Visual Analysis of Big Data Related Job Recruitment Information Based on 51job

Jingjing Shen(✉) and Shuyan Yu

Yuanpei College, Shaoxing University, Shaoxing, China
1157121536@qq.com, ysy@usx.edu.cn

Abstract. In such an era of big data, the accumulation of data leads to a sharp increase in the demand for big data-related positions, and a large number of recruitment information is published on recruitment websites. The mining and analysis of these recruitment information will help those engaged in related fields to understand the current situation of the industry and make relevant predictions. Based on 51job, this paper uses various technologies of Python to carry out visual analysis on the job information related to big data major. The random forest algorithm of machine learning and XGBoost algorithm are used to train the salary prediction model, optimize the selection of features, and compare the models, and finally get a model with high accuracy for prediction. Based on the research above, this paper mines out the information of big data-related positions, and provides a comprehensive and intuitive analysis of the industry employment market for big data practitioners.

Keywords: Major in Big data · Recruitment information · Python · Visual analysis · Machine learning

1 Introduction

1.1 Foreword

Big data has become a hot research topic in today's society, and more and more people have noticed the importance of big data. With the development of Internet informatization, all walks of life have generated and accumulated a large amount of data, which will be of great value to all walks of life. Big data is the oil of the 21st century, and whoever holds the data will hold the password of the future trend [1]. At present, the scale of big data market continues to expand, and the demand for talents in big data-related industries is also increasing [2]. This paper makes a visual analysis of the job information related to big data to help the employees to understand and choose the jobs related to big data.

1.2 Background and Current Situation at Home and Abroad

Domestically, the Outline of the 13th Five-Year Plan for National Economic and Social Development of the People's Republic of China was released, which takes big data as

D. Kumar et al. (Eds.): IEIT 2023, AHSSEH 10, pp. 211–225, 2023.
https://doi.org/10.2991/978-94-6463-230-9_27

a basic strategic resource, comprehensively implements actions to promote the development of big data, accelerates the sharing, opening up, development and application of data resources, and contributes to industrial transformation and upgrading and innovation in social governance. Therefore, at present, all walks of life are in the digital construction, the digital planning and construction is bound to produce a large number of data, at the same time, a large number of big data professional related personnel demand. According to the Statistics of The China Council of Commerce, the talent gap of basic data analysis in China will reach 1.5 million in 2021, and it is still growing at a compound rate of more than 20% every year, which is a very large talent gap. In China, the vacancy of big data majors is huge. In foreign countries, all countries attach great importance to big data research and launch their own research plans from the national strategy level [3]. According to Gartner, more than 75% of big data positions are unfilled. From the current situation of foreign countries, there is also a great demand for positions related to big data. With the development of Internet informatization, more data will be generated and accumulated in all walks of life, and more big data professionals are needed to mine the value of these data. The social demand for positions related to big data will only increase, and the vacancy of positions related to big data will only increase [4]. Therefore, it is very necessary to carry out visual analysis of big data related job information. Although there are relevant literatures that use Python technology to crawl and analyze job recruitment information on the website, there is no analysis of recruitment information related to big data. Therefore, python technology can be used to carry out visual analysis of recruitment information related to big data based on 51Job website.

1.3 Design Ideas

At present, although relevant literature on visual analysis of job information can be found, there is no complete visual analysis of big data related job information [5]. Moreover, a lot of visual analysis only visualizes the data in charts, but a complete visual analysis needs to produce a predictive model. Therefore, this paper adopts network crawler technology, Python data cleaning technology, Python data visualization technology and machine learning to make a complete visual analysis of recruitment information related to big data major. In terms of model prediction, salary is currently the most concerned issue for most job seekers, but it is very sensitive to ask salary during employment, so we choose the model to predict salary by training features. The research idea is to use the web crawler to obtain the current job information related to big data in 51Job website, use Python data cleaning technology to process the repeated values and outliers of the data obtained by crawling, and conduct data analysis according to the job field. Python data visualization technology is used to visually present the analyzed data in charts, and some results can be used as the basis for job-seekers to choose positions related to big data. Data analysis is used to extract features, conduct feature analysis and feature selection, and build models based on features. Train a model that can predict the average salary of a job based on characteristic data. The model allows job seekers to know in advance the average salary of a job.

