# Group 2

## US Census 1994 Income Data

Exploration Data, Machine Learning Model, Optimization, Visualization

AI Bootcamp Class with Suzanna Ayash

March 2025

# Contributors

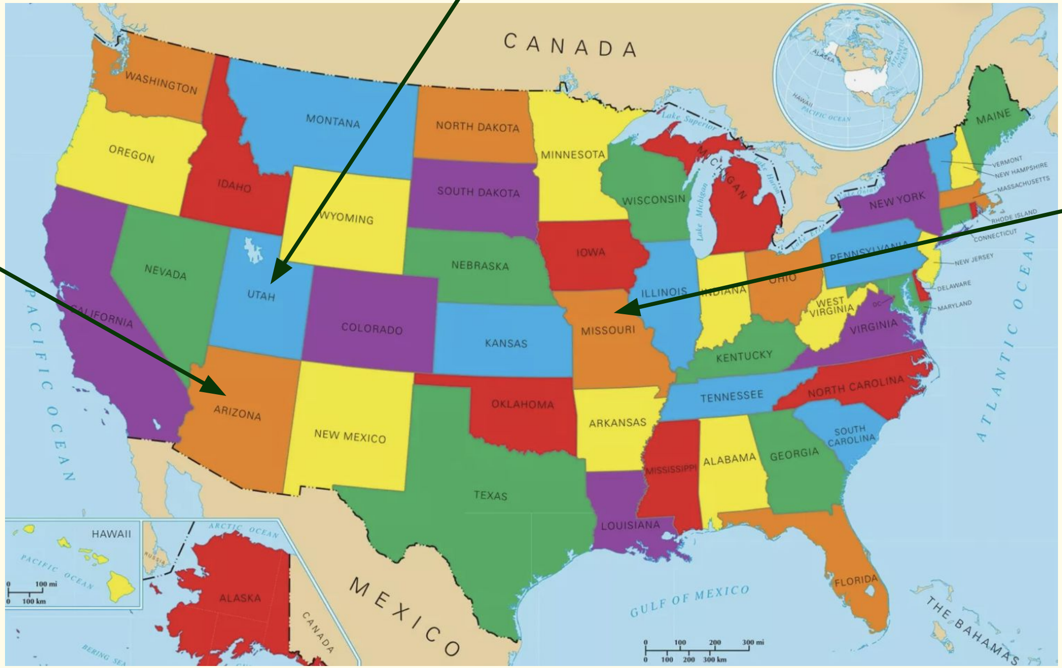
Patrick McCourt

Ingrid Blankevoort

Spencer Gerritsen

Vijay Srinivasula

Matt Le



# Income Census Data is important! Why?

- Economic planning and resource allocation
- Identify Socio/Economic inequalities
- Helps business sell by income segments
- Helps government create laws and policies

# AI Can Help Us with Analyzing Income Census Data. How?

## AI Tools

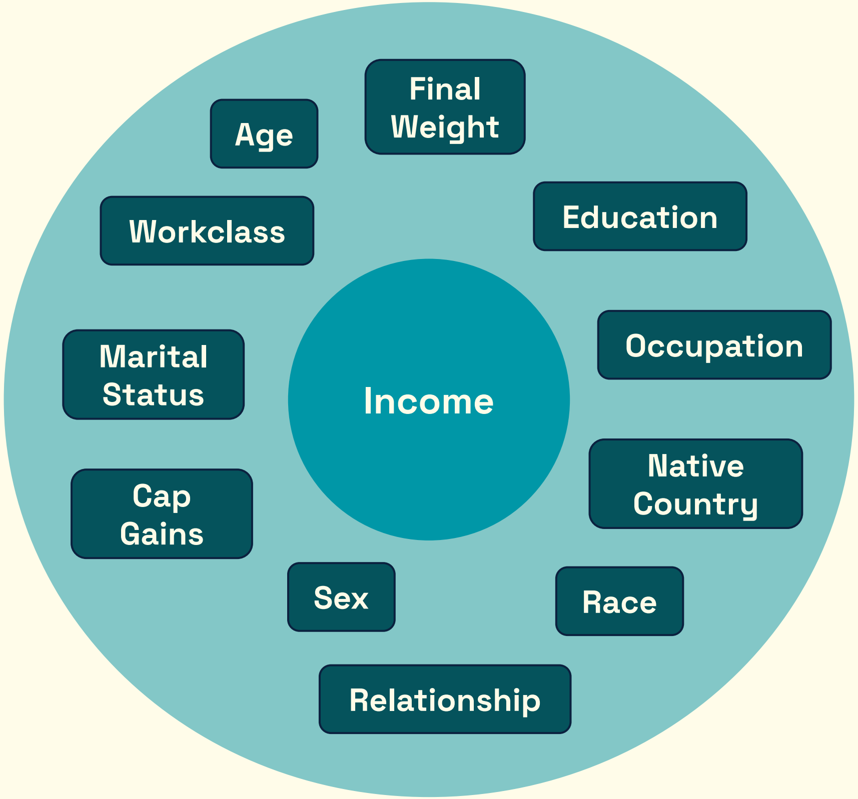
SciKit-Learn
Imbalanced-Learn
Pandas
Random Forest
MatPlotLib
& More

