

**Module 1 Project: Understanding Income Inequality**

**ALY 6020**

Name: Kavitha Palthayvath

Instructor’s name: Justin Grosz

Due Date: 07/18/2021

**Abstract**

Unbiased payment system is one of the biggest problems every country faces with their citizens. This paper tries to leverage the dataset provided of US citizens to analyze the factors that lead to inequality in pay and leverages Predictive analysis concepts to predict the class of salary range a citizen would belong to by knowing a few basic attributes of his personal life. These attributes include age, education, marital status, occupation, relationship, race, sex, capital standing, and hours of work. Since this data is a raw data, we took careful consideration into taking the right steps to clean the data before thorough analysis and used KNN algorithm to be able to classify citizens based on their earning abilities and capital standing.

**Introduction**

In this project we are going to work on census dataset about attributes of US citizens. We will build a model to see accuracy of classification of salary (low income to high income) and understand attribute (occupation, gender, race, capital loss, capital gain) to develop policies in the US. These features form a basis for every one of us as humans to decide if a person should be classified to a high salary category or a lower one. We can give this power to a machine learning model to predict the classification and help us formulate new policies and find reasons for inequality in pay rate.

**Analysis**

**Part 1: Data Cleaning**

On first examination of the dataset, I found that this data covers various aspects of a person, from educational history, to work experience and the financial standing. On loading the dataset, I found that I need to provide right header to the columns so that I can refer to these columns in an easier manner. After this step I moved on to analyzing continuous variables and the outliers in them to clean the data. I did this by plotting box and whisker plot, and a scatter plot. As seen in Figure 1 and Figure 2.

From the plots I concluded that the age is rightly distributed and the average age of people under census is about 35-38 and does not need any change. Moving onto the financial aspect I created a box and whisker plot, and scatter plot for Capital Loss and found outliers. I considered this as an outlier because this distribution of value zero across all age group will not help us in analysis and contribute towards making prediction. I removed the outliers by replacing all rows where capital loss is zero with the mean value of capital loss. As seen in Figure 3 and Figure 4.

The same was seen for capital gain, but the outlier lied at the value of 100000 for every age group. Which can be seen in Figure 4 and Figure 5. I replaced the outliers with the mean value for capital gain. Moving onto the categorial variables, I chose to calculate the frequency distribution of each column to find any values in these categorial variables that are missing or do not make sense. On calculating the distributions, I found 2 columns (work class, occupation) which had ‘?’ as one of the category values. This indicated missing data, to address this outlier I replaced this ‘?’ with the respective mode of each column, thus concluding the data cleaning.

**Part 2: Designing Model**

Since we are dealing with a lot of categorial value, it becomes implicit that we convert this into numeric value with the help of an encoding method, I chose to use One Hot Encoding method using the get\_dummies function. Furthermore, the dependent variable, which is the salary, needed to be converted into a binary decision, hence I encoded salary <=50K as 0 and >50K as 1.

After taking care of encodings, I decided to find the correlation matrix to find which columns have a strong correlation with the salary column. The outcome from the analysis of the correlation matrix was that the following variables made strong correlation with the dependent variable in the order of strongest affinity as seen in Table 1.

As per the correlation matrix we see that **marital status, relationship being husband, capital gain, age, hours per work, gender being male, occupation being manager** are among the top contributing factors for deciding the salary classification and hence the inequality in the salary.

After the analysis of correlation table, I divided the dataset into training and test set (80:20) and used KNN algorithm with k=3 to fit the model and calculate the **RMSE (Root Mean Square Error) value for training and test data set, which came out to be 0.253, and 0.343 respectively**. This shows that our model is good as compared to the real values with the difference in RMSE of 0.09 contributed by the prediction accuracy of the model.

To check if the model is learning right or not, I plotted the Seaborn scatterplot for age and work class which have strong correlation with salary for both predicted y values and test y values. And it came out to be same which is indicative of that the model is learning correct. These visualizations are two-dimensional views of a multi-dimensional dataset. If you play around with them, it will give you a great understanding of what the model is learning and, maybe, what it’s not learning or is learning wrong (Korstanje, 2021) .

