### **PROJECT AND TEAM INFORMATION**

**Project Title**

* InsightIncome: Predicting User Income from Social Media Behavior and Demographics

*(This is a suggestion, feel free to change it!)*

**Student/Team Information**

* **Team Name:** [User to Fill]
* **Team #:** [User to Fill - Mentor needs to assign]
* **Team member 1 (Team Lead):** (Last Name, name: student ID: email, picture): [User to Fill]
* **Team member 2:** (Last Name, name: student ID: email, picture): [User to Fill, if applicable]

### **PROJECT PROGRESS DESCRIPTION (35 pts)**

Project Abstract (2 pts)

This project aims to develop a machine learning model capable of predicting an individual's income level based on their social media usage patterns, demographic information, and other related attributes. The primary goal is to explore the relationships between these features and income, and to build a robust predictive tool. The project involves data collection, extensive exploratory data analysis (EDA) to uncover insights, feature engineering to create meaningful predictors, model development using a Random Forest Regressor in R, and rigorous evaluation to assess its performance. The ultimate outcome is a functional model that can provide income estimates, supported by an understanding of the key factors influencing these predictions.

Updated Project Approach and Architecture (2 pts)

The project follows a structured machine learning workflow using R:

1. **Data Collection:** Utilized a dataset named 'Social Media Dataset.csv' (2002 observations, 15 variables) containing user demographics, social media activity, and income.
2. **Exploratory Data Analysis (EDA):** Performed comprehensive EDA using R libraries such as tidyverse (including ggplot2 for plotting and dplyr for data manipulation) and corrplot to understand data distributions (e.g., average age ~32 years, average income ~$63,472), relationships between variables, and identify patterns through visualizations.
3. **Data Cleaning & Preprocessing:** Verified data integrity; the dataset had no missing values. Categorical variables (Gender, Education.Level, etc.) were encoded by converting them to factors and then to integers. Numerical features were scaled (centered and scaled) using the preProcess function from the caret package.
4. **Feature Engineering:** Created new features in R to enhance predictive power, including Platform\_Count, binary indicators for usage of specific platforms (Instagram, Facebook, etc.), Age\_Group, Usage\_Intensity, and combined platform usage features (e.g., Uses\_Instagram\_Facebook, Professional\_Platforms).
5. **Model Development:** Employed a randomForest model from the randomForest package in R. The model was trained with specified parameters (e.g., ntree = 100) and the importance = TRUE argument to assess feature importance.
6. **Model Evaluation:** Assessed the model's performance on a held-out test set (80/20 split) using R² score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), calculated using functions from the caret package or standard R functions.
7. **Libraries Used:** tidyverse (for ggplot2, dplyr, readr, forcats, stringr, tibble, lubridate, purrr), caret (for preprocessing, data splitting, and evaluation metrics), randomForest (for the regression model), corrplot (for correlation visualization), and gridExtra (for arranging plots). Model and preprocessing objects were saved as .rds files.

**Tasks Completed (7 pts)**

| **Task Completed** | **Team Member** |
| --- | --- |
| Data Collection and Initial Setup in R | [User to Fill] |
| Comprehensive Exploratory Data Analysis (EDA) using R | [User to Fill] |
| Data Cleaning (Verification) and Preprocessing in R | [User toFill] |
| Extensive Feature Engineering in R | [User to Fill] |
| Random Forest Model Training (with specified parameters) in R | [User to Fill] |
| Model Evaluation and Performance Metrics Calculation in R | [User to Fill] |
| Saving of Trained Model and Preprocessing Objects as .rds files | [User to Fill] |

**Challenges/Roadblocks (7 pts)**

* **Feature Selection/Engineering:** Iteratively identifying and creating the most impactful features from the available data in R to maximize model performance while avoiding overfitting or noise.
* **Parameter Selection for RandomForest:** While not an exhaustive grid search, selecting appropriate parameters for the randomForest function (like ntree, mtry) still requires consideration and potentially some experimentation to optimize performance.
* **Package Management & Conflicts:** Ensuring all R packages are installed and managing potential function masking between packages (e.g., dplyr::filter() vs stats::filter(), dplyr::combine() vs randomForest::combine()) requires attention, as noted in the script output.
* **Model Interpretability:** While Random Forest provides feature importance, deeply understanding the nuanced interactions and the precise "why" behind individual predictions remains a complex aspect of such models.

