2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) | 978-1-6654-9246-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/MLISE57402.2022.00081

Employee Salaries Analysis and Prediction with Machine Learning

Guanqi Wang
School of Computing, Faculty of Engineering
University of Leeds
Leeds, United Kingdom
sc20gw@leeds.ac.uk

***Abstract*—The starting point of this article is to find a suitable method of salary prediction to find a job. Firstly, this paper will introduce the content and usage of different regression models in machine learning. After understanding the methodology that will be used, it is pointed out that the goal of this study is to find the correlation between the salaries of employees and different influencing factors, and to use regression model to predict the salary and screen out the most effective method. In the following experiment, the correlation table between salary and influencing factors was drawn to show which factors had stronger correlation. Secondly, the R2 and RMSE values are used to analyze the results of 5 different regression models (including Multiple Linear Regression, Ridge Linear Regression, Elastic-Net Regression, Lasso Linear Regression and Polynomial Regression) to select a model with the best performance. The results show that graduates can give priority to further improve their academic background before seeking jobs, and it is more suitable to use polynomial regression model to solve this problem.**

Keywords-Machine learning; Linear regression; Polynomial regression; Salary prediction

I. INTRODUCTION

When we college students graduate from university, we will be faced with the problem of finding a job. It is more effective for graduates to evaluate themselves and demand a better fit based on the salary offered by these jobs. Therefore, it is important for graduates to find a way to predict what salary they will receive and then compare the predicted salary with the actual salary offered, which can tell them more about themselves and increase the likelihood of a successful job search.

Recently, machine learning has made great progress in various tasks such as regression analysis. Regression prediction model plays a very good role in time series prediction task. Regression analysis is a set of statistical technique for estimating the relationship among dependent variables. In statistical models, variant variables have reasonable relation and correlatively influence the result. The focus of univariate regression is analyzing the relationship between dependent variables and independent variables, which always try to derive the linear relationship equation among different variables. Another different regression model which solves the relationship between one dependent variable and more than one independent variable is called multilinear regression, is more common in mathematical problems.

Regression models which use a single independent variable is called univariate regression analysis, while another regression models that use multiple independent variables is called multivariate regression analysis [1]. The relations between a dependent variable and an independent variable are analyzed when using univariate regression analysis, finally we can get an equation which represent the linear relations among dependent and the independent variables. To sum up, the regression models which only controlled by one variable is called single regression analysis and the models with more than one independent variable, however, is known as multivariate regression analysis.

In this study, we are trying to use regression algorithms of machine learning to get the prediction of a specific person's salary through analyzing his or her personal information. Our objectives are exploring the relationship between salary and different factors and predicting one specific employee's salary through machine learning algorithm. In order to get a better salary's forecast, this paper will include several different regression models to multidimensional analyze the data set. Different algorithms will have different performance for a particular data set, so our methods compare the performance of each regression models and find out which algorithm is best for salary forecasting by comparison.

II. RELATED WORK

A. Employee Salaries Analysis and Prediction

Salary benchmarking gives an impartial idea of competitive salaries and allows organizations to make informed decisions. Salary benchmarks provide data points, whether it is worth it or not to pay an employee above the average salary. It also helps understand the holistic remuneration packages offered by employers. Recently, more and more people are paying attention to whether the salary benchmark matches their internship and mathematical methods make great progress in analysis and prediction tasks. Hence, machine learning methods can use to deal with the problem of employee salaries analysis and prediction.

B. Machine learning and Regression

Machine learning methods nowadays are more and more widely used in various fields to solve difficult problems which are too difficult to deal with based on other simple computer methods. Machine learning algorithm mainly includes supervised learning algorithm and unsupervised learning algorithm. The linear regression is one of the simplest and most

978-1-6654-9246-1/22/$31.00 ©2022 IEEE
DOI 10.1109/MLISE57402.2022.00081

common supervised machine learning algorithms which needs to give supervisory signals to fit the trend of the diversity variables. Linear regression can be used for projection of continuous algebraic variables and predictive analysis by mathematical methods. Galton [2] first suggested the concept of linear regression in 1896 and from then on, regression models developed gradually. Linear regression is a mathematical evaluation method, mainly used to evaluate and quantify the relationship between various variables involved in the problem. In fact, univariate regression analyses are useful in data analysis, but they cannot be utilised to take into account the results of other factors in the analysis. As a result, partial correlation and regression are tests that allow scientists to examine the impact of confusions by analysing the relationship between two variables [3]. Linear regression [4] is a typical mathematical research tool that allows you to quantify and simulate expected effects using numerous input variables. It is a data analysis and modelling technique that develops linear relationships between dependent and independent variables.

