IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Job Prediction System Using Machine Learning

Aarshey Srivastava
Student, Dept. of ISE
B.M.S.College of Engineering Bangalore, India

Anuska Devkota
Student, Dept. of ISE
B.M.S.College of Engineering Bangalore, India

Anmol Joshi
Student, Dept. of ISE
B.M.S.College of Engineering Bangalore, India

Amritanshu Amrit
Student, Dept. of ISE
B.M.S. College of Engineering Bangalore, India

Aveek Bose
Student, Dept. of ISE
B.M.S. College of Engineering Bangalore, India

Dr. M Dakshayini
Professor, Dept. of ISE
B.M.S. College of Engineering Bangalore, India

Abstract—

This paper discusses a well-defined comprehensive framework of machine learning, which it focuses on predicting the job-role for an individual based on his or her technical skills, personality, and professional competencies. It has proposed the application of such a system based on data, which involves features related to proficiency in core technical domains, soft skills, and personality attributes to guide people accordingly in making decisions about their career choice.

This predictive model uses feature scaling for varied ranges of input data and incorporates a variety of supervised learning algorithms: Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Tree classifiers. A soft-voting ensemble model is also implemented using these classifiers for better accuracy and robustness of the prediction. The models are both trained and tested using standardized metrics such as accuracy, classification report, confusion matrices, and some error scores in the form of MAE, MSE, R^2.

The system has been proven effective through detailed comparative analysis of individual classifiers and the ensemble model. Confusion matrix visualizations and model accuracy comparisons are presented for validation of performance. The study also gives real-time prediction capabilities where users can input personalized data to receive job role recommendations with associated confidence scores. Further, a large-scale probabilistic analysis is performed using randomized synthetic data to study the overall probability distribution of roles predicted across job categories. Our findings are that the ensemble Voting Classifier is consistently more accurate than any individual model. This work, therefore, offers a scalable data-driven solution for career guidance that can be used in education, workforce development, and recruitment systems. This is also another possible direction for future research in which besides these features, work experience, certifications and industry-specific requirements might add to the sophistication of the model.

INTRODUCTION

Such development in technology, along with the changing nature of job markets, has created an immense demand for personalized career guidance systems. It is only by close analysis of the technical skills and the soft skills associated with the person along with his personality that the right identification of a suitable job role can be made for an individual. Traditional methods have been subjective for a long period of time with career counselling whereas the machine learning operates using data and promises to deliver objective, accurate, as well as scalable solutions to prediction of careers.

This paper proposes a machine learning-based framework for predicting job roles by using a rich dataset with technical and behavioural features. The attributes include proficiency in areas such as cybersecurity, networking, programming, and data science and personality dimensions such as openness, conscientiousness, and emotional range. All these diverse features can then be integrated so that the proposed system can holistically assess an individual's profile and give suitable career paths.

Preprocess the data:

Feature scaling, label encoding are performed for getting data acceptable to ML algorithms. Here are four supervised models implemented, and those are- Random Forest, KNN, SVM, and Decision Trees. Now, using Voting Classifier will further enhance accuracy as well as credibility to the model, using strength from the respective individual models. Each model is assessed by the accuracy, mean absolute error, mean squared error, and R^2 scores and visualizations, such as confusion matrices and accuracy comparison charts. The framework also includes a user-interactive prediction system where a person can input his or her attributes and get the job role with corresponding confidence levels. Finally, a large-scale probabilistic analysis is carried out using synthetic data to study role distributions over several job categories. The results of the ensemble model are uniformly higher than any individual classifier: good prediction and robustness performance. This is an excellent application system for counseling careers, such as for the guidance of students, job seekers, and professionals making career changes in the workforce. It also potentially opens up some opportunities for educational establishments, recruitment, and workforce planning applications. This study leverages the power of machine learning to bridge the gap between individual potential and career opportunities in an increasingly competitive job market.

I. LITERATURE SURVEY

Paper 1: Student Career Prediction Using Machine Learning

Background and Motivation

It is one of the most difficult yet most important decisions that students have to make because they do not have enough knowledge about the job market and their personal strengths. Inefficient and unsatisfactory career choice happens due to a poor career choice. The present study tries to solve these problems using machine learning algorithms to predict an appropriate career for students based on their skills, interests, academic performance, and personal attributes. The ultimate

end is to provide effective career advice to students and enhance the results of employability.

Methodology

Data Set:
A dataset was culled from school records and various student records and comprises attributes that include academic record, interests, programming skills, and extracurricular activities. About 15,000 records were accrued.

Data Preprocessing:

Cleaning:
Removing null values, replacing missing data, and removing outliers.

