DATAIKU TECH ASSESSMENT

Stavros Emmanouilidis

AGENDA

Executive Summary

Exploratory Data Analysis

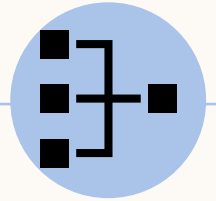
Data Preparation

Data Modeling

Summary & the Future

Questions

EXECUTIVE SUMMARY



SITUATION

- US Census Bureau collects demographic and economic data for strategic planning.
- Data is used to allocate funding for public services.
- A sample dataset of ~300,000 individuals is provided for analysis.

COMPLICATION

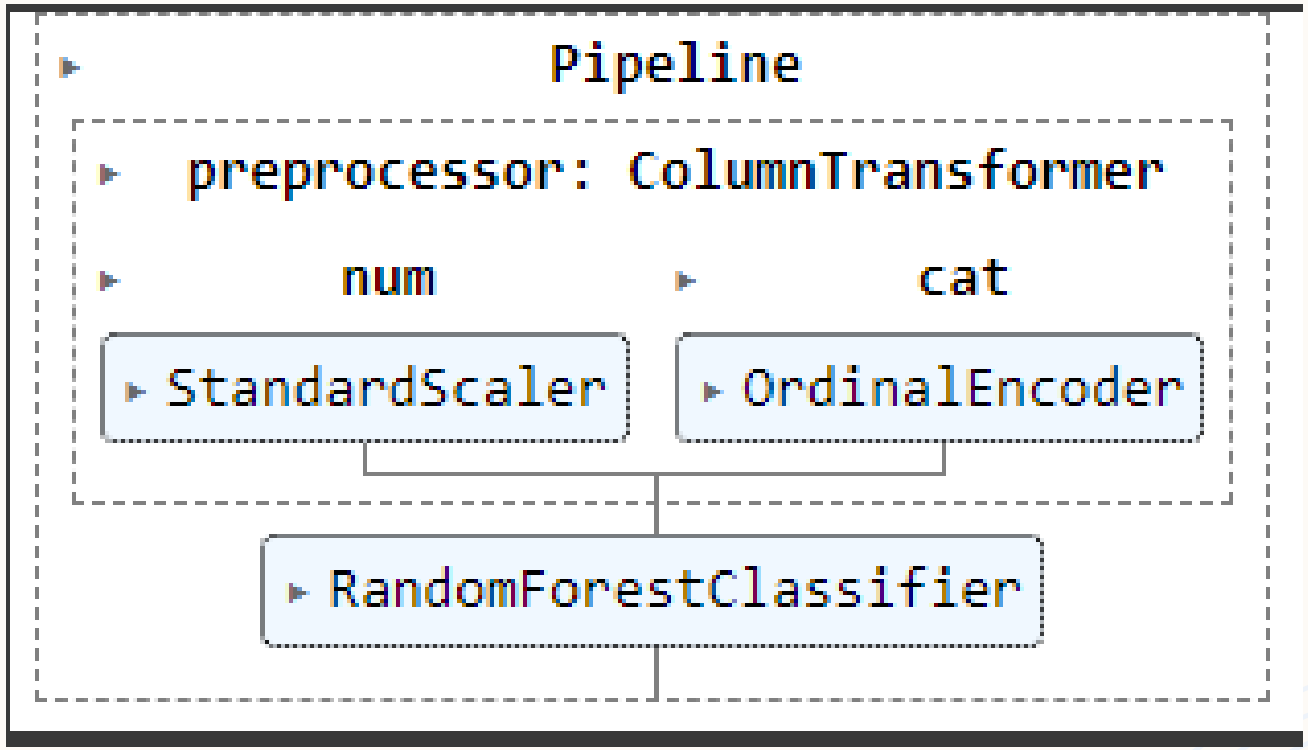
- Need to analyze extensive data to identify characteristics that correlate with income level.
- Goal is to understand what factors are associated with making more or less than $50,000 per year.
- Important for understanding economic disparities and policy planning.

QUESTION

- What characteristics are associated with a person making more or less than $50,000 per year in the sample dataset?

- Trained a Random Forest Classifier that achieved the following Test set performance
  - Accuracy: 0.9538
  - Precision: 0.7276
  - Recall: 0.4064
  - F1-Score: 0.5215
  - ROC AUC: 0.9340

- Over 90% accuracy!
- However 93% of the instances belong to the `-50000` class.