C. Regression Models

According to the regression model, the dependent variable [5] was predicted by fitting the given independent variable. The value of dependent variable "y" is estimated by regression analysis through the range of a large number of independent variable values "x". Regression [6] can be simple linear regression or multiple regression. This paper mainly discusses various methods of multiple linear regression and polynomial regression to find the most suitable prediction model.

D. Multivariate Linear Regression (MLR)

MLR is a statistical method that uses multiple explanatory variables to predict the outcome of a response variable. The goal of MLR is to model the linear relationship between independent variable X and dependent variable Y and analyze it as [7]. The basic model of MLR is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

And we translate the formula to the formula matrix as:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

The parameters can be fitted and solved by matrix operation and inverse operation.

Other linear regression models. Ridge regression [8] is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. When linear regression models have highly correlated independent variables, we need to create a ridge regression estimator (RR), as its variance and mean square estimator are often smaller than the least square estimators previously derived. Compare to the simple regression models, the ridge regression estimator is shown as following:

$$\hat{\beta}_{ridge} = \left(X^T X + k I_p\right)^{-1} X^T y$$

Lasso regression is another regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model. The lasso regression model is easily extended to other statistical models such as generalised linear models, generalised estimating equations, and proportional hazards models [4] by relying on the form of the constraint and having a number of interpretations. The cost function for Lasso (least absolute shrinkage and selection operator) regression can be written as:

$$\sum_{i=1}^{M} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{M} (y_i - \sum_{j=0}^{p} w_j \times x_{ij})^2 + \lambda \sum_{j=1}^{p} |w_j|$$

E. Polynomial regression

Polynomial Regression [9] is a type of linear regression that estimates the relationship as an nth degree polynomial. It is a special case of Multiple Linear Regression. Because Polynomial Regression is sensitive to outliers, the existence of one or two of them can have a negative impact on the results. When the connection between the data is linear, the simple linear regression procedure works. However, if we have non-linear data, linear regression will be unable to create a best-fit line and will fail in this case. There is a lot of nonlinear connection data in real problems that linear regression models can't fit well. To solve this problem, we use polynomial regression, which identifies the curvilinear relationship between independent and dependent variables.

III. METHOD

In order to deal with this analysis and prediction task, we are trying to use the regression model. We compare the performance of different models including Multivariate Linear Regression, ridge linear regression, lasso linear regression, Elastic-Net regression and polynomial regression. Specially, the last model is non-linear model which may be more suitable for the salary prediction problem. And we design specific optimization objectives based on the task we need to solve.

IV. OBJECTIVES

A. Relationship between salary and different factors

There are several distinct factors that may have an influence on the salary including gender, age, and their education background (whether they get PhD degree or not). The first step is to investigate the correlation between these factors and employee salaries.

B. Predict one specific employee's salary through machine learning algorithm

Linear regression is an algorithm which based on supervised learning model. [10] It shows that a dependent variable (y) is determined by an independent variable (x). And multiple linear regression is based on the simple linear regression. However, it contains more than one explanatory variable which can directly influence the response variable. [11] Because there are there different influencing factors in this dataset, we will need to use this multiple linear regression model to do the prediction.

V. EXPERIMENTS

As we design the special objectives and choose models which needed to compared, we need to choose suitable datasets with experiments, so that we can build a best regression model to predict the employee salaries. We design the process of

experiments and evaluate the models with the metrics like R2, RMSE, and so on.