- Reduce Data Collection & Processing Time
- Data Quality and Validation Improvement
- Identifying Patterns
- Demographic Insights
- Real-time populating tracking
- Automate Predictive Analytics

# Overview of the 1994 Income Census Data

- 14 Total features
- 48K rows with over 1K unique values
- Target is Income
- Greater or less than $50K

*Too much 'chaos' in this data...let's talk 'Data Prep!'*



# Data Preparation - Overview

## Visualization

Discover meaningful relationships between each feature and income, possible imbalances

## Clean

Remove duplicates, redundant columns (ie: relationship, education-num, native-country)

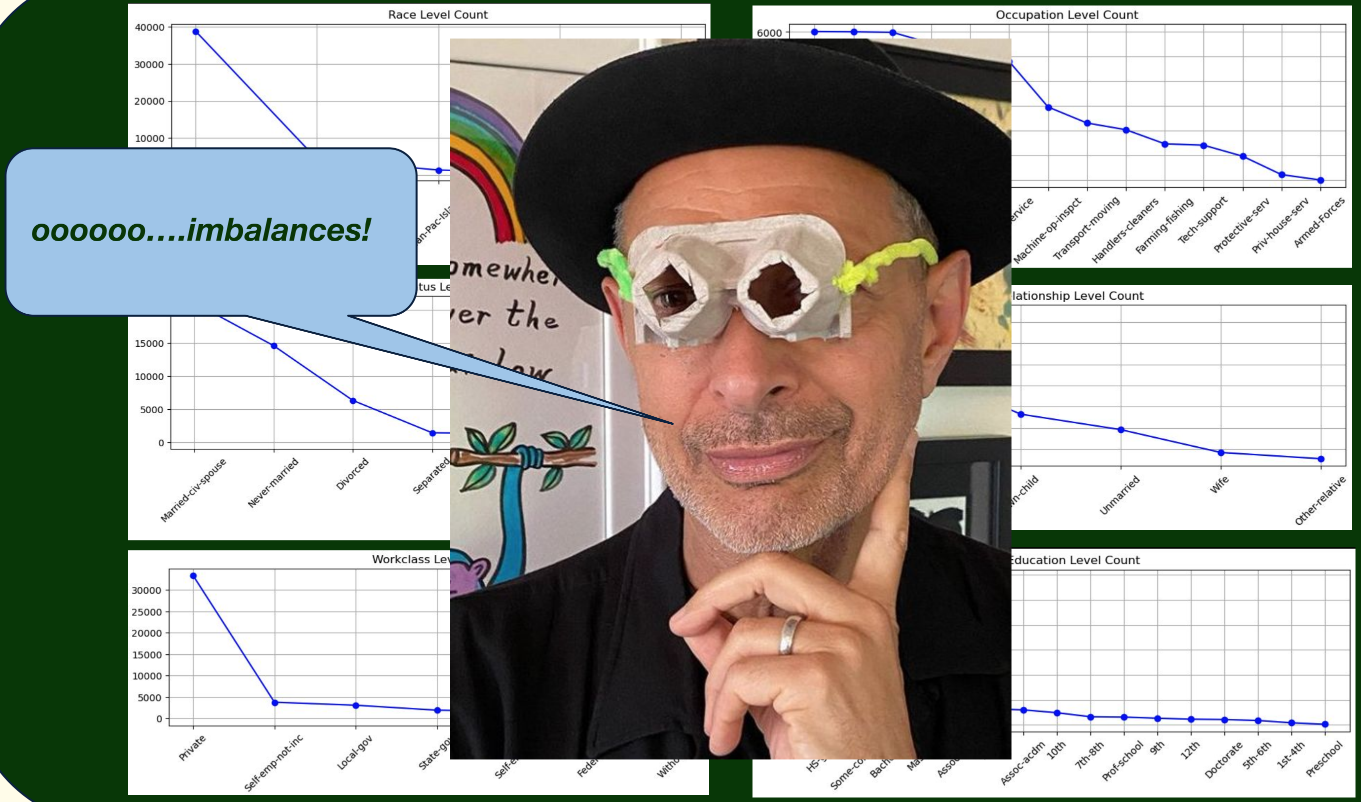
## Feature Engineering

Create new 'assets' by taking 'Capital Gain' minus 'Capital Loss.'

