AGENDA

INTRODUCTION

Adult Census Dataset

The Adult census data is a dataset containing information on individuals in the United States, including their demographic and socio-economic characteristics, and whether their income is above or below $50,000 per year.

Problem Statement

Can we predict an individual's income based on their characteristics?

Objectives

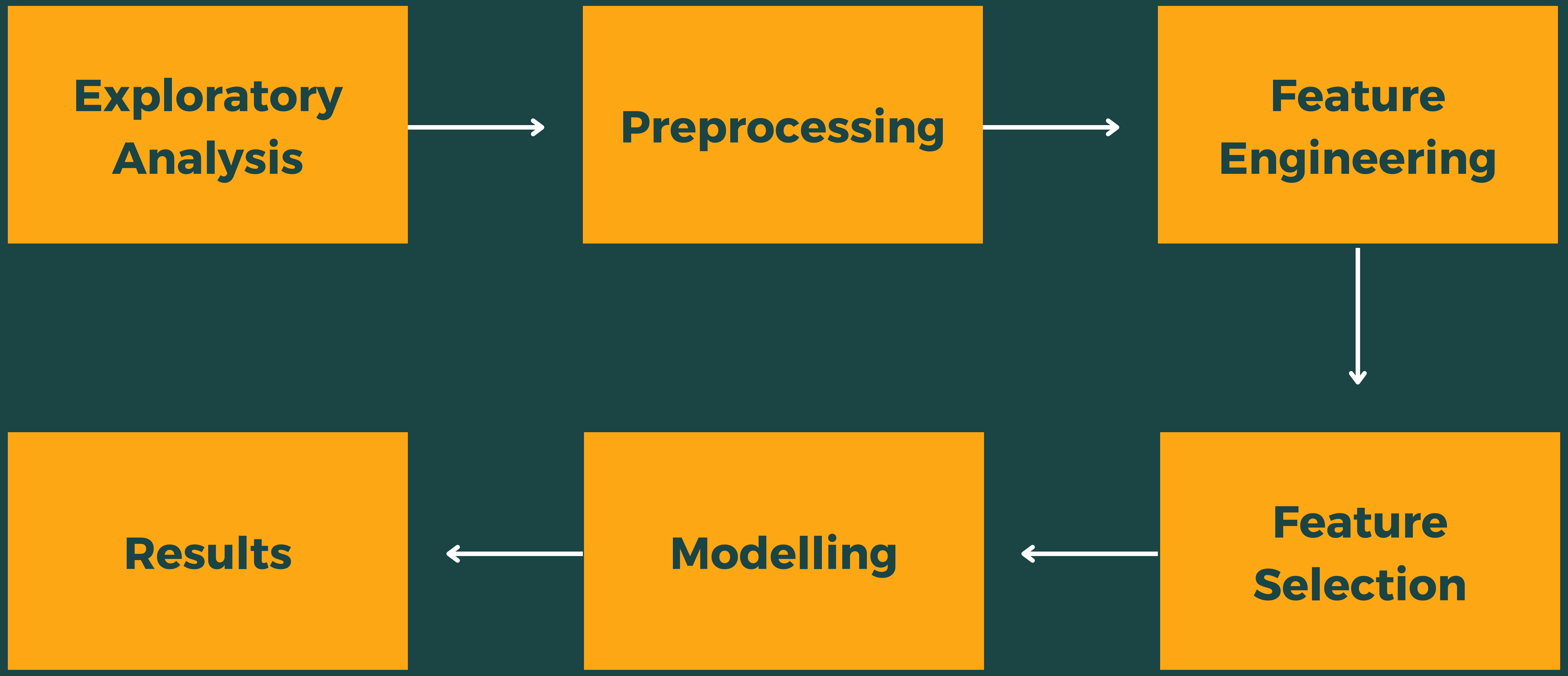
It is significant because it can help researchers better understand the factors that contribute to income inequality in Society.

Solution

Demonstrate an effective methodology for predicting income from the Adult census data and provide insights into contributing factors.

