Salary Prediction Based on the Resumes of the Candidates

Yanming Chen[1,*] and Xinlong Li[2]

[1]Shantou University, Shantou, China
[2]Nanyang Technological University, Nanyang, Singapore

Abstract. This paper aims to build a salary prediction model based on the resumes of candidates in a recruitment environment. Point-biserial correlation analysis and random forest feature importance ranking methods are employed for the paper to conduct feature selection after the dataset is cleaned and preprocessed. Then, OLS linear regression is adopted to analyze the features selected, and three different models, including random forest regression, decision tree, and ridge regression, are applied in prediction experiments, helping obtain results to be compared and analyzed based on RMSE and MAE. Finally, a stacking ensemble method can be used to integrate and fuse different models to build the final salary prediction model. This model definitely has practical reference significance for both candidates and recruiters.

1. Introduction

Salary prediction raises heated discussion worldwide. Among the prediction model, algorithms including multiple linear regression, random forest, neural network, decision tree, etc., are frequently used. Many scholars have conducted a lot of research on salary prediction, such as using neural networks to predict wages based on workers' skills [1], predicting the per capita wages of urban mining units based on grey theory [2], predicting job salaries based on random forest algorithm [3], and conducting employment salary forecast via KNN algorithm [4]. However, these studies considered heavily based on some specific salary influencing factors and recruitment requirements or certain industries, which lacked generality. This paper hence conducts multiple regression model and stacking ensemble experiments on resumes of candidates from different industries and positions, aiming to explore and establish a reasonable and effective salary prediction model for reference considering different candidates’ resumes.

During the job-seeking and recruiting processes, various information presented in the resumes of candidates is always difficult to be comprehensively analyzed and allocated with reasonable weight when deciding salaries due to many reasons, including the recruiters' subjective feelings, the imperfect salary system, etc., leading to unreasonable and opaque salary situations [5]. Therefore, for candidates, a sound salary prediction model can help them clarify their positioning and make choices regarding the company, and even observe their future development trends by adjusting model variables. From the perspective of recruiters, a salary prediction model is beneficial for improving recruitment and salary standards, as well as for providing more reasonable salaries to attract and discover talents.

2. Materials and methods

2.1 Dataset used in the study

Since India is a large emerging economy deeply involved in globalization, and its salary level is to some extent linked to the international salary level, this paper hence selects candidates in India as the core research subject, and dataset records were obtained from kaggle.com.

The resumes of the candidates included in the dataset come from different cities such as Mumbai, Delhi, and Kolkata, as well as different industries such as IT, Marketing, and Management. In addition, there are different requirements for working hours and different positions. The dataset contains a total of 500 observations, each representing the monthly salary corresponding to a candidate’s resume in US dollars. The salary range in the dataset shows a difference from $1510 to $6991.

Based on those resume information, a salary prediction model can be established. There are eleven variables in this dataset. The variable named 'Job Title' only serves an explanatory purpose which has no substantive meaning, so it is not included in the scope of the model calculation. Therefore, the variable named 'sal' serves as the dependent variable, and the nine variables other than 'Job Title' 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.

[*] Corresponding author: 21ymchen@stu.edu.cn

Table 1. The proportion of two categories of variables

	Categorical	**Numerical**
Number	7	2
Proportion	77.78%	22.22%

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

Table 2. Basic information about numerical variables

	Longitude	**Latitude**	**Sal**
Count	473	473	500
Mean	75.9884	21.0897	4224
Min	-79.0305	-8.1237	1510
Max	121.0977	46.3144	6991
Missing values	27	27	0

Table 3. Basic information about categorical variables

	Count	**Unique**	**Freq**	**Missing**
Experience Required	500	76	38	0
Key Skills	500	473	22	0
Role Category	463	59	130	37
Location	489	101	71	11
Functional Area	489	40	119	11
Industry	489	53	187	11
Role	486	138	96	14

In the following data processing, the essay will transform the categorical variable 'Job Experience Required' into two columns of numerical variables.