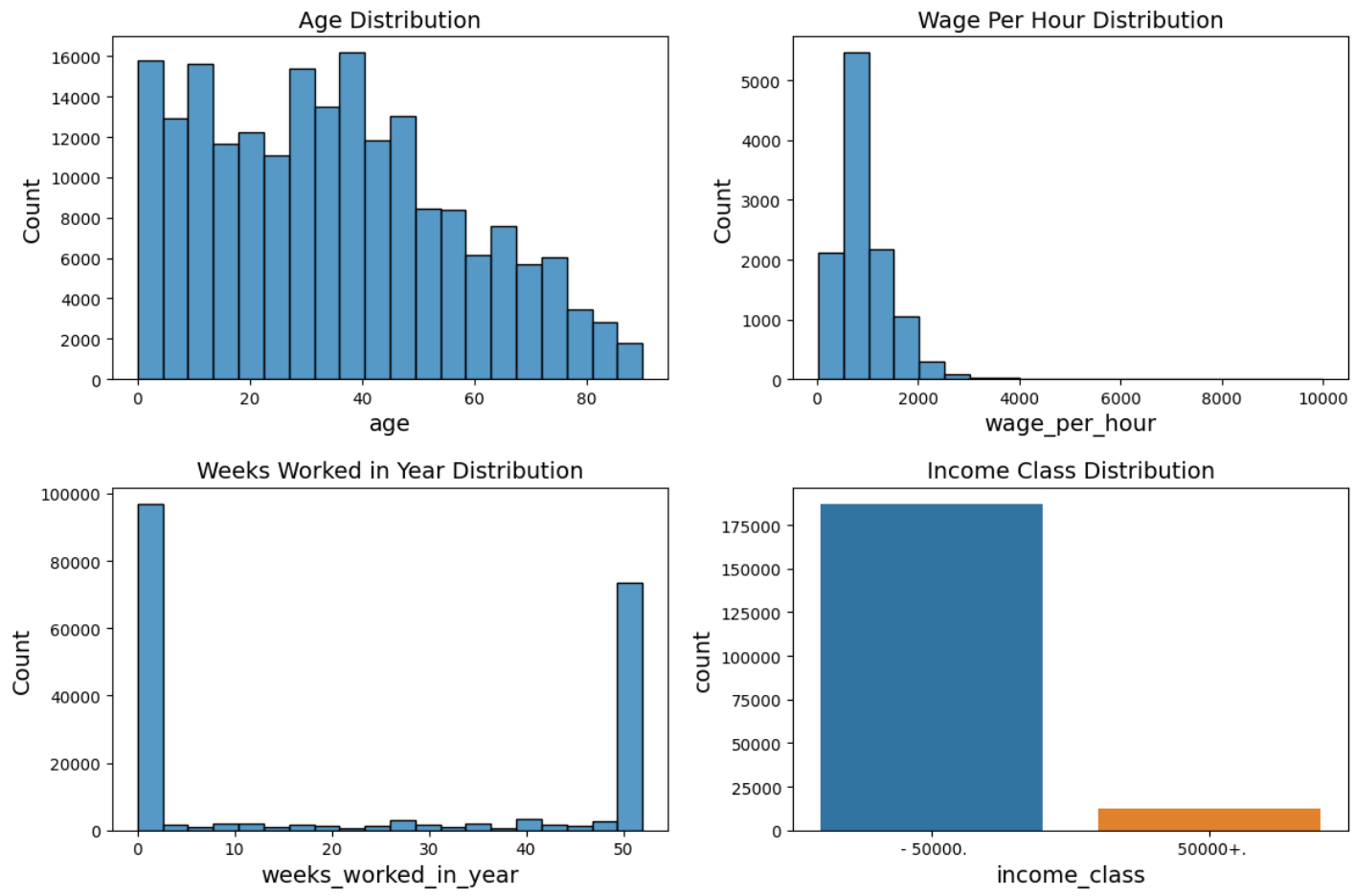
EXPLORATORY DATA ANALYSIS

Age Distribution: Most people in the dataset are between 15 and 60 years old. We can see a decline in the number of people as age increases, which is expected.