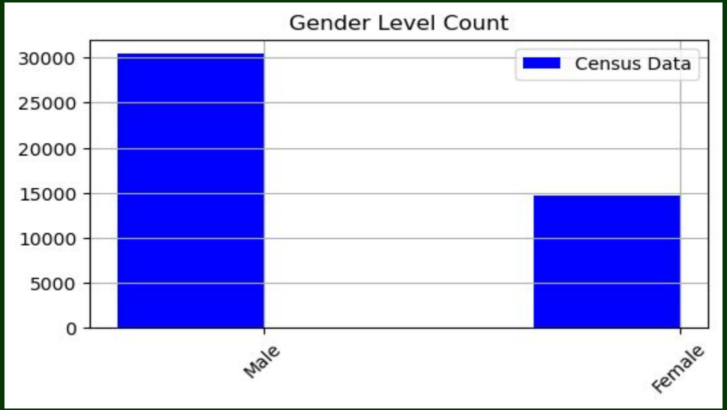
## Encoding & Scaling

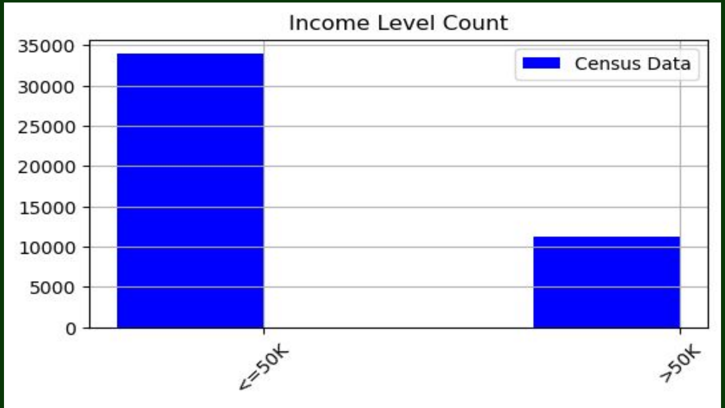
Binary assignment, numerical categorization, and post-split scaling new 'assets' column.

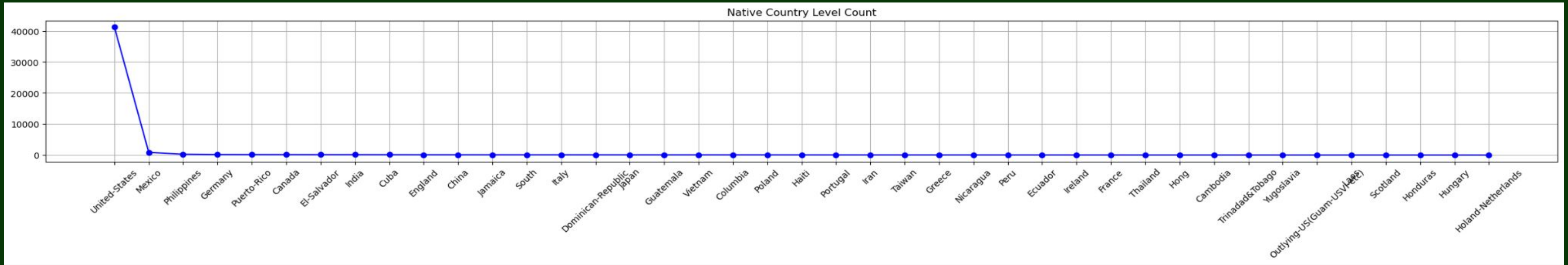
# Data Preparation - Visualization



# Data Preparation - Visualization (cont'd)



# Categorical Encoding

| LEGEND | Age | Workclass | Education | Marital Status | Occupation | Race | Sex | Hours per week | Income |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Less than 30 years | Private sector | Preschool - 6th grade | Single, Divorced, Never Married, Widowed | Blue-collar | White | Male | Part time for less than 30 hours/week | <= 50K |
| 1 | 30 to 50 years | Self Employment | 7th - High School grad | Married | White-collar | Black | Female | Full time for 30-40 hours/week | > 50K |
| 2 | Above 50 years | Government Worker | Some College, Assoc, Bachelors Degree | N/A | | Asian and Pacific Islander | N/A | Over time for above 40 hours/week | N/A |
| 3 | N/A | | Masters, Doctorate, Prof School | N/A | | American Indian, Eskimo | N/A | N/A | N/A |
| 4 | N/A | | | N/A | | Other | N/A | N/A | N/A |

| | age | workclass | fnlwgt | education | marital-status | occupation | race | sex | hours-per-week | income | assets |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 41254 | 41254 | 41254 | 41254 | 41254 | 41254 | 41254 | 41254 | 41254 | 41254 | 41254 |
| mean | 0.88 | 0.418 | 187263.564 | 1.635 | 0.504 | 0.549 | 0.162 | 0.326 | 1.166 | 0.253 | 1037.596 |
| std | 0.702 | 0.734 | 105039.504 | 0.64 | 0.5 | 0.498 | 0.514 | 0.469 | 0.656 | 0.435 | 7629.923 |
| min | 0 | 0 | 13492 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -4356 |
| 25% | 0 | 0 | 115803 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 50% | 1 | 0 | 176728 | 2 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 75% | 1 | 1 | 234640.75 | 2 | 1 | 1 | 0 | 1 | 2 | 1 | 0 |
| max | 2 | 2 | 1490400 | 3 | 1 | 1 | 4 | 1 | 2 | 1 | 99999 |

**Now we work with 41,254 records out of 48,840 original records**

## Variance Inflation Factor (VIF)

- All features VIF values ranged from 1.018 to 1.334 which means NO collinearity detected!
- Importantly, all of these features also had low VIF scores, meaning they weren't collinear with each other

| | feature | VIF |
|---|---|---|
| 3 | marital-status | 1.337080 |
| 6 | sex | 1.314780 |
| 4 | occupation | 1.304685 |
| 2 | education | 1.271688 |
| 0 | age | 1.143666 |
| 7 | hours-per-week | 1.126601 |
| 1 | workclass | 1.060613 |
| 8 | assets | 1.024331 |
| 5 | race | 1.017819 |

## Coefficients for Feature Importances

We found that Marital Status was the strongest predictor, followed by Education, Age, and Occupation.

