**Final Capstone Project Report**  
**Project Title:** Predicting Income Level Using the Adult Census Dataset  
**Course:** CSIS 503 - Data Science & Analytics  
**Institution:** Osiri University  
**Submitted On:** August 25, 2025  
**Instructor:** Dr. Noble Anumbe  
**TA:** Sebastian Boscan Villalobos  
**GitHub Repository:** [https://github.com/](https://github.com/Amblessed01)[Amblessed01/csis503-capstone-income-prediction](https://github.com/Amblessed01/csis503-capstone-income-prediction)

### ****Executive Summary****

This project, conducted by Group 9, applies the complete data science lifecycle to predict whether an individual earns above or below $50,000 annually using the U.S. Adult Census Dataset. We leveraged logistic regression, random forest, and XGBoost classifiers, achieving a highest ROC-AUC score of **0.968** with XGBoost. Key insights indicate that **education level**, **capital gains**, and **hours worked per week** are strong predictors of income level. Interpretability was ensured through SHAP, enhancing the transparency of our models. The findings have practical applications in workforce development, policy planning, and financial behavior analysis.

### ****Project Overview****

Group 9 embarked on this capstone project to demonstrate our readiness as data-driven problem solvers. Using the U.S. Adult Income Census data, we developed machine learning models to classify individuals’ income level and derived actionable socio-economic insights.

### ****Dataset:****

* Source: UCI Machine Learning Repository (Adult Census Dataset)
* Records: ~48,842 entries
* Features: 99 demographic and socio-economic attributes

### ****Project Objectives****

* Apply a full data science pipeline: data wrangling, EDA, modeling, and interpretation.
* Predict income level (>$50K or <=$50K) via binary classification.
* Evaluate and compare multiple models.
* Interpret feature contributions using SHAP.

### ****Data Wrangling and Cleaning****

* Extracted datasets from UCI and converted to CSV.
* Combined adult.csv (training) and adult\_test.csv (testing) into df\_combined.
* Cleaned income labels and stripped whitespaces.
* Added a source column to track origin.
* Created target variable is\_high\_income:
* 1 if income > $50K, 0 otherwise.
* Converted age to numeric, removed invalids, and dropped null rows.

### ****Feature Engineering****

* Binned age: 18-30, 31-45, 46-60, 60+
* One-hot encoded categorical features
* Standardized numerical variables
* Addressed class imbalance using SMOTE

### ****Exploratory Data Analysis (EDA)****

* Visualized distributions for: age, education\_num, hours\_per\_week, capital\_gain, capital\_loss
* Boxplots: income group comparisons
* Correlation matrix and pairplots: feature relationships

Key Observations:

Higher education\_num, capital\_gain, and hours\_per\_week correlate with higher income

### ****Predictive Modeling****

**Models Used:**

* Logistic Regression
* Random Forest Classifier
* XGBoost Classifier

**Evaluation Metrics:**

Accuracy, Precision, Recall, F1-score, ROC-AUC, Confusion Matrix

**Model Performance Comparison:**

| **Model** | **Accuracy** | **F1 Score** | **ROC-AUC** |
| --- | --- | --- | --- |
| Logistic Regression | 0.85 | 0.78 | 0.88 |
| Random Forest | 0.87 | 0.80 | 0.90 |
| **XGBoost** | **0.89** | **0.84** | **0.928** |

**XGBoost Confusion Matrix:**  
precision recall f1-score support

0 0.91 0.95 0.93 7407

1 0.82 0.69 0.75 2360

accuracy 0.89 9767

macro avg 0.87 0.82 0.84 9767  
weighted avg 0.89 0.89 0.89 9767

### ****Model Interpretability (SHAP)****

* SHAP Summary Bar Plot: Top feature impacts
* SHAP Beeswarm Plot: Individual prediction influences

**Top Predictors Identified:**

* education\_num
* capital\_gain
* hours\_per\_week
* Age
* capital\_loss

Notably, capital\_loss had a high SHAP value (~0.127), suggesting its relevance to investment-related income patterns.

### ****Insights & Implications****

* Higher education and capital gains are linked to higher income.
* Work hours and age are consistent predictors.
* Financial investment behavior (capital\_loss) significantly impacts income.

Implications span:

* **Policy-making**: Data-informed decisions on education and job training.
* **Business/HR**: Tailored recruitment based on predictive socio-economic indicators.

### ****Reproducibility:**** requirements.txt file for package consistency

Clean, modular Jupyter notebook:

* Data loading & cleaning
* EDA
* Modeling
* SHAP explanations

### ****Technical Deliverables****

### Jupyter Notebook: capstone\_pipeline.ipynb

* requirements.txt
* Visual outputs: charts, plots, confusion matrices
* SHAP visualizations
* Professional README.md for GitHub documentation

### ****Group 9 Members & Roles****

| **Member Name** | **Role Description** |
| --- | --- |
| ThankGod Israel | Project Lead, Data Cleaning, EDA, Modeling Lead, Hyperparameter Tuning, Final Report |
| Oni Akintunde Julius | Visualizations, EDA Charts, Feature Distributions, Correlation Heatmaps |
| Victory Madu | Quality Assurance (QA), Code Review, Testing Pipeline Consistency |
| Chikezie Amarachi | GitHub Repository Setup, Version Control, File Management |
| Doris Akachukwu | Slide Creation, Presentation Design, Team Presentation Prep |
| Mary Paschal Iwundu | Data Acquisition, Model Evaluation |
| Daniel Getaye Tareke | Data Cleaning, EDA & Material Sourcing |

### ****Conclusion****

Our project effectively demonstrates mastery of the data science process, from raw data to predictive insights. By focusing on transparency, interpretability, and real-world relevance, we showcase our preparedness to apply data science solutions to complex socio-economic problems.