A. Datasets

We now compare the performance of different regression models with the employee salaries datasets. It's important for graduates to find a way to predict what salary they will receive and then compare the predicted salary with the actual salary offered, which can tell them more about themselves and increase the likelihood of a successful job search, so we choose a dataset closely related to the salary of the job. The data consists of Salaries of Employees, including their gender, age & PHD degree. The dataset is downloaded from UCI Machine Learning Repository. The datasets have 100 instances with four different class attributions. The dataset has very limited features and samples, however, we build regression models to capture all the patterns in the dataset, also maintain the generalizability of the model in our experiments.

1) Data Preparation

After examining the dataset, I observe that there are 100 samples with 4 features, among which Gender and PhD are categorical features, which are already encoded (for the PhD feature, 1 denotes the employee has a PhD degree; for the Gender feature, the number of male and female employees is exactly the same). While Age and Salary are numerical features.

TABLE 1. THE ATTRIBUTIONS AND FEATURES OF THE DATASET

	Salary	Gender	Age	PhD
0	140.0	1	47	1
1	30.0	0	65	1
2	35.1	0	56	0
3	30.0	1	23	0
4	80.0	0	53	1

2) Data exploration

This table shows the statistics of this salary dataset. Besides, we explore the distribution of the salary data by plotting a histogram of the target variable. Through the histogram we can learn that the salary data are normally distributed, with a mean of approximately 50. Similarly, we can also learn the distribution of the age feature (whose mean=46.9, Standard deviation=15.3). Also, the analysis shows that 40% of employees are PhDs.

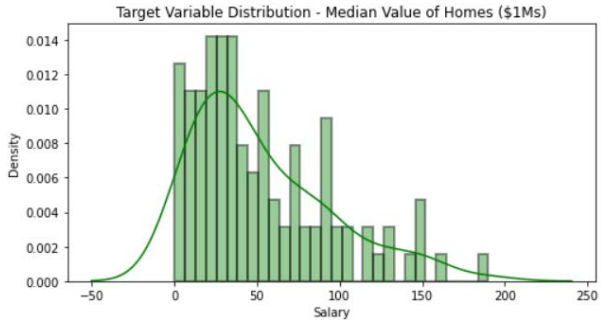

Figure 1. Target Variable (salary data) Distribution

B. Data preprocessing

1) Data preprocessing

Before the train-test split, I check if there are any duplicates in the dataset. If there are several empty elements, we can convert categorical columns to numeric and find that there is a empty elements. Therefore, we drop the outlier data and get the final processed data. Since categorical features are already encoded, and there are no null values, no further preprocessing is needed. Next, 80 % of the data is used for training and 20 % is used for testing. Following that, we standardize features by removing the mean and scaling to unit variance. (Because features with different units do not contribute equally to the analysis and might end up creating a bias.)

2) Understating the correlations between the features

Intuitively, we know that an employee with a PhD degree is likely to have a higher salary. Besides, different gender and age also influence the level of salary, thus we need to understand the correlations between the features. I use this correlation matrix to show the pairwise correlation of features. This matrix shows that salary is indeed highly correlated to this feature. Age feature also has a positive correlation with the target value because senior employees tend to have higher income. On the other hand, there is a low correlation between gender and salary.

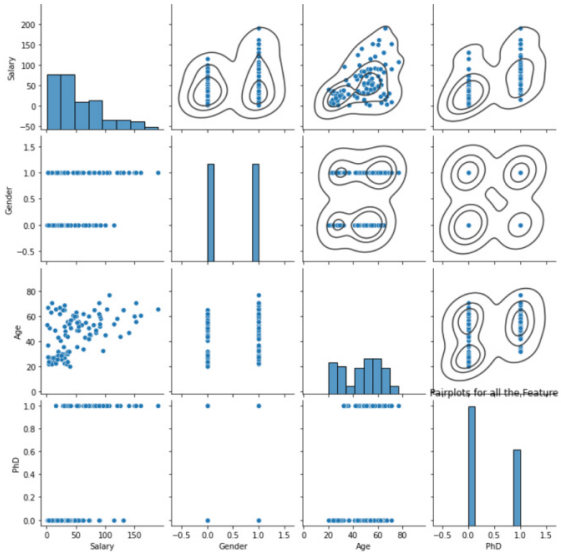

Figure 2. The correlation of between various features

3) Data manipulation

When we split the data into training and testing, we can get the 80% of the training data and 20% testing data while we need to standard the data. With the features scaling, we get the final standardization of the data as the figure shown.

