

Research Article

Prediction of Factors Influencing the Starting Salary of College Graduates Based on Machine Learning

Ping Wang,1,2 Wensheng Liao,1 Zhongping Zhao,3 and Feng Miu4

1School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, China
2The Key Laboratory of Financial Intelligence and Financial Engineering of Sichuan Province, China
3Student Career Planning and Career Guidance Center, Southwestern University of Finance and Economics, Chengdu, China
4School of Artificial Intelligence and Law, Southwest University of Political Science & Law, Chongqing, China

Correspondence should be addressed to Feng Miu; miu_feng@swupl.edu.cn

Received 19 April 2022; Revised 24 May 2022; Accepted 9 June 2022; Published 26 July 2022

Academic Editor: Kuruva Lakshmanan

Copyright © 2022 Ping Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

It is an important deployment of the Party Central Committee and the State Council to fully promote the employment of college graduates with higher quality, and salary is an important indicator of quality measurement. This paper takes the cross-sectional data of the employment of graduates from a financial and economic university in 2020 as the sample; whether the actual starting salary is a high salary as the dependent variable; and human capital, social capital, labor market as the explanatory variables and uses R to establish a logistic regression model to analyze the determinants of the high salary of graduates. Five machine learning methods, SVM, naive Bayes, CART, random forest, and XGBoost, are used to predict whether graduates can get a high starting salary, compare the advantages and disadvantages of various methods horizontally, optimize the parameters at the same time, and further enhance the performance of the model. Based on the employment data of graduate students in a university of finance and economics in 2020, this paper makes an empirical study. The study shows that academic qualifications, professional disciplines, employment regions, employment industries, the nature of employment units, gender, and whether they have served as student cadres have a significant impact on whether graduates can get “high salaries.” The main factors affecting the starting salary of graduates are the accumulation of human capital and social capital, but the segmentation of labor market is also the main reason affecting the starting salary of graduates. The prediction results of several models show that the integrated models have better performance than single models, and the XGBoost model is the best, which can help predict whether graduates get high salary.

1. Introduction

Employment is the most important issue for the whole society. Affected by the superposition of many factors such as the COVID-19 and the downward pressure on the economy, the employment situation is more complex and severe. In particular, the employment of university students has always been a social hot spot. As socialism with Chinese characteristics enters a new era, the main social contradiction is between the people’s growing needs for a better life and unbalanced and insufficient development. Employment is an important part of a better life. People not only hope to have a job but also have greater expectations for the quality of employment. The Fourth Plenary Session of the 19th CPC Central Committee proposed to “improve the promotion mechanism conducive to fuller and higher quality employment,” which provides a fundamental basis for achieving fuller and higher quality employment. In November 2013, the Ministry of Education issued the notice of the general office of the Ministry of Education on preparing and publishing the annual report on the employment quality of college graduates, requiring colleges and universities to publish the annual report on the employment quality of college graduates from 2013. In the reports from many universities,
the expression of the indicator of “employment quality” is different, but “employment rate,” “employment satisfaction,” and “starting salary of graduates” are the three indicators mentioned most. One of the most important indicators for a job to provide a decent life is salary, which may be the most direct embodiment of high-quality employment.

So what factors affect the job search results of graduate students? How to improve the employment quality of graduates? How to effectively improve the employment competitiveness of college graduates of finance and economics? Exploring these problems can not only provide an important reference for college students’ career planning and job hunting ability preparation but also help the administrative department of education make scientific decisions on the employment of college graduates. Therefore, this study has important practical significance. In the past, most scholars analyzed employment from the perspectives of human capital and social capital and lacked consideration of the key factor of labor market. This paper integrated the relevant theories of human capital, social capital, and labor market, constructed the influencing factor model, combined qualitative and quantitative research, established the logistic regression model by using R, cooperated with the four machine learning methods of SVM, CART, random forest, and XGBoost, compared the advantages and disadvantages of various methods horizontally, optimized the parameters at the same time, further improved the performance of the model, analyzed the determinants of some graduates’ employment and high-salary attainment, predicted whether graduates can obtain high starting salary, and made an empirical research based on the employment data of post-graduates in a financial university in 2020 through the method of empirical research.

2. Related Research

2.1. Human Capital and Job Seeking Behavior

2.1.1. The Formation and Development of Human Capital Theory. William Petty in the 17th century was the first to regard human capital as a capital that can bring income. He first incorporated the idea of human capital into the calculation of national income, reflecting the statistical role of “human” value in income. Since then, the thought of human capital also appeared sporadically in the theoretical research of classical economics. It was not until Schultz formally put forward a more complete concept of human capital in the 20th century. In the wealth of nations, Adam Smith of the classical economic school also regarded human ability as capital and believed that individual capital could have a positive impact on labor production. There are three main points about the definition of the concept of human capital: (1) human capital is condensed in the workers with labor ability, and the quality and ability of workers are the basic elements of human capital; (2) human capital needs to be invested to reflect economic growth. According to the economic growth research of classical economics, capital is investment, and the accumulation of capital needs to be reflected in the cost of money, time, or energy. (3) Human capital is condensed in the knowledge, experience, and skills of workers. Therefore, human capital is accumulated through the effective allocation of education and research, scientific skills, and health. It is not only the manpower condensed in people but also the capital of economic growth.

2.1.2. Evaluation Indicators of Human Capital. The indicators to measure human capital include health, education, scientific research, and other categories. Summarizing the previous human capital research of domestic scholars, Yue and Zhang [1] summarized the political outlook (party member), academic cadre experience, academic degree (minor or second degree), scholarship, academic achievement, and qualification certificate as human capital variables when studying the influencing factors of job hunting of college graduates. Zhao et al. [2] considered that the employment flow of graduates from highly educated and famous universities is more obvious in view of the imbalance of regional employment of college graduates. On this basis, they proposed to include the type of colleges, majors, and educational levels into the variables of human capital. Li used education level, working years, training times, and vocational skills to reflect the human capital of urban workers [3].

