Regression Vs Tree Based Model

Md Mominul Islam

2/3/2022

Table of Contents

QUESTION: ORDINARY LEAST SQUARES REGRESSION VS TREES

Source: Bierens, H.J., and D. Ginther (2001): "Integrated Conditional Moment Testing of Quantile Regression Models", Empirical Economics 26, 307-324

Objective: Understand and compare how error is measured within the context of OLS and Tree regression when the dependent (target) variable is a continuous number.

Using the US Wages data, we have compared tree based model and OLS regression model and created the following-

Used Summary statistics and exploratory plots of wage, education, and experience that show a good profile of the distribution of the values as well as the relationships among the variables.

A linear regression model using lm() with wage as dependent (target) variable and education and experience as independent (predictor) variables.

A tree model using rpart() with wage as dependent variable and education and experience as independent variables.

A set of ‘scored’ data. We created predicted values for each of the two models for all data in the model dataset and appended these two new columns to the dataset and export this to a file either csv or Excel.

Calculated the SSE (sum of squared error) for both the OLS and the tree model. At the end, we decided which model is a better 'fit'.

Answer:

The US wages dataset, with its 2000 rows and 10 columns, was utilized in this study. Samples from the 1988 Population Survey were used to calculate the weekly wages for male employees in the US. All the dataset's predictors were utilized in this study to see whether they could be used to predict individual weekly salaries.

Summary Statistics

At first, we have the summary statistics for the US wage data.

Parameters	Wage	Education	Experience
Minimum Value	50.39	0.00	-2.00
First Quartile	308.64	12.00	8.00
Median	522.32	12.00	15.00
Mean	608.12	13.11	18.41
Third Quartile	783.48	16.00	27.00
Maximum Value	7716.05	18.00	59.00

- From the output, we can infer that the average wage of the people in the United States is USD 608.12, the average education is 13.11 years, and the average experience is 18.41.

- The middle most value of a variable in a data is its median value. From the output depicted in the table, we can infer that the median wage of the people in the United States is USD 522.32 which is off from the mean value, the median education is 12 years, and the median experience is 15.
- Skewness is a measure of symmetry, or the lack of it, for a real-valued random variable about its mean. The skewness value can be positive, negative, or undefined. In a perfectly symmetrical distribution, the mean, median, and the mode will all have the same value. However, the variables in our data are not symmetrical, resulting in different values of the central tendency.

Wage	Education	Experience
3.7872164	-0.6967118	0.6568895

Highly skewed distribution: Wage value is highly skewed as greater than +1.

Moderately skewed distribution: Education and Experience are moderately skewed.

Correlation Matrix

In statistics, we're often interested in understanding the relationship between two variables.

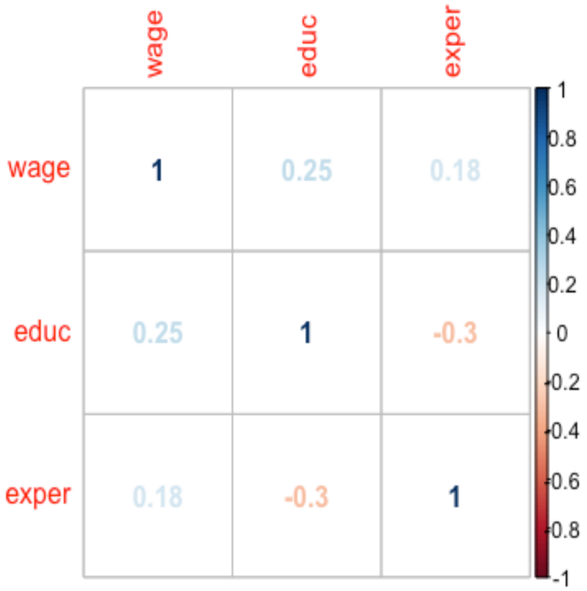
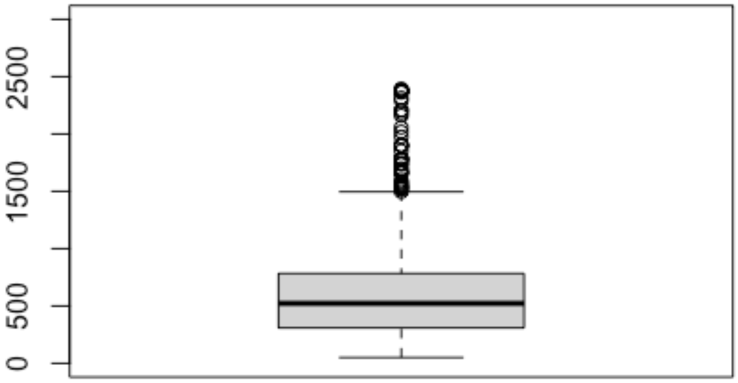