2 Technical Analysis

2.1 Web Crawler Technology

Web crawler is an application or script that can extract web content according to certain rules. Its working principle is to capture URL, then perform DNS resolution, download the web page to the local library, and then parse and obtain web page information [6]. The advantages of web crawler technology are summarized as four points. Firstly, it can save time. Web crawler can quickly obtain information of multiple web pages without manually collecting data. Secondly, large scale data can be obtained. More data can be obtained by using web crawlers than by manual collection. Third, accurate data can be collected, and the possibility of introducing errors with the data crawled by web crawlers is very low. Fourth, you can collect structured data, and you can write code to retrieve data content in a fixed format.

Compared with many web crawler technologies, xpath is the most appropriate to use with Google Chrome to crawl 51Job. Among them, the developer tool of Google Browser can help extract request headers. The advantage of xpath is that it can obtain multiple nodes in the web page and directly obtain key information to form formatted information for saving.

2.2 Python Data Cleaning Technology

Python data cleaning deals with duplicate values and outliers using the Collections, Pandas, and Seaborn libraries [7]. The problem of repeated values is caused by the crawling of multiple same data. Therefore, the collection library can count the number of occurrences of each data to check the existence of repeated values. The Seaborn library can present data to check the outliers of data. The PANDAS and DROP_duplicates methods are used to view and process duplicated values, and the DROP_outlier method is used to process outliers. Python data cleaning technology can solve the problem of duplicate data and outliers, ensuring that data will not be abnormal in subsequent processing steps.

2.3 Python Data Visualization Techniques

Python data visualization is the diagramming of data using the matplotlib and Seaborn libraries. Matplotlib library provides a complete 2 d graphics and limited support 3 d graphics, and seaborn library is used to create the information rich and attractive statistical database, it offers a variety of functions, such as built-in themes, palette, functions, and tools, to achieve the single factor and double factors, the linear regression, data matrix and statistical time sequence of visual, To further build complex visualizations [8].

2.4 Machine Learning

Machine learning is mainly based on a large amount of data and certain algorithm rules so that computers can learn autonomously like human beings, and through continuous training, can improve the behavior of intelligent decision-making. The advantage

of machine learning is that different problems can be transformed into problems that machine learning can solve, and the resulting model of machine learning can be used to process large amounts of data. In this prediction, two algorithms are mainly used, namely random forest algorithm and XGBoost algorithm.

Random forests with integrated learning thought, integrating multiple decision trees into a forest to predict the results, and training set is divided into several new training set, and then makes every believe training set to build a model, but do our different, unrelated, finally in the prediction of the multiple model for integration, integration way is, the classification problems, The minority subservient to the majority is used for integration, and the mean value is used for integration when regression problems are encountered.

Xgboost is a gradient lifting algorithm based on decision tree, which is to add decision trees through feature splitting. Finally, many trees are obtained, and each tree will correspond to a node, and each node will have a score. Finally, all the corresponding scores are added up to the final predicted value result [9–11]. In practice, two strategies are used, namely the global strategy and the local strategy. The global strategy is reflected in that each feature will determine a global split point set, while the local strategy is reflected in that each split point needs to be re-selected.

3 Implementation of Visual Analysis

3.1 Crawl Data

The crawler technology is used to crawl the recruitment information related to big data on 51Job website. The xpath method in the crawler technology is mainly used. First, Google Browser is used to log in 51Job website, and the developer tools of Google Browser are used to check the identifiers such as cookie and UserAgent in the webpage. And through the search function and page turning function of the website to check the CHANGE of URL, according to these information use crawler to crawl the data, use the URL analysis, the job information and page number to be searched respectively with key, page control, cookie, useragent and other identifiers into the request header. To obtain the response object of the web page, xpath is used to parse the response object to obtain the standard web page result. Finally, the data is formatted anda CSV file is generated using pandas.

After the above series of crawling operations, a total of 54950 pieces of data were climbed, which mainly obtained the job information related to big data in 12 fields, including job name, company name, lowest salary, highest salary, working place, company nature, release time, job category and company benefits.

3.2 Data Cleaning

1) View and process repeated values

First of all, the Counter function of the Collections library was used to check the repeated values of the data. Here, the URL of the post was statistically checked. As can be seen from the results, each data only appeared once without repeated values, so there was no need to redo the data.