So, we can trust that their influence is statistically sound and not inflated due to overlap with other variables.

```
[(0.34918395491982024, 'marital-status'),
 (0.17711134368003462, 'education'),
 (0.12974789081621338, 'age'),
 (0.11501608825273132, 'occupation'),
 (0.09513362543523164, 'hours-per-week'),
 (0.06546328706006603, 'sex'),
 (0.03469961356030286, 'workclass'),
 (0.03364419627559988, 'race')]
```

So 11 features and a target binary classification....what model to organize this ‘chaos?’

Let’s run an accuracy test on all!

# Accuracy Score and Model Coefficients

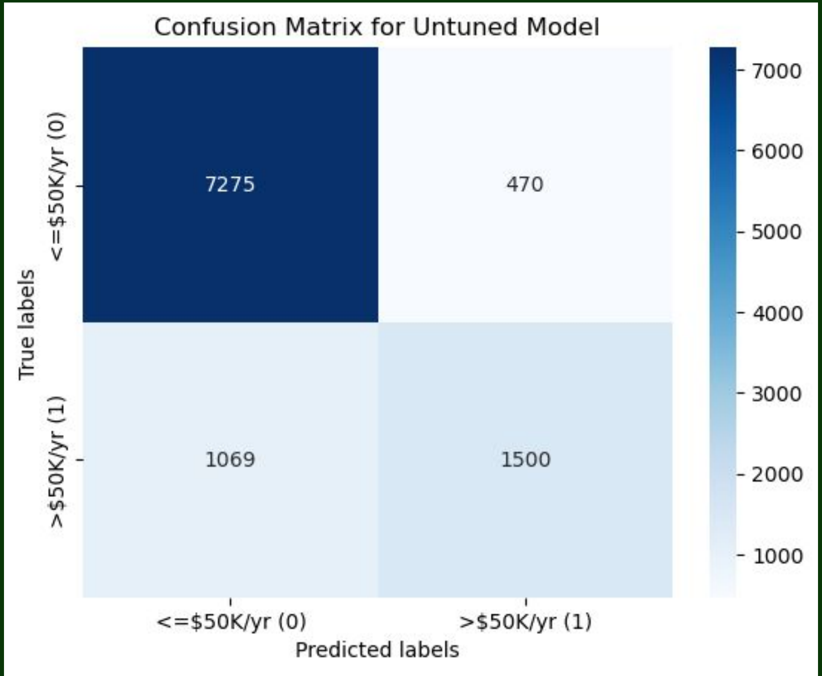
```
XGBClassifier: Train Accuracy = 0.8622, Test Accuracy = 0.8555, Balanced Accuracy = 0.7736, F1 Score = 0.6779
GradientBoostingClassifier: Train Accuracy = 0.8546, Test Accuracy = 0.8508, Balanced Accuracy = 0.7616, F1 Score = 0.6609
ExtraTreesClassifier: Train Accuracy = 0.8667, Test Accuracy = 0.8424, Balanced Accuracy = 0.7561, F1 Score = 0.6487
RandomForestClassifier: Train Accuracy = 0.8667, Test Accuracy = 0.8481, Balanced Accuracy = 0.7653, F1 Score = 0.6631
DecisionTreeClassifier: Train Accuracy = 0.8667, Test Accuracy = 0.8492, Balanced Accuracy = 0.7668, F1 Score = 0.6657
KNeighborsClassifier: Train Accuracy = 0.8118, Test Accuracy = 0.7900, Balanced Accuracy = 0.7092, F1 Score = 0.5652
AdaBoostClassifier: Train Accuracy = 0.8392, Test Accuracy = 0.8351, Balanced Accuracy = 0.7535, F1 Score = 0.6409
LogisticRegression: Train Accuracy = 0.8293, Test Accuracy = 0.8271, Balanced Accuracy = 0.7348, F1 Score = 0.6135
```

