INCOME CLASSIFICATION

DATA SCIENTIST INTERVIEW PRESENTATION

Joseph Ekpenyong
05 June 2023

AGENDA

01. PROBLEM STATEMENT
Context

02. DATA EXPLORATION
Data and EDA

03. PREPROCESSING AND MODELING
Data modeling lifecycle

04. CONCLUSION
Results and insights

EXECUTIVE SUMMARY

INCOME INEQUALITY

Plays a significant role in determining individuals' opportunities and quality of life

COMPLEX FACTORS

Understanding factors that contribute to different income levels can be complex and challenging

MACHINE LEARNING

We leverage machine learning to uncover underlying patterns that impact income levels

FINDINGS

We learn that investment income, sex, age, education, and occupational roles are associated with income class

CONTEXT

DATA

US Census
Bureau

ACCESS

Information is publicly
available

QUESTION

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

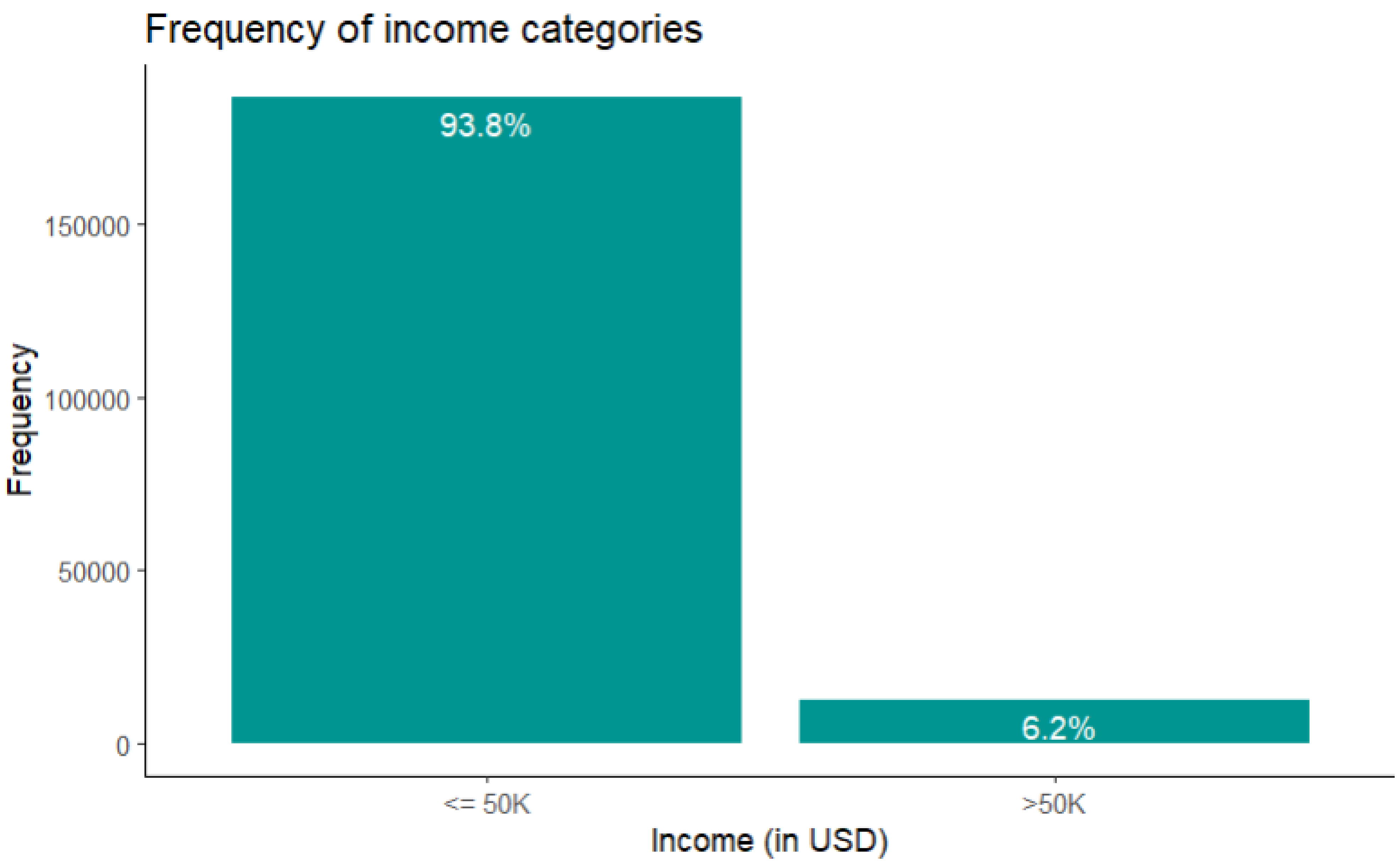
DATA SUMMARY

- Sample dataset from US Census Archive
- Unit of observation is an individual
- 300,000 observations and 42 variables
- Demographic, socioeconomic and employment-related attributes
- Binary target variable: income (<=50K or >50K)

