CSE 578 - Data Visualization
Project milestone 1
Course Project Progress Report

Kedar Sai Nadh Reddy Kanchi
School of Computing and Augmented Intelligence
Arizona State University, Tempe
kkanchi@asu.edu

Problem Statement

UVW College, a local college is looking to bolster enrollment and for this it has chosen salary as a key demographic to determine criteria for marketing its degree programs. Therefore, for UVW College to get it's results, there are three things to be done:

- To develop marketing profiles of individuals using data supplied by the United States Census Bureau, with a focus on $50,000 as a key number for salary.
- To identify the factors that determine the individual's income.
  - There are many key variables that must be assessed for individuals making less than and more than $50,000, including age, gender, education status, marital status, occupation, etc.
- To develop an application to predict the income of an individual based on different values of the input parameters so that UVW College can tailor their marketing efforts when reaching out to the individuals.

Progress So Far

- I have procured the data from the United States Census Bureau links provided in the project description file and loaded into the pandas data-frame.
- I have conducted an initial analysis of the data to comprehend its structure.
- I have executed data quality checks to identify and address missing or incorrect values in the data by removing all the rows that contained "?" in any of the columns. This helps to guarantee the accuracy of the model that will be developed from the data.
- I then conducted individual feature analysis on all the features against the class variable to comprehend distributions, and relationships between them.
- From the conducted uni-variate analysis, I identified the most relevant features that affect the class variable.
- Finally, I have made a plan on how to conduct the analysis between the above identified relevant features to better understand to understand the relationship between themselves and with the class variable in combination before moving on build a machine learning model to predict an individual's income based on these identified features.

Background

- The first thing I did was to download the data set from the United States Census Bureau, that is required for this project.
  - The data-set includes many key factors such as age, gender, education status, marital status, occupation, etc that can be utilized in predicting an individual's income.
- Now, after downloading the data sets, I have conducted an initial analysis of the data to comprehend its structure.
  - There are a total of 32,561 rows.
  - The last column is the result column, Class - which has two classes/categories - above 50k (">50k") and below 50k ("<=50k")
    - Out of these 32,561 individuals in the data-set, 24,720 earn a salary less than $50,000 while the remaining 7,841 individuals earn a salary greater than $50,000.
  - Apart from the resultant column at the end, the data set has a total of 14 features. And out of these 14 features, 8 features are having categorical data.
- Now, coming to the data in the data-set, the data is in the form of a comma-separated values. Also, there are a total of 4,262 missing values which is represented by "?" in the data.
- The data-set is loaded into a pandas data-frame.
  - This is because Pandas, enables quick and easy analysis of the data at hand.
  - Pandas also has various in-built pre-processing techniques at hand for cleaning the data-set so that visualisations and models can be built.

As a result of the background work accomplished thus far, I have a comprehensive understanding of the data and its structure. I am now ready to proceed with the next steps of data Pre-Processing/Cleaning and model selection.

Summary of the Tasks Completed Till Date

Loading Data

- The data-set comprising 32,561 data points were loaded into a pandas data-frame.
- A header row was added to the data-frame to contain feature names. These features are as follows: age, workclass,

fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, and class.

Data Pre-Processing/Cleaning

- I executed the command, "df.isna().sum().sum()", to check if there are other missing values other than the ones represented by "?" in the data.
  - But the command "df.isna().sum().sum()" resulted in a value of 0, which means that there are no other missing values in the whole data-set apart from the ones represented by "?".
- So, now the following step for me was to handle the missing values in the data-set represented by "?".
  - There were 4,262 instances of values represented by "?" out of the total 32,561 rows which constitutes to approximately 13% of the complete data.
  - Since, the rows containing the missing values, constitutes to a very less percentage, I have decided to remove these 4,262 rows from the data-set.
  - Therefore, for each column, I extracted all the rows that did not contain "?".
  - After removing the rows, that contained "?" values, the total number of rows in the cleaned data-set are 30,162. So, this means that effectively only 2,399 rows have been removed as opposed to 4,262 rows. So, the number of removed only constitute to 7% of the initial data-set.

Individual Feature Analysis

Now, I processed to perform analysis, uni-variate analysis on each of the continuous and categorical columns to understand their distributions and relationships with the class variable. This analysis helps to pinpoint any potential issues with the data, such as outliers. This process helps to guarantee the accuracy of the model that will be developed from the data.

To carry out this uni-variate analysis, each feature was analyzed individually using data exploration techniques to identify patterns. Data visualization tools such as pie charts, bar charts, histograms, and box plots were used to accomplish this. To better comprehend the underlying distribution, other statistical techniques like mean, median, and standard deviation were applied when necessary.

As a result of the Individual Feature Analysis, I have reached the preliminary conclusion that the most important features are age, education-number, marital-status, occupation, and capital-gain.

In addition to the five features mentioned above, less significant or redundant features, I believe capital-loss, fnlwgt, and education are less significant or redundant features.

- For example, the education feature was discovered to be redundant in other features such as education-num, because education-num represents the number of years of completed education while education represents the highest education completed in categorical form.

Therefore, such redundant features might not required for building the marketing profiles in the future alongside the five mentioned above.

Issues Encountered and Resolutions

- **Issue-1:** The skewness of the data may make generating precise marketing profiles difficult. This is because 76% of the population earns less than $50000 per year. As a result, the question of whether the entire data-set should be included when creating marketing profiles arises.
  - **Solution:** The most effective way to address this issue is to sample the data so that there is an equal number of people earning less than and more than $50,000.
- **Issue-2:** The data-set also revealed gender bias. This is because 67% of the people in the data-set are men and 33% are women. This bias may make marketing UVW College to female audiences difficult.
  - **Solution:** Again, this can be solved by selecting an equal number of men and women who fall into the less than and greater than $50,000 categories, allowing both genders to be effectively compared in terms of salary while also taking other important factors such as education-number and marital-status into account.

Tasks Yet to be Completed

- Individual plots were created to provide an understanding of how the data is distributed and how much it affects the salary; however, it would be a better approach to go over the pros and cons of various types of plots to see which data visualizations would be more suitable to model the salary. In order for this to happen:
  - We must conduct multivariate analysis with improved plots based on the five identified features that influence the class/salary variable.
- Following the identification of the important features, additional data pre-processing is required to remove redundant feature columns and convert categorical variables to numerical variables for machine learning computation. This is because, given the remaining characteristics, a suitable machine learning model must be built to predict the salary.
- Then, given the features, a machine learning model must be implemented to predict the salary of any individual.
  - ∗ I'll compare several machine learning models to see which one produces the best results when it comes to predicting an individual's income based on data using different sci-kit-learn libraries in Python, along with multiple classification models, including SVM, Logistic Regression, SGDC, Gaussian Naive Bayes, Decision Tree, and Ensemble Methods (Random Forest).
  - ∗ The data will be pre-processed and divided into training and testing sets to begin. The models will then be trained using training data and tested using testing data. The accuracy, precision, and recall of each model will be measured, and the model that produces the best results will be chosen for further investigation.