Data Scientist Technical Assessment

Christina Schmitz

Introduction & Background information

The United States Census Bureau leads the country's Federal Statistical System; its primary responsibility is to collect demographic and economic data about America to help inform strategic initiatives.

Every ten years, the census is conducted to collect and organize information regarding the US population to effectively allocate billions of dollars of funding to various endeavors (e.g., the building and maintenance of hospitals, schools, fire departments, transportation infrastructure, etc.).

Additionally, the collection of census information helps to examine the demographic characteristics of subpopulations across the country.

Introduction & Data

I have been provided a sample dataset from the US Census archive containing detailed but anonymized information for ~300,000 individuals.

This archive contains three files:

census_income_learn.csv (data for model training).

census_income_test.csv (data for model testing).

census_income_metadata.txt (metadata for both datasets).

Introduction & Problem Statement

For this technical assessment, I have been tasked with identifying characteristics that are associated with a person making more or less than $50,000 per year.

Task Details

Binary Classification Task

(Less Than $50,000 – More Than $50,000)

Exploratory Data Analysis

Binary Classification Task

Income is our target variable

0 is is Less Than $50,000

1 is More Than $50,000

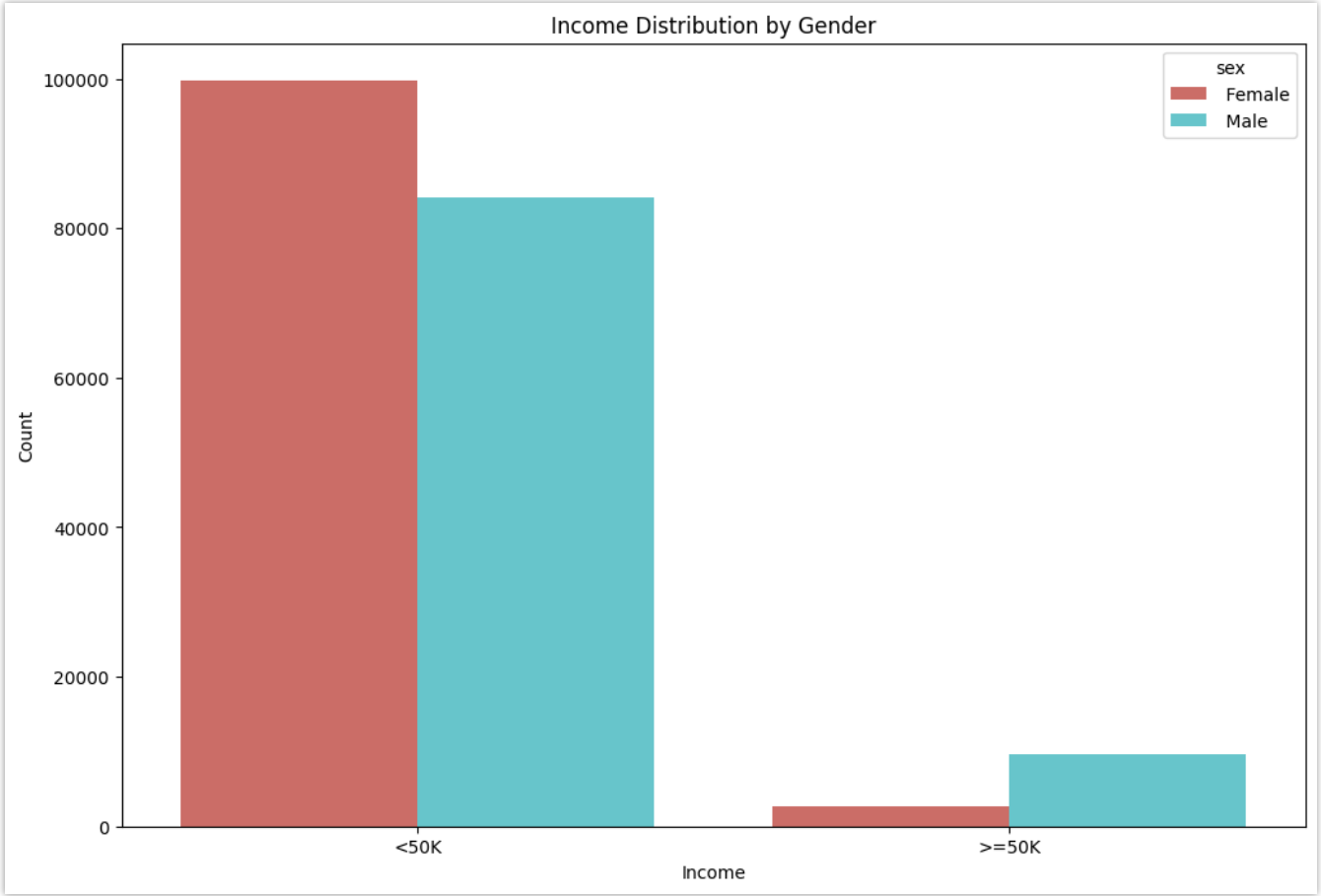
There are 42 features in total, 35 nominal, 8 continuous including the label column.

Exploratory Data Analysis: Income Distribution by Gender

This plot displays the distribution of income by gender.

It shows that a significantly higher number of females earn less than $50k compared to males.