Figure: Correlation Matrix

From the matrix above, we can infer that wage and education are highly correlated compared to experience and wage.

Exploratory Data Analysis

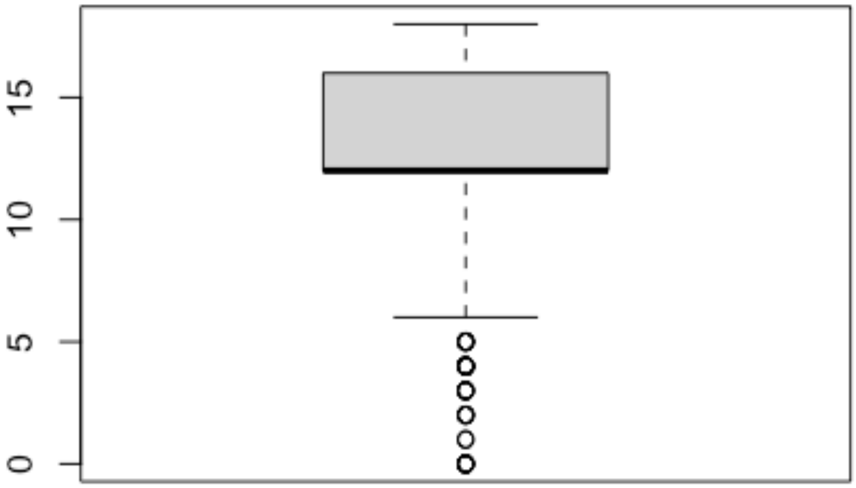
Boxplot: A boxplot is a standardized way of displaying the distribution of data based on a five number summary which can tell us about our outliers and what their values are. Moreover, it can also tell us either the data is symmetrical or not.



Boxplot for US Wages

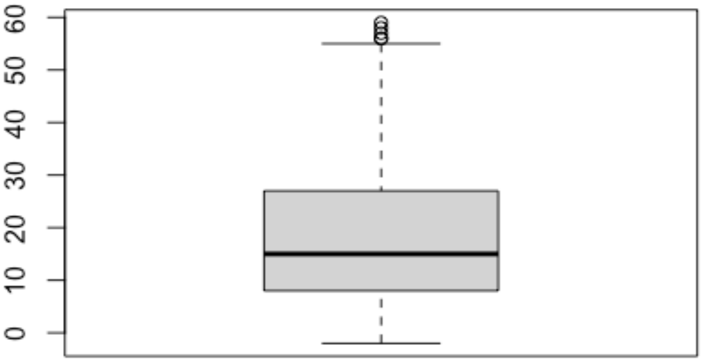
Interpretation of the box plot (alternatively box and whisker plot) rests in understanding that it provides a graphical representation of a five number summary.

The box encompasses 50% of the observations. So, we can say that 50% of the people have wage in between USD 0-1000. There are some extreme values in our data set.