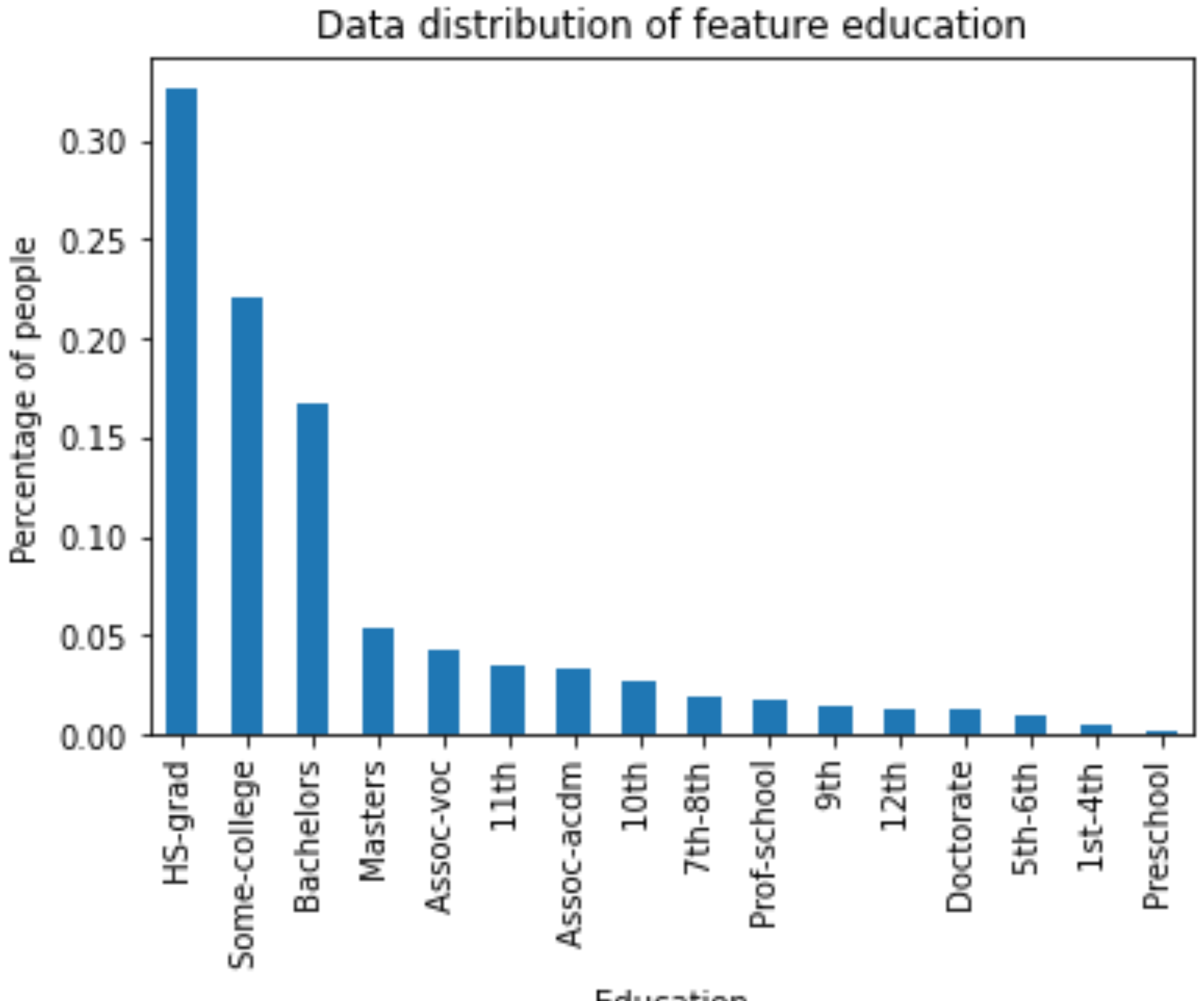
Back to Agenda

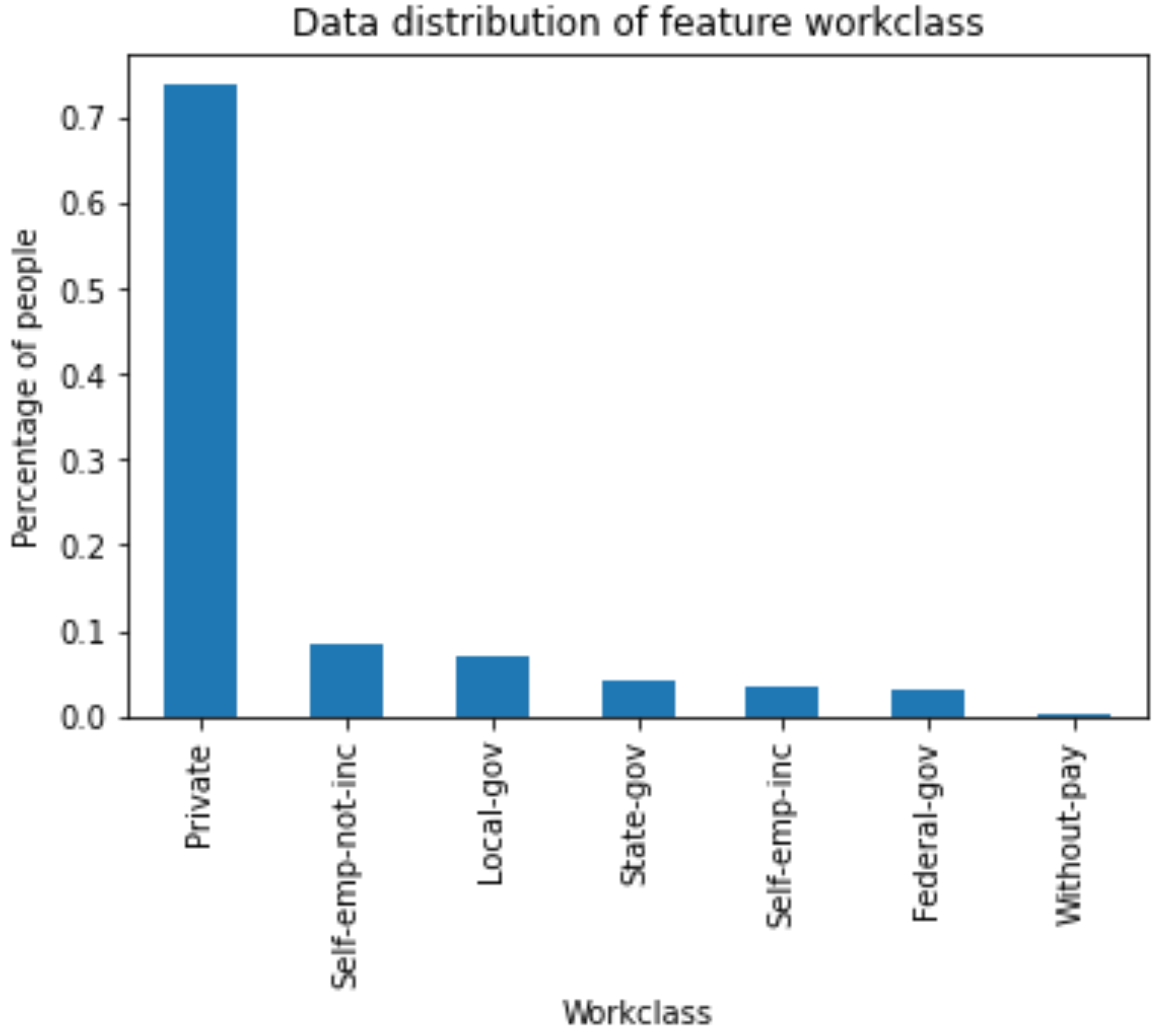
METHODOLOGY

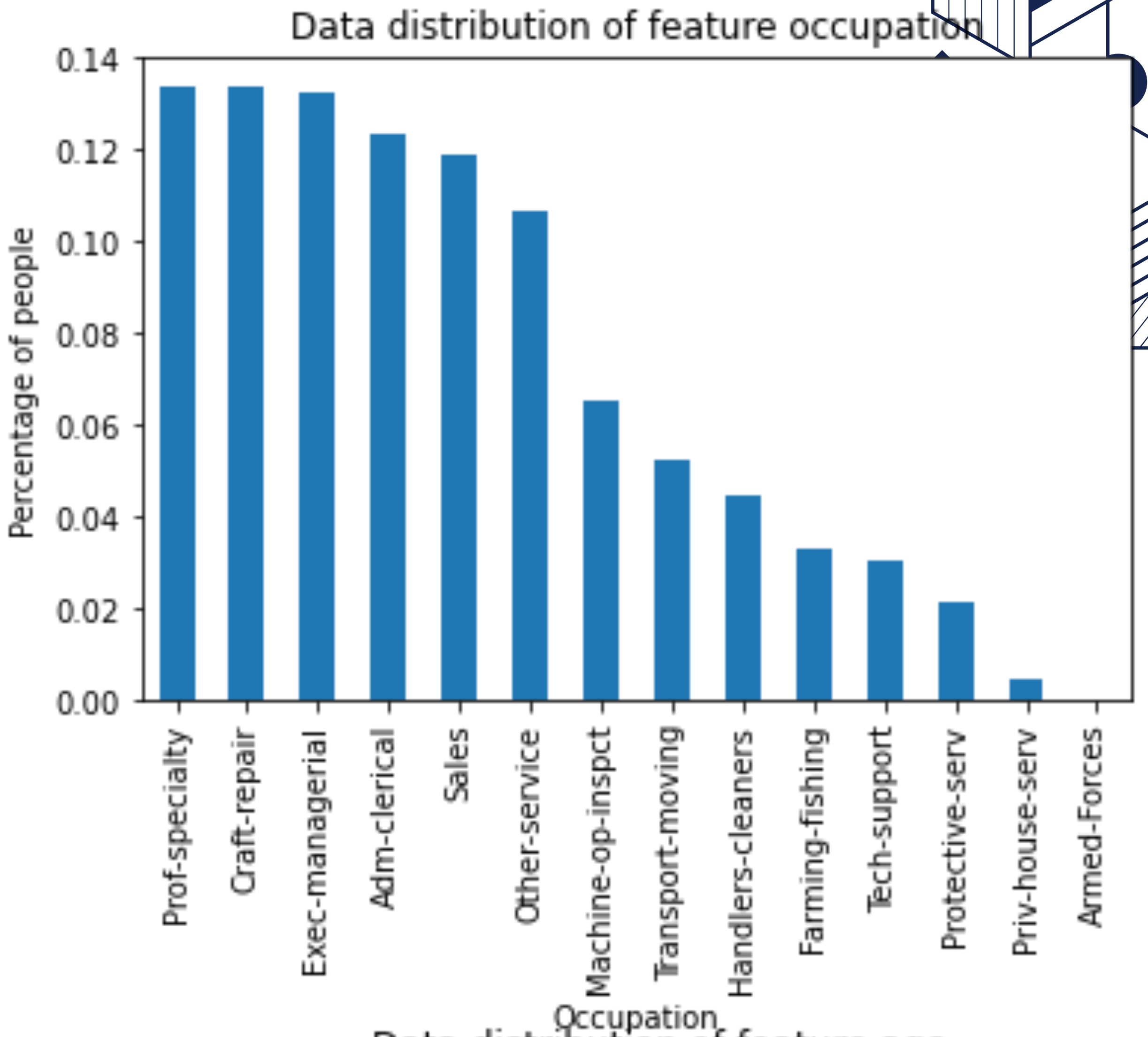


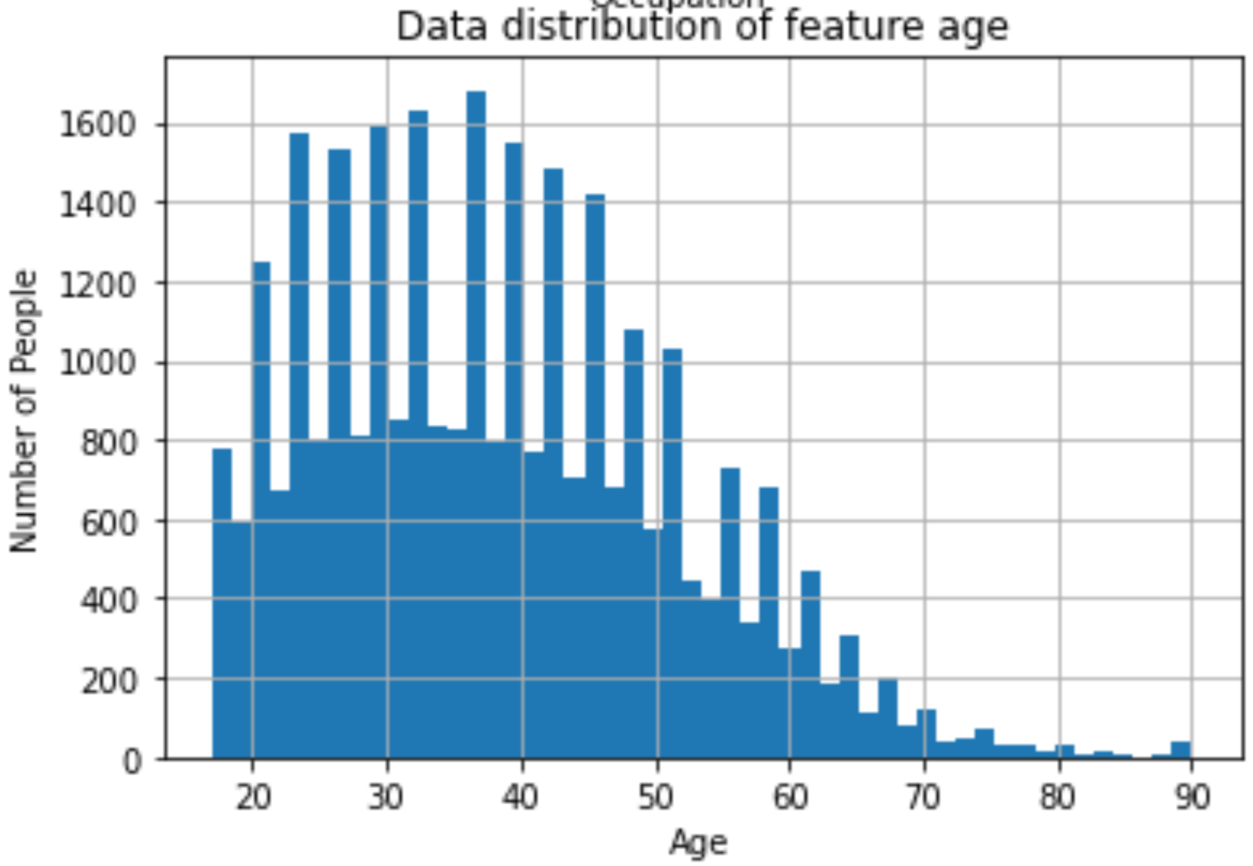
EXPLORATORY ANALYSIS

IMBALANCED CLASSES IN ADULT CENSUS DATA









PREPROCESSING

Handling Missing Values

The Adult census data contains missing values, which need to be handled before the data can be analyzed. Rows with missing data were dropped and some redundant columns were dropped.

