Salary Prediction with Analyzing Affected Elements by using Pearson Correlation

Zelin Huang*

University of California, Irvine, California, United States

*Corresponding author: zelinh3@uci.edu

Abstract. This paper proposes a model for predicting salaries based on the 1994 Census Database to predict whether the final salary is over 50 thousand dollars or not. In order to do this, the researchers will identify various factors that may influence wages and analyze their impact on salary. To find the most accurate model, the researchers will use 32561 records, each containing 15 elements, and apply three different machine learning algorithms: Random Forest, Decision Trees, and Logistic Regression. These algorithms will be evaluated using a 5-fold cross validation method, which is a commonly used technique for measuring the performance of machine learning models. The goal of this study is to provide guidance on wage levels for individuals entering the workforce, as well as to understand how various factors may affect an individual's salary. This information could be useful for job seekers, as well as for employers looking to attract top talent. By identifying the factors that influence wages, the model could also potentially help policy makers and other stakeholders to address issues of wage inequality and fairness in the workforce. Overall, the model developed in this study has the potential to provide valuable insights and guidance for a wide range of individuals and organizations.

Keywords: Salary predicting; Pearson correlation; random forest; decision trees; logistic regression.

1. Introduction

Salary is one of the most important criteria that graduating students look for in their future jobs. They want to get a job that pays well and aligns with the skills that they have. However, it is difficult for students to find the best match between the salary that they want and the skills that they have. If there were a system or a model that can predict the salary that students can get with the skills that they have, it will help students to apply for the right kinds of jobs and motivate them in their future careers. According to Khongchai and Songmuang, they build a salary prediction system which uses gender, faculty, program, type of work, job training, certificate, and GPA as the input for the system, and then output the predicted salary [1]. And they also send out 50 questionnaires to the students who are about to graduate. The result of the salary prediction system shows that it can be used as a tool to manage and set stable goals, and it will enhance the academic performance of the students because they will want to increase their chances of getting a higher salary [1].

There have been several studies on the development of salary prediction systems. Another study also mentioned that using a machine learning system, it could help an adult realize if he could get through safely if he suddenly ran into financial problems [2]. And the research points out that machine learning is one of the most effective ways to design a model that can predict salary.

Moreover, prediction systems can also help students identify which areas they are good at and which areas they need to work on, and the system will use the details of the students to come up with a prediction. And the system will empower students to pay attention to their skills and develop them accordingly [3]. It means that the salary prediction system does more than predict salary; it can help students identify what they need to improve. They can understand what they should do to increase their chances of getting higher salaries. As Khongchai and Songmuang explain, many students do not understand what they should do that can help them excel academically, and it affects their chances of getting high-paying jobs; a salary prediction system can motivate them to do better because they will be motivated by other successful students as the system can provide them with role models [1]. It means that salary prediction systems can be used to encourage students to strive for the best and develop future goals. The skills and knowledge that they have will determine what salary they can

get in the future, which can be extremely important information because using this information, they can develop a career plan for themselves. Those who do not have career plans tend to get frustrated with college and are likely to drop out, which can put their career in jeopardy, as explained by Khongchai and Songmuang. Therefore, a salary prediction system can help them get back on track and show them how they can improve their likelihood of succeeding in their future careers.

Furthermore, there are many other ways that can illustrate the importance of prediction by using machine learning. B. Bian, Q. Yuan and H. Zhang have conducted deep research about the financial valuation of early retirement to help people to find optimal retirement strategy [4]. Machine learning can also be implemented in the medical area, where it depends on the type of cancer, using the best fit models to predict the likelihood of a person having cancer [5, 6]. Besides that, it can also pre-processed the image data like the CT-scan image so that it would be more clear for the doctor to analyze it [7]. Another article proposed using machine learning into loan applications for the bank to lower the risks of people failing to repay the loan from the applicant dataset [8]. Mohammadi and Wang have a detailed study of how to use machine learning to develop the manufacturing process [9].