In contrast, more males earn above $50k than females.



Exploratory Data Analysis: Income Distribution by Education Level

This histogram indicates the income distribution across different education levels.

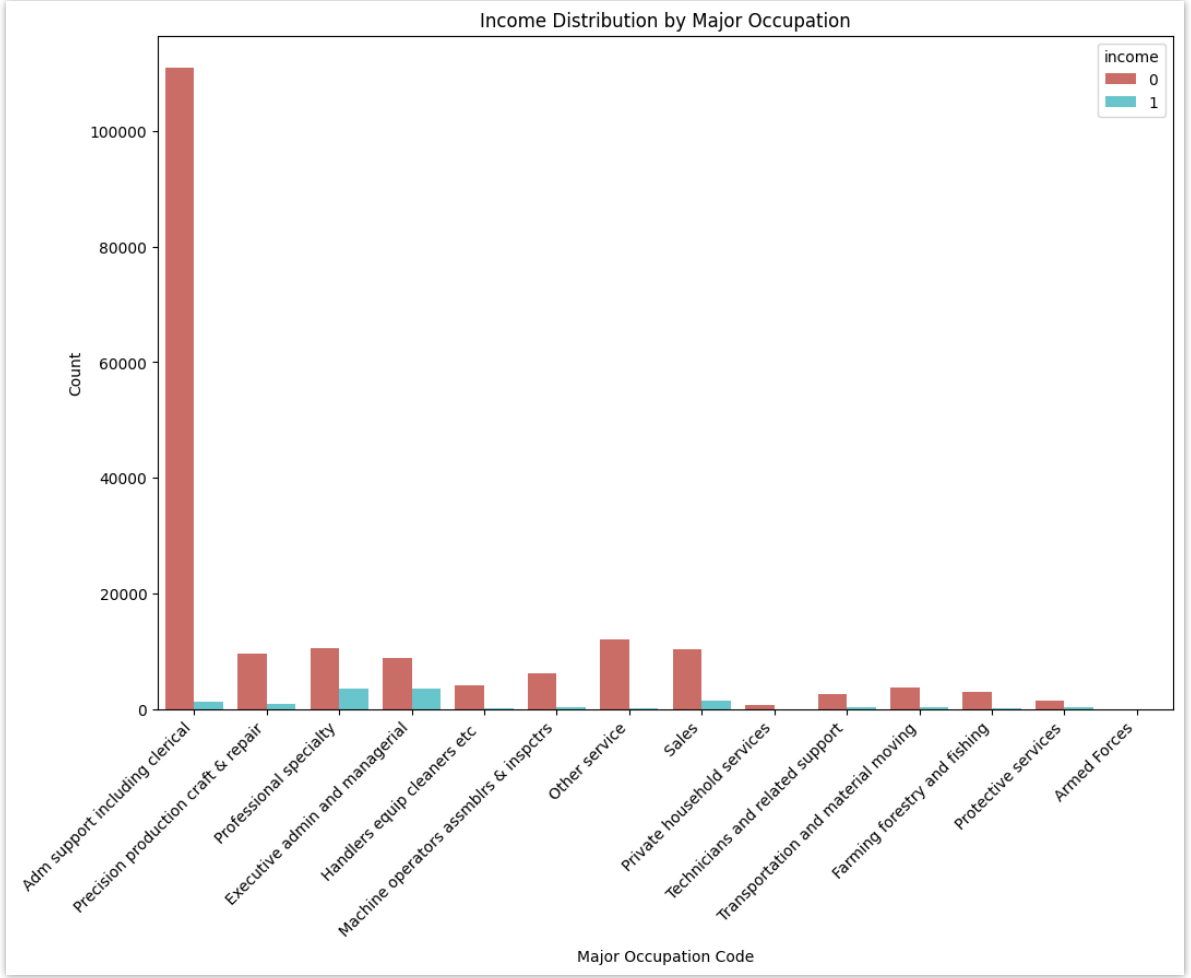
It appears that individuals with higher education levels, particularly those with master's and professional degrees, have a greater number of persons earning above $50k.

Conversely, those with less than a high school education predominantly earn below $50k.

Exploratory Data Analysis: Income Distribution by Major Occupation

This bar chart categorizes income based on major occupation codes.