Categorical Variables

The Adult census data contains several categorical variables, such as race and occupation, that need to be converted into numerical variables before they can be used in predictive models. This can be done using techniques such as one-hot encoding or label encoding.

Feature Engineering

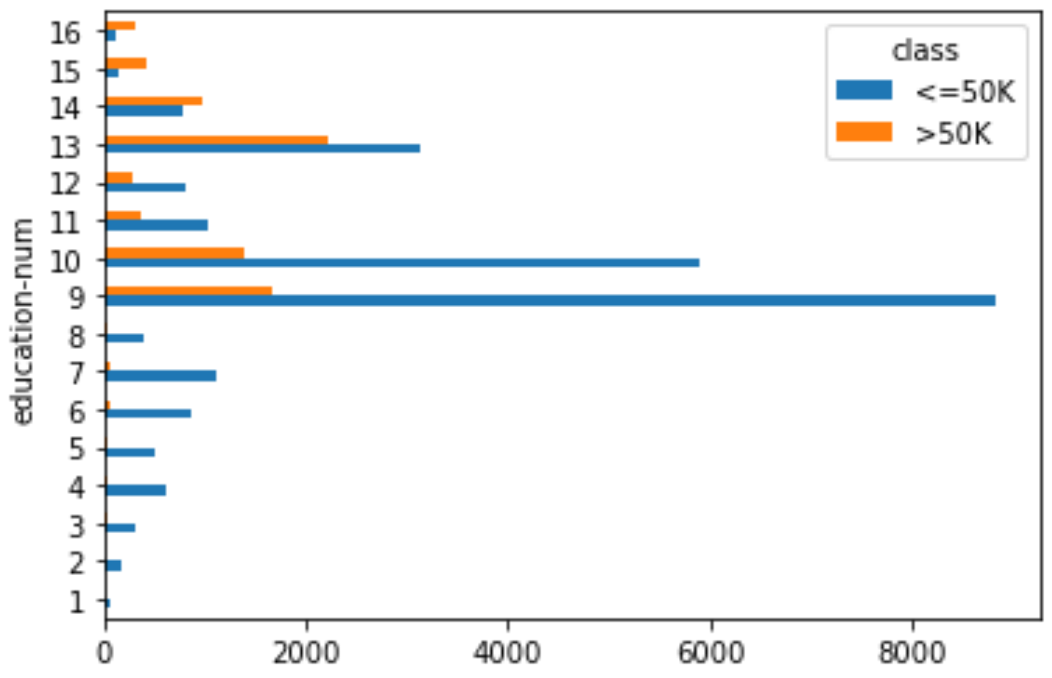
We modified some features in the data to reduce complexity and make more balanced