Boxplot for Education

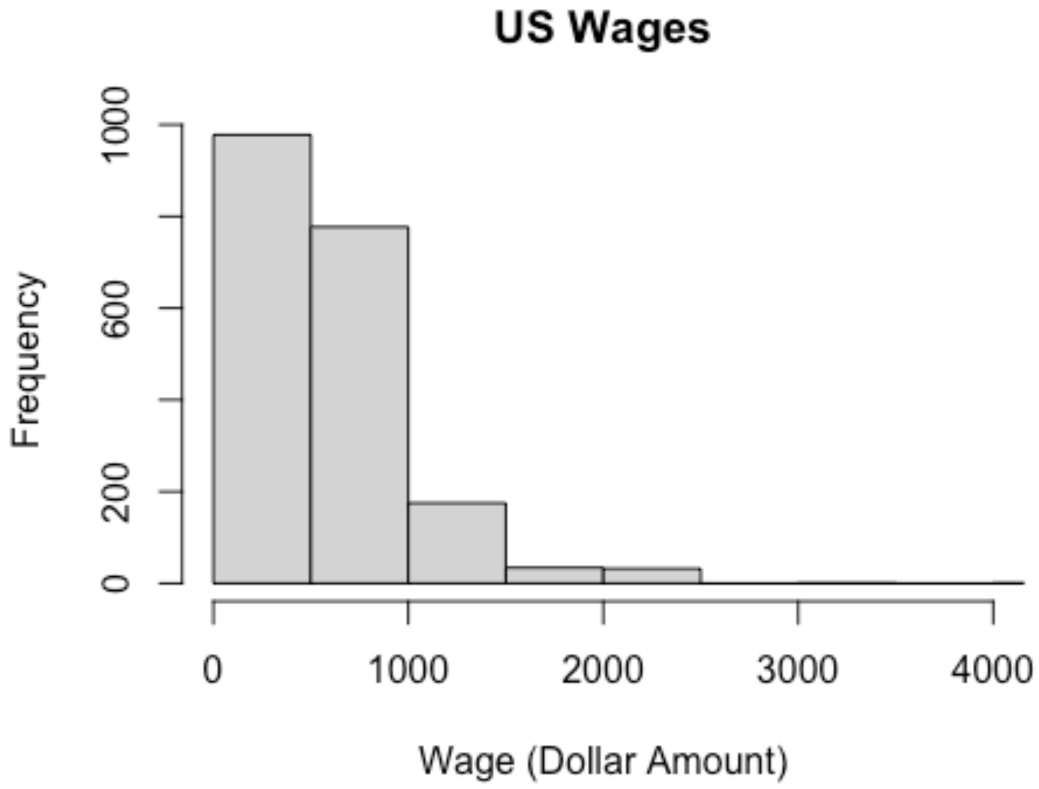
If we look at the boxplot for education, most of the data are beyond the median value which defines data is skewed.



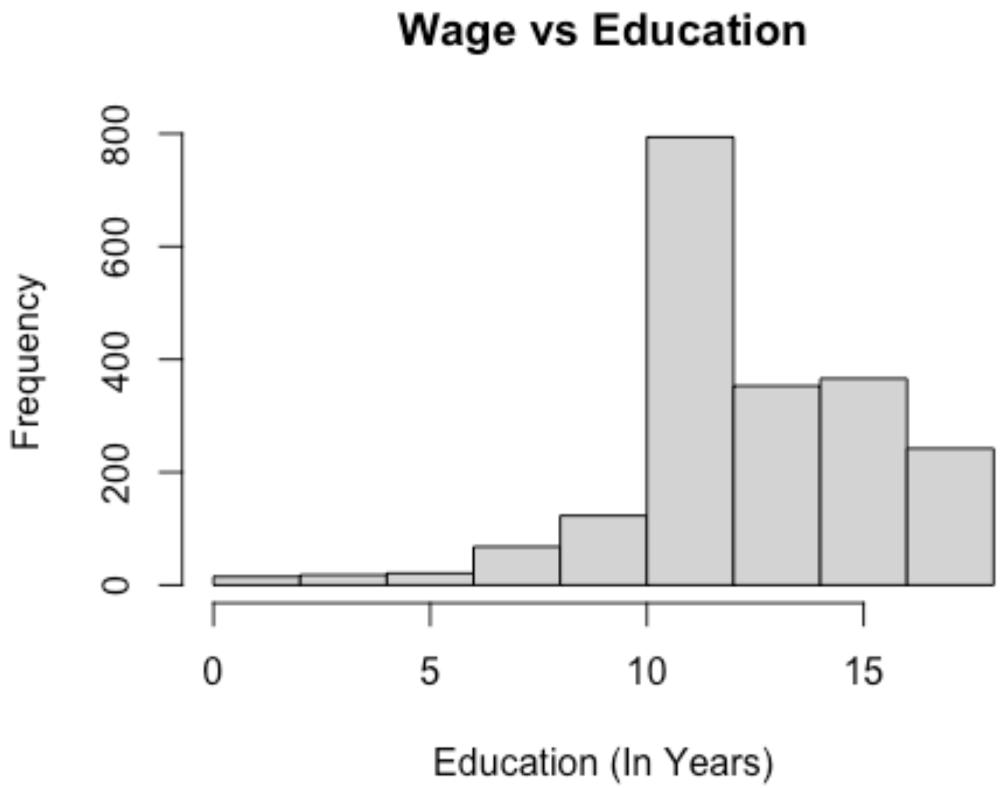
Boxplot for Experience

If we look at the boxplot for experience, most of the data are beyond the median value which defines data is skewed with presence of extreme values in them.