To conclude the characteristics of the prediction model, one of the most important factors of having an accurate model is that the dataset has to be realistic [10]. Because without a reliable data source, no matter how accurate the model can fit in, it is meaningless. In this paper, the introduction is about why it is significant to use machine learning and how it can apply to real life. In method, based on the 1994 Census Database, it will talk about what algorithm is used for predicting if one's salary is over 50k. In the result and discussion, it will explain the deciding factor of the experiment result and the suggestion learned from the result.

2. Method

In the paper, the method used for studying the relation between variables is Pearson Correlation Coefficient. The feature of Pearson Correlation Coefficient is to describe the linear correlation of two variables. The formula for how to calculate the relation is shown as Fig. 1. N is how many pairs for x and y. R is the Pearson Coefficient that indicates the association between x and y. In addition, the range of r is either positive, negative or zero. The maximum value of r is positive 1 and the minimum value is negative 1. The closer the value approaches to 1, it means that the stronger association between two variables. In contrast, if the value approaches negative 1, it means two variables are strongly irrelevant. Whereas if r is equal to 0, then two variables have no linear correlation, but it could be another type of correlation such as exponentially. Since in this paper, the main idea is to find a relation to indicate what is the direct factor that may have influence on one's salary. Thus Pearson Correlation works fine here.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \quad (1)$$

First of all, in order to calculate the Pearson Coefficient, both x and y should be numbers. Thus, the original dataset needs to do some data preparation. Table 1 describes each columns' information. Since some variables are objects, it needs to convert to integer type.

Table 1 shows the data set has a total of 15 distributes, and 9 of them are objects. Which means that half of the inputs cannot be used with Pearson Correlation if it cannot convert into integer type. Thus, by sorting the number of times of one input variable into multiple categories so that it transfers from object type into integer type. For example, the sex is object type without data preparation, but by dividing it into categories, it only has two types of variables which are men and women. By doing so, calculating the repeated times of each variable so that it can be implemented for the Pearson correlation.

Table 1. The variables of the dataset

Column	Non-Null Count	Data type
age	32561 non-full	int64
workclass	32561 non-full	object
fnlwgt	32561 non-full	int64
education	32561 non-full	object
education-num	32561 non-full	int64
marital-status	32561 non-full	object
occupation	32561 non-full	object
relationship	32561 non-full	object
race	32561 non-full	object
sex	32561 non-full	object
capital-gain	32561 non-full	int64
capital-loss	32561 non-full	int64
hours-per-week	32561 non-full	int64
native-country	32561 non-full	object
salary	32561 non-full	object

After doing so, the following figure 1 shows the heat map of the Pearson correlation for salary. By using the heat map to show the relation between two variables, it is much more direct and simpler to see the dependence of how each element can affect the salary. It shows that the redder it is, the more positive correlation it is. The bluer it is, the more negative correlation it is. It is clear that attributes are strongly related to itself, hence the diagonal line value is all equal to 1. To study the relation between salary and other factors, the only thing it needs to be consider is the last row of the heat map where it indicates the pearson correlation coefficient.