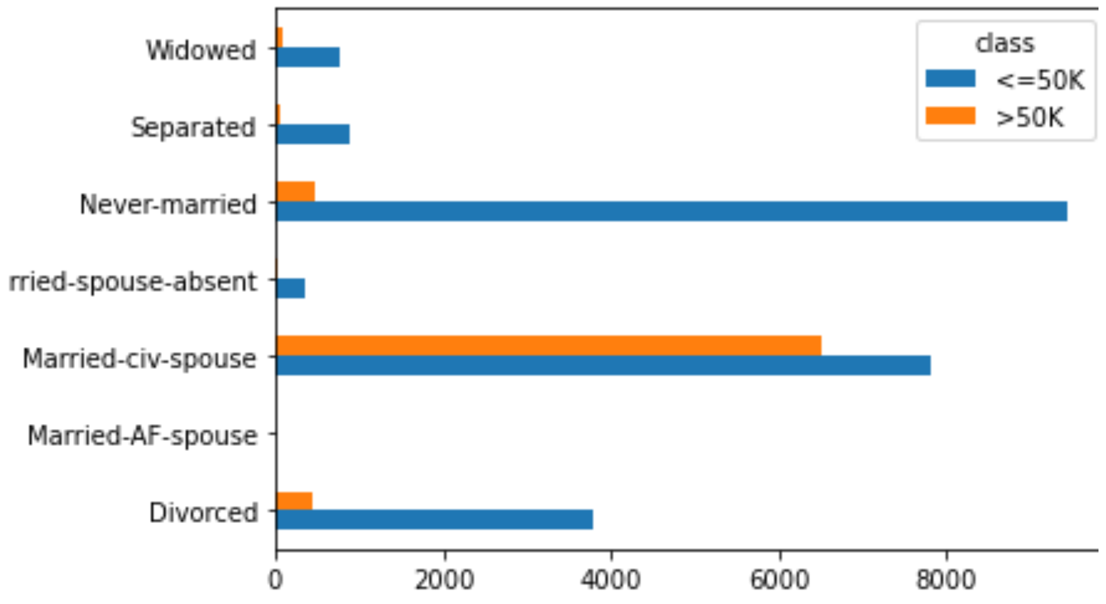
Feature Selection

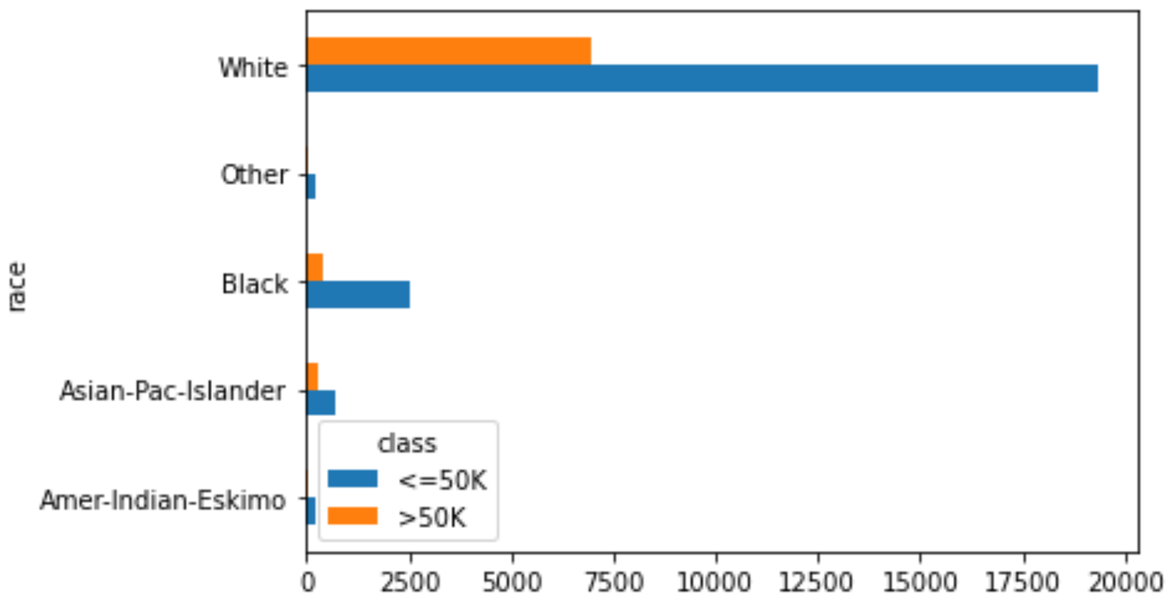
Some Algorithms such as Random Forest, Chi-square Feature Selection were done to select the most important features for the model.