Wage Per Hour Distribution: Most people have a wage of $0 per hour, which likely means they are not currently employed or their primary income is not from hourly wages. For those who do earn an hourly wage, the distribution is positively skewed.

Weeks Worked in Year Distribution: A significant number of people have worked zero weeks in the year, likely indicating unemployment or other forms of non-employment (e.g., students, retirees). Many others have worked the full 52 weeks in a year.

Income Class Distribution: The majority of people in the dataset earn less than $50,000 per year. This shows that the target variable is imbalanced, which is something to consider during modelling.



DATA PREPARATION

DATA PREP STEPS

HANDLING MISSING VALUES

Although the dataset doesn't have explicit NaN values, some entries like "Not in universe" might act as placeholders for missing or inapplicable data.

CATEGORICAL ENCODING

Many machine learning algorithms require numerical input and output variables. We'll need to convert the categorical variables to a format that could be provided to machine learning algorithms.

FEATURE ENGINEERING

New features were created based on existing ones to better represent the underlying problem to the machine learning models.

DATA NORMALIZATION

Some machine learning algorithms are sensitive to the scale of input variables. We'll normalize the features to make them comparable.

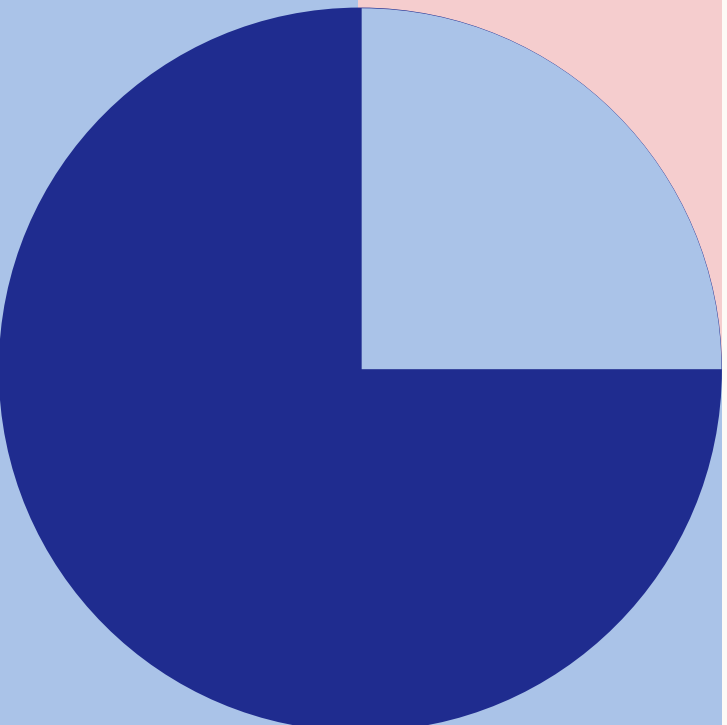
HANDLING MISSING VALUES

Several columns have missing values:

- class_of_worker: 100,245 missing
- enrolled_in_edu_inst_last_wk: 186,943 missing
- major_industry_code: 100,684 missing
- major_occupation_code: 100,684 missing
- member_of_labor_union: 180,459 missing
- reason_for_unemployment: 193,453 missing
- full_or_part_time_employment_stat: 123,769 missing
- ... and so on.

- To handle missing values I decided to **replace them with a separate category labelled `Unknown`**, since when the value is missing it might be for a reason that is meaningful to the research question.

- For example, `**reason_for_unemployment**` being missing likely indicates that the person is not unemployed, which is itself informative.



DATA NORMALIZATION

- 7 numerical features
- Standardize features by removing the mean and scaling to unit variance.
- The standard score of a sample x is calculated as:

$$z = (x - u) / s$$

where u is the mean of the training, and s is the standard deviation of the training

CATEGORICAL ENCODING

- 29 categorical features
- Used Ordinal Encoding which encodes categorical features as an integer array.
- It's the simplest form of Categorical Encoding.

