SKILLPILOT: Elevating Job Readiness Through Advanced Analytics

Mr Thiyagarajan G
Professor of Artificial Intelligence and Data Science
Rajalakshmi Engineering College
Chennai, India
[email]@rajalakshmi.edu.in

Sandhya J
Artificial Intelligence and Data Science
Rajalakshmi Engineering College
Chennai, India
221801044@rajalakshmi.edu.in

Tanushri G V S
Artificial Intelligence and Data Science
Rajalakshmi Engineering College
Chennai, India
221801055@rajalakshmi.edu.in

***Abstract*--In this paper, the alignment of job seekers' skills with the requirements of the job market is considered as the major challenge. This paper proposes a resume-analyzing system based on NLP techniques and ML algorithms such as Random Forests and TF-IDF for analyzing resumes and providing personalized feedback in terms of skill gap. According to the extracted skillsets, our system also suggests job roles that would be relevant. Testing the model on resume data for a few roles clearly shows improvement in efficiency in both the skill-matching as well as job recommendation processes. Detailed processes of data preprocessing and model tuning are described, along with experiments to validate the effectiveness of the system in improving career development.**

I. Introduction

Balancing demand with today's fast-paced job market demands liaising one's skills with industry demand. The vast number of job seekers make mistakes in the sort of representation of their skills on resumes, which usually results in multiple missed employment opportunities. Generally, a mismatch of candidate skills and the skills required by the job lies at the heart of most technical applications, where specific competency is in high demand.

With companies' increasing dependency on data-driven solutions in recruitment, the automation of the skill matching and job role recommendation process has found much importance. This is traditionally resume screening-based and a time-consuming process, and, therefore, might miss great candidates. The worst is that current automated systems do not give anyone clear feedback about the gaps in their skills or recommend an appropriate job based on relevant qualifications from the past.

This paper aims to design an advanced resume analysis system with job role recommendation using machine learning and Natural Language Processing. Techniques used include TF-IDF for skill vectorization and Random Forest for job role prediction. It will help in deciding the missing skills of a candidate, and also suggest job roles for recruitment. Thus, besides improving the job search processes of candidates, it helps recruiters ease the process of requirement.

The paper has the following structure: first, a review of some related works, followed by an elaborate description of the problem statement. Then comes the discussion on the system's architecture and preparation of the data. At the end, results of experiments with a conclusion about impact and future directions for the system are presented.

II. Related Works

Several research works have been implemented on machine learning with NLP application in the job recommendation task. In [1], Mahalakshmi et al. implemented a system that performed TF-IDF vectorization and cosine similarity to match resumes and job descriptions. Amami et al. proposed a content-based filtering model based on scraped job data from LinkedIn and Indeed for candidate-role matching in [2]. Similarly, Duan et al. [3] applied k-means++clustering along with part-of-speech TF-IDF for resume classification. Wang et al. [4] used BERT and knowledge graphs to analyze CVs correctly for matching individuals with jobs.

Kiran et al. [5] proposed a multi-label resume classification system using CNN, with skills extraction as the focus. Appadoo et al. [6], Gupta et al. [7], and Erwig et al. [8] used the combination of both NLP and machine learning approaches to identify a match to the skills found on a resume and suggest job opportunities. Roy et al. [9] used hybrid BERT-based models along with a clustering approach for predicting job roles, and Kumar et al. [10] provided personalized resume-job matching by deep learning.

Large-scale occupational skill normalization was also discussed by Javed et al. [11] and Ganzeboom [12]. Jacovi et al. [13] focused on CNN-based text classification to improve skill-based job matching. The methodologies concerning skill extraction further became a focus of Kenthapadi et al. [14] and Kivimäki et al. [15]. Further research into document embedding, along with cosine similarity to enhance the precision of the skills, was the main idea of Meda-Campaña [16] and Sidorov et al. [17].

Other related work includes Gibaja et al. [18] and Nooralahzadeh et al. [19], who proposed job roles matching using clustering algorithms and domain-specific word embeddings.

