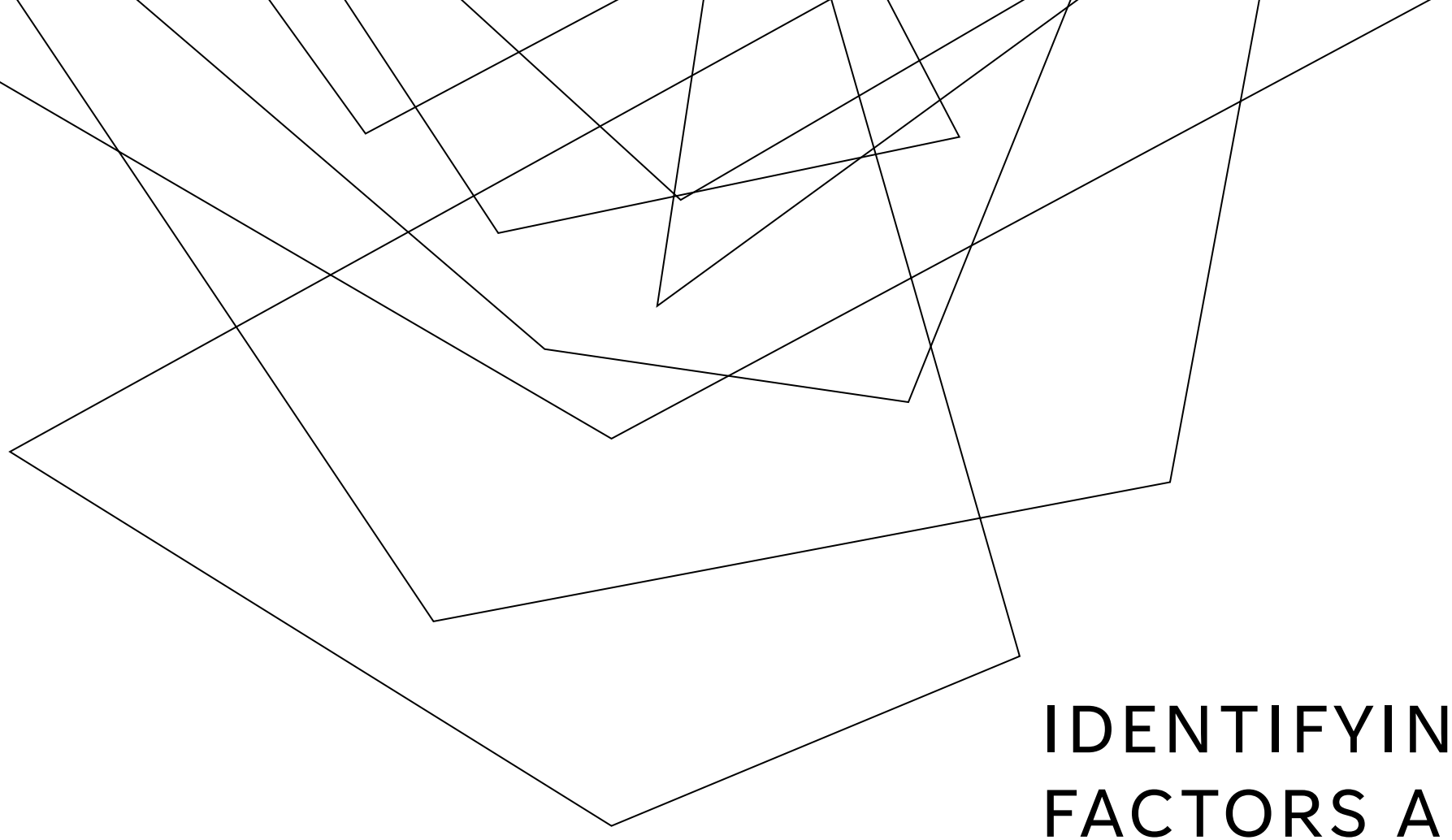


IDENTIFYING FACTORS AFFECTING PERSONAL INCOME

Nick Ross for Dataiku

CONTEXT

SITUATION

The US Government aims to allocate federal funding to ensure their strategic priorities are adequately supported.