**XGBClassifier and GradientBoostingClassifier gave the top 2 best model scores**

Ingrid

# Confusion Matrix For Gradient Boosting Classifier

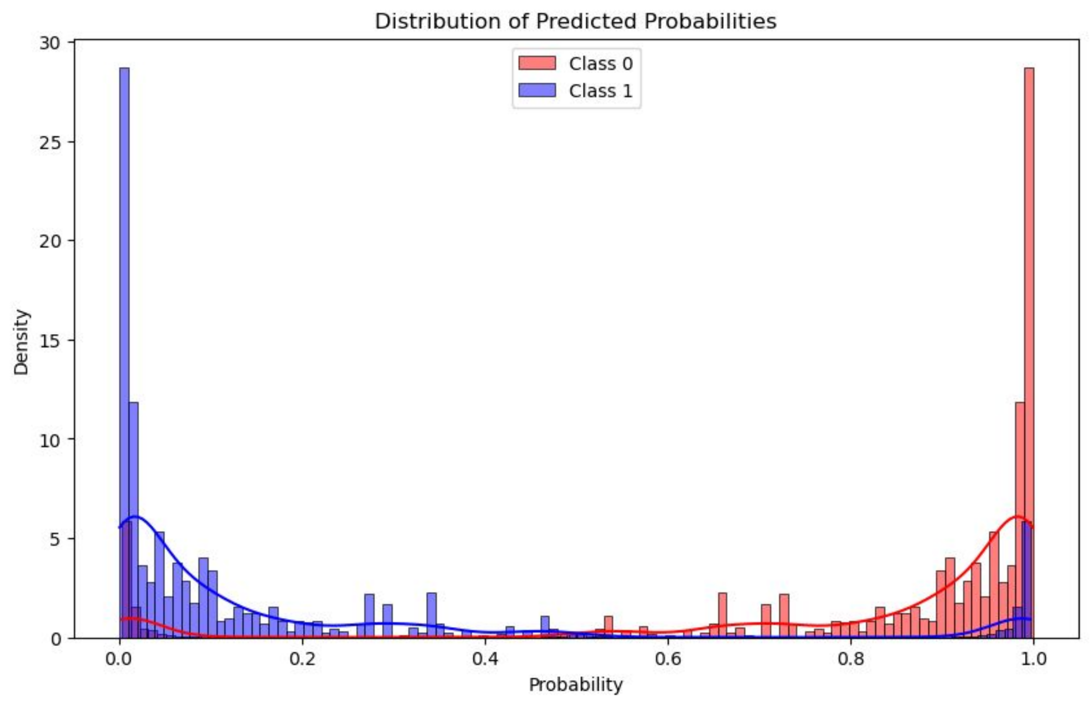
**(Train Acc=0.8546, Test Acc=0.8508, Balanced Acc Score=0.7616, F1 Score=0.6609, ROC AUC=0.87 )**

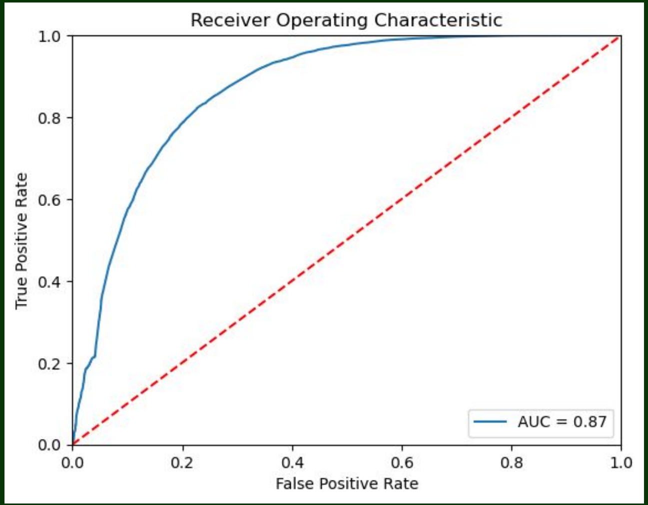


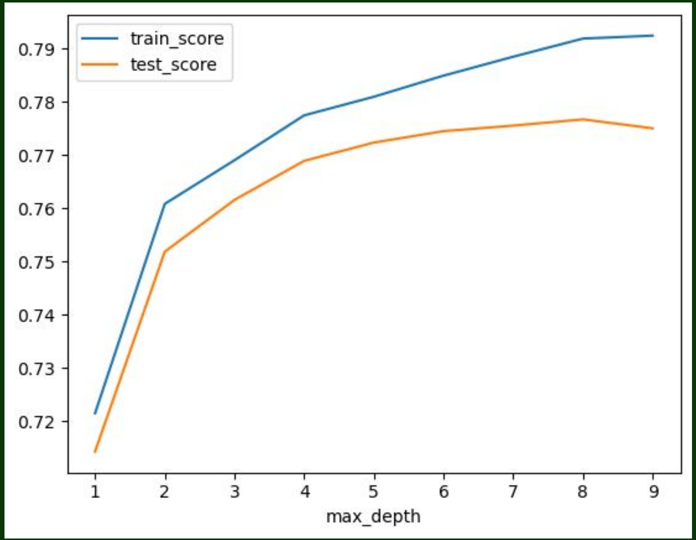

| Metric | Count | Meaning |
| --- | --- | --- |
| True Negatives (Top Left) | 7275 | Correctly predicted low income |
| False Positives (Top Right) | 470 | Mistakenly predicted high income |
| False Negatives (Bottom Left) | 1069 | Missed actual high income |
| True Positives (Bottom Right) | 1500 | Correctly predicted high income |

