Individual Contribution Report

CSE 578 – Data Visualization

Pranav Sanjay Toggi

My Contribution and Reflection

- To identify key attributes that influence salary, I established the core strategy of the team for individual analysis of the attributes. The essence of it is as follows,
  - Since, income is a binary attribute, we label the income levels '>50K' and '<=50K' as 1 and 0 respectively.
  - While analyzing a categorical attribute, we group the data frame by that attribute and extract the mean of the income.
  - Assuming the consensus data is representative of the population, the extracted means of the income are indicative of the probability of earning an income greater than $50K for each category in the attribute.

- Analysis of the Native Country attribute-

  Since the Native country attribute has 41 categories, overfitting and high complexity is a likely issue to be encountered in the predictive model. To counter this, as a preprocessing step, I grouped the native countries into corresponding global regions such as 'Asia-East', 'South-America', 'East-Europe', 'West-Europe'. Apart from countering overfitting, this coarsening of data made the interpretation of the visualizations, easier to comprehend.

- Analysis of the Education attribute-

  This attribute provides the education level of the individual and as demonstrated in Team50's Executive Report, is a strong factor that influences the Income of an individual. On doing an analysis on all factor level of this attribute, it was apparent to me that all education levels below High School Graduate, in other words, a Dropout, had the same probability of earning an income greater than $50K. Thus, it made logical sense to bin together those education levels into the 'Dropout' education level. This provides a cleaner stacked bar chart visualization to highlight the influence of Education on the Income.

- Analysis of the Age attribute-

  The Age attribute was considered as a categorical variable. To analyze this attribute a histogram graph of Income vs Age was plotted. As expected, the probability of earning an income greater than $50K grew steeply with age due to skill gained through experience until around the age of 50. Following which, the probability showed a downward trend with increase in age which could be due to depletion in health with age.

- Analysis of the Marital Status attribute-

  The analysis of Marital Status was a relatively straightforward one given that it had only 7 categories. Apart from generating a stacked bar chart to represent every level's influence on income, the order of the labels was also represented based on the magnitude of their influence on income. As expected, the individuals with spouses had a higher chance of earning an income greater than $50K.

Team Overview

The team developed classification strategies by doing an exploratory data analysis (EDA) on the data supplied by the United States Census Bureau. The development of strategy was made after individually analyzing the different attributes and how they influence the income class. Based on the analysis, 4 key attributes were chosen to be having the strongest influence on the income of an individual, namely, Education, Age, Occupation and Relationship. Since all chosen attributes were categorical variables, the standard deviation of proportions among the categories of the attributes was computed and compared among the attributes to determine the degree of influence on the income class. For example, if the standard deviation is zero for a particular attribute, it implies that the categories of that attribute have the same impact on the income and thus the attribute as a whole has no influence on income.

The EDA done by our team aims to equip the XYZ organization with crucial factors to consider in developing marketing profiles. These profiles will be utilized by XYZ's clients, the UVW College, to efficiently advertise and market their degree programs to bolster the college's enrollment numbers.

Lessons Learned

By embarking on this team project on data visualization, I have learned many lesson throughout the way, especially about the appropriateness of the kind and design of visualization for a particular use case. The best one being about how to not express unnecessary information in a visualization. For example, while analyzing the Marital Status attribute, a mere histogram chart of Marital Status vs Income is not good enough because it gives the unnecessary information about the frequency of the income levels for each category. The ideal approach would be that of a stacked bar chart which only highlights the proportion of the income levels for each category and which is the only factor that shows its influence on the income.

Assessment

This team project has definitely contributed to my development as a professional in the field of computer science and information technology. It has tested my every aspect of working as team member of a team with a common goal.

The project has taught me the significance of allowing maximum flexibility and liberty to team members in exploring different directions in the initial phases of development. There may be overlapping of focused effort, but I believe it is a crucial and foundational stage in establishing an agreed upon strategy to tackle obstacles.

Future Applications

Skills developed when undertaking this project were,

- Calculating correlations between numeric attributes and visualizing a correlation matrix to understand how the attributes relate with one another.
- Quantifying relationships between categorical attributes and the class label using standard deviation.
- Aggregating categories with the same influence level to extract a better attribute for analysis.
- Choosing the ideal visualization based on the information to be showcased to the target audience.

I believe the above mentioned skills in Exploratory Data Analysis (EDA) will be of significant use in any future Data science projects where predictive modelling is to be done using categorical data. The EDA skill in general is of vital importance to anyone wanting to make sense of any BigData in a particular domain for example, social media analysis and Know Your Customer (KYC) systems. The skills learned in this project can translate directly to being successful at embarking on large scale projects where “knowing the data your dealing with” is at the crux of it.