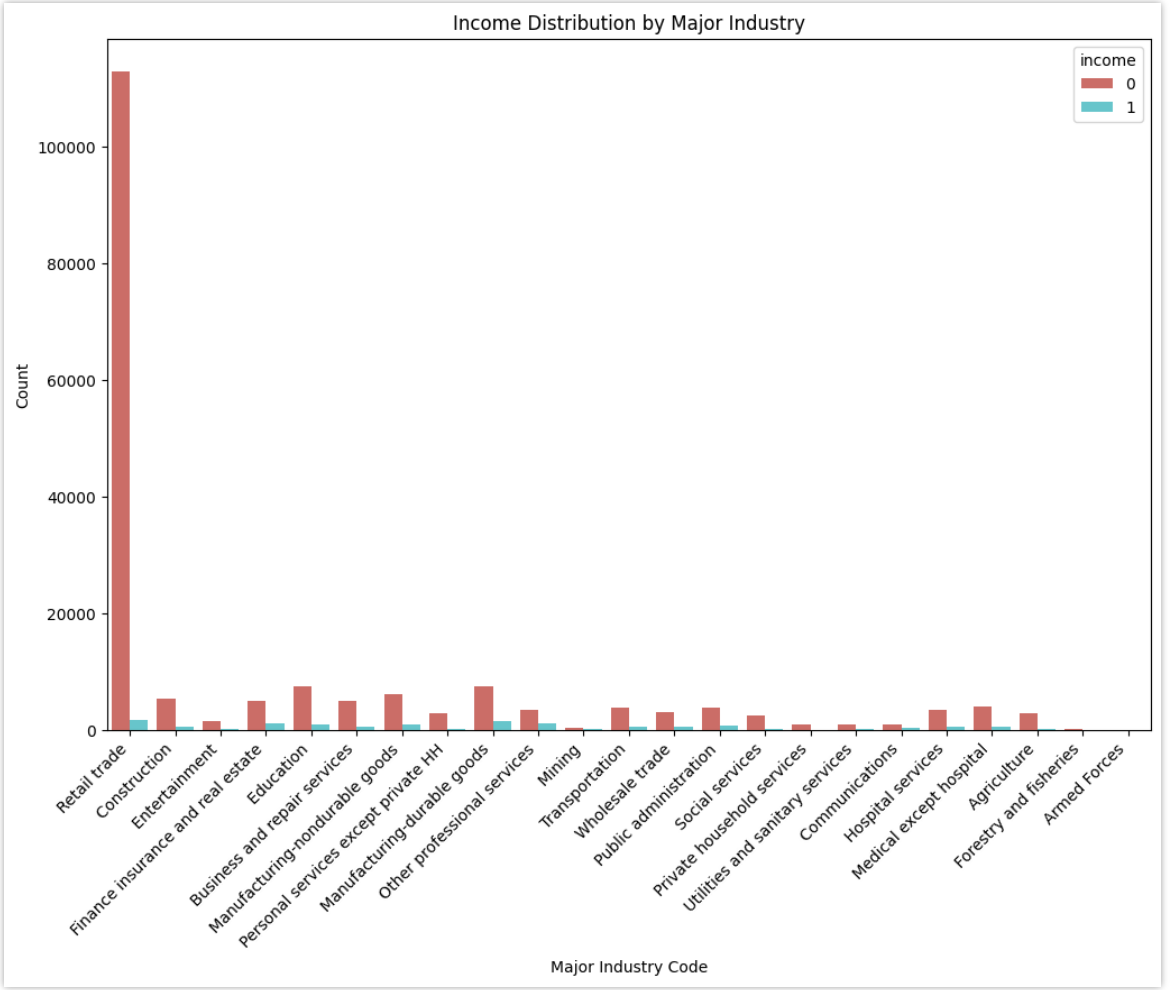
Individuals in executive, managerial, and professional specialty occupations tend to have a higher count of earners above $50k compared to other occupations.



Exploratory Data Analysis: Income Distribution by Major Industry

This plot breaks down income distribution across various industries.

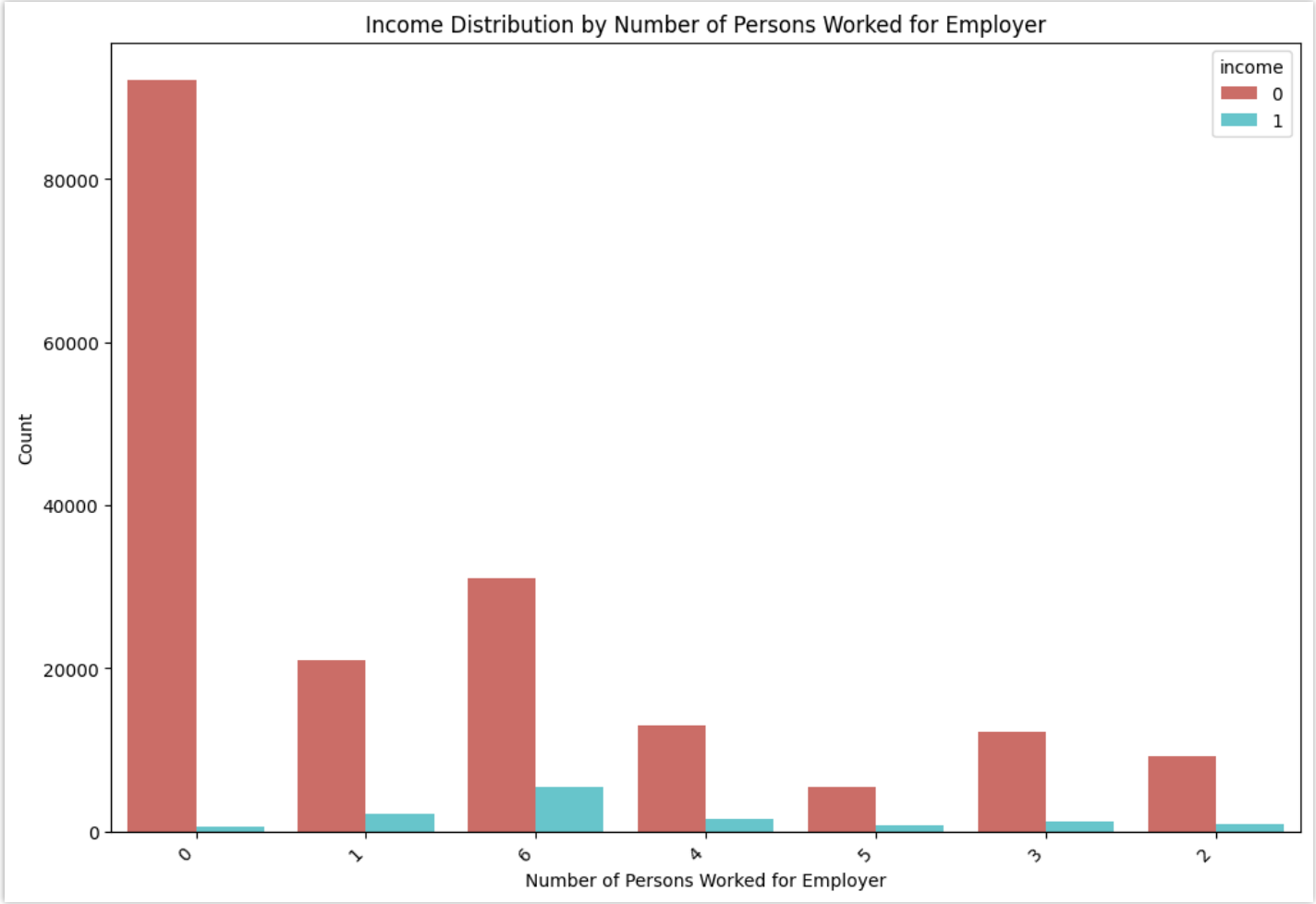
Similar to the occupation distribution, industries like professional services and finance have higher counts of individuals earning more than $50k.



