SCHOOL OF
SCIENCE &
TECHNOLOGY

PREDICTING HIGH INCOME USING CLASSIFICATION MODELS

Contents

Executive Summary

This report details the comprehensive analysis carried out to try to predict whether a US citizen earns more than 50K USD per year. Using the Adult dataset, the study involves a thorough exploratory data analysis (EDA), feature engineering, and model evaluation. Four models were compared within a single integrated pipeline: a pruned Decision Tree, a Bagging classifier, a Random Forest, and an XGBoost classifier. Although the XGBoost classifier demonstrated the highest accuracy and the highest accuracy (improving by less than 2% compared to the Decision Tree and Random Forest), considerations around computational expense suggest that a pruned decision tree might be preferred for larger datasets, while boosting methods are more advantageous when working with moderately sized data. The study findings highlight the **importance of education, hours of work per week, and having stock investments as drivers of higher income,** possibly shifting educational and financial policies and governmental projects to the future.

Key Findings

- **Important predictors:** Capital Gain, Years of Education, Age, Number of Hours worked per Week, and capital loss are the most relevant features when predicting whether a citizen will have an income higher or lower than $50K per year, sidelining variables such as country of birth, marital status, and gender.
- **Income Profile Insights:** The models consistently revealed that individuals with more advanced education levels and those who work more hours per week generally have a higher likelihood of exceeding the 50K income threshold. Occupation type also plays a significant role, as certain professions correlate more strongly with higher earnings
- **Performance Comparison:** XGBoost yielded the highest overall accuracy. However, the performance margin over the pruned Decision Tree and Random Forest was relatively small (less than 2%). This underscores that while ensemble methods may offer the best performance, simpler models can still provide strong insights with lower computational overhead.
- **Practicality vs. Performance:** The pruned Decision Tree is fast and resource-efficient, making it appealing for large datasets or real-time scenarios. Conversely, XGBoost may be better suited for smaller or more manageable datasets where even slight gains in predictive accuracy can significantly impact decision-making.

Data Preprocessing and Feature Engineering

Data Loading and Transformation

- **Dataset Import:** The dataset used for this analysis (adult-all.csv) was imported with a clearly defined set of headers. The transformation process began by mapping the

target variable, income_over_50K, to a binary format (1 for ">50K" and 0 for "<=50K").
- **Recoding of Categorical Features:** The birth_country variable was recoded to a binary indicator where a value of 1 signified “United-States” and 0 represented all other countries. This transformation was vital in ensuring that the model could focus on domestic income patterns.

Data Cleaning

- **Handling Missing Values:** The dataset contained missing values, represented by “?”. These were replaced with a recognized null marker and then dropped, as the missing data comprised less than 10% of the overall entries. This ensured that the models were trained on reliable and complete data.
- **Column Pruning:** The variables fnlwgt and education were dropped. The former was not relevant for predictive classification, and the latter was redundant because education_num already captured education level.

Final Dataset Preparation

After performing the above transformations and cleaning steps, the dataset was ready for model training. The final features included all remaining variables that were deemed predictive of the target outcome, ensuring that the data was in a consistent and model-friendly format.

Model Development and Evaluation

Overview of the Modeling Pipeline

The modeling process involved an iterative evaluation of four different classifiers integrated into a single pipeline. The pipeline was designed to allow for direct comparison between the models on the same dataset, ensuring that the evaluation metrics were directly comparable.

Models Evaluated

1. **Pruned Decision Tree Classifier:**
    - **Approach:** A baseline decision tree was built and then pruned to prevent overfitting.
    - **Advantages:** Quick training time and low computational cost.
    - **Findings:** Delivered competitive accuracy values, making it a strong candidate for large-scale applications.

2. **Bagging Classifier:**
    - **Approach:** An ensemble method using bagging was applied to the pruned decision tree to reduce variance and improve stability.

