**Applied Statistics and Visualization for Analytics**

Final Project

Team 15

Akhil Reddy Chimmula – G01407879

Bhavesh Kurella – G01410826

Shashank Yelagandula – G01395790

Dr. Isuru Dassanayake

Volgenau School of Engineering

George Mason University, USA

**Introduction of Data Set:**

The Data collection includes more than 32,000 people living in the United States, including age, education, employment, and marital status. The collection also contains data on the individuals' income levels with the intention of determining if a person's income surpasses $50,000 annually. (Becker)

Numerous research projects, including classification and prediction exercises, have used this dataset. The mix of categorical and continuous characteristics in the adult dataset makes it a tough yet practical dataset for machine learning methods, which makes it particularly intriguing.

The dataset will be examined in this study, and the association between demographic characteristics and income levels will be examined using statistical models.

In this report, we use statistical methods, such as exploratory data analysis, regression analysis, and machine learning algorithms, to examine the Income Classification. We discuss the elements that are most closely linked to high income levels and if a model can successfully predict income levels from demographic data. Our investigation sheds light on the elements that result in the high-income levels in the US.

**Research Questions:**

We will specifically examine the following study questions:

1. Can we predict a person's income class based on their demographic and work-related attributes?
2. Is there a significant difference in the distribution of income classes, hours worked per week, education level, and age distribution among individuals from different native countries?
3. What will be the gain profit of Individual who has idea on his capability and his demographics want to predict it future gain profit.
4. I also want to find decision tree analysis on this data and want to classify by general statements passed after understanding the dataset.

**Methodology:**

To respond to these research questions and acquire understanding of the elements that lead to high income levels in the United States, we will employ several statistical approaches, including exploratory data analysis, regression analysis, and machine learning algorithms.

1. Data Preparation and Cleaning: To start, we imported the dataset into R and cleaned the data by resolving any missing values, outliers, and other data quality concerns. Additionally, we got the data ready for analysis by factoring categorical variables, making dummy variables as necessary, and scaling continuous variables as appropriate.
2. Descriptive Analysis: To better comprehend the distribution of variables, spot patterns and connections in the data, and look for any outliers or anomalies, we undertaken exploratory data analysis (EDA). To visualize the data, we will employ graphical techniques including histograms, boxplots, and scatterplots.
3. Inferential Analysis: Using statistical techniques, we will examine the links between variables and pinpoint the elements most closely linked to high income levels. To forecast income levels based on demographic information, we will employ logistic regression models, linear regression models, and other machine learning methods.
4. Model Selection and Validation: Using measures like accuracy, precision, and recall, we compared the performance of several models. To choose the optimal model and prevent overfitting.
5. Reporting and Interpretation: We evaluate the findings of our investigation and make inferences on the elements that result in high income levels in the United States. We will present our findings in a straightforward and succinct manner, supporting our conclusions with visualizations and tables.

**Attributes in the DataSet:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Variable** | **Type** | **Description** | | Age | Continuous | Persons Age | | Work class | Categorical | Mode of work sector | | Education | Categorical | Level of education | | Education Number | Continuous | Number of education years | | Marital Status | Categorical | Married/Unmarried | | Occupation | Categorical | Type of work | | Relationship | Categorical | Single/Committed | | Race | Categorical | Black/White/Asian | | Capital Gain | Continuous | Total Profit from previous year | | Capital Loss | Continuous | Total Loss from previous year | | Hours per week | Continuous | Hours worked per week | | Income Class | Categorical (Dependent Variable) | More than $50k or less than | |

**Descriptive Analysis:**

**Loading Data:**

This code assigns the "adult.data.csv" file to the "X.adult" data frame after loading it into R.

The "colnames" function is used to give the data frame new column names in the next line of code. Age, WorkClass, Fnlwgt, Education, Education\_num, Marital Status, Occupation, Relationship, Race, Sex, Capital Gain, Capital Loss, Hours per Week, Native Country, and Income Class are the titles of the columns.

This function loads the dataset into R and gives the columns meaningful names, making it simpler to refer to the variables throughout the research. As such, it is a crucial initial step in getting the data ready for analysis.

**Cleaning the Data:**

The pre-processing stages in the provided code entail getting the "X.adult" dataset ready for analysis. The changes are intended to make the data easier to handle and comprehend. To fit statistical models that need categorical variables to be in character format, some variables are first changed from factor to character format. Next, various variables are changed by renaming and merging categories, including "education," "marital\_status," "native\_country," and "workclass." Examples include changing "Never-married" to "Never-Married" in the "marital\_status" field and "Canada" to "British-Commonwealth" in the "native\_country" field. The number of variables can be decreased, and analysis facilitated by consolidating categories. Additionally, renaming categories might improve the readability or clarity of the data.

