Portfolio Report

Feature Selection Using Data Visualization

CSE 578 Data Visualization

Islam Hamed
Ira A. Fulton School of Engineering
Arziona State University
ID: 1221217088

I. INTRODUCTION

Data Visualization provides us with clear insight about what the data means, by giving descriptive chart or graph of data. This way makes the data for viewer mind more informative and natural, accordingly finding trends, patterns, and outliers within big data easier, faster and more efficient [1]. Also, data visualization helps decision maker in organization to have the right decision by viewing interactive charts and graphs of data. Many insights data visualization can provides, for examples:

- Correlations in Relationships: between independent variables.
- Trends Over Time: we need to have past and present information to be able to predict the future.
- Frequency: how often this trend happened.

Accordingly, in this project we answered customer's questions by data visualization. The customer improves marketing profiles and market sigma information using data. These data are used by many enterprises for marketing purpose. The customer is a College that tries to support and increase the enrollment. So, the college selected the income feature as demographical key to define the criteria for marketing its degree programs. By the United States Census Bureau provided data, we developed the marketing profiles, and determined $50,000 as threshold for income (salary). The data has many features that can affect the income class (>$50K, or <=$50k), including marital status, native country, occupation, education, and gender.

For instance, if the higher percentage of income <=$50k, has features of age is less than 40 years old, gender is male, marital status is single, and education is bachelor's degree, the college can promote to this demographic with tuition fees, program focus, and even ground or online programs suitable for this demographic.

In order to reach this enrollment target, a model has been developed. This model can find the features that affect the income label for individuals. This model used United States Census Bureau data. Also, this model able to classify the individual's income based on different features inputs.

The Datasets:

The dataset consists of 14 attributes (6 continuous and 8 nominal) and income class. The attributes detail as follows: [2]

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
15. class: >50K, <=50K

II. SOLUTIONS

We used different visualization charts to represent different attributes data type (continues and categorical). Also, we used each attribute independently from others in respect of income label. The used visualization charts are stacked bar chart, bar chart, scatter plot, mosaic plot, histogram and pie chart. In order to get the best interpretation form attribute's

visualization, some attributes have many different visualization charts. The used charts for the attributes as follow:

- Stacked Bar Chart: used to visualize and analyze Education-num attribute. Bars represent the two-income class (>50k, <=50k) verse number of individuals with specific education num. It was better to deal with education number values as categories especially it has a few distinct values, and it was easier to plot and analyzed. Example shown in Figure 1.

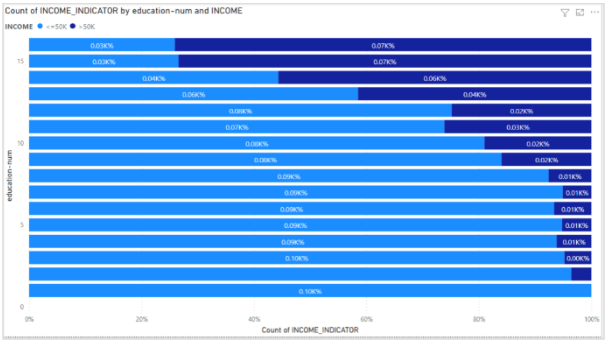

Figure 1 Stacked Bar Chart. Education Num vs Income Class

- Unstacked Bar Chart: used to visualize and analyze Education, Occupation, Work-class and Marital-status attributes. Each class label was represented by different color verse percentage of number of individuals in each category to have consistent charts and bars, which make the interpretation easier and more accurate. Example shown in Figure 2.

