Term Project Milestone 3 - Census Income Prediction

Team Name: Analytics

Team Members: Kesav Adithya Venkidusamy, Meenakshi Shankara, Sherry Kosmicki

Bellevue University

DSC630-T301 Predictive Analytics (2227-1)

 Professor Fadi Alsaleem

07/17/2022

**Census Income Prediction**

**Milestone 2 - Data Selection and Project Proposal (Week 2)**

### **Information about Dataset**

**Name**: Census Income

**About this dataset:**

This data was extracted from the [1994 Census bureau database](http://www.census.gov/en.html) by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0))

**Abstract:** Predict whether income exceeds $50K/year based on census data

|  |  |
| --- | --- |
| **Data Set Characteristics** | Multivariate |
| **Attribute Characteristics** | Categorical, Integer |
| **Associated Tasks** | Classification |
| **Number of Instances** | 48842 |
| **Number of Attributes** | 14 |
| **Missing Values** | Yes |
| **Area** | Social |
| **Date Donated** | 5/1/1996 |

**Attribute Information:**

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Feature Type** |
| age | Age of the person | Continuous |
| workclass | Work class of the person | Discrete |
| fnlwgt | Final Weight | Continuous |
| education | Education of the person | Discrete |
| education-num | Number of years the person | Continuous |
| marital-status | Marital status of the person | Discrete |
| occupation | Occupation of the person | Discrete |
| relationship | Relationship of the person to the family | Discrete |
| race | Race of the person | Discrete |
| sex | Sex of the person | Discrete |
| capital-gain | Capital Gain | Continuous |
| capital-loss | Capital Loss | Continuous |
| hours-per-week | Hour the person worked for a week | Continuous |
| native-country | Native country | Discrete |
| Income | Income of the person | Target |

### **What types of model or models do you plan to use and why?**

A logistic regression model will be used on the dataset to determine which features are mostly related or correlated to our target which is “Income” of the person. Logistic regression is a statistical analysis method used to predict a binary outcome such as yes or no based on prior observation of the data set. Here, “Income” feature present in the dataset has only binary values: whether the income of the person is less than or equal to 50K per year or greater than 50K per year. So, this feature will be used as target for the model. This model falls under supervised learning as the data is well labelled and has a target variable, a column in the data representing values to predict from other columns in the data. Sometimes supervised leaning is called predictive modeling. Supervised learning allows collecting data and product data output from previous experience.

Under supervised learning, this dataset falls under classification model as it reads the input and generates an output that classifies the input into two categories: one having income less than or equal to 50K per year and another with income greater than 50K per year.

In addition to the logistic regression model with 1 target and 14 features, another logistic model will be used with only 5 best features where the 5 features are selected based on highest chi-squared statistics.

### **How do you plan to evaluate your results?**

We plan to calculate the accuracy, precision, recall and F1 score of both the logistic regression models with 14 features and best 5 features. We will also verify the performance of the model by visualizing confusion matrix which is a 2\*2 table that shows the predicted values from the model vs. the actual values from the test dataset. We will plot ROC curve to determining the best cutoff value for predicting whether a new observation is a "failure" (0) or a "success" (1).

The Area Under the ROC curve (AUC) metric is evaluated to see how well a logistic regression model classifies positive and negative outcomes at all cutoffs. The value can range from 0.5 to 1. The result is considered excellent if AUC value is between 0.9-1, good for the AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8 and poor for the AUC values 0.6-0.7 and failed for the AUC values between 0.5-0.6.

### **What do you hope to learn?**

Working as a group on this course project will help us to understand how to work with others in the same field of study. We can learn each other's strengths and highlight them in the process. For those that have areas they would like to improve, we can capitalize on learning from our teammates. Learning possibilities are available if we communicate our project progress.

During the review process of the income prediction dataset, we can determine if there are any statistical patterns or predictors based on visualizations. The modeling process can help us define the best model to use with our data based on accuracy, precision, and recall. Once we have trained our model new data can be applied to predict what income category a person is part of.

