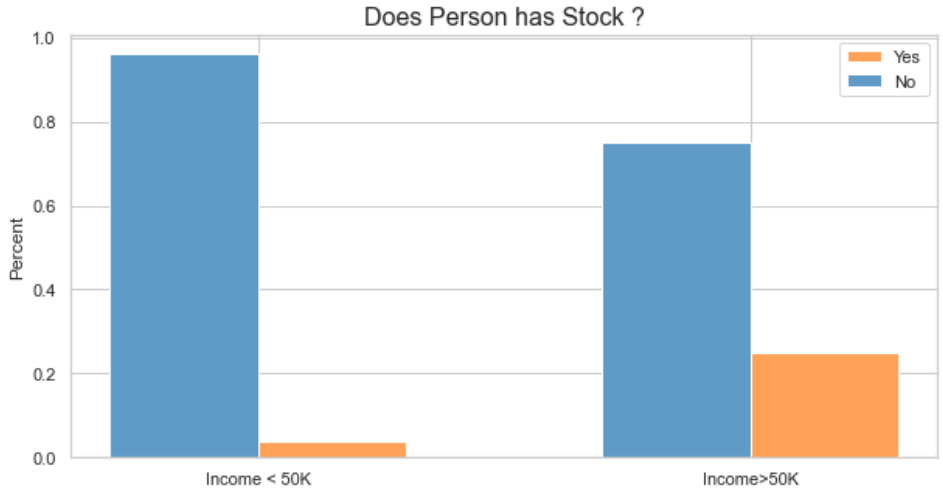
EXPLORATORY DATA ANALYSIS – Marital Status VS Income