    - **Advantages:** Enhances robustness over a single decision tree by aggregating multiple models.
    - **Findings:** Delivered worse performance compared to the basic pruned decision tree.

3. **Random Forest Classifier:**
    - **Approach:** Building on the bagging method, a Random Forest was employed with hyperparameter tuning (via grid search) to optimize performance.
    - **Advantages:** Demonstrated improved performance in terms of accuracy compared to the bagged and pruned tree model.
    - **Findings:** Achieved results nearly on par with the more complex XGBoost classifier, with a lower computational cost than boosting.

4. **XGBoost Classifier:**
    - **Approach:** An XGBoost classifier was trained with hyperparameter tuning via cross-validation.
    - **Advantages:** Delivered the best overall performance, optimizing accuracy, although at a higher computational expense.
    - **Findings:** Outperformed the other models but at a higher computational cost. For organizations that can handle heavier computation, the slightly improved metrics can yield better business outcomes

Evaluation Metrics

- **Accuracy:** This was the primary metric for assessing model performance. Accuracy provided a straightforward measure of correct classification, comparing models directly.
- **Precision:** Offered insight into the models' capacity to correctly identify positive cases (earning >50K) and avoid misclassifying negatives, which may be especially useful for targeted interventions (e.g., career counseling programs aimed at lower-income individuals).

Detailed Comparison

The iterative process revealed that while XGBoost provided the marginally best performance, the differences in accuracy among the pruned Decision Tree, Random Forest, and XGBoost were minimal. This suggests that for extremely large datasets, where computational efficiency is paramount, the pruned Decision Tree might be preferred over the more computationally intensive boosting method. Conversely, when the prediction task is on a smaller scale, the slight performance edge of XGBoost can be fully leveraged.

Key Findings from Predictions

From a broader perspective, the models strongly indicate that:

1. ***Educational Attainment Matters*:** The higher the education, the greater the chance of surpassing 50K. This suggests **that investment in education or upskilling programs** can significantly influence an individual's earning potential.
2. ***Longer Work Hours*:** Individuals who work more hours per week have an increased likelihood of higher earnings, highlighting potential policy discussions around **work-hour regulations and wage structures**.
3. ***Occupation Type**:* Certain occupations correlate more strongly with high income (e.g., managerial or professional roles). Organizations aiming to bolster employee growth might **design career-path programs that lead workers toward these roles**.
4. ***Capital gains and losses***: The study showed that the people in the sample with capitals gains and losses had a higher probability of making over 50K. This is a little ambiguous, as it could very well be that high-income subjects can invest in the stock market with greater frequency than low-income ones, but also opens the doors to considering whether **financial education could serve as a tool to drive income higher.**

Conclusions and Recommendations

Business Insights

- ***Strategic Investment in Education**:* Policymakers and social programs could **prioritize educational grants or scholarships**, as incremental gains in education correlate with higher incomes. Also, it'd be worth seeing if financial education could benefit here.
- ***Corporate Hiring and Training*:** Companies may focus on upskilling their existing workforce, especially in roles that tend to pay above the 50K threshold, boosting overall economic well-being.
- ***Workforce Mobility*:** For those seeking higher pay, transitioning into in-demand occupations could be a key driver. **Vocational training or targeted certification programs** may facilitate these moves.

Model Recommendations

- ***High-Volume, Real-Time Predictions*:** A **pruned Decision Tree** or **Random Forest** is recommended due to faster processing times and strong overall accuracy. These methods are cost-effective and straightforward to maintain.
- ***Focused, High-Stakes Applications*:** Where every fraction of a percent in predictive accuracy matters—such as specialized recruitment or advanced analytics—the **XGBoost** classifier justifies its computational cost by offering superior performance.
- ***Ensemble Flexibility*:** Organizations can adopt a multi-model approach, running XGBoost for key decisions while defaulting to a pruned Decision Tree for large-scale tasks, ensuring a balanced approach.

