Team Members

Name	**Contribution**	**Role**
Carl Patel (Team Leader)	EDA and Feature Selection	Train Models
Marcello Battista	EDA and Feature Selection	Train Models
Benjamin Tang	Dataset Description	API and Frontend Creation
Makhtar Thioune	Literature Review	Analyse Models
Jordan Tecson	Early Plan For Model Development	Model Plotting

Dataset Description

The dataset represents the Census Income from the UCI Machine Learning Repository. There are 14 features, including demographic (age, sex, race), socioeconomic attributes (workclass, education, occupation, hours-per-week), and financial variables (capital-gain, capital-loss). There is 1 target variable (income) that shows whether or not the individual's income exceeds $50,000/year. There are missing values that are represented with a "?" in the dataset for workclass, occupation, and native country, which means that the individual either didn't answer the question, withheld information for privacy purposes, or didn't fit into the given categories.

Variable Name	**Type**	**Description**	**Missing**
Age	Integer	Individual's age	No
Workclass	Categorical	Private, self-employed, never-worked, etc	2799
fnlwgt (final weight)	Integer	The number of people the census believes the row represents	No
education	Categorical	Bachelor's, some college, high school, etc	No
education-num	Integer	Years of school	No
marital-status	Categorical	Married, Divorced, Widowed, etc	No
occupation	Categorical	Tech-support, sales, etc	2809
relationship	Categorical	Wife, Husband, relative, etc	No
race	Categorical	White, Asian, etc	No
sex	Binary	Female, Male	No
capital-gain	Integer	Capital gains (in US dollars)	No
capital-loss	Integer	Capital losses (in US dollars)	No
hours-per-week	Integer	Number of hours worked per week	No
native-country	Categorical	US, Canada, China, India, etc	857
income	Binary	>50K, <=50K	No

Exploratory Data Analysis (EDA)

For the EDA we made plots for each feature compared to income. By doing this we were able to visualize how each feature compared to a person having >50k of annual income. Here is some analysis of those plots:

- **Income Distribution Bar Chart:** Most individuals earn <=50K, which could bias models toward the majority class unless addressed.
- **Age, Hours per Week, and Capital Features by Income:** High earners tend to be older, work longer hours, and are more likely to report capital gains or losses. This shows clear separation between income groups and could be strong candidates for training.
- **Education by Income:** Individuals with higher education levels are far more likely to earn >50K. Since education-num is a numeric representation of education level, it may be more useful for prediction than a classifier.
- **Workclass and Occupation by Income:** Government, private sector, and professional roles are associated with higher income, while service and manual labor roles are more associated with the <=50K group.
- **Marital and Relationship Status by Income:** Married individuals are more likely to be high earners. Relationship status also reflects household role and can contribute to income prediction.
- **Sex by Income:** Males are more likely to earn >50K, indicating a gender income gap that may need to be carefully considered.
- **Correlation Heatmap:** Education-num with capital gain and hours-per-week both have a moderate positive correlation with income, while fnlwgt is so close to zero that it has little relevance and may be excluded from training.

Feature Selection

Based on the EDA, the most predictive features selected for modeling income are age, capital-gain, capital-loss, hours-per-week, education-num, workclass, marital-status, occupation, relationship, and sex. These features show strong separation between income classes, for instance, high earners tend to be older, work more hours, and often have capital gains or losses. Categorical variables like occupation and marital-status show meaningful patterns, with professionals and married individuals more likely to earn >50K. Features like fnlwgt, native-country, and race were excluded due to weak correlation with income or missing too much data

Early Plan For Model Development

Our plan is to develop a logistic regression model using the Adult dataset from the UCI Machine Learning Repository to predict whether an individual earns more than $50K annually. The development process is as follows:

1. **Load the Data:** Fetch the Adult dataset directly from the repository.
2. **Clean the Data:** Remove entries with missing values and drop irrelevant features.
3. **Split Features and Target:** Separate the dataset into input features and income (output).
4. **Preprocess the Data:** Standardize numerical features and one-hot encode categorical features to train the model.
5. **Train-Test Split:** Split the processed dataset into training and testing subsets to evaluate model generalizability.
6. **Train with Hyperparameter Tuning:** Use grid search to optimize hyperparameters and train the logistic regression model on the training set.
7. **Evaluate Model Performance:** Assess the model using metrics such as accuracy, precision, recall, f1-score, and support.

Literature Review

Across disciplines such as economics, sociology, and public policy, a substantial body of research has explored the relationship between socioeconomic factors and individual income levels. Numerous studies have identified key variables that are strongly associated with income. Abdullah (2015) found an increase in income levels as education increases; Rothwell and Massey (2014) found a strong link between growing up in a poor neighborhood and lower income levels; and Akee et al. (2019) observed differences in income between races, with Blacks, American Indians, and Hispanics all having lower income levels than Whites and Asians on average. These factors are often interrelated and have been shown to significantly influence an individual's earning potential. Given these correlations, it is reasonable to explore whether socioeconomic factors can be used to predict income levels.

Modeling techniques such as logistic regression provide a framework to assess the influence of multiple factors on a target variable. For example, Rahman (2013) utilized logistic regression to estimate the likelihood of poverty in Bangladeshi households. Through his regression analysis, he was able to draw a link between gender makeup of the household, age of household head, occupations of the household, and education of the head of the household. Moreover, Bhushan and Narta (2014) employed a logistic regression model to link financial literacy and the socioeconomic conditions of individuals. Notably, women tended to have lower levels of financial literacy compared to men; an increase in education was associated with higher financial literacy; and one's occupation played a role in determining an individual's financial literacy. These results show logistic regression's ability to model binary socioeconomic factors, highlighting its relevance in classifying income levels based on socioeconomic variables.

Works Cited

Abdullah, A., Doucouliagos, H. and Manning, E. (2015), Does Education Reduce Income Inequality? A Meta-Regression Analysis. *Journal of Economic Surveys,* 29: 301-316. https://doi.org/10.1111/joes.12056

Akee, R., Jones, M.R. & Porter, S.R. (2019), Race Matters: Income Shares, Income Inequality, and Income Mobility for All U.S. Races. *Demography* 56, 999–1021. https://doi.org/10.1007/s13524-019-00773-7

Bhushan, Puneet & Narta, Salma. (2014). Logistic Regression Model for Predicting Financial Literacy. *Commerce and Management Explorer*. 4. 142-153.

Rothwell, J. T., & Massey, D. S. (2015). Geographic Effects on Intergenerational Income Mobility. *Economic Geography*, 91(1), 83–106. https://doi.org/10.1111/ecge.12072