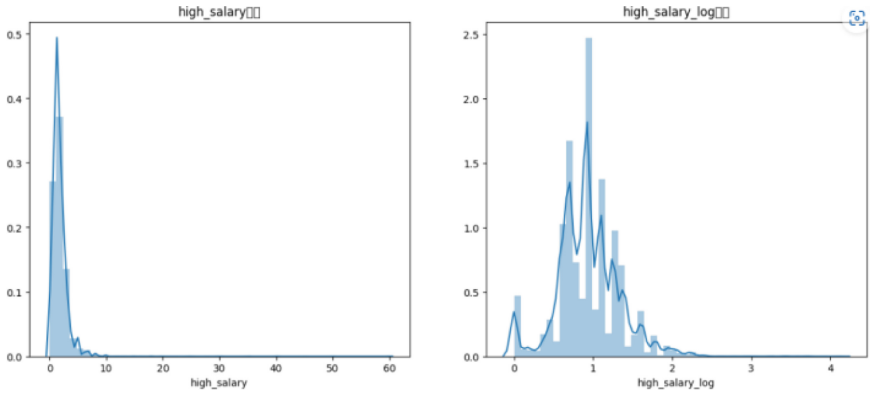

Fig. 1. Comparison of the highest salary outlier before and after processing.

2) View and handle outliers

Aggregate through the groupby(by) method of DataFrame object. The by parameter refers to the specific fields to aggregate. It can be a string, but only one field can be aggregated; It can also be a list, which can aggregate multiple fields. Then perform statistical operations through the agg(func) method of DataFrameGroupBy object, where func parameters can be a method, string, list and dictionary, and the method returns a DataFrame object. If a method is passed, it can implement user-defined statistics and carry out relevant operations according to the passed method; If a string or list is passed in, its value can be statistics related operations such as mean and sum, which can be applied to each column; If you only want to perform statistical operations on the specified columns, you can pass in a dictionary in the form of {'field name': 'statistical According to the types of fields in the acquired data, the lowest salary LOW_salary and the highest salary high_salary data need to be checked for outliers. Seaborn library is used to present the Gaussian distribution map of the data distribution to check whether there are outliers. According to the ICONS presented, it can be seen that the data are concentrated in one place, so there are outliers. Low_salary_log, high_salary_log, low_salary_log, Then seaborn library is used to present gaussian distribution and box graph of low_salary_log and high_salary_log data. Comparing observations before and after outliers shows that the data becomes evenly distributed, as shown in Fig. 1 and Fig. 2.

3.3 Data Visualization Analysis

1) Visual presentation of job categories

According to the number of job category, to present word cloud of all job categories, can be seen from the word cloud computer software, Internet, e-commerce, computer services, such as the job category for large data of related professional demand is bigger, when job seekers in the job, can consider these jobs categories of recruitment, as shown in Fig. 3.

According to the statistics of the average minimum salary and average maximum salary of the job category, a scatter heat map is presented for the job category. From the location of the scatter points in the figure, it can be seen that the current general salary

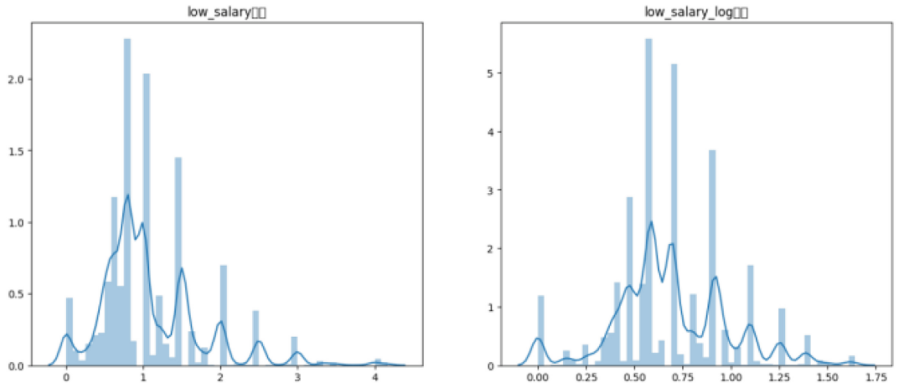

Fig. 2. Before and after processing of minimum wage outliers.

Fig. 3. Word cloud display of job categories.

range is about 10,000/month to 20,000/month. Between months, compared with other industries, it can be seen that the overall salary of big data professional-related positions is at a higher level. Therefore, if job seekers choose to engage in positions related to big data, there is a high probability that they will get a good salary, as shown in Fig. 4.

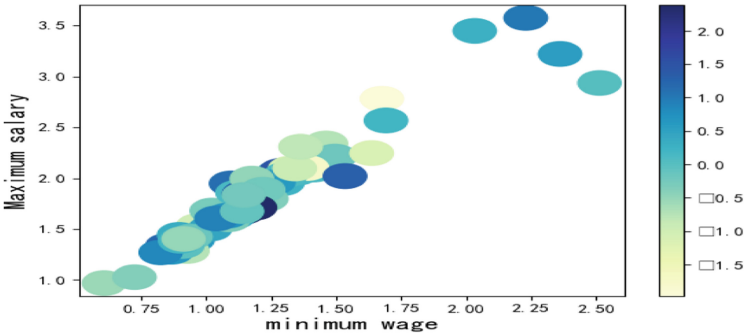

Fig. 4. A scatter heat map showing the highest and lowest salaries for job categories.