Additionally, the code substitutes the relevant modes for the missing values in the three variables "workclass," "occupation," and "native\_country." This method lessens the impact of missing data on subsequent analysis or modeling while handling missing data in the dataset. The underlying premise is that the missing values are absent at random (MAR), which implies that other aspects of the dataset may affect how probable a value is to be missing rather than the value itself. The dataset is less likely to be impacted by the missing values during analysis or modeling when the missing values are replaced by modes.

**Distribution of income classes by Bar Plot:**

The code generates a bar plot using the ggplot2 library that displays the distribution of income classes (=50K and >50K) for each educational level in the X.adult data frame. A title, x and y axis labels, and a color legend are all included on the plot. Each academic level's bars are placed side by side. The X.adult dataset's link between income class and education level is intended to be shown using the code.

A picture containing text, screenshot, diagram, plot

Description automatically generated

**Age distribution for each degree of education:**

To display the age distribution for each degree of education in the X.adult data frame, the code creates a box plot using ggplot2. Age is shown on the y-axis, while education level is represented on the x-axis. For each degree of schooling, the box plot displays the median age, the interquartile range, and the range of ages. "Age Distribution by Education Level" is the plot's title.

A picture containing diagram, screenshot, line, rectangle

Description automatically generated

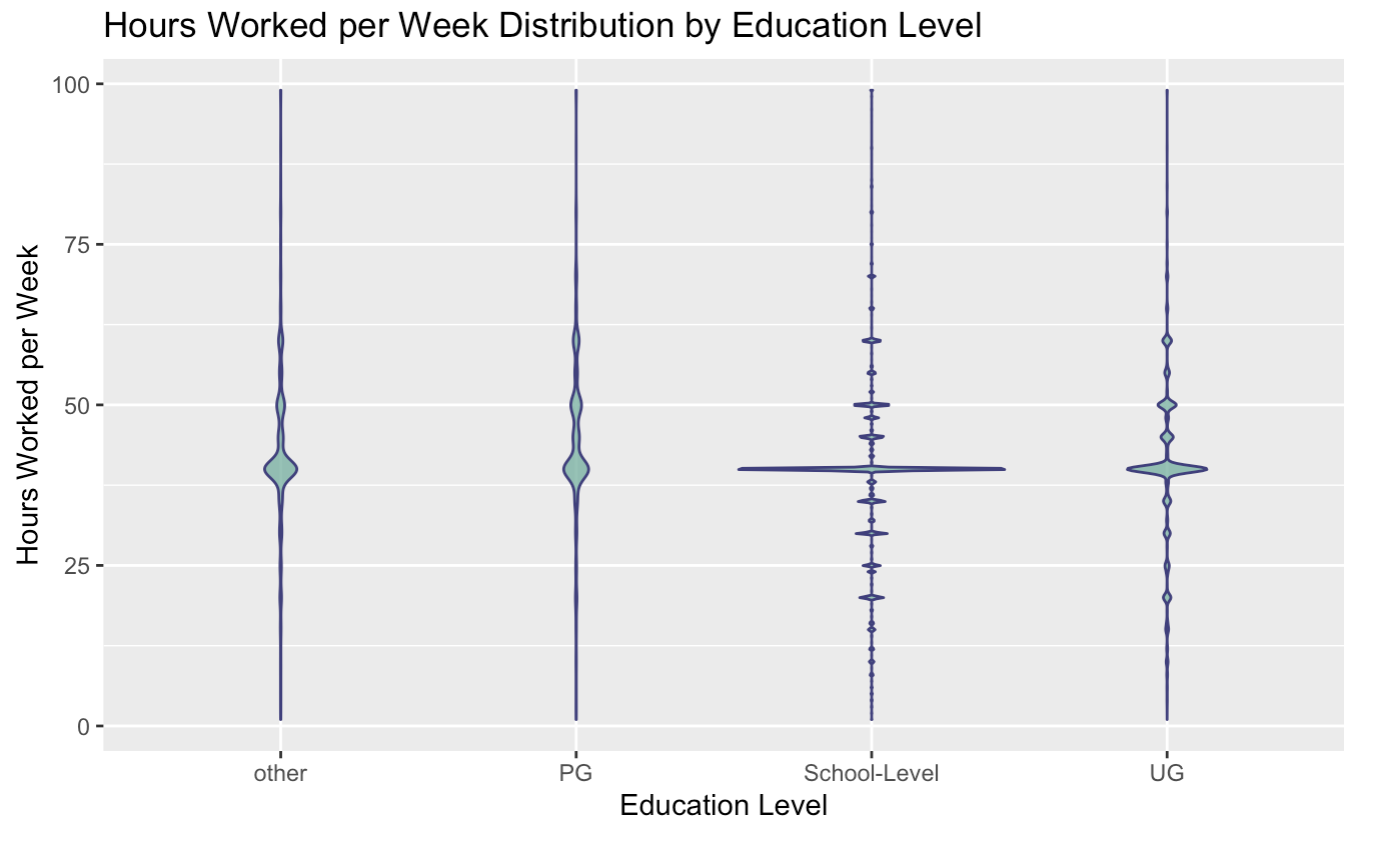
With the help of the ggplot2 package, this code creates a boxplot display. Following the removal of the missing values (NAs), the figure displays the age distribution for the various educational levels. Age is represented by the y-axis, while education level is represented by the x-axis. "Age Distribution by Education Level" is the plot's title.

