ONLINE RECRUITMENT PLATFORM WITH CANDIDATE SELECTION USING DATA MINING

Suvan Sadashiv Shirsat1, S. Ganesh Kumar2, Kartik Hemant Sawant3

1Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Chennai, India, ss9990@srmist.edu.in
2Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Chennai, India, ganeshkl@srmist.edu.in
3Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Chennai, India, kh2907@srmist.edu.in

ABSTRACT

This research is aimed at identifying an adequate strategy for candidate selection for an online recruitment platform utilizing Data Mining. The principal point is to streamline the recruitment cycle and get the most ideal outcome for all potential parties involved. While choosing a candidate, marks or CGPA shouldn’t be the only consideration as that does not straightforwardly translate to real-world employment abilities. So, while recruiting, numerous parameters would be considered to choose the most ideal candidate.

Keywords: recruitment platform, Bayes’ theorem, data mining, Artificial Neural Network

I. INTRODUCTION

On-campus placements is an activity that is conducted in all engineering colleges. The companies arrive at the colleges to select the best freshers that are fit for a job at their organization. Campus recruitments usually go in various stages. The first stage is the eligible candidates registering for the placement process. The candidates are then shortlisted according to various factors which traditionally involves ranking based on Cumulative Grade Point Average (CGPA). Next, the other rounds like coding and interviews are further carried out. The first round is the most important and relatively out of control of the students as the shortlisting criteria generally do not consider the various skills that the students might possess. Hence students who have employable skills are ignored in favor of students who are highly skilled and can directly start working without training. Training employee results in the employer bearing all the costs. This could be prevented by recruiting the correct candidates.

Various methods for recruiting the correct candidates have been proposed. Although not all methods can guarantee that the best candidate would always be selected, we can certainly try to achieve the best possible result in that direction.

In India, a million students apply for campus placements each year. Generally, only around 48% of eligible students are placed. Many colleges have rules and regulations regarding accepting multiple offers. This is usually done so that few students don’t take all the offers. This rule is common in private engineering colleges where there are a large number of students and the college takes on the responsibility of placing such a high number of students. The rule generally states that candidates after getting an offer from a company can only apply for those companies where they can get two times the Cost to Company (CTC). This leads to a few candidates not applying for low paying jobs or getting stuck with a low paying job.

This could be prevented if we use a strategy that can determine how much CTC can the candidate expect based on their abilities. This would help the companies get a good pool of candidates for the right profile.

In the past few years, there have been emerging trends in using predictive analytics in numerous fields. One such utilization of predictive analytics is in the field of candidate shortlisting for recruitment purposes. Modern-day
recruitment conducted on campuses around the world involves various processes. Initial shortlisting is done according to company norms which usually factor heavily on CGPA and academic records. The shortlisting is further narrowed down through aptitude and Multiple-Choice Questions (MCQ) tests and later by programming tests and finally, the selected candidates undergo an interview process. This tedious workload can be greatly reduced if the company can obtain a shortlist of best-fit candidates directly through predictive analytics who can then be interviewed.

Usually, colleges and organizations have a large amount of student/employee data in their databases, also the candidates enter a lot of information into online recruiting platforms. To filter this data and extract useful information and patterns from it we can make use of data mining methodologies and techniques. These patterns can be further used to predict suitable candidates from the test data set. Certain jobs required a set number of skills that are relevant to that field. The software industry demands a high level of technical and programming skills with relevant knowledge of the already established as well as new and upcoming technologies. The performance during the job and the output of the software development process are largely dependent on knowledge possessed by the developers about a certain topic. The whole data mining process is described as shown in Fig 1.

These qualities cannot always be accurately determined in the old tedious process of recruitment, but by using ANN classifiers and identifying the relevant attributes, records can be analysed and compared. This would give us a comprehensive ranking according to the most important attributes needed for the specific job role. These selected attributes are then ranked according to their importance. Good, average, and poor classifications will be used to grade a particular attribute of the candidate and software company evaluation approach. After recognizing the important attributes that would have a large effect on the output of the neural network, they can be fed into the ANN that would give an output after going through one or more layers that have been predefined. The structure of an artificial neural network is depicted in Fig 2.
II. STATE OF THE ART (LITERATURE SURVEY)

A. Data Mining

Various techniques like data mining can be used for gathering and extracting useful information from a huge set of unfiltered data. The research was carried out to get the profile data of candidates by using these data analytics techniques during the recruitment cycle. To improve recruitment significant discoveries like statistical means were proposed. Recommender systems could be built by developing these strategies for solving problems like variances in the job market. These new methods can be deployed in a production environment to test in real-world use cases.