2.1.3. The Impact of Human Capital on Job Hunting. The problem of urban employment will be restricted by human capital factors. The choice of work, job maintenance, or job change are inseparable from the accumulation of human capital. Yue and Bai [4] believe that graduates’ personal human capital accumulation is loved by the labor market, and human capital is still an important factor affecting graduates’ job search results, starting salary, and job satisfaction. The results show that academic achievement, party members, and foreign language certificates are significant factors affecting graduates’ employment, while other factors such as computer qualification certificate, second degree, and student cadre experience are not important factors affecting college graduates’ job search results. Zheng [5] finds that the human capital improved by the study of professional courses and comprehensive quality courses during the period of higher education has a positive and significant relationship with the starting salary of graduates, which makes them easier to find a job successfully. Song et al. believe that the school situation of graduates, such as student achievement and whether they are student cadres, is one of the most important factors for employers [6], because the achievement can reflect the candidate’s potential control over the work to a certain extent, which will also improve the expected salary level. It is generally believed that the situation of college students receiving scholarships represents their learning ability and mastery of professional knowledge. Therefore, graduates with excellent academic performance have more competitive advantages and can increase their confidence in the expected salary.

2.2. Social Capital and Job Seeking Behavior

2.2.1. Formation and Development of Social Capital. In fact, there are still a large number of employment problems of
college graduates that cannot be fully explained by human capital. For example, the parents of graduates can help introduce jobs and so on. Such resources obtained through social networks are social capital. Sociologists were the first to study social capital. With the deepening of research, social capital includes the standardization and tightness of the network between various social subjects in sociological research and economic research. The French sociologist Bourdieu was the first to study. He constructed the theory of social capital from a microperspective and believed that social capital was a collection of actual or potential resources. Bourdieu’s important definition of social capital has two parts. One is that this kind of capital is based on mutual understanding or mutual knowledge, because it can establish social relations only when it knows each other or knows each other’s existence; second, members of social networks have resources that can be exchanged with each other. Social capital is a kind of resource that both sides can provide each other. Putnam puts forward that social capital is the basic feature of a social network from the perspective of social structure and social organization, and its trust, network, and norms are important features. At the same time, the performance of social capital is also different under different social mechanisms. In western society, weak relationship is used by job seekers to realize work allocation, while in eastern society, strong relationship is more important and can even directly affect the acquisition of work.

2.2.2. Evaluation Indicators of Social Capital. Compared with the indicators of human capital, the indicators of social capital in previous studies are basically the same. Liu [7] quantified the social characteristics of graduates as specific indicators, such as the network scale, network diversity, network density, and the total reputation score and career impact score of people based on graduates’ job hunting help. The research shows that among the elements of social capital, network scale, total score of professional reputation, and network density play a positive role in the job hunting of physical education graduates, and the number of units and total score of professional reputation play a negative role in the job hunting of physical education graduates. Under the influence of COVID-19, the competition for job hunting becomes more and more intense, and the importance of social capital is more and more apparent. Wang [8] believes that the indicators of social capital factors are categorical variables, which mainly use the following categories to describe the use of social capital by college students: (1) family economic situation, difficult or not difficult; (2) more social resources in urban students than in township students; and (3) employment information sources, which come from friends and relatives, internships, and others (in addition to obtaining recruitment information from schools and from the media). Qiao et al. [9] provide a quantitative method for studying the measurement of social capital of Chinese college students. Social capital is divided into strong relationship capital and weak relationship capital, and strong relationship capital is divided into talent strong relationship capital and acquired strong relationship capital. When measuring strong relationship capital, Zhao’s social capital method of measuring individual level—nomination generation method—is used, and the person’s social status, reputation, and wealth are evaluated with a five-level scale. At the same time, the network scale was used to measure weak relationship capital, and the subjects were asked to search for people who might help online, and the degree of power, reputation, and wealth were evaluated with a five-level scale.

2.2.3. The Impact of Social Capital on Job Hunting. The impact of social capital on job search is mainly reflected in job search channels. Job search channels refer to the ways and specific ways to obtain jobs. Most studies divide job search channels into three categories: formal channels, informal channels, and government institutional arrangements. The use of social capital is mainly through informal channels, which refers to the employment opportunities and work flow channels obtained by relying on various social relations. Zhong and Wang [10] believed that in the stage of obtaining job search information, if college students obtain employment information provided by social networks such as teachers, classmates, relatives, and friends from their families and schools, this information is difficult to obtain on public platforms, which can have a positive impact on college students’ job search results. Chen [11] found that in the same situation, graduates with strong social capital are more likely to determine employment, social capital has a significant impact on the job search results of poor students, and graduates with strong social capital are more likely to flow into the main labor market.

2.3. Labor Market and Job Seeking Behavior. The impact of the labor market on college graduates’ job hunting is reflected in that graduates’ personal employment intention is the main labor market. When entering the main labor market is not smooth or failed, college graduates tend to wait or continue to compete, resulting in partial unemployment. The reason for the above phenomenon is that the labor market is divided into primary and secondary markets, and most of the main market industries are monopoly industries and public sectors. This kind of industry is characterized by high salary level, good welfare, and large rigid demand, which attract quite a lot of fresh graduates. The main reason why college graduates favor the labor market is the employment cognition brought by higher education, and some experts classify this kind of unemployment as “knowledge unemployment” [12]. Knowledge unemployment is the embodiment of the inability to allocate human capital effectively and reasonably. It refers to the state that the labor force of the people with higher education is not used up. In real life, graduates who have received higher education choosing to avoid employment due to difficulties in entering the main labor market, or graduates who are temporarily unemployed to prepare for the iron rice bowl of the public system, are the main sources of knowledge unemployment.