EXPLORATORY DATA ANALYSIS – Has Investing

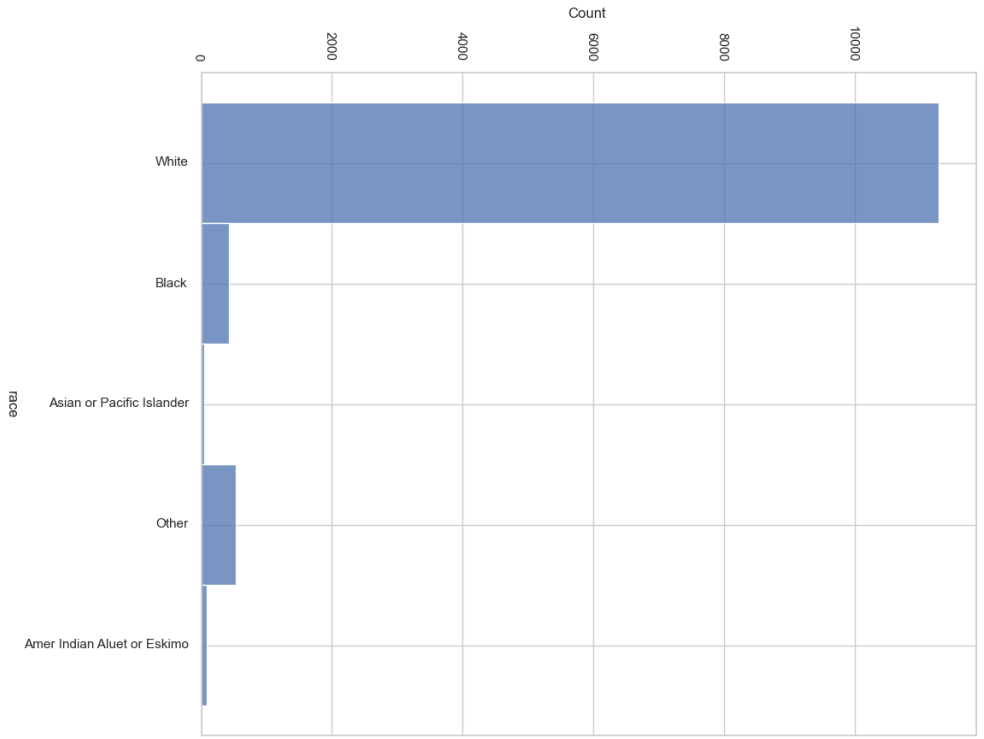




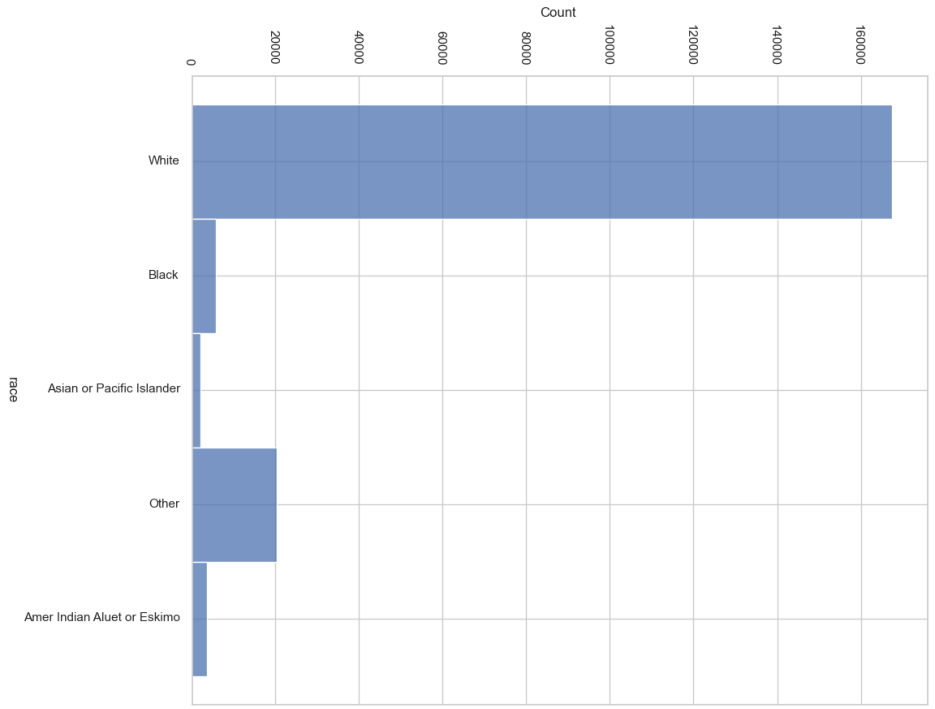


EXPLORATORY DATA ANALYSIS – Race VS Income

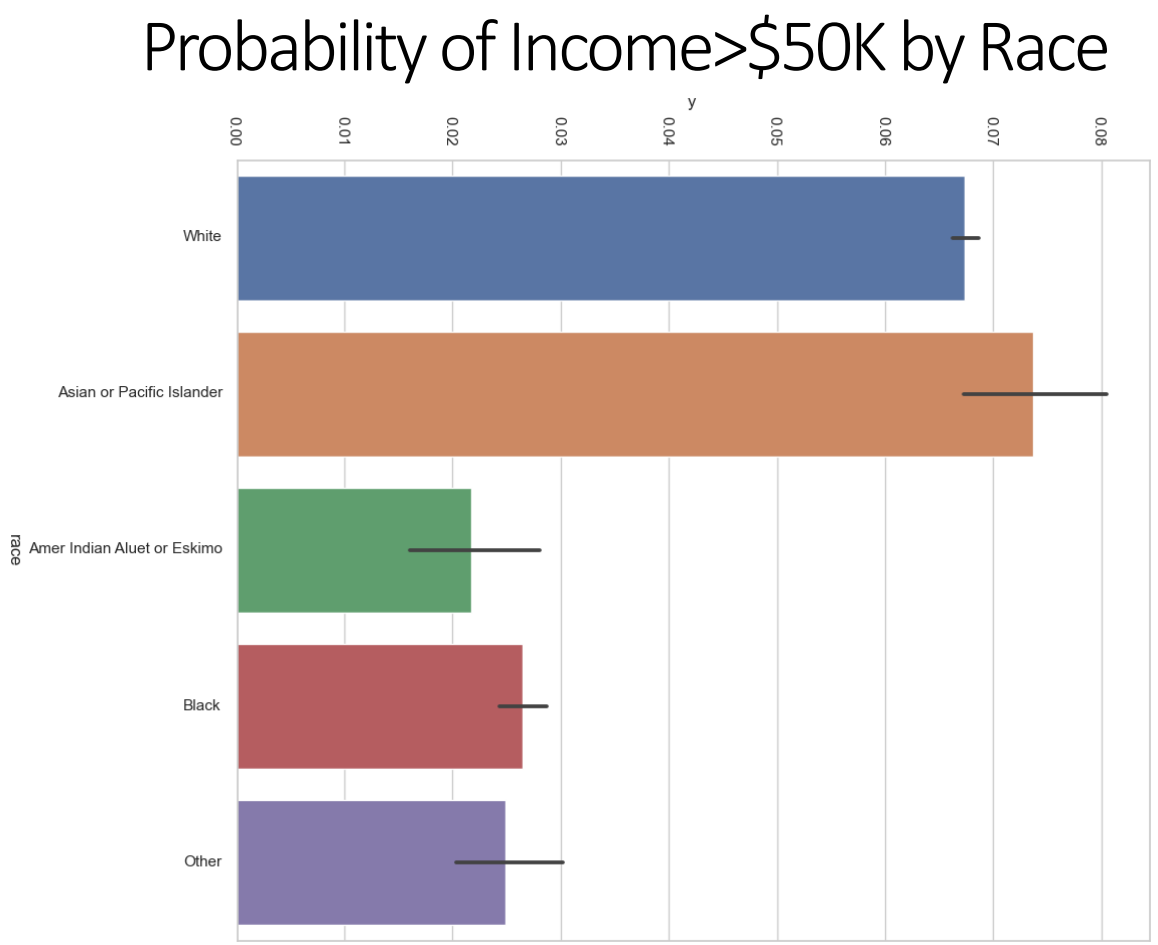
Total Race Distribution



Race Distribution where Income > $50K



EXPLORATORY DATA ANALYSIS – Race VS Income



DATA PREPARATION – Categorization

- Encode categories by using LabelEncoder for 28 nominal features out of 35.
- Other 7 nominal features were **encoded ordinally** by researching the features on Current Population Survey Handbook
- **New features** are created (feature engineering)
- Some features are **normalized** for easier training.

