Group Name: Group 9
Course: CSIS 503 – Data Science & Analytics | Instructor: Dr. Noble Anumbe
Date: Friday, August 22, 2025
Group Members: ThankGod Israel, Oni Akintunde Julius, Victory Madu, Chikezie Amarachi, Doris Akachukwu, Mary Paschal Iwundu, Daniel Getaye Tareke

1. Project Title
Predicting Income Levels Using the U.S. Adult Census Dataset

2. Introduction & Objective
This project investigates whether demographic and socio-economic characteristics can predict if an individual earns more than $50,000 annually. The goal is to apply the end-to-end data science lifecycle—from data cleaning to model interpretation—to build accurate and transparent predictive models.

The dataset is relevant because income prediction plays a role in policy planning, workforce development, and recruitment strategies. Our objective was to compare machine learning models, address class imbalance, and identify the most influential predictors of income.

3. Dataset Description
- Source: UCI Machine Learning Repository – Adult Census Dataset
- Size: ~48,842 records, 14 attributes
- Key Variables: age, education_num, occupation, hours_per_week, capital_gain, capital_loss, income (target)
- Cleaning: Removed missing/invalid entries, standardized data types, stripped whitespace from labels, and created a binary target variable (is_high_income).

4. Methodology
We applied the following steps:
- Exploratory Data Analysis (EDA): Plotted distributions of age, education_num, hours_per_week, capital_gain/loss; generated correlation heatmaps and boxplots.
- Feature Engineering: One-hot encoding of categorical variables, standardization of numeric variables, binning age groups.
- Imbalance Handling: Used SMOTE to oversample minority class.
- Models: Logistic Regression, Random Forest, and XGBoost.
- Tools/Libraries: Python, pandas, matplotlib, seaborn, scikit-learn, imbalanced-learn, XGBoost, SHAP.
- Reasoning: Logistic Regression was chosen as a baseline for interpretability; Random Forest was applied to capture non-linear patterns and feature interactions. XGBoost was chosen for its strong performance and high predictive accuracy on structured data, while SHAP was included to ensure model transparency and explainability.

5. Key Findings & Results
Model Performance (Test Set):

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 | 0.93 |

- Top Predictors: education_num, capital_gain, hours_per_week, age, capital_loss.
- Insights from SHAP: Capital gains strongly increase income probability; education and working hours remain consistent predictors; capital loss revealed relevance to financial investment behaviors.

6. Insights & Recommendations

- Our findings suggest that education and financial investments significantly influence income levels.
- Policy-makers could prioritize education and job training programs to reduce income inequality.
- Businesses and HR departments can tailor recruitment by considering socio-economic factors that predict higher income potential.

7. Limitations

The dataset has some limitations:

- It is limited to U.S. demographics from the 1990s, reducing generalizability.
- Class imbalance required oversampling (SMOTE), which may introduce bias.
- The dataset lacks contextual variables such as geographic location (e.g., urban vs. rural differences) and qualitative factors like soft skills (e.g., leadership, communication), which also impact income but are not captured here.

8. Reference & Acknowledments

- Dataset: UCI Machine Learning Repository – Adult Dataset.
- Libraries: pandas, scikit-learn, seaborn, matplotlib, imbalanced-learn, XGBoost, SHAP.
- All technical deliverables, including the Jupyter Notebook (group9_capstone.ipynb), requirements.txt, charts, SHAP visualizations, and README.md, are available in our GitHub repository: https://github.com/Amblessed01/csis503-capstone-income-prediction.
- Group Roles:

Member Name	Role Description
ThankGod Israel	Project Lead, 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 & Material Sourcing