Salary Prediction Based on the Dual-Adaboosting System

Yanming Chen[1, a*], Kunye Zhan[2, b], Songguang Lin[3, c], Ken Yao[2, d]

{21ymchen@stu.edu.cn[a,*], 2021040486@email.szu.edu.cn[b] , 2065964884@qq.com[c] , 2021020086@email.szu.edu.cn[d] }

Student, Shantou University[1,*], Student, Shenzhen University[2] , Student, Sun Yat-sen University[3]

The authors Kunye Zhan, Songguang Lin and Ken Yao have made equal contributions to this paper and are therefore co-authors as second authors.

Abstract. This paper aims to build a salary prediction model based on the Dual-Adaboosting System which is an improvement method of the Adaboosting algorithm. Adaboosting feature importance ranking method is employed for the paper to conduct feature filtering after the dataset is cleaned and preprocessed. Afterward, regression prediction models were established using different methods to help obtain comparative and analytical results based on RMSE and MAE, serving as a control experiment. Finally, we conducted experiments using the Dual-Adaboosting system and obtained the final salary prediction model. This model definitely has practical reference significance for human resources management and the improvement of the Adaboosting algorithm.

Keywords: Salary Prediction; Adaboosting Regression; Dual-Adaboosting System; Machine learning; Ensemble learning; Randomforest; Feature Filtering

1 Introduction

Salary prediction is a topic of global interest and generates a lot of discussion. Lasso regression, Random forest, Ridge regression, Decision tree, Adaboosting and other machine learning algorithms are commonly employed in salary prediction models. Several researchers have dedicated extensive effort towards studying salary prediction, such as predicting the per capita wages of urban mining units based on grey theory [1], predicting job salaries based on random forest algorithm [2], conducting employment salary forecast via KNN algorithm [3], and application of machine learning in human resource management [4]. However, the widespread used regression prediction algorithms can not break through the lower limit of error when faced with complex salary datasets.

This paper proposes an optimization algorithm called Dual-Adaboosting system and uses this algorithm to predict the salaries of candidates in the data science field from the United States. At the same time, a control experiment is set up to test the rationality and effectiveness of the new algorithm. This method can deal with more complex and varied salary data for salary prediction, and can be better applied to human resource management. Moreover, this new algorithm can be applied to other regression prediction fields.

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334428

2 Theoretical foundation

Machine learning is the scientific field focusing on the ways that machines analyze on historic data and produce meaningful conclusions automatically [5]. Ensemble learning algorithms are one type of machine learning algorithms. Adaboosting is the typical example of Ensemble learning. The Dual-Adaboosting system used in this paper is an optimized innovation based on the Adaboosting regression.

Adaboosting regression is an ensemble method that combines multiple weak regressors. In the process of combination, the training sample weights are adjusted based on the previous round's prediction results and the prediction errors of each weak regressor are weighted. Finally, a more accurate strong regressor is obtained through iteration [6]. The flowchart of the Adaboosting regression is shown in **Figure 1.**

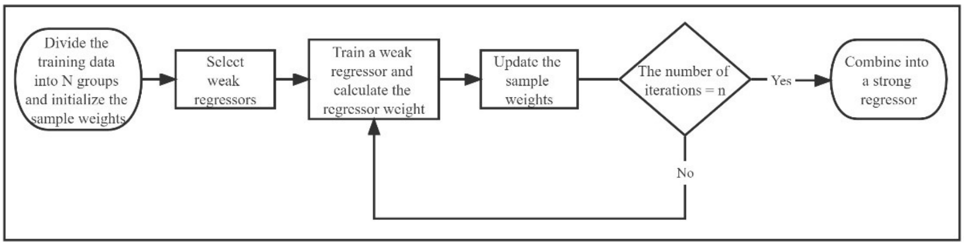
Divide the training data into N groups and initialize the sample weights
Select weak regressors
Train a weak regressor and calculate the regressor weight
Update the sample weights
The number of iterations = n
Yes
Combine into a strong regressor
No

Fig. 1. The flowchart of the Adaboosting regression

3 Materials and Methods

3.1 Dataset used in the study

Since America is a large economy deeply involved in globalization, and its salary levels are largely correlated with international salary levels, this paper hence selects the salary in the US data science field as the core research subject, and dataset records were obtained from kaggle.com.

The dataset contains a total of 742 observations, with each observation representing an annual salary (in thousands of US dollars) corresponding to multiple columns of candidates' information. The salary range in the dataset shows a difference from \$13.5k to \$254k and the salary distribution is shown in **Figure 2**.