Technical Annex

1. Data Inspection

- **Data Source:** The Adult dataset provided demographic and employment information for a large sample of US individuals. Key fields included age, education_num, occupation, marital_status, hours_per_week, and income_over_50K.
- **Initial Structure:** Before any transformations, the dataset included columns representing personal details (e.g., race, sex, birth_country) and employment variables (e.g., occupation, hours_per_week) with potential missing data indicated by "?".

2. Data Preprocessing and Cleaning

- **Mapping and Cleaning:**
  - The column income_over_50K was converted to a binary format: 1 for those earning above 50K, 0 for those at or below 50K.
  - Missing values denoted by "?" were converted to Pythonic null values and dropped since they accounted for less than 10% of the dataset.
- **Column Removal:**
  - fnlwgt was removed, as it represented sampling weights irrelevant to the predictive task.
  - education was dropped due to redundancy with education_num (numeric encoding of education level).
- **Categorical Adjustments:**
  - birth_country was simplified to a US vs. non-US binary indicator.

3. Feature Engineering

- **Primary Predictors:** Based on domain knowledge and exploratory analysis, education_num, hours_per_week, occupation, and marital_status emerged as key features.
- **Handling Categorical Variables:** Most categorical features (e.g., marital_status, occupation) were label-encoded or handled natively by tree-based models.
- **Data Split:** For model validation, the dataset was typically split into training and testing sets (e.g., 70%/30%) to evaluate out-of-sample performance.

4. Model Selection and Tuning

- **Pruned Decision Tree:**
  - **Rationale:** Provided a transparent baseline. Pruning was performed to control overfitting, ensuring the tree remained interpretable while maintaining strong predictive power.
- **Bagging (Bootstrap Aggregation):**
  - **Description:** Combined multiple decision trees trained on different bootstrap samples to reduce variance.
  - **Hyperparameters:** Number of estimators and sample sizes were tuned via grid search.

- **Random Forest:**
  - **Description:** Extended bagging by introducing random feature selection, further reducing correlation among trees.
  - **Hyperparameters:** Number of trees, maximum depth, and minimum samples per leaf were optimized with grid search.

- **XGBoost Classifier:**
  - **Description:** An iterative boosting algorithm that adds trees to minimize a defined loss function, well-known for high performance in structured data.
  - **Hyperparameter Tuning:** Learning rate, maximum depth, gamma, and n_estimators were iteratively optimized via cross-validation.

5. Evaluation Metrics

- **Accuracy:** Used to measure the proportion of correctly classified instances, straightforwardly indicating how well the model distinguishes high-income vs. non-high-income individuals.
- **Precision:** Indicated how many of the individuals predicted to be above 50K earned this amount.

6. Implementation and Practical Use

- **Business Application:** The selected model can be integrated into decision-making systems, such as applicant screening tools or targeted marketing campaigns.
- **Scalability:**
  For large-scale, real-time predictions, more computationally efficient models like pruned Decision Trees or Random Forests are favored. For targeted or smaller datasets, XGBoost's marginally better accuracy could be invaluable.