2.2 Method

In this paper, variables are first divided into numerical variables and categorical variables. After data cleaning and preprocessing of numerical variables, multiple categorical variables are reclassified, and virtual variables are generated using the point-biserial correlation algorithm for feature selection. Since there is some correlation between variables, random forest feature importance is used for additional screening 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, and a stacking ensemble method is used to integrate different models to obtain the final salary prediction model.

2.2.1 Data cleaning and data preprocessing

Missing and outlier values for the location and longitude/latitude are first processed. The 27 missing values in longitude and latitude are filled using the global longitude and latitude query system based on their corresponding locations. Similarly, the 11 missing values in the location are filled based on their corresponding longitude and latitude. The occurrence of outlier values in longitude and latitude is shown in the Table 4.

The two outlier values, '-79.030572, -8.123729' and '121.0977529, 14.6719732', are located in other countries. The corresponding 'Location' for '-79.030572, -8.123729' is 'Gurgaon', while the corresponding 'Location' for '121.0977529,14.6719732' is the 'NCR region'. Therefore, the global longitude and latitude query system can be used to fill in these outlier values, and upon observing the dataset, it is found that the variables corresponding to '11.048029,46.314475' are mostly missing values, so these 11 rows of data could be directly deleted.

Table 4. Distribution of outliers in latitude and longitude variables

Longitude	**Latitude**	**Occurrence Number**
-79.0305	-8.1237	1
11.0480	46.3144	11
121.0977	14.6719	9

There exists a certain correlation between the four variables, 'Role Category', 'Functional Area', 'Industry', and 'Role', and the 'Job Title' variable also has some reference value for these four variables. Therefore, missing and outlier values in these four variables can be filled in and modified by referring to the data before and after.

The research removes the unit 'yrs,years' from the variable' Job Experience Required 'and converts it into two columns of variables -' Minimum time requirement 'and' Maximum time requirement '.

The 'Min-Max Rescaling' method is used to perform linear transformation on the variables 'Minimum time requirement', 'Maximum time requirement', 'Longitude', and 'Latitude', mapping their feature values to the interval [0, 1]. The formula for Min-Max Rescaling is as follows (1) :

$$X' = (X - X\_min) / (X\_max - X\_min) \qquad (1)$$

Where X is the original feature value, X_min and X_max are the minimum and maximum values of the feature, respectively, and X' is the transformed feature value.

In many categorical variables within the dataset, each row of data is composed of different combinations of categories. After tokenization, there are hundreds or even thousands of different categories, and these categories have a significant correlation, inclusion, and overlap, which is not conducive to building regression prediction models. Therefore, the data needs to be reclassified.

Firstly, the categories in the variable 'Role Category' are reclassified, and the reclassified categories are divided into 17 types: 'administrative department', 'editorial department', 'education', 'engineering department', 'finance department', 'hr', 'ir', 'it', 'marketing', 'middle management', 'planning department', 'public relations', 'quality department', 'r&d', 'service department', 'top management', and 'other'. Each row of data in the reclassified variable 'Role Category' is one of these 17 types or a combination of them.

Next, the categories in the variable 'Industry' are reclassified, and the reclassified categories are divided into 40 types: 'sales', 'biotech', 'finance', 'management', 'bpo', 'interior design', and so on. Each row of data in the reclassified variable 'Industry' is one of these 40 types or a combination of them.

The paper uses the same reclassification method for variables 'Role 'and ' Functional Area '.

There are 101 categories in the variable 'Location', which can be simplified into 13 categories, including Mumbai, New Delhi, Kolkata, Chennai, Noida, etc. The Noida category in this classification includes both Noida and Greater Noida.

2.2.2 Point-biserial correlation algorithm for feature selection

Taking the variable 'Role Category' as an example, 17 dummy variables are generated for the 17 categories of 'Role Category', and then point-biserial correlation analysis is performed based on 'sal', yielding the partial results shown in Table 5.

