Adult Income Prediction – Personal Project Report

Goal:

To assess and demonstrate hands-on skills in data ingestion, cleaning, feature engineering, model building, evaluation, and pipeline design using Python libraries like Pandas and Scikit-learn.

Task:

Build a binary classification model to predict whether an individual earns more than $50K/year based on demographic and work-related attributes from the UCI Adult dataset.

1. Data Ingestion & Cleaning

I started by importing the dataset (in CSV format), then used Pandas to explore and clean it:

Missing Values:

The dataset had several missing entries represented by '?', particularly in Workclass, Occupation, and Native-country.

- Numerical Features (e.g., Age, Capital-Gain): Replaced missing or outlier values using median imputation to prevent skewing.
- Categorical Features: Retained missing entries by labeling them "Missing" to preserve patterns rather than drop data.

2. Feature Engineering

To enhance predictive power, I created the following engineered features:

- Capital-Total = Capital-Gain - Capital-Loss
  Helped the model understand net financial value instead of treating gain/loss separately.

- Age-Group: Binned ages into ranges (e.g., 25–35, 36–50)
  Useful for capturing lifecycle trends in income potential.

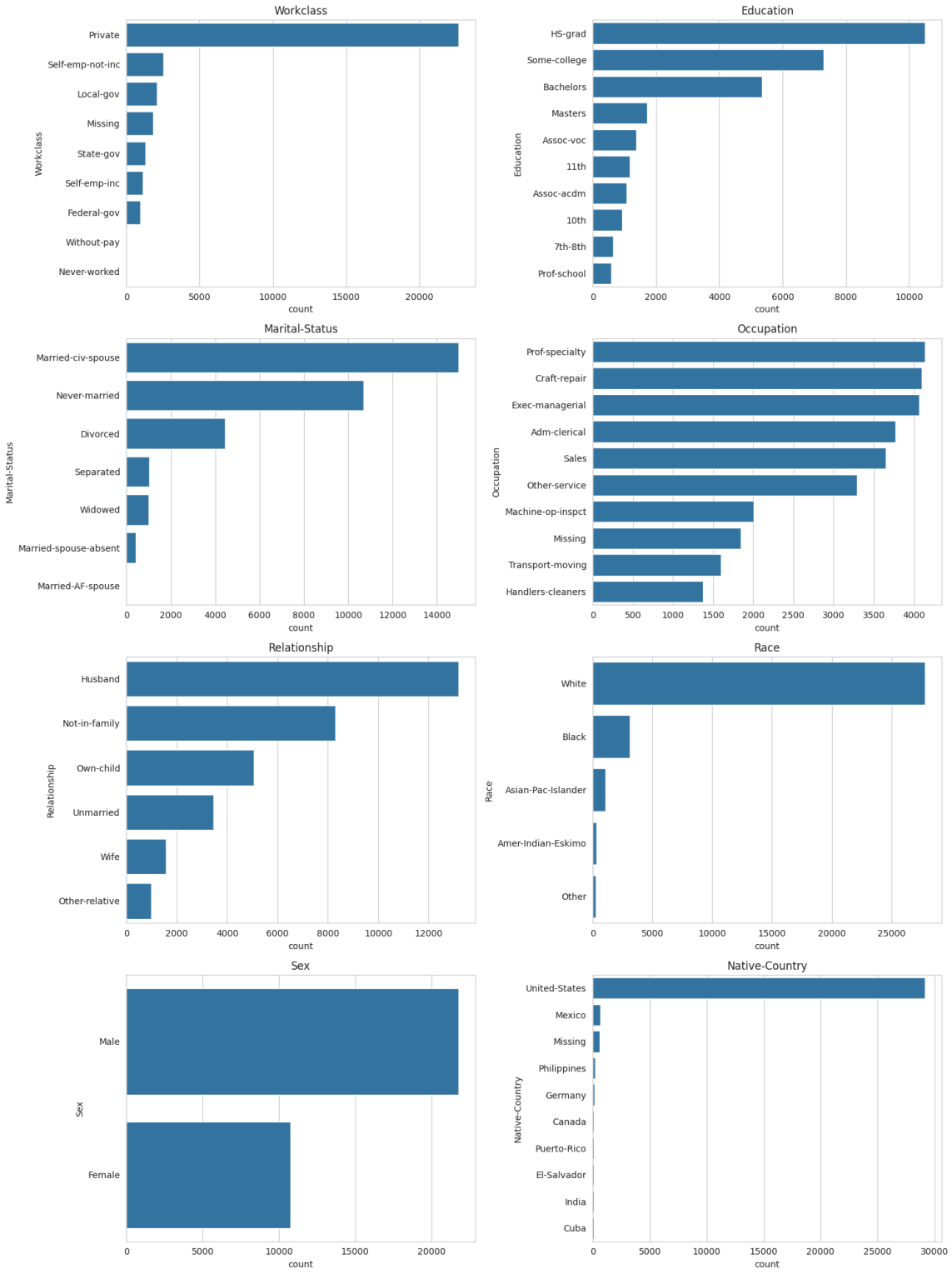

Fig. 1: Categorical features

This step improved model interpretability and performance during training.

