Project on salary classification

Yuntian Xu

University of California, Irvine, Los Angeles, 90045, United States

yuntiax@uci.edu

Abstract. In this project, The results use three different machine learning algorithms to approach salary classification. The analyzed data used many different variables such as education level, age, and work-class to label each person into two categories, one with a salary greater than 50k and the other with a salary less than or equal to 50k. First of all, this work uses a single decision tree model to visualize data because it is more concise and understandable, and then by using the support vector machine method, the result becomes more accurate. After building two different models, The accuracy was found to be about 86.32%, which is relatively high and reliable. However, higher accuracy may be more persuasive. So, this project uses another model which is the random forest model. This algorithm is considered a highly accurate method because of the number of decision trees that participated. This model explained 87.03% of the accuracy of my result. According to my models, if a person desires a wage increase, that person should do his best to improve his education level, and he needs to have a stable marriage situation and be able to start his own business as much as possible between the ages of 20 to 60.

Keywords: salary classification, machine learning, decision tree model.

1. Introduction

Traditional economics recognizes that money can buy happiness because it occupies a very important place in society, both for individuals and for companies. This is because people need to use the money to exchange necessities of life and companies need to use the money to motivate employees. However, depending on the actual situation, money is not the only thing that measures happiness. According to the article “Money and Happiness: Rank of Income, Not Income, Affects Life Satisfaction ” written by Christopher J. Boyce, Gordon D. A. Brown, and Simon C. Moore, “The correlation between money and happiness is often small, but effect sizes are larger in low-income developing economies [1] and even small correlations can reflect substantial real differences in happiness [2].” So it can be concluded that although money has a small correlation with happiness, it is a tool to improve it.

For companies, money is a fantastic tool for increasing employees’ motivation. “when asked directly about the importance of pay, people tend to give answers that place somewhere around fifth in lists of potential motivators. In contrast, meta-analytic studies of actual behaviors in response to motivational initiatives nearly always show pay to be the most effective motivator [3].” So pay is usually the most motivating thing for employees, and then it will be important for the company to distribute the pay.

These two reasons are the background of why this project needs to be done.

Based on the two reasons presented above, This work will help individuals and companies. Individuals, are able to use the results provided by the project to see what individuals need to improve to get a higher salary and thus a better quality of life. Companies, are able to make better salary allocations so that companies can operate more efficiently.

This program is clearly divided into four parts, namely, introducing the data, cleaning the data, visualizing the data, and building the model. The data introduction and data cleaning are the preparatory work for data visualization and model building. And the first two parts will be addressed in the next section. The process of data visualization and model building will be explained in sections 4 and 5.

There are very many difficulties in this project. There are a lot of string variables in the given data, which is not good news for data visualization and model building, so these data need to be fixed with the label encoder tool in python, and there are a lot of variables in each column, and these data need to be sorted before making charts.

2. Data preparation

The salary classification data contains 32561 rows and 15 columns, Each column is a person's basic information, for example, the age column records the age of each person, and the work-class column records the attributes of a person's workplace, such as working for the government or owning a company. Another example is the column of marriage status which records whether a person is in a marriage. Most of the columns are in the form of strings, for example, the education column records whether each person's education level is high school, bachelor's, master's or doctoral. According to the basic information in the first fourteen columns, the last column is divided into 50k or less than or equal to 50k. This last column is also the most important one, because the last column needs to be equaled to y_train or y_test when we train data. This is the main outline of salary classification data.