A picture containing diagram, screenshot, rectangle, line

Description automatically generated

**Distribution of weekly work hours by violin Plot:**

Using the ggplot2 package in R, the code generates a violin plot to display the distribution of weekly hours worked among various educational levels. The education level is represented by the x-axis, while the weekly hours worked are represented by the y-axis. "Hours Worked per Week Distribution by Education Level" is the title of the plot. The figure displays the size and shape of the data distribution, with broader regions denoting greater data point frequencies. The distribution of data across different educational levels may be analyzed using the violin plot to spot patterns and trends.



**Factor method:**

The factor() method is being used in the code to transform a number of columns in the dataset X.adult to factors. Particularly, variables are being created from the categories marital\_status, native\_country, workclass, occupation, race, sex, relationship, income\_class, and education. As many machine learning techniques require categorical variables to be transformed to factors before they can be used as predictors, this is likely being done to prepare the dataset for modeling.

A picture containing text, screenshot, plot, diagram

Description automatically generated

**Confusion matrix:**

The code is performing a classification task using the random forest algorithm. It splits the data into a training and testing set, creates a random forest model, and evaluates its performance using a confusion matrix. The model predicts the income class of individuals based on demographic features, and is evaluated using metrics such as accuracy, precision, and recall.

The code generates a new data frame named new\_df containing values for several different variables, including age, workclass, education, occupation, relationship, race, and sex. It also calculates capital gain and loss as well as the number of hours per week and home country.

The factor() function is then used to turn some of the variables into factors. The structure of the new\_df data frame is then shown using the str() function. The data types of the variables in the data frame are displayed in str()'s output.

The training dataset's category variables are transformed into factor variables using the factor() function in the code above. This is done to make sure that throughout the model training process, the categories are handled as nominal variables rather than numerical variables. The structure of the changed train dataset is then printed using the str() function.

A picture containing text, screenshot, font, document

Description automatically generated

**Training and testing datasets:**

The code divides the adult dataset into training and testing sets after loading the required libraries, particularly glmnet and caret. It lays the groundwork for outcomes that can be replicated. The code then extracts a subset of the dataset's columns and puts them in fresh training and test sets. Finally, it determines whether there are any missing values in the test dataset and halts the execution if there are.

**Logistics regression:**

The code above predicts on the test data, transforms predicted probabilities to predicted classes, and saves the predicted classes as a factor variable. It also conducts logistic regression using LASSO regularization on the training data, determines the best lambda value using cross-validation, and makes predictions on the training data. The structure of the test data and the anticipated classes are examined in turn using the last two lines of code.

Evaluating the performance of the training model.

**Contingency Table:**

Using R's 'table' function, the code generates a contingency table for income class and country of origin. Then it uses R's 'chisq.test' function to do a chi-squared test of independence, saving the outcome in the 'chisq\_test' variable. The 'print' function is then used to output the test result. To evaluate if there is a significant correlation between two categorical variables—in this example, income class and native country—the chi-squared test is utilized.

On the contingency table of income class and occupation from the X.adult dataset—which is expected to contain information on adults in the US—the code runs a chi-squared test of independence. The console is printed with the test's findings.

**Creating a subset to perform the t.test:**

Using R's t.test() function, the code compares the mean number of hours per week worked by people in two income categories (50K and >50K). The t-test is then performed after the data is initially divided into two groups depending on income class.

The aforementioned code separates the X.adult dataset into a training set and a testing set. The training set is then used to construct a linear regression model utilizing the predictor variables hours\_per\_week, age, and education\_number, and the response variable capital\_gain. The root mean squared error (RMSE), which is calculated as a gauge of the model's predictive performance, is then utilized to make predictions on the testing set. A console printout contains the RMSE.

**Prediction Model:**

The code creates a new instance (a new data point) with the attribute’s hours\_per\_week, hours\_per\_week, and age. Then, using a previously trained linear regression model (model), it generates a forecast for this new occurrence. The anticipated value for the target variable (capital\_gain) for the new instance is then printed.