Table 5. Partial point-biserial result between 'Role Category' and 'Sal'

	Correlation	Pvalue	Correlation(abs)
Top management	0.253	1.35e×10-8	0.253
Service department	-0.176	9.04e×10-5	0.176
Other	-0.124	0.006	0.124
Quality department	0.066	0.139	0.066
IT	-0.056	0.211	0.056
Middle management	0.051	0.260	0.051

In point-biserial correlation analysis, p value < 0.05 indicates higher significance. From Table5, it can be seen that only the dummy variables for 'top management', 'service department', and 'other' in 'Role Category' have higher significance. Therefore, these three dummy variables are selected as features of 'Role Category'.

Similarly, features can be selected for 'Industry', 'Role', and 'Functional Area'. However, since there are hundreds of dummy variables in 'Key Skills' with a p value < 0.05, both p value and frequency of occurrence should be considered in the process of selecting features for 'Key Skills'. The final selected features for these four variables are shown in Table 6.

Table 6. Screened feature of four variables

Variable	**Feature**
Industry	'other', 'biotech', 'sales'
Role	'business development manager' 'software developer' 'it-software'
Functional Area	'it software - application programming and maintenance', 'r&d', 'creative', 'bpo', 'bi', 'accounts', 'front office'
Key Skills	'recruitment', 'customer service', 'python', 'planning', 'javascript', 'excel', 'oracle', 'jquery', 'ajax', 'ms sql', 'c#', 'marketing executive', 'business development'

2.2.3 OLS Regression

The essay selects variables with higher significance and performs OLS regression with 'Sal' [6], removing the 'it-software' column in 'Role' as the baseline, which mainly refers to basic IT personnel, and removing the 'Other' column in 'Location' as the baseline. The 'Other' variable has 8 observations and represents some smaller and underdeveloped cities. The results are shown in the Table 7.

Table 7. Partial OLS Regression result

	Coef	**Std err**	**P>|t|**	**0.025**	**0.975**
Top Management	1818	329	0.000	1171	2465
Developmen t Manager	962	290	0.001	390	1533
Planning	908	435	0.037	53	1763
Customer Service	1029	375	0.006	290	1768
Service Department	-1586	407	0.000	-2385	-786
Mumbai	1017	1113	0.036	-1169	3205
Noida	-932	408	0.023	-1734	-130
Pune	1567	851	0.046	-106	3240
Ahmedabad	2294	1187	0.054	-40	4628

From the OLS Regression Result, it can be seen that salaries for industries and positions such as 'top management', 'business development manager', 'planning', and 'customer service' are significantly higher than that given to basic IT personnel. On the other hand, salaries for basic personnel in 'service department' such as front desk personnel are significantly lower than that for IT personnel. Comparing small cities in the 'Other' category, salaries in larger cities or developed areas such as 'Mumbai', 'Noida', 'Pune', and 'Ahmedabad' are generally higher.

2.2.4 Feature re-screening according to the importance of random forest features

The paper deals with the variable location since not all 13 dummy variables are suitable to be put into the prediction model. Hence, the salary distribution and average salary of different locations are studied and concluded in the following Table 8.

Based on the above data, this paper encodes the variables " Other; Ahmedabad, Kolkata, Chennai, Chandigarh, Bengaluru, Hyderabad, Gurgaon, Pune, New Delhi; Mumbai, Navi Mumbai, and Noida" as 0, 1, and 2 respectively. Here, 0 represents cities with lower average salary levels, 1 represents cities with normal average salary levels, and 2 represents cities with higher average salary levels.