Once again, browsing the whole dataset, there are some duplicate rows in the dataset, so in this step of data cleaning, the first thing is to drop these duplicate rows, after completing this step, the size of the dataset becomes 32537 rows and 15 columns. The second step of cleaning up the data is to handle the missing values, this step is quite significant because not handling the missing data will lead to biased results of the model building. But for this dataset, all the missing data appear in the form of question marks, and these question marks are all string variables, so the first step is counting the columns where the question marks appear, and finally, there are only three columns have question marks, and these columns are work-class, occupation and native-country. The next step is to replace the question mark with the most common value in each column, which are private, prof- specialty, and United-State. The reason why replacing the missing data instead of dropping them is that the missing data only appears in one row, and usually, only one variable is missing while the others are not, so if these rows are dropped directly it will lead to biased results. The last step of preparing the data is to classify the variables in each column, in this data, there are several columns with many variables, for example, in the work-class column there are eight variables, which are private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. The high number of these variables can affect the process of data visualization. They can cause the graphics to become complex and difficult to understand, so these columns need to be classified. The classification of the categories Self-emp-not-ine, Self-emp-ine as self-emp, Federal-gov, Local-gov as gov, and Without-pay, Never-worked as unemp was completed for all other columns.

3. Data visualization

After completing the data preparation, data visualization will be used as a tool to aid in the understanding of the data. As Antony Unwin noted in his article 'Why is Data Visualization Important? What is Important in Data Visualization?', a picture is worth a thousand words. The relationship between each variable and salary will be examined, and Figure 3 illustrates the identified quantities that have a relationship with salary [4].

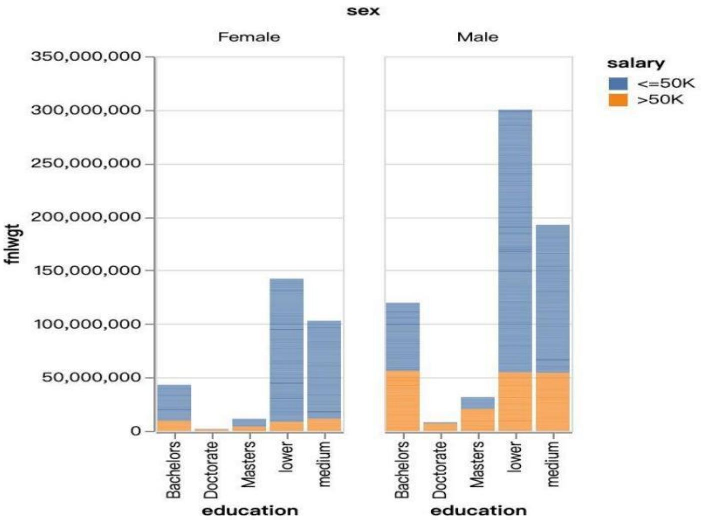

Figure 1. Relationship between education and salary.

Figure 1 illustrates the relationship between the level of education, number of people, and salary. The horizontal axis shows the level of education, the vertical axis shows the number of people, and blue represents a salary less than or equal to 50k while orange represents a salary greater than 50k. The data is also distinguished by gender. The figure shows that as the level of education increases, the likelihood of obtaining a higher salary also increases. Specifically, the bar for Ph.D. indicates almost no blue, while the proportion of orange in the bars for master's and undergraduate degrees is much larger than blue. This finding is consistent with other academic articles, such as 'Higher Education, Mental Ability, and Screening' by Paul J. Taubman and Terence J. Wales, which highlights that census data and related studies show a positive correlation between earnings and education level, with the social rate of return to education at least equal to the return on other societal investments [5].

