CSE578 Portfolio Report

Wenzhe Zheng
Arizona State University
Tempe, American
wzheng41@asu.edu

***Abstract*—Data visualization is the most important part of learning big data. After the data is cleaned up, it is presented to the audience in various graphs or tables, such as: Bar Chart, Pie Chart, QQ Plot, Mosaic Plot, Scatter Plot, Parallel Coordinate Plot, Star Plot, Box Plot, Line Chart, Histograms, Heatmap, etc. In this project, we will use these charts and data predictive analysis together, and at the end we will predict the future trend.**

***Index Terms*—Bar Chart, Pie Chart, Q-Q Plot, Mosaic Plot, Scatter Plot, Parallel Coordinate Plot, Star Plot, Box Plot, Line Chart, Histograms, Heatmap, Decision Tree, KNN [1], PCA (deduce all features into two features) and apply classification method, Linear Regression, Navie Bayes, SVM [3]**

I. INTRODUCTION

This project is a team project of CSE578. The purpose of this project is to exercise our ability to visualize and analyze data. In this project, the data we use is the income data provided to us by the United States Recognition Census. Our goal is to use US$50,000 the income threshold helps UVW University to help them find suitable candidates, thereby increasing the enrollment rate, so we have to determine which attributes have the greatest impact on personal income and the relationship between all attributes and income. We will use many machine learning classification methods,to predict the income of a person by inputting important characteristics.

Ṫhis project is divided into two parts. The first part is called: *Exploratory Data Analysis*. This part uses a variety of graphics to support data visualization to find the relationship between various characteristics, trends, and variables. To this end, we thoroughly train the entire database, perform various data prepossessing steps, such as separating elements and features, transforming categorical variables into 0 and 1, and dividing the data into training and testing data sets, and then transforming them into standardized Digital form and implement visualization. According to the visualization that has been carried out, we will analyze the relationship between various variables and characteristics and the proportion of income.

İn the second part of the project, we called: *Results Analysis*. In this part, we used several data classification models such as Naive Bayes, SVM [3], Decision Tree, Logistic Regression, and KNN to predict the relationship between variables.

A. About the dataset [1]:

- Data Source:US Adult Census
- Label:”>50K” and ”<50K”
- RangeIndex: 32561 entries
- Data Training:Remove unknown value (’?’)
- Memory Usage: 3.7+ MB
- Number of data in dataset after removing unkown value: 30162

B. Project Package Used:

TABLE I
PACKAGES

Language	Python
IDE	Jupyter Notebook
Packages	Pandas, Numpy, Matplotlib, Sklearn, Seaborn
Data Visualizations	Bar Chart, Pie Chart, Box Plot, Mosaic Plot, Scatter Plot, Heatmap
Prediction Models	Naive Bayes, SVM [3], Decision Tree, Logistic Regression, KNN [1]

In the following section called Solution, I will explain in detail the relationship between each graph and the results obtained after using each model to analyze the data.

II. SOLUTION

A. Phase One: Exploratory Data Analysis

Since we created a correlation heatmap(in Figure 1) based on the whole 14 attributes and the income label, we could assume that the top 8 darkest squares would be the attributes we wanted. However, the correlation can be negative so even though Relationship has a large negative effect on Income, it still is a key factor to be considered. Therefore, we assumed that we found the top 8 largest absolute values of the income row. After this operation we can have a sorted list of attributes which is [Education Num, Relationship, Age, Hours Per Week, Sex, Capital-Gain, Marital-Status, Capital Loss]

From this Heatmap, we have selected several interesting data and analyzed them. The data are as follows:

1)Bar Plot of Education-Num VS Income: From the heatmap visualization we can see that the Education-Num attribute plays the most important role when it contributes to the Income

