Research on the Effect of Age and Experience on Salary based on Linear Regression and Tree-Based Train

Han Gao*

Natural Science University College London, London, NW1 1TP, United Kingdom

*Corresponding author: zczlhg2@ucl.ac.uk

Abstract. This paper examines the current salary of the employed based on the economic environment affected by the epidemic so as to have a better understanding of the economic market and to facilitate the employed to find ways to increase their salary according to their situation. This paper uses linear regression and tree-based train to investigate the effect of age and years of experience on salary. It is fi nally found that there is no obvious linear relationship between age and years of experience on salary, but according to the tree-based train it is found that the value of salary increases gradually with the increase of both. In other words, the larger the age, the more experience the employee has, the higher the salary will be. To sum up, the employed person should accumulate his/her work experience to get higher salary.

Keywords: Salary, Covid-19, Linear-regression,Tree-based Train.

1. Introduction

The international economy has changed drastically in recent years due to a number of factors, which has led to huge changes in the form of the economy of many countries, in which there have been huge fluctuations in the incomes of workers. Henderson reports on the Clinical Research Industry Salary Epidemic that has led to home quarantine, and the inflation of recent years that has led to significant layoffs and wage reductions[1]. Bansal and Narang used linear regression to make predictions of employee salary and house price and compared single linear regression and multiple linear regression and concluded that MLR is more accurate than SLR[2]. Yasser and Suliman use machine learning to make predictions about employee wages, where it is verified that MLR does not provide the best fitting model, while the non-linear machine learning model improves the model fit better[3]. Nesreen and Fang used tree-based train to predict employee turnover and applied tree-based methods such as random forest at the end verified that the probability of retaining an employee is extremely small when salary is lower than that of contemporaneous market competitors.[4]. Wu and Liu consider the best linear unbiased estimator of regression coefficients in the singular linear model[5]. Mayr et al. explains the application of boosting in machine learning[6]. Wu et al. use bagging to predict occupants' thermal comfort in buildings[7]. Emine and Abdulhamit try to compare bagging and boosting[8]. Lebanov et al. use random forest to Classify the chemicals in essential oils[9].

Therefore, this research will focus on the income of employed people in recent years. For the specific study this research will select a company from the United Kingdom as the sample target. This research will design a prediction model to predict the wages earned by the employees using the age and years of employment as the main factors affecting the wages, that will provide a good understanding of the income status of the employees in the current economic situation. In this paper, multiple-linear regression and tree-based train will be used to test the data. Multiple-linear regression will clearly show the correlation between the influencing factors and the results, such as the use of t-value, coefficient value and so on [10]. In tree-based train, we will adopt various tests such as random forest, boosting which iterates the sample set several times and divides different weights according to the previous result, and bagging which selects different parts of the sample set independently and compares them with each other several times in order to compare the optimal results.

2. Data and Method

2.1. Data

The data used in this paper contains 343 samples of a company in America and seven influential factors about salary. And this study selects age and years of experience as the factors that affect salary. The data set is from Kaggle (Table 1).

Table 1. Data set

Variable	Mean
Age	19.215817694369974
Years of experience	5.515415549597855
Salary	100577.34584450402

2.2. Method

In this paper, we will mainly use two model prediction methods, one is multiple-linear regression and the other is tree-based train.

2.2.1 Multiple-linear Regression

Multiple-linear regression is based on simple linear regression, which is a very simple supervised learning method, but it is still very effective in statistical applications [10]. simple linear regression very simply assumes that there is a linear relationship between a variable and the dependent variable, which can be expressed as Eq:

$$\mathrm{Y} = \beta_0 + \beta_1 X \tag{1}$$

to represent. The relationship between X and Y can be shown in equation (1), which is also known as regression Y on X. β_0 and β_1 denote the intercept and slope in the novel model, respectively. And the slope of them can show the degree of influence of the corresponding factors on the results. multiple-linear regression is the result of simple linear regression with the addition of the variable X. The result of multiple-linear regression is the result of simple linear regression with the addition of the variable X.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{2}$$

Equation (2) shows the simplest form of the multiple-linear regression who consists of multiple influences X. All values of β except β_0 indicate how much the respective corresponding value of X affects the linear model. For example, $\beta_1$indicates how much the factor X_1 affects the trend change of the linear model [11]. Based on the market logic that most companies generally prefer to hire experienced employees, this paper argues that the two main factors age and years of experience in the selected sample data should have a certain positive correlation with salary, i.e., the older the age, the more experienced the person is, the more salary he or she reaps.