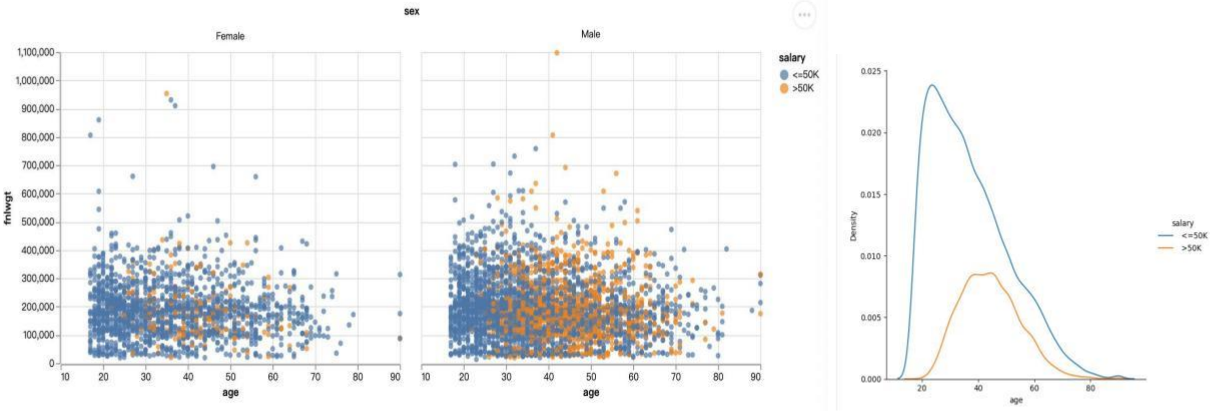

Figure 2. Relationship between age and salary.

Figure 2 displays two dot plots and one line plot, all of which have age on the horizontal axis and the number of people on the vertical axis. The plots use two different colors to distinguish wage levels. The figures clearly show that the peak years of wage growth for both men and women are between the ages of 20 and 60. This finding aligns with the biological perspective that individuals are typically stronger in terms of cognitive and physical ability when they are young. Therefore, this outcome is not unexpected.

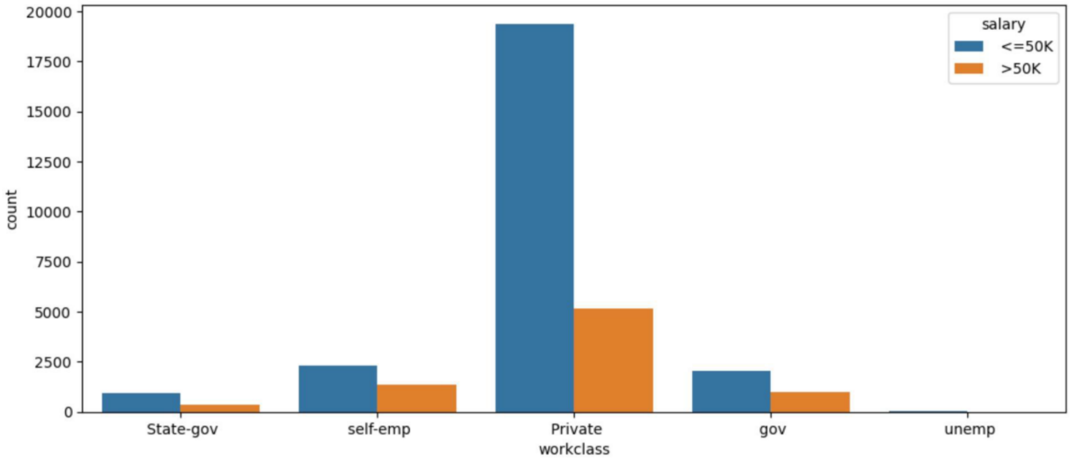

Figure 3. Relationship between work-class and salary.

Figure 3 presents a graph with the horizontal axis representing the type of work, the vertical axis representing the number of people, and different colors used to distinguish salary levels. The data indicates that the number of people working in private companies is the highest, and the number of people earning more than 50k is also the highest. However, the graph reveals that less than half of the people working in private companies earn more than 50k. Conversely, more than half of the self-employed individuals earn more than 50k. Therefore, it can be concluded that starting one's own business can provide greater opportunities for higher salaries.

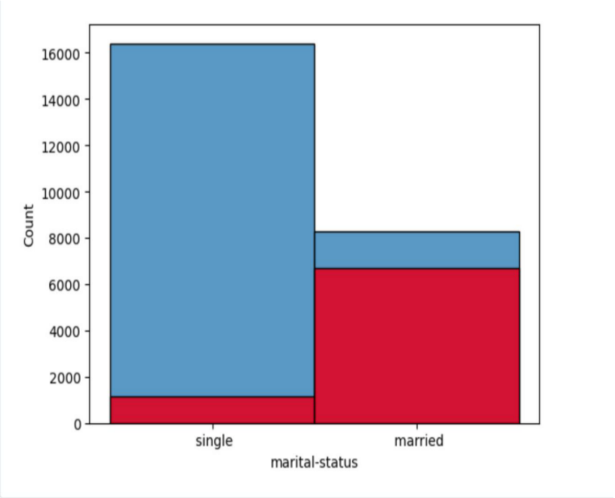

Figure 4. Relationship between marital-status and salary.