According to the statistics of the average salary and the average maximum salary of the position category, the position category is presented in a box diagram. It can be seen that the number of position categories between 10000 and 20000 is the largest, as shown in Fig. 5 [12].

According to the statistics of the top 10 job categories, a circular pie chart is presented for job categories. As can be seen from the pie chart, computer software, real estate, and Internet/e-commerce positions account for a large proportion of jobs, and these three job categories can be chosen first if job seekers want to get more job opportunities, as shown in Fig. 6.

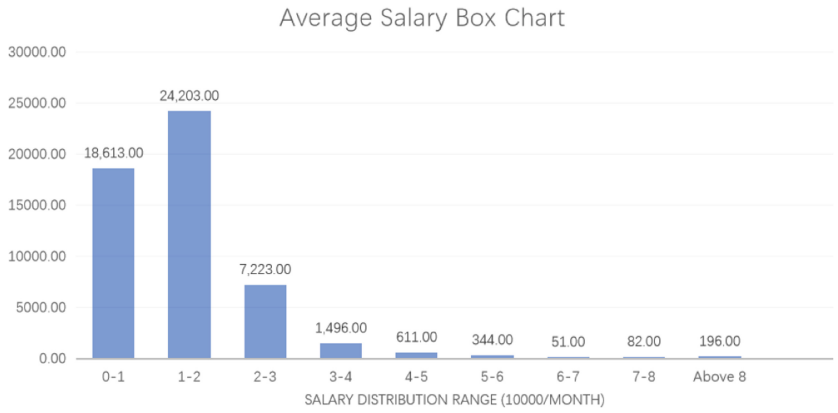

Fig. 5. Box chart display of average salary of each category.

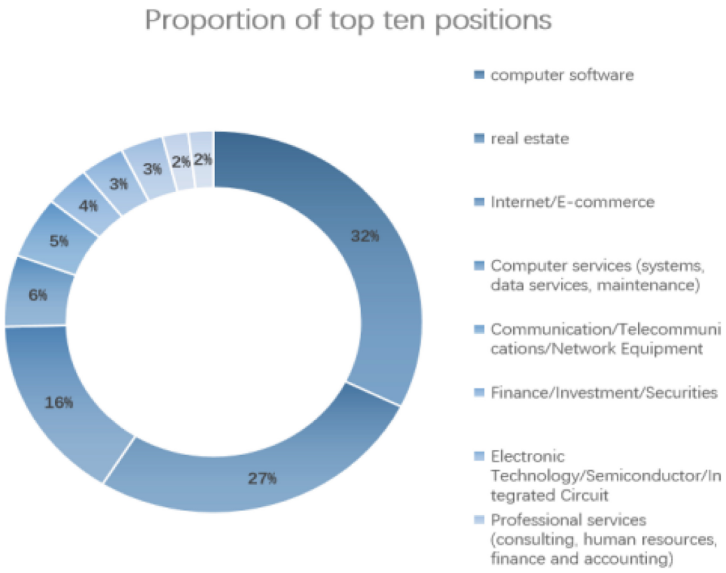

Fig. 6. Pie chart showing the top 10 job categories in proportion.

According to the statistics of the top 10 jobs with average minimum salary and the highest average salary, a bar chart is presented for job categories. According to the bar chart, it can be seen that the average minimum salary of the top ten job categories is above 14,000 yuan/month, and the average maximum salary of the top ten job categories is above 22,000 yuan/month, as shown in Fig. 7 and Fig. 8.

2) Visual representation of urban job distribution

According to the statistics of the number of jobs related to big data major in each city, the urban job distribution of China map heat map is presented. As can be seen from the chart, Shanghai is the city with the largest demand for positions related to big data majors, followed by Guangdong, Beijing, Jiangsu, and Zhejiang, Sichuan, and Hubei. It can be seen that Beijing, Shanghai and Guangzhou, as first-tier cities, have a more urgent demand for posts related to big data, as shown in Fig. 9.