TABLE 2. STANDARDIZATION ON TRAINING SET

	Gender	Age	PhD
count	7.900000e+01	7.900000e+01	7.900000e+01
mean	-3.372829e-17	-7.149696e-17	-2.810691e-17
std	1.006390e+00	1.006390e+00	1.006390e+00

	Gender	Age	PhD
min	-9.627197e-01	-1.658134e+00	-8.469896e-01
25%	-9.627197e-01	-1.049611e+00	-8.469896e-01
50%	-9.627197e-01	1.994625e-01	-8.469896e-01
75%	1.038724e+00	9.040681e-01	1.180652e+00
max	1.038724e+00	1.993004e+00	1.180652e+00

TABLE 3. STANDARDIZATION ON TESTING SET

	Gender	Age	PhD
count	20.000000	20.000000	20.000000
mean	0.138074	0.257112	-0.238697
std	1.021572	0.837216	0.953321
min	-0.962720	-1.209748	-0.846990
25%	-0.962720	-0.328992	-0.846990
50%	1.038724	0.519738	-0.846990
75%	1.038724	0.856027	1.180652
max	1.038724	1.608674	1.180652

C. *Feature Extraction*

When we get the features of the data, we first check the correlation of different features based on correlation matrix as the figure shown below. There seems to be strong multi-correlation between the features and we try to fix these features and find the key factor of salary [12].

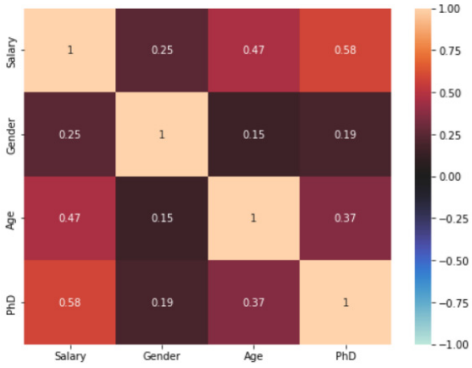

Figure 3. The correlation matrix of features

We can fix these multicollinearities with three techniques: Manual Method - Variance Inflation Factor (VIF) [13], Automatic Method - Recursive Feature Elimination (RFE) [14], Feature Elimination using PCA Decomposition [15]. We show the variation tendency of the explained variance of components and explained variance as we use PCA decomposition methods and find the key factor of salary increase.

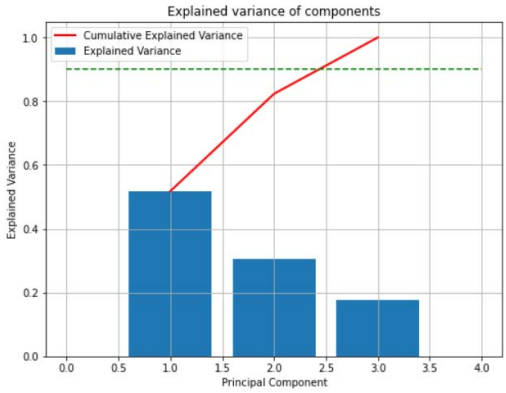

Figure 4. The variation tendency of explained variance

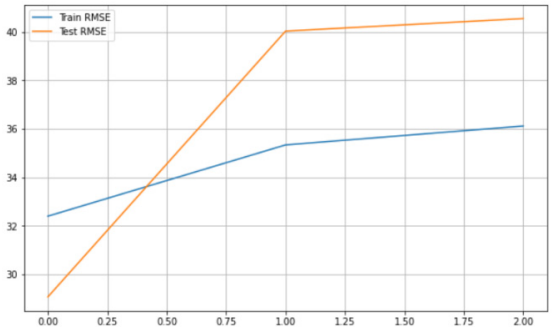

Figure 5. The performance of the model dropped with the feature

We drop two features by above methods and compare the RMSE with the complete one. It can be seen that the performance of the models is quietly comparable upon dropping features using VIF, RFE & PCA Techniques. Comparing the RMSE plots, the optimal values were found for dropping most features using manual RFE Technique.