Exploratory Data Analysis: Income Distribution by Number of Persons Worked for Employer

This plot illustrates the distribution of income based on how many different employers individuals have worked for.

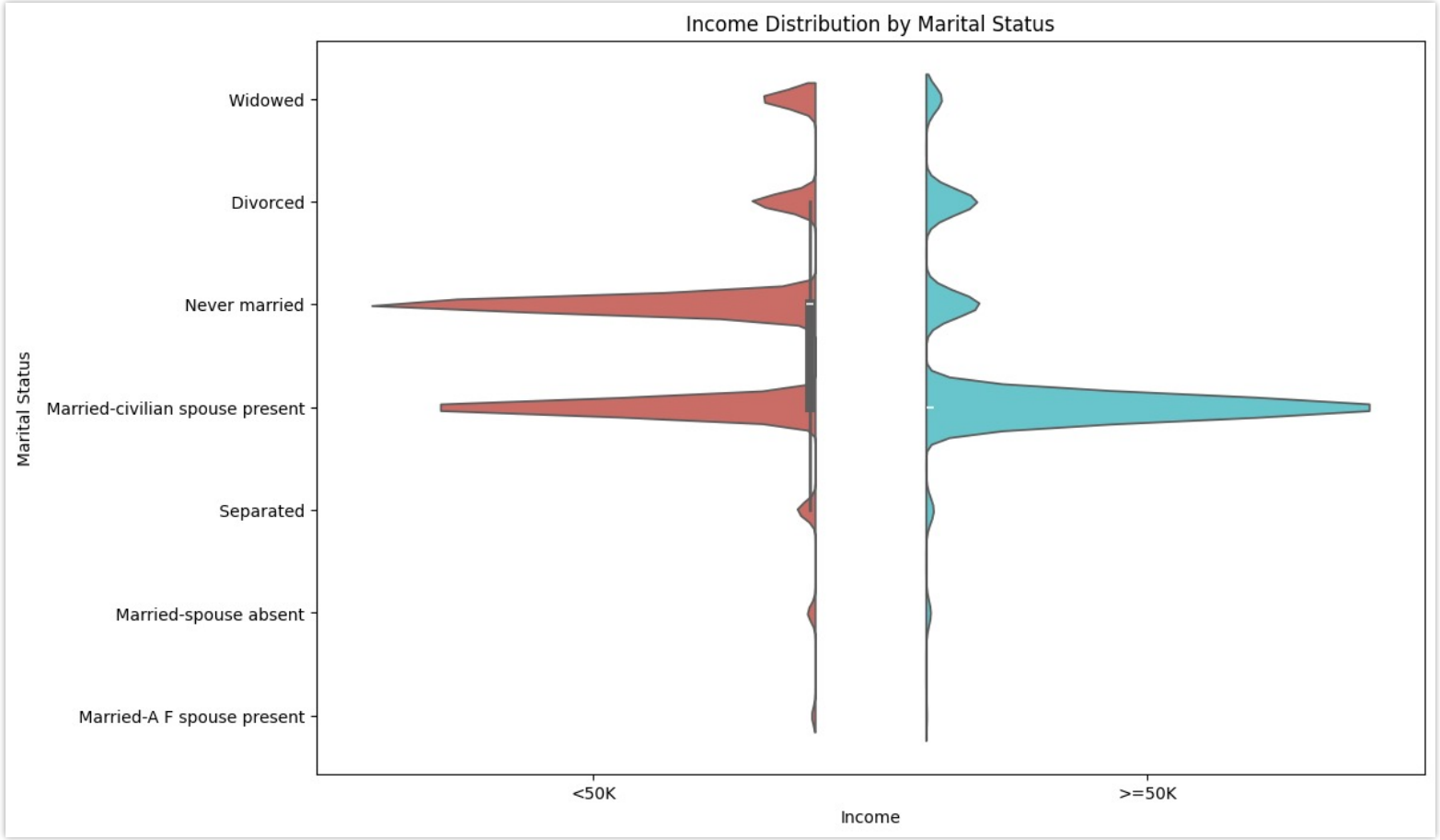
It shows a general trend that individuals who have worked for more employers are more likely to earn above $50k, though the majority earn below $50k.