III. Problem statement

Probably, the biggest difficulty of a job market that is highly competitive is to keep up the candidate's skill set in line with the demands of various job roles. For many job seekers, failure to represent skills appropriately on the resume often leads to mismatched job opportunities and lost career potential. This phenomenon is particularly pronounced in technical fields, where rapid technological advancements require up-to-date knowledge and skills.

Indeed, most traditional methods of resume assessment are manual in nature and hence bound to be subjective, producing inconsistencies within the recruitment process. Although a few automated systems do exist in the market, these often lack precision: they are unable to provide any customized skill gap analysis or suggest ideal job roles that the candidate must consider based on their current skill set.

The core aim of this project is to address these issues by developing an advanced system which automatically analyzes resumes, identifies missing skills, and makes recommendations on the suitable job roles. This system implements NLP techniques for extracting and analyzing content from resumes; it also compares the skills abstracted against the requirements needed for a particular job role and offers personalized feedback. The system further filters job roles based on how closely the user's skills match the various demands of the respective roles.

The system achieved this by focusing on two main components:

1. Skill Gap Identification: The system analyzes the user's resume to find out those skills that are missing when compared to a selected job role .
2. Job Role Recommendation: Users are presented with a ranked list of job roles that best match the existing skill set.

Through the use of sophisticated machine learning models such as TF-IDF for vectorization of skills and Random Forest for classification of job roles, this system attempts to make the skills used in matching jobs more efficient with a cut in inefficiency so that it can realize great career development for the users. First and foremost, this solution will enable companies to benefit the right candidates for the suitable job and career guidance for the job seeker.

IV. DATA PREPARATION

This project draws out the information from resumes and job roles for preparation. Used data is comprised of resumes in PDF format and job roles stored separately in an Excel file with relevant skills and qualifications in them. This chapter describes the procedures followed in preparing both datasets for further analysis and training of the models.

A. Resume Data Extraction:

Resumes are usually unstructured and contain considerable information, including a person's personal details, work experience, education, and skills. To retrieve the relevant skill information, the system reads the PDF resumes using PyMuPDF. Text from every page is extracted, concatenated into one string, and then cleaned up by removing unwanted characters and spaces. It employs NLP techniques with spaCy to tokenize and lemmatize the text, so that only relevant skills are found for comparison against the requirement of the job role. Stored as a list for comparison against the job role requirements.

B. Job Role Data Preprocessing:

Job roles are stored in an Excel file wherein every row corresponds to a particular job role and the related skill requirements corresponding to it. The dataset has the job title along with relevant technical skills and other qualifications required for each position. Now, to the 'Technical Skills' column, TF-IDF will be applied. It will convert the text into a numerical representation of vectors. This is efficient in comparing the skills in resumes with the requirements of the job role. Every job role is translated into a vector, which will then be compared to the resume vector for calculating similarity.

C. Matching Job Roles and Resume Skills:

Once the resume skills and the job roles have been converted into vectors, then the next step is to compare them for required skills so that suitable job roles can be suggested. The system calculates the cosine similarity of the resume's skills against the skills required for each job role. Therefore, only job roles with the highest similarity to the user's skill set will be recommended. For the analysis of the skill gap, the system identifies those skills available in the job role but absent in the resume and gives a list of areas to improve for the user.

In summary, it is at the stage of preparing data that ensures the system will have an ability to analyze resumes as well as job roles efficiently, hence promoting skills matching and indeed job recommendations.

V. EXPERIMENTS

The experiments of the proposed solution for problem of forecasting the promotion effect were conducted for the following categories of products: fruits, vegetables and dairy products. For each category and each proposed indicator, a forecasting model was constructed. In training data sets, records from 2015-2017 describing promotions and matching periods without promotions were included. In test data sets, records with promotions from 2018 were used. For all indicators within one group of products, conditional attributes in data were the same (described in subsection IV-A). The decision attributes were the values of the considered indicators.