VI. RESULTS ANALYSIS

In order to get better predictions of salaries, this paper will include several distinct regression models to multidimensionally analyze the dataset. For a specific dataset, different algorithms will have different performance, so the following step is to solve which algorithm is most fitted to the salary prediction after comparison. Let us now try building multiple regression models and compare their evaluation metrics to choose the best fit model both training and testing sets. We compare the performance of different models including multivariate linear regression, ridge linear regression, lasso linear regression, elastic-net regression and polynomial regression. Finally, we use the metrics of R2 scores and RSME to evaluate the performance of different models and choose 2nd order polynomial regression as it gives the optimal training & testing scores

A. *Simple Linear Regression*

Linear regression is an algorithm which based on supervised learning model. [10] It shows that a dependent variable (y) is determined by an independent variable (x). In this way, once we get the parameter of the independent variable (x), the dependent variable (y) can be predicted by computing $y_i = \beta_o + \beta_1 x$. [16] And this is called Simple Linear

Regression. Using simple linear regression methods, we can easily find that it can't fit well for the salary data. Therefore, we are trying to use multiple linear regression and other more complex methods to predict the trend of the salary.

B. Multiple Linear Regression

Multiple linear regression is based on the simple linear regression. However, it contains more than one explanatory variable which can directly influence the response variable. [17] The formula can be concluded as below:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

There are three different explanatory variables in this model. Through the process of salary data to be predicted, we can get three coefficients respectively. The rest thing is to use MLR.intercept_ to get the intercept to draw the plot. And we can get coefficient of the multiple linear regression is *[6.00966484 15.37959586 16.08049821]*.

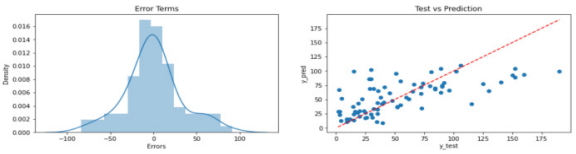

Figure 6. The error of MLR and MSE of prediction with test data

C. Ridge Linear Regression

Ridge regression[8] is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. And we use ridge regression model to fit the data. And we can get the results and find the coefficient of the regression model is [5.97388749 15.2565531 15.94225001]. And the intercept of the regression Model was found to be 53.91392405063291. The following is the difference between the actual data and the prediction data:

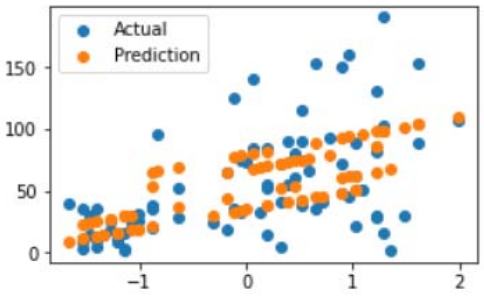

Figure 7. The difference between the ground truth and prediction

D. Lasso Linear Regression

Lasso regression model is easily extended to other statistical models including generalized linear models, generalized estimating equations, proportional hazards models[11]. When we evaluate the performance of the Lasso Regression Model, we also can get the coefficient of the regression model was [5.19043287 14.77073727 15.49812163].

E. Elastic-Net Regression

The cost function of elastic network regression algorithm combines the regularization methods of Lasso regression and ridge regression, and uses two parameters λ and ρ to control the size of penalty term. And the loss function is defined as follow:

$$Loss(w) = \left(\sum_{i=1}^{N} (y_i - w^T x_i)^2 + \frac{\lambda(1-\rho)}{2} \|w\|_2^2 \right)$$

It can be seen that when ρ = 0, its cost function is equivalent to ridge regression, when ρ = 1, its cost function is equivalent to Lasso regression. Just like Lasso regression, there is absolute value in the cost function, which is not differentiable everywhere, so it is impossible to get the analytical solution of W directly by taking the derivative directly, but we can still be solved by using coordinate descent method [18]. And we also evaluate the performance of this method using R2, RSS, MSE and RMSE metrics shown in the Table 4.