DATA EXPLORATION

OBSERVATION

The target variable is highly imbalanced



DATA EXPLORATION

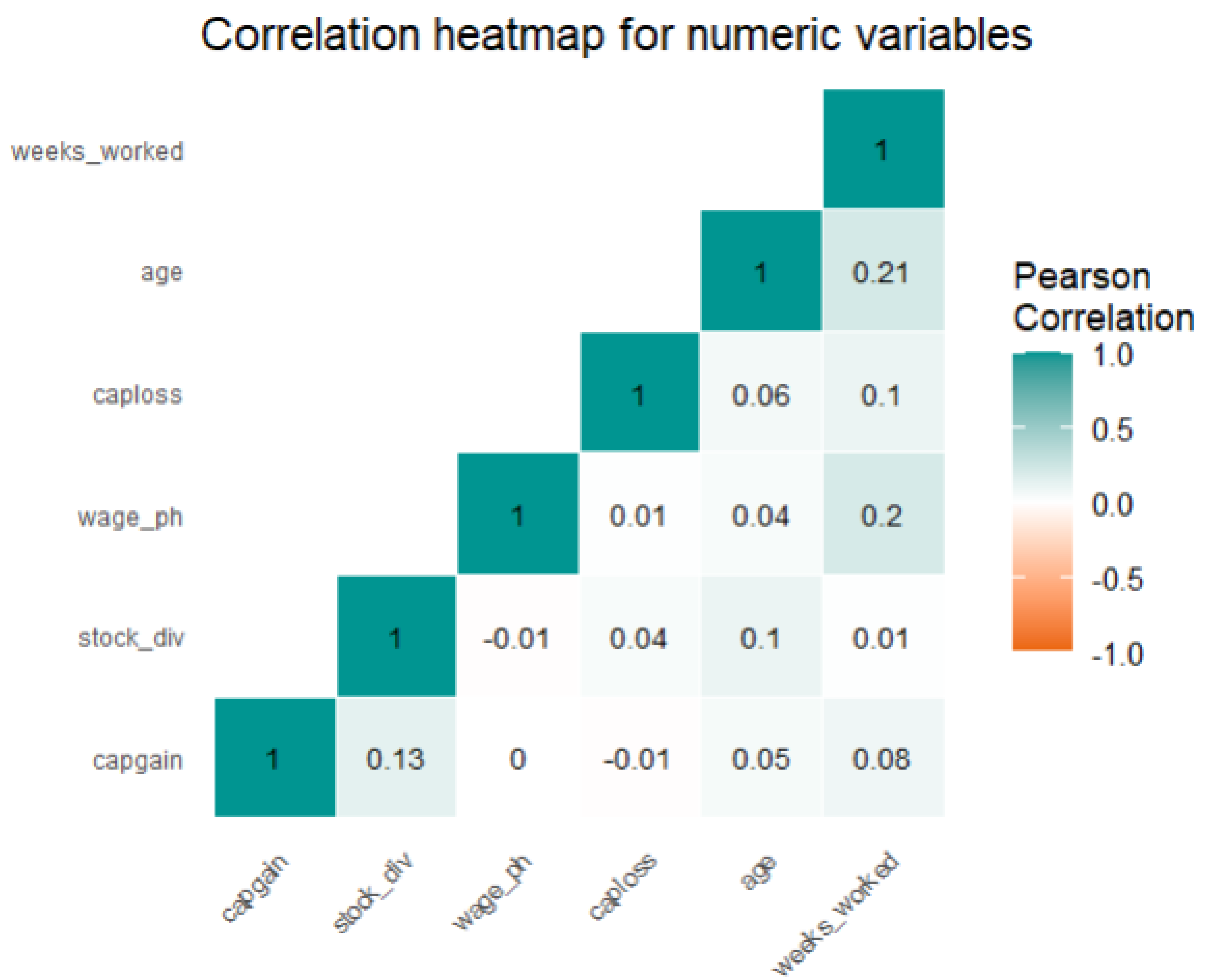
OBSERVATION

Children (persons under 15) do not work

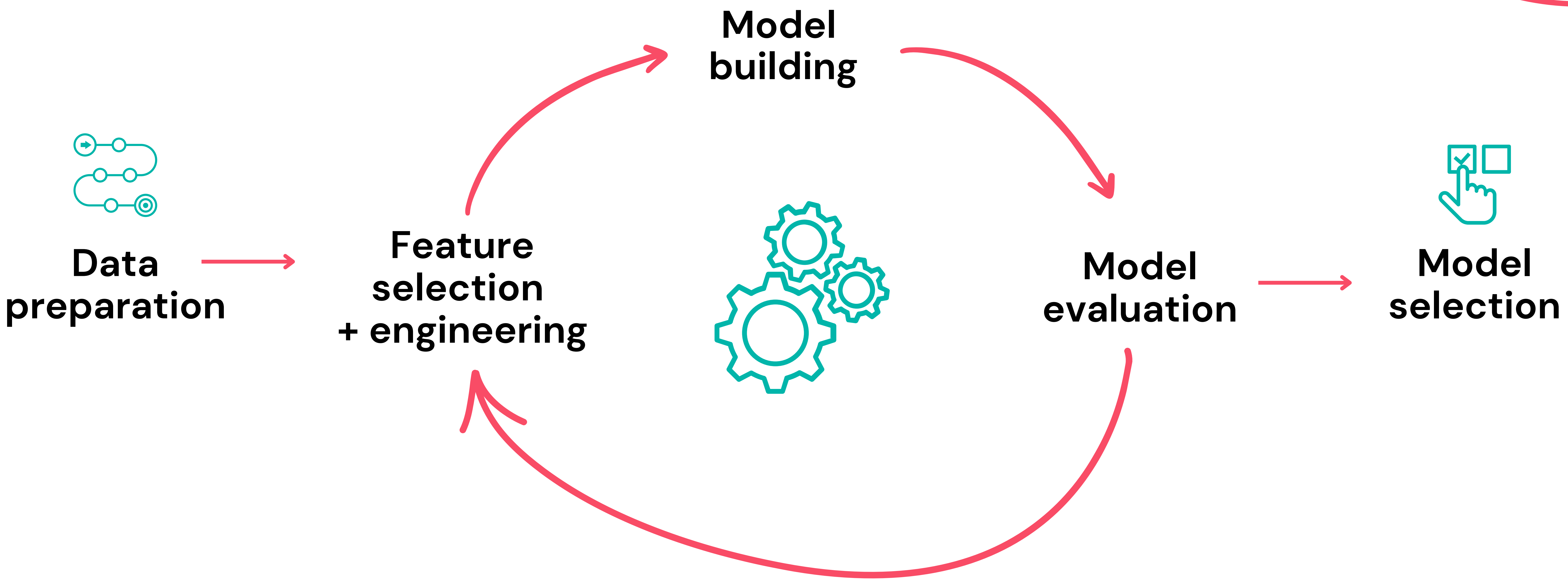
DATA EXPLORATION

OBSERVATION

There is weak-to-no correlation among numeric variables



OVERVIEW OF MODELING STEPS



DATA PREPARATION

Outliers are transformed later

Train-test split: 2 to 1

Missing values: Hispanic origin DROP MISSING

Outliers: capital gain, stock dividends, wage per hour
