COURSE PROJECT PROGRESS REPORT

Vasanta Prayaga

PROBLEM STATEMENT

To develop an application to predict the income of an individual, based on the factors used in developing marketing profiles on people, using data provided by the United States Census Bureau. The focus is kept on $50,000 as the key number for salary.

PROGRESS MADE AND BACKGROUND WORK COMPLETED

- Imported the United States Census Bureau dataset that was provided in a Data frame into Jupyter Notebook
- The data set has 32,561 rows and 15 columns
- The data was in a comma separated format
- Some of the data missing was represented with '?'
- Assumptions were made about conversion of original data, final weights, similar demographic characteristics should have similar weights etc.
- There were no headers/column names for the data
- The names file had a description of the columns
- Started data cleaning by adding column names to the data frame to make it more meaningful
- As part of the data cleaning, also addressed '?' values by creating a new data frame without rows having '?'. They couldn't be imputed as it didn't make the data meaningful
- Calculated how many people had salary below 50K and how many had above 50K
- Started working on finding correlation between different attributes and salary
- Used this correlation to pick the factors that influence the salary the most and come up with user stories/ questions that need to be answered by the analysis
- These factors are then being used to develop the marketing profiles of people.
- Did research on which plots would give more insight into which factors influenced the salary the most
- Did research on how to find the correlation between the different factors
- Did research on univariate, multivariate and machine learning analysis needed to be implemented in the project

SUMMARY OF THE TASKS COMPLETED TO DATE

- Completed analyzing of what the dataset contains and came up with a problem statement with the objective for the project which is a prediction task to determine whether a person makes over 50K a year
- Cleaned the data by removing rows with missing data as they couldn't be imputed with average or any other without loosing the meaning to the data
- Added column headers to the data frame to make it more meaningful and easier to code with

- Using this clean data set, I started doing a univariate analysis to see the correlation between the different attributes given and the salary.
- Also, analyzed the relationship between the other factors and salary specifically for salary less than 50K and a separate one for more than 50K and found there were more earning less than 50K
- Listed the 5 user stories/ factors for the project, that affect the salary the most after exploring and analyzing the data
- Completed the analysis and plots using all the factors and got insights from the visualizations
- The factors age, education, marital status, occupation, and sex are some with strong influence on the salary
- Started working on multivariate analysis occupation- age and education-age

ISSUES ENCOUNTERED THUS FAR

- Figuring the best ways to analyze multivariate data and infer from the results
- Difficulty understanding the implementation of machine learning in this project and still researching

TASKS YET TO COMPLETE AND PLAN OF APPROACH

- Complete multivariate analysis with the other factors and see their influence on the salary
- Implementation of machine learning
- Try to get more information from research, during office hours, discussion forums and peers
- I will try finish the multivariate analysis by this weekend to have enough time to implement the machine learning to finish the application to predict the income of a person based on the different factors identified.
- Then finish the project by testing of making salary predictions based on the profile