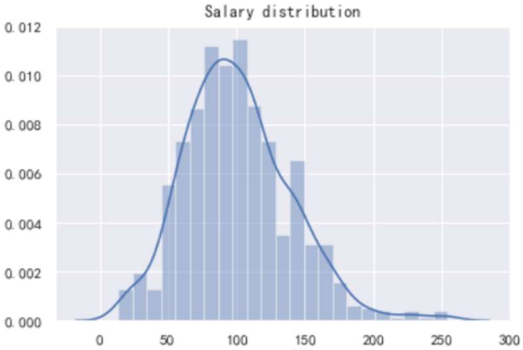

Fig. 2. Salary distribution of the dataset

Based on the information of the dataset, a salary prediction model can be established. There are 28 variables in this dataset. Some variables are only used for explanatory purposes or have no substantive meaning, so they are not included in the scope of the model calculation. After removing these variables, there are 18 variables left in the dataset. Therefore, the variable named 'salary' serves as the dependent variable, and the other 17 variables are used as independent variables. The independent variables in this dataset are a mix of numerical and categorical variables, and the distribution of the numerical and categorical variables is shown in **Table 1.**

Table 1. The proportion of two categories of variables

Type	Number	Proportion
Categorical	8	47.06%
Numerical	9	52.94%

The basic information of numerical and categorical variables is shown in **Table 2** and **Table 3**, respectively.

Table 2. Basic information about numerical variables

	Count	Min	Max	Mean
Rating	742	-1	5	3.62
Company age	742	-1	276	46.59
Python	742	0	1	0.53
Spark	742	0	1	0.23
Aws	742	0	1	0.24
Excel	742	0	1	0.52
Same state	742	0	1	0.56
Hourly wage system	742	0	1	0.03
Provided insurance	742	0	1	0.02

Table 3. Basic information about categorical variables

	Count	Unique	Top	Freq
Job title	742	264	Data scientist	131
Job location	742	200	New York	55
Headquarters	742	198	New York	52
Company size	742	9	1001 to 5000	150
Type of ownership	742	11	Private company	410
Industry	742	60	Biotech	112
Sector	742	25	Information technology	180
Company revenue	742	14	$10+ billion	124

3.2 Methods

In this paper, variables are first divided into numerical variables and categorical variables. After data cleaning, numerical variables are scaled and categorical variables are transformed into multiple dummy variables. Then, Adaboosting feature importance ranking is used for feature filtering to reduce the risk of overfitting. Afterward, various methods such as random forest regression, decision tree, and ridge regression are used to build regression prediction models, serving as control models. Finally, the Dual-Adaboosting method is used to establish the final salary prediction model.

Data cleaning and preprocessing.

Processing of missing values and outliers

There are no missing values in this dataset, but there are some outliers, such as the appearance of -1 for company age. For numerical variables with outliers, we fill them with mean values, while for categorical variables with outliers, we fill them with mode values.

Processing of numerical variables

In order to reduce data errors and improve algorithm performance, the "min-max rescaling" method is used to perform linear transformation on the variables "Rating" and "Company age" to map their feature values to the interval [0, 1]. The formula for "min-max rescaling" is as follows (1) :

$$X' = \frac{X-Xmin}{Xmax-Xmin} \tag{1}$$

Where X is the initial feature value, Xmax and Xmin are the maximum and minimum values of the feature, respectively, and X' is the transformed feature value.

Processing of categorical variables

In this paper, all categorical variables are directly transformed into dummy variables, which are 0-1 variables. As a result, the entire dataset has become numerical variables.

Adaboosting feature importance ranking for feature filtering.

In previous studies, the commonly used feature selection methods were point-biserial correlation analysis and random forest feature importance ranking method [7]. However, when facing complex and diverse data, the performance of point-biserial correlation analysis is not satisfactory. Through experimental comparison, it is found that Adaboosting feature importance ranking method performs better than random forest feature importance ranking method on this dataset. Therefore, we choose Adaboosting feature importance ranking for feature filtering.

The specific procedure is as follows: The dataset is split into training and testing sets in a 7:3 ratio. Then, the variables in the dataset are inputted into the Adaboosting regression model. Finally, the Adaboosting feature importance ranking is generated, and features with an importance value less than 0.001 are filtered out. This process is shown in **Figure 3**.