- Discriminatory features like `race`, `Hispanic origin`, `sex` were dropped

FEATURE ENGINEERING

Idea	Engineered Feature
Interaction Features: Create new features by combining two or more variables	**net_capital_gain**: Represents the net capital gain, calculated as capital gains – capital losses
Polynomial Features: Create new features by raising an existing feature to a power. Could help capture non-linear relationships	**age_squared**: A polynomial feature representing the square of the age.
Categorical Aggregations: Aggregate numerical variables based on categorical variables.	**avg_wage_per_industry**: The average wage per hour for each industry. **avg_wage_per_occupation**: The average wage per hour for each occupation.
Binning: For numerical variables we could create bins to transform them into categorical variables. This could help capture non-linear relationships.	**age_bin**: Binned age categories.
Boolean Features: Create new features that indicate the presence or absence of a certain condition.	**has_capital_gain_or_loss**: A boolean feature indicating whether a person has any capital gains or losses.

DATA MODELING

Models trained using 5-fold cross-validation

- LogisticRegression
- RandomForestClassifier
- GradientBoostingClassifier

Based on performance I chose the RandomForestClassifier model for the task.

Random Forest Balanced is an attempt at dealing with class imbalance.

Random Forest Combined used a combination of the original features as well as the engineered ones.

Fine tuning was also attempted but did not offer a major improvement.

	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.946322	0.722933	0.218947	0.336102	0.914026
Random Forest	0.953038	0.719918	0.398159	0.512741	0.932968
Gradient Boosting	0.953689	0.751199	0.379422	0.504185	0.942203

	Accuracy	Precision	Recall	F1-Score	ROC AUC
Random Forest Balanced	0.951073	0.708931	0.358989	0.476624	0.930751

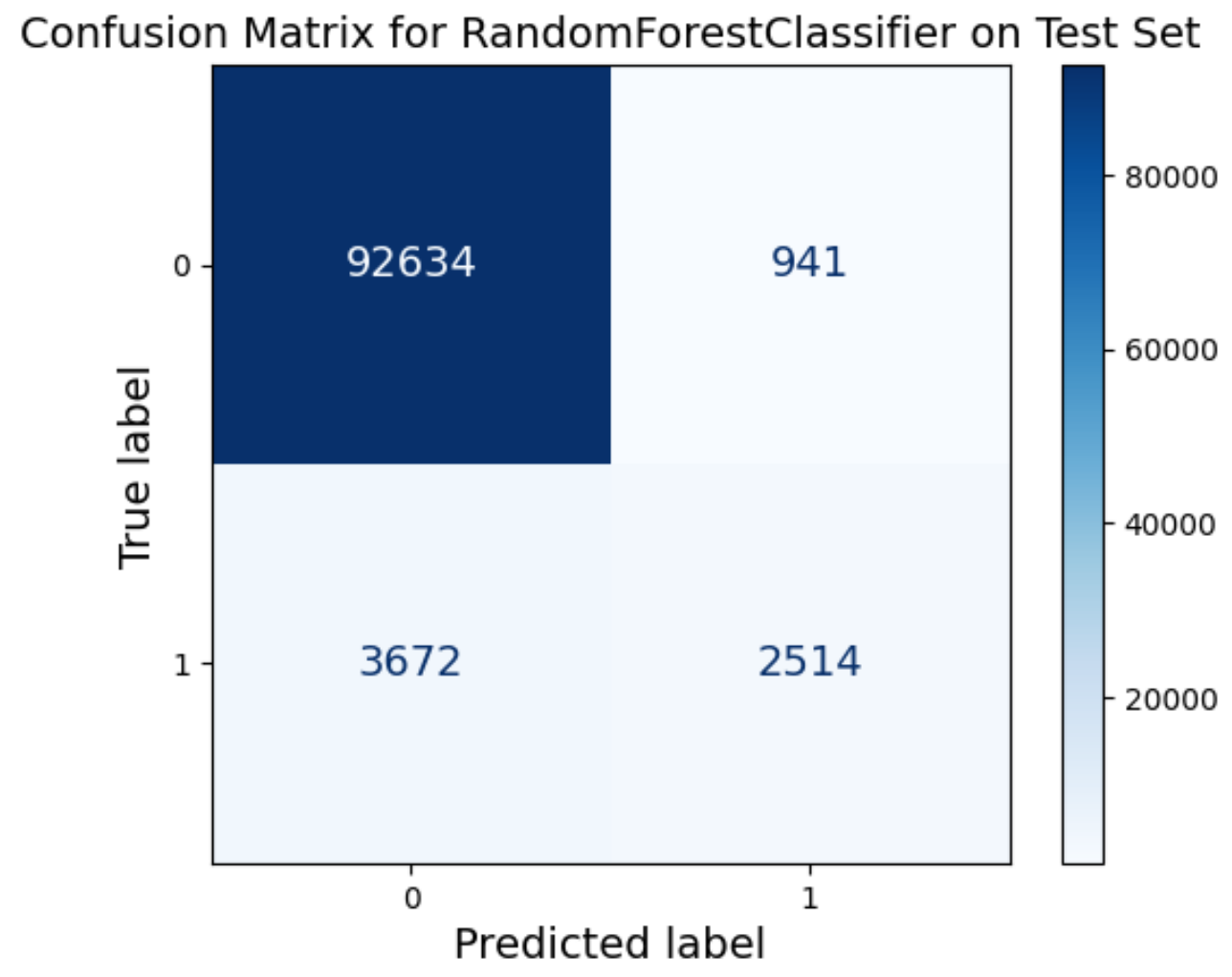
	Accuracy	Precision	Recall	F1-Score	ROC AUC
Random Forest Combined	0.953669	0.731963	0.399855	0.517184	0.933487