Figure 4 displays a graph with the horizontal axis representing the number of marriages and the vertical axis representing the number of counts. Prior to creating the graph, the marriages were divided into single and married categories to aid in interpretation. Additionally, a new column named salary_num was generated by encoding the salary column as 1 for salaries above 50k and 0 for salaries below 50k. In the graph, the color red represents 1 and blue represents 0. The graph clearly illustrates that individuals with a positive marital status tend to have higher salaries.

There is also a paper named “Does Marriage Matter?” written by Linda J. Waite shows that because two people can share the cost of living together when they are married, and because they can share a home, a vehicle, and joint property, people with higher marriages can focus more on their work [6].

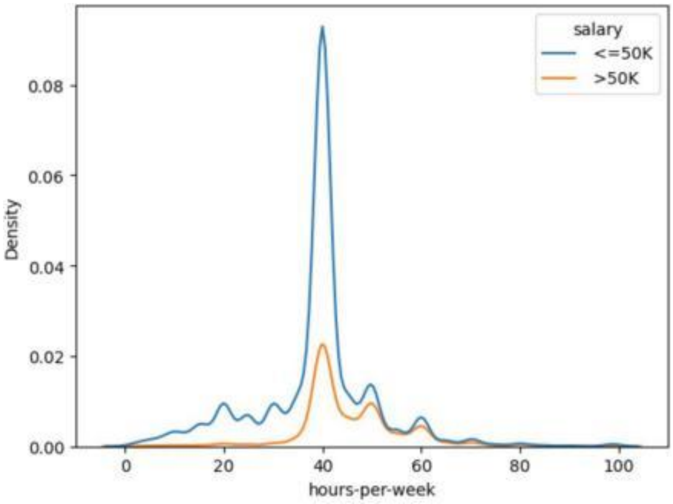

Figure 5. Relationship between hours-per-week and salary.

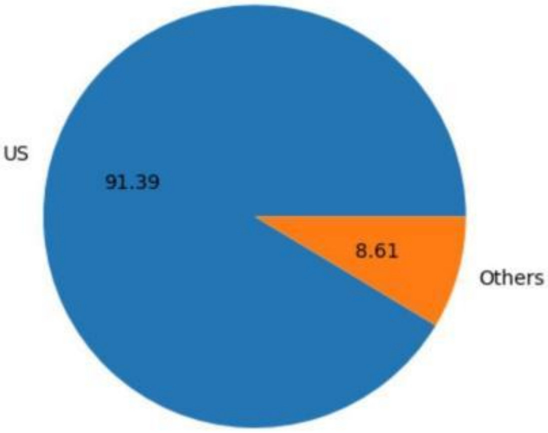

Figure 6. Headcount ratio to country.

Figure 5 displays a graph with the horizontal axis representing the hours of work per week and the vertical axis representing the density. Figure 6 is based on the native-country column, where countries other than the United States are grouped as 'other.' Both graphs show that a majority of people tend to work 40 hours per week, likely due to legal restrictions in the United States. Therefore, the impact of working hours on wages cannot be fully considered. However, in reality, individuals may choose to work longer hours for higher pay.

4. Models

After going through the data visualization, There are three different models can be used in this project. The first model is the decision tree model because this model is one of the most widely used tree-like algorithms, which is applied to the classification problem and can handle a vast volume of information [7]. This model is very well suited for my data. Before building the model, This data needs to be further processed, and by understanding the data, most of my data is in the form of string variables, which is not friendly for building the decision tree model. So the first step is to convert these string values into integer values. And the Label Encoder method in python is able to complete this step. This method can normalize labels and transform non-numerical labels. Except for the salary column, all other columns become numerical columns, and then use train_test_split to fit the data. And I use 90% of the training size to do this.