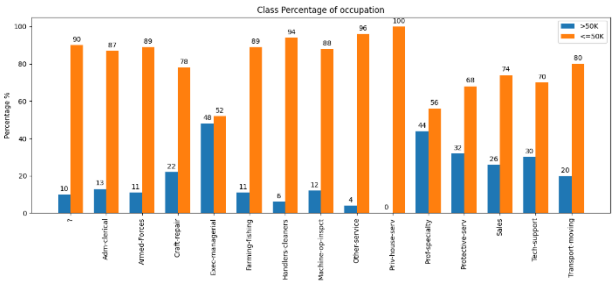

Figure 2 Bar Chart. Occupation Rate vs Income Class

- Scatter Plot: used to visualize and analyze Capital-gain and Capital-loss. Each individual's attribute value is represented by point verse the class label. Example shown in Figure 3.
- Mosaic Plot: used to visualize and analyze Gender and Relationship. Each attribute category was represented by rectangle (each rectangle width represents the rate of individuals with this category verse others) and each rectangle divided to two smaller rectangle that represent the rate of individual with respect to income class label. Example shown in Figure 4.

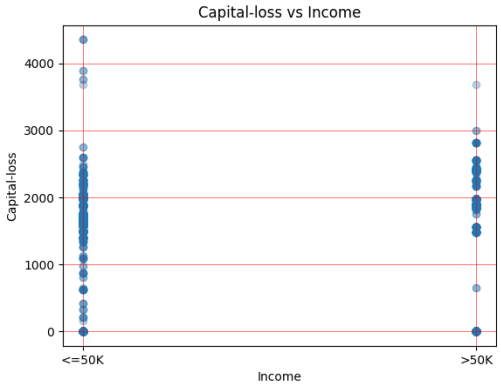

Figure 3 Scatter Plot Capital-loss vs Income Class

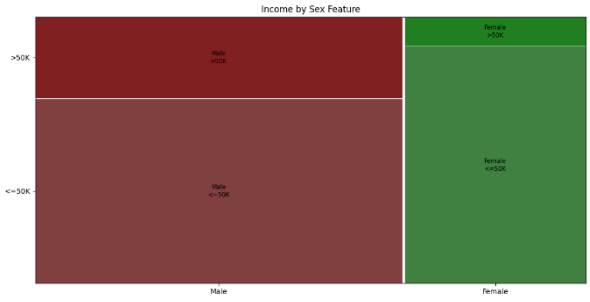

Figure 4 Mosaic Plot. Gender vs Income Class

In order to have a balanced graph, which is easier and more accurate to analyze and interpret, we used in most of the charts the rate of individuals in categories vs the income label, instead of the occurrence of individual in categories. Accordingly, the scale value is from 0 to 100 %. Also, we type the value of each category above its bar to be more readable and accurate.

III. RESULTS

After visualization, analysis and interpretation, we select four attributes, which are Education-num, Education, Capital-gain, and Capital-loss, as the most effect on income label. Based on their visualizations and our analysis and interpretation, these attributes can classify the income as >50k, or <=50k.

1. Education Num: shown in Figure 1. It represented by stacked bat chart. Looking to the chart, we can see that increasing the education num value, increases the income (>50k). So, we have a threshold here, if the individual has a certain education num value, we can know its income by checking in which class this value belongs.
2. Education: shown in Figure 5. It represented by bar chart. The education's values (Prof-School, Masters and Doctorate) are much more likely to

have income >50k based on the figure. Other values may have the other income class <=50k.

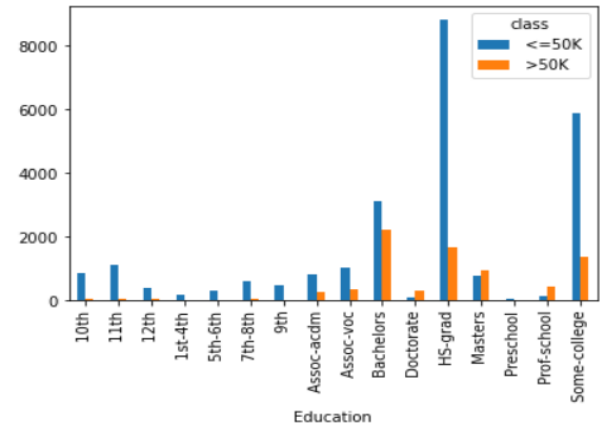

Figure 5 Bar Chart. Education vs Income Class

3. Capital-gain: shown in Figure 6. According to this chart, capital gain less than 10k is more likely to have income <=50k. Also, if the capital gain greater than 10k, the individual income more likely to be >50k. So, this attribute can classify the income based on individual's capital gain.