A RandomForestClassifier using the default parameters was trained on the full training set, and evaluated on the test set.

Test set metrics:

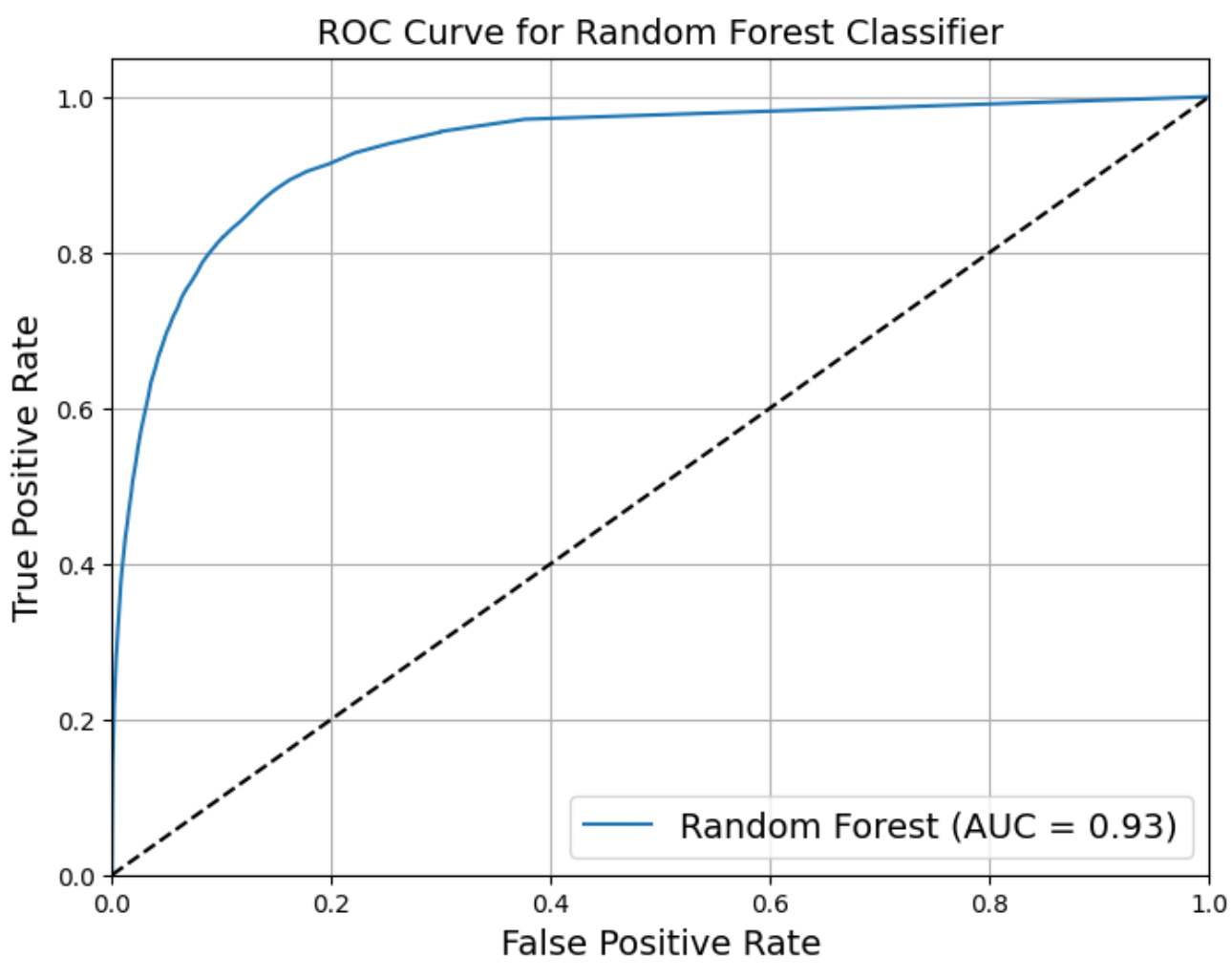
- Accuracy: 0.9538
- Precision: 0.7276
- Recall: 0.4064
- F1-Score: 0.5215
- ROC AUC: 0.9340