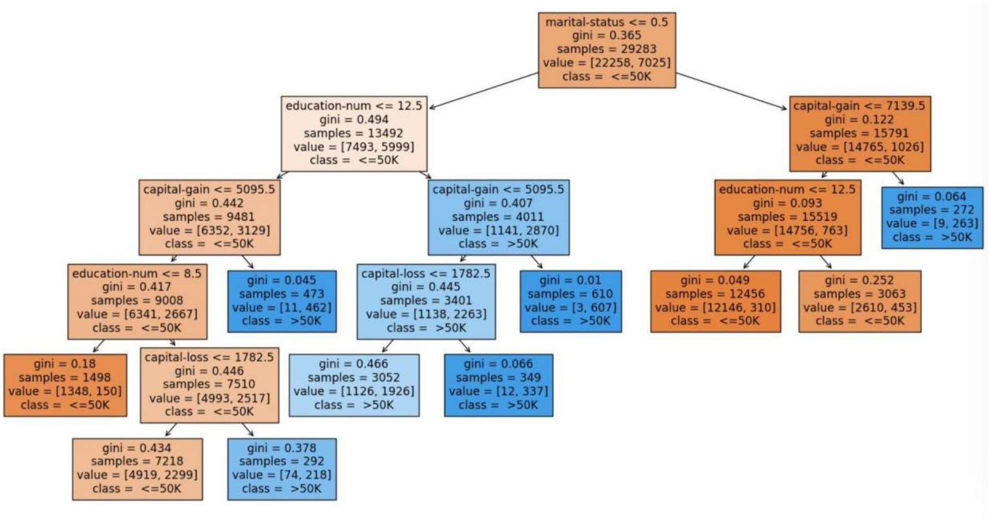

Figure 7. Decision tree based on salary classification data.

After fitting the data, this project uses the method in sklearn to build the decision tree, as shown in figure 7, which is a smaller decision tree because it can show a clearer picture. The picture helps to categorize the boxes according to the first row of each box step by step from top to bottom. Then this project performed an accuracy test with a result of 86.32%, a result shows very successfully in performing the classification task for the data.

The second choice of model is the support vector machine algorithm. This model is currently one of the most popular algorithms. It leans by example to assign labels to objects [8]. This project expects a higher accuracy rate in this model. The usage of this model is very similar to the decision tree, and after applying the model, the accuracy of both models is very similar, both are around 86.32%.

This accuracy needs to be further improved, and the random forest model can do a better job. The only difference between this model and the decision tree is that this model uses many decision tree models at once to improve the accuracy. It is largely accepted that the performance of a set of many weak classifiers is usually better than a single classifier given the same quantity of train information [9]. But before applying this model, using feature importance to process this data again can improve accuracy because choosing the right set of features for classification is one of the most important issues in designing a good classifier [10].

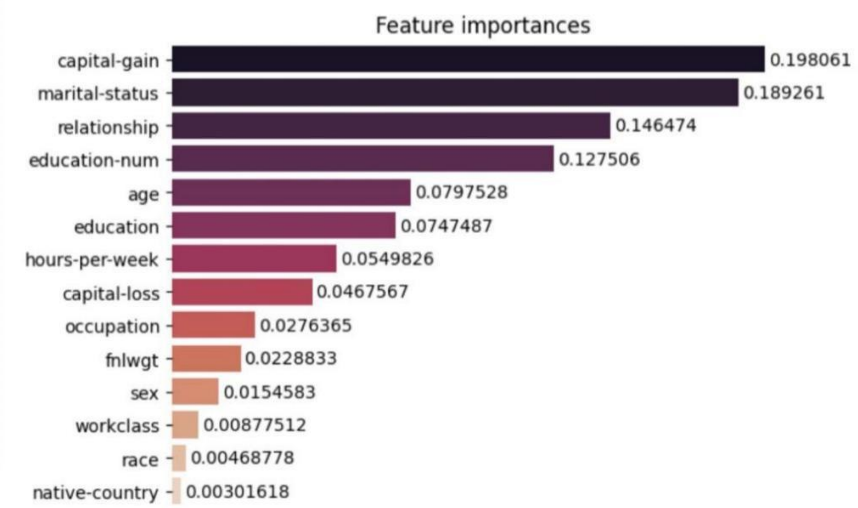

Figure 8. Feature Importance about salary classification data.