Encoding:
One-hot encoding was applied in converting categorical to numerical formats.

Organization:
The data was organized for suitability with machine learning models.

Machine Learning Algorithms:
In the paper up for review, some algorithms were evaluated as in the following aspects:
Support Vector Machines (SVM)
Decision Trees
Random Forest
Adaboost Classifiers

Neural Networks Among the tested above, the Random Forest stood out with 89.06% accuracy, so it is best to use the algorithm in forecasting students' direction in life.

Results

Research findings are summed up as the following:

Algorithm Performance:
Random Forest gave the best accuracy in prediction among all algorithms as 89.06%.

Important Features:
In this study, it was noted that programming ability, academic proficiency, and out-of-class extracurricular activities were more important for students' career paths predictions.

Practical Applications:
Based on this model, educational organizations can provide personalized career guidance services and identify learners who require remedial support for their career pathway.

Machine learning can be an effective method of predicting the future career paths of students by evaluating different personal and academic aspects. This paper offers a model that institutions may use to better provide career guidance to students. The future research in this area should address model interpretability and adaptability in the changing dynamics of the job market.

The study illustrates how machine learning may revolutionize education and direct students toward rewarding careers.

Paper 2: Student Career Prediction Using Decision Tree and Random Forest

Background and Motivation

The study stresses the importance to be urgently called for in choosing individual career lines and national growth. In today's highly competitive world, now is the apt time to gauge the student potential and assist him or her with his or her suitable career options. The paper on career choice forecasting using questionnaires is very inadequate, and it is thus sensible to opt for computing techniques based on machine learning.

Methodology

The research uses a machine learning framework, specifically DT and RF classifiers, to predict whether undergraduate students will pursue higher education based on various attributes.

Dataset:

Data was collected from educational institutions, including attributes like age, parents' education, health conditions, and family economic status. Preprocessing steps included handling missing data and removing irrelevant information to enhance accuracy.

Machine Learning Algorithms:

Decision Tree (DT):
The classifier uses decision trees for classification by recursively partitioning the data set based on information gain and entropy.

Random Forest (RF):
This classifies by forming multiple decision trees and then using their output. Its accuracy and strength are much greater than a stand-alone DT classifier.

Implementation:
Both classifiers were implemented in the Python programming environment.

The models were trained and tested on the processed data to compute the accuracy of prediction by the system.

Conclusion

The results that were obtained from the research are as follows:

Performance Metrics:
DT was correct 91%.

RF performed better than DT and got 93%.

Feature Selection:
Features such as family economic status, education of parents, and student health were the predictors of the career outcomes.

Utility:
It is practical utility for educational institutions and recruiters to discover talent and train it further.

Conclusion:
The study from the research concluded that machine learning, in this case RF, is a reliable and efficient method in predicting students' likely career paths. The insights brought forth can be used by educational institutions to improve the training programs and assist recruiters find the appropriate candidates.

This research underlines the transformative potential of machine learning in education and calls for further advancements to address evolving challenges and improve model scalability and interpretability.

Paper 3: Job Shifting Prediction and Analysis Using Machine Learning

Background and Motivation

With an increasingly volatile employment environment, employee turnover becomes a huge challenge for organizations. Organizations cannot retain their best talent and are inefficient in planning their workforce. The critical need for this research is the prediction of job shifts and, more importantly, it presents ML as a good solution to the problem. It leverages structured datasets and advanced ML algorithms to present actionable insights to mitigate workforce disruption for organizations.

Methodologies:

Data:
The training and testing dataset was derived from Analytics Vidhya. It is supposed to contain categorical and numerical features about employees, city development index, experience, education level, the size of companies, and last new job.

Data Preprocessing:

Missing Values:
Missing values were encoded using imputation techniques.

Categorical attributes:
Categorical attributes were transformed into numeric form by using one-hot encoding.

Feature selection:
Feature selection focused on attributes most relevant to predicting job shifts.

Algorithms:
The paper evaluates several machine learning algorithms, including:

Logistic Regression
Random Forest
Gradient Boosting (XGBoost, AdaBoost, CatBoost)
Decision Trees
Naive Bayes
SGDClassifier

Among these, CatBoost, an algorithm developed by Yandex, demonstrated superior performance due to its ability to handle categorical data effectively without extensive preprocessing.

Performance Metrics:
Model accuracy was measured with the ROC-AUC score, which CatBoost acquired at 0.6867.

Results

The main findings of the research are:

Visualization :
Exploratory data analysis showed that patterns such as the connection between experience, company size, and likelihood of job shifts existed.
Algorithm Performance : CatBoost performed a better outcome than the other models, proving it's the right choice for this problem.