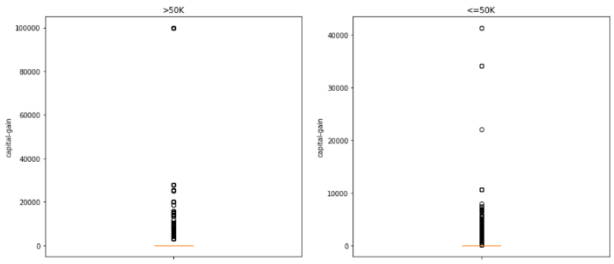

Figure 6 Box Plot. Capital-gain vs Income Class

4. Capital-loss: shown in Figure 3. According to the chart, there are range of capital loss where both class (>50k and <=50k) belong. On the other hand, there are ranges where only one class belong. So, this attribute can classify the income based on individual's capital loss.

The other attributes, according to their visualization; cannot classify both class labels (>50k and <=50k). Because the higher rate of individuals in the category belongs to one class (<=50k), as shown in example Figure 2. All class <=50k bars (orange colored) is higher than the other class >50k (blue colored). However, these attributes can support the prediction of <=50k class, as there are some categories have 100% rate of <=50k class. Accordingly, any individual with this category value has income <=50k.

IV. CONTRIBUTIONS

I worked on analysis and interpretation four attributes (Capital-loss, Native-country, Sex, and Occupation). In this stage, every attribute's visualization was independently from others. I was exploring whether the attribute's visualization can be helpful in my classification task in general or not, regardless other visualizations.

Capital-loss was the fourth one of the best attributes. As, it is really helpful to classify the income label (>50k, or <=50k). Occupation, and sex attributes may have an important insights.

One of the most important criteria, that I depended on when analyzing visualization and taking a decision, is if this attribute has values that can classify both income labels (>50k, and <=50k) or not. For instance, if there are two values from this attributes, one classify label >50k by 65%, and another classify label <=50 by 75%. So, if we think of two labels as class 1 (>50k) and class 0 (>=50k), we need to have attributes that able to classify both classes as true positive and true negative.

V. LESSONS LEARNED

I have encountered to new technology and techniques to visualize and manipulate the dataset. I used different visualization charts for different attributes properties such as continues and nominal.

I used Matplotlib [3], it is a complete library for generating static, animated, and interactive visualizations in Python. That provides many charts such as stacked, unstacked and grouped bar chart, line chart, scatter plot and pie chart.

Also, I used Pandas [4], it is a powerful and simple to use open-source data analysis and data processing tool, built on top of the Python.

In addition to statsmodels [5], it is a Python module that offers classes and methods for the evaluation of many different statistical models. I used it to plot mosaic chart.

Finally, in machine learning, data-preprocessing is an important technique ahead of using raw data. feature selection is a data-preprocessing technique, it uses to avoid the curse of dimensionality. I used to choose representation features or attributes based on their description, which is not accurate and not enough indicator if this feature is representative one. On the other hand, attribute visualization is a great indicator. It is helpful, quick and accurate one. As An Alternative of using some expensive computational techniques such as entropy, data visualization is more efficient and effective in respect of optimization.

VI. REFERENCES

[1] "analytiks," 10 June 2020. [Online]. Available: https://analytiks.co/importance-of-data-visualization/#:~:text=Data%20visualization%20gives%20us%20a,outliers%20within%20large%20data%20sets..

[2] US Census Bureau, "US Census Bureau," 05 May 1996. [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names.

[3] J. Hunter, "matplotlib," 08 May 2012. [Online]. Available: https://matplotlib.org/.

[4] W. McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython," 02 March 2021. [Online]. Available: https://pandas.pydata.org/docs/.

[5] J. Perktold, "statsmodels.graphics.mosaicplot.mosaic," 21 Feb

2019. [Online]. Available: https://www.statsmodels.org/stable/generated/statsmodels.graphics.mosaicplot.mosaic.html.