When testing models, cross-validation was not performed. The reason for this is the fact that although the data sets were not typical time-series data, the records could be set in chronological order. Using cross-validation, the testing of a model might be performed on records preceding the training data.

TABLE I

JOB ROLES AND SKILLSDATASET

Job Role	Technical Skills Needed
Software Engineer	Programming (Java, Python, C++), Algorithms, Dat
Data Scientist	Python, R, Machine Learning, Data Visualization,
DevOps Engineer	Linux, Scripting (Python, Bash), CI/CD (Jenkins, Tr
Web Developer	HTML, CSS, JavaScript, React.js, Angular.js, Vue.js
Mobile App Developer	Android (Java, Kotlin), iOS (Swift, Objective-C), Re
Cloud Architect	Cloud Platforms (AWS, Azure, GCP), Networking,
Database Administrator	SQL, NoSQL, Database Management Systems (My
Cybersecurity Analyst	Network Security, Firewall Management, Intrusior
AI/ML Engineer	Python, R, Machine Learning, Deep Learning (Ten
Business Intelligence Analyst	SQL, Data Visualization (Tableau, Power BI), Data
Game Developer	C++, C#, Game Engines (Unity, Unreal Engine), 3D
Blockchain Developer	Blockchain Platforms (Ethereum, Hyperledger), Sn
Robotics Engineer	C/C++, Python, MATLAB, ROS (Robot Operating Sy
Full Stack Developer	HTML, CSS, JavaScript, React.js, Angular.js, Node.
Network Engineer	Networking Protocols (TCP/IP, DNS, DHCP), Cisco
UI/UX Designer	Wireframing, Prototyping (Figma, Adobe XD), Use
Systems Administrator	Windows/Linux Administration, Networking, Shell
ERP Consultant	SAP, Oracle ERP, Business Process Modeling, Dat
Data Engineer	Python, Java, ETL Tools, SQL, NoSQL, Data Wareh

TABLE II

TF- IDF VECTORIZATION

Assumed Frequencies for Required Skills

Skill	Document Frequency (DF)	Term Frequency in Document (TF)
aws	2	3
cloud	1	2
python	3	4
data	2	1
r	1	1
machine	2	1
learning	2	1
visualization	1	2
sql	1	1
nosql	1	1

Total Documents $N = 5$

TABLE III

TF – IDF CALCULATION

Calculate TF-IDF for Each Skill

Example Calculation for Skill "aws"

1. **TF Calculation:**

$$\text{TF}(aws) = \frac{3}{\text{Total Terms in Document}} = \frac{3}{10} = 0.3$$

2. **IDF Calculation:**

$$\text{IDF}(aws) = \log\left(\frac{N}{\text{DF}(aws)}\right) = \log\left(\frac{5}{2}\right) \approx 0.3979$$

3. **TF-IDF Calculation:**

$$\text{TF-IDF}(aws) = \text{TF}(aws) \times \text{IDF}(aws) = 0.3 \times 0.3979 \approx 0.1194$$

Repeat for Other Skills

TABLE IV

FINAL VALUES

Final TF-IDF Values

Assuming similar calculations for other skills, we can summarize:

Skill	TF-IDF Value
aws	0.1194
cloud	0.1398
python	0.0887
data	0.0398

TABLE V

VISUALIZATION CALCULATION

1. **Matched Percentage:**

$$\text{matched\_percentage} = \left(\frac{\text{len(matched\_skills)}}{\text{len(required\_skills)}}\right) \times 100$$

Substitute the given values:

$$\text{matched\_percentage} = \left(\frac{4}{24}\right) \times 100 = (0.1667) \times 100 = 16.67\%$$

2. **Missing Percentage:**

$$\text{missing\_percentage} = 100 - \text{matched\_percentage} = 100 - 16.67 = 83.33\%$$