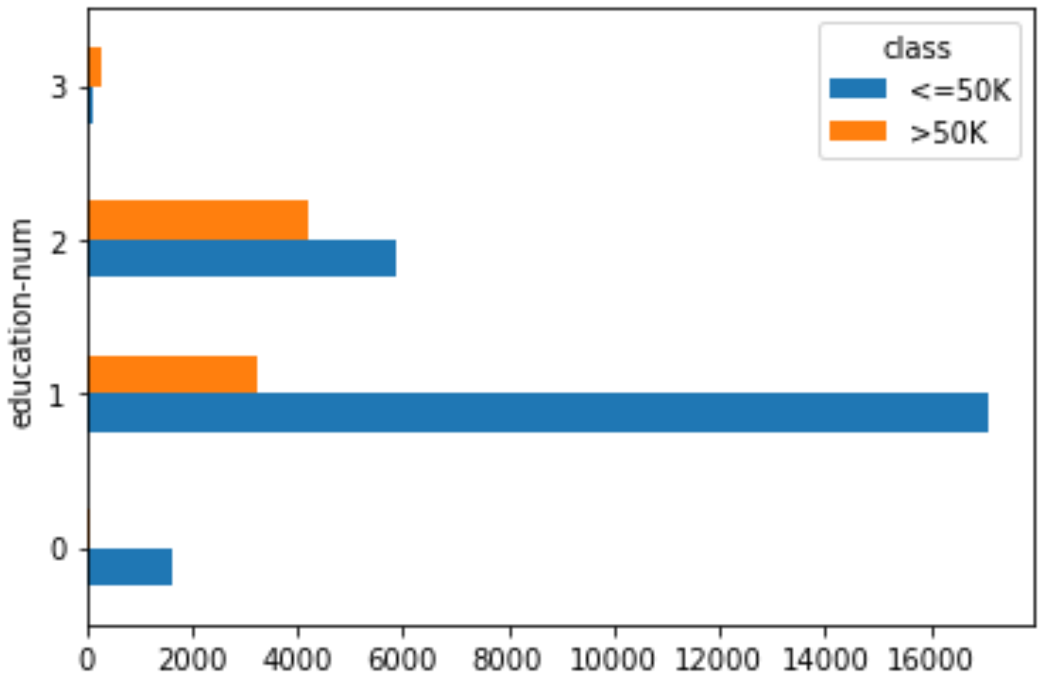
Back to Agenda

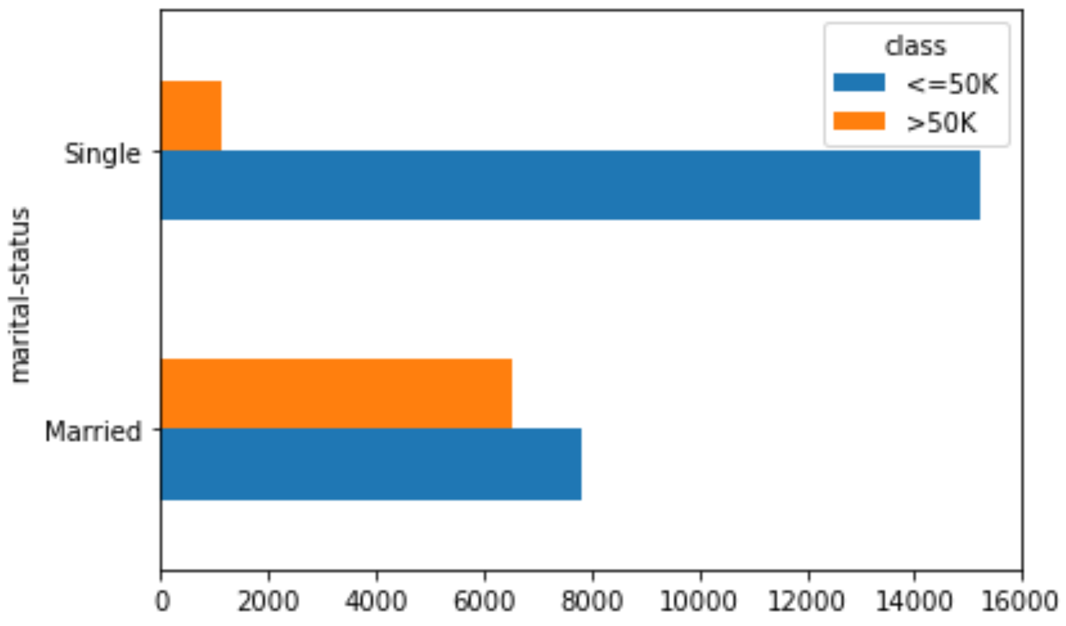
FEATURE ENGINEERING

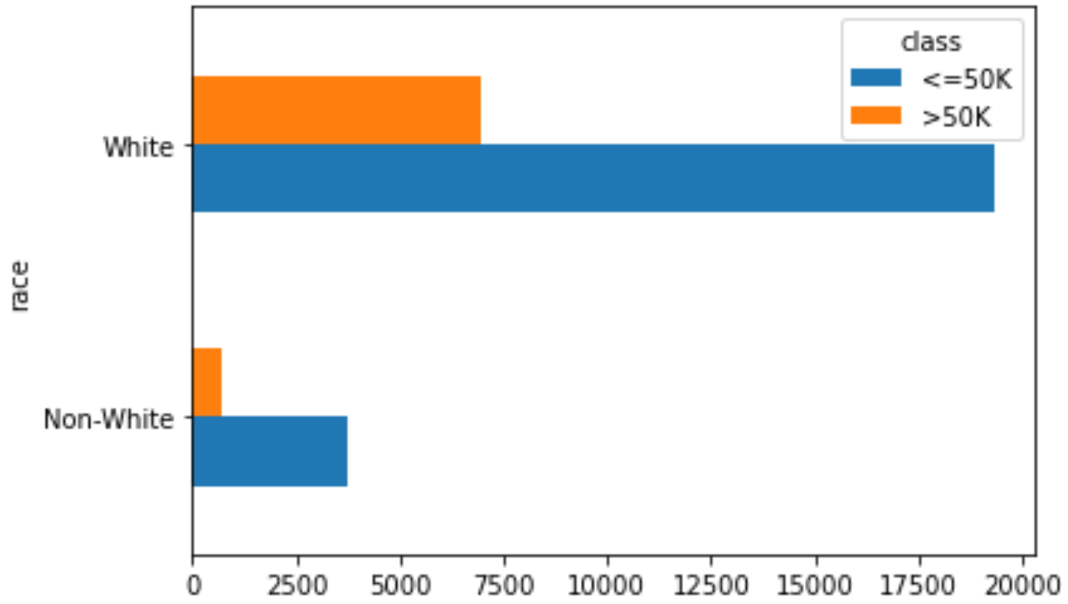






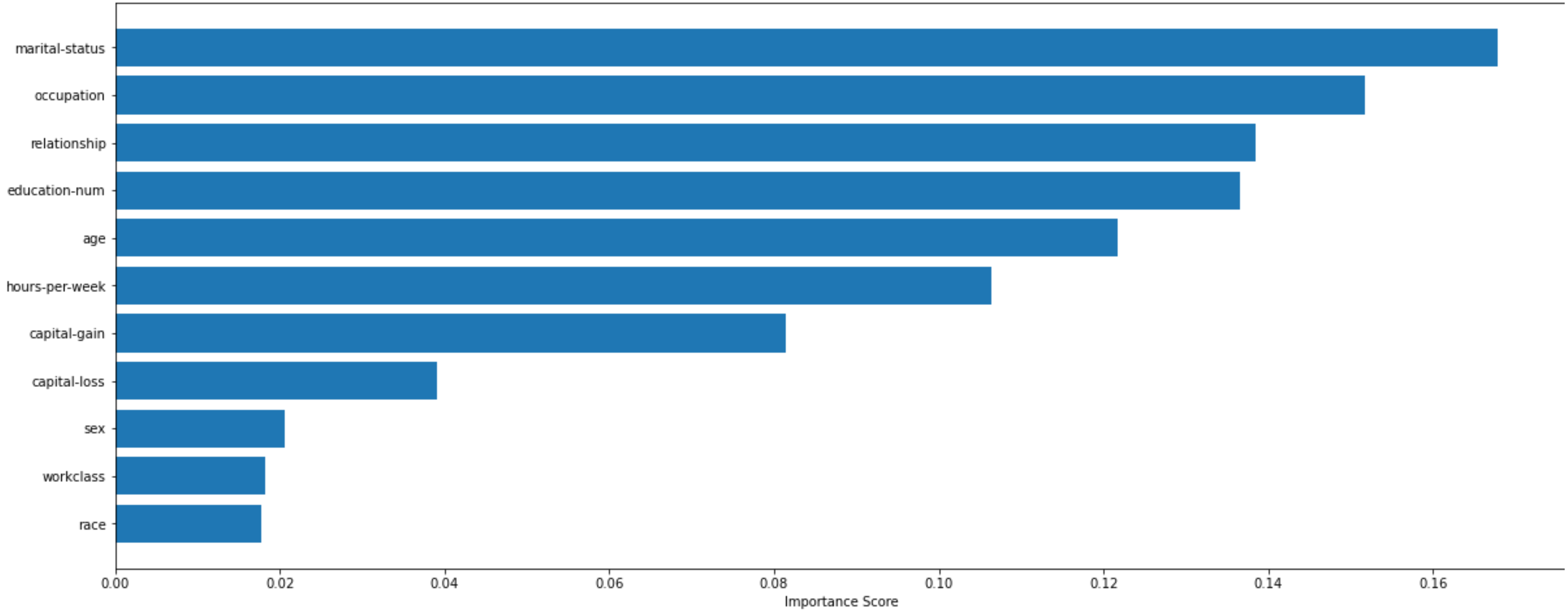






FEATURE SELECTION

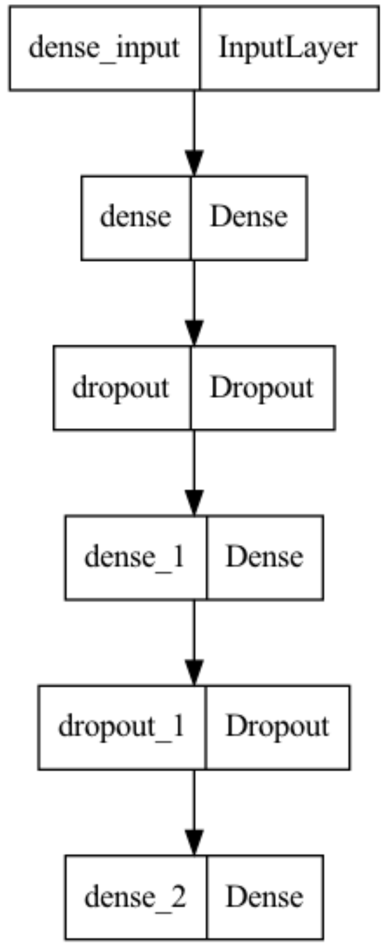
Based on the Random Forest feature importance tests, 3 common features with the least importance are sex, race, and work class.

