Prediction of factors influencing Income

Subhash Chandra Bose Vuppala
Svuppal6@asu.edu

Abstract*—This project aims to develop targeted marketing profiles based on income as a key demographic in the decision-making. We used the United States Census Bureau data to look at various features that can affect a person's income, including age, sex, education, occupation, capital gain, hours per week and marital status, etc. The key target is determining if an individual has an income above or below $50,000 with relevant attributes. In order to achieve this, we plan on using visualizations to find correlations between the individual features from the dataset and the income of a person.

Keywords—visualization, correlation, mosaic plot, bar chart

I. Introduction

Data preprocessing is the process of fixing or removing incorrect data from the dataset. It is vital for data processing to be done correctly not to negatively affect the end product or data output. Data processing takes raw data and transforms it into a more readable format, giving it the type and meaning that computers need to interpret. There are many stages in the data processing. They include data collection, data preparation, Processing, and Interpretation. In data collection, we collect data from available sources, which have data warehouses. It is important that the data sources used are reliable and well-built for the data obtained to be of the highest possible quality. After data collection, we move to the data preparation stage, and this stage is also called the preprocessing stage. In this stage, raw data is cleaned up, and information is diligently checked for any errors. This stage aims to eliminate redundant, incomplete, or incorrect data and create high-quality data. Machine learning algorithms are used to process the data, but the process can differ slightly depending on the data source. The interpretation stage is when data is finally available by people who are not data scientists. It is interpreted and readable, and it is often presented in graphs, videos, photographs, plain text, and other visual aids.

Data visualization is an versatile field that deals with the graphic representation of data. It is an efficient way of communicating when the data is vast. Data visualization tools understand trends, outliers, and patterns in data by using visual elements including charts, graphs, and maps. Some standard visualizations are bar charts, pie charts, mosaic plots, box and whisker plots, stacked bar charts, parallel coordinates, etc. If you want to display a distribution of data points or compare metric values across various subgroups of your data, a bar chart is used. Mosaic plots aid in the visualization of relationships and the comparison of classes. When attempting to determine the composition of something, a pie chart is better used for categorical data. A box and whisker plot is a visual representation of a collection of data on an interval scale. Individual data elements are plotted using parallel coordinates through a variety of performance measures.

II. Description of Solution

After the initial inspection of the dataset, we noticed that some fields contain the value '?', which is equivalent to a null or undefined value. To fix this issue we identify which features has the undefined values and those were work class, occupation and native country. In the next step we identify the mode or most frequent value of these features and fill the null values with the value that is obtained by mode or frequent value. We also pivoted on qualitative columns like education level, marital status, relationship status, etc. and vectorized the new columns that is Preschool, Doctorate, Husband, Wife, etc. to 0 and 1. We use 1 if value was present in the original, pivoted column and 0 if not present.

Before doing analysis, I, along with my team, used python's pandas' libraries to store and manipulate the data as needed for the models. To find the features that are best predictors of income and the accuracy of the predictions, we fit the data to a neural network and found the best features as age, fnlwgt, capital gain, capital loss, and hours per week. We found the feature importance by using the Extra Trees Classifier and Select K best. The top features of the Extra Trees classifier include fnlwgt, age, capital gain, whereas Select K best includes capital-gain, educational-num, relationship, age, etc. Then, I used Spearman's feature correlation (fig-1) and Pearson's feature correlation (fig-2) to find the correlation between the features.

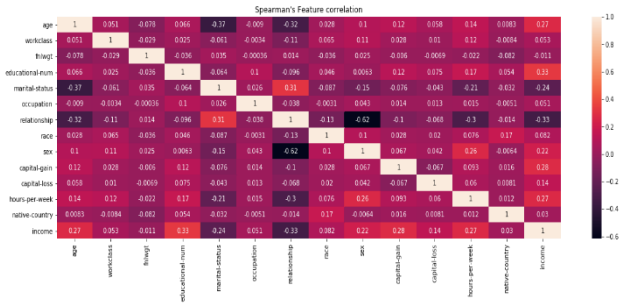

Fig-1 Spearman's Feature Correlation

We used exploratory data tools to visualize the data and found out correlations between the factors. The factors which we used include gender, occupation, age, hours per week, relationship status, education level, marital status, sex, native country, and race.