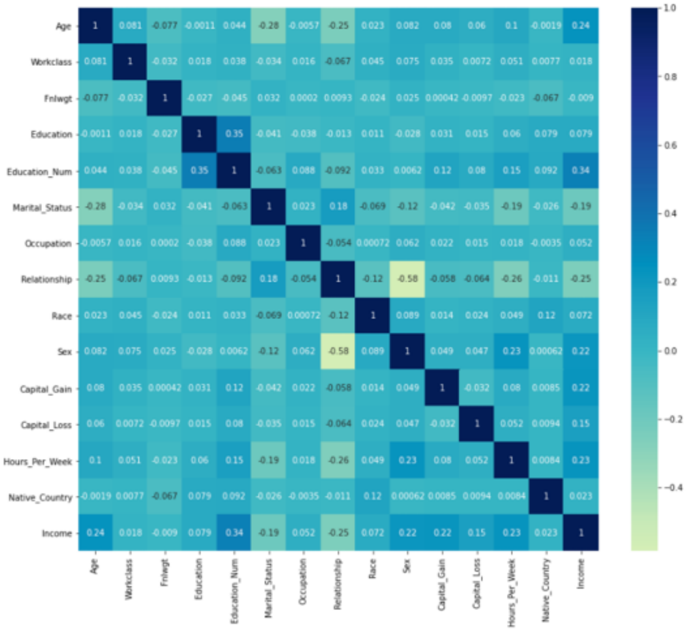

Fig. 1. Data Visualization Heatmap

label and it makes sense because the higher education a person gets, the more likely he/she will get high income. From Fig 1 we can see that if people get higher than level 13 education level, they have a higher chance to get over 50K of income.

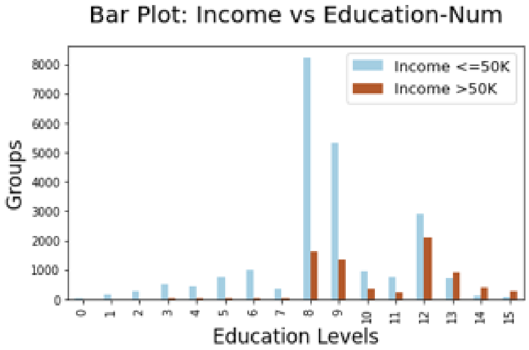

Fig. 2. Bar Plot of Education-Num VS Income

*2)Box Plot of Education-Num VS Income:*A Box Plot can always show the percentage of a group. As Fig 3 shows, there is a 75% of chance that a person gets under Education Level 10 when the income is under 50K while only a 25% of chance that a person gets under 10 when the income is over 50K. Besides, Education Level over 12 is considered an outlier when the income range is under 50K while there are no outliers of high education level in the income range of over 50K. This means that the higher education one gets, the more likely he/she has a high paid job.

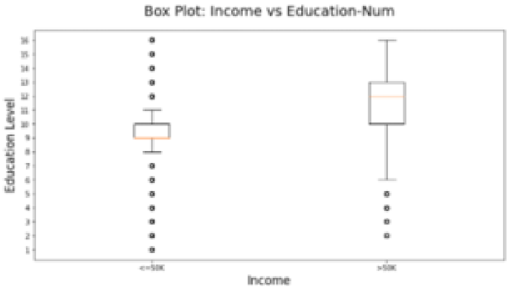

Fig. 3. Box Plot of Education-Num VS Income

*3)Mosaic Plot of Sex VS Income:*A Mosaic Plot is suitable for Sex VS Income plot because there are only 4 regions that should be visualized. As the plot shows, it infers that if one gets over 50K, the person is more likely to be a Male. Besides, there is more Male data than Female in the original dataset.

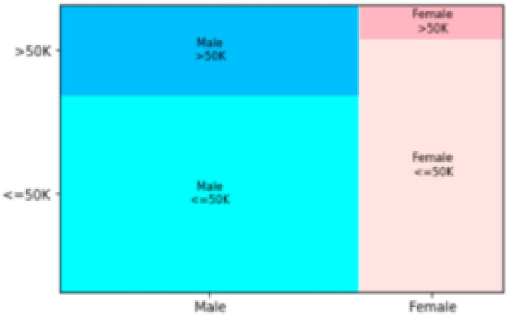

Fig. 4. Mosaic Plot of Sex VS Income

*4)Bar Plot of Marital Status VS Income:*Based on the heatmap, the marital status shows very low coefficient of income, however, it is an attribute that is always been asked or considered in real job seeking situations. So it is still a key factor of income which makes the plot meaningful. Bar plot is suitable with marital status VS income which can show clearly the number of people earning different incomes in different marital status. Based on the plot, we can see that no matter which marital status is, the number of people earning less than 50K is larger than the number of people who earn more than 50K. However, a civilian spouse can have much higher probability to earn more than 50K than other marital status.