New skills are needed by the different organizations in the job market and to satisfy their needs a methodology was proposed that considers the trends in the job market. Machine learning strategies can be used and the whole process is automated. Systems to recommend jobs to applicants have been directly talked about where various boundaries are utilized for the same.[2]

Another important factor is to get data from resumes. Multiple problems to process the resumes and determine the group of suitable resumes from a huge collection of characterized lists of resumes exist. A grading system can be used to grade the resumes so that the accurate resume has been selected that fits all the criteria. Resumes can be selected based on the characterized work prerequisites followed by featuring the novel highlights. The distinctiveness of a resume includes the skills and other factors that are mentioned like work experience, job duration, etc. If the important factors are picked then this would assist the managers to go through the resumes a lot faster and choose the correct candidate based on their particular trait. The information from the résumés was removed utilizing the VisualText analyzer. Contrasting with other methods where some resumes were excluded due to external factors, this method would include all relevant resumes and exclude the extremely unnecessary results.[3]

B. Big data approach

Big data is an emerging field and its techniques can be widely used for various purposes. Previous research has been conducted in the domain to apply big data for recruitment purposes. A web-based mining and predicting framework was brought forward for customized job and candidate recommendations using big data. To reduce the external load while handling a huge collection of information a tree-based approach could be utilized that is based on contextual learning that can group it in clusters. This model is additionally versatile for the latest additions including terminating the elements. To ensure the customized suggestion for various clients, a context space was worked from the clients' unequivocal data and dynamic partition in each round. In this way, the compensations from comparable clients would be utilized to grade the tree structure and subsequently solve any further problems that occur.[4]

In another paper, a strategy for campus recruitment fair effectiveness assessment dependent on big data analysis is proposed. The statistical analysis of the big data statistics of the campus recruitment fair is developed. The sample regression analysis strategy is utilized to do the campus recruitment fair effectiveness data of the campus recruitment fair. The statistical attributes of employment fairs depend on the average mutual information clustering method for pattern recognition and feature screening, and the adequacy assessment of campus recruitment fairs dependent on big data analysis is realized. The simulation results show that the strategy is utilized to assess the viability of campus recruitment fairs.[5]

C. Naïve Bayes classifier

"Naïve Bayes classifiers" are a group of straightforward "probabilistic classifiers" which are established using Bayes' theorem with tight separation of concerns between the characteristics. A predictor model was implemented that proposes a proper job position to an applicant dependent on their resume. Resume assessment was done dependent on different perspectives, for example, academic performance and professional experience, education, projects, awards, etc. This classifier can be successfully used to inspect the resumes of various candidates and determine if they are a good fit for a particular position. A 3-stage architecture can be used to prepare the resumes and afterward use the Bayesian classifier to characterize, figure, and anticipate the employability of a particular candidate. In the present time, individuals regularly land low compensation grade positions despite having a ground-breaking profile. This model validates the profiles with their normal compensations. [6]

A decision-making system could be used to evaluate if the candidate is fit for a specific role. In a corporate organization, the decision to hire a particular candidate is taken by the HR department while considering the needs.
demands, and expectations of the organization head. Surveys are used to evaluate the candidate based on the supervisor’s requirements. The first step in the process is to take a gander at work hours required for the job, achievements, abilities, and a decent point of view, which would form the underlying premise for determination. The administrators would carry out the next phase of the process using the surveys conveyed by the human resource department. The dissemination of surveys would obtain evaluation performance metrics. Summary of the evaluation results is handled using the Naïve Bayes formula and outputted as per the measures set by the HR Section.[7]

D. Web based approach and other considerations

Previous studies have been done on college recruitment websites. One such research was by Lan in 2012. The paper focuses on creating a web-based data mining tool. The framework is hypermedia and contains different highlights and assets of WWW. The database design is critical in its working. The paper clarifies the issues which ought to be appended to significance to the cycle of database design. Besides, it expresses the detailed design for the arrangement of college talent recruitment, respectively from the database needs examination, database setup, database table structure design, and different perspectives. The actual working results of the framework show that the plan of the database accomplishes great outcomes. Each function indicator of the framework accomplishes the expected outcome following three months of commissioning.[8]

Another paper by Harris in 2017 defines a comparison of ranked records produced by chief recruiting experts which can be created by three search techniques. A gamified environment setup with crowd workers could be the first technique. Another approach is the searching technique and lastly a strategy which consolidates the searching technique with the weighted feature methodology. These strategies can be assessed in different fields whether a certain type of knowledge is required or not. This examination results that the crowdsourcing technique is most suitable for all kinds of roles that exist in the industry independent of the type of knowledge that is necessary for the particular job. [9]