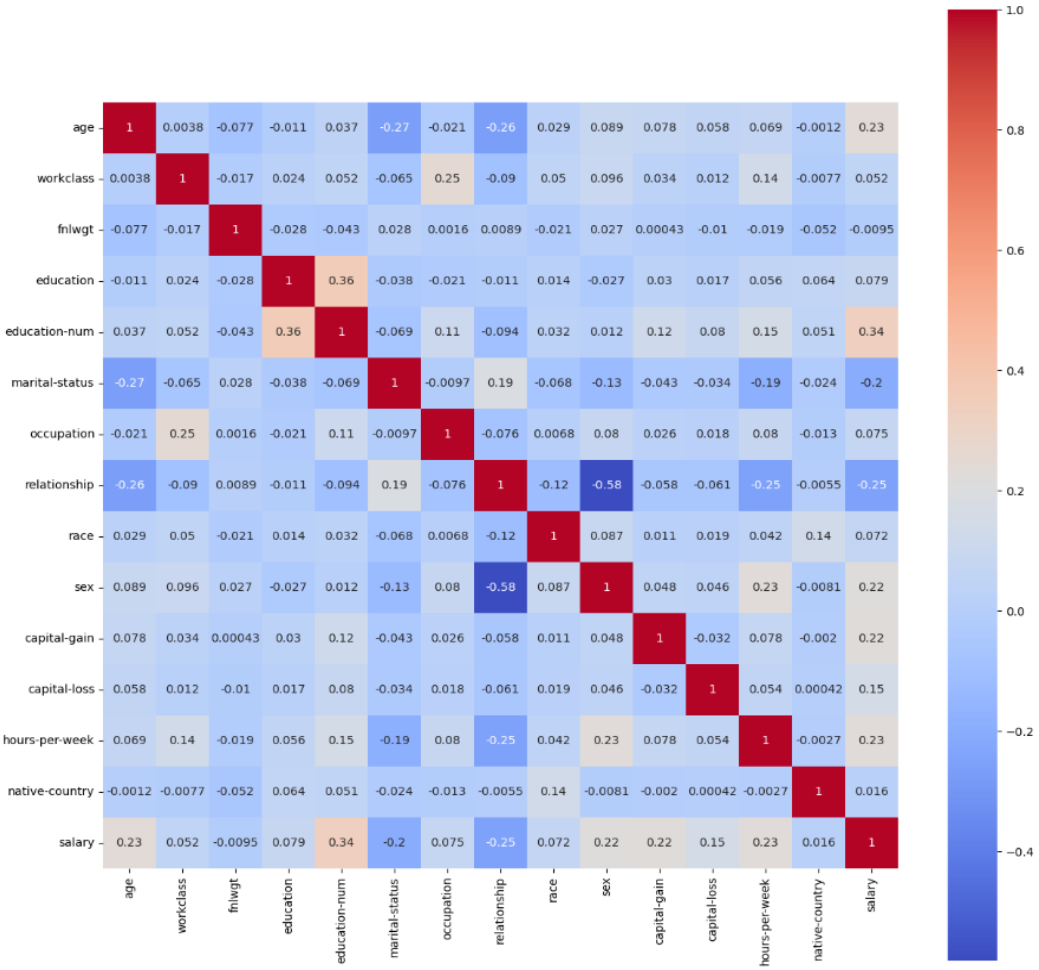

Figure 1. The Pearson Correlation heat map

3. Results and Discussion

3.1 Results

As the heat map is shown in Fig. 1, it is clearly shown that the age, education-num, sex, capital-gain, capital loss, hour-peer-work has great positive correlation with salary. Workclass, education, occupation, race, native-country has almost no correlation with salary. And fnlwgt, marital-status, relationship has negative correlation with salary. Among the positive correlation, education-num, age and hours-per-week are the top three effect elements for the salary.

The Fig.2 is a kernel density estimate plot which is for visualizing the distribution of the elements. The yellow line indicates the density curve of salary higher than $50k. The blue line indicates the salary with a density curve of salary that is less than $50k. From the graph, it can be divided into three parts. First part is from 0 to 7.5 years of education. In this phase, since the blue line shows the density of people's salary is less than 50k is much higher than the yellow line where it is almost closer to 0. The second phase is between 7.5 to 12.5 years of education. And the third phase is between 12.5 to 17.5 years. From the last two phases, it is a good illustration of how the number of years of education can affect one's salary. One's salary is higher than 50k has increased dramatically compared to the first phase. Furthermore, in the third phase, the yellow line is even higher than the blue line, which indicates that more people have a salary over $50k than salary is lesser than $50k.