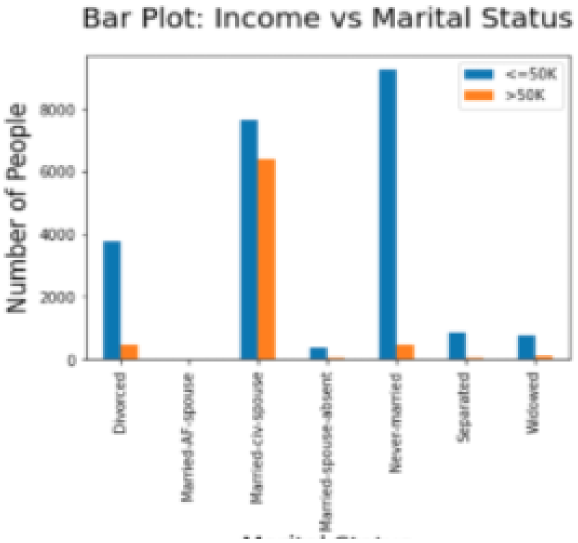

Fig. 5. Bar Plot of Marital Status VS Income

*5)Pie Plot of Workclass VS Income:*Based on the heatmap, the workclass has a low coefficient of income. But based on the real situation, workclass is always a determinant of how much income you can earn. The plot shows that a private has both very high probabilities to earn less than or more than 50K compared to other work classes. For other work classes, they all have higher probabilities to earn more than 50K.

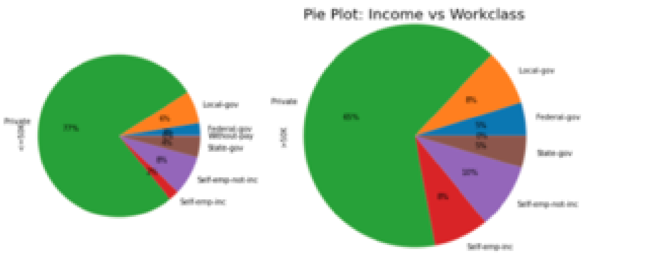

Fig. 6. Pie Plot of Workclass VS Income

*6)Scatter Plot of Age and Capital-Gain VS Income:*From the heatmap visualization we can see that the Capital-Gain attribute plays the most important role.It can be seen from this chart that capital gains are not closely related to age, but capital gains are closely related to income. Only when the income is greater, the capital investment and capital gains can be greater.

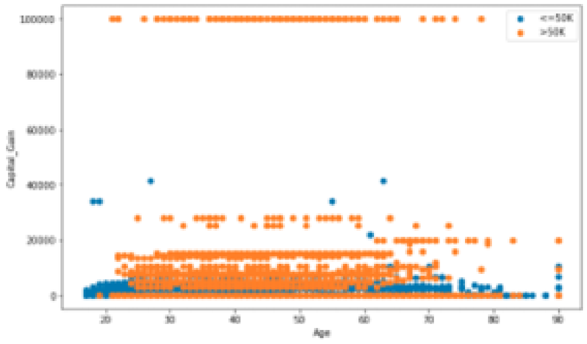

Fig. 7. Scatter Plot of Age and Capital-Gain VS Income

7)Bar Plot of Hours-Per-Week VS Income: Based on the heatmap, the Hours-Per-Week attribute also occupies a certain proportion of the income. After all, any work requires the work of working hours, and only work can be rewarded. We can see from the figure that, regardless of their income, those who work 36-45 hours a week account for the highest proportion of all people, indicating that 36-45 hours a week work is recognized by everyone and the best reporting ratio can be derived.

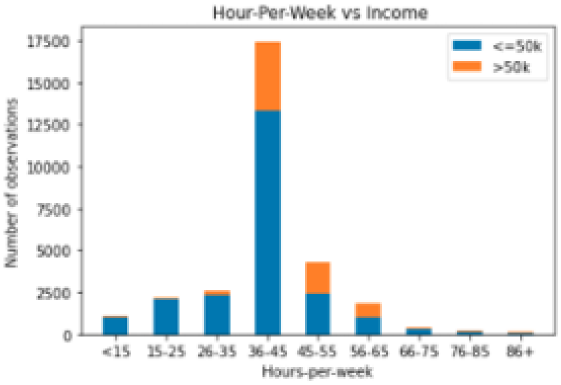

Fig. 8. Bar Plot of Hours-Per-Week VS Income

*8)Line Plot of Education VS Income:*Education is one of the most significant factors that influence people's earnings. As the heatmap visualization shows, education has a strong correlation with income. It's worth noting in the below graph that people with higher education levels, such as master's, professional school, and doctorate degrees, are more likely to earn higher wages than those with lower education. Also, we can see that people with lower educational levels have a lower chance to get higher salaries (¡=50 K).