The results on the test set remain consistent with the performance during training, which indicates good generalisation performance.



Looking at the ROC curve, we may think that the classifier is really good.

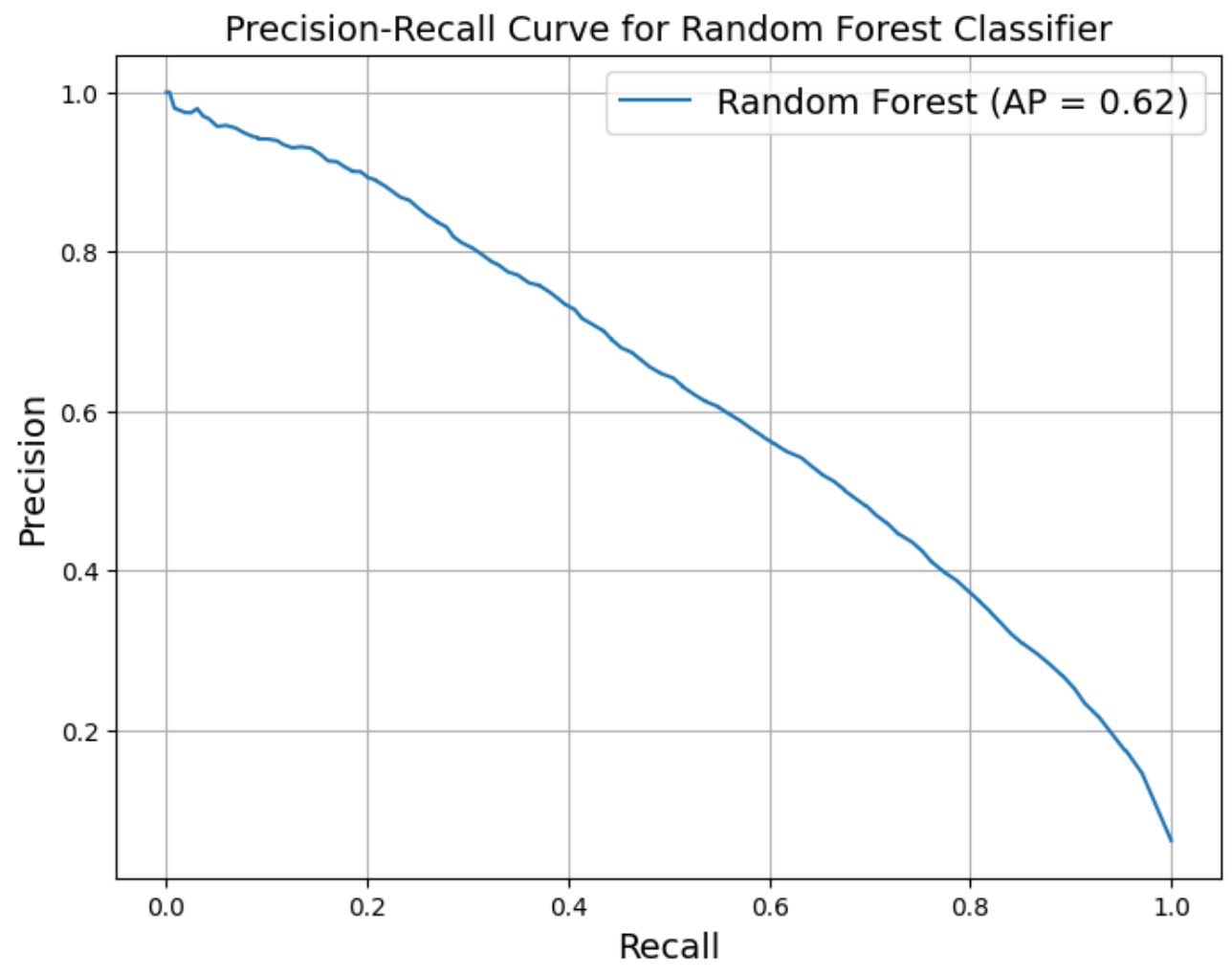
However, this is mostly because there are few positives (+50000) compared to the negatives (-50000).