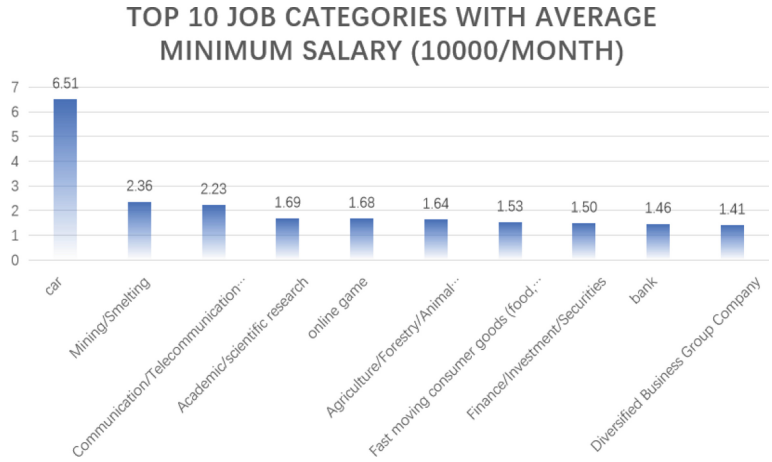

Fig. 7. Bar chart showing the top 10 job categories in minimum wage.

Fig. 8. A bar chart showing the top 10 highest paid job categories.

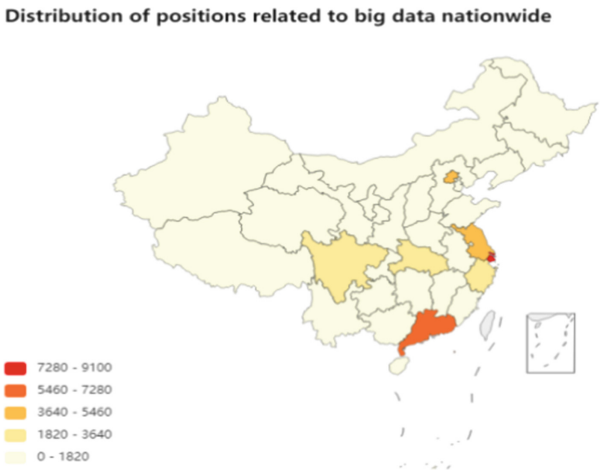

Fig. 9. China map heat map showing the number of big data jobs in each city

3) Visual representation of the nature of the company.

According to the statistics of the number of positions related to the big data major of the top ten companies, the company nature is presented in a bar chart [13]. As can be seen from the bar chart, private companies account for the largest proportion, accounting for 66.15%, followed by listed companies and state-owned enterprises, as shown in Fig. 10.

4) Visual representation of the number of positions.

According to the statistics of the number of positions, the number of positions is presented in a bar chart and a circular sector chart. As can be seen from the bar chart and circular pie chart, big data development engineers are the most in demand, accounting for 48%. Big data analysis engineers followed, accounting for 15 percent, as shown in Fig. 11 and Fig. 12.

Salaries are presented in a double line chart based on the average highest and lowest salaries for the top 10 positions. As can be seen from the chart, the salaries of these paid positions are relatively average and very stable, as shown in Fig. 13.

5) Visual representation of daily job postings

According to the statistics of the number of posts published daily, the line chart of the number of posts published daily shows that the posts related to big data are on the rise. Therefore, it can be seen that the social demand for posts related to big data is

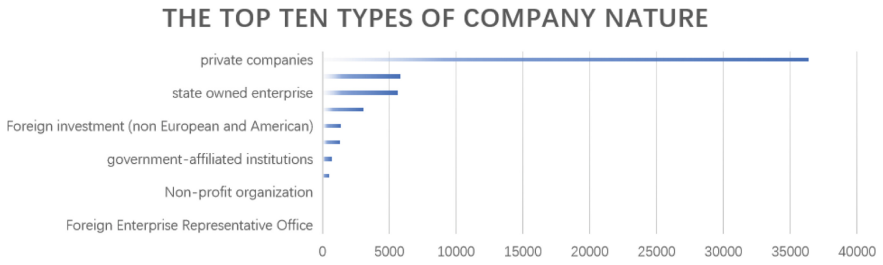

Fig. 10. A bar chart showing the top ten companies by number of jobs.

