Project Progress Report

1.Problem Statement

Data Mining and Machine Learning are becoming more useful to find the hidden patterns and correlations between different parameters in the dataset, which lead to the prediction of difficult to predictable occurrences. These patterns and correlations are crucial to visualizing a large dataset. The dataset used in this project was an adult income dataset, a common unbalanced machine learning dataset.
This project aims to create a profile for a college marketing team using a dataset containing demographic information. This profile will help the college to achieve its enrollment target. This report will find the correlations between the different demographic variables of a person. With these findings, an impactful and informative profile will be created to help recommend a primary demographic group for college enrollment.

2.Recent Developments Made

The project was started on Tuesday, May 27th, with some primary planning and resource collections. The primary work was the exploration of data given for the project. After careful observation and exploration of the data set, it was clear that some of the features of the data could be used for marketing aspects. After the initial observation of the dataset, several critical features were found which will be used in the project.

Fourteen variables were combined with ordinal, category, and numerical data types in the given dataset. There were some missing values in the data set, and question mark symbols (?) were used to indicate those missing values. The data set had 48842 rows with 3620 missing values and had 45222 blank rows.

The system used for this project needed Python to download and install Jupyter notebook along with Python matplotlib, seaborn libraries, and NumPy library for visualization. Some preliminary graphs were created to help break down each demographic category based on their salaries. The preliminary graphs were exploratory and were created at a demographic stage. A stacked bar chart was created for each sector to analogize the salary buckets (<=50k,>50k), which was the throughput of the only demographic values. From the preliminary findings, about 75% of the population was in the "<=50K" group, and about 25% of the population was in the ">50K" group.

3.Completed Tasks and Approaches

This project is currently at its very preliminary stage. The visualization, which is one of the significant parts of this project, has not been created yet. Nothing can make the data visualization harder at this point of the project, and the visualization will be created at the end of the project. Then it will be added to the final deliverables. Currently, the brainstorming part is taking the most attention, along with gathering the things necessary to complete this project. The following features were identified after the initial brainstorming and data analysis:

a. Gender
b. Race
c. Age
d. Marital Status
e. Education
f. Occupation
g. Hours of Work per week

These features will provide major insights for this project. The project was divided into two parts: a univariate and multivariate analysis. The univariate analysis visualizes the critical observation of the features found in the initial research, and the multivariate analysis visualization helps create the final profile for marketing. The projection of the works completed so far looks pretty good, and all the works for this project will be finished before the due date. Therefore, the project will undoubtedly meet the due date:

4.Challenges Faced

The data set provided for this project was filled with regular demographic data without separating the distinguishing features. Separating them and putting them under the different groups was a challenge initially. The data such as education, age, and gender are more common in the demographic data set, but salary buckets were not easy to compare since they were not common in the given data set. It was also challenging to create a visualization of several demographic features to compare using an extensive data set with several variables. When the data collection started from the given datasets, it was found that there was a combination of categorical and numerical features columns in the data. So, for visualization, the categorical and numerical data were separated, and different visualization was created for the different features. Then the crosstable function was used from the panda's data frame to make individual visualization regarding the income and gender and the income and race.

5.Summary of the Work to be done

According to the schedule of completion of the project, the completion of work is ahead of schedule. The amount of work done so far is good enough to accomplish the project goal. The next job is to explore the data set again to produce a list of all the visualizations which will be added to the final project report.

This week's target is to start writing the Python code based on the user stories identified. This Python code will then be used to visualize the various data to provide more insights into the data. The data visualization is well underway and will be added to the final project report.