*(These are common challenges. Please tailor them to your specific experience.)*

**Tasks Pending (7 pts)**

| **Task Pending** | **Team Member (to complete the task)** |
| --- | --- |
| User Interface (UI) Development for Model Interaction (e.g., using R Shiny if planned) | [User to Fill] |
| Deployment of the Predictive Model (if planned) | [User to Fill] |
| Further Model Refinement/Testing with New Data | [User to Fill] |
| Preparation of Final Project Report and Presentation | [User to Fill] |

Project Outcome/Deliverables (2 pts)

The key outcomes and deliverables of this project include:

1. A trained and evaluated machine learning model (RandomForest) capable of predicting income based on social media and demographic data, saved as income\_regressor\_model.rds.
2. Saved preprocessing objects (preprocess\_params.rds) and factor level information (factor\_levels.rds) for consistent data transformation in R.
3. A comprehensive set of data visualizations and insights derived from the Exploratory Data Analysis (saved as PNGs in 'charts' directory).
4. An R script (main.R) detailing the entire workflow from data loading to model prediction.
5. (Future) A user interface allowing users to input data and receive income predictions, potentially developed using R Shiny.

Progress Overview (2 pts)

The project has made significant progress using R for its analytical core. The data analysis, feature engineering, model development, and evaluation phases are complete. The core predictive engine is functional, achieving an R² of approximately 0.98 on the test set, and has been tested. The primary remaining component is the development of the user interface, which is currently pending. Documentation and preparation for the final report are ongoing. Overall, the project is largely on schedule for its analytical components.

**Codebase Information (2 pts)**

* **Repository Link:** [User to Fill - e.g., GitHub link]
* **Branch:** [User to Fill - e.g., main, development]
* **Important Commits:**
  + Initial data loading and EDA setup in R.
  + Completion of feature engineering and preprocessing steps in R.
  + Implementation and training of the Random Forest model in R.
  + Final model evaluation and saving of R model artifacts (.rds).  
    (User to provide specific commit messages or hashes if available/relevant)
* The primary codebase is contained within the main.R script.

**Testing and Validation Status (2 pts)**

| **Test Type** | **Status (Pass/Fail)** | **Notes** |
| --- | --- | --- |
| Model Performance (R² Score) | Pass | Achieved an R² score of **0.9817** on the test set, indicating the model explains approximately 98.17% of the variance in income. |
| Model Performance (MAE) | Pass | Mean Absolute Error of **$1758.80**, indicating the average prediction error. |
| Model Performance (RMSE) | Pass | Root Mean Squared Error of **$2607.36**, penalizing larger errors more. |
| Categorical Encoding Validation | Pass | Factor encoding successfully transformed categorical data for the R model. |
| Feature Scaling Validation | Pass | preProcess function from caret successfully centered and scaled numerical features. |
| Data Integrity (Missing Values) | Pass | Initial dataset confirmed to have no missing values. na.omit() was included as a safeguard but did not alter the dataset. |

**Deliverables Progress (2 pts)**

* **EDA and Insights:** Completed. Visualizations and key findings documented.
* **Feature Engineering:** Completed. New features integrated into the model.
* **Predictive Model:** Completed. Trained and evaluated. Model artifacts saved (.rds files).
* **Codebase (main.R):** Completed. Script covers the end-to-end process in R.
* **User Interface (UI):** Pending. This is the next major development task.
* **Final Report & Presentation:** In Progress.