COMPLICATION

The effective allocation of funds relies on deep understanding of the demographic characteristics of US subpopulations.

QUESTION

How can we use US census data help us to understand demographic trends in the US population and inform the effective allocation of funding?

SOLUTION

Machine learning techniques can be used by the US Government to understand the factors behind income inequality.

POTENTIAL BENEFITS OF THIS WORK

UNDERSTAND KEY DRIVERS

Identify the primary drivers of income inequality and prioritise areas for intervention

PREDICT POLICY IMPACTS

Enable data-driven decision making by forecasting the effects of policies on income distribution across demographics

UNCOVER SYSTEMIC DISPARITIES

Identify systemic issues in demographics such as gender and race, enabling focused support for marginalized groups

CONTENTS

EXPLORATORY DATA ANALYSIS

Informing tactics for answering the solution

DATA PREPARATION

Preprocessing and feature engineering

DATA MODELLING

Testing candidate models

MODEL ASSESSMENT

Evaluating candidate models

RESULTS

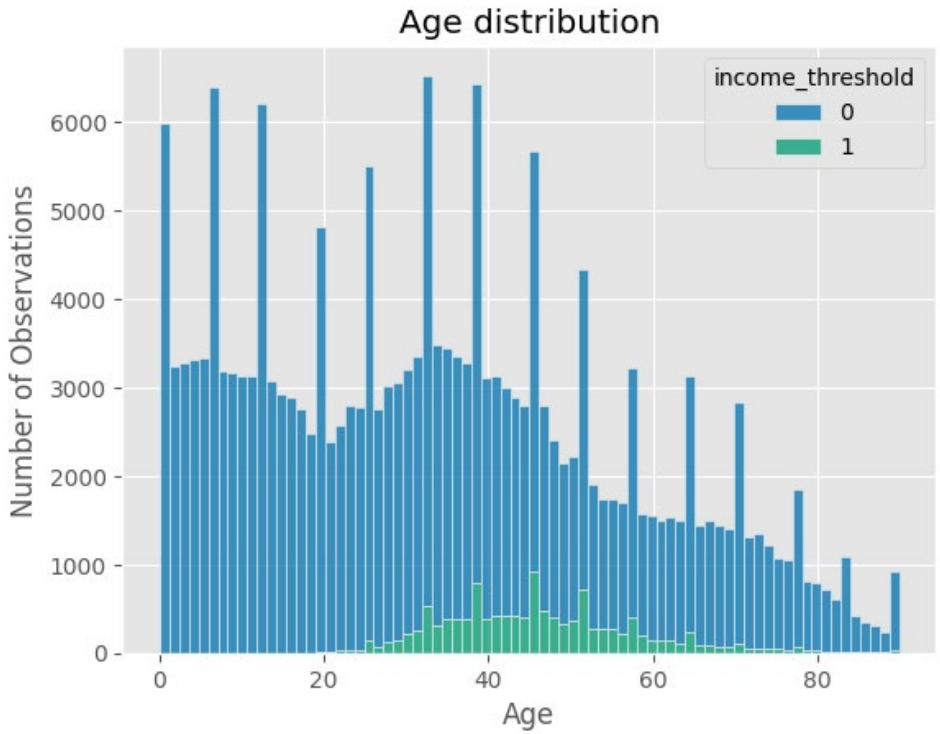
Feature importance

POSSIBLE EXTENSIONS

Additional avenues of research that could prove useful

EXPLORATORY DATA ANALYSIS

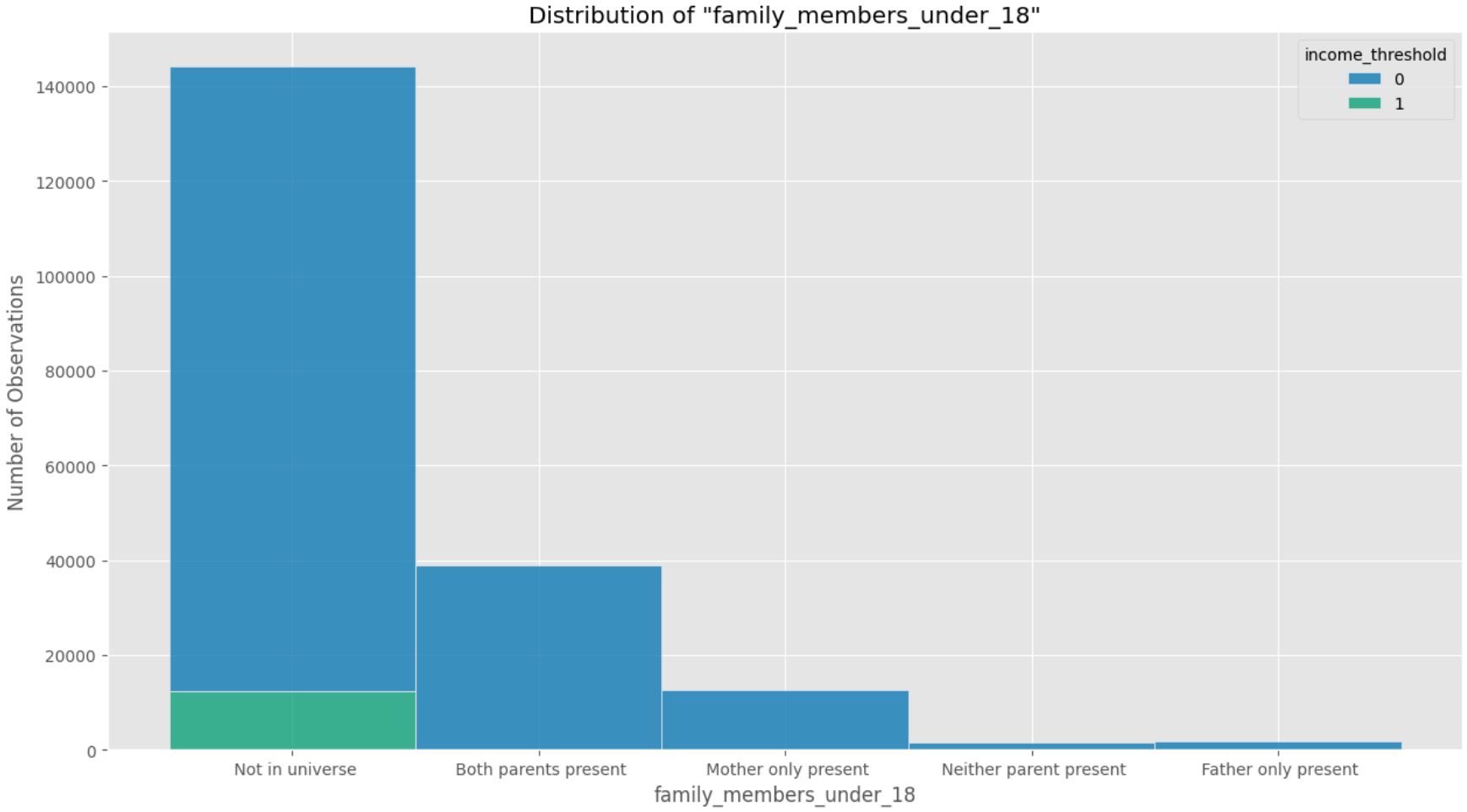
Dataset: US Census data for 1994 - 1995
Variables: 40 demographic measures, 2 survey description variables (year and instance weight)
Target variable: Income threshold (binary feature representing above or below $50,000 in income)



Histograms were used to understand the distribution of different features and how the target variable is distributed within that feature.