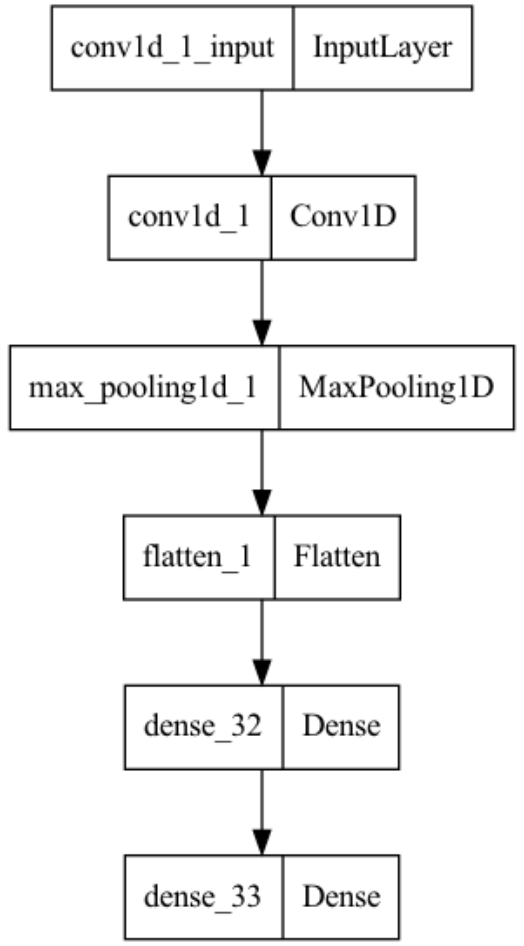


Back to Agenda

MODELLING

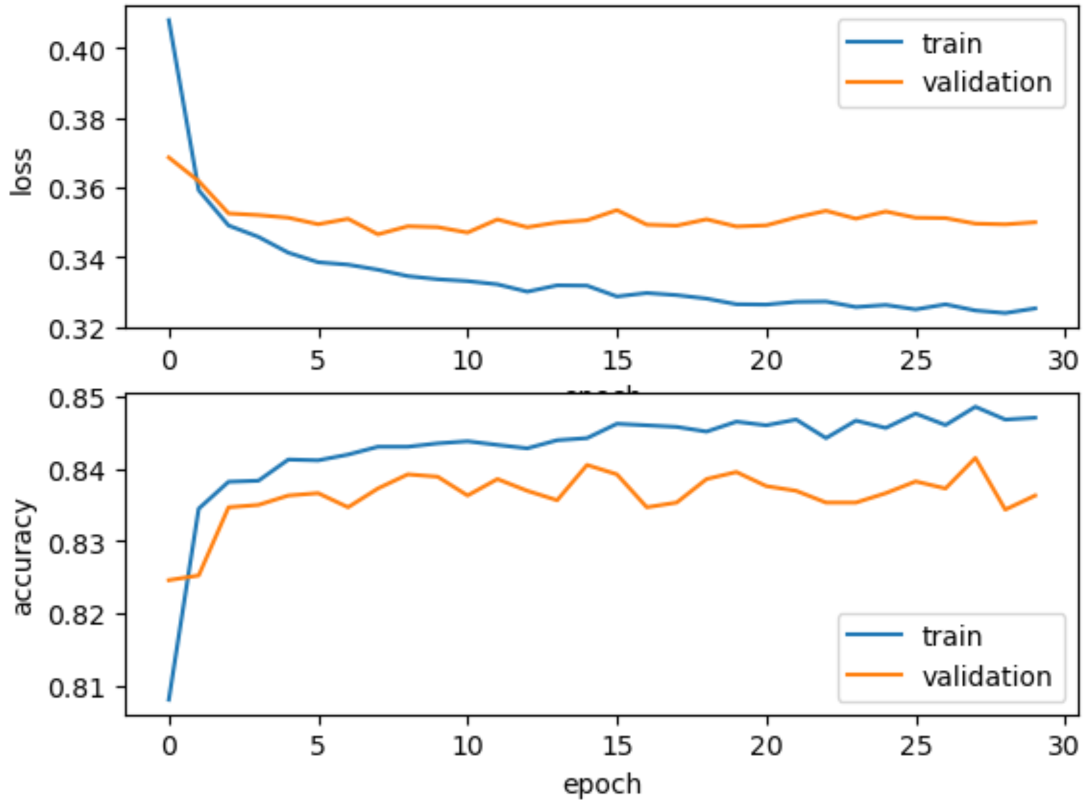
- Various neural network models were explored. DNN and CNN were the most common architectures.
- Model performance was evaluated using accuracy, roc-auc score, precision, recall, and F1 score.
- Models were optimized through hyperparameter tuning and feature selection.
- Models were trained using GPUs on Google Colab and Kaggle to utilize parallelisation.