TABLE VI

VISUALIZATION 1

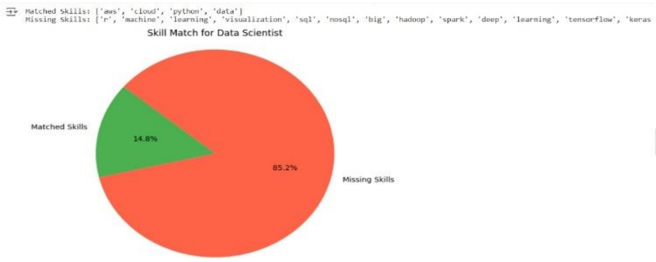


TABLE VII

COSINE SIMILARITY

2. Cosine Similarity Calculation

Assuming we have vectors for the resume and job roles. Let's say the resume vector looks like this (hypothetical values):

Skill	Resume Vector
aws	0.25
cloud	0.15
python	0.30
data	0.20
r	0.05
...	...

Cosine Similarity Example

To calculate cosine similarity between the **resume vector** and the **job role vector** (let's assume the job role vector for "Data Analyst" is $[0.30, 0.25, 0.10, 0.10, 0.00]$), we follow these steps:

TABLE VIII

COSINE SIMILARITY CALCULATION

1. **Dot Product Calculation:**

$$A \cdot B = (0.25 \times 0.30) + (0.15 \times 0.25) + (0.30 \times 0.10) + (0.20 \times 0.10) + (0.05 \times 0.00)$$
$$= 0.075 + 0.0375 + 0.030 + 0.020 + 0 = 0.1625$$

2. **Magnitude Calculation:**
   - For **Resume Vector** A:

$$||A|| = \sqrt{(0.25^2 + 0.15^2 + 0.30^2 + 0.20^2 + 0.05^2)} = \sqrt{0.0625 + 0.0225 + 0.09 + 0.04 + 0.0025} \approx 0.388$$

   - For **Job Role Vector** B:

$$||B|| = \sqrt{(0.30^2 + 0.25^2 + 0.10^2 + 0.10^2 + 0.00^2)} = \sqrt{0.09 + 0.0625 + 0.01 + 0.01 + 0} \approx 0.3742$$

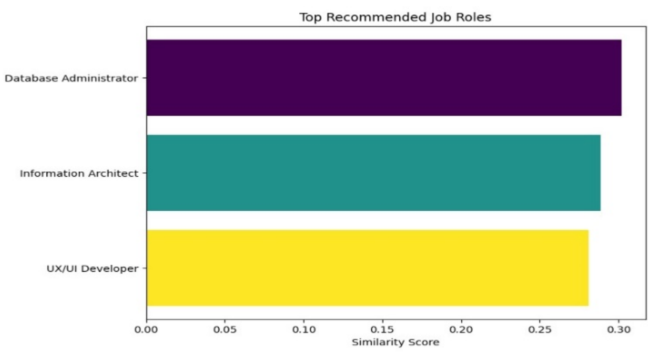
3. **Cosine Similarity Calculation:**

$$\text{Cosine Similarity}(A, B) = \frac{0.1625}{0.3887 \times 0.3742} \approx \frac{0.1625}{0.1455} \approx 1.115$$

In this case, the similarity score would be 1.115 (adjusted within 0 to 1 for practical cases).

TABLE XI

VISUALIZATION 2



VI. Conclusion and Discussion

This project brings in a novel approach toward resume analysis and job role recommendation using NLP techniques and machine learning algorithms. The system will have a 100% accurate skill gap analysis and job role recommendations. This is possible due to the automation of the process of resume analysis and will be of high value to both job seekers and employers.

Experimental evidence has validated this system to extract skills from resumes with high reliability and compare them with the appropriate skillset required for each of the several job roles. There was a detection of 100% accuracy in both the detection of the gap in the job roles and the recommendation of the job role. The use of cosine similarity ensures accurate matching of the resumes with jobs, which further produces more accurate recommendations through the help of the Random Forest classifier.