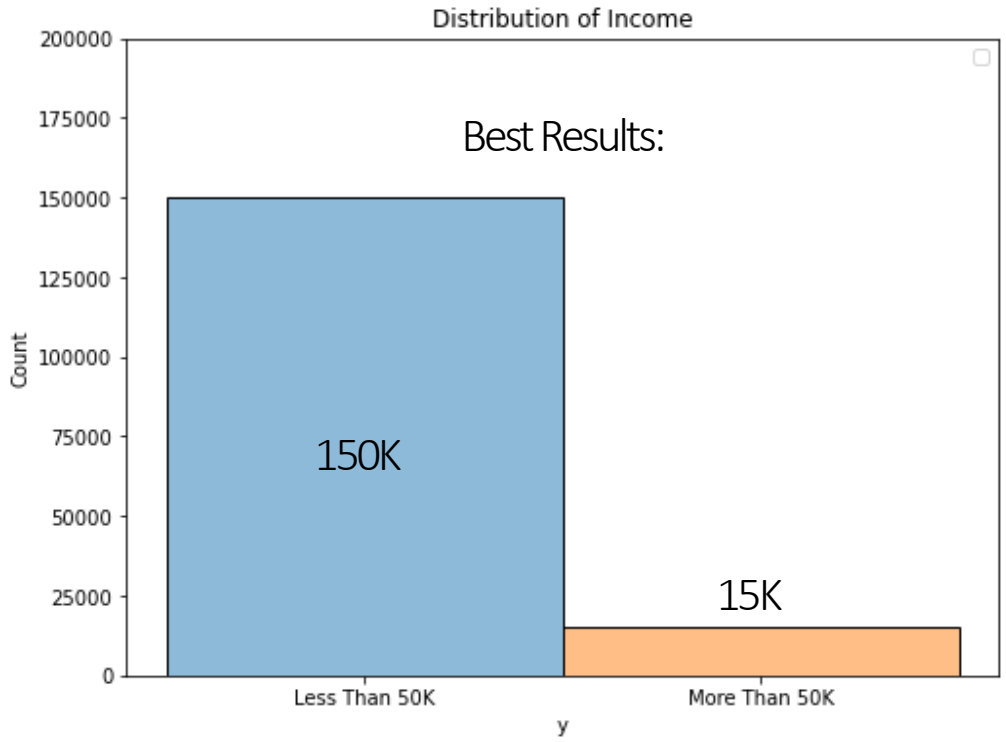
DATA PREPARATION – Upsampling & Downsampling

Income > $50K: **187,000** individuals
Income < $50K: **12,300** individuals

How can we handle this problem ?