2.2.2 Tree-based Train

Tree-based train will divide the data into several zones according to the characteristics and magnitude of the values and use the mean, mse and other characteristics of the observed data to predict which zone the data will be assigned to. On top of the basic tree-based train method bagging and boosting methods are used to improve the accuracy of the model, where the bagging method selects the test dataset multiple times from the sample data, but each time the test set is selected independently, which greatly reduces the variance of the model by comparing multiple independently selected sets, whereas the boosting data are compared to a set of independently selected datasets. boosting data, on the other hand, will conduct the next round of selection and testing based on the results obtained from the test sets that have already been selected, and assign different weights to the values of different regions each time to adjust the overall model bias[11]. Most of the 40 per cent of clinical research industry personnel who were studied through Henderson, Lisa. Applied Clinical Trials did not receive

additional bonuses. And 30 percent of those who received bonuses earned less than $10,000 on average. In addition, 75 per cent of industry personnel received no non-cash compensation such as cars, stock options or mobile phones [4]. Therefore, this paper speculates that the number of employees receiving lower wages in the tree-based train prediction should account for a majority of about 60% or more.

3. Results

3.1. Correlation Analysis

From Table 2 we can get the correlation between the two main factors of our study and the salary value which includes t value, p value, standard error and the correlation coefficient. According to the value of the correlation coefficient, both age and years of experience have a positive correlation with salary. But the very small p-values allow us to reject this assumption.

Table 2. Relevant Coefficient

	Coefficient	Std error	t	p
Intercept	-18700.0000	17100.000	-1.096	0.274
Age	1885.8292	632.537	2.981	0.003
Years of experience	4853.8434	681.935	7.118	0.000

3.2. Scatter Plot of Linear Regression

Figures 1 and 2 give the distribution between age and salary and years of experience and salary, assuming that there is a correlation between age and salary and years of experience then the distribution of points in the graph should show some regularity. However, it can be seen from Figure 1 and Figure 2 that these points are biased towards a random discrete distribution, which means that they do not show some correlation. This indicates that there is no significant correlation between age and salary and years of experience.

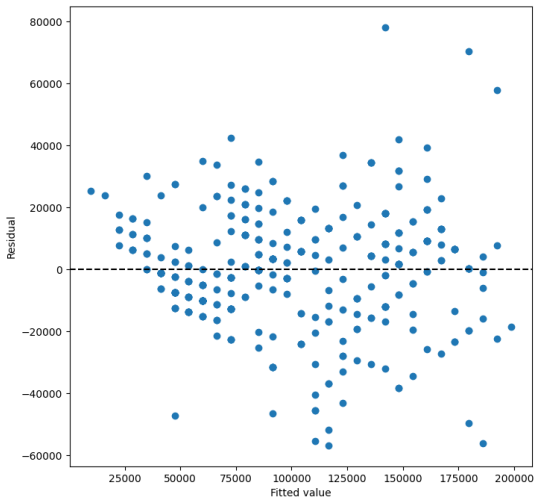


Fig. 1 Age and Salary

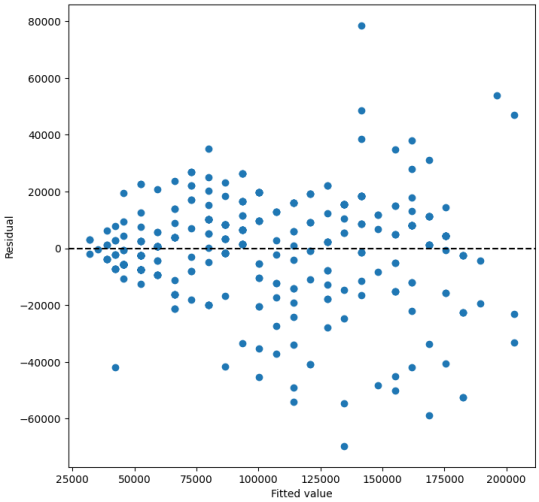


Fig. 2 Years of Experience and Salary

3.3. Results of Tree-based Train

Figure 3 shows the results we obtained using tree-based train. Here we first try to use fitting classification and make the continuous variable salary binary. Here we set salary greater than 100000 as high and set max_depth to 3. And from this we can see that years of experience is a more critical factor because the first branch is judged by whether the years of experience is greater than 8.5. And

whether the age is greater than 41.5 and years of experience is greater than 6.5 become the second division of the salary standard.