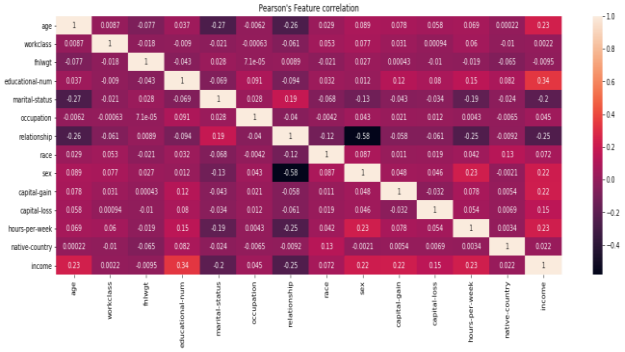

Fig-2 Pearson's Feature Correlation

III. RESULTS

A. *Age Vs Income*

- We use Box and Whisker Plot to visualize (fig 3) age and its relationship with Income.
- People with age between 35 and 50 are likely to have a salary above 50K.

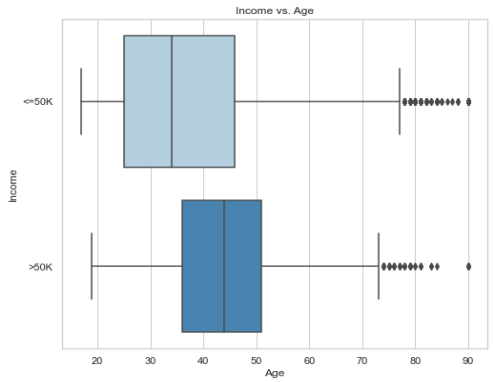

Fig-3 B&W plot of Age Vs Income

- There is a good portion of people whose age between 35 and 45 with the salary below 50K.
- Ages between 25 and 35 are more likely to have income of 50K and below.
- We can find some outliers between age 72 and 90 that have income above 50K.

B. *Relationship Status Vs Income*

- A bar chart (fig 4) may be used to compare values within the same function to see if any trends or clustering appear.
- To show how each sub-category performed against the average we use a mean line.
- The more divided the values are, the better the function would be at predicting income level.

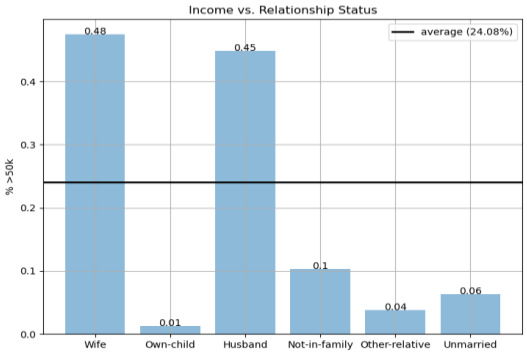

Fig-4 Bar chart of Relationship Vs Income

- Individuals that are Wife or Husband are far more likely to have an income more than 50k.
- Individuals with Own-child, Not-in-family, other-relative, or unmarried are significantly likely to have less than 50k salary.

C. *Native Country Vs Income*

- To compare numerical data across multiple categorical variables we use stacked bar chart (fig 5).
- More detail can be condensed into a single visualization for different categories of data than a simple bar chart.

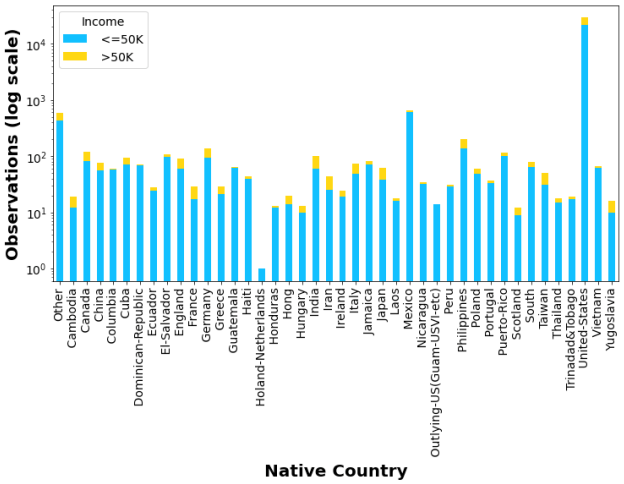

Fig-5 Stacked Bar Chart of Native Country Vs Income

- All countries have much more observations that have less than 50K income as label.
- As data is highly skewed we use log scale in visualization.