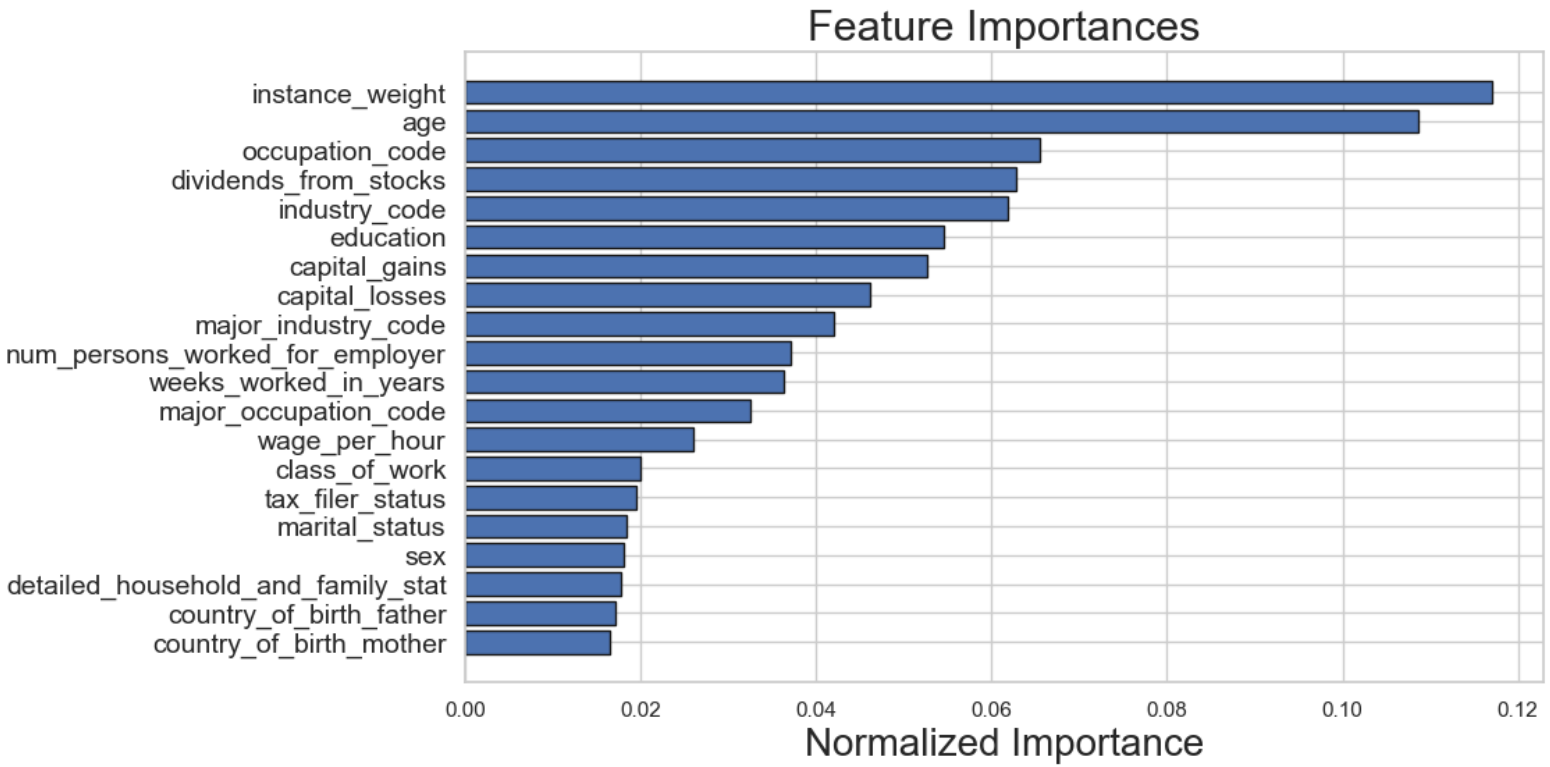
Upsampling the minority class & **downsampling** the majority class