*Your 1st model is good!!*

# GradientBoosting Model



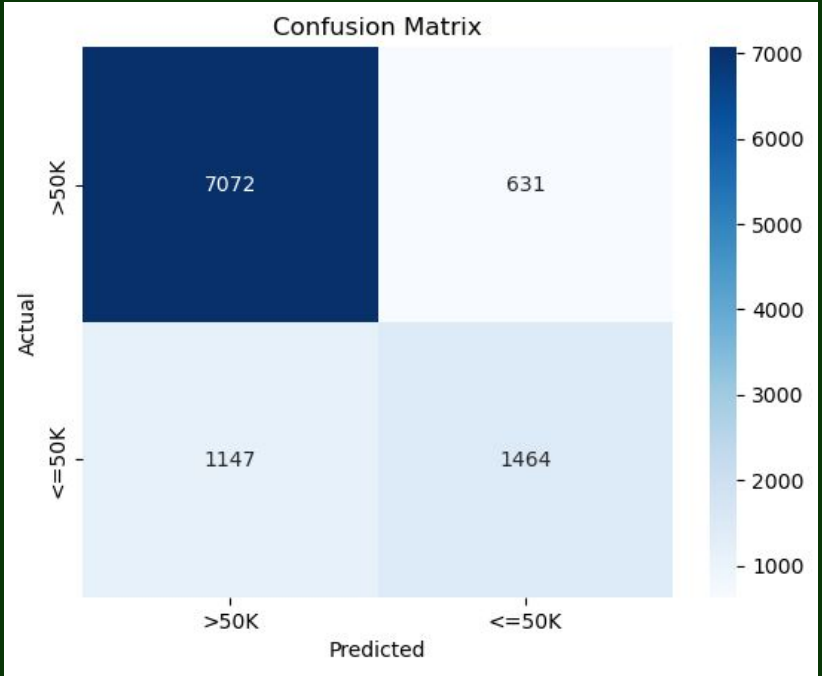



Ingrid

# Confusion Matrix For XGBoost Classifier

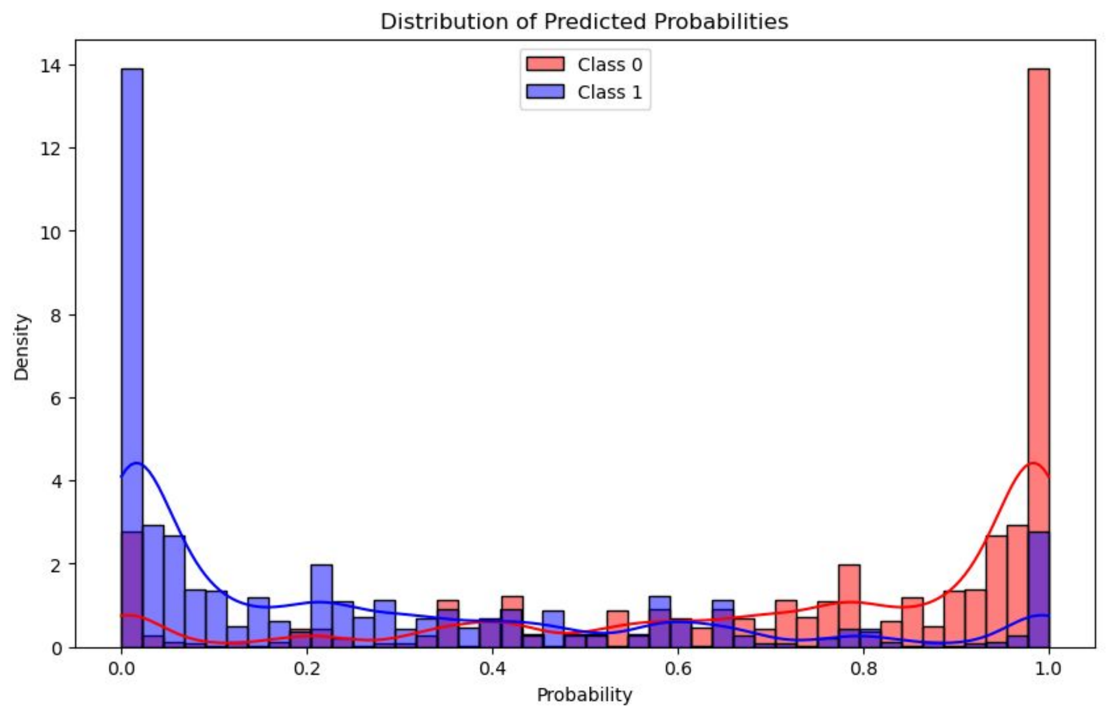
**(Train Acc=0.8622, Test Acc=0.8555, Balanced Acc Score=0.7736, F1 Score=0.6779, ROC AUC=0.92)**