3.4. Developed Tree

In Figure 4 we try to go a step further with the fitting regression model, for which we set the test data to take up 30% of the overall data. This time the data is different according to the first branch indicates that whether the age is greater than 40.5 becomes the most critical factor affecting salary. The next two splits are based on years of experience, and the data show that the greater the years of experience, the higher the salary. And based on the final predicted data distribution salary values will be on the bimodal side, with very few people getting salaries around 120,000 near the middle. The number of people getting low salaries (<=100000) while being polarised is less than the number of people getting high salaries(>=150000).

3.5. Fitting Regression Model

To verify the accuracy of the model a new linear model including mse and data number was built. From this it can be seen that both test data and train data, their mse levelled off when the number reached 1600.

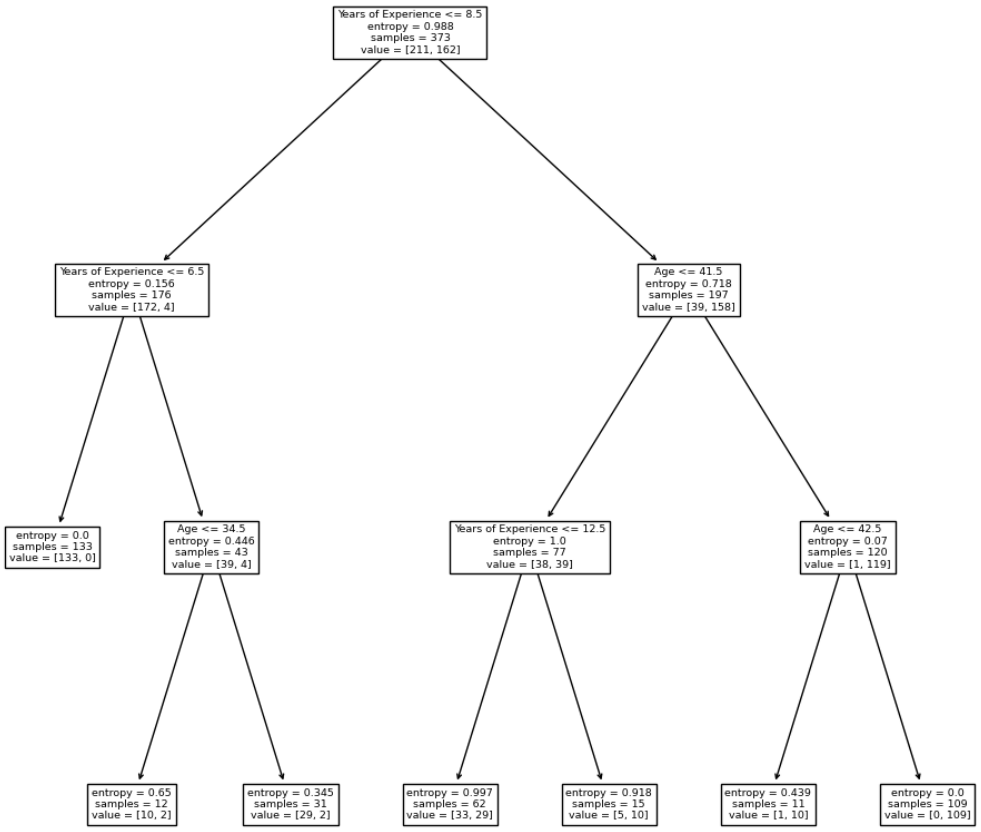


Fig. 3 Basic Tree

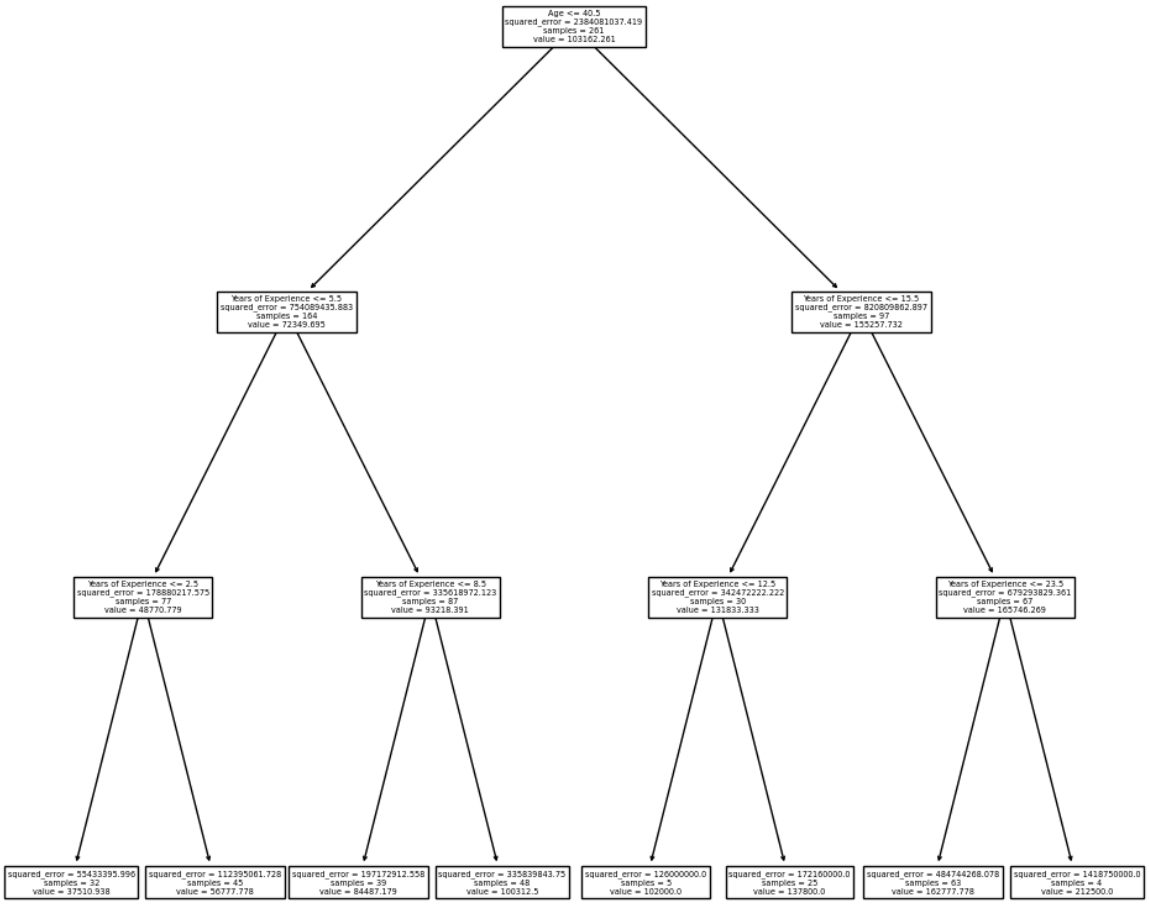


Fig. 4 Developed Tree

3.6. Accuracy

First this research calculates the relevant mse of the basic tree to be about 308224812 after getting the relevant data of the basic tree. Then this research also attempted to calculate the most possible mse value that he could have based on the data of the basic tree to be about 249344029. This represents the basic tree and the square root under the best prediction is about 17556 and 15790. This means that in the case of the basic tree there will be a gap of 17556 between the predicted data and the real salary, and this gap may be minimised to 15790. Next we will try to reduce the mse of our model in different ways, including bagging and boosting, to get it as close as possible to the best mse.

The mse value is 296363469 after using bagging, we can find that the mse value of bagging is not much different from that of the basic tree, that is to say, bagging does not make a good adjustment to the accuracy of our model.

After verifying the lifting of the model by bagging This research tries another method boosting to lift the model. In this experiment, the new function changes the max depth of the basic tree from 3 to 5. The new function tries to improve the accuracy of the model by using more branches and more classifications. However, based on the final mse value, it can be seen that a higher max depth value does not lower the mse value but raises it to about 339423612. This suggests that increasing the amount of max depth is not a good way to improve the accuracy of the model here.

The final two functions attempt to change the value of λ in boosting. And after continuously scaling up and scaling down the value of λ we find that the mse value obtained by the model increases when the value of λ is raised and when the value of λ is adjusted to 0.001 we get the mse value which is very similar to the best mse. Thus boosting can effectively improve the model accuracy and achieve the best mse value when the λ value reaches 0.001.