DATA PREPARATION – Upsampling & Downsampling



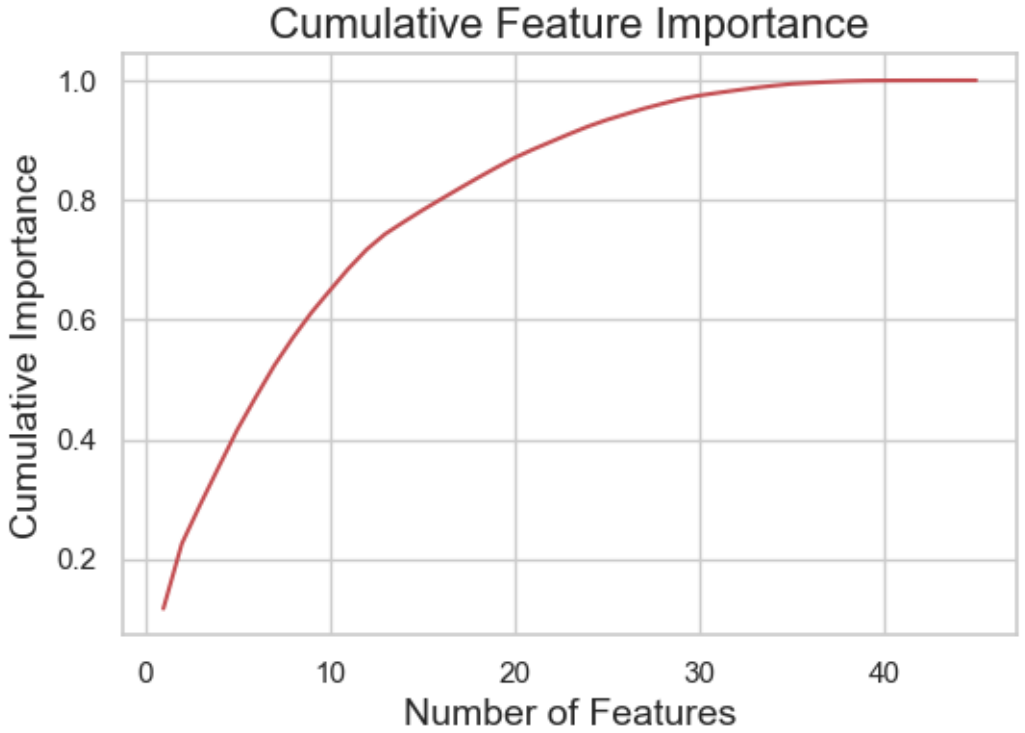
DATA PREPARATION – Feature Importance

Used LightGBM to extract Feature Importance

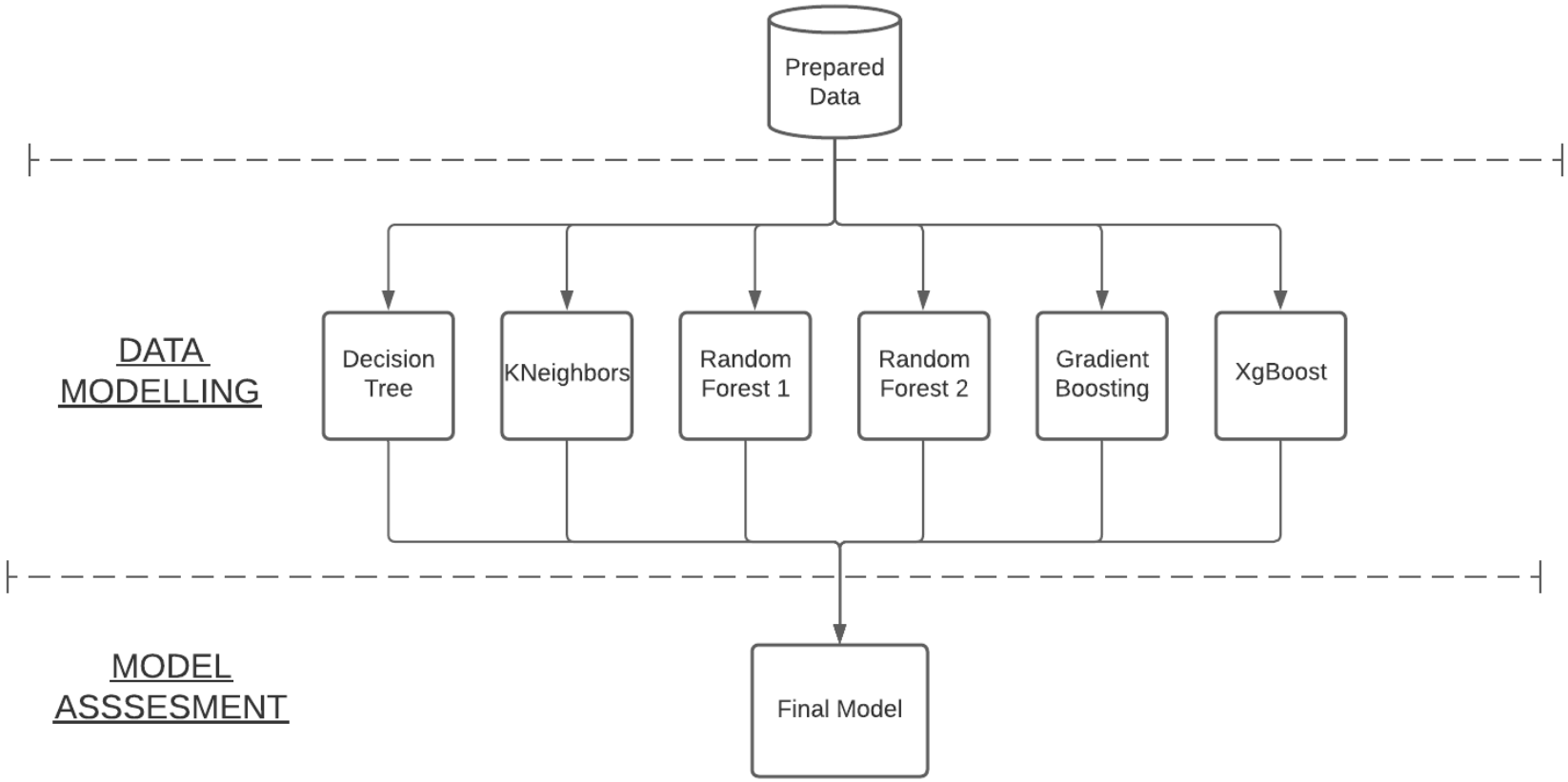


DATA PREPARATION – Cumulative Feature Importance

Removed 22 features from 45 since they weren't adding any value to the model



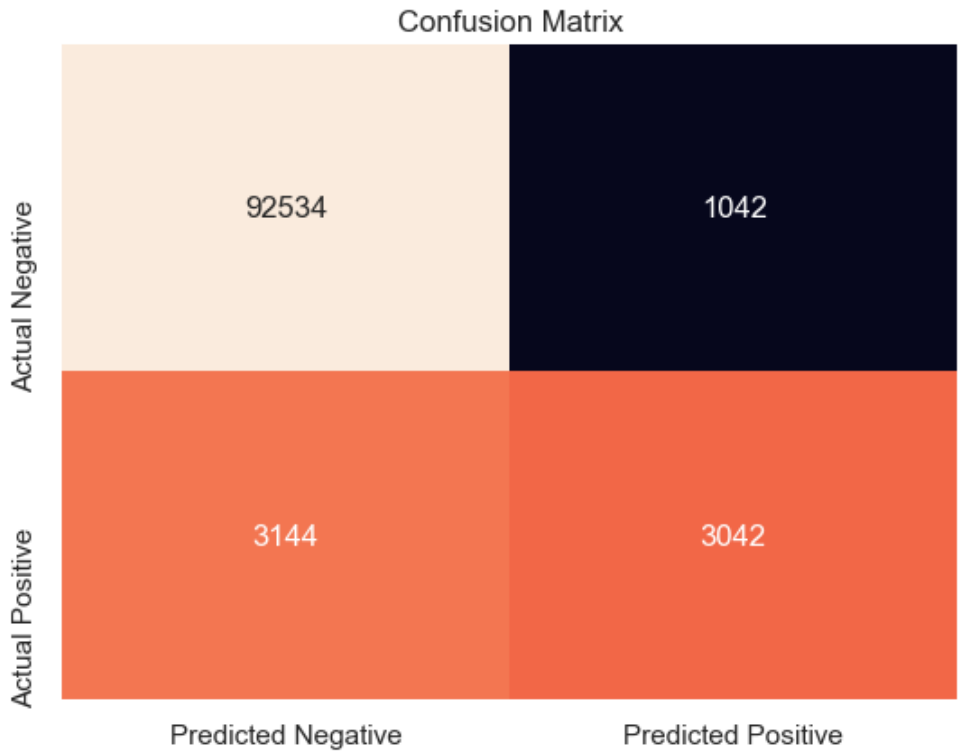
DATA MODELLING – Final Model Diagram



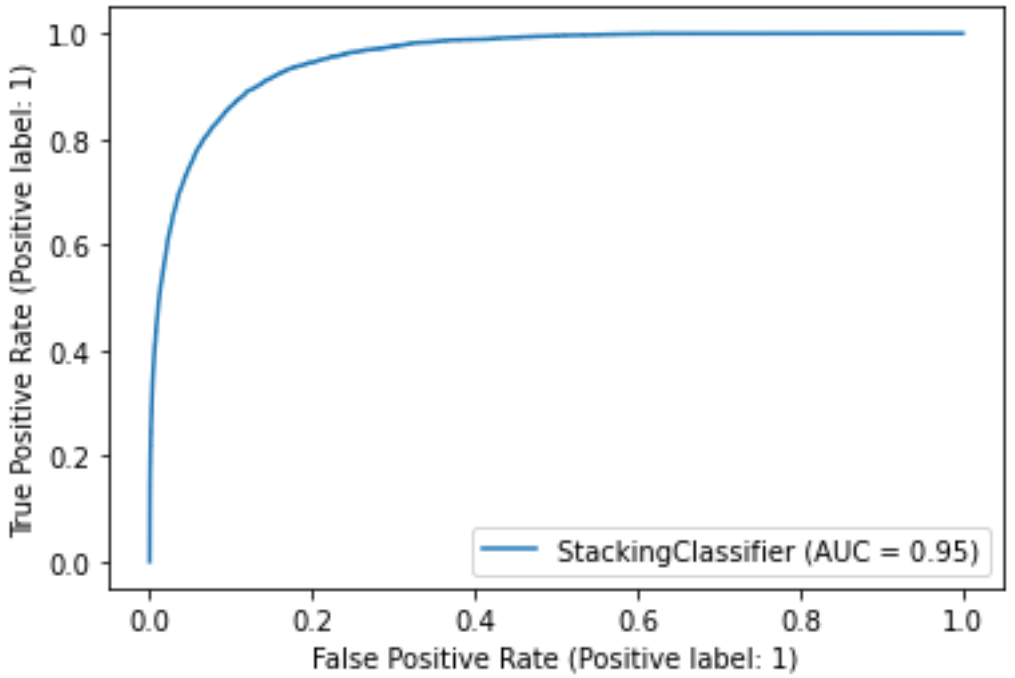
DATA MODELLING – Ensemble Learning

- 6 different model has been trained.Each one of them might be good at capturing some of the data.
- **6 predictions** for each example.
- On top of that 6 models, one linear regression model which decides the final prediction. This is called **“Stacking”**
- Disadvantage: **Takes long time to compute** prediction for 6 different models.

MODEL ASSESSMENT – Confusion Matrix / ROC CURVE



ROC and Precision/Recall Graph



MODEL ASSESSMENT

	PRECISION	RECALL	F1-SCORE	SUPPORT
INCOME < $50K	0.97	0.99	0.98	93576
INCOME > $50k	0.75	0.48	0.59	6186
ACCURACY	-	-	0.96	99762
MACRO AVG	0.86	0.74	0.785	99762
WEIGHTED AVG	0.95	0.96	0.958	9976

RESULTS– Quantitivate

- 97% of the time, individuals who has less income are predicted correctly.
- 75% of the time , individuals who has more income are predicted correctly.