Exploratory Data Analysis: Income Distribution by Marital Status

A detailed view of income distribution by marital status using a dual histogram format.

It shows that married individuals, both with civilian and armed forces spouses present, have a higher proportion earning above $50k, while never-married and divorced individuals predominantly earn below $50k.



Initial Data Preparation:

- **Column Naming:** Initially, there were 42 columns in both our training and test datasets. To enhance clarity and interpretability, I renamed the columns, which were initially represented by numbers.
- **Duplicate Removal:** As part of the cleaning process, I removed all duplicate entries from the datasets, ensuring data integrity and reducing redundancy.

Initial Data Preparation:

- **Target Variable Mapping:** I applied a mapping method to our target variable, transforming it into a binary classification problem. Specifically, I assigned the label 0 to incomes below $50,000 and the label 1 to incomes above $50,000.
- **Value Transformation:** I cleaned the data further by removing white spaces from values such as 'Not in universe' and transforming ambiguous values like 'Not in universe', '?', and 'NA' into NaN (Not a Number), ensuring consistency and accuracy in the dataset.

Initial Data Preparation:

These initial steps laid the foundation for subsequent feature engineering and model training processes, ensuring that the data was well-prepared for analysis and modeling.

Best Practices in Data Preparation & Feature Engineering:

1. Handling Missing Values:

- I deleted columns with high percentages of missing values, applying a threshold of 51% for removal. This decision was made according to best practices, which recommend removing columns with a high percentage of missing values to avoid introducing bias or noise into the model.
- I replaced missing values with the mode, the most frequent value in each column. This approach aligns with best practices in imputation techniques, ensuring that missing values are filled with plausible estimates.

Best Practices in Data Preparation & Feature Engineering:

2. Dealing with Correlated Features:

- I calculated the correlation matrix for numerical features and removed highly correlated columns with a threshold of 0.51. This action was taken in line with best practices to address multicollinearity, which can lead to unstable model estimates and reduced interpretability.

Best Practices in Data Preparation & Feature Engineering:

3. Eliminating Irrelevant Numeric Columns:

- I removed less relevant features based on domain knowledge, initial data exploration, and feature importance analysis. This practice is recommended to streamline the model and improve its performance by focusing on the most informative features.

Best Practices in Data Preparation & Feature Engineering:

There are the columns I deleted during my data preparation and feature engineering process:

- Columns with more than 51% of NaN values: 'enroll_in_edu_inst_last_wk', 'member_of_a_labor_union', 'reason_for_unemployment', 'tax_filer_stat', 'region_of_previous_residence', 'migration_code_change_in_msa', 'migration_code_change_in_reg', 'migration_code_move_within_reg', 'live_in_this_house_1_year_ago', 'migration_prev_res_in_sunbelt', 'family_members_under_18', 'fill_inc_questionnaire_for_veteran's_admin'.

- Highly correlated numeric columns with income: 'veterans_benefits', 'num_persons_worked_for_employer', 'detailed_occupation_recode', 'weeks_worked_in_year'.

- Less relevant features based on domain knowledge and initial data exploration: 'state_of_previous_residence', 'country_of_birth_father', 'country_of_birth_mother'.

These deletions align with best practices in feature selection and data cleaning, focusing on retaining only the most relevant and informative features for model training and evaluation. At the end of the feature engineering, there were 25 columns left in my train and test dataset.

Best Practices in Data Preparation & Feature Engineering:

24 features left in my train and test dataset:

'age', 'class_of_worker', 'education', 'wage_per_hour', 'marital_stat', 'major_industry_code', 'major_occupation_code', 'race', 'hispanic_origin', 'sex', 'full_or_part_time_employment_stat', 'capital_gains', 'capital_losses', 'dividends_from_stocks', 'federal_income_tax_liability', 'detailed_household_and_family_stat', 'detailed_household_summary_in_household', 'migration_code_change_in_msa', 'migration_code_change_in_reg', 'migration_code_move_within_reg', 'live_in_this_house_1_year_ago', 'country_of_birth_self', 'citizenship', 'own_business_or_self_employed'.

+ 1 target: Income.

Best Practices in Preprocess Data & Data Modelling:

1. Data Preprocessing:

- I defined categorical and numerical features based on the training data and set up a **ColumnTransformer** to preprocess the data. This practice aligns with best practices in preparing the data for modeling, ensuring consistency and compatibility between the training and test datasets.

Best Practices in Preprocess Data & Data Modelling:

2. Model Training:

- I set up a pipeline that included preprocessing steps followed by machine learning model training. This approach is consistent with best practices for building robust and reproducible machine learning workflows.