### **Access any risks with your proposal**

One of the earliest challenges we might face is during the data preparation step of the model building. Identifying the correct features that contribute to the target, planning on how to handle the missing values, deciding the next steps if the data is imbalanced to name a few. The way to mitigate these issues would be creating various visualizations to identify correlations. To mitigate data imbalance, we may choose to over-sample or under-sample the dataset. We may also need to go back to research other relevant supplement datasets to strengthen the cause.

During the model building phase, we may face the challenge of finding the right models for our project in terms of accuracy. As a mitigation plan, we will identify at least 3 models to train the data and calculate the accuracy with multiple methods.

We will also update/upgrade/change the course of our project based on the feedback received during peer reviews.

### **Identify a contingency plan if your original project plan does not work out.**

If for any unforeseen reasons, we are unable to continue working on this dataset, we have considered a backup dataset and we will perform a high-level analysis in parallel. This will help us quickly shift to the new data set without too much loss of time.

Backup Dataset - <https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019>

### **Include anything else you believe is important.**

**Why income prediction is important**

Income prediction is important for a variety of areas in the private and nonprofit sectors. One critical area this affects is marketing, where income segmentation of the population is an extremely important tool. Businesses may make different variations of their items designated for certain subgroups of the population, and these subgroups often include the income of individuals. Income prediction also helps to identify those individuals who are of a lower income that may need the most assistance, who some nonprofits strive to identify and assist. The ability to predict the income of individuals from this information has far-reaching impacts for every industry.

**Data consideration for logistic regression**

Following are some of the important points to be considered while choosing data set for logistic regression.

* The response variable should be binary
* The features present in the dataset should be independent to one another
* Make sure the data represent the population of interest
* Collect enough data to provide necessary precision
* Measure variables as accurately and precisely as possible
* If the model does not fit the data, then the results can be misleading. In the output, use residual plots, diagnostic statistics for unusual observations, and model summary statistics to determine how well the model fits the data

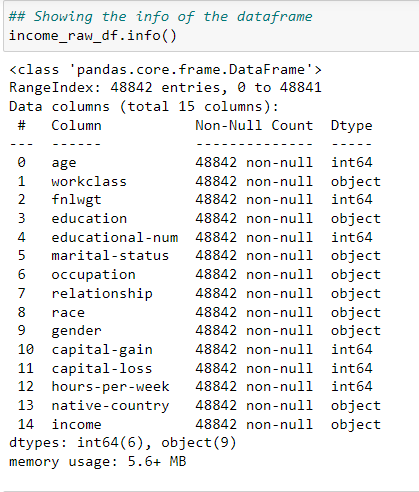
**Milestone 3 - Data Selection and Project Proposal (Week 6)**

### **Will I be able to answer the questions I want to answer with the data I have?**

The problem statement of this project is to identify the dataset feature(s) which are mostly related to or affecting the income of household. With dataset having total number of records as 48842, we would be able to predict or answer our problem statement. The dataset consists of 15 features of which 6 are numerical and rest all are categorical with “income” being the target. The target variable income contains 2 values <=50K and >50 which would be subsequently converted to 0 and 1 respectively. The details are shown in figure 1.

Among 14 features, we see missing/null values present only for below features. I have given the percentage of missing values for each of the feature in table 1. We noticed that the values where ‘workclass’ is missing, also has ‘occupation’ missing. While trying to identify the extra rows where ‘occupation’ is missing, we observed the workclass is ‘Never-Worked’. Since the percentage of null values present in these features is low, the rows will be removed from the dataset.

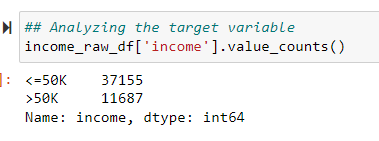
|  |  |  |
| --- | --- | --- |
| **Feature Name** | **# Of missing Values** | **Percentage** |
| workclass | 2799 | 5.7% |
| occupation | 2809 | 5.8% |
| native-country | 857 | 1.8% |



**Table 1: Features with null values and percentage**

**Figure 1: Features and dtypes**

The income column is our target variable with 2 values - ‘<=50K’ and ‘>50K’. The count of these values is 37155 and 11687 respectively, suggesting that people with income higher than 50K are significantly less, and our data set is kind of imbalanced considering the target variable. However, we will evaluate the outcome and apply filter to the dataset, if required.