Insights:

More likely to quit work people from smaller companies.

Gender differences in turnover trends: males tended slightly more likely to remain rather than females.

Generally, more education and experience predicted lower turnover.

.

Conclusion

Career choice is one of the most important decisions of every student nowadays as they do not have enough knowledge about job markets and where he or she has the skills. This means that wrong career choices lead to dissatisfaction and inefficiency. The research will aim at solving this problem by predicting the appropriate career paths for students based on their skills, interests, academic performance, and personal attributes using machine learning algorithms. In the end, it will guide students in making the right career choice and improve the employability outcome.

II. IMPLEMENTATION

The job prediction system follows a systematic approach to provide personalized career recommendations based on an individual's skills and personality traits. The implementation process is categorized into five major components: data preprocessing, model training, job prediction logic, system deployment, and validation.

1. Data Preprocessing

Data preprocessing is a critical step to prepare raw data for machine learning models. The following steps were performed:

Dataset Composition:

The dataset will have technical skills (such as programming, data science) rated 1–6 and personality traits (such as openness, conscientiousness) rated 0–1. The target variable represents the job role.

Feature Scaling:

There are two different techniques applied to rescale the range of features that have been developed for standardizing feature ranges.
Technical Skills: There was a proportionate adjustment between 1-6 and adjusted to 0-10 scales using StandardScaler to represent an intensity of skills.

Personality Traits:
Rescaled from 0–1 to 0–10 directly using StandardScaler.

This ensures that all features are on a comparable scale so that no feature dominates others while training the model.

Label Encoding:

The Role variable, which is a variable for job categories, was encoded into numerical values using LabelEncoder. This is a transformation that converts categorical labels into integer representations required by machine learning models.

Data Splitting:

The dataset was split into training (80%) and testing (20%) sets using train_test_split to evaluate model generalization.

2. Model Training

The system employs multiple supervised machine learning models for the classification of job roles, trained on the preprocessed data:

Individual Models:

Random Forest Classifier:

This is an ensemble-based model that builds multiple decision trees and combines their predictions. Known for its robustness, it identifies important features contributing to predictions.

K-Nearest Neighbors (KNN):

A distance-based algorithm where it predicts the label based on the majority class of the nearest neighbors.
Optimize the number of neighbors to achieve better accuracy

Support Vector Machine (SVM):

It is a kernel-based classifier and is used for a complex nonlinear relationship in data, and hence the RBF kernel is utilized. Probabilities are enabled in this model so that the predictions made are integrated with the ensemble model. Decision Tree Classifier:

A tree-based model, scalable for interpretable and hierarchical decision-making.

It's simplicity and fast training make it an essential baseline model.

Ensemble Model - Voting Classifier

Uses the soft voting mechanism to combine predictions of all the above models

Probability outputs from different models are aggregated to make a decision

This ensemble reduces variance and improves accuracy in general, as the strengths of each model are exploited.

3. Job Prediction Logic

This module produces predictions for specific users from their input profiles:

Input Format:

Scaled and transformed input features that conform to the preprocessed train data format for that particular model. Users input their scores for technical and personality traits in the form of a structured dictionary.

Prediction Workflow :

The output of each trained model is the user's job role with a confidence score (i.e., the highest probability in the prediction).

The Voting Classifier aggregates predictions from all individual models to produce the final recommendation.

Output:

The predicted job role along with the confidence

score for each model is presented.

This ensures transparency, allowing users to compare results across models.

4. System Deployment

To make the job prediction system accessible and user-friendly, the following steps were planned for deployment:

Interactive User Input:

A Python function was designed that takes in user inputs, makes predictions on the roles, and prints model-specific predictions along with their confidence levels.

Inputs are automatically validated and preprocessed to match the format of the training data.

Visualization:

Confusion Matrices:
Heatmaps for each model show the performance on all job roles

Accuracy Comparison:
Bar plots for the accuracy of individual models as well as the ensemble Voting Classifier

Scalability:

This can be done through the RESTful API in deployment with a web or mobile application to integrate with real-time predictions.

Wide accessibility can be guaranteed by hosting the application on cloud servers with a scalable architecture at the back-end (Flask, FastAPI).

5. Validation

This step verifies that the system is precise, reliable, and robust based on a range of evaluation metrics and analyses, including:

Performance Metrics:

Accuracy- this is the fraction of correct predictions for job roles.

Mean Absolute Error (MAE): Average prediction error.

Mean Squared Error (MSE): Larger prediction errors are penalized.

R^2 Score: Measures how