TRANSFORM LATER

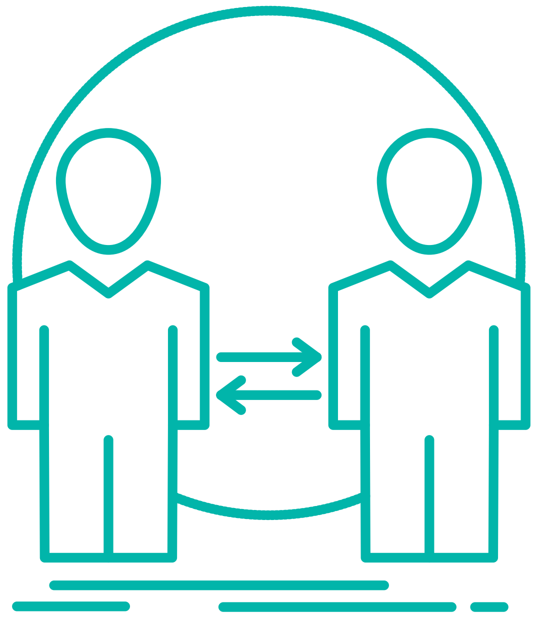
FEATURE SELECTION + ENGINEERING

Apply to train and test sets

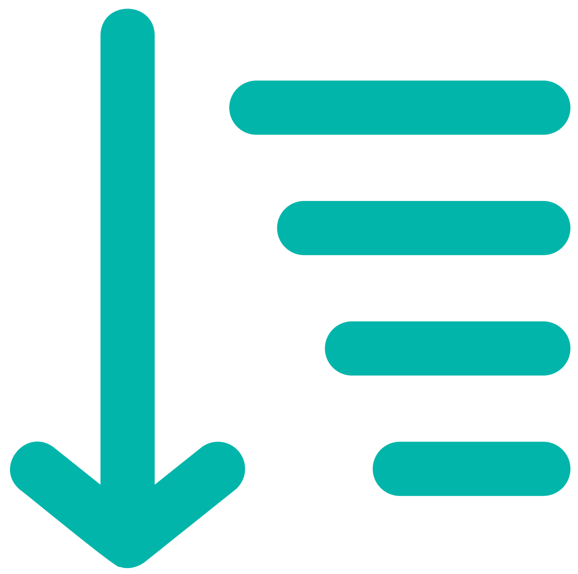
1. Investment income
2. Drop columns
3. Map categorical variables
4. One-hot encode
5. Standardize numerical variables

TARGET CLASS IMBALANCE

Undersampling
to address
class
imbalance



Duplicates



Drop children

THIS IS A CLASSIFICATION PROBLEM

Logistic Regression

- Linear classifier
- Predicts probabilities
- Interpretatable

Random Forests

- Ensemble learning method
- Handles complexity better
- Resistant to overfitting

MODEL EVALUATION (with undersampling)

Logistic Regression

```
Confusion matrix:
 [[90930  2646]
 [ 3692  2494]]

Classification Report:
              precision    recall  f1-score   support

    - 50000.       0.96      0.97      0.97     93576
     50000+.       0.49      0.40      0.44      6186

    accuracy                           0.94     99762
   macro avg       0.72      0.69      0.70     99762
weighted avg       0.93      0.94      0.93     99762
```