### **What visualizations are especially useful for explaining my data?**

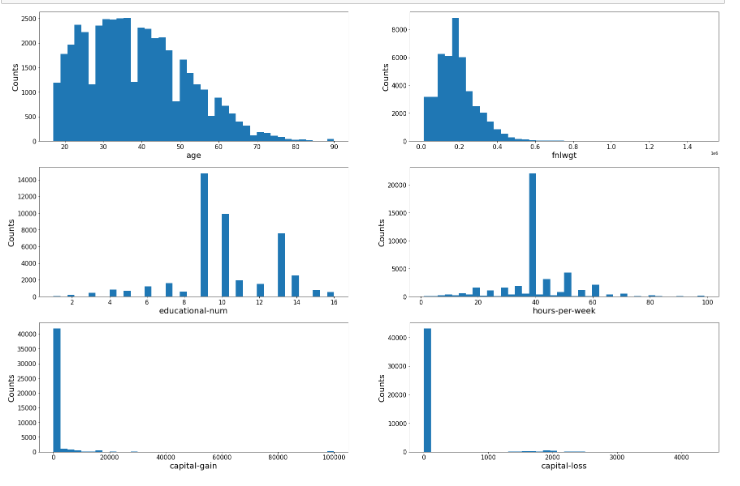
As mentioned before, the dataset contains 6 numerical features and 8 categorical features as follows, and ‘income’ feature being the target variable.

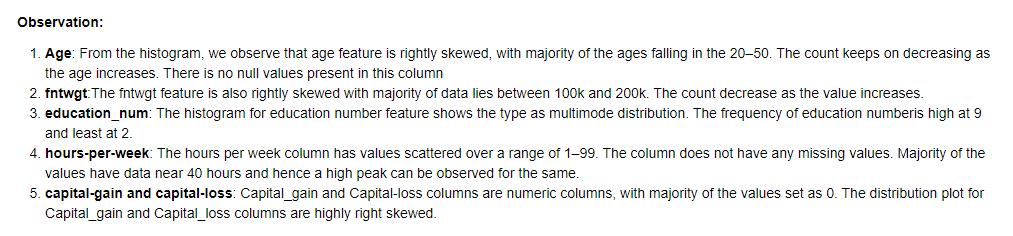
Numerical: age, fnlwgt, educational-num, hours-per-week, capital-gain, capital-loss

Categorical: workclass, education, marital-status, occupation, native-county, relationship, race, gender

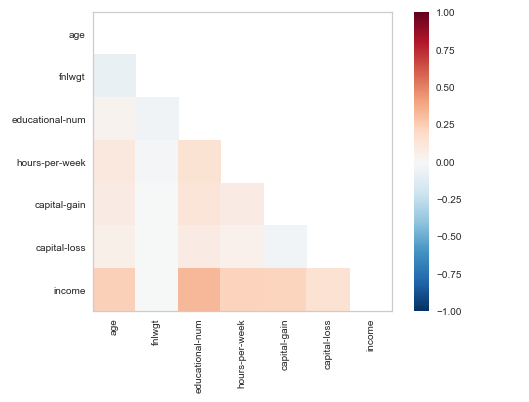
So, we have used several visualization charts to analyze the useful information present in the data. The following are the visualizations used based on nature of the feature.