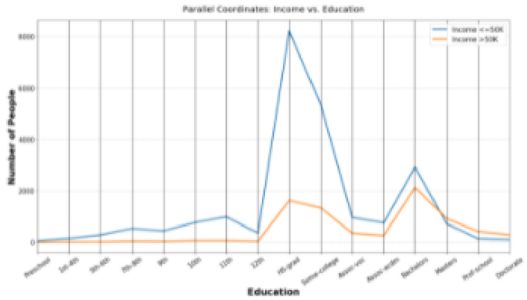

Fig. 9. Line Plot of Education VS Income

B. Phase Two: Results Analysis

In this part, we used NaiveBayes, SVM [3], LinearSVM, LogisticRegression, KNN [1], DecisionTree to train the data, and got different accuracy, the data is as follows:

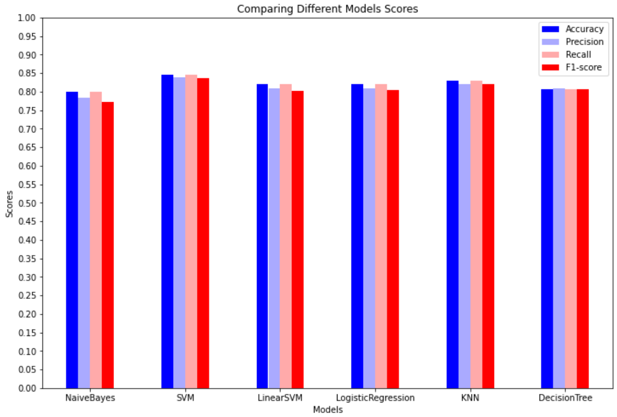

Fig. 10. All models accuracy rate and other comparisons

From the above figure, we can conclude that the accuracy rates of SVM [3] and LogisticRegression are very high, respectively 84.5 and 84.4.

Based on the above demonstrations of various pictures and our research using six machine learning models for predictive analysis, we have summarized the six most important features for predicting whether personal income is higher or lower than 50K. They are:

1)Education-num
2)Captial-gain
3)Age
4)Sex
5)Marital-Status
6)Capital-Loss

This project was done by ourselves but had not been reviewed by other teams. Besides, we did not have a chance to review other teams' results so the final exploratory analysis and predictive analysis were not authorized by either the marketing department (our customer) or other authorities. Therefore, our team's plan is to get someone that has experience in this marketing and business analysis field and let the professional people check on the final results in order to acknowledge the real world marketing system.

III. CONTRIBUTIONS

This is a team project. I have the honor to complete the analysis and sorting of this project with Wei Xin, Zian Zhang, Xinyi Liu, DongAo Ma, Mohammad Reza Hosseinzadeh Taher. Thank them very much. I have also participated a lot in this project. Things like:

- I drew Scatter Plot of Age and Capital-Gain VS Income(Figure 7) and Bar Plot of Hours-Per-Week VS Income(Figure 8)
- Organize a zoom meeting. Since our team members have all returned to their own countries due to the epidemic, we have three time zones. I used the meeting planner website to find the time when we can unify the meeting.
- Create task lists and assign them during discussions
- Create Github and share with group members
- Participate and discuss with the group members how to train data
- Use SVM machine model to analyze and predict data
- Pass the request of the member's merge code on github, and combine the code
- View the progress of other team members' data visualization
- Participated in and compiled project execution reports and system reports

IV. LESSONS LEARNED

The following list is what I learned from this course

- Learned how to train data more scientifically
- Learned a variety of graphics to better express the situation of the data
- Create task lists and assign them during discussions
- Create Github and share with group members
- Learned to use machine learning models such as NaiveBayes, SVM, LinearSVM, Lo-gisticRegression, KNN [1], Decision Tree to predict data
- Learned to express mine opinions more effectively in group discussions
- Learned how to efficiently carry out team work and manage other people's works on github

REFERENCES

[1] Harrison, O. (2019, July 14). Machine learning basics with the k-nearest Neighbors ALGORITHM. Retrieved May 03, 2021, from https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[2] https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

[3] andhi, R. (2018, July 05). Support vector machine - introduction to machine learning algorithms. Retrieved May 03, 2021, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47