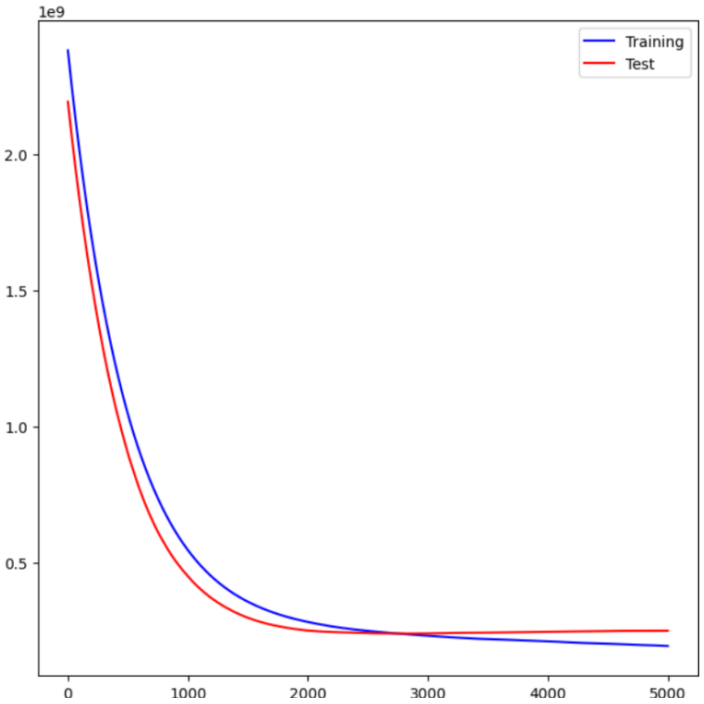


Fig. 5 Fitting Regression Model

4. Discussion

The results from the RESULTS show that with age up to 40 and years of experience of 8.5 are the two key values that determine whether or not an employee's salary reaches the high salary level that we have set. This result shows that the current economic market favours older employees and employees with more work experience, which is in line with the screening mechanism of senior talents, but is not conducive to the vitality and circulation of the economic market, and is prone to vicious competition for ineffective cumulative work experience, which leads to the solidification of the market's economic structure. If the market can allocate more wages to young employees with excellent skills, it may help to avoid people blindly accumulating more work experience.

5. Conclusion

In this paper, multiple-linear regression and tree-based train are used to analyse the salaries and age and salaries and years of experience of the employed persons respectively. The results of the multiple-linear regression test show that there is no clear linear relationship between either age or years of experience and salary. The results of tree-based train show that higher age or years of experience corresponds to larger salary in the overall scope. The results of this paper are beneficial to the understanding of the salary of the employed in the epidemic, but the selection of only two influencing factors and the data of only one year of the employed limit the research of this paper.

References

[1] Henderson, Lisa. Applied Clinical Trials. Salary Survey: The Age of COVID. MultiMedia Healthcare Inc.2022:12-15

[2] Bansal, U; Narang, A; Sachdeva, A; Kashyap, I. Empirical analysis of regression techniques by house price and salary prediction. IOP Publishing.2021

[3] Matbouli, Yasser T; Alghamdi, Suliman M. Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations. MDPI AG.2022

[4] El-Rayes, Nesreen; Fang, Ming; Smith, Michael; Taylor, Stephen M. International Journal of Organizational Analysis. Predicting employee attrition using tree-based models.Emerald Group Publishing Limited.2020

[5] Wu, Jibo; Liu, Chaolin. The best linear unbiased estimator in a singular linear regression model. Springer Nature B.V.2018

[6] A. Mayr, H. Binder, O. Gefeller , M. Schmid. The Evolution of Boosting Algorithms From Machine Learning to Statistical Modelling. Thieme.2014

[7] Wu Zhibin, Li Nianping, et al. Using an ensemble machine learning methodology-Bagging to predict occupants' thermal comfort in buildings. ScienceDirect. 2018:117-127

[8] Yaman, Emine; Subasi, Abdulhamit. BioMed Research International. Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification. Hindawi Limited.2019

[9] Leo Lebanov a b, Laura Tedone a, Alireza Ghiasvand a c, Brett Paull. Random Forests machine learning applied to gas chromatography – Mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils. 2020

[10] L. Hu, Z.G. Zhang, A. Mouraux G.D. Iannetti. Multiple linear regression to estimate time-frequency electrophysiological responses in single trials. ScienceDirect. 2015:442-453

[11] Gareth James, Daniela Witten , Trevor Hastie , Robert Tibshirani. An Introduction to Statistical Learning.Springer New York, NY. 2013