2.4. Human Capital, Social Capital, Labor Market, and Job Seeking Behavior. Social capital can affect the accumulation of human capital to a certain extent. Living in a family with rich social capital will have a positive impact on the level of
education, foreign language skills, and other interests, so as to increase the accumulation of human capital. At the same time, graduates with rich human capital are more likely to choose units with more social status and resources for employment and have stronger motivation to enter the main labor market. According to the employment situation of college graduates in 2019, Yue et al. [13] put forward an empirical study on the job hunting behavior and starting salary of college graduates combined with the theory of human capital, social capital, and labor market segmentation. The research results show that human capital and social capital have a significant impact on the job hunting behavior of college graduates, and the employment and institutional units in big cities are highly attractive to graduates.

In summary, in the research of analyzing the employment situation of college graduates, a large number of studies have shown that demographic characteristics, human capital, social capital, and labor market have a significant impact on graduates’ job search results. In the COVID-19 stage, the economic growth is slow. Some industries are developing hard, such as tourism and foreign trade. This makes the problem of college students’ employment more and more prominent. Moreover, the gross enrollment rate of higher education in China has exceeded 50% and has entered the popularization stage of higher education. Based on this, solving the employment problem of college graduates is the top priority. Therefore, it is of great practical significance to explore the employment characteristics and influencing factors of employment quality of college graduates in the new era.

3. Study Design and Variable Definition

3.1. Research Framework and Assumptions. Through the comprehensive analysis of the above literature, this paper studies the impact of human capital, social capital, and labor market on the starting salary level of college students, analyzes the regression results, and gives conclusions and suggestions. The above independent variables and dependent variables were analyzed by using R language and logistic regression.

In terms of human capital, there are self-influencing factors, discipline categories, achievements, political outlook, etc. In terms of social capital, starting from the family income of college students, the education level of parents, and the location of family income, comprehensively consider the family economic situation, family location, household registration type, and practice platform, as well as the source of obtaining employment information. With the help of the theoretical point of view of the influencing factors of college graduates’ starting salary, exploring the relevant factors of college students’ starting salary, we can find that they do not exist in isolation but interact with each other. Therefore, the impact of college graduates’ starting salary cannot be explained by a single model or theory. The diversity of influencing factors determines the diversity of college students in the process of career exploration.

Based on the discussion of the literature, this study infers the following assumptions:

H1: there is a positive and significant relationship between the human capital promoted by the study of professional courses during the period of receiving higher education and the starting salary of graduates.

H2: there is a positive and significant relationship between family social capital and graduates’ starting salary.

H3: there is a significant relationship between regional differences in employment and the starting salary of graduates.

H4: there is a significant relationship between employment industry differences and graduates’ starting salary.

H5: there is a significant relationship between student work and graduates’ starting salary.

H6: there is a significant relationship between gender and starting salary of graduates.

3.2. Source and Composition of Sample Data. This paper takes the 2020 graduate students of a financial college in Sichuan as the research object. The original data comes from the data collection of the employment management system platform of the management functional department of a financial university. The data filling time is from October 2019 to December 31, 2020. There are 2604 graduate student samples in 2020. The data used mainly involve the basic situation of graduates, number and scale, employment rate, graduation destination, regional flow direction, scholarship, computer certificate, whether to be a class cadre, etc. Excluding the unfulfilled employment recommendation form, as well as the data of no employment, going to school, and going abroad and other data, 1244 valid samples were finally obtained.

3.3. Definition and Processing of Variables. In this study, the explanatory variable wage is defined as the starting salary of graduates, which is the salary data reported by the students themselves in the employment scheme of some financial college graduates. The salary grades are mainly divided into 0-3000 RMB, 3000-4000 RMB, 4000-5000 RMB, 5000-6000 RMB, 6000-7000 RMB, 7000-8000 RMB, 8000-9000 RMB, 9000-10000 RMB, 12000-15000 RMB, and more than 15000 RMB. According to the announcement of Chengdu Municipal Bureau of Statistics on the average wage of employees in all urban units in 2020 issued on June 24, 2021, the average wage of employees in all urban units in Chengdu was 83556 RMB and the average monthly wage was 6963 RMB in 2020. Since the university where the sample data of this study were obtained is located in Chengdu, we define 50% (about 10000 RMB) above the average wage level as high salary, which is assigned as 1, and below this level as ordinary salary, which is assigned as 0.

The independent variables of this study come from four dimensions: human capital, social capital, labor market, and other control variables [14]. The main descriptions are as follows:

(1) Human Capital Factors

(a) Discipline categories and qualitative variables: the categories are divided into engineering,
economics and management, law, literature, history, philosophy, education, and agronomy. The original data are assigned 1–4, respectively.

(b) Political outlook, nationality, educational background, and qualitative variables: the original data are assigned 1–0 according to whether they are party members, ethnic minorities, doctors, and masters.

(c) In terms of academic level, scholarships are divided into 1–2, 3–5, and more than 5 levels, with 0–3 points. Achievements in science, technology, and competition are divided into 4 levels: 1–2, 2–3, and more than 3, with 0–3 points. In terms of scientific research and published papers, published ones are assigned 1 and unpublished ones are assigned 0.

(d) If the English ability reaches a certain level (passing CET-4 and CET-6 tests in China or TOEFL, IELTS, or other English tests of higher level), the variable is assigned 1; otherwise, it is assigned 0. Regarding computer level, having a computer certificate is assigned 1 and having no certificate is assigned 0. With regard to work internship experience, if there is work internship experience, the value is 1, and if there is no work internship experience, the value is 0.