A paper by Saat in 2011 proposed a decision support system (DSS) framework which can profile potential candidates as per the choice standards and ready to assess the suitability of candidates to be designated dependent on profile matching. The definition of DSS is a framework used to help administrative leaders in semi-organized and unstructured decision circumstances. The framework can be utilized to check the suitability of the candidates and create candidates' profiling reports dependent on the profile similarity between the job profile and the candidates’ profile. The rationale of this paper is to profile the candidates and give more data and direction to assist questioners with being all the more reasonable, objective and predictable in their interviewers and afterward settling on the recruiting choice.[10]

E. LinkedIn and other social media platforms

LinkedIn is a professional networking site where potential employees and companies can create their profiles. Professionals can network with others and seek employment opportunities from their connections. Professionals can include their resume in their profile with their skills, experience, projects, etc.

LinkedIn information can be collected by utilizing LinkedIn 2.0 API that is available for analysis and this data can then be filtered by removing the unwanted data. Further, data can be standardized and sorted by geolocation data of the users that can be extracted from location provider services. At that point, this data set can be clustered and sorted by job title, organization, and geo-location areas utilizing various available clustering algorithms. By breaking down information from LinkedIn, we can likewise plan a predictive framework that suggests the favored job for any individual based on his range of abilities and experience.[11]

To check the qualifications and assess the aptitude of potential candidates a framework can be mechanized and deployed as a web application. The companies would be allowed to post job notifications and the candidates who are interested in that particular job would apply. An online form was used to fill in the resume details which can include multiple characteristics. Along with the resume, the candidates would provide their Twitter username which would be used to fetch the account information of the individual. Using the fetched tweets and other information, the emotional aptitude of the candidate can be assessed. This allows two sources of data; one is Twitter and the other would be the resume provided by the candidate. To evaluate the character traits and the individual, the Five-Factor Model is utilized. Various strategies including machine learning can be used to grade the candidates and
rank them according to the suitability that is defined by the employer who posted the job notification. This would lead to the best and easy selection of candidates from a large pool.[12]

Another paper talks about the effects of personality, emotion, interests, and education on individuals' professional movement. They used clients' data from multiple platforms. From ones' LinkedIn pages, one assembled their professional experience as well as educational background. With the profession stages very much characterized, they process Pearson correlation coefficients to examine the effect of personality, emotion, interests, and education on individuals' career movement across various industries, and career stages. The outcomes show that individuals who climb quickly in their careers share comparable qualities and interests. Their perception additionally uncovers that for every industry and each stage there are interesting attributes that assist individuals with climbing quicker. A promising precision demonstrates the solid segregation of character highlights on career progression prediction tasks.[13]

Using skill appraisal techniques, a framework was proposed which used regression analysis. It is all-inclusive and gives a strong and highly effective apparatus to judge various kinds of evaluation strategies. Also, to evaluate the skills new strategies are brought forward, known as "authority score (AS) and credit-based authority score (CAS)". LinkedIn which is a popular professional connection website uses an endorsement counting strategy that does not consider a wide range of abilities unlike the ones proposed. This was inferred from the testing output. The LinkedIn strategy was compared with the new strategies and the results favored the new strategies. The evaluation of skills is important to judge the credibility of the candidate. Moreover, a structure to assess the evaluation methods was proposed from the regression model, and hence regression functions show a highly suitable fit that agrees with the gathered dataset remembering the outcomes for the exhibition of the skill appraisal strategies. Hence the exact forecast can be made of the skills of the candidates that have applied for a particular job.[14]

III. PROPOSED WORK

The proposed technique would be used to determine if a candidate is suitable for a job or not. This technique would consist of 3 steps. The attributes to be used are listed in Table 1 with their description. The first three attributes are the traditional attributes that are used for judging candidates. The language candidate was included as there is a wide opinion among the public regarding multi-lingual people being more capable with employability probability. The output parameter would be a Boolean that would tell if the candidate is placed or not.