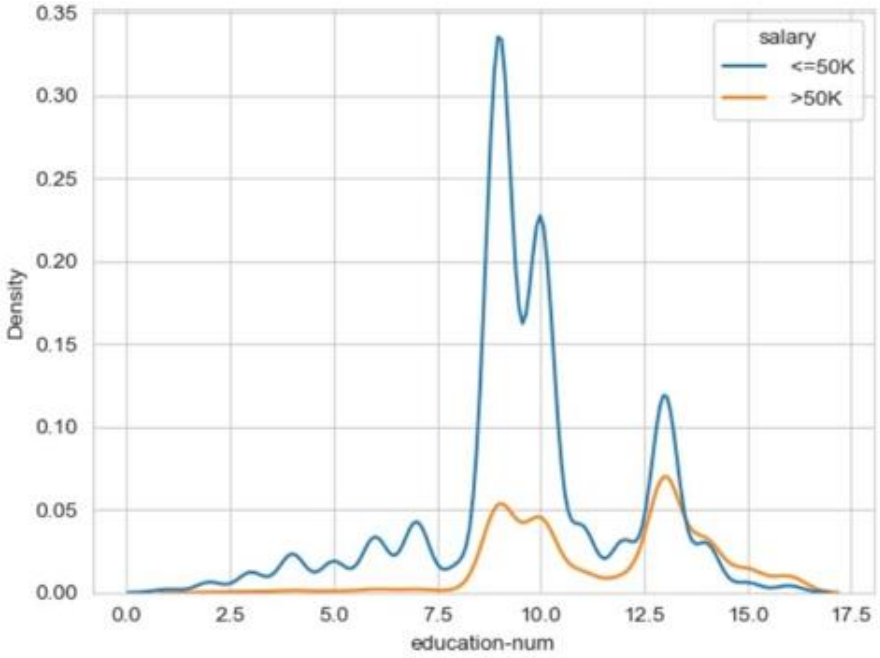

Figure 2. Education-num versus Salary density plot

Fig. 3 shows the kernel density estimate plot with x-axis is the age, and the y-axis is the density. The relationship can also be divided into three stages, the first stage is from 0 to 20 years old, the second stage is from 20 to approximately 70 years old, and the rest belong to the third stage.