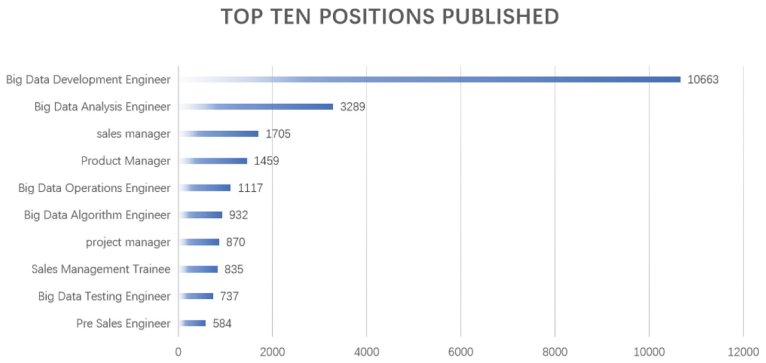

Fig. 11. Bar chart showing the top 10 positions by number of posts.

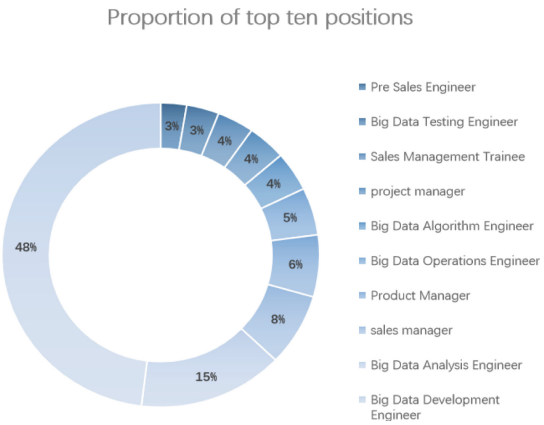

Fig. 12. A pie chart showing the share of the top ten jobs by number of posts.

increasing day by day. Big data is the trend of future development, and the development of big data is infinite and getting better and better, as shown in Fig. 14 [14].

6) Visual presentation of welfare statistics

The fulI in the data is presented in the word cloud. It can be seen from the word cloud that most big data-related posts have post benefits such as paid annual leave, five insurities and one housing fund, employee travel, year-end bonus, performance bonus, regular physical examination and professional training. It can be seen that due to the increase in job demand, companies have put forward a lot of benefits to attract talents to join, as shown in Fig. 15.

3.4 Characteristics Analysis

1) Characteristics of the processing

Before selecting features, the average salary of the target prediction field should be selected first. Therefore, the average salary of each position is calculated according to the highest and lowest salary, and a new field average salary (AV_salary) is generated.

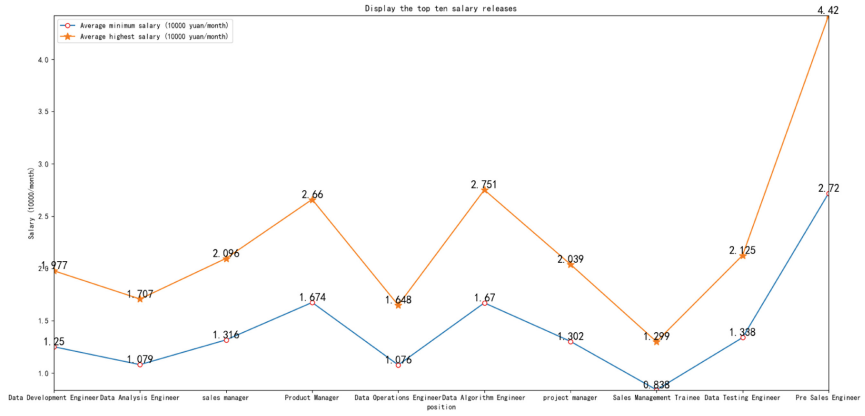

Fig. 13. A double line chart showing the highest and lowest salaries for the top 10 jobs.

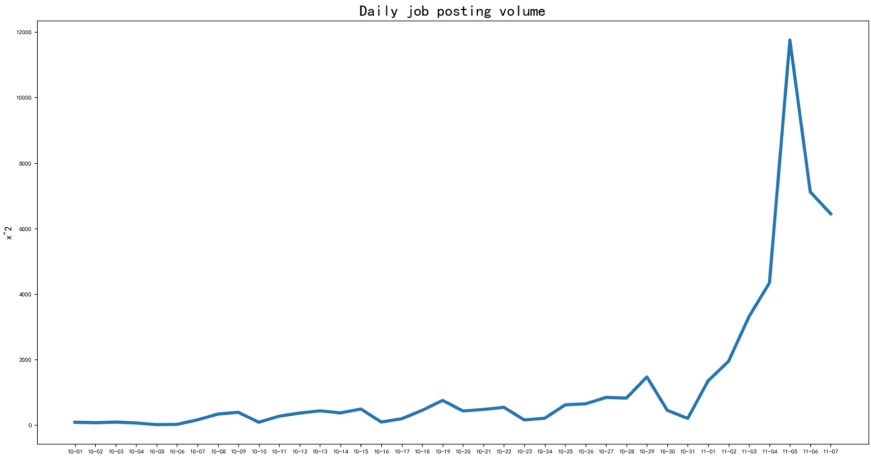

Fig. 14. Daily post number line chart display.