Table 1. Attributes

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 10</td>
<td>Marks scored by the student in 10 class</td>
</tr>
<tr>
<td>Class 12</td>
<td>Marks scored by the student in 12 class</td>
</tr>
<tr>
<td>UG Marks</td>
<td>Marks secured by the student in UG course</td>
</tr>
<tr>
<td>Experience</td>
<td>Number of months of experience working in a company during internships listed on the candidates’ LinkedIn profile</td>
</tr>
<tr>
<td>Skills</td>
<td>Number of relevant technical skills listed on the candidates’ LinkedIn profile</td>
</tr>
<tr>
<td>Certifications</td>
<td>Number of relevant certifications listed on the candidates’ LinkedIn profile</td>
</tr>
<tr>
<td>Projects</td>
<td>Number of projects listed on the candidates’ LinkedIn profile</td>
</tr>
<tr>
<td>Languages</td>
<td>Number of literary languages listed on the candidates’ LinkedIn profile</td>
</tr>
</tbody>
</table>

The first step is data cleaning and normalization. The data is normalized by taking into account the maximum and minimum values of a particular attribute. If the normalized value is greater than 0.5 then it would be classified as good and if less than 0.5 then it would be categorized as bad. The formula to calculate the normalized value \(X_{\text{normalized}}\) is as follows:

\[
X_{\text{normalized}} = \frac{(X - X_{\text{minimum}})}{(X_{\text{maximum}} - X_{\text{minimum}})}
\]  

(1)
The second step involves applying Bayes’ theorem on each attribute so that we get the probability for good and bad with respect to the final result. Subsequent to applying Bayes’ theorem the outcomes were utilized extract the prognostic attributes that have the highest probability of the candidate being good. A 2x2 confusion matrix is used to get the data probability of TP. The odds that a phenomenon will happen are demonstrated by an estimation of somewhere in the range of 0 and 1. Therefore, the most relevant attributes are chosen that have a probability higher than 0.5. Bayes’ theorem is defined as:

$$P(A|C_i) = \frac{P(C_i|A) \cdot P(A)}{P(C_i)}$$  \hspace{1cm} (2)

The probability of the True Positive value in the confusion matrix would be considered to determine if the attribute is relevant enough to be considered as an input parameter for the artificial neural network classifier. The value of TP is given in Table 2.

Table 2. 2x2 confusion matrix

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Actual Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>True positive</td>
</tr>
<tr>
<td>Negative</td>
<td>False negative</td>
</tr>
</tbody>
</table>

The final step involves implementing an Artificial neural network classifier. The ANN classifier would have a neuron for each of the prognostic attributes in the input layer followed by a hidden layer and the final output layer with a single neuron with a Boolean value. If a person is placed then the final output would be true and false if the person has not been placed.

The proposed architecture would be implemented as a web application using the MERN stack to be used for the college recruitment process. Artificial Neural Networks (ANN) would be implemented using the Synaptic.js library.

The dataset for this project was taken from different sources. The dataset of students with their marks and their placement status was taken from university sources. This data was then combined with the data that was extracted from their LinkedIn Profiles. This dataset then went through different data mining techniques which gave us the most relevant attributes that can be used in an Artificial Neural Network to get the output.

There would be two types of users, a company and a student. A company can register on the web portal and post a job update. The student can create their own profile and add all their details. They can then apply for a job post using their pre-filled data. This data would be stored in a MongoDB database. After the deadline, the company can generate the list of candidates ranked based on their compatibility.
For authentication, JWT token-based authentication is used which has several advantages over other methods. The user’s password would be stored in the database in an encrypted form using libraries like Bcrypt.

To show the list of ranked candidates data mining techniques would be implemented in the backend and the list generated would be sent as a response from the API to the frontend. In the backend, there would be a pre-trained Artificial Neural Network classifier that would give the final output for a particular candidate.

IV. IMPLEMENTATION

The entire proposed system would be implemented using JavaScript for easy migration after testing to a web application server with Node.js running in the backend.

A. Data normalization

Data normalization is implemented by taking into consideration the minimum and maximum values for a particular attribute.“Neural-data-normalizer”JavaScript library was used to get the normalized values. This library also provides us with meta data so that it can be used later to normalize singular set of values during testing.

B. Extracting prognostic attributes using Bayes’ theorem

The values in Table III are determined as follows. The likelihood of attribute Class 10 marks being ‘good’ and the student being placed is signified as \( P(A = \text{Good}|C_{\text{Good}}) \). After using Bayes’ theorem, this is represented by:

\[
P(A = \text{Good}|C_{\text{Good}}) = \frac{P(C_{\text{Good}}|A = \text{Good}) \cdot P(A = \text{Good})}{P(C_{\text{Good}})}
\]

To find \( P(C_{\text{Good}}|A = \text{Good}) \), we use the conditional probability formula:

\[
P(C_{\text{Good}}|A = \text{Good}) = \frac{P(C_{\text{Good}} \cap A = \text{Good})}{P(A = \text{Good})}
\]

Hence \( P(C_{\text{Good}}|A = \text{Good}) = \frac{29}{95} = 0.31 \). The probability of \( P(A = \text{Good}) \) is \( \frac{96}{240} = 0.4 \). Similarly, the probability of \( C_{\text{Good}} \) is \( \frac{35}{240} = 0.145 \). Therefore, the value for \( P(A = \text{Good}|C_{\text{Good}}) \) is:

\[
P(A = \text{Good}|C_{\text{Good}}) = \frac{P(C_{\text{Good}}|A = \text{Good}) \cdot P(A = \text{Good})}{P(C_{\text{Good}})} = \frac{0.31 \times 0.4}{0.145} = 0.855
\]

The remaining attribute values in Table III were determined in similar way as above.