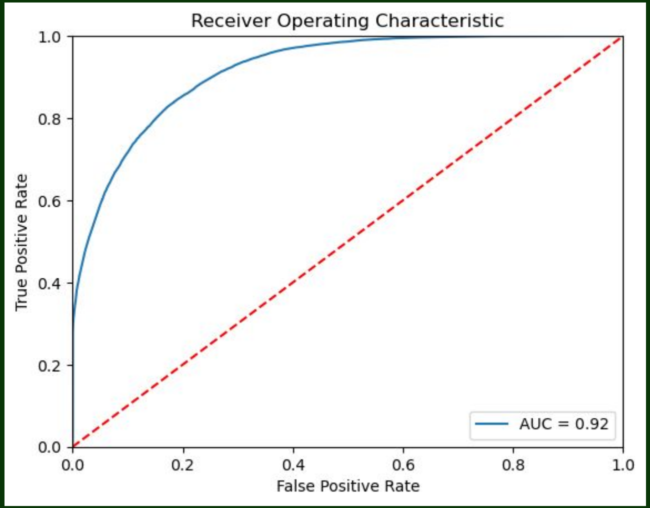


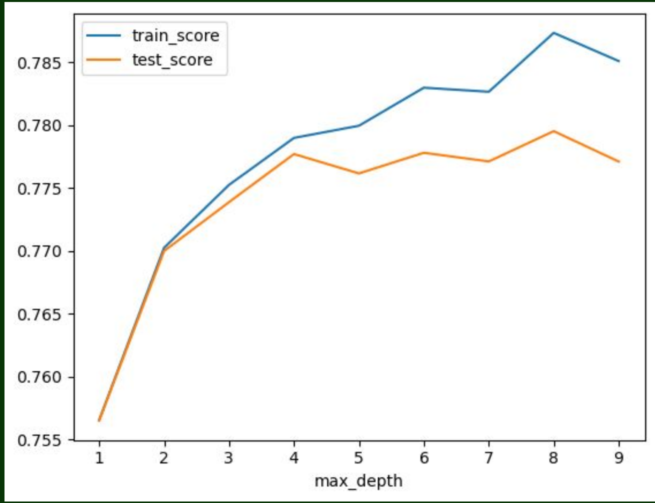

| Metric | Count | Meaning |
|---|---|---|
| True Negatives (Top Left) | 7072 | Correctly predicted low income |
| False Positives (Top Right) | 631 | Mistakenly predicted high income |
| False Negatives (Bottom Left) | 1147 | Missed actual high income |
| True Positives (Bottom Right) | 1464 | Correctly predicted high income |

***Your 2nd model is good too, which one is better?***

# XGBoost Model



# Which Model Wins?

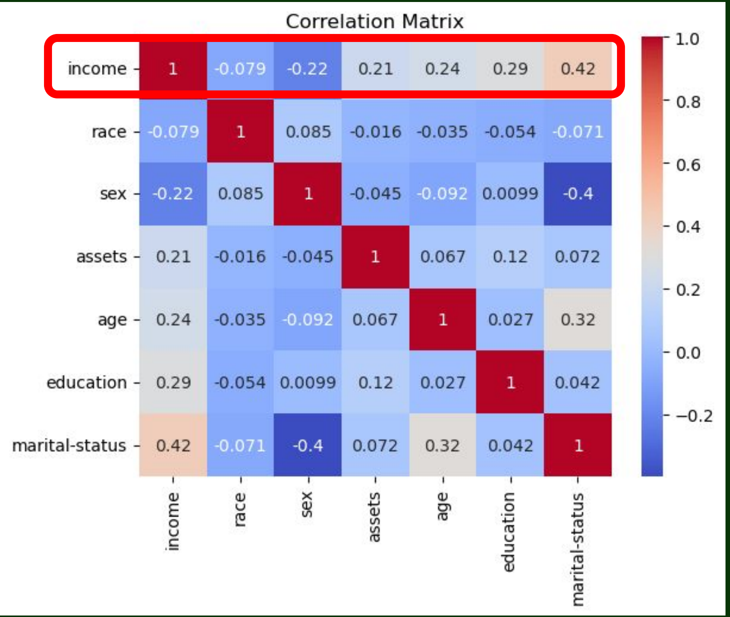
| XGBoost | Vs. | Gradient Boosting |
|---|---|---|
| ✔ (86% vs 85%) | Train and Test Accuracy | |
| ✔ (77% vs 76%) | Balanced Accuracy | |
| ✔ (68% vs 66%) | F1 Score | |
| ✔ (92% vs 87%) | ROC AUC | |
| | Precision for >$50K | ✔ (76% vs 70%) |
| | Overall accuracy correctness | ✔ (85% vs 83%) |
| ✔ (68% vs 66%) | F1 Score | |
| WINNER | Conclusion | RUNNER UP |