This histogram of age shows significant peaks at regular intervals, which should be addressed by feature engineering.

EXPLORATORY DATA ANALYSIS



This analysis also revealed one feature that was incorrectly labelled as ‘family_members_under_18’ when it likely represented the feature ‘presence_of_parents’.

This feature can be dropped as the question was not applicable to any of the respondents in the target class as it was only asked to under 18-year-olds.

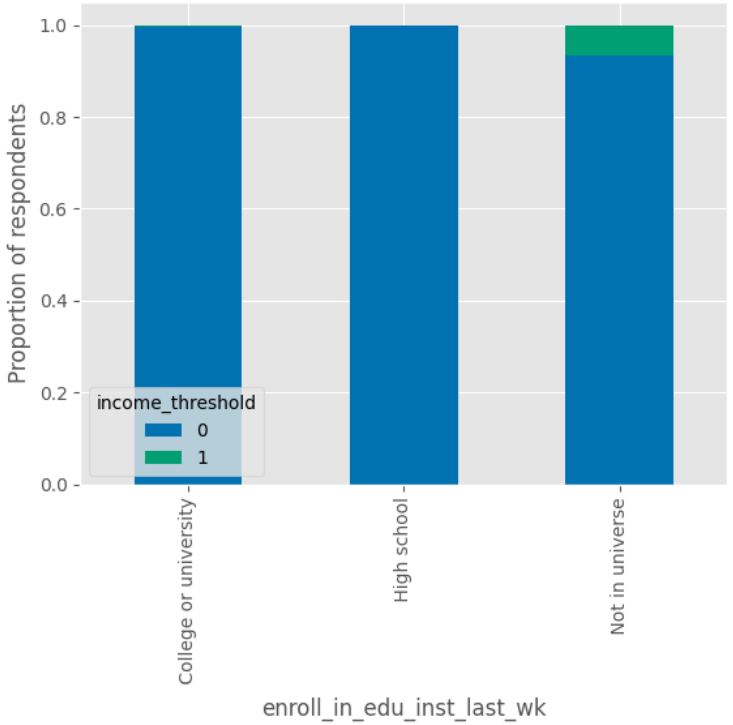
EXPLORATORY DATA ANALYSIS

Features including the value **‘not in universe’** were explored to decide whether to treat these values as NA.

Analysis showed that this value **should be treated as it’s own category** as the histograms demonstrated **predictive power** when visualizing the distribution of the target class in the overall proportion of each class.

Features containing actual NA values were noted as either suitable for imputation or treated as their own category:

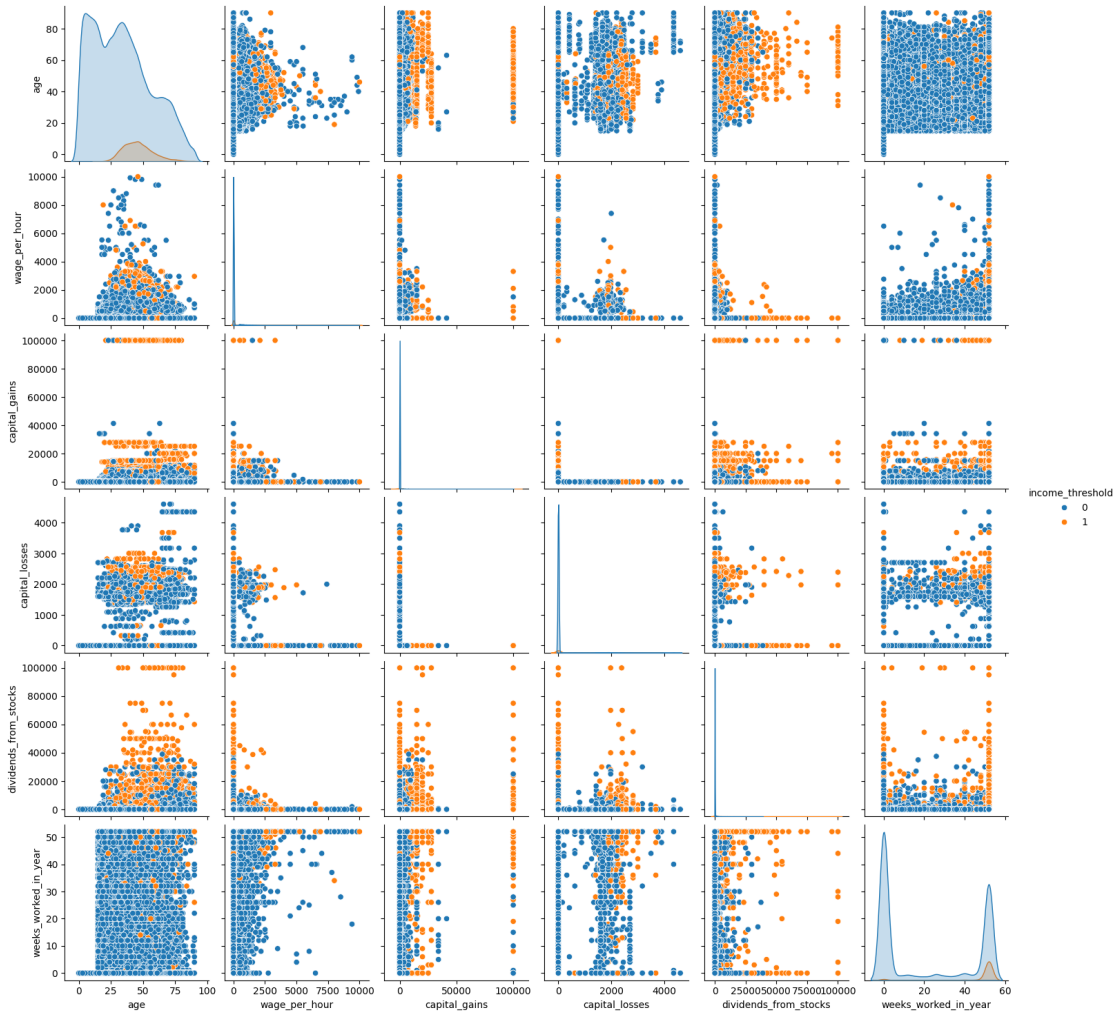
- **Hispanic_origin**: impute with most frequent (0.4% of dataset)
- **country_of_birth**: impute with most frequent (~6% of dataset)
- **migration codes**: treat '?' as own category



EXPLORATORY DATA ANALYSIS

Pair plots were used to explore the relationship of the target variable with different combinations of numeric variables.

By inspection, several combinations of numeric variables were found to contain significant signal in identifying the target variable which validated their inclusion in the model.



DATA PREPARATION

Based on the insights gained in the exploratory data analysis, the following feature engineering and preprocessing steps were performed:

Dropping features that are unlikely to help prediction
family_members_under_18, fill_inc_questionnaire_for_veterans_admin, instance_weight, year

Creating new feature
Classifying age into child, working age and retirement age based on US standards

Binning
Simplifying features into categories whilst preserving predictive performance

Scaling
Using StandardScaler to standardize the distribution of variables

One-Hot Encoding
Encoding categorical features into a binary column per feature value

Resampling
Using SMOTE to increase the number of observations of the minority class (Respondents earning above $50k)

DATA MODELLING

Three candidate models were selected for testing as they are well known and quick to train under the hardware limitations for this project.

- **Logistic Regression Classifier**

A linear model that predicts the probability of a binary outcome by applying a logistic (sigmoid) function to a weighted sum of input features.

- **Random Forest Classifer**

An ensemble method that trains multiple decision trees and combines their predictions.