D. Education level Vs Income

- We use Bar chart (fig 6) to visualize Education level and its relationship with Income.
- Income with more than 50k is highly correlated with having a specialized, secondary education that is Prof-School, Doctorate, Masters.

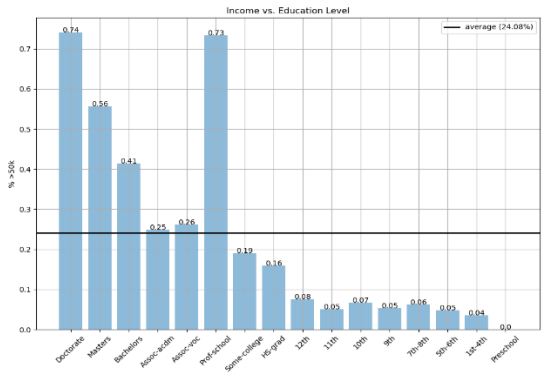

Fig-6 Bar Chart of Educational level Vs Income

- Those with degrees less than a bachelor's are statistically likely to have a salary of <50k.

E. Hours per week Vs Income

- We use Box and Whisker Plot to visualize (fig 7) Hours per week and its relationship with Income.
- Most of the people who work 40 to 50 hours per week have income above 50K.

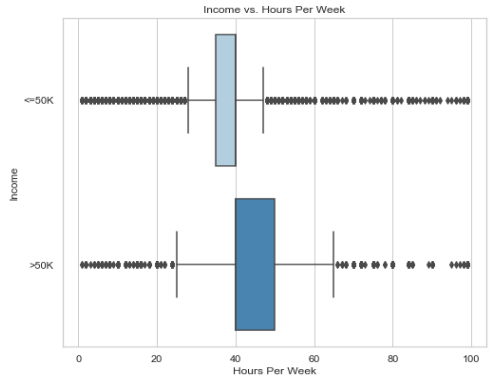

Fig-7 B&W plot of Hours per week Vs Income

- Most of the people who work 35 to 40 hours per week have income below 50K.
- There are some outliers who work either less than 30 hpw or more than 60 hpw and have income above 50K.

F. Race Vs Income

- A pie chart (fig 8) takes categorical data from a statistical sample and breaks them down by group, showing the percentage of individuals that fall into each group.
- Race feature is highly skewed.

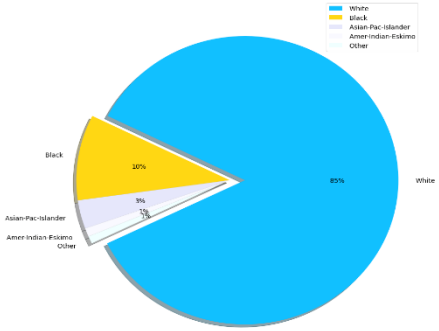

Fig-8 Pie Chart of Hours per week Vs Income

- 85% of sample is 'White', 10% of sample is 'Black' and 5% of observations for other races.

G. Marital Status Vs Income

- We use Bar chart to visualize (fig 9) Marital Status and its relationship with Income.
- People with Income greater than 50k are more statistically probable to have a marital status of married and present.

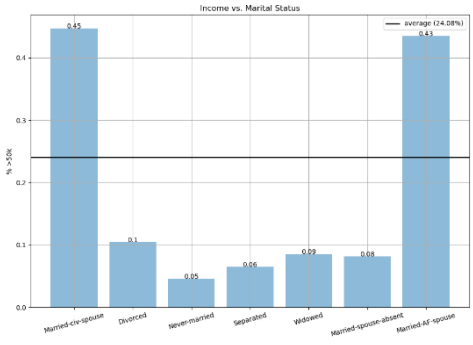

Fig-9 Bar chart of Marital Status Vs Income

- Income less than 50k is highly correlated with being unmarried or being married but not present.