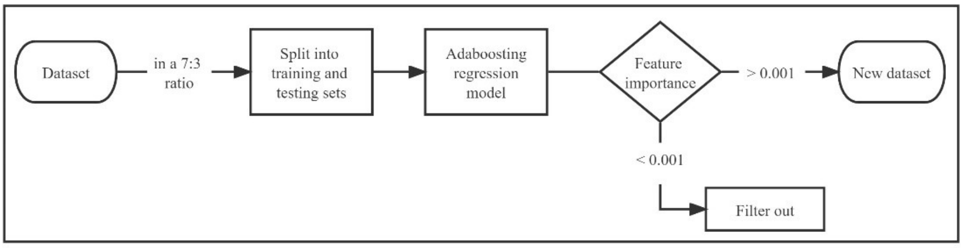

Fig. 3. Process of Adaboosting feature filtering

The partial feature importance ranking generated by the Adaboosting regression model is shown in **Table 4**.

Table 4. The top7 features of the Adaboosting feature importance ranking

Feature	Importance
GRM Actuarial	0.090
Machine Learning Scientist	0.054
San Francisco	0.048
Rating	0.046
Company age	0.042
Python	0.035
Hourly wage system	0.031

Model building

This paper uses an innovative optimization algorithm called Dual-Adaboosting, in modeling and compares the results with those obtained by other traditional algorithms, confirming the feasibility of the Dual-Adaboosting system.

First, we split the dataset that has undergone feature filtering into training and testing sets. Then, we use random grid search and Bayesian search to regulate parameters of the Adaboosting model and fit the first Adaboosting model [8]. We chose decision tree as the base model in the Adaboosting model. Afterward, we calculate the error between each predicted value and the true value of the first Adaboosting model training set to obtain the residual. We use the independent variable X in the training set and residual/2 which has a better fitting effect than residual to fit the second Adaboosting model, the reason for using residual/2 is shown in formula (2) :

$$Y_{train} - \frac{Y_{train}+Y_{pred}}{2} = \frac{Y_{train}-Y_{pred}}{2} = \frac{Residual}{2} \tag{2}$$

In this formula, Y_{train} refers to the dependent variable Y in the training set, and Y_{pred} refers to the predicted value of the first Adaboosting model for the training set.

Finally, the predicted salary value of the testing set is equal to the predicted salary value of the first Adaboosting model testing set plus the predicted salary value of the second Adaboosting model testing set. The whole flow chart of the Dual-Adaboosting system is depicted in **Figure 4.**

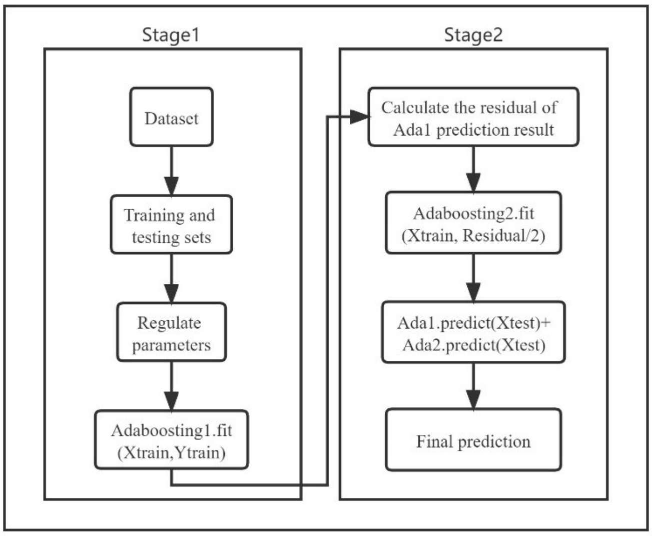
Stage1
Dataset
Training and testing sets
Regulate parameters
Adaboosting1.fit (Xtrain,Ytrain)
Stage2
Calculate the residual of Ada1 prediction result
Adaboosting2.fit (Xtrain, Residual/2)
Ada1.predict(Xtest)+ Ada2.predict(Xtest)
Final prediction

Fig. 4. The whole flow chart of the Dual-Adaboosting system

4 Experiments & Results

4.1 Experiment environment

The dataset comes from a public database named kaggle.com. This experiment was done in python 3.8.0, and the configuration of the computer is shown in **Table 5.**

Table 5. The configuration of the computer

Hardware	Hardware model
CPU	Intel core i7 CPU 2.90 GHZ
RAM	40.0 GB

4.2 Experiments and results

Firstly, comparative experiments are conducted using Lasso regression model, Ridge regression model, random forest regression model, SVR model [9], GradientBoosting model [10] and Adaboosting model. The results are shown in **Table 6.**

Table 6. Experimental results of different regression models

	Training (RMSE)	Testing (RMSE)	Testing (MAE)
Lasso	32.99	35.77	28.38
Ridge	19.27	26.44	19.64
Randomforest	31.09	36.48	28.23
SVR	38.01	41.05	31.76
GradientBoosting	37.24	40.45	31.46
Adaboosting	3.18	20.67	10.52

Secondly, we conducted experiments using the Dual-Adaboosting system, and the result is shown in **Table 7.**

Table 7. Experimental result of the Dual-Adaboosting system

	Training (RMSE)	Testing (RMSE)	Testing (MAE)
Dual-Adaboosting	2.83	17.57	8.36

From the experimental results, the Dual-Adaboosting system can reduce the risk of overfitting to some extent, and the performance of the traning and testing sets are better than that of the other models.