TABLE 4. MODEL EVALUATION COMPARISON MATRIX

	Train-R2	Test-R2	Train-RSS	Test-RSS	Train-MSE	Test-MSE	Train-RMSE	Test-RMSE
Multiple Linear Regression (MLR)	0.438976	0.187421	82906.397947	16896.121689	1049.448075	844.806084	32.395186	29.065548
Ridge Linear Regression (RLR)	0.438947	0.188384	82910.692873	16876.106630	1049.502441	843.805332	32.396025	29.048328
Lasso Linear Regression (LLR)	0.437902	0.199770	83065.222504	16639.346979	1051.458513	831.967349	32.426201	28.843844
Elastic-Net Regression (ENR)	0.408930	0.194910	87346.603756	16740.417611	1105.653212	837.020881	33.251364	28.931313
Polynomial Regression (PNR)	0.484207	0.167835	76222.312031	17303.383906	964.839393	865.169195	31.061864	29.413759

F. Polynomial Regression

When we choose polynomial regression to deal with this task, we first need to know about the order of the polynomials. We compare the RMSE based on different orders as Figure 3. And we can choose 2nd order polynomial regression as it gives the optimal training & testing scores.

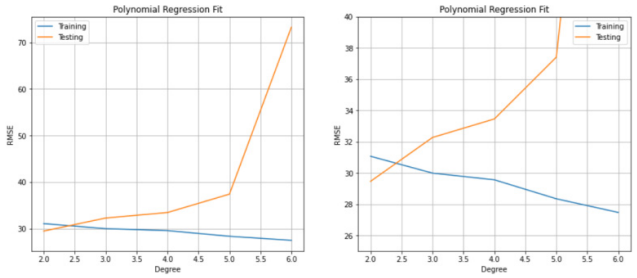

Figure 8. The relationship between RMSE and the Degree

We evaluate each regression models and compare RMSE and R2-Scores to check which one can best fit this prediction. And the following is the table of model evaluation comparison matrix and the histogram comparison of RMSE scores of each model. Besides, as shown at Figure 9, we also compare the R2-Scores of different regression models including linear regression and non-linear regression shown as Figure 10.

Through the comparison of RMSE and R2-Scores, we can find that in Employee Salaries data, polynomial regression is the best regression model to fit this prediction. The main reasons for our analysis are as follows: The factors affecting salaries are complex and changeable, and the traditional linear regression model cannot reflect the changing trend of the curve, so polynomial regression can better predict salaries.

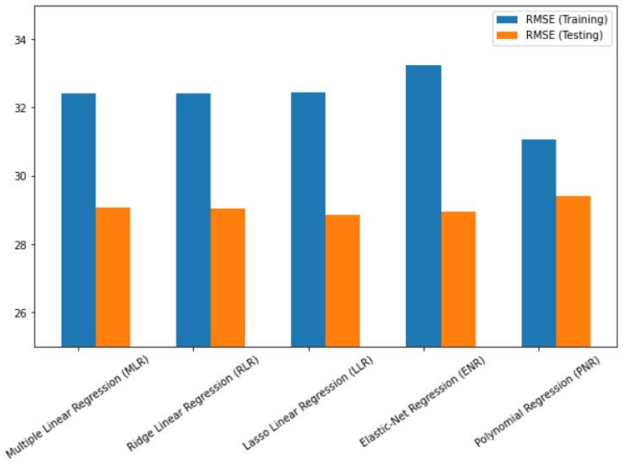

Figure 9. The histogram comparison of RMSE scores of each model

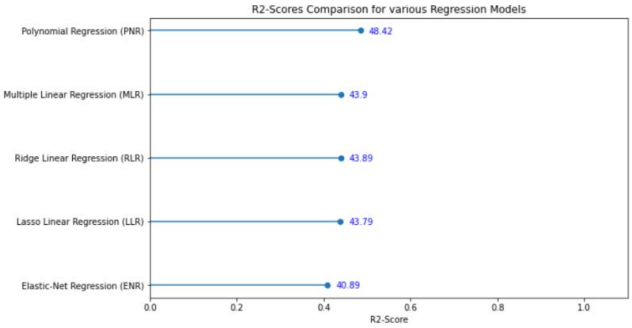

Figure 10. the R2-Score comparison between different regression methods

VII. CONCLUSION

The correlation matrix clearly shows the strong relationship between education background and the salary. So before graduates plan to find a job, it is also a good choice to study further and improve their competitiveness. This will help them get a better work and receive higher salary.