Then generate more new data type fields based on the relevance of the other fields to the job. The new field city_job_rate is generated according to the proportion of the number of jobs in the city. The new field cn_rate is generated according to the proportion of the company nature. The new field job_rate is generated according to the proportion of the number of jobs. A new field, fuli_num, is generated based on the number of benefits. This data is then outliered using the drop_outlier() method to generate new fields with _log.

2) Characteristics analysis

According to the characteristic data, correlation analysis was conducted to form a correlation thermal diagram for presentation. From the chart, you can see the correlation of each characteristic to salary based on the data in each overlapping square. According to the correlation coefficient analysis of the relevant thermal map, it can be seen that the

Fig. 15. Word cloud display of job benefits.

proportion of the number of jobs has the greatest correlation with the average salary, followed by the number of benefits, followed by the proportion of the number of jobs in the city, the proportion of the nature of the company and the proportion of the job category. Therefore, the five characteristics selected are the proportion of job category (class_rate_log), the proportion of city job number (city_job_rate_log), the proportion of company nature (CN_rate_log), the proportion of job number (job_rate_log) and the number of benefits (fuli_num) _log), as shown in Fig. 16.

3.5 Build Models

The pandas library is used to divide characteristic fields and target fields into training data (80%) and test data (20%). First of all, the random forest is used to build the model, and the model is trained according to the training data, and the trained model is saved as Rf.pkL. Then, the XGBoost algorithm is used to build the model, and the trained model is trained with the training data, and the trained model is saved as RF.PKl.

3.6 Predict and Verify Accuracy

Then get_BEST_model_AND_accuracy () was used to compare the predicted results with the accurate results. The accuracy of the random forest model and XGBoost was up to 99.99% and 99.85% respectively. By contrast, the model trained by random Forest is more accurate in salary prediction and should be retained for salary prediction. Therefore, the average salary is predicted using the random forest model, and the resulting average salary is put into a new field, pre_SALARY, and all data is exported to a CSV file using PANDAS 'TO_CSV. By comparing the predicted salary data with the actual salary data, it can be seen that the salary of the positions related to big data is generally on the rise, which means that the salary of the positions related to big data will continue to increase, as shown in Fig. 17.

Using data to train the model, plot_learning_curve() is used to present the learning curve of the model. It can be seen that the accuracy of the model is increasing during the training process, as shown in Fig. 18.

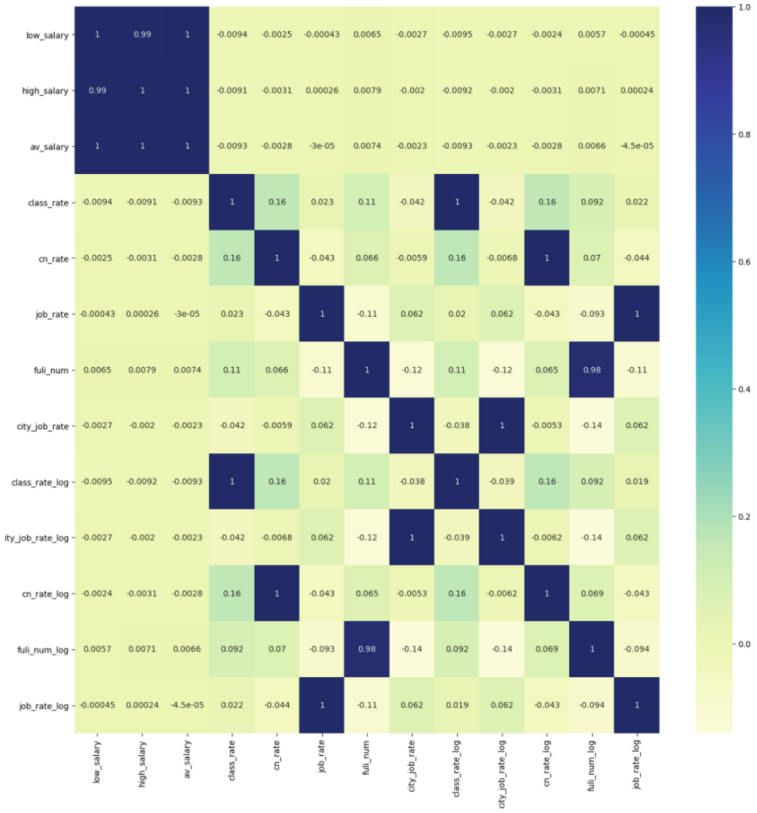

Fig. 16. The correlation of the features is shown in a thermal map.

```
0          1.883649
1          1.121612
2          1.534028
3          1.673250
4          1.651366
              ...
52711      0.486783
52712      1.154432
52713      1.338462
52714      0.699386
52715      0.699386
Name: pre_salary, Length: 52716, dtype: float64
```

Fig. 17. Average salary prediction results of random forest.