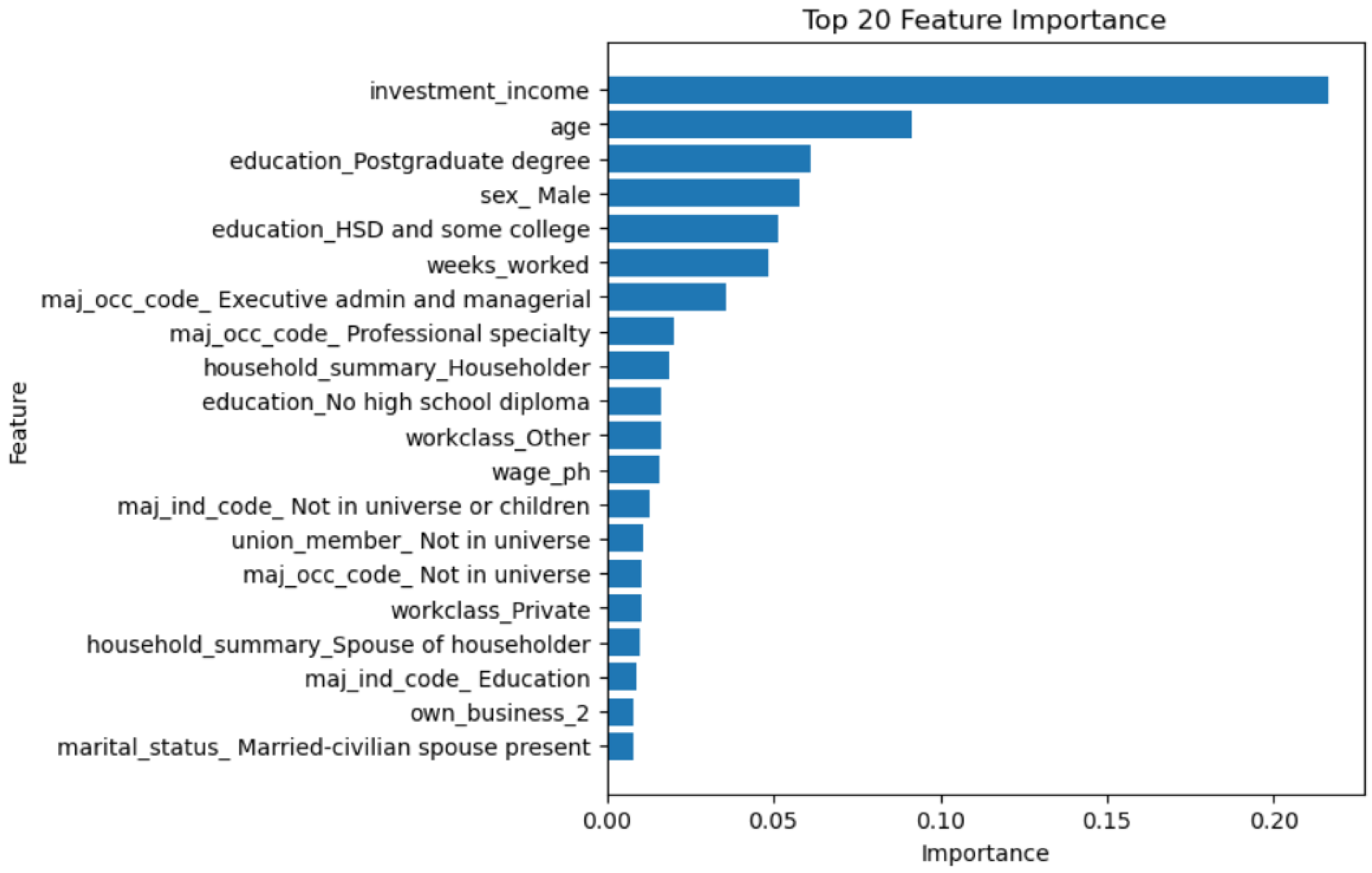
Random Forest

```
Confusion matrix:
 [[89745  3831]
 [ 3718  2468]]

Classification Report:
              precision    recall  f1-score   support

    - 50000.       0.96      0.96      0.96     93576
     50000+.       0.39      0.40      0.40      6186

    accuracy                           0.92     99762
   macro avg       0.68      0.68      0.68     99762
weighted avg       0.92      0.92      0.92     99762
```

FINAL MODEL: Random Forest

OBSERVATION

Investment income, age, education, and being an executive are all important characteristics associated with income class.



FUTURE WORK

Feature Engineering

Spend more time exploring relationships between variables

Best Parameters

Use grid search to select the best parameters for a model

Robust Metrics

Build more robust model evaluation metrics

What would we do if we had more time and computing power?

CONCLUSION

RESULTS AND FINDINGS

What characteristics are associated with a person making more or less than $50,000?

- **Investment Income** provides insights into an individual's financial portfolio and wealth accumulation.
- **Sex or gender** disparities exist in income levels and should be approached with caution.
- **Occupational roles** can be associated with higher incomes.
- **Age** correlates with work experience and seniority level in a job.
- **Education** can provide individuals with the knowledge and skills required for higher-paying professions.