- **Gradient Boosted Classifer**

An iterative ensemble method that builds decision trees sequentially, where each tree aims to correct the errors of the previous one.

Sklearn pipelines were used to package preprocessing steps and ensure reproducibility.

Mlflow was used to log model parameters and metrics to ensure reproducibility in addition to unlocking easier collaboration with other data scientists if required in the future.

Once the best performing model was identified, Optuna was used to optimize it's hyperparameters.

MODEL ASSESSMENT

The performance metrics for this binary classification problem were selected as:

- **F1 score**

The harmonic mean of precision and recall – a measure of the balance between false positives and false negatives.

- **Area under the Receiver Operator Curve**

A measure of a model's ability to distinguish between classes.

Each model was tested with and without feature binning and resampling to determine whether these preprocessing steps were worthwhile.

The parameter for each model were set to common defaults.

RESULTS – METRICS

Logistic Regression

Binning	SMOTE: Enabled		SMOTE: Disabled	
	F1 Score	ROC_AUC	F1 Score	ROC_AUC
Binning: Enabled	0.423	0.937	0.464	0.938
Binning: Disabled	0.449	0.949	0.514	0.946

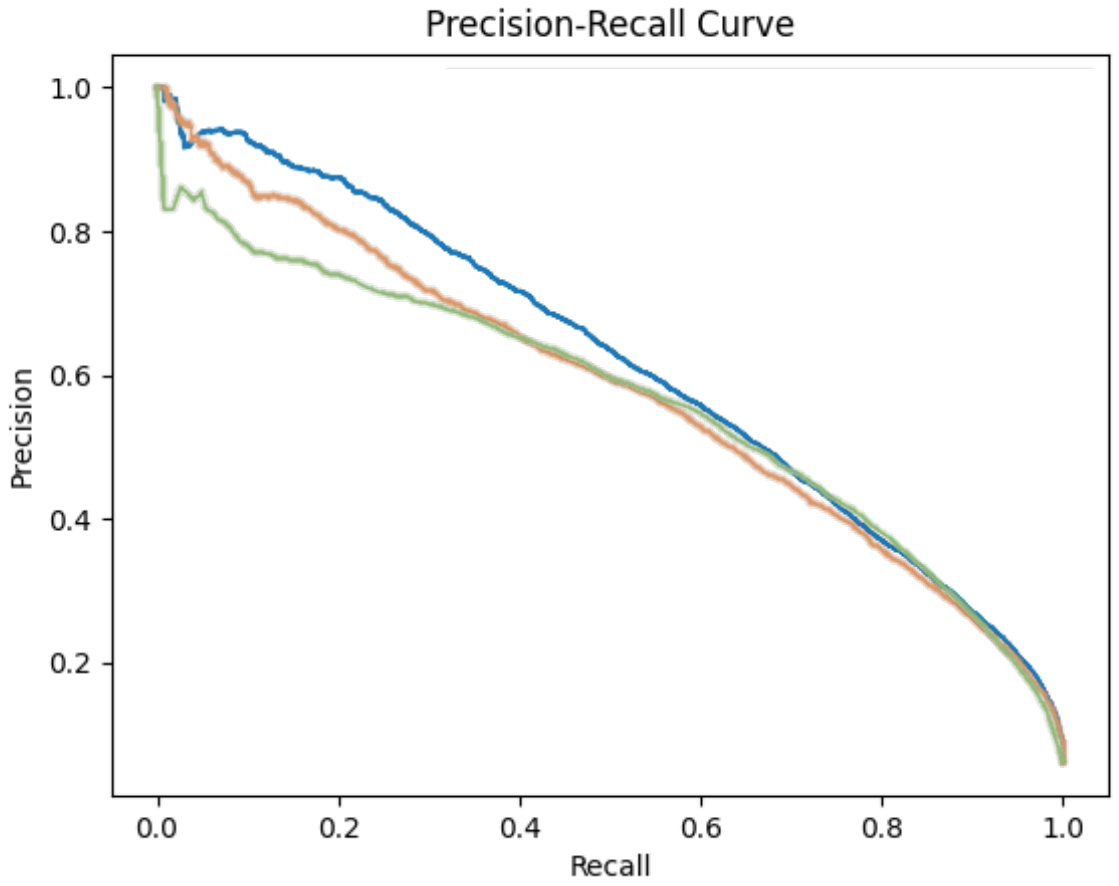
Random Forest

Binning	SMOTE: Enabled		SMOTE: Disabled	
	F1 Score	ROC_AUC	F1 Score	ROC_AUC
Binning: Enabled	0.500	0.924	0.497	0.924
Binning: Disabled	0.529	0.937	0.511	0.939

Gradient Boosted

Binning	SMOTE: Enabled		SMOTE: Disabled	
	F1 Score	ROC_AUC	F1 Score	ROC_AUC
Binning: Enabled	0.495	0.934	0.506	0.942
Binning: Disabled	0.534	0.947	0.528	0.947

RESULTS – PLOTS



Logistic Regression

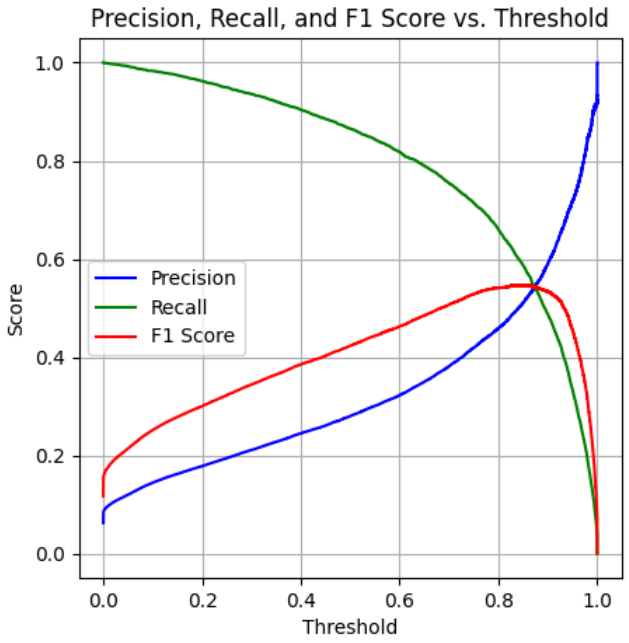
Random Forest

Gradient Boosted

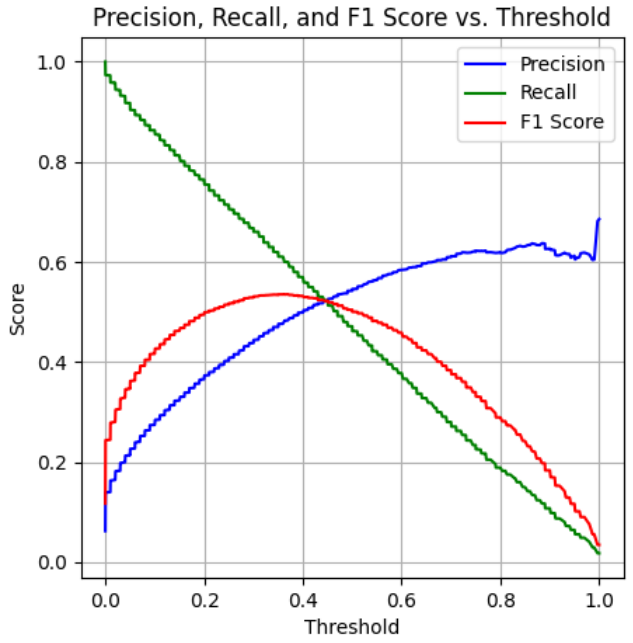
Conclusion: These models are significantly underperforming.