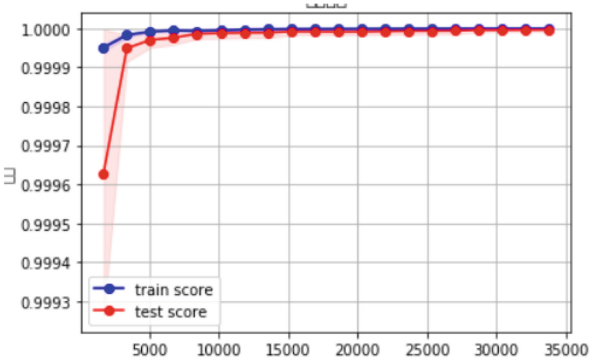

Fig. 18. Random forest model learning curve display.

4 Conclusion

On the basis of data analysis, this paper uses diversified charts to visually present the analyzed data. At the same time, this paper trains a model that can predict the average salary of positions related to big data by using several characteristics related to big data major. The model's prediction accuracy has reached 99%.

As the society attaches more and more importance to big data, positions related to big data will become more and more important, and there is an increasing demand for positions related to big data in all walks of life [15]. Professional related information visualization analysis of large data, can help agencies or personnel to understand the current big data industry present situation, for workers, can help employees to choose professional related data, for large data related to education research institutions, professional can be learned from all walks of life to solve large data of related situation, forecasts the trend.

References

1. X.Y.Zhou and Y.L.Yin, Knowledge Structure Analysis of Big data Management talents based on domestic market demand, Information Science,vol. 35, no. 1, pp.29-34, 2017.
2. C.L.Zhong Z.Y. Cao and X.L.Bai, Analysis of Consumption Behavior of Campus Cards Based on Python.China Academic Journal Electronic Publishing House,2019.
3. W.L.Ning and H.X.MAO, Content crawling of 51Job website based on Python crawler technology,Network and Communication Technology,no.4, pp.47–49,2021.
4. X.C.Qian,The development and industrial opportunities of big data, Internet of Things Technology,no.10, pp.84-86,2013.
5. G.N.Ni,Development and Application of Big Data,Information Technology and Standardization, no.9, pp.6-9,2013.
6. B.Yuan,Application status and Development Trend analysis of Big Data industry,China New Communications,no.24, pp.75-76,2014.
7. T.Y.Luo,Research on scale prediction of network public opinion hot spots, journal of information, vol. 35, no. 10, pp. 181–184+145,2016.

8. S.MAO and H.X.MAO,Recruitment Information Crawling and analysis based on 51Job.com -- Taking Python technical positions as an example,Network Security Technology and Applications, no.4, pp.47–49,2021.
9. B.Huang , H.Chen, Z.J.Fang, M.S.Wang and W.Z.Liu, Analysis of COVID-19 Hot Topics based on Microblog, Journal of Wuhan University, vol. 66, no. 5, pp. 425-432, 2020.
10. T.X.Xie and Z.Wei, Research on Job Characteristics and Talent Demands of China's Internet Finance——2017 Content Analysis Based on Recruitment Website Data, Jiangsu Science and Technology Information, no. 18, pp. 6-8, 2017.
11. X.M.Wang and X.Y.Zhong,Research on Job Characteristics and Talent Demands of Logistics and Supply Chain Finance——Based on Analysis of Recruitment Information from Three Recruitment Websites, Logistics Engineering and Management, vol. 43, no. 2, pp. 166-170, 2021.
12. C.Q.Zong,Text data mining,Beijing: Tsinghua University Publishing House,2021.
13. E.Mathers, Python Programming From Beginning to Practice 2nd Edition, Beijing: People's Posts and Telecommunications Press, 2021.
14. W.Chen, Data Visualization, Beijing: Electronic Industry Press, 2019.
15. W.McKinney, Data Analysis with Python, Beijing: Machinery Industry Press, 2018.