- 99% of the individuals who has less income are caught. (So it is obvious if you have less money)
- Only **60%** of the individuals who has more income are caught.

- The accuracy of the model is 96%
- The macro average accuracy is **79%**

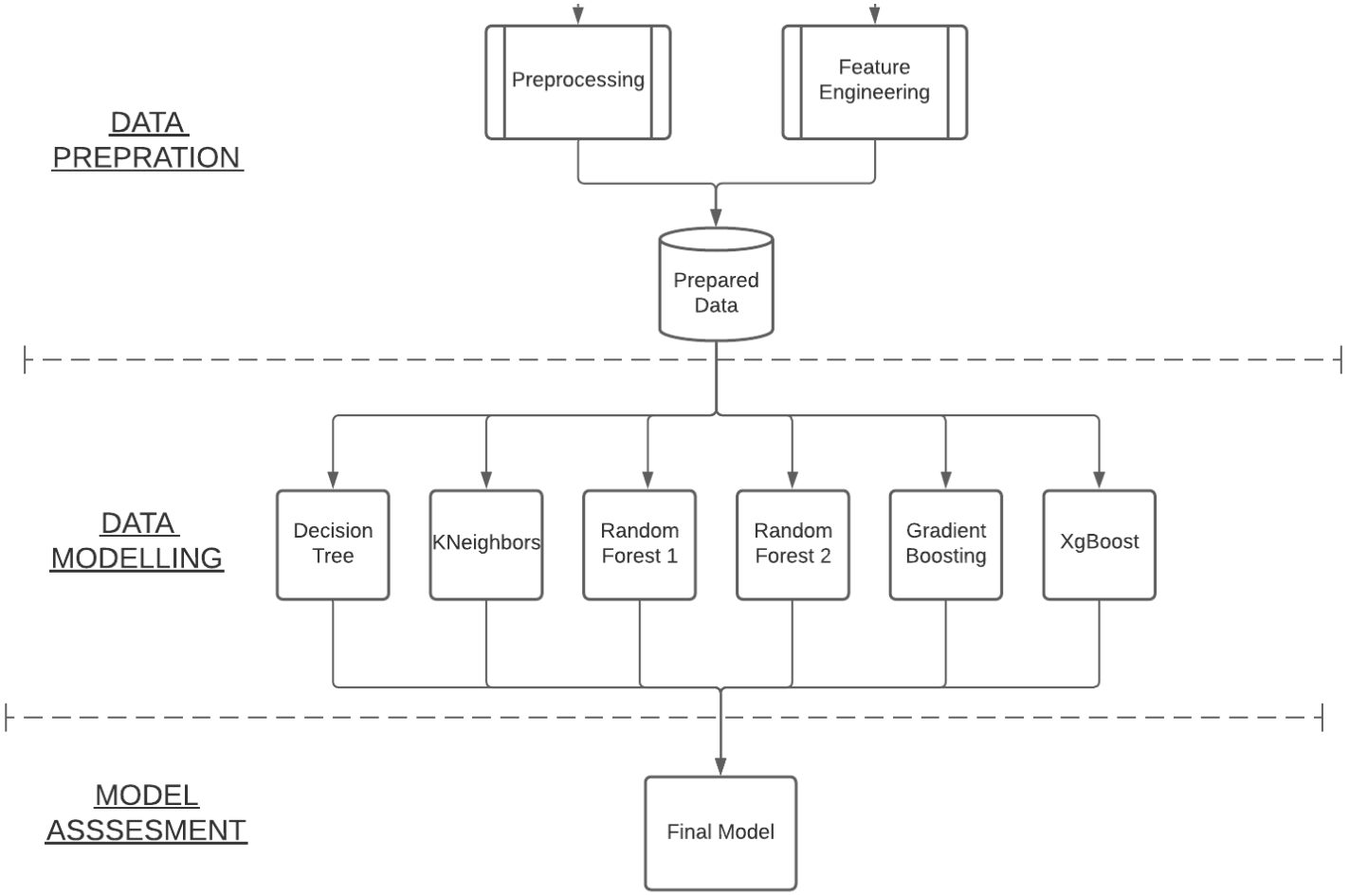
RESULTS - Qualitative

- Most of the individuals who have more income than $50K are **white** but as a ratio **Asian or Pacific Islanders have higher probability**.
- The gender inequality is way higher between **high earners**.
- Having capitals correlates with **high earning** which indicates these individuals are **more willing to take risk** and invest their money.
- **Marrried individuals** have more chance to earn >$50K than **never married** people.
- **Age** is an indicator for predict the individuals with **more income**.