RESULTS – PLOTS

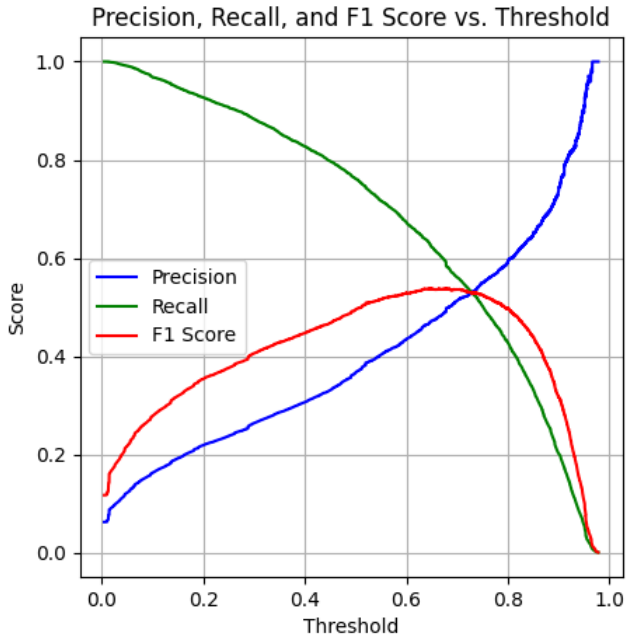
Logistic Regression



Random Forest



Gradient Boosted



These plots show **sub-optimal performance** for all three models (Optimal F1 score < 0.6).

Hyperparameter optimization for the gradient boosted classifier using Optuna found negligible performance increase.

RESULTS – FEATURE IMPORTANCE

Feature	Importance
weeks_worked_in_year	0.341395
dividends_from_stocks	0.185886
age	0.087759
sex = Female	0.052087
Education = Bachelors degree	0.050123
num_persons_worked_for_employer = 6 (1000+ people)	0.029747
education = High school graduate	0.025898
education = Masters degree	0.023155
sex = Male	0.022637
major_occupation_code = 'Professional specialty'	0.019531

Extracted from the GradientBoostedClassifier model by aggregating each feature's contribution to improving the model (reduction in impurity).

REVISITING OBJECTIVES

1. UNDERSTANDING KEY DRIVERS

The top five predictors (according to this research) of individuals making more than $50,000 per year:

- Number of weeks worked in a year
- Age
- Income from stock dividends
- Sex
- Education

Possible questions :

- What are the underlying reasons for individuals to work part-time, and can they be supported to work full time to increase their earnings potential?

- Should the US educate more of it’s citizens on financial market participation to increase their earnings potential?

REVISITING OBJECTIVES

2. PREDICTING POLICY IMPACTS

An individual's sex was an important feature in determining earnings potential.

Further questions:

- To what extent does policy aimed at different sexes (e.g. legislation on reproductive rights) affect earnings potential for those demographics?

REVISITING OBJECTIVES

3. UNCOVER SYSTEMIC DISPARITIES

The sex of an individual is a top 5 factor for predicting income.

Out of 506 total features, the 20th most important feature for predicting income was whether not the individual is white.

Further questions:

- How can discriminated sub-populations be better supported to achieve equity in their earnings potential when compared with their non-discriminated against equivalents?

POSSIBLE EXTENSIONS

Acquire more data

Additional observations, particularly for rare classes in categorical features will likely improve model performance.

Test additional feature engineering techniques

Measuring the efficacy of different combinations of binning variables and the creation/removal of features may improve model performance.

Test more advanced models

E.g. SVC, XGBoost and Neural Networks were not tested as part of the project due to hardware / time limitations. It's likely that a more powerful model will perform better (at the cost of additional compute).

Test different resampling techniques

The effect of of different permuations of SMOTE and under and oversampling techniques could be studied.

Explore different feature importance techniques

E.g. SHAP, linear model coefficients