3. Model Building & Evaluation

I built and compared three classification models using Scikit-learn and XGBoost:

- Logistic Regression – as a baseline model.
- Random Forest Classifier – for handling categorical/numerical mix.
- XGBoost Classifier – known for strong performance on tabular data.

Evaluation Metrics Used:
Accuracy

F1 Score (for handling class imbalance)

ROC AUC Score

Model	ROC AUC	Accuracy	F1 Score
XGBoost	0.92	87.2%	0.72
Random Forest	0.91	86.1%	0.70
Logistic Regression	0.89	85.2%	0.65

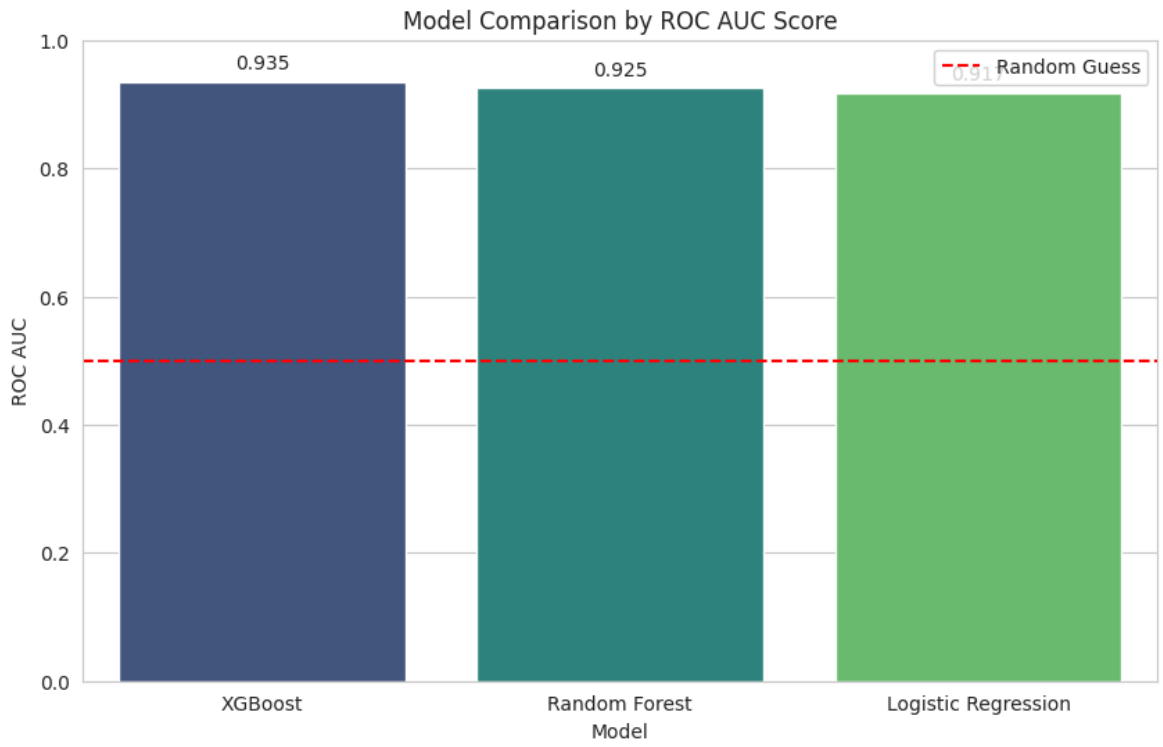

Fig. 2: Model Comparison by ROC AUC Score

XGBoost performed best with strong generalization, better feature handling, and minimal overfitting.

4. Key Insights from Feature Importance

Top predictors for high income included:

- Capital-Gain: Investment returns matter more than salary alone.
- Education Level: Higher education → better pay opportunities.
- Marital Status (Married-civ-spouse): Indicative of dual income stability.
- Age (35–50): Correlates with peak career earning phase.
- Hours-per-week: Productivity plays a role but plateaus beyond 45–50 hours.

5. Conclusion

This project helped me put into practice all core steps of a real-world data science workflow—from data wrangling and transformation to advanced modeling and performance tuning. The structured pipeline approach and deep model evaluation using metrics like ROC AUC and F1 helped ensure the solution was robust, reproducible, and explainable.