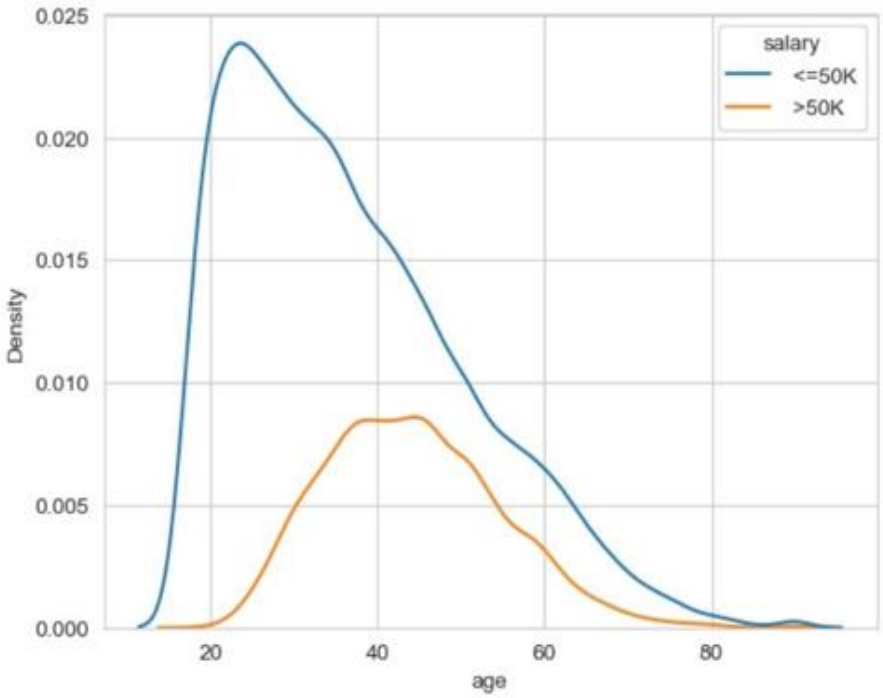

Figure 3. Age versus Salary density plot

Fig. 4 shows the kernel density estimate plot with x-axis is the working hour per week, and the y-axis is the density. From the plot, it can be divided into two stages, the first part is from 0 to 40 hours working per week, the second part is from over 40 hours working per week.

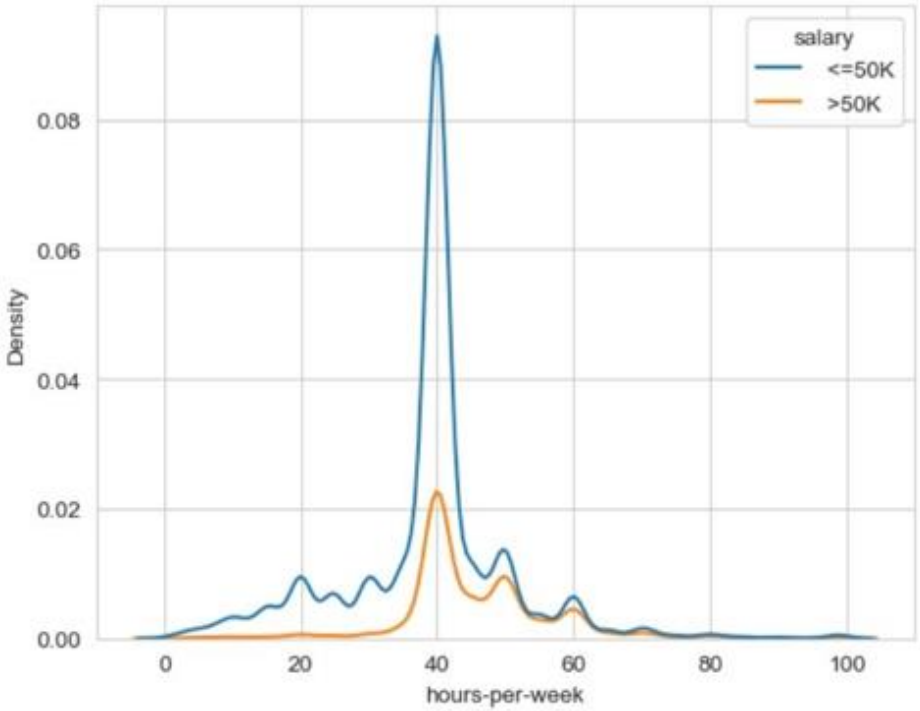

Figure 4. Hours-per-week versus Salary density plot

In Table 2, by comparing the model between Random Forest, Decision Tree, and Logistic Regression by five-fold cross validation, the test size is set as 20% and random state as 42. The accuracy for Random Forest is 84.74%, for Decision Tree is 81.18%, for Logistic Regression is 82.57%. The Random Forest shows the best predicting result.

Table 2. Accuracy Table

Model	accuracy1	accuracy2	accuracy3	accuracy4	accuracy5	Mean
Random Forest	84.124	84.781	84.781	85.119	84.904	84.742
Decision Tree	81.406	80.789	81.035	81.649	81.050	81.186
Logistic Regression	82.204	82.371	82.816	82.493	83.001	82.577

3.2 Discussion

According to the result section, it has been clear that some elements would effectively influence the salary. From the education years aspect, if someone has a long term of education, they will have more salary because the longer they stay in the school the more knowledge they can learn. At this period, someone can possibly find an area of interest that encourages him to continue learning. The above heat map has shown that salary has a negative correlation with occupation. In this case, combining these two factors, it interprets that the aim for getting a longer term of education is for self-improvement. If someone has achieved a very high level of his occupation, no matter what job it is, the salary should be higher than those who have less education than him due to the professional skill he has learned from the school and the experience he gained from years of practice. In the other words, high-paying jobs require a certain kind of skill, and that skill is not what humans are not born with. For instance, to be a programmer, it requires years of learning the basic concept, and beyond that, the fluency of writing codes needs frequent practice. That is also the reason why the density of salary over $50k for those who have over 13 years of education is higher than those who don't.