(2) Social Capital Factors

(a) With reference to the urban household registration, the household registration nature and qualitative variable are assigned 1; otherwise, they are assigned 0.

(b) If the family is not in financial difficulties, the family economic status variable is assigned 1; otherwise, it is assigned 0.

(3) Labor Market Factors

(a) Employment area and qualitative variable: areas include Beijing, Shanghai, Guangzhou, and Shenzhen, other new first-tier cities, and other cities. The original data are assigned 1–4, respectively.

(b) Nature of employment unit and qualitative variable: employment units include party and government organs, state-owned enterprises, foreign-funded enterprises, military forces, scientific research and education institutions, and other units. The original data are assigned 1–5, respectively.

(c) The nature of employment industry and qualitative variables: the first category includes information industry and financial industry, the second category includes government and public management, the third category includes education, scientific research and technical services, culture and entertainment, health, and social work, and the fourth category includes construction, real estate, and other industries. The original data are assigned 1–5, respectively.

(4) Control Variables

(a) Gender and qualitative variables: gender is divided into men and women, and the original data are assigned 1–0, respectively. Although some studies classify gender as human capital, this paper believes that gender is a preassigned personal attribute, so it does not belong to human capital or social capital in employment.

(b) Student work and qualitative variables: whether to be a student cadre or not, the original data are assigned 1–0, respectively. There have always been arguments on whether college students serving as student cadres belong to human capital or social capital. Some scholars believe that serving as student cadres belongs to human resources [15], because serving as student cadres can improve students’ comprehensive quality, communication ability, and overall planning ability, and planning ability. Some scholars believe that being a student cadre will increase the social communication opportunities of college students and broaden the breadth and depth of their social network, so as to increase the access to information and enhance the possibility. Therefore, in the data analysis, this paper takes whether students have done student work as the control variable. The description of the indicators is shown in Table 1.

Since the computer can only recognize numbers, the character data described in Table 1, such as “political outlook” and “nation,” are coded. The subsequent individual models need that the target variable must be a factor type, so the “wage” field is converted to a factor type. In addition, because the value of input variables will directly affect the output results, the machine learning model will automatically give different weights to different values, so it is unscientific to directly use coded values for discrete values. Based on the data set obtained by the above method, this paper uses the “get_dummy” method to encode the discrete numerical columns “majorType,” “area,” “nature,” and “industry,” so as to avoid the influence of numerical size on the model.

3.4 Data Exploration and Descriptive Analysis. Firstly, the corrrplot package is used to analyze the correlation between variables. The results are shown in Figure 1.

As shown in Figure 1, it can be preliminarily seen that gender, professional disciplines, nature of employment unit, employment industry, employment area, and educational background are strongly related. In other words, the starting salary is likely to be determined by gender, professional disciplines, the nature of employment units, employment
Table 1: Definition of explanatory variables.

<table>
<thead>
<tr>
<th>Category</th>
<th>Specific variables</th>
<th>Variable meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human capital</td>
<td>majorType</td>
<td>Four categories: engineering, economics and management, law, literature, history, philosophy, education, and agronomy</td>
</tr>
<tr>
<td>Baseline personal information</td>
<td>Political</td>
<td>Party member = 1, else = 0</td>
</tr>
<tr>
<td></td>
<td>Nation</td>
<td>Han nationality = 1, else = 0</td>
</tr>
<tr>
<td></td>
<td>Education</td>
<td>Master, doctor</td>
</tr>
<tr>
<td>Academic level</td>
<td>Scholarship</td>
<td>None = 0, one or two = 1, three to five = 2, more than five = 3</td>
</tr>
<tr>
<td></td>
<td>Race</td>
<td>None = 0, one or two = 1, two to three = 2, more than three = 3</td>
</tr>
<tr>
<td></td>
<td>Paper</td>
<td>Published papers = 1, else = 0</td>
</tr>
<tr>
<td>Job hunting ability level</td>
<td>English</td>
<td>CET6 and CET4 certificate or above, TOEFL, IELTS = 1, else = 0</td>
</tr>
<tr>
<td></td>
<td>Computer</td>
<td>Have computer certificate = 1, else = 0</td>
</tr>
<tr>
<td></td>
<td>Experience</td>
<td>Work internship experience</td>
</tr>
<tr>
<td>Social capital</td>
<td>Residence</td>
<td>Town registered residence = 1, else = 0</td>
</tr>
<tr>
<td>Family situation</td>
<td>Economy</td>
<td>Family financial not difficult = 1, else = 0</td>
</tr>
<tr>
<td>Labor market</td>
<td>Area</td>
<td>Four employment area categories: Chengdu, Chongqing, the first-tier cities of Beijing, Shanghai, Guangzhou, and Shenzhen, other new first tier cities, and other cities (reference group)</td>
</tr>
<tr>
<td>Employment company</td>
<td>Nature</td>
<td>Five categories: party and government organs, state-owned enterprises, foreign-funded enterprises, scientific research and education institutions, and others (reference group)</td>
</tr>
<tr>
<td></td>
<td>Industry</td>
<td>Five industry categories: the first category includes information industry and financial industry, the second category includes government and public management, the third category includes education, scientific research and technical services, culture and entertainment, health, and social work, and the fourth category includes construction, real estate, and other industries (reference group)</td>
</tr>
<tr>
<td>Other control variables</td>
<td>Gender</td>
<td>Male = 1, female = 0</td>
</tr>
<tr>
<td></td>
<td>Cadre</td>
<td>Served as a student cadre = 1, no = 0</td>
</tr>
</tbody>
</table>

industries, and employment areas. This is consistent with our previous literature review.

