CSE 578 Course Project Progress Summary

By Kexin Kathy Tong

Problem statement

The problem I am trying to solve is to determine a demographic profile for UVW to market its degree programs based on criteria of targeting individuals making more and less than $50k in salary. In this report, I will identify important factors that contribute to a person's income, and build a prediction model to predict whether the income of a person is >$50k or <$50k given a sample profile.

Progress made so far

I set up my Jupyter Notebook environment and downloaded all the required libraries and the dataset. I have uploaded the data into my notebook, and cleaned the dataframe for further analysis. I have created functions to build out single variable and multivariable analysis for both categorical and numerical data, and summarized some findings below that answers these key questions.

Specifically, the user stories I want to identify are:

- Is age a reliable indicator in determining income?
- Is education a reliable indicator in determining income?
- Do occupation, race, sex and marital status help us determine income?
- Is there any meaningful relationship between age, education and salary?
- Are there any meaningful relationships between age, capital gain/loss and salary?

Specific tasks I have completed to date

I started off importing the data into a dataframe, adding the column names, and cleaning the data of any rows with '?' variables by removing these rows from the dataframe. I then converted the string columns into categories so they can eventually be mapped into numbers for the prediction models in future steps. Now our cleaned data frame has 22k rows with >$50k income and 7k rows with <$50k income.

I built a function to summarize numeric variables by looking at its breakdown of the number of variables with >$50k and <$50k income and finding out the min, max, average, standard deviation of total and each income category. I also added a histogram plot to see if the distribution of salary varies greatly for the particular variable, and added a box plot to see what the median and outliers look like. I ran this single variate analysis function through all the numeric variables and found that education, marital status, occupation and relationship may be important factors in determining incomes.

Next, I built a function to evaluate categorical variables by summing the number of datapoints for each category, along with the breakdown into >$50k income and <$50k income. I also added a pie graph to visualize total data points per category, a histogram for category count by income and a mosaic plot to see

the breakdown of each category by income. I ran this single variate analysis function through all the categorical variables and found that age. capital_gains and education_num are important factors in determining incomes, while the others do not look to be a meaningful factor in determining income.

Lastly, I built a function for multivariate analysis between two or more numeric variables by plotting the scatter matrix of each variable against each other, colored by income. If there are any two variables with two distinct clusters of the two income levels this could identify a strong candidate profile of income level.

Issues faced so far and plan on how to resolve them

One issue I faced is finding a suitable metric to determine whether a categorical variable is important for determining salary. After plotting pie graphs, mosaic plots, histograms, converting each category into a numerical number and running the numeric analysis function, I realized that histogram and mosaic plots are most useful because it is easy to tell whether breakdown of income differs from category to category, and thus I chose to use these plots in my visualization.

Tasks yet to complete and how I plan to solve them

My next task is to analyze the relationship between two or more categorical variables, and find a meaningful set of factors that determine salary. I will do this by building a function that takes in two or more factors to build a parallel coordinate plot after scaling each factor, and build a mosaic plot using two categorical factors to see if there are notable differences between income.

Then, I will add legends, x and y-axis labels, title and formatting for each plot that best answers the user stories I identified above and write my conclusion for each in the report. As an example, the kde histogram plot can easily show that age is an important factor in income because the distributions between >$50k and <$50k are very different.

Lastly, I want to build a prediction model that tests 2-3 machine learning models using the important features I identified above to find an algorithm that predicts the salary amount of the testing data. The algorithms I want to test will be a subset of decision tree, k-nearest neighbor, regression and random forest. I will use a 80-20 split for testing and training data to evaluate the performance of each model, using accuracy, precision and recall.