From figure 8, it is clear that the native-country column and race column are the two columns with the lowest impact on the data, so dropping these two columns can help to build the random forest model. This project uses RandomForestClassifier when building the model, and this program applied 1000 decision trees to build the random forest algorithm. In the end, the accuracy increased to 87.03%, which is a very high and reliable accuracy.

5. Conclusion

This project explores the factors that affect wage levels and visualizes the data in a way that is good for understanding the data. Three models were used to test the accuracy, and the final result was 87.03%, which is a relatively high and reliable accuracy. According to my models, if a person desires a wage increase, that person should do his best to improve his education level, and he needs to have a stable marriage situation and be able to start his own business as much as possible between the ages of 20 to 60.

References

[1] "Salary Classification", Kaggle, published by AYESSA https://www.kaggle.com/datasets/ayessa/salary-prediction-classification

[2] Boyce, J Christopher, Brown, D.A. Gordon and Moore C. Simon. "Money and Happiness: Rank of Income, not Income, Affects Life Satisfaction", published on Psychological Science.https://dspace.stir.ac.uk/bitstream/1893/12866/1/BoyceBrownMoore_PsychScience.pdf

[3] Rynes, L. Sara, Gerhart, Baarry and Minette, A. Kathleen. "The Importance of Pay In Employee Motivation: Discrepancies Between What People Say And What They Do." https://download.clib.psu.ac.th/datawebclib/e_resource/trial_database/WileyInterScienceCD/pdf/HRM/HRM_2.pdf

[4] Unwin, Antony. "Why is Data Visualization Important? What is Important in Data Visualization?", published on Jan 31, 2020, Updated on Feb 02, 2020. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Why+isDataVisualization+Important%3FWhatis+ImportantinData+Visualization%3F&btnG=

[5] Taubman, J. Paul and Wales, J. Terence. "Higher Education, Mental Ability, and Screening". http://kumlai.free.fr/RESEARCH/THESE/TEXTE/INEQUALITY/Segment/OK%20Higher%20Education%20Metal%20Ability.pdf

[6] Waite, J. Linda. "Does Marriage Matter?". Published on Nov 1995 and published by Population Association of America. https://www.researchgate.net/profile/Linda-Waite/publication/14281103_Does_Marriage_Matter/links/5849c62008ae5038263d89f6/Does-Marriage-Matter.pdf

[7] Jijo, Taha. Bahzad and Abdulazzez, Mohsin. Adnan. "Classification Based on Decision Tree Algorithm for Machine Learning". Published on Journal of Applied Science And Technology Trends. Publish on Mar 24, 2021. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Classification+Based+on+Decision+Tree+Algorithm+for+Machine+Learning&btnG=

[8] Noble, S. William. "What is a Support Vector Machine?". Published at Natural Biotechnology on Dec 2006. https://www.ifi.uzh.ch/dam/jcr:00000000-7f84-9c3b-ffff-ffffc550ec57/what_is_a_suppor t_vector_machine.pdf

[9] Oshiro, Mayumi. Thais, Perez, Santoro. Pedro and Baranauskas, Augusto. Jose. "How Many Trees in a Random Forest?". https://www.researchgate.net/profile/Jose-Baranauskas/publication/230766603_How_Many_Trees_in_a_Random_Forest/links/0912f5040fb35357a1000000/How-Many-Trees-in-a-Random-Forest.pdf

[10] Perner, Petra. "Improving the accuracy of decision tree induction by feature preselection." Published on Nov 30, 2010. //www.tandfonline.com/doi/pdf/10.1080/088395101317018582