7. Graphs

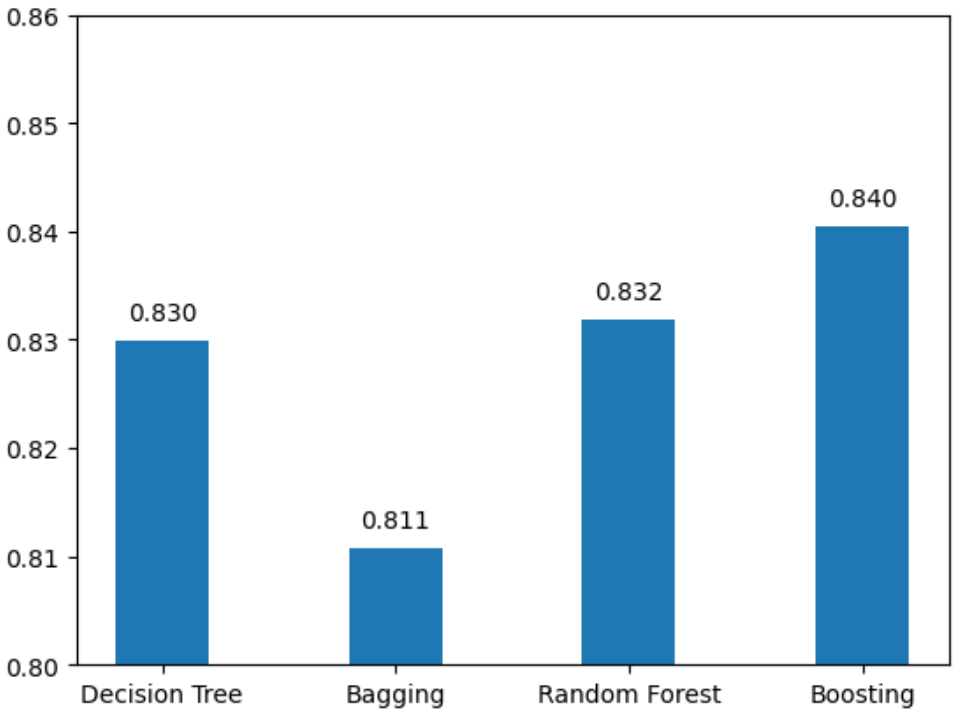

Figure 1. Accuracies for the different models used.

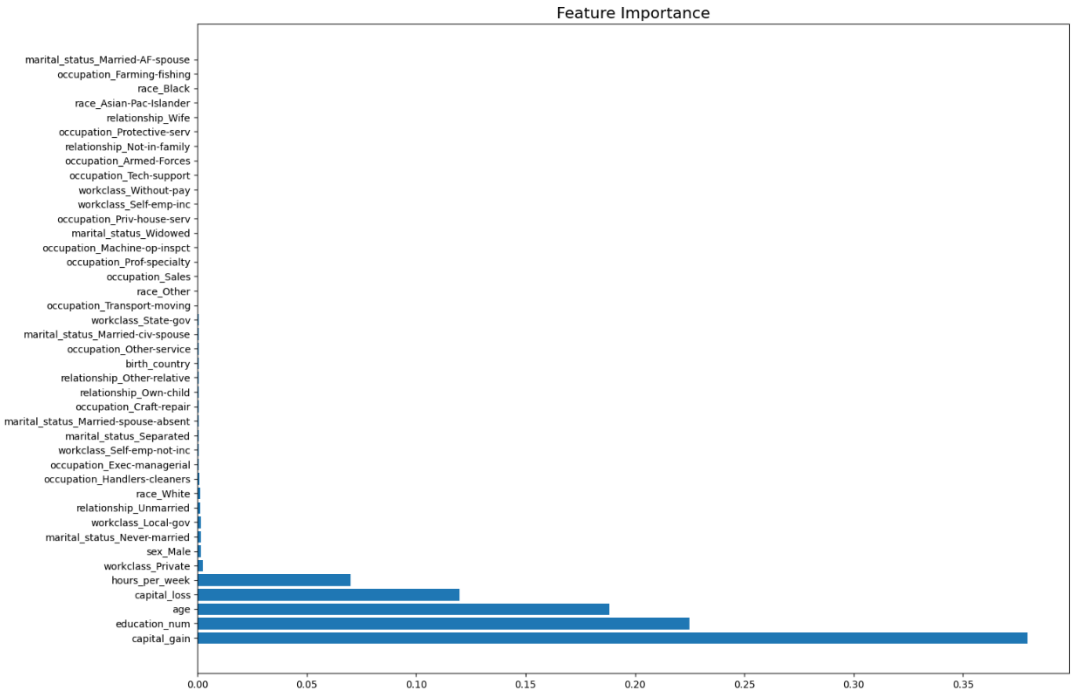

Figure 2. Pruned Decision Tree Feature Importance.

