**Module 1: K Nearest Neighbor (KNN Model)**

ALY6040 – Data Mining Applications, Northeastern University, Boston

Professor Justin Grosz

CRN: 20387

02/26/2023

Submitted by: Abhinav Jain

**Introduction**

The purpose of the study was to examine the relationship between numerous factors, including occupation, education, gender, and race, and the degree of prosperity among US people. The objective was to create a KNN classifier-based model that can distinguish between people with low and high earnings.

In order to create laws and programs that benefit the underprivileged and advance equality, it was critical to define financial success and provide equal possibilities for all residents. The study's dataset, which has 48,000 records and 15 variables, was employed. The data was cleaned before being subjected to an exploratory data analysis process utilizing several approaches, including correlation matrices, heat maps, bar charts, and boxplots. These methods were used to analyze the data and find any trends or connections between factors that could affect income levels.

The KNN model, which offers a deep knowledge of the commercial issues associated with unequal distribution, was finally adopted. It may be able to combat income disparity and advance equal opportunities for all citizens by understanding the elements that lead to financial success and using this information to design laws and initiatives. A report is a useful tool for scholars and policymakers trying to understand how different socioeconomic and personal characteristics affect income levels in the US.

**Data Cleaning**

Initially, the index column was initially absent from the dataset, according to the data library allocated names to the relevant columns, and removed the "fnlwgt" variable because it was irrelevant to the analysis. Many cells that included the question mark "?" in three characteristics that were categorical variables that couldn't be determined by the central tendency (mean, median, or mode) led to the decision to remove the records from the dataset. These cells were discovered throughout the analysis. They were the variables "Workclass," which has 2799 missing values, "Occupation," which has 2809 values, and "Native country," which has 857 categorical values.

Age, sex, salary, hours worked per week, and education was taken into account for the analysis, although sex and salary needed to be transformed into dummy variables (0 and 1 form). On the other hand, age, education numbers, and hours per week were also converted into 0 and 1 form with case statements such as age greater or equal to 75 converted to 0 and less than 75 converted to 1, education numbers less than 10 convert to and education number greater than or equal to 10 convert to 1, and hours per week converted on the working hours less than 40 to 0 and if working hours were more than or equal to 1.

**Exploratory Data Analysis (EDA)**

The data shows that there was a gender pay gap, with a higher proportion of men earning over $50,000 compared to women. Age, education number, and working hours per week have the strongest positive correlations with wages. The weakest correlation was found between capital gain and capital loss. The boxplot revealed that the 38-50 age range had the highest earners while the 26-47 age range had the lowest earners, indicating that education level was a factor in determining wages.

**Chart

Description automatically generated**In this bar graph, we can see that the maximum pay received by men is actually more than that of women, with 22% of men receiving more than $50,000 and 46% receiving less than $50,000. Just 5% of women earn more than $50,000, compared to 27% of males who earn less than $50,000.

Graphical user interface, application

Description automatically generatedThe correlation matrix also shows that age, the number of education credits, and the number of working hours per week have the strongest correlations, all of which are greater than 0.1. If we look at the lowest correlation, which was -0.032 between capital gain and capital loss.

**Chart, box and whisker chart

Description automatically generated**In the boxplot, the outlier was above the age of '77' having less than an equal to $50,000 wage, more than $50,000 was drawn by the 38–50 age range, while less than $50,000 was drawn between the age range of 26–47, demonstrating that those with greater education had higher compensation.

**Analysis Model Building - Nearest Neighbors (KNN)**

The dataset contains information about individuals, including their age, education number, working hours per week, and salary. The dataset needs to be cleaned and preprocessed before performing any analysis. After the cleaning process, the key performance indicators (KPIs) are identified, where age, education number, and working hours per week are selected as the independent variables, and salary is chosen as the dependent variable.

To evaluate the performance of a KNN model on this dataset, it is divided into two parts, the training set, and the test set, using an 80/20 ratio. The training set was used to build the model, and the test set was used to evaluate its performance. The variables were then transformed into binary form, where each variable was either 0 or 1, to facilitate the KNN implementation.

Finally, the KNN model was trained using the training dataset, and its performance was evaluated on the test dataset with an accuracy of 36%.

Therefore, k = 212 is used to make predictions using the KNN algorithm for a dataset with N = 45222 samples. The value of k was calculated as the square root of N, which is a common heuristic for selecting an appropriate value of k.

The K-nearest neighbors (KNN) algorithm determines the class of a new data point by considering the class of its k-nearest neighbors in the training dataset. In this case, k is set to 212. The selection of k is a critical parameter in KNN because a small value of k may result in overfitting, while a large value of k may lead to underfitting.

To achieve good performance in KNN, it is important to select an appropriate value of k. To determine the optimal value of k, we can plot the error rate versus the k-value graph. In this case, we observed that the error rate was 0.5 when the k-value was 25, 33, 38, 40, and 48. These values of k indicate that the algorithm is not performing well, as the error rate is too high.

**Interpretation**

According to the statistics, there was a gender wage inequality, with a bigger percentage of males than women making above $50,000. The strongest positive relationships were between age, education level, and weekly working hours and pay. The relationship between capital gain and loss was shown to be the least significant. The boxplot showed that the greatest earnings were between the ages of 38 and 50, while the lowest earners were between the ages of 26 and 47, demonstrating that education level affected salaries.

The research shows that the existing model for pay prediction based on personal attributes is not particularly accurate, proving that there are additional elements than personal traits that affect income. The fact that gender seems to be a major factor in income inequality emphasizes the need for more research on this problem. Although they may not be as relevant as gender, other personal characteristics including age, country and familial position may nevertheless have an effect on income.

**Recommendation:**

We can collect more comprehensive data on household income distribution, and demographic characteristics based on the nature of work, revenue, expenses, and employment of workers and their earnings. This data can be used to develop more effective policies and programs to reduce income disparities and improve economic opportunities for all individuals, regardless of their background or demographic characteristics.

**Conclusion**

The research shows that the model for predicting salary based on personal traits was only 35% accurate, indicating that there are factors beyond personal traits that impact income. The fact that more than 50% of men earn over $50,000 yearly suggests that gender may play a significant role in income disparity. As a result, other traits such as nationality, family status, and age become irrelevant.

Therefore, we need to further refine the selection of k to find a better value that results in lower error rates. In general, we should select k based on the characteristics of the dataset, such as the size, distribution, and complexity of the data. Additionally, cross-validation techniques can be used to evaluate different values of k and select the optimal one that results in the lowest error rate.

**Appendix:**

**Box Plot:**

**Chart

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**Error Rate vs K value:**

**Shape

Description automatically generated**