Use the ggplot package to draw a density map to explore the relationship between technology, competition, scholarship, and whether to get a high starting salary. The results are shown in Figures 2 and 3, indicating that these two have no significant impact on whether to get a high starting salary.

The ggplot package is used to draw a percentage stack diagram to explore the relationship between gender, employment industry, employment region, education, and starting salary. The results are shown in Figure 4.

As shown in Figure 4, male graduates are more likely to get a high salary, more than 25%. In terms of employment industry, information industry, financial industry, construction industry, and real estate industry are related to high salary, up to 25%, and the probability of government and public management getting high salary is less than 5%. In terms of employment areas, the proportion of high salary in Chengdu and Chongqing is about 35%, which may be related to the fact that the sample source is Sichuan and the main place of employment is Chengdu and Chongqing. The probability of getting a high salary for a doctor’s degree is higher than that for a master’s degree graduate, indicating that education is the main factor affecting a high starting salary.

4. Empirical Research Process and Test Results

4.1. Interpretation Model Establishment

4.1.1. Logit Regression Model. Linear regression is one of the most basic statistical models in social research, which is usually used to measure the linear relationship between numerical variables. However, when the explanatory variable or the explained variable is a classified variable, such as the dependent variable which is a continuous variable with spacing measurement level, violating this article will lead to very unreasonable estimation results that are no longer applicable. It is necessary to consider setting dummy variables or selecting other classification models [16].

A logistic model is a model that studies multiple discrete variables and can carry out logistic regression analysis. It is one of the models of a discrete selection method. Its probability expression is explicit, and the model has fast solution speed and convenient application. The model can compare and test the prediction results, which overcomes the limitation that the previous model can only predict the data in
the sample. Therefore, this study selects the logistic model to study the factors affecting the starting salary of college graduates in finance and economics and find the main factors affecting the starting salary of college students.

There are three kinds of logit regression: binary logit regression, multiclassification logit regression, and ordered logit regression (also known as ordinal regression). The difference between these three methods lies in the data type of dependent variable $y$, as shown in Table 2.

Since the dependent variable starting salary in this study mainly distinguishes whether it is “high salary,” the binary logistic regression model is used. The specific model is shown as follows:

$$
\logit \left( p_i \right) = \ln \left( \frac{p_i}{1-p_i} \right) = \alpha_i + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n,
$$

$$
P(Y_i = 1|X_i) = \frac{1}{1 + e^{\alpha_i + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n}} \quad i = 1, 2, \ldots, n,
$$

where $Y$ indicates whether the starting salary of graduate students is high, $X$ is the independent variable, $\alpha_i$ is the intercept of the model, and $\beta_i$ is the coefficient in front of the explanatory variables in the model.

4.1.2. Establishment of the Logit Regression Model Affecting the Starting Salary of Postgraduates. In the regression model, sometimes the introduction of some variables cannot improve the performance of the model but may damage the performance. Therefore, it is a necessary step to reduce the complexity of the model and select model variables. In this paper, stepwise regression is used to select appropriate variables and analyze the influence of independent variables on dependent variables.

The basic idea of stepwise regression is to introduce variables one by one. When each variable is introduced, the selected variables should be tested one by one. When the original introduced variables become no longer significant due to the introduction of later variables, they are eliminated. This process is repeated until neither significant variables are selected into the equation nor insignificant independent variables are removed from the regression equation. R implemented stepwise regression using AIC. AIC can be expressed in two ways. For the first type, see

$$
AIC = \log L - M,
$$

where $\log L$ is the likelihood value of logarithmic operation, also known as log likelihood. $M$ is the number of model parameters. The larger the AIC, the better. The idea of this method is to reward the fitting of the model (the greater
the log likelihood, the better) and punish the complexity of the model (the less the model parameters, the better). Because the log likelihood is often negative, the AIC obtained is also negative. For better analysis, AIC is also represented by

$$AIC = \log L - M.$$  \hspace{1cm} (3)

The smaller the AIC, the better here. AIC used in R is the second kind. R uses the GLM function and step function to implement stepwise logistic regression. The step function is as follows:

```r
step(object, scope, direction = c("both", "backward", "forward"), trace = 1).
```

This paper uses backward stepwise logistic regression. A backward search is to build a full model first; that is, use all variables to build a model. Then, gradually remove a variable and calculate the AIC value of the model at that time. Select the model corresponding to the smallest AIC. Repeat the above steps until removing any variables does not make the AIC of the model smaller. This is the final model.

This paper selects three influencing factors (human capital, social capital, and labor market) as explanatory variables for regression analysis, makes statistical description analysis on dependent variables and independent variables, and comes to the conclusion that whether the starting salary of postgraduates is “high salary” has different degrees of correlation with human capital, social capital, and labor market [17]. In order to further analyze the influence relationship between variables, R will be used to solve and test the model, and the optimal basic regression equation will be selected by integrating economic theory and statistical test until the best estimation form of the comprehensive model is obtained.

The regression model is established by using the impact of R on human capital, social capital, and labor market on employment starting salary. The specific results are shown in Table 3. The likelihood ratio (LR) statistic of the model is 175.54, so it is at a given significance level $\alpha = 0.05$; then, reject the original hypothesis, indicating that the coefficients of all explanatory variables in the discrete selection model are not all 0, indicating that the selection of independent variables is reasonable. Mcfaddentr2 = 0.2019815, indicating that the fitting accuracy of the model is ideal and has passed the statistical correlation test.