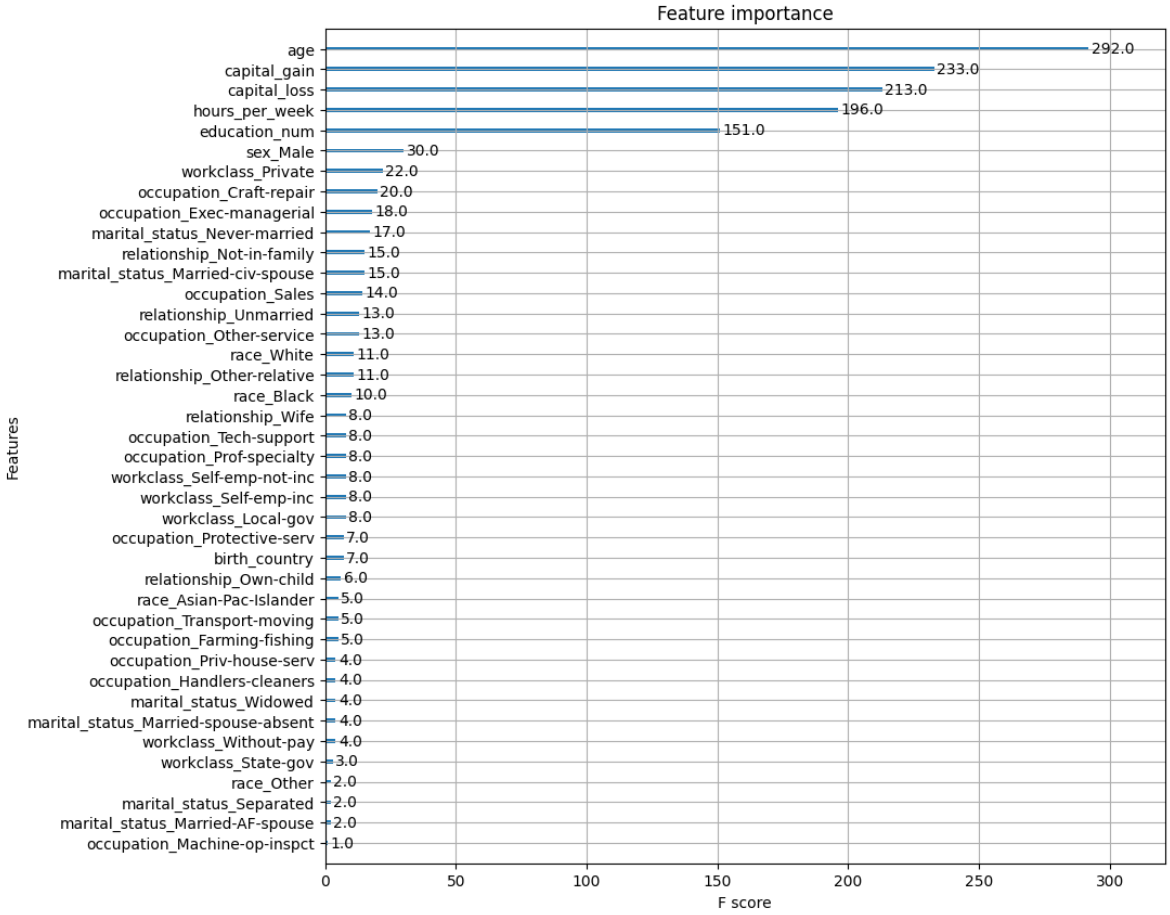

Figure 3. Boosting Feature Importance.