Besides that, the plot of age and work hour per week can be combined to emphasize how these two elements can affect salary. Apparently, one's age would be a preliminary factor that influences salary. At the younger age, without professional skills, there are barely jobs to do even if it only relates to physical labor. That is the reason why the density of salaries less than $50k is much higher. In consonance with the previous statement, with years of study, one could enhance the skills and knowledge. From Fig.3, it can be seen that starting at age 20, the density of salary higher than $50k is gradually increasing. The density tends to be steady from age 38 to 45. From Fig. 4, the density of salary over $50k dramatically reduces once the age is over 40. In the manner of a heat map, it also shows positive correlation between age and work hour per week. Obviously, from the point of view of physical health, human health will decline significantly which causes restriction of working hours. Without the ability to work, the situation turns back to the preliminary stage of being less qualified for some job.

4. Conclusion

In conclusion, this paper proposes a way to illustrate what element could affect salary by using Pearson Correlation. For further improvement, salary prediction systems can help students to find the correlation between their skills and their future career by determining their future salary. There are different ways of creating a salary prediction model using machine learning, and many studies have shown that the decision tree and the Random Forest can be very effective. In salary prediction systems, the details of the student are fed into the system as the input, and the system uses information from job requirements to provide the output, which is the salary prediction. It not only helps to determine salary but also shows the students what they can do to improve their chances of getting better jobs. Since it shows a correlation between their skills and the salary that they can get, this information can be used to motivate students. Not only that, by adding more relevant factors into the model, it may have further improvement and helps not only the student, but also the company to see what salary is appropriate for hiring someone.

References

[1] Khongchai P, Songmuang P. Random Forest for Salary Prediction System to Improve Students' Motivation. 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2016, 637 - 642.

[2] Pawar L, et al. Optimized Features Based Machine Learning Model for Adult Salary Prediction. 2022 IEEE International Conference on Data Science and Information System (ICDSIS), 2022, 1 - 5.

[3] Kumar, Nikhil, et al. Campus Placement Predictive Analysis using Machine Learning. 2nd International Conference on Advances in Computing, Communication Control and Networking, IEEE, 2020, 214 - 216.

[4] Bian B, Yuan Q, Zhang H. Financial valuation and optimal strategy for retirement benefits in a jump diffusion model. 2009 IEEE International Conference on Control and Automation, Christchurch, New Zealand, 2009, 2233 - 2236.

[5] Sruthi G, et al. Cancer Prediction using Machine Learning. 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddha Nagar, India, 2022, 217-221, doi: 10.1109/ICIPTM54933.2022.9754059.

[6] Bah A, Davud M. Analysis of Breast Cancer Classification with Machine Learning based Algorithms. 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), Istanbul, Turkey, 2022, 1 - 4.

[7] Tumuluru P, et al. Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms. 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2022, 349 - 353.

[8] Mohammadi P, Wang Z J. Machine learning for quality prediction in abrasion-resistant material manufacturing process. 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Vancouver, BC, Canada, 2016, 1 - 4.

[9] Bouvier V, Dreyfus-Schmidt L. Performance prediction under dataset shift. Piscataway: The Institute of Electrical and Electronics Engineers, Inc. (IEEE), 2022.

[10] Bansal U, Narang A, Sachdeva A, et al. Empirical analysis of regression techniques by house price and salary prediction. IOP Conference Series: Materials Science and Engineering, 2021.