H. Age, Capital Gain and Hours per Week Vs Income

- A parallel coordinate plot (PCP) is a visualization (fig 10) used to plot multivariate data, to check the relationships between them that is if they are correlated.
- For this scenario, PCP has been used to check the relationship between three variables, namely, hours-per-week, age, and capital-gain.

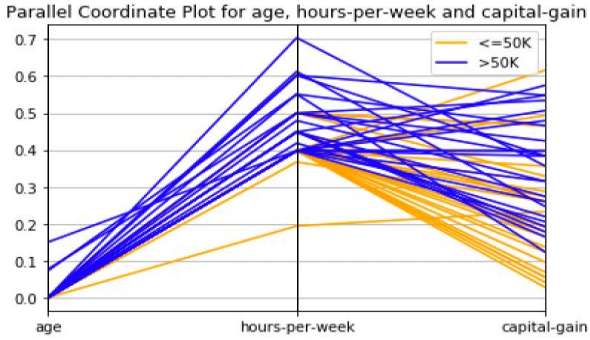

Fig-10 Parallel Coordinate Plot

- In the PCP plots above, the orange lines represent the salary values in the range <= 50K and the blue lines represent the salary values in the range > 50K.
- Individuals at an older age are likely to earn a higher salary than individuals at a younger age.
- There are always exceptions to every case, for example, an individual at a younger age can earn a salary greater than 50K as well.
- Individuals at an older age are more likely to work more hours per week than individuals at a younger age.
- Individuals that work a large number of hours are more likely to earn a salary greater than 50K.

IV. Individual Contributions

To analyze the provided data and clean it up, such as fixing missing data, I used data preprocessing techniques. My team and I used the pandas libraries in Python to store and manipulate the data for the models. We used a neural network to suit the data to find the best predictors of income and the accuracy of the predictions, and the best features were era, fnlwgt, and capital gain, capital loss and hours per week. My visualizations are given below.

A. Occupation Vs Income

- The categorical feature from the dataset that has been used in the visualization is Occupation (fig 11), and its relationship with Salary.
- For most categorical data, the dispersion of the two classes is exceptionally skewed indicating that this component can be utilized to recognize among the two classes.
- Individuals with occupations such as "Exe-managerial", "Prof-specialty" are more likely to earn 50K income.

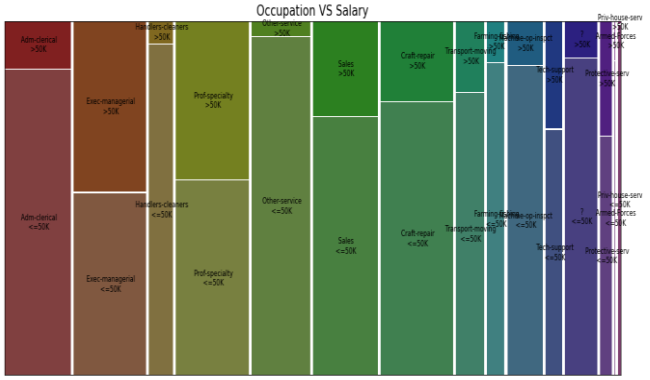

Fig-11 Mosaic plot of Occupation Vs Income

B. Sex Vs Income (Bar Chart)

- We use Bar chart to visualize (fig 12) Sex and its relationship with Income.
- Males with income greater than 50K are high in number when compared to females with income greater than 50K.