4.2. Establishment of the Prediction Model. After preprocessing the data, this paper successively uses support logic regression (logit), naive Bayes, support vector machine (SVM), regression tree model (CART), random forest, and XGBoost to establish models and records the performance of the training set and test set.
SVM maps vectors to a higher dimensional space, in which a maximum interval hyperplane is established. It is a strong classifier with high prediction accuracy. CART is a widely used decision tree learning method, which can be used for both classification and regression. It assumes that the decision tree is a binary tree, and the values of internal node features are “yes” and “no.” Such a decision tree is equivalent to recursively bisecting each feature, dividing the input space, that is, the feature space, into finite units, and determining the predicted value or category on these units. The naive Bayes method is a corresponding simplification based on the Bayesian algorithm. It is characterized by a simple algorithm and less estimation parameters. A random forest is a classifier containing multiple decision trees, and its output category is determined by the mode of the category output by individual trees. It has the advantages of high accuracy and not easy to overfit. XGBoost is an implementation of a boosting algorithm. The main advantage is that it is easy to use in various competitions, and it is also widely used in industry. The above models are widely used, so this paper selects these methods to predict whether graduates can get high salary.

SVM maps vectors to a higher dimensional space, in which a maximum interval hyperplane is established. It is a strong classifier with high prediction accuracy. CART is a widely used decision tree learning method, which can be used for both classification and regression. It assumes that the decision tree is a binary tree, and the values of internal node features are “yes” and “no.” Such a decision tree is equivalent to recursively bisecting each feature, dividing the input space, that is, the feature space, into finite units, and determining the predicted value or category on these units. The naive Bayes method is a corresponding simplification based on the Bayesian algorithm. It is characterized by a simple algorithm and less estimation parameters. A random forest is a classifier containing multiple decision trees, and its output category is determined by the mode of the category output by individual trees. It has the advantages of high accuracy and not easy to overfit. XGBoost is an implementation of a boosting algorithm. The main advantage is that it is easy to use in various competitions, and it is also widely used in industry. The above models are widely used, so this paper selects these methods to predict whether graduates can get high salary.

Table 2: Classification and comparison of logit regression models.

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Dependent variable</th>
<th>Research method</th>
<th>Research scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative data/categorical data</td>
<td>Categorical data (2 groups)</td>
<td>Binary logistic regression analysis</td>
<td>Influence relationship between data</td>
</tr>
<tr>
<td>Quantitative data/categorical data</td>
<td>Categorical data (multigroup and unordered)</td>
<td>Multiclass logistic regression analysis</td>
<td>Influence relationship between data</td>
</tr>
<tr>
<td>Quantitative data/categorical data</td>
<td>Categorical data (multigroup and ordered)</td>
<td>Ordinal logistic regression analysis</td>
<td>Influence relationship between data</td>
</tr>
</tbody>
</table>

Figure 4: Relationship between gender, employment industry, employment area, educational background, and starting salary.
The calculation methods of each indicator are as follows:

\[
\begin{align*}
\text{precision} &= \frac{TP}{TP + FP}, \\
\text{recall} &= \frac{TP}{TP + FN}, \\
\text{accuracy} &= \frac{TP + TN}{TP + FN + FP + TN}, \\
F1 &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.
\end{align*}
\]

Precision refers to the prediction results. It indicates how many samples are real in the positive prediction. Recall is based on our original sample. It indicates how many positive examples in the sample have been predicted correctly. When we evaluate, of course, we hope that the higher the precision of the search results, the better. Meantime, the higher the recall, the better. But in fact, the two are contradictory in some cases. Therefore, precision and recall indicators sometimes have contradictions, which requires comprehensive consideration. The most common method is F1 measure, which evaluates an indicator by calculating the F1 value.

4.3. Preliminary Evaluation of the Prediction Model. The preprocessing of all models is the same, and the only difference is that the models used are different. After the model is run, the "confusionMatrix" is used to get the table of confusion matrix and calculate each index. After preprocessing, 1244 pieces of data remain in this data set. After dividing the data set, there are 879 pieces of data in the training set, of which 0