From Table III we can see that the ‘Languages’ attribute had probability 0.265 which is less than 0.5 hence it was not considered as a relevant attribute. The prognostic attributes were determined to be: Class 10, Class 12, UG, experience, skills, certifications and projects. These attributes would then be fed into the next step which is the ANN classifier.
Table III. Result of applying bayes’ theorem on the dataset

| Attributes   | Attribute values | Probability P(A|C<sub>i</sub>) |
|--------------|------------------|-------------------------------|
|              | Good             | Bad                           |
| Class 10     | Good             | 0.855                         | 0.532                        |
|              | Bad              | 0.195                         | 0.468                        |
| Class 12     | Good             | 0.821                         | 0.357                        |
|              | Bad              | 0.179                         | 0.646                        |
| UG CGPA      | Good             | 0.749                         | 0.622                        |
|              | Bad              | 0.251                         | 0.378                        |
| Experience   | Good             | 0.691                         | 0.311                        |
|              | Bad              | 0.309                         | 0.689                        |
| Skills       | Good             | 0.641                         | 0.408                        |
|              | Bad              | 0.359                         | 0.592                        |
| Certifications | Good          | 0.693                         | 0.471                        |
|              | Bad              | 0.307                         | 0.529                        |
| Projects     | Good             | 0.703                         | 0.592                        |
|              | Bad              | 0.297                         | 0.408                        |
| Languages    | Good             | 0.265                         | 0.772                        |
|              | Bad              | 0.735                         | 0.228                        |

C. Artificial Neural Network Classifier

The neural network classifier is executed using the “Synaptic.js” JavaScript library which utilizes generalized LSTM-like algorithm. It was built for second order recurrent neural networks.

The number of prognostic attributes derived from the previous step was 7 so we would have 7 neurons in the input layer. The next would be a hidden layer contains another set of neurons. The learning rate of the classifier was set to 0.1 with outputs repeated 10,000 times. This would train the neural network to predict the final class of any test input given to it.

D. Web Application

The whole system was implemented as a web application using MERN stack. It uses REST principles for all APIs written in Express.js and Node.js. The frontend was built using React framework. The database used for this project is MongoDB which is a NoSQL database. There are three collections: jobs, users and profiles. The ‘jobs’ collection stores all the list of jobs and registered candidates for that particular job. The ‘users’ collection stores the authentication data of the users with password encrypted using Bcrypt JavaScript library. The ‘profiles’ collection stores the profile data about each candidate.

The authentication system has two types of users: students and companies. The students can view and register for various jobs as per their choice. The companies would be able to post new jobs and after certain number of applicants they can generate the shortlist that would give them the details about the candidates with their expected compatibility.
V. RESULTS DISCUSSION

We explored the sufficiency of the proposed technique by contrasting it with different algorithms and assessed the anticipated result and anticipated result by the ANN classifier. The dataset was split into two sets, 85% training while the remaining 15% was used for testing.

The proposed technique is compared with other algorithms and techniques like Naïve Bayes, Random Forest and SVM. The final results were added to a confusion matrix and then we tested the results for three standard indicators: accuracy, precision and recall.

A confusion matrix is used for results evaluation which gives us the values that could be used in the three performance indicators. The formulae are as follows:

\[
Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)
\]

\[
Precision = \frac{TP}{TP+FP} \quad (6)
\]

\[
Recall = \frac{TP}{TP+FN} \quad (7)
\]

In the above equations: TP = True Positive, FP = False Positive, TN = True Negative, and FN = False Negative.

From the results in Table IV, we can see that the proposed technique has a higher performance in all three indicators as compares to the other three algorithms.
VI. CONCLUSION

We can conclude from the performance evaluation that the proposed strategy has a superior performance when contrasted with other previous algorithms. This research could be improved further by using more recruitments and can be promptly utilized for recruitment in college on-campus placements. The students would be judged on various factors and not just the marks. This would assist the students with a significantly higher level of skills but low CGPA to get equal opportunities in the recruitment cycle. Utilizing this process is gainful for the students in landing great positions and the companies to get the most suitable candidates.

REFERENCES