Best Practices in Preprocess Data & Data Modelling:

3. Model Evaluation:

- I split the data into training and test sets, trained the model on the training data, and evaluated its performance on the test data using various metrics. This evaluation methodology adheres to best practices, providing insights into the model's predictive power and generalization ability.

Best Practices in Preprocess Data & Data Modelling:

4. Feature Importance:

- I calculated and visualized the feature importances of the trained model, which helps in understanding the relative importance of different features in predicting the target variable. This practice is recommended for interpreting model outputs and identifying key drivers of the target variable.

Best Practices in Model Assesment:

I implemented three different models (Random Forest, Logistic Regression, and Decision Tree) to predict whether income exceeds $50k or not, and provided detailed results for each, including feature importance from each model.

Let's analyze each model's performance and interpret the feature importance outputs.

Best Practices in Model Comparison and Interpretation:

1. Random Forest

• Performance: High accuracy (95.19%) and AUC Score (0.9316), indicating strong classification performance, particularly in distinguishing between the two income classes.

• Classification Report: High precision and recall for the "less than $50k" class. However, for the "greater than $50k" class, although precision is moderately high, recall is quite low (0.39), indicating many false negatives.

• Feature Importance: Major features include 'detailed_household_summary_in_household', 'age', and 'dividends_from_stocks'. This suggests that household context, age, and investment returns are key determinants in predicting higher income levels.

Best Practices in Model Comparison and Interpretation:

2. Logistic Regression

• Performance: Similar to Random Forest with slightly higher AUC (0.9379) but slightly lower accuracy (95.07%).

• Classification Report: Similar trends to Random Forest, with even lower recall for the "greater than $50k" class.

• Feature Importance: Dominated by financial variables ('dividends_from_stocks', 'capital_gains', 'capital_losses') and education ('education_Doctorate degree'). Logistic regression coefficients indicate the strong influence of these features on the likelihood of having a higher income.

Best Practices in Model Comparison and Interpretation:

3. Decision Tree

• Performance: Lower accuracy (93.02%) and significantly lower AUC (0.7157) compared to the other models.

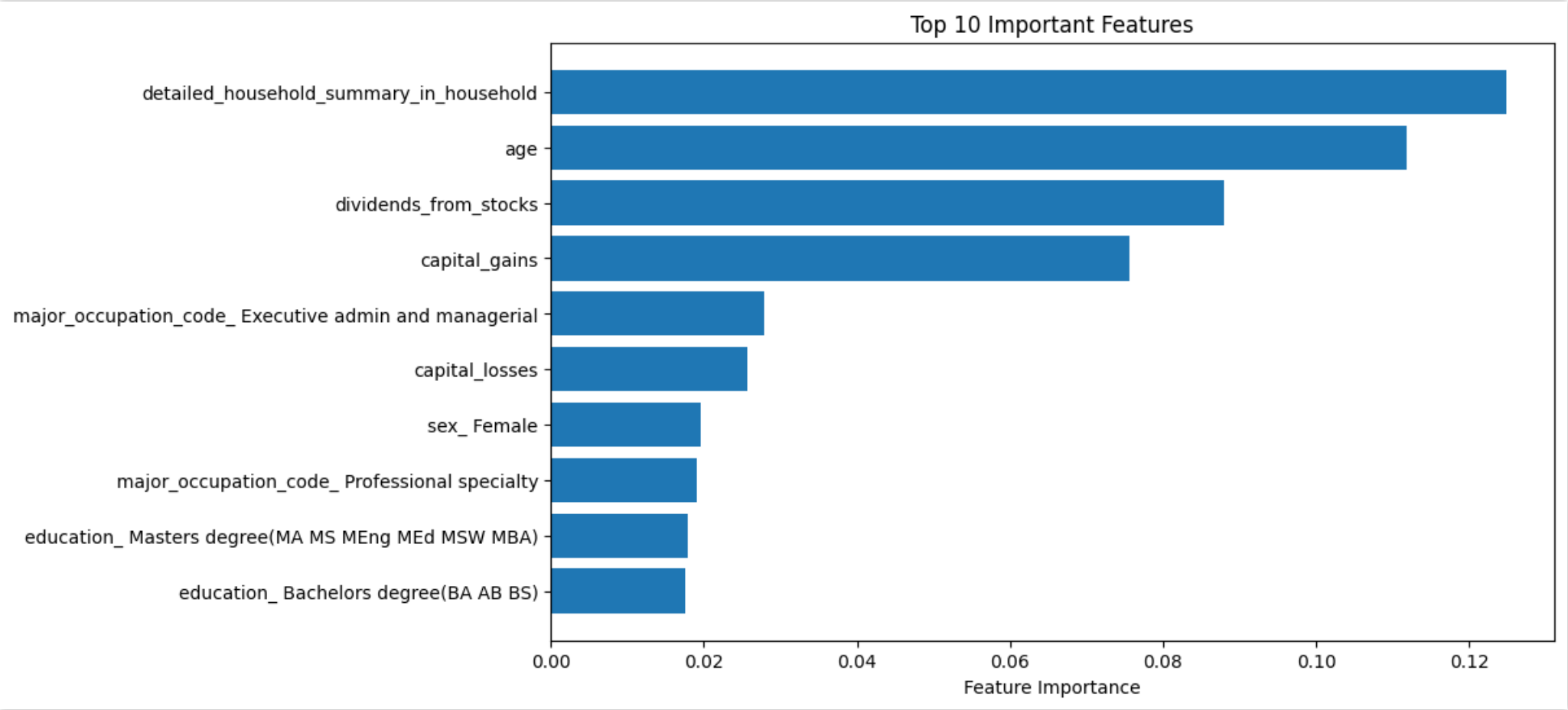
• Classification Report: Similar precision for the "less than $50k" class as the other models but much lower recall and precision for the "greater than $50k" class.