The Precision/Recall (PR) curve makes it clear that **the classifier has a lot of room for improvement.**

The curve could really be closer to the top right corner.

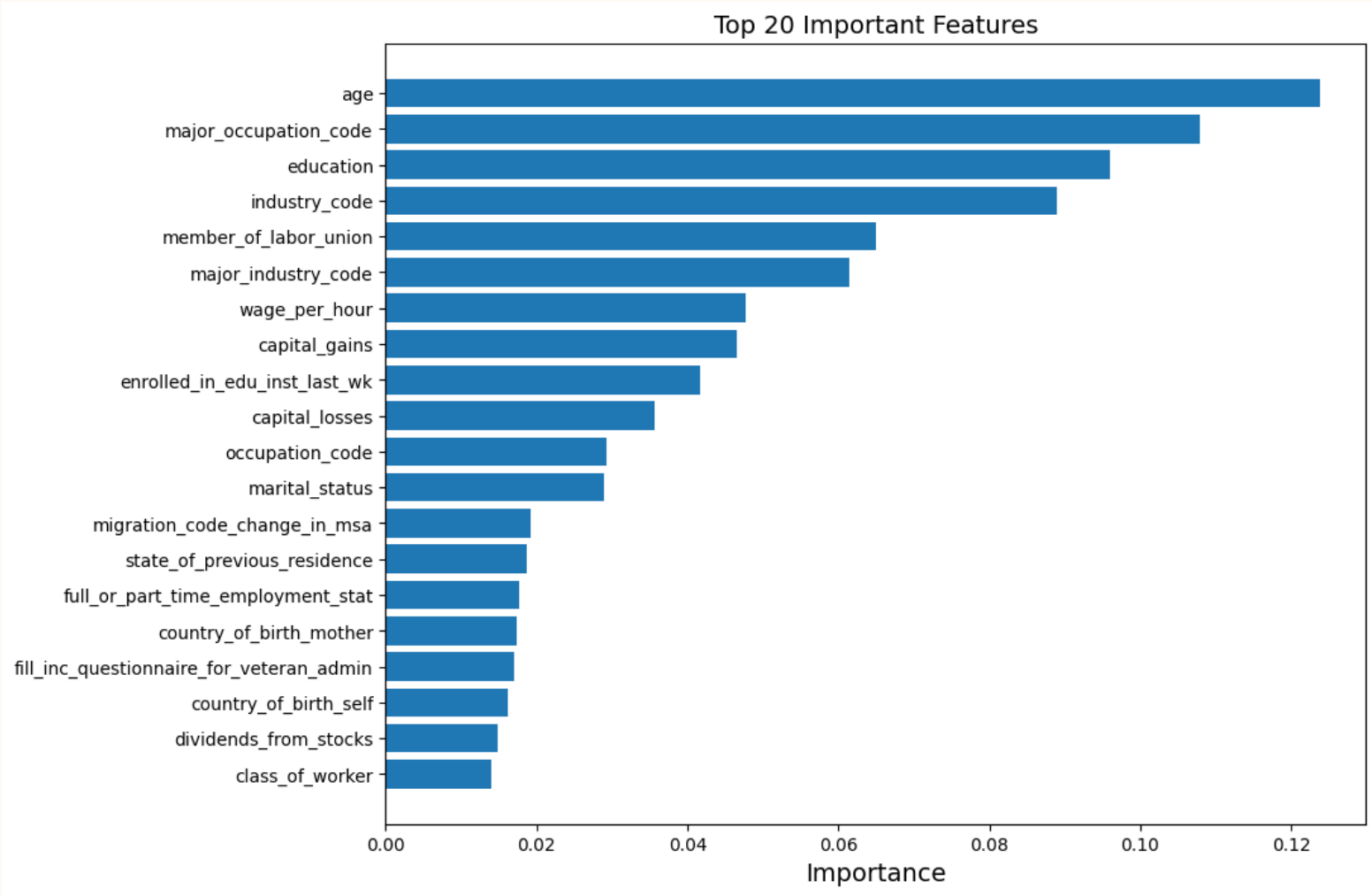
The PR curve is preferred **when the positive class is rare.**