Due to the intercorrelation among variables of 'Role Category', 'Role', 'Industry', and 'Functional Area', and the large number of features selected by the variable 'Key Skills', this paper uses the random forest feature importance method to further screen the features in order to reduce the risk of overfitting [7]. The specific procedure is as follows: The screened features are transformed into dummy variables and combined with numerical variables to be input into the random forest regression model. The dataset is then split into training and testing sets in a 7:3 ratio. The model is run multiple times with different splits of the dataset into training and testing sets, and each run generates a random forest importance ranking. Finally, the variables with consistently low importance rankings across multiple runs are filtered out.

Table 8. The average salary of different locations

Location	Average salary
Mumbai	4453
Navi Mumbai	4790
New Delhi	4110
Noida	4536
Pune	4120
Gurgaon	4155
Hyderabad	4337
Bengaluru	4058
Chandigarh	4347
Chennai	4072
Kolkata	4374
Ahmedabad	4407
Other	2708

3. Results & discussion

3.1 Experiment environment

The training and testing datasets come from a public database named kaggle.com. This experiment was done in python 3.7.0, and the configuration of the computer is shown in Table 9.

Table 9. Hardware environment to implement the method

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

3.2 Results & discussion

First, the optimal combination of 'max_depth' and 'n_estimators' in the random forest regression model is determined using the methods of random grid search and grid search. Then, experiments are conducted using random forest regression, decision tree, and ridge regression models, and the results are shown in Table 10.

Table 10. Experimental results of different models

	Random Forest	Decision Tree	Ridge
Training (RMSE)	624	1330	1372
Training (MAE)	500	1049	1126
Testing (RMSE)	1520	1603	1480

Second, the methods of stacking are used to integrate the models [8]. The first layer model selects stochastic forest regression, decision tree, and ridge regression, and the second layer model selects the decision tree as the integration and fusion model. The results are shown in Table 11.

Table 11. Experimental results of stacking

	Training	Testing
RMSE	1361	1362
MAE	1076	1115

From the experimental results, it can reduce the risk of overfitting to some extent, and the performance of the test set is better than that of the single model. Therefore, the model that is integrated with stacking is ultimately selected as the salary prediction model.

4. Conclusions

The stacked model is adopted as the salary prediction model. It can effectively assist candidates or recruiters in predicting salaries based on resumes. This model also can be extended to practical applications such as company selection and recruitment standards development. However, further in-depth research is needed to improve the model's accuracy. Because the limited number of observed values (only 500) in the dataset affects the model's accuracy.

References

1. J. Otto, C.D. Han, and T. Stella, "Using Neural Networks to Predict Wages Based on Worker Skills," Studies in Business and Economics, **16**, pp.95-108. (2021)
2. C.P. Lang, T. Deng ,C.L. Zhang, and Z.D. Xu, "Prediction of per capita wages of mining urban units based on grey theory," Industrial Minerals & Processing, **16**, pp.18-22. (2022)
3. Y.C. Peng, J. Zhang, and Z.S. Qin, "Job salary prediction based on random forest algorithm, "

Intelligent Computer and Applications, **11**, pp.67-72. (2021)

4. J.Y. Zhang, and J.Y. Cheng, “Study of Employment Salary Forecast using KNN Algorithm,” In: 2019 International Conference on Modeling, Simulation and Big Data Analysis. Wuhan, pp.175-179. (2019)
5. G. Tang, “ Game analysis and Countermeasure Research on information asymmetry in enterprise recruitment,” Marketing Circles, **33**, pp.150-151. (2021)
6. S. M. Wang, J.Y. Li, W.J. Guo, H.L. Yang, and Y. Fei, “An empirical analysis of dialect and regional economic growth based on OLS regression model,” E3S Web of Conferences, **253**, pp.02057. (2021)
7. W.J. Wu, and J.X. Zhang, “ Feature selection algorithm of random forest based on fusion of classification information and its application, ” Computer Engineering and Applications, **57**, pp.147-156. (2021)
8. Z.K. Liang, “Research and empirical analysis of vehicle and cargo matching Model based on Stacking Integrated Learning,” Modern Computer, **28**, pp.46-50. (2022)