<table>
<thead>
<tr>
<th>Human capital factors</th>
<th>Coefficient</th>
<th>Z statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional discipline: engineering (majorType_gx)</td>
<td>1.8936</td>
<td>2.7984</td>
<td>0.0051**</td>
</tr>
<tr>
<td>Professional discipline: economics and management (majorType_jg)</td>
<td>0.6250</td>
<td>1.4206</td>
<td>0.1554</td>
</tr>
<tr>
<td>Professional discipline: law (majorType_fl)</td>
<td>-0.0589</td>
<td>-1.037</td>
<td>0.9174</td>
</tr>
<tr>
<td>Political</td>
<td>-0.2314</td>
<td>-1.2894</td>
<td>0.1988</td>
</tr>
<tr>
<td>Nation</td>
<td>0.8160</td>
<td>1.8247</td>
<td>0.0681</td>
</tr>
<tr>
<td>Education</td>
<td>-2.3252</td>
<td>-2.8155</td>
<td>0.0049**</td>
</tr>
<tr>
<td>Scholarship</td>
<td>0.0979</td>
<td>0.9823</td>
<td>0.3260</td>
</tr>
<tr>
<td>Race</td>
<td>-0.0743</td>
<td>-0.3356</td>
<td>0.7372</td>
</tr>
<tr>
<td>Paper</td>
<td>-0.0708</td>
<td>-0.4037</td>
<td>0.6865</td>
</tr>
<tr>
<td>English</td>
<td>1.2508</td>
<td>1.0676</td>
<td>0.2857</td>
</tr>
<tr>
<td>Computer</td>
<td>-0.6040</td>
<td>-1.0116</td>
<td>0.3117</td>
</tr>
<tr>
<td>Experience</td>
<td>-0.3400</td>
<td>-1.5925</td>
<td>0.1113</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social capital factor</th>
<th>Coefficient</th>
<th>Z statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residence</td>
<td>0.1828</td>
<td>1.0614</td>
<td>0.2885</td>
</tr>
<tr>
<td>Economy</td>
<td>0.4916</td>
<td>1.1415</td>
<td>0.2536</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Labor market factors</th>
<th>Coefficient</th>
<th>Z statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment areas: Chengdu and Chongqing (area 1)</td>
<td>1.0121</td>
<td>5.2027</td>
<td>0.0001***</td>
</tr>
<tr>
<td>Employment area: north, Shanghai, Guangzhou, and Shenzhen (area 2)</td>
<td>-0.4532</td>
<td>-2.0592</td>
<td>0.0395*</td>
</tr>
<tr>
<td>Employment area: new first-tier cities (area 3)</td>
<td>-0.6648</td>
<td>-2.5076</td>
<td>0.0122*</td>
</tr>
<tr>
<td>Nature of employment unit: party and government organs (nature_dz)</td>
<td>-2.1515</td>
<td>-1.9278</td>
<td>0.0539</td>
</tr>
<tr>
<td>Nature of employment unit: state-owned enterprise (nature_gq)</td>
<td>-1.2605</td>
<td>-7.2174</td>
<td>0.0001***</td>
</tr>
<tr>
<td>Nature of employment unit: three capital enterprises (nature_sz)</td>
<td>-1.1793</td>
<td>-3.5156</td>
<td>0.0016**</td>
</tr>
<tr>
<td>Nature of employment unit: scientific research and education institutions</td>
<td>-1.9485</td>
<td>-3.1840</td>
<td>0.0015**</td>
</tr>
<tr>
<td>Employment industry: real estate and construction (industry_fdc)</td>
<td>0.8063</td>
<td>2.6646</td>
<td>0.0077**</td>
</tr>
<tr>
<td>Employment industry: government and public administration (industry_gg)</td>
<td>-0.3992</td>
<td>-0.3516</td>
<td>0.7251</td>
</tr>
<tr>
<td>Employment industry: finance and information (industry_xxjr)</td>
<td>0.5318</td>
<td>2.3945</td>
<td>0.0166**</td>
</tr>
<tr>
<td>Employment industry: education, scientific research, and technical services (industry_wsjy)</td>
<td>0.3750</td>
<td>0.7003</td>
<td>0.4837</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other factor</th>
<th>Coefficient</th>
<th>Z statistic</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>0.6125</td>
<td>3.5804</td>
<td>0.0003***</td>
</tr>
<tr>
<td>Leader</td>
<td>0.4516</td>
<td>2.3381</td>
<td>0.0194*</td>
</tr>
</tbody>
</table>

Note: *P ≤ 0.05, **P ≤ 0.01, and ***P ≤ 0.001.
is 707 and 1 is 172. The test set has 365 pieces of data, of which 0 is 295 and 1 is 70. The results are shown in Table 5.

4.4. Improvement of the Prediction Model. The F1 of the model on the training set and test set are all not ideal. The number of tags can be obtained by using the table function. By checking the model, it is found that this data set is a very unbalanced data set. You can see that there are 172 positive samples (1) and 707 negative samples (0) in the training data (bysdata.train). Therefore, the unbalanced data set needs to be processed.

To solve the sample unbalance, resampling is performed. According to the sampling method, it is divided into undersampling, oversampling, and combined sampling.

Undersampling reduces the proportion of most samples in the training data, so that the proportion of positive and negative samples is approximately 1:1. Oversampling increases the proportion of a few samples in the training data, so that the proportion of positive and negative samples is approximately 1:1. Combined sampling reduces most samples and increases the proportion of a few samples, so that the proportion of positive and negative samples is approximately 1:1. The results are shown in Table 6.

4.5. Analysis of Empirical Research Results. As shown in Table 6, on the whole, although the integrated model costs more time, its prediction performance is better than that of the single model. From the comparison of integrated models, the XGBoost method has better overall prediction effect (0.925), followed by the random forest method (0.825). The possible reason is that the XGBoost model corrects the error caused by the previous sequence. In the single models, the prediction effect of pruning decision tree (CART) is better than others. It can be seen that the classifier constructed by the tree model can achieve good results in predicting whether graduates get high salary.

To sum up, the following are shown:

1. The impact of different academic degrees and majors on the starting salary of graduates

In terms of the impact of human capital on obtaining high salary, the estimated parameters of education are significantly negatively correlated, indicating that under the condition of keeping other variables unchanged, a doctoral degree has a greater chance of obtaining high salary than a master’s degree, and high education level has a positive impact on graduates’ starting salary. This finding is consistent with the view of human capital theory. At the same time, the actual starting salary of engineering graduates is higher than that of graduates of other disciplines, and the opportunity ratio of agronomy majors in literature, history, philosophy, and education is more significant.

2. The impact of different family economic conditions on graduates’ starting salary

In terms of social capital, it has no significant impact on whether graduates can obtain high salary. Compared with rural residence registration and family economic status, urban residence registration is not sensitive to the impact of obtaining high salary, which is different from the analysis of other authors that “students from rural areas originally had a low living and consumption level, and under the guidance of money concept, they expected a low starting salary [18].”

3. Impact of different labor markets on graduates’ starting salary

(i) In terms of the choice of employment areas, the estimated parameters of employment places in Chengdu and Chongqing are positive, and the P value is very significant, indicating that under the control of other variables unchanged, the employment places in Chengdu and Chongqing have a higher chance of obtaining high wages than other non-first-tier cities, which is consistent with our daily cognition [19], indicating that there is a wide employment space in economically developed areas, and the opportunity of obtaining high wages is correspondingly large. At the same time, we can see that graduates have a strong momentum of obtaining high salaries in Chengdu and Chongqing among the new first-tier cities.