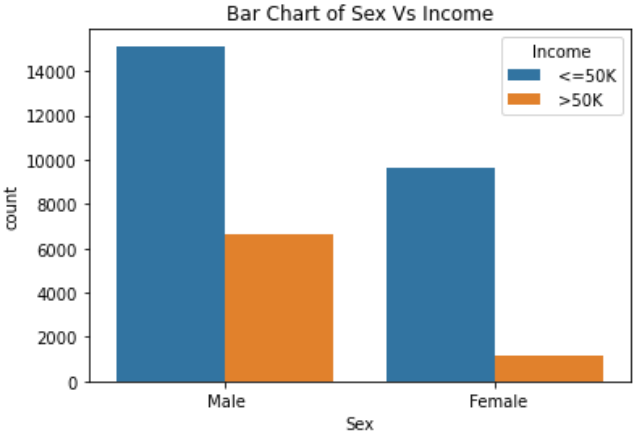

Fig-12 Bar chart of Sex Vs Income

- Males with income less than 50K are high in number when compared to females with income less than 50K.

C. Sex Vs Income (Mosaic plot)

- A mosaic plot (fig 13) is used to check how categorical data are related to each other.
- The categorical feature from the dataset that has been used in this visualization is Gender, and its relationship with Salary (<=50K or >50K)
- The color coding is done such that the smaller proportion has a lighter color, and the larger proportion has a darker color respectively.

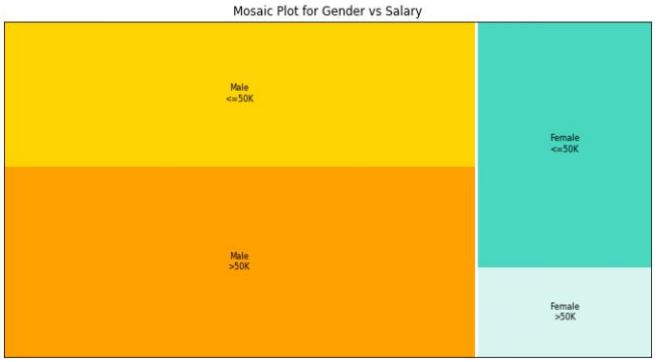

Fig-13 Mosaic Plot of Gender Vs Income

- A large proportion of males earn more than 50K salary.
- A small proportion of females earn more than 50K salary.

D. Marital Status Vs Salary

- The categorical feature from the dataset that has been used in the visualization is Marital Status (fig 14), and its relationship with Income.
- From the Mosaic Plot we can conclude that Married-civ-spouse are more likely to earn more than 50K income.

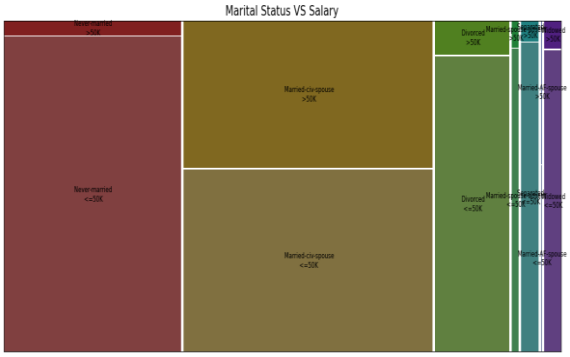

Fig-14 Mosaic Plot Marital Status Vs Income

V. Lessons Learned

- I learned the necessity of Data exploration and Visualization, which resulted in several inferences that helped solve the main task.
- Apart from the concepts I learned during the course and assignments, I went beyond doing my research to learn visualization tools and find patterns in the data.
- While doing the visualizations, I got a chance to explore seaborn and plotly, which made my work easy to plot mosaic plots and count plots.
- I learnt about violin plot, count plot and heat maps.
- A correlation matrix is used to determine the relationship between two variables and a variable with high correlation should be focused.
- Tableau is a powerful visualization tool that converts the raw data into a very easily understandable format.
- Finally, this project helped me to work on multivariate analysis, specifically on categorical data. After completing the project, I have had the opportunity to reflect on the importance of Data Visualization towards drawing meaningful conclusions from the dataset.

VI. References

[1] https://matplotlib.org/stable/gallery/index.html

[2]https://seaborn.pydata.org/#:~:text=Seaborn%20is%20a%20Python%20data,can%20read%20the%20introductory%20notes.

[3] https://www.tableau.com/learn/articles/data-visualization

[4] https://docs.python.org/3/tutorial/index.html

[5]https://www.talend.com/resources/what-is-data-processing/