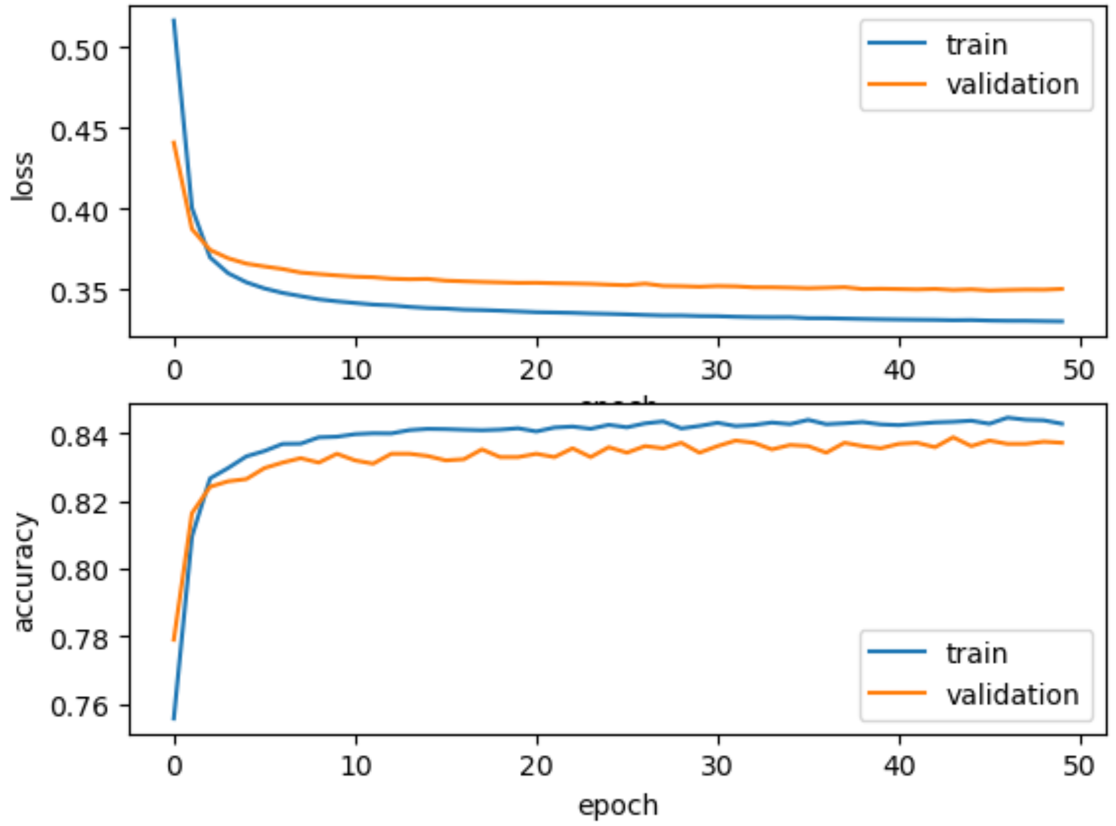




Dense Neural Network Architecture

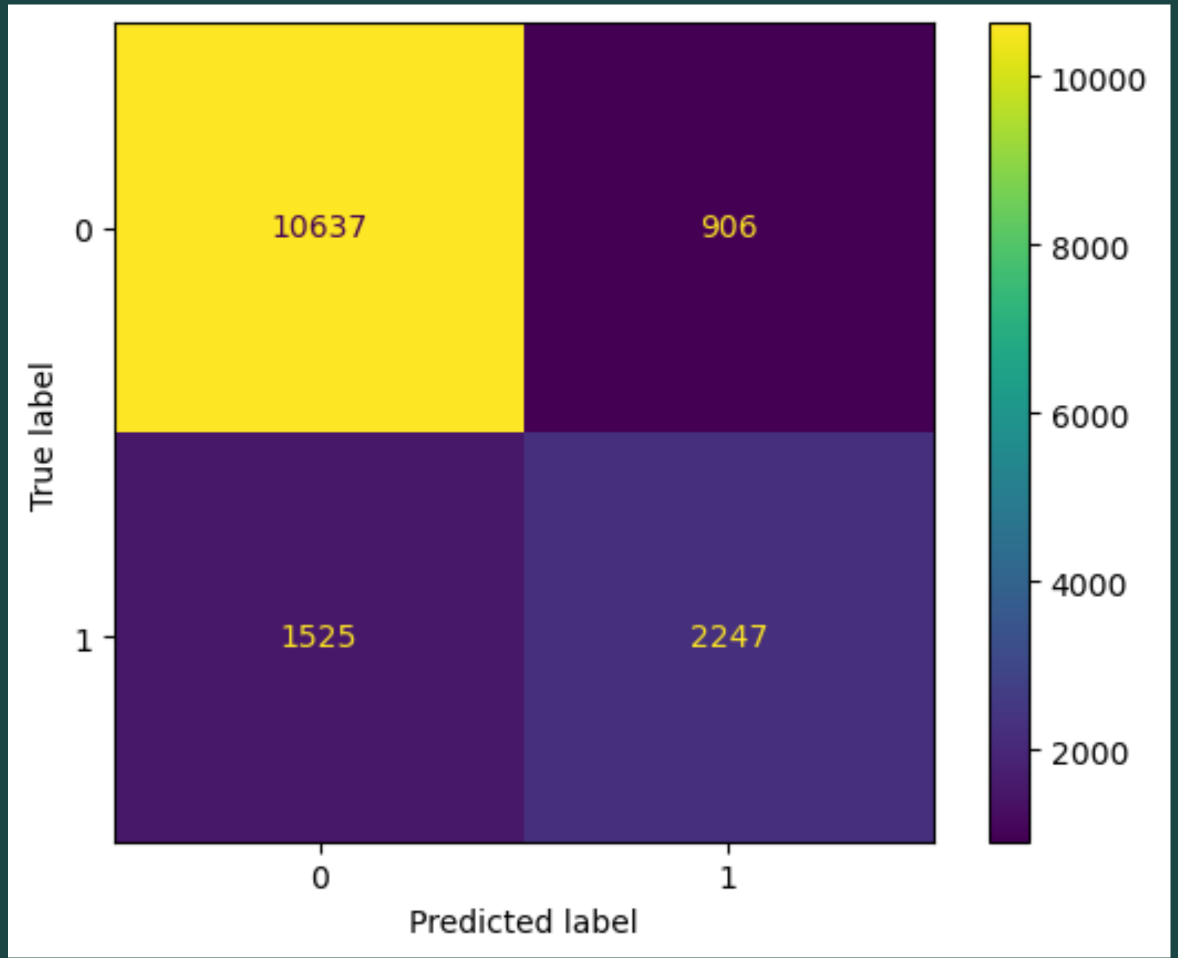
Convolution Neural Network Architecture

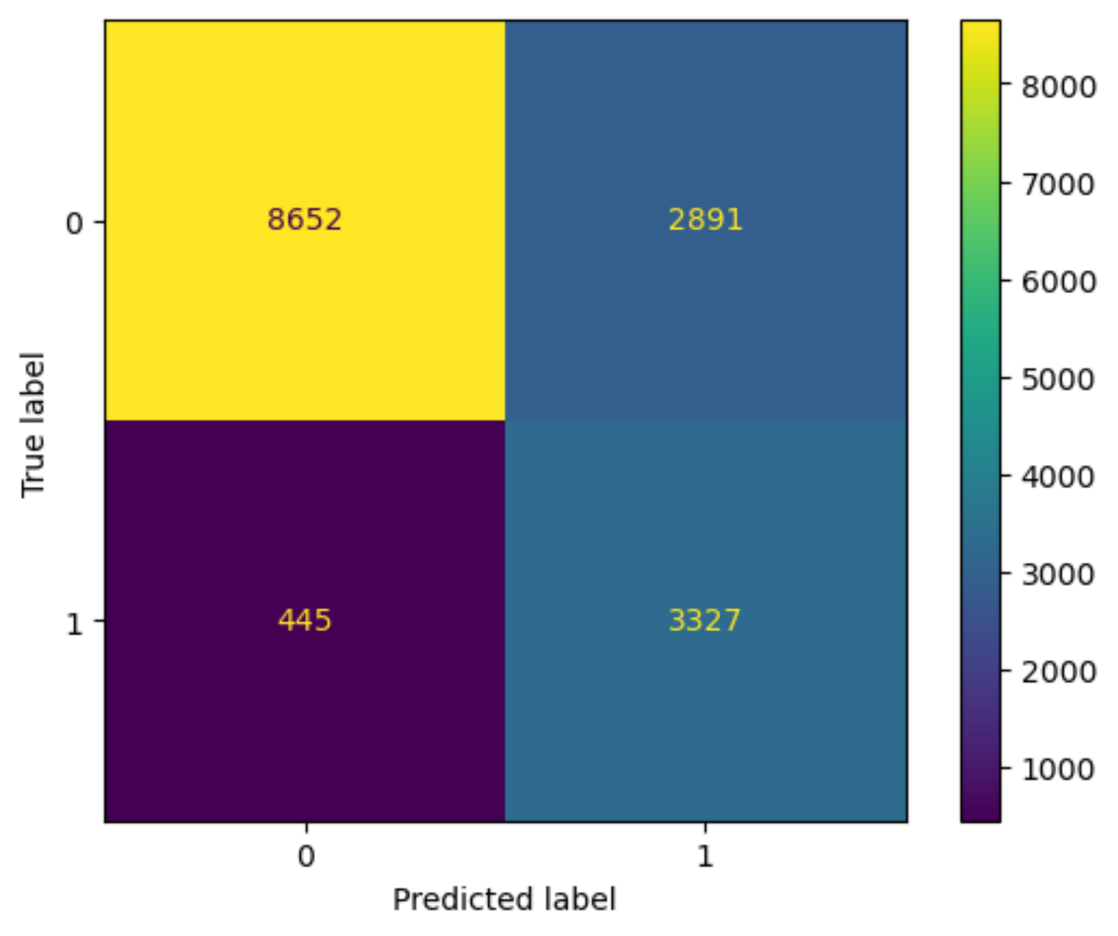




RESULTS FOR IMBALANCED TRAINING SET

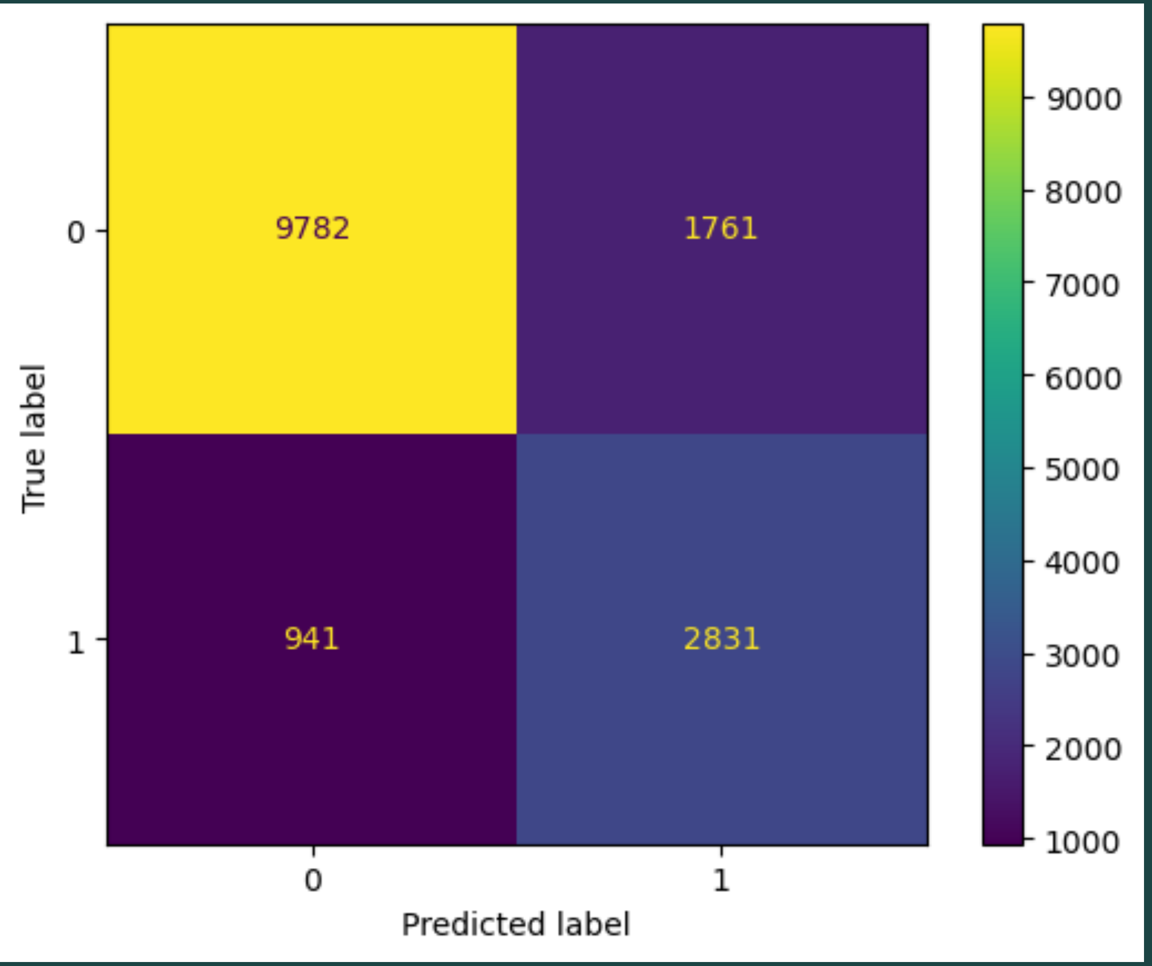
Default Threshold (0.5)	Best Threshold = 0.212586
Accuracy Score: 0.85	Accuracy Score: 0.78
Precision: 0.71	Precision: 0.53
Recall: 0.59	Recall: 0.88
F1: 0.64	F1: 0.66
AUC: 0.75	AUC: 0.81