5 Conclusions

In this paper, we propose the innovative optimization algorithm Dual-Adaboosting system and applie it to salary prediction. Compared with various traditional algorithms, this method effectively reduces errors in salary prediction and has good application prospects. However, there are also aspects that need improvement for this approach. For example, its running time is longer than traditional algorithms, and the parameter tuning process is more complex.

References

[1] C.P. Lang, T. Deng ,C.L. Zhang, and Z.D. Xu. (2022) “Prediction of per capita wages of mining urban units based on grey theory,” Industrial Minerals & Processing, 16, pp.18-22. 10.16283/j.cnki.hgkwyjg.2022.01.004.

[2] Y.C. Peng, J. Zhang, and Z.S. Qin. (2021) “Job salary prediction based on random forest algorithm,” Intelligent Computer and Applications, 11, pp.67-72. https://navi.cnki.net/knavi/journals/DLXZ/detail?uniplatform=NZKPT.

[3] J.Y. Zhang, and J.Y. Cheng. (2019) “Study of Employment Salary Forecast using KNN Algorithm,” In: 2019 International Conference on Modeling, Simulation and Big Data Analysis. Wuhan, pp.175-179. 10.2991/MSBDA-19.2019.26

[4] H.L. Huang. (2022) “The application of machine learning in the field of human resource management,” Human Resources Development, 23, pp.92-93. 10.19424/j.cnki.41-1372/d.2022.23.031.

[5] N.A. Jalil, H.J. Hwang , and N.M. Dawi. (2019) “ Machines Learning Trends, Perspectives and Prospects in Education Sector,” In: Education and Multimedia Technology 2019. pp.201-205. 10.1145/3345120.3345147.

[6] D. Wang, H.L. Cheng, W.N. Ding, D.J. Li, and H.N. Liu. (2022) “ The application of SVR and AdaBoosting algorithms in the interpretation of porosity in fractured carbonate reservoirs,” Progress in Geophysics, pp.1-13. https://kns.cnki.net/kcms2/article/abstract?v=Mio27DFCfpCf3Ki-mr2w_pVtX8kQSdT09Y1L_b5rIknezY-fOUgm1E4ByDlU1JSE1K-zWcoCCRjHkgr2Suwc4vkUCfvCHZ6iL3d1Qxqb3bTfV5tQfJYvKOJxeVXGo4DAUVS9kT-Gmdg=&uniplatform=NZKPT&language=CHS.

[7] W.J. Wu, and J.X. Zhang. (2021) “Feature selection algorithm of random forest based on fusion of classification information and its application,” Computer Engineering and Applications,57, pp.147-156. https://kns.cnki.net/kcms2/article/abstract?v=Mio27DFCfpBlVw-heCD5ebhI-j3UfIZHMKZBI30jEAy76X7s_jNK3etnRKlncfHPSSv0T5xq0Sr1Xhqp2y27iol_otFiGfteDedVng2iHoIGsgsNzvyUtKbXLvFCv-Pg_dbxAttXmVE=&uniplatform=NZKPT&language=CHS.

[8] L. Liu, J. Liang, L. Ma, H.L. Zhang, Z. Lin, and S. Liang. (2022) “ Gas Pipeline Flow Prediction Model Based on LSTM with Grid Search Parameter Optimization,” Processes, 11, pp.63-63. 10.3390/PR11010063.

[9] S.Y. Zhong, Y. Zhang, J. Dai, and J. Qian. (2023) “ Improved grey wolf optimization algorithm based on SVR prediction,” Computer Measurement & Control, pp.1-11. https://kns.cnki.net/kcms2/article/abstract?v=Mio27DFCfpD9mI_L9isjZApv7tvlJ88IaOHgJk4LEtsDOMlVIuZhliMNeIr0NzvJ_tW4wBwrrfWEOdo_33NW-GtMkrCYX-lHSinrrJy5EH2Xs2wj2imH4rSd778QeFdxkplRZiptf30=&uniplatform=NZKPT&language=CHS.

[10] D.P. Luca, and F. Nicola. (2023) “ Energy Consumption Forecasts by Gradient Boosting Regression Trees,” Mathematics, 11, pp.1068-1068. 10.3390/MATH11051068.