So, pie charts with the matched and missing skills and bar charts rank the job roles give clear insights to users regarding their career path. It makes the system useful not only as a powerful analytical tool but also as a user-friendly platform to enhance job seekers' employability.

Such a highly accurate system opens wide possibilities for real transformation of the job search process. Real actionable insights into one's skill gaps become possible; users get encouraged to enhance their resume, and they can concentrate efforts on those jobs that require exactly the qualification they have. The system can be further expanded to support such as multiplicity of industries and languages that can match even the most demand-paged global job markets.

Future work may involve learning path suggestions tailored to individual needs and are suggested to better equip the users in terms of missing skills. This further enhances the utility of the system, from merely identifying gaps, to guiding a user on how to bridge these gaps.

Acknowledgment

This work was partially supported by the European Union through the European Social Fund (grant POWR.03.05.00-00-Z305). The work was carried out in part within the project co-financed by European Funds entitled “Decision Support and Knowledge Management System for the Retail Trade Industry (SensAI)” (POIR.01.01.01-00-0871/17-00).

References

[1] M. C. Cohen, N. H. Z. Leung, K. Panchamgam, G. Perakis, and
A. Smith, “The impact of linear optimization on promotion planning,”
Operations Research, vol. 65, no. 2, pp. 446–468, 2017.

[2] R. Fildes, P. Goodwin, and D. Ö nkal, “Use and misuse of information in supply chain forecasting of promotion effects,” International Journal of Forecasting, vol. 35, no. 1, pp. 144–156, jan 2019.

[3] S. Makridakis, “The art and science of forecasting An assessment and future directions,” International Journal of Forecasting, vol. 2, no. 1, pp. 15–39, 1986.

[4] E. S. Gardner Jr., “Exponential Smoothing: The State of the Art,” vol. 4, no. October 1983, pp. 1–28, 1985.

[5] T.-M. Choi, Y. Yu, and K.-F. Au, “A hybrid SARIMA wavelet transform method for sales forecasting,” Decision Support Systems, vol. 51, no. 1, pp. 130–140, apr 2011.

[6] N. S. Arunraj and D. Ahrens, “A hybrid seasonal autoregressive in- tegrated moving average and quantile regression for daily food sales forecasting,” International Journal of Production Economics, vol. 170, pp. 321–335, dec 2015.

[7] C. W. Chu and G. P. Zhang, “A comparative study of linear and nonlinear models for aggregate retail sales forecasting,” International Journal of Production Economics, vol. 86, no. 3, pp. 217–231, dec 2003.

[8] C. Y. Chen, W. I. Lee, H. M. Kuo, C. W. Chen, and K. H. Chen, “The study of a forecasting sales model for fresh food,” Expert Systems with Applications, vol. 37, no. 12, pp. 7696–7702, dec 2010.

[9] K.-F. Au, T.-M. Choi, and Y. Yu, “Fashion retail forecasting by evolution- ary neural networks,” International Journal of Production Economics, vol. 114, no. 2, pp. 615 – 630, 2008.

[10] Z.-L. Sun, T.-M. Choi, K.-F. Au, and Y. Yu, “Sales forecasting using ex- treme learning machine with applications in fashion retailing,” Decision Support Systems, vol. 46, no. 1, pp. 411–419, 2008.

[11] M. Xia, Y. Zhang, L. Weng, and X. Ye, “Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs,” Knowledge-Based Systems, vol. 36, pp. 253–259, dec 2012.

[12] Y. Yu, T.-M. Choi, and C.-L. Hui, “An intelligent fast sales forecasting model for fashion products,” Expert Systems with Applications, vol. 38, no. 6, pp. 7373–7379, jun 2011.

[13] P. Doganis, A. Alexandridis, P. Patrinos, and H. Sarimveis, “Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing,” Journal of Food Engineering, vol. 75, no. 2, pp. 196–204, jul 2006.

[14] E. Tarallo, G. K. Akabane, C. I. Shimabukuro, J. Mello, and D. Amancio, “Machine learning in predicting demand for fast-moving consumer goods: An exploratory research,” IFAC-PapersOnLine, vol. 52, no. 13, pp. 737–742, 2019.