***Histogram****:* Histogram is used to identify the distribution of numerical features present in the dataset.

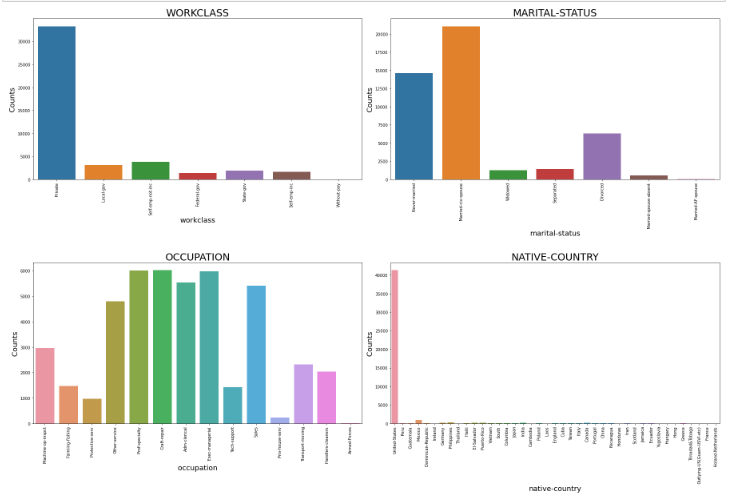


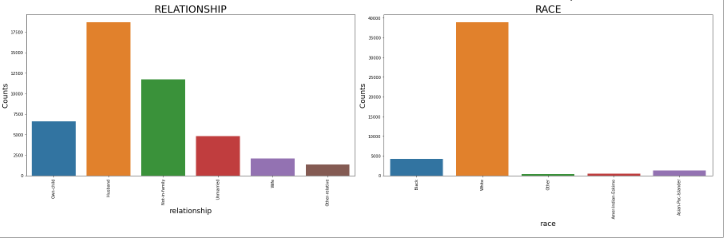


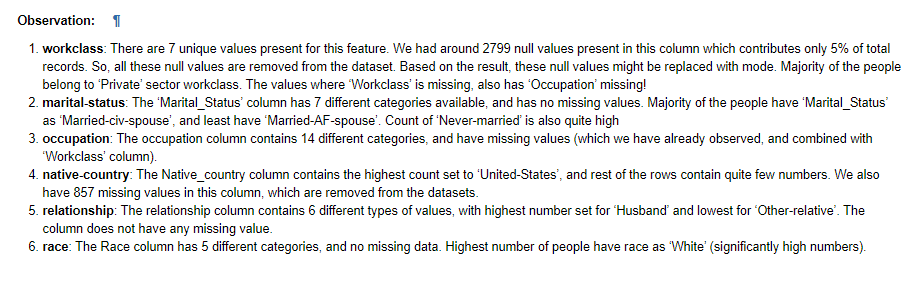
***Heat Map****:* Heat map has been created to understand Pearson’s correlation between the target variable ‘income’ and other numerical variables. We observed that all numerical features have positive correlation with the target except fnlwgt feature. Among the features having positive correlation, age and education-num features are having high value. The details are shown in the below heat map chart.