# Correlation Matrix - Income vs. Features

- Income level decreases
  - Race & Sex

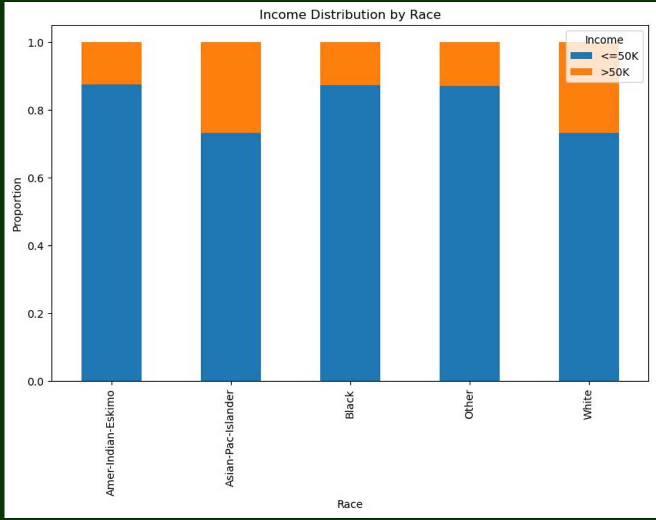
- Income level Increases
  - Assets, Age, Education and Marital Status

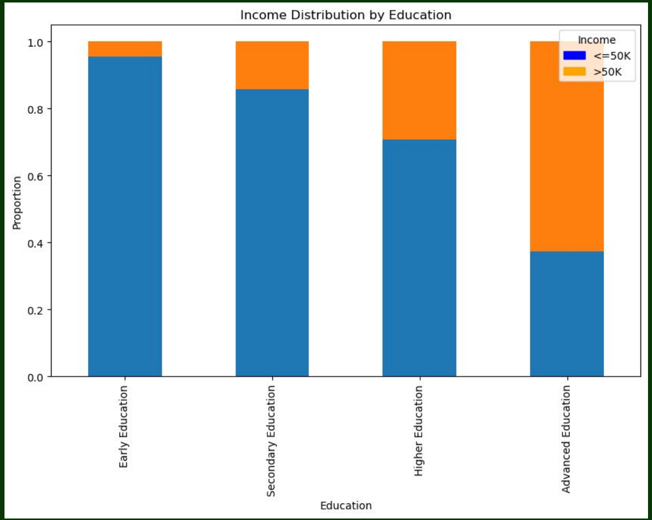
- Max correlation - Marital Status
- Min correlation - Race

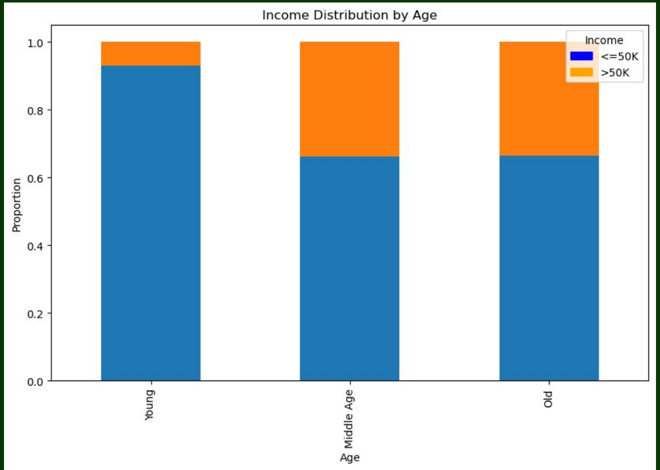


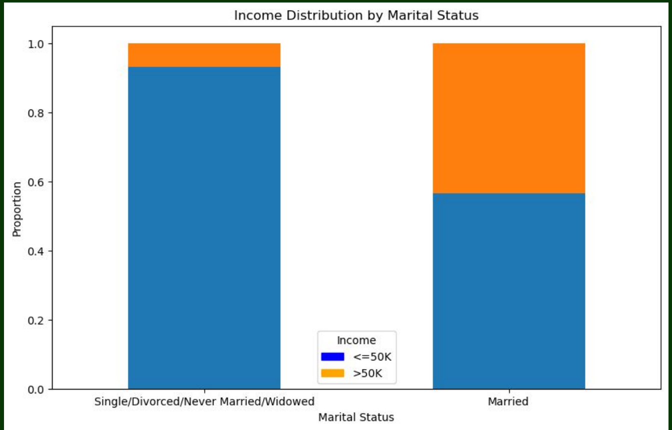
# Granular look at Income Distribution by Feature

- Race
- Education
- Age
- Marital Status



# Conclusion

**Raw Census Data** → **Cleaned the Data** → **Encoded, Feature Engineering** → **Model Evaluation** → **XGBOOST Model (83% Classification Accuracy)**

## Model Prediction - factors driving people making > $50K/year based on the 1994 census data

- Race = White and Asian/Pacific Islander
- Education = College Degree or Higher
- Age = 30-50 years old
- Marital Status = Married

# Future Work

Fine Tuning Model to improve F1 Score

Look at Income Balances

Explore demographic features for income < $50K

Cross granular feature analysis (i.e. - race/education vs income)

Census data is complicated!

…but it's worth analyzing to plan for a less chaotic future!

# Questions?


# Thank You