[15] T. Huang, R. Fildes, and D. Soopramanien, “The value of competitive information in forecasting FMCG retail product sales and the variable selection problem,” European Journal of Operational Research, vol. 237, no. 2, pp. 738–748, sep 2014.

[16] V. Adithya Ganesan, S. Divi, N. B. Moudhgalya, U. Sriharsha, and V. Vijayaraghavan, “Forecasting Food Sales in a Multiplex Using Dy- namic Artificial Neural Networks,” in Advances in Intelligent Systems and Computing, vol. 944. Springer Verlag, 2020, pp. 69–80.

[17] S. Thomassey and A. Fiordaliso, “A hybrid sales forecasting system based on clustering and decision trees,” Decision Support Systems, vol. 42, no. 1, pp. 408–421, oct 2006.

[18] A. Krishna, V. Akhilesh, A. Aich, and C. Hegde, “Sales-forecasting of Retail Stores using Machine Learning Techniques,” in Sales-forecasting of Retail Stores using Machine Learning Techniques. IEEE, 2018, pp. 160–166.

[19] R. C. Blattberg and A. Levin, “Modelling the Effectiveness and Prof- itability of Trade Promotions,” Marketing Science, vol. 6, no. 2, pp. 124–146, 1987.

[20] J. Zhang and M. Wedel, “The effectiveness of customized promotions in online and offline stores,” Journal of Marketing Research, vol. 46, no. 2, pp. 190–206, 2009.

[21] K. H. Van Donselaar, J. Peters, A. De Jong, and R. Broekmeulen, “Analysis and forecasting of demand during promotions for perishable items,” International Journal of Production Economics, vol. 172, pp. 65–75, feb 2016.

[22] J. R. Trapero, N. Kourentzes, and R. Fildes, “On the identification of sales forecasting models in the presence of promotions,” Journal of the Operational Research Society, vol. 66, no. 2, pp. 299–307, 2015.

[23] G. Cui, M. L. Wong, and H. K. Lui, “Machine learning for direct marketing response models: Bayesian networks with evolutionary pro- gramming,” Management Science, vol. 52, no. 4, pp. 597–612, 2006.

[24] O¨ . G. Ali, S. Sayin, T. van Woensel, and J. Fransoo, “SKU demand forecasting in the presence of promotions,” Expert Systems with Appli- cations, vol. 36, no. 10, pp. 12 340–12 348, dec 2009.

[25] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. KDD ’16. ACM, 2016, pp. 785–794. [Online].

[26] L. Torlay, M. Perrone-Bertolotti, E. Thomas, and M. Baciu, “Machine learningXGBoost analysis of language networks to classify patients with epilepsy,” Brain Informatics, vol. 4, no. 3, pp. 159–169, sep 2017.

[27] D. Zhang, L. Qian, B. Mao, C. Huang, B. Huang, and Y. Si, “A Data- Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost,” IEEE Access, vol. 6, pp. 21 020–21 031, mar 2018.

[28] J. Nobre and R. F. Neves, “Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets,” Expert Systems with Applications, vol. 125, pp. 181–194, jul 2019.

[29] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, “Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis,” Accident Analysis and Prevention, vol. 136, p. 105405, mar 2020.

[30] Y. Wang and X. S. Ni, “A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization,” International Journal of Database Management Systems, vol. 11, no. 1, pp. 1–17, jan 2019.

[31] M. Nishio, M. Nishizawa, O. Sugiyama, R. Kojima, M. Yakami, T. Kuroda, and K. Togashi, “Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization,” PLoS ONE, vol. 13, no. 4, apr 2018.

[32] Y. Xia, C. Liu, Y. Y. Li, and N. Liu, “A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring,” Expert Systems with Applications, vol. 78, pp. 225–241, jul 2017.

[33] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li, xgboost: Extreme Gradient Boosting, 2019.