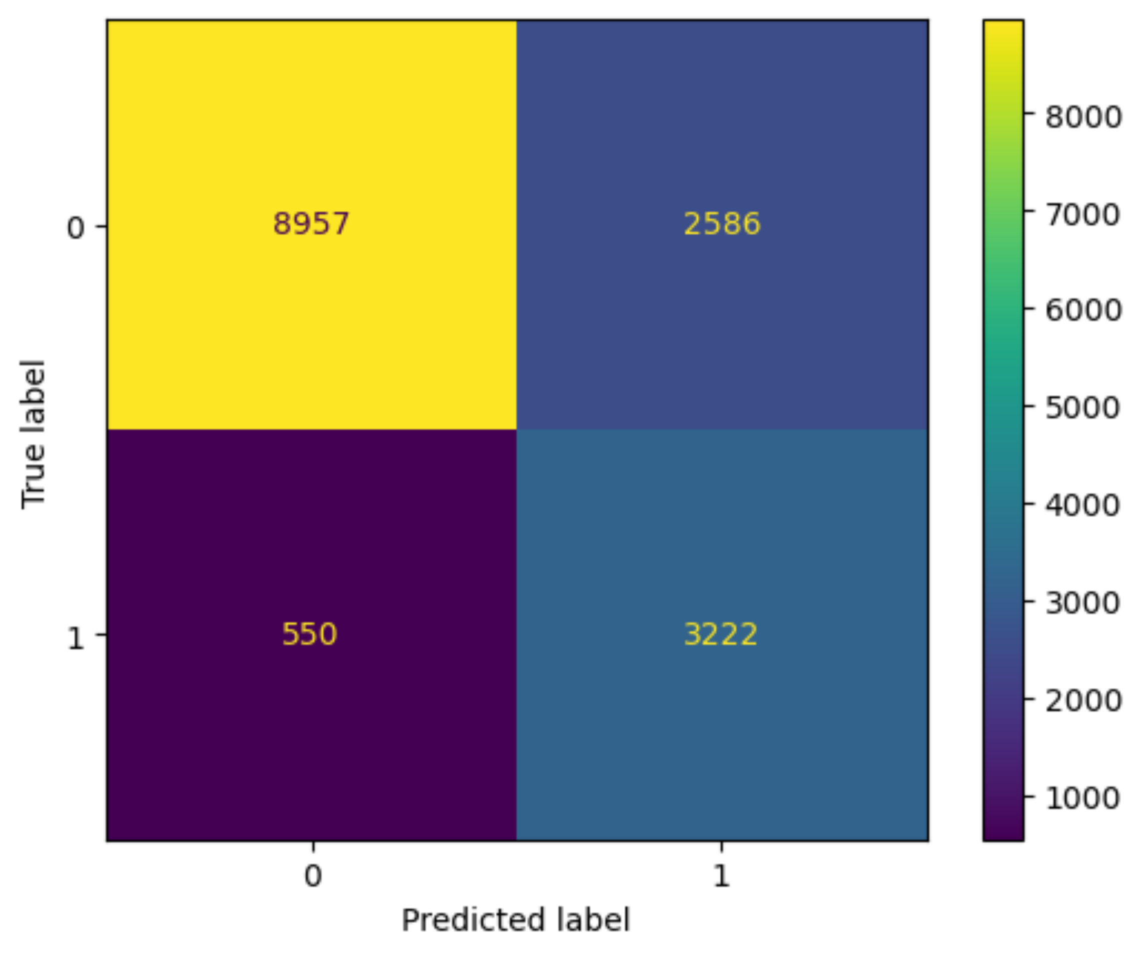




RESULTS FOR OVERSAMPLED TRAINING SET

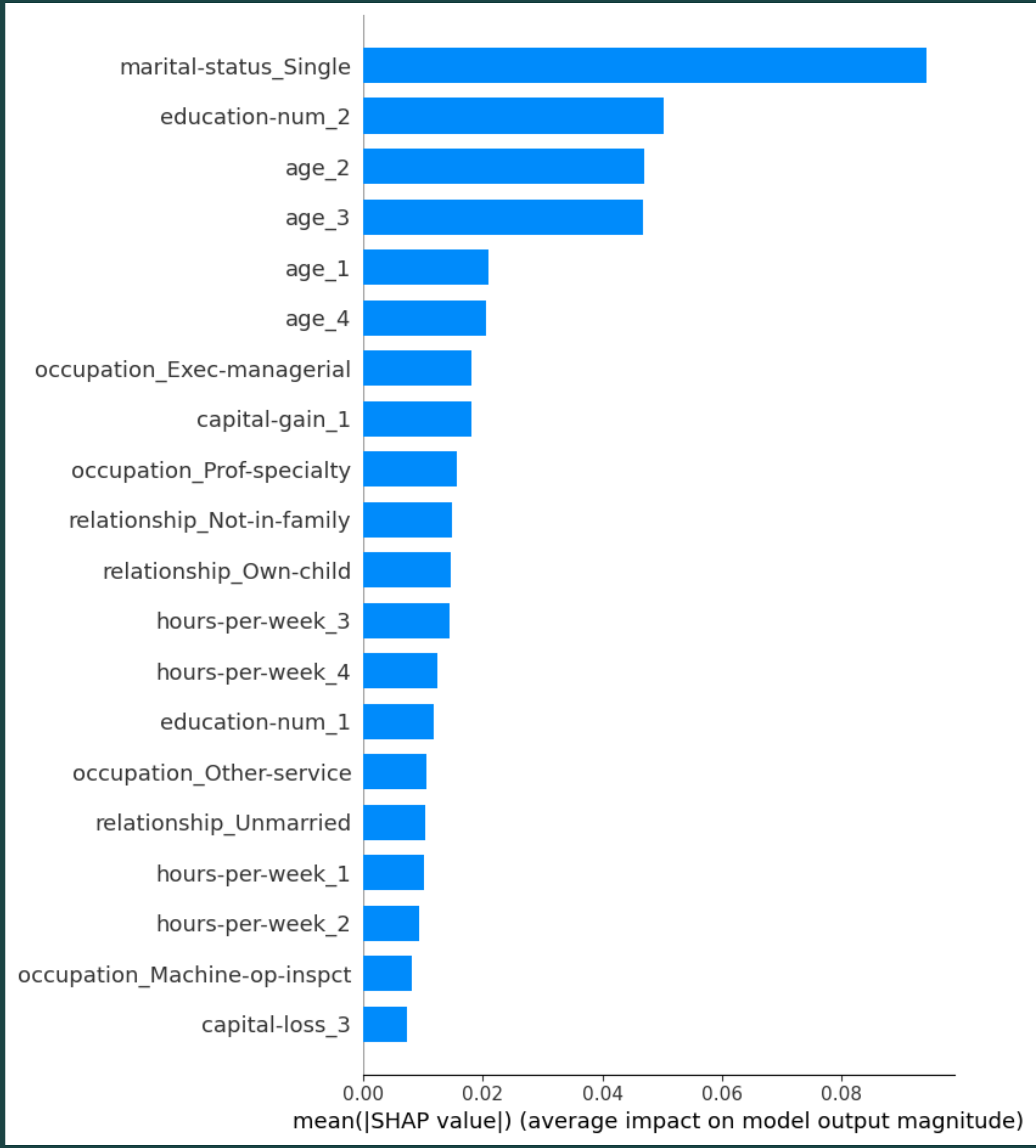
Default Threshold (0.5)	Best Threshold = 0.405082
Accuracy Score: 0.82	Accuracy Score: 0.79
Precision: 0.61	Precision: 0.55
Recall: 0.75	Recall: 0.85
F1: 0.68	F1: 0.67
AUC: 0.80	AUC: 0.82





CONCLUSION

Main Problem	Can we predict an individual's income based on their characteristics?
Why is it important?	The factors that contribute to it can help us design policies to address income inequality.
Conclusion	We can observe that factors such as marital status, education, age and occupation are important in determining income.



MODEL INTERPRETATION USING SHAP VALUES