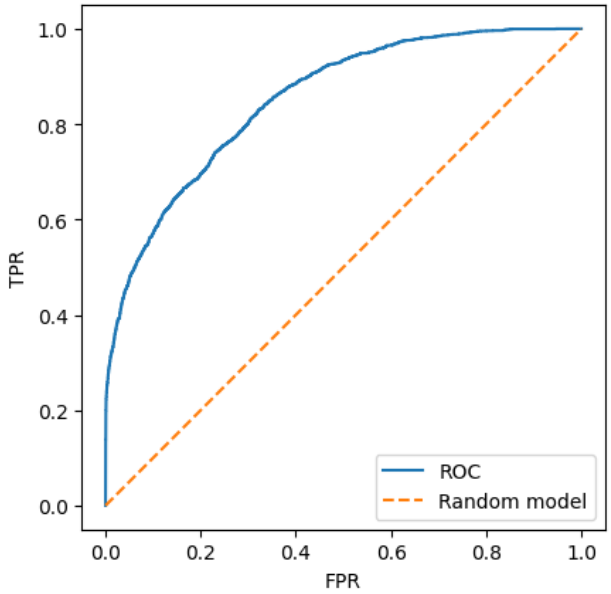

Figure 4. Random Forest ROC Curve

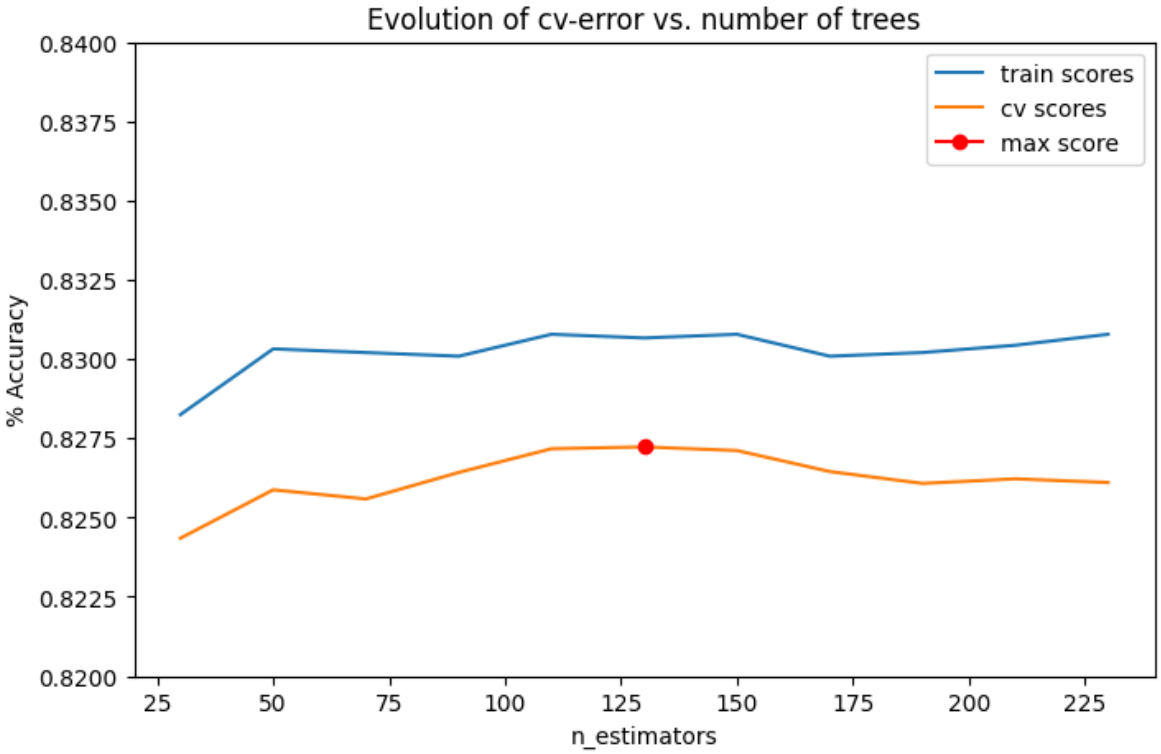

Figure 5. Random Forest Cross Validation - Trees vs Accuracy.