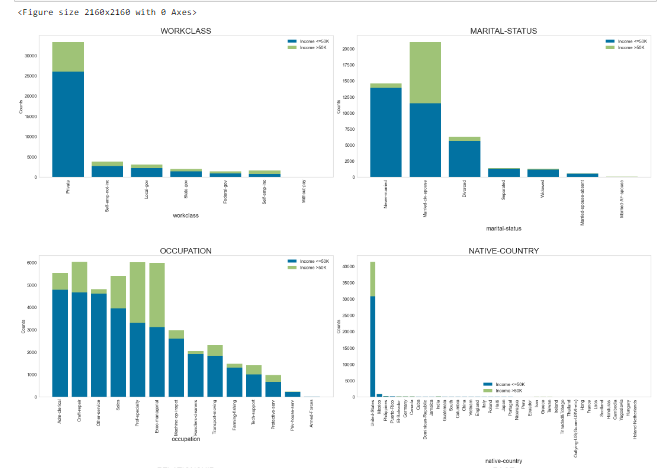


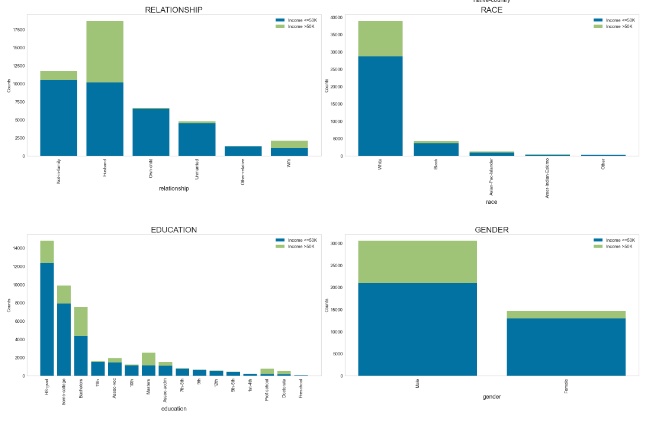
***Bar and Stacked Bar Graph:*** Bar graph has been plotted for all categorical features to understand the distribution of data among unique values. Stacked bar chart has been plotted to compared those earning less than or equal to 50K (represented as 0) and greater than 50K (represented as 1) for all the categorical features.

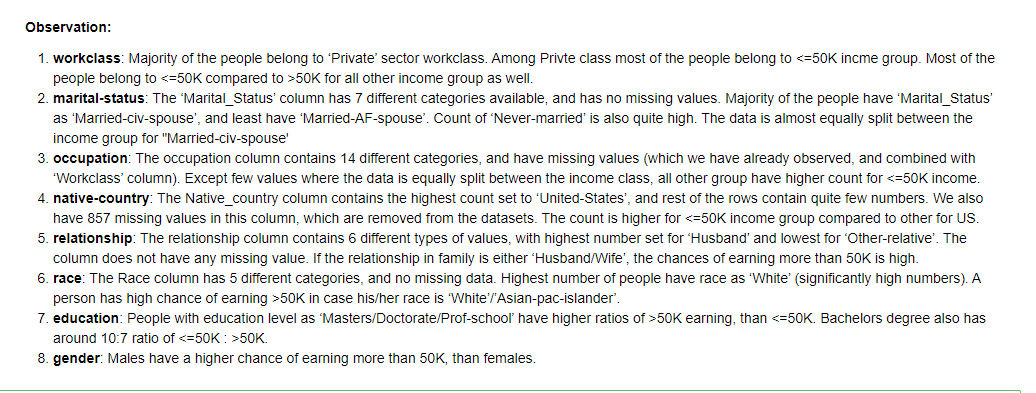












### **Do I need to adjust my model/evaluation choices?**

In Milestone 2 it was determined we would use logistic regression on the dataset. This analysis method is specific to a binary target variable which is our income determination amount. Once we’ve completed this step, calculating accuracy, precision, recall and F1 scores should help determine quantify model performance. The ROC curve metric will also be used to evaluate the logistic regression model where we hope to see values between 0.9-1.

The group has discussed creating a decision tree model to evaluate the data in addition to logistic regression. Based on categorical variables such as workclass and marital-status in this dataset, we can take them and group the values into a broad summarization of the values and use them for evaluating the decision tree.

The dataset will work well with this evaluation since a decision tree can handle both numerical and categorical values. The other benefits of using a decision tree are it works with the outliers and the missing values currently present in our dataset. We will have to determine the best way to split the variables into subgroups. This will help build efficiencies in the evaluation done by the decision tree. Additional data gathered for the purpose of continuing the evaluation of a person's potential income could strengthen the data model over time. We will then determine the accuracy of the decision tree and avoid an opportunity for overfitting.

### **Are my original expectations still reasonable?**

The original expectations are to find accuracy in the model building process to predict the income of an individual based on the gathered variables. The predictive data within the dataset has value and is functional based on the visual evaluations and data munging steps during the EDA process. If the accuracy of our models does not meet our needs, we may consider other model options such as Naïve Bayes classifier.

We will continue to meet once or twice weekly as a team to continue our discussion of the project. Each of us has a personal stake in the project. We will gather information and define the best possible steps to complete the project using the CRISP-DM process model.

### **Reference:**

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. John Wiley & Sons, Inc.

<https://www.census.gov/en.html>

<https://www.kaggle.com/datasets/uciml/adult-census-income>

<https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression>

<https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/before-you-start/data-considerations/>