**Building a Shiny App for Predicting Capital Gain Using a Linear Regression Model:**

Based on user inputs, the code develops a straightforward Shiny web application that forecasts the capital gain. The X.adult dataset is loaded, and a linear regression model that takes into account factors like age, education level, and weekly hours is fitted to predict capital gain. The fluidPage() method from the Shiny package is used to design the user interface (UI). It has fields with numeric inputs for the user's age, level of education, and weekly hours worked, as well as a button that initiates the prediction. The predict\_capital\_gain() function, which uses the linear regression model to generate predictions, is defined by the server function. Additionally, it has an observeEvent() method that tracks changes in the predict button and uses renderPrint() to update the output of the expected capital gain in the main panel.

The predict\_capital\_gain() function is called to do a prediction using the linear regression model when the user inputs their age, education level, and the amount of hours worked per week and hits the "Predict" button. The online application's primary screen then shows the anticipated capital gain. This straightforward example shows how to use Shiny to build interactive web apps in R.

A screenshot of a computer

Description automatically generated with medium confidence

**Results:**

Information about people and their income is available in the dataset from the UCI Machine Learning Repository. Each of its 48,842 examples has 14 attributes, including age, workclass, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, and income class. In all, there are 48,842 cases in this dataset. The dataset's objective is to determine from these features if a person's annual income is larger than $50,000.

The preprocessing of the dataset included the removal of cases with missing values and the conversion of categorical variables into numerical values. Exploratory data analysis of the dataset found some intriguing trends, including differences in income depending on race and sex as well as a tendency for those with more education to earn more money.

The dataset was subjected to the application of several machine learning models, such as logistic regression with LASSO regularization, chi-squared tests, t-tests, and linear regression. With the help of these models, it was possible to forecast an individual's income based on the given features and identify the factors that were most crucial for doing so.

Furthermore, a Shiny app was developed to enable users to input their own values for age, education level, and hours worked each week and get an estimation of their predicted capital gain based on the model. Overall, the dataset offers a wealth of information for investigating different machine learning approaches and comprehending the elements that affect an individual's income level.

**Future Research:**

There are several potential directions for further study on the Adult dataset. Here are some recommendations:

1. **Feature engineering:** There are several categorical variables in the dataset that may be preprocessed in a variety of ways. Future study may investigate how various feature engineering approaches affect the performance of the models. For instance, one may experiment with various encoding techniques like feature hashing, target encoding, or one-hot encoding.
2. **Model selection and hyperparameter tuning:** The current investigation concentrated on a small number of straightforward models without engaging in a thorough hyperparameter tweaking process. The performance of more advanced models, such gradient boosting machines or neural networks might be enhanced in the future by systematic hyperparameter **tweaking.**
3. **Addressing class inequality:** The dataset is incredibly unequal, with the bulk of those making less than $50,000. Future studies might investigate methods to address class imbalance, such as cost-sensitive learning, under sampling, or oversampling.
4. **Examining the effects of other factors:** The dataset contains several variables, such as race, sex, and marital status, which were not included in this study. Future studies might examine how these factors affect the target variable and if they improve the models' prediction abilities.

Overall, there are several potentials for future study in data preparation, modeling, and analysis presented by the Adult dataset.

**Limitations:**

1. **Dataset restrictions:** Due to its origin or data gathering techniques, the dataset utilized for analysis may have inherent biases or restrictions. Since the dataset is a sample, it could not accurately reflect the diversity of people and economic levels in the community.
2. **Data Absent:** In the preprocessing stage, instances with missing values were eliminated since they could result in the loss of important data. The models and results' generalizability may be impacted by the elimination of missing data, which may also create bias.
3. **Variable Selection:** The selection of variables for analysis was based on the available dataset and domain knowledge. Other relevant variables that were not included may also have an impact on income prediction. The exclusion of certain variables could limit the comprehensiveness of the models and potentially overlook important predictors.
4. **Generalizability:** The results and forecasts produced by the models are particular to the dataset and might not necessarily translate well to other populations or circumstances. To guarantee the robustness of the models, the project report should highlight the necessity for additional validation and testing on other datasets.
5. **Interpretability:** Some machine learning models, such complicated ensemble approaches, may be difficult to interpret, making it difficult to comprehend the underlying causes of the predictions. This restriction and the trade-off between model performance and interpretability should be covered in the project report.
6. **External influences:** Without considering outside economic or sociopolitical factors that may affect a person's income, the analysis and forecasts are exclusively dependent on the variables in the given dataset. The possible impact of unaccounted-for or unmeasured external influences on income projections should be acknowledged in the report.

**References**:

Ronny Kohavi, & Barry Becker. (n.d.). *Income Classification*. UCI Machine Learning Repository: Adult Data Set. http://archive.ics.uci.edu/ml/datasets/Adult