Future Plans

- **More data** from people who earn more than $50K for balancing out the imbalance data.
- **More features** from Census Data
- Up-to-date Data (they are from 94/95)
- Hyperparameter Tuning
- Model Interpretability by additive explanations (SHAP takes some time and computing power)

Summary


DATA PREPRATION
Preprocessing
Feature Engineering
Prepared Data
DATA MODELLING
Decision Tree
KNeighbors
Random Forest 1
Random Forest 2
Gradient Boosting
XgBoost
MODEL ASSSESMENT
Final Model

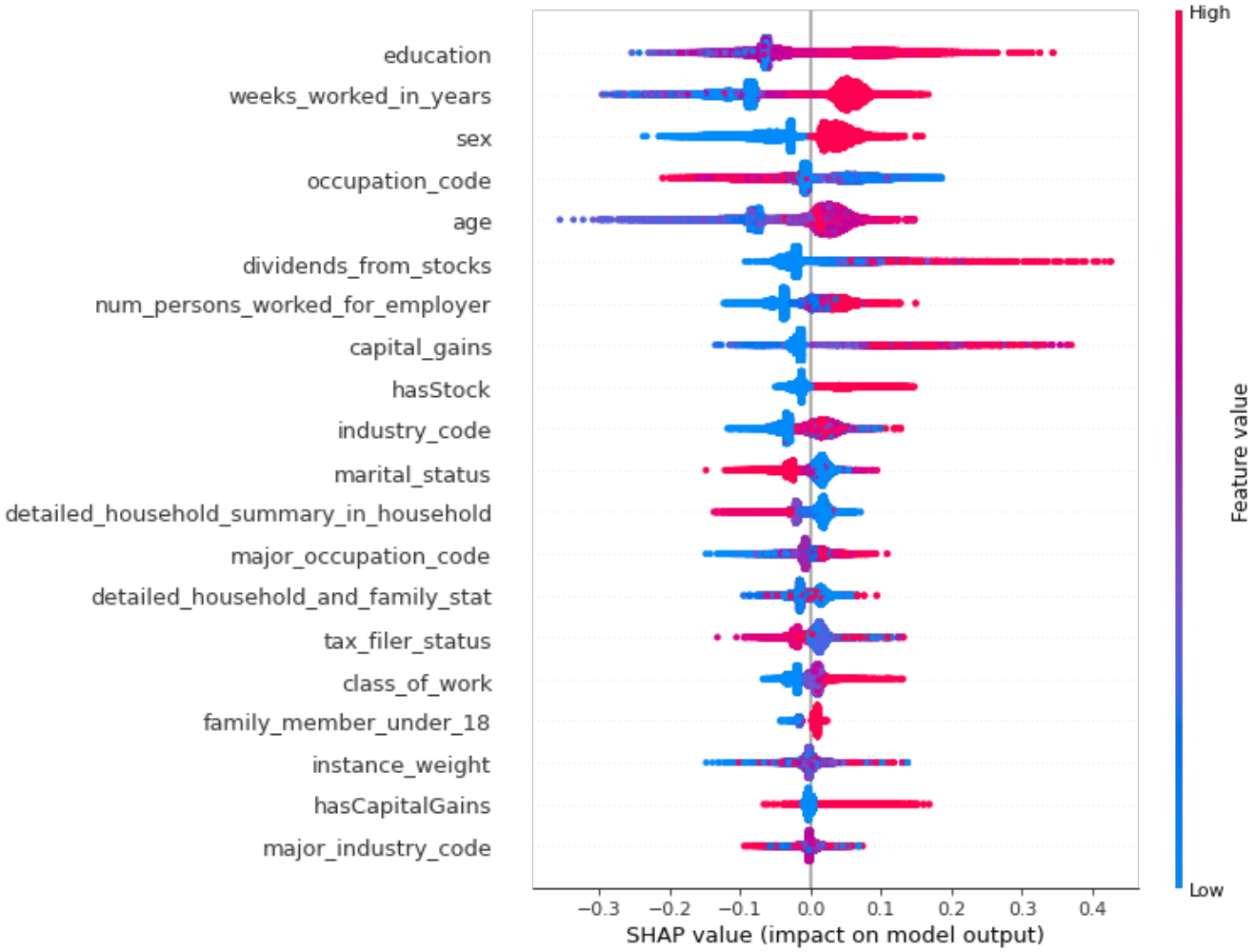
Thanks for Listening
Q & A

F. Ege Hosgungor

Appendix A – All Selected (22) Features

- Age
- Class of Worker
- Industry Code
- Occupation Code
- Education
- Wage per hour
- Marital Status
- Major Industry Code
- Major Occupation Code
- Sex
- Capital Gains
- Capital Losses
- Dividends from Stocks
- Tax Filer Status
- Detailed Household and family Statistics
- Detailed Household summary in household
- Instance weight (Ignore)
- Number of persons worked for employer
- Country of Birth Father
- Country of Birth Mother
- Own Business or Self Employed
- Weeks Worked In Years

RESULTS – Model Interpretation (SHAP Graph)



Appendix B – What is SHAP

- SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model.

- Used for analyzing the results of Model.

- Takes too much time to compute shap value (Not optimized that much yet)

- But has promising results