(ii) In terms of the nature of employment units, the estimated parameters are negative and the P values are significant, indicating that the salary of state-owned enterprises, scientific research and education institutions, and foreign-funded enterprises has no great competitive advantage over other enterprises (mainly private enterprises), but the work of state-owned enterprises and institutions is relatively stable, which is related to the current mentality of graduates to seek stability in work and do not blindly pursue high-paying employment.

(iii) In the employment industry, graduates who worked in the real estate industry and finance and information industry are more likely to receive high salaries than those in other industries. The high salary in the real estate industry has always been obvious to all. Even under the premise of the economic downturn in the past two years, the salary level still occupies an advantageous position like an aristocrat. According to the 2019 national salary report released by
Tencent CAITONG and other institutions in 2019, the real estate industry ranks second only to the Internet industry, surpassing the financial industry. It can be seen that the salary of the real estate industry has not decreased due to the bad market situation, and it is still strong in the era when real estate development is praised as bad. In the context of "new finance," the existing financial disciplines and majors respond to the challenges brought by the development of science and technology, society, and economy. The traditional finance and information industry are combined, and the high salary also reflects the social demand for compound talents with both financial skills and professional quality to a certain extent.

Finally, gender has a great impact on salary [20]. Male students have a higher probability of high salary than female students, which also shows that in terms of salary, female students’ expected salary and actual starting salary are lower than those of male students. Graduates who have served as student cadres are significantly more likely to receive high salaries than graduates who have not served, which is also consistent with the employment quality report released by relevant colleges and universities every year and social media reporting.

### 5. Conclusion and Discussion

Through regression and logistic analysis, this paper makes an empirical study on the influencing factors of human capital, social capital, and labor market on the employment starting salary of finance and economics colleges and universities, puts forward a test on how to improve their competitiveness and obtain a high salary, and forecasts whether they can obtain a high salary. The main conclusions of the study are as follows: firstly, education has a great impact on whether to obtain a high starting salary. For example, doctors have a greater chance of obtaining a high salary than
masters, which reflects that personal knowledge and skills are the measurement standard of college students’ salary. Secondly, in terms of labor market, the nature of employment area, employment industry, and employment unit all have a significant impact on whether graduates can obtain high salary. Specifically, Chengdu and Chongqing have a higher chance of getting high salary than other non-first-tier cities. The salary offered by private enterprises for recruiting target talents is more competitive. There is a greater chance of getting high salary in the financial industry and information industry. Engaging in student work also has a significant impact on getting high salary. In addition, gender affects the starting salary of graduates, and the probability of female college students getting high salary is lower than that of men. Although gender is a preendowed personal attribute, which is neither human capital nor social capital in employment, its impact on employment results is still worthy of attention. Finally, the impact of social capital on whether to obtain high salary is not significant. This paper has the following practical guiding significance.

First, embrace “new finance” and enhance the human capital of college students. Human capital refers to the accumulation of knowledge and skills obtained by workers through investment in education, training, and practical experience. This paper verifies that personal human capital will affect the starting salary of graduates and have an important impact on job search results, especially academic education and professional disciplines. Therefore, college students should pay attention to improving their comprehensive quality. In the context of new finance and economics, in response to the urgent needs of the development of science and technology, society, and economy for the compound talents of engineering and technology majors, college students of finance and economics should make career positioning in advance, make career planning, and strive to study engineering majors to improve their employability.

Second, aim at the labor market and find employment in innovative enterprises in the new first-tier cities and the “new finance” industry. According to media reports, there will be about 9.09 million college graduates and 800 thousand overseas job seekers in China this year. Their flow represents the attraction and vitality of a city to a certain extent [21]. The employment report of college graduates in 2021 from multiple recruitment platforms shows that new first-tier cities such as Chengdu, Hangzhou, Xi’an, and Chongqing have sprung up and even tend to catch up with and surpass Guangzhou and Shenzhen. The “four innovations” of the Ministry of Education and the “new finance and economics” innovation proposed by us are innovation based on historical inheritance and problem-oriented innovation. Its most prominent characteristics are problem orientation and “intersection, communication, and blending”, of which the most key is blending [22]. We can say that this is the fourth innovation of knowledge production mode in the new century and new era. Therefore, college students should be guided by the social labor market, actively seek employment in relevant regions and fields, and choose companies with strong innovation ability and research cutting-edge issues for employment.

Third, improve the comprehensive quality of college students and enhance their sense of self-efficacy. The difficulty of college students’ employment and the obstacles encountered in their career development are closely related to self-efficacy. Self-efficacy is an important factor affecting career decision-making and judgment. Colleges and universities can improve the employment situation of college students by effectively intervening in self-efficacy. To improve students’ communication ability, overall planning ability, and planning ability, Bandura’s research shows that the main factors affecting the formation of self-efficacy are the success or failure experience of personal behavior, alternative experience or imitation, verbal persuasion, and emotional arousal. Therefore, in colleges and universities, the self-efficacy of college students can be improved through the following activities: first, by improving personal human capital, students can experience successful experience and obtain ability transfer; second, let graduates realize that social resources do not play a big role in finding “good” jobs for graduates as rumored, and reasonably adjust expectations.

Finally, we should pay attention to the employment of female college students. Focus on the career development and employment of female college students by adding a gender perspective to the course of college students’ career planning; meanwhile, various career planning and employment guidance activities are carried out to improve the abilities of female college students in self-cognition, career choice, job search skills, and so on.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Key Laboratory of Financial Intelligence and Financial Engineering of Sichuan Province.

References