The cumulative feature importance scores by category are as follows:

Occupational	30.1%
Other	21.9%
Demographic	20%
Educational	13.7%
Financial	9.7%
Geographic	4.3%

Features related to occupation and industry account for 30% of the model's decision-making, indicating the importance of job-related factors in income.



SUMMARY & FUTURE

SUMMARY

Objective

- To identify characteristics associated with an individual earning more or less than $50,000 per year using U.S. Census data.

Methodology

1. Exploratory Data Analysis (EDA)
   - Identified missing values and data types.
   - Visualized the distribution of the target variable and key features.
2. Data Preprocessing
   - Handled missing values by labeling them as 'Unknown'.
   - Encoded categorical variables.
3. Feature Engineering
   - Generated new features like net_capital_gain, total_income, age_squared, avg_wage_per_industry, and avg_wage_per_occupation.
4. Model Building
   - Utilized Logistic Regression, Random Forest, and Gradient Boosting models.
   - Employed cross-validation for robust performance assessment.
5. Model Evaluation
   - Metrics used: Accuracy, Precision, Recall, and F1 Score.
   - Random Forest showed the highest accuracy and F1 score on the training dataset.

RECOMMENDATIONS & FUTURE IMPROVEMENTS

- Address Class Imbalance

  The target variable is imbalanced; resampling techniques like oversampling/under sampling, or SMOTE can improve performance.
- Hyperparameter Tuning

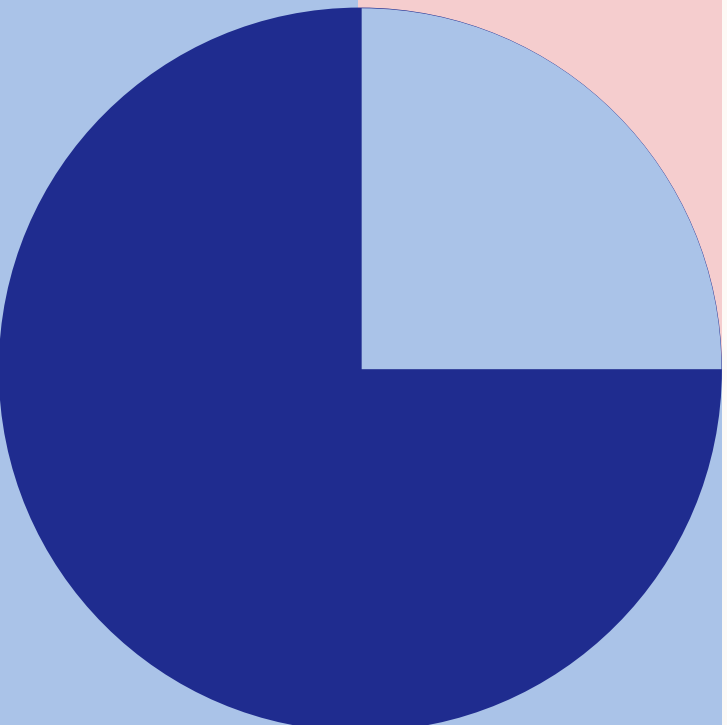
  Fine-tune model parameters for better predictive power.
- Additional Feature Engineering

  Further engineered features can provide more insights and improve model performance.
- More Comprehensive EDA

  Correlation analysis, outlier detection, etc.
- Advanced Models

  Experiment with ensemble methods or neural networks.
- Real-world Validation

  Test the model's performance on new, real-world data to assess its generalizability.



QUESTIONS?

THANK YOU

Stavros Emmanouilidis