A picture containing scatter chart

Description automatically generated

A picture containing text, screenshot

Description automatically generated

Fig 7: Scatter plot for y pred Fig 8: Scatter plot for y test

Now to optimize KNN to choose the right value of K to have the most accurate model for prediction I used GridSearchCV library to find the right value of k for which we get the best performing model, I got the value to be 9 when testing for k = 1 to 10. When using k=9 we got the RMSE value as 0.317 which is lesser error then k=3 which we had initially used.

**Conclusion**

I do not believe that this model is accurate enough to be used in real world. It considers very few points and does not have enough maturity to make real life classification decisions with the amount and variety of data given to it. It needs to be nurtured as the per the need of using it and tested more thoroughly, addressing the data it needs to be made perfect and retrain the model accordingly. On performing a thorough data cleaning process and analysis it shows that data cleaning is an integral part of analysis, which if not done thoroughly can lead to wrong predictive analysis. For instance, if I would have left occupation variable as ‘?’ in the outliers we would have missed the fact that it has a strong correlation with determining the salary. In sum, the model does perform well after doing a thorough cleaning of the data and using the right value of K but it will not scale well as the data fed to it increases or decreases, there will be change in correlation and also a possible saturation in the error rate which depends on the value of K.

**References**

Korstanje, J. (2021, April 7). *The k-Nearest Neighbors (kNN) Algorithm in Python*. Retrieved from realpython: https://realpython.com/

**Appendix**

Chart, box and whisker chart

Description automatically generatedChart, scatter chart

Description automatically generated

Fig 1: Box Plot for Age Fig 2: hours\_per\_work vs ageChart, scatter chart

Description automatically generated

Table

Description automatically generated with medium confidence

Fig 3: Box plot for capital loss Fig 4: Scatter plot for age vs capital loss

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

Fig 5: Box plot for capital gain Fig 6: Scatter plot for age vs capital gain

|  |  |
| --- | --- |
| **Variable Name** | **Correlation Value** |
| marital\_status\_Married-civ-spouse | 0.4458525188656926 |
| relationship\_Husband | 0.40379123269651457 |
| capital\_gain | 0.3085352392779874 |
| age | 0.23036946784752046 |
| hours\_per\_work | 0.22768676056081139 |
| sex\_Male | 0.21462803456392784 |
| occupation\_Exec-managerial | 0.21093807533663944 |
| education\_Bachelors | 0.18037141111141938 |
| education\_num\_13 | 0.18037141111141938 |
| education\_Masters | 0.17418363542414983 |
| education\_num\_14 | 0.17418363542414983 |
| education\_Prof-school | 0.15462740992779914 |
| education\_num\_15 | 0.15462740992779914 |
| workclass\_Self-emp-inc | 0.13959621645343182 |
| education\_Doctorate | 0.12647298849503588 |
| education\_num\_16 | 0.12647298849503588 |
| relationship\_Wife | 0.12048366458910562 |
| occupation\_Prof-specialty | 0.11149217797224786 |
| race\_White | 0.08370967320210619 |
| capital\_loss | 0.07274637421619425 |
| workclass\_Federal-gov | 0.06211221735802704 |
| workclass\_Local-gov | 0.03457640206623019 |
| native\_United-States | 0.03255124635918786 |
| workclass\_Self-emp-not-inc | 0.0271898842151328 |
| occupation\_Protective-serv | 0.024872687521949412 |
| occupation\_Sales | 0.023977170366918144 |
| native\_India | 0.02236098486870131 |
| occupation\_Tech-support | 0.020950847758864893 |
| native\_England | 0.015653140725874086 |
| native\_Canada | 0.015319591678050953 |
| native\_Taiwan | 0.013751420815088606 |
| workclass\_State-gov | 0.013618838149645188 |
| race\_Asian-Pac-Islander | 0.01258709274396638 |
| native\_France | 0.011888351835930034 |
| native\_Japan | 0.011052155722558677 |
| native\_Iran | 0.010890057280989119 |

Table 1: Correlation Matrix with Salary