Though the multiple linear regression model performs an imperfect result, it proves the regression model is a useful direction to deal with this dataset. Thus, according to our experiments' results, we can try more different regression methods to find which algorithm could best fit this prediction situation and then we could get a more precise prediction with reference value.

As there is more and more important for graduates to find a way to predict what salary they will receive and then compare the predicted salary with the actual salary offered, we can use machine learning methods to analyze and predict it. Through the experiments, we can easily find that 2nd order polynomial regression model is the best model to fit the salary tendency and predict how many salary graduates will get when they choose different jobs.

REFERENCES

[1] Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). Using multivariate statistics (Vol. 5, pp. 481-498). Boston, MA: pearson.

[2] Galton, F. (1886). Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

[3] Sulaiman, M. A. (2020). Evaluating Data Mining Classification Methods Performance in Internet of Things Applications. Journal of Soft Computing and Data Mining, 1(2), 11-25.

[4] Jiang, Y., He, Y., & Zhang, H. (2016). Variable selection with prior information for generalized linear models via the prior LASSO method. Journal of the American Statistical Association, 111(513), 355-376.

[5] Roopa, H., & Asha, T. (2019). A linear model based on principal component analysis for disease prediction. IEEE Access, 7, 105314-105318.

[6] Kavitha, S., Varuna, S., & Ramya, R. (2016, November). A comparative analysis on linear regression and support vector regression. In 2016 online international conference on green engineering and technologies (IC-GET) (pp. 1-5). IEEE.

[7] Zhang, Z., Li, Y., Li, L., Li, Z., & Liu, S. (2019, May). Multiple linear regression for high efficiency video intra coding. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1832-1836). IEEE.

[8] Hilt, D. E., & Seegrist, D. W. (1977). Ridge, a computer program for calculating ridge regression estimates (Vol. 236). Department of Agriculture, Forest Service, Northeastern Forest Experiment Station.

[9] Ostertagová, E. (2012). Modelling using polynomial regression. Procedia Engineering, 48, 500-506.

[10] Tzanis, G., Katakis, I., Partalas, I., & Vlahavas, I. (2006, July). Modern applications of machine learning. In proceedings of the 1st annual SEERC doctoral student conference–DSC (Vol. 1, No. 1, pp. 1-10).

[11] Tranmer, M., & Elliot, M. (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), 1-5.

[12] Yuan, K. H., & Hayashi, K. (2006). Standard errors in covariance structure models: Asymptotics versus bootstrap. British Journal of Mathematical and Statistical Psychology, 59(2), 397-417.

[13] Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. Quality engineering, 14(3), 391-403.

[14] Chen, X. W., & Jeong, J. C. (2007, December). Enhanced recursive feature elimination. In Sixth International Conference on Machine Learning and Applications (ICMLA 2007) (pp. 429-435). IEEE.

[15] Hu, C., Wang, J., Zheng, C., Xu, S., Zhang, H., Liang, Y., ... & Xu, W. (2013). Raman spectra exploring breast tissues: Comparison of principal component analysis and support vector machine-recursive feature elimination. Medical physics, 40(6Part1), 063501.

[16] Robert Tibshirani., Trevor Hastie., Daniela Witten., & Gareth James. An Introduction to Statistical Learning: With Applications in R.

[17] Tranmer, M., & Elliot, M. (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), 1-5.

[18] Wright, S. J. (2015). Coordinate descent algorithms. Mathematical Programming, 151(1), 3-34.

[19] M, Yasser, H. (2021). Employee Salaries Datatset (Version 1) [Data set]. https://www.kaggle.com/yasserh/employee-salaries-datatset