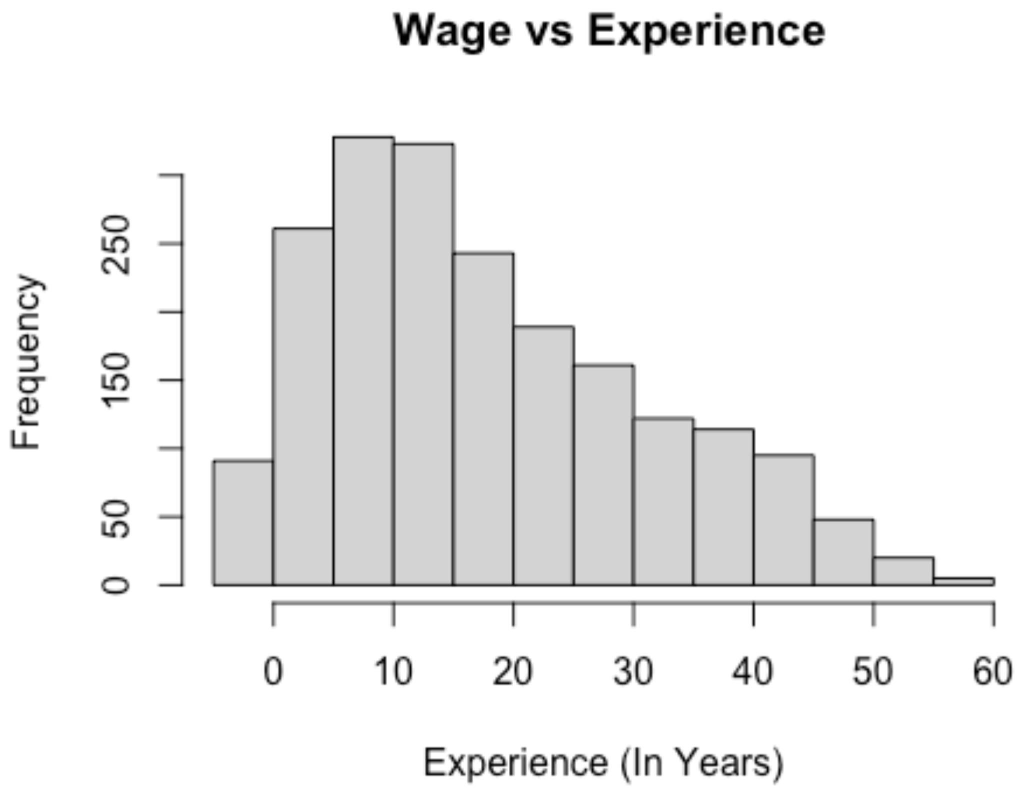
Histogram



If we look at the histogram for the wages, we can see similar scenario as we have seen before. Most of the data are on the left side of the histogram which can be inferred as data is right skewed.

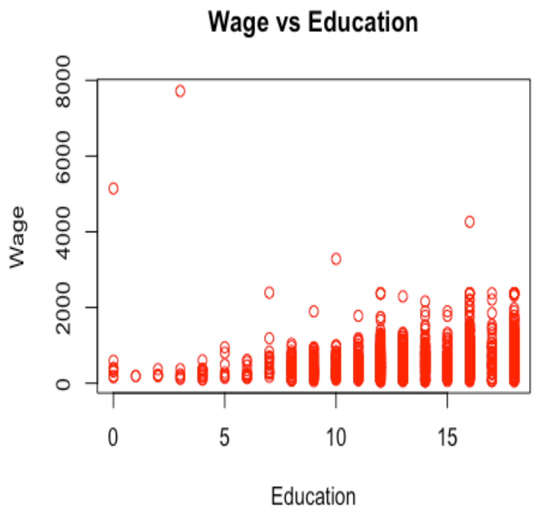


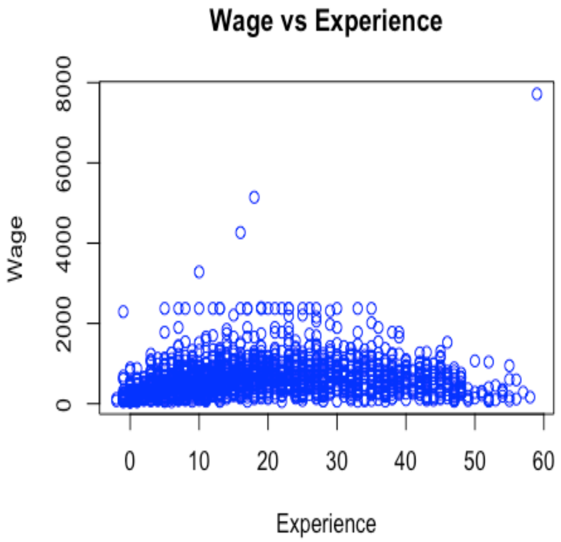
If we look at the histogram for the education, we can see similar scenario as we have seen before. Most of the data are on the right side of the histogram which can be inferred as data is left skewed.



If we look at the histogram for the experience, the data is moderately skewed to the right.

Scatterplot





Interpretation: Significantly positive correlation exist between weekly wages and years of education as an increase in years of education is observed to cause an increase in an individual's weekly wages. However, Years of experience appears not having significant impact on weekly wages. Most individuals with the lowest weekly wages have the lowest years of education but highest years of experience. Individuals with significant years of education have high weekly wages even though they have lower years of experience.

Regression Model

The general form of linear regression model is defined as:

$Y = \beta_0 + \beta_1 X_1 \ldots\ldots + \beta_n X_n + \epsilon$

here,

β_0 is the intercept

β_1, β_n are the co-efficient of predictor variables

$x_1, x_2 \ldots x_n$ are the predictor variables

Y is the target variable

∈ is the error term,

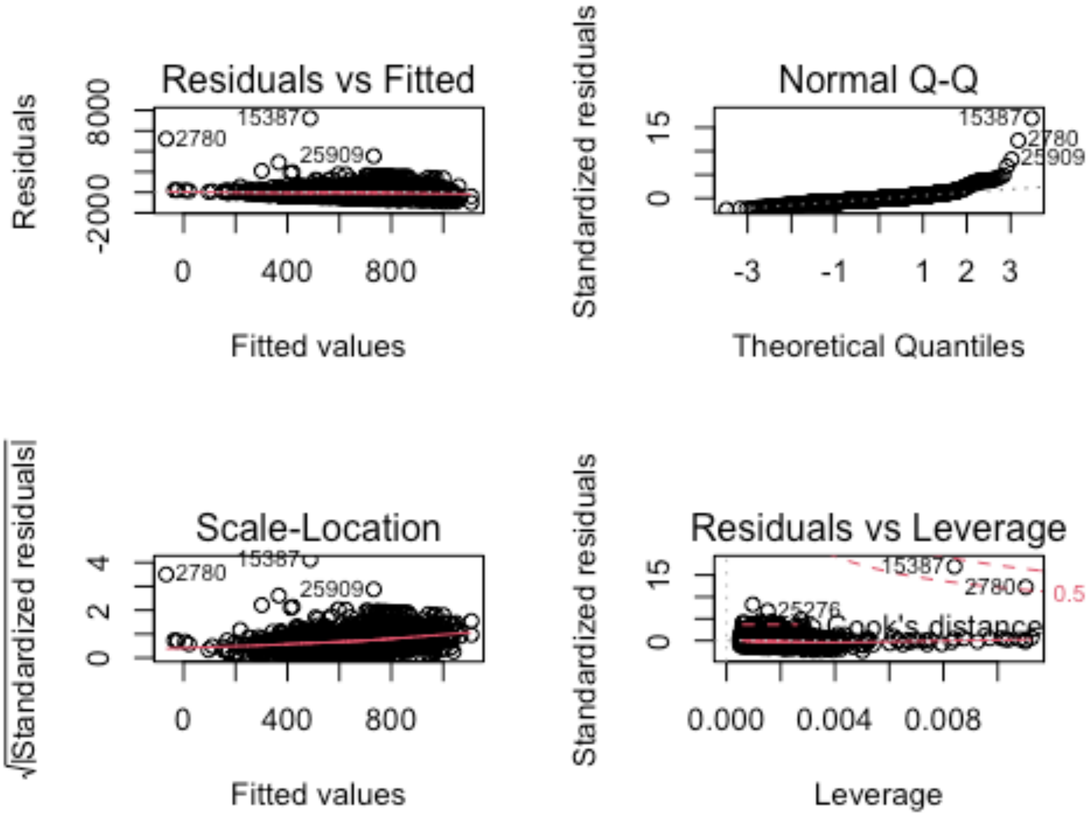
Our regression model is defined in the following way:

$$\mathbf{wage} \sim -\mathbf{242.799412} + \mathbf{51.175268(educ)} + \mathbf{9.774767(exper)} + \boldsymbol{\epsilon}$$

	Estimate	**Std. Error**	**t-value**	**Pr(>|t|)**
Intercept	-242.7994	50.6816	-4.791	1.78e-06 ***
Education	51.1753	3.3419	15.313	2e-16 ***
Experience	9.7748	0.7506	13.023	2e-16 ***

Initially to build our model, we have selected wage as Target variable, and education and experience as predictor variables. After modeling, we found education and experience are significant in our data set with p-value less than the $\alpha = 0.05$.

From our regression model, we calculated SSE and the value is 365568644.



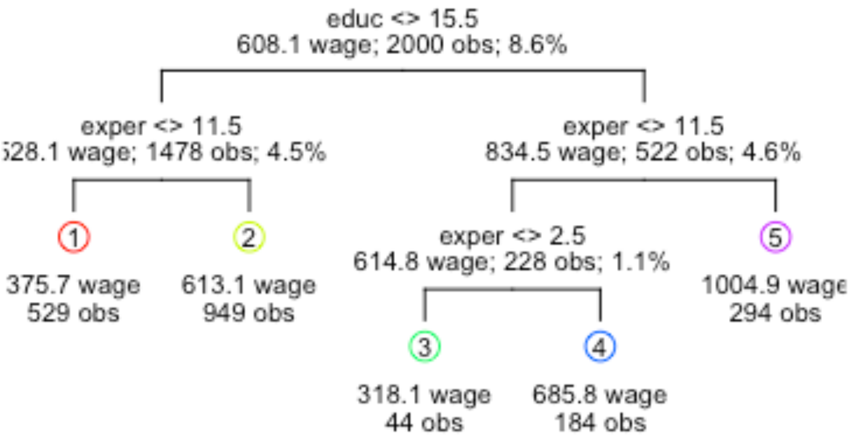
1. Residuals vs Fitted: This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable, and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. Here, we found equally spread residuals around a horizontal line without distinct patterns, that is a good indication we don't have non-linear relationships.

2. Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. We can tell the type of distribution using the power of the Q-Q plot just by looking at the plot. From QQ plot, we can say that data is normally distributed with some outliers like 15387, 2780 and 25909.

3. This plot helps us to find influential cases if any. Not all outliers are influential in linear regression analysis (whatever outliers mean). Here, outliers like 15387and 2780 are not influential in our model.

Tree Based Model

Wage, Education and Experience are the dependent and independent variables in a tree regression model in R-Studio. From our tree-based model, the education variable is contained in the top node, which is divided into two nodes initially. The break between these two nodes occurred once again, this time into two nodes. We end up with a total of seven nodes at the conclusion of the process. The tree diagram of the tree regression model is seen in the image below.



The sum of square error (SSE) measures the overall difference between actual data and the values predicted by an estimation model. Sum of square error for a model is calculated as:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

From our two models, we can compare our SSE and the value was

Model	SSE
Regression Model	365568644
Tree Model	342982890

Conclusion

Given the factors in the dataset, the tree model, with an SSE of 34298289, was superior at predicting weekly earnings for individuals. SSE 365568644 of OLS linear regression was equivalent to this model's performance.

END