• Feature Importance: Like Random Forest, it shows a mix of demographic and financial variables. Unique in placing significant importance on 'wage_per_hour', a feature less prominent in the other models.

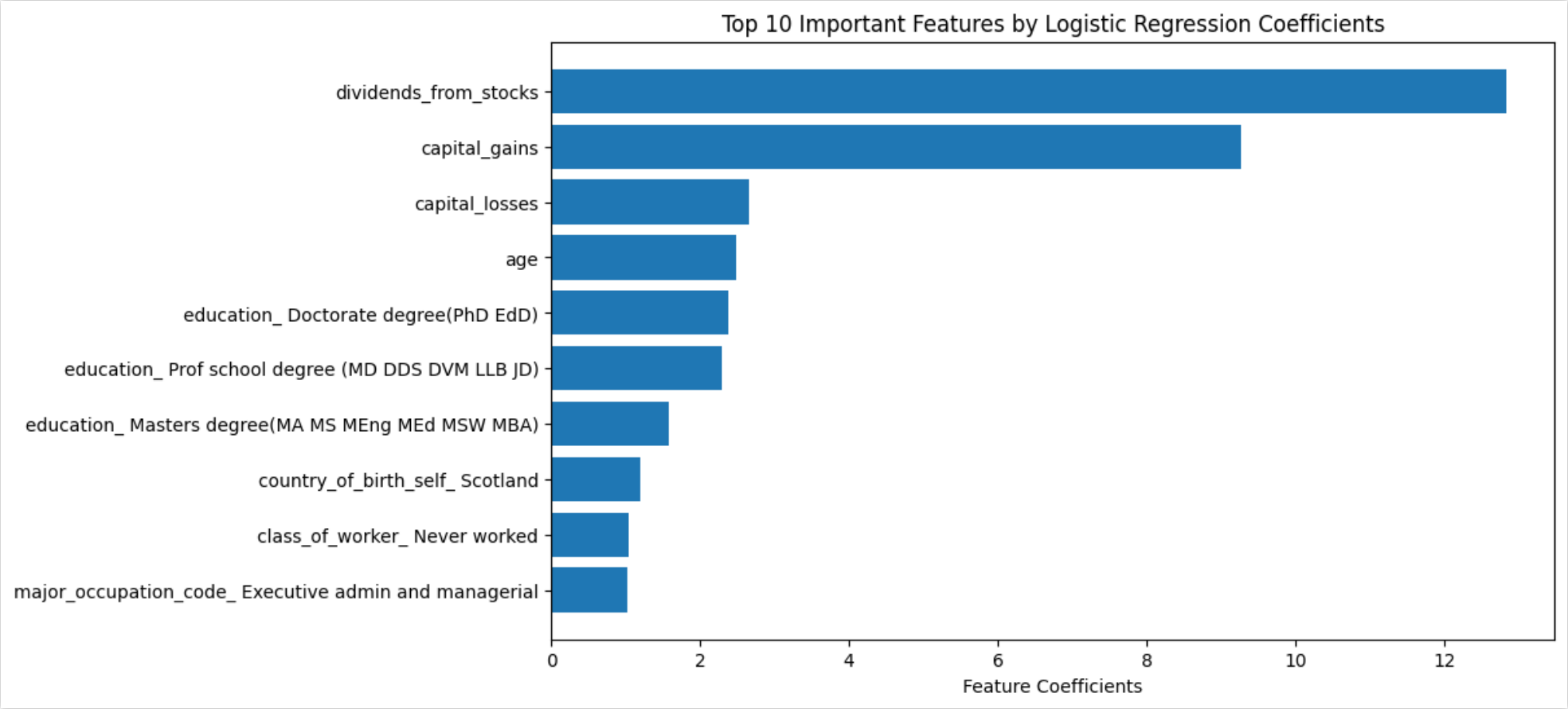
Metrics Explained

- Random Forest, Logistic Regression, and Decision Tree were utilized to classify income levels (above or below $50k).
- Accuracy measures the overall correctness of the model, indicating a high level of correct predictions across the models, especially in Random Forest and Logistic Regression.
- AUC (Area Under the Curve) assesses the model's ability to discriminate between the classes, with Logistic Regression performing slightly better than Random Forest and significantly better than Decision Tree.
- Classification reports reveal better precision and recall for the lower income class across all models, but particularly low recall for the higher income class in Decision Tree.

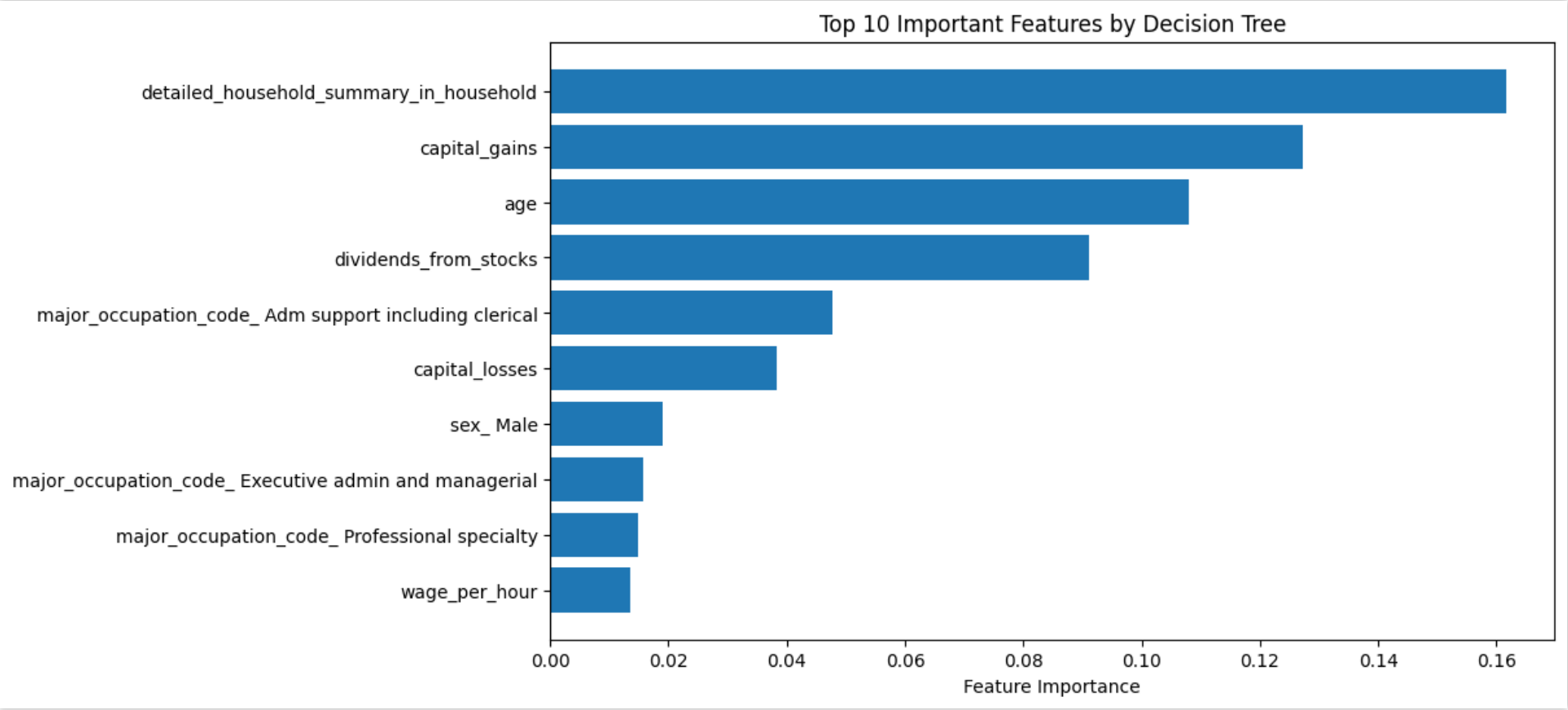
Random Forest: Feature Importance



Logistic Regression: Feature Importance



Decision Tree: Feature Importance



Key Feature Insights

• Financial Indicators like dividends, capital gains, and losses were pivotal in all models, indicating a strong link between financial health and higher income brackets.

• Demographic Factors such as age and household details were also significant, suggesting their importance in income dynamics.

• Occupational and Educational Attributes were crucial, especially in Logistic Regression, highlighting the impact of job type and education level on income.

• The Gender variable in the Decision Tree model indicated potential disparities in income based on gender.

Recommendations

• Model Selection: Given its balanced performance and interpretative ability, Random Forest is recommended for deployment, though Logistic Regression also offers valuable insights due to its interpretability.

• Class Imbalance: Addressing the low recall for the higher income class through techniques like SMOTE or adjusting class weights could improve model fairness and accuracy.

• Feature Engineering: Investigating interactions between key features such as age and education could unveil deeper insights and enhance model performance.

Future Improvements

- Enhanced Feature Selection: Implementing advanced feature selection could refine model inputs, particularly improving Logistic Regression and Decision Tree outcomes.

- Model Tuning: Hyperparameter tuning can optimize model settings to enhance accuracy and reduce overfitting.

- Hybrid Models: Combining model strengths through ensemble techniques or stacking could leverage their unique advantages for better overall performance.

- Regular Model Updates: Continuously updating and validating the models with new data will ensure they remain accurate and relevant over time.

Conclusion

For the task of classifying income levels, the Random Forest model stands out due to its high accuracy, excellent AUC score, and relatively balanced performance across income classes.

However, all models benefit from improvements in feature engineering, handling class imbalance, and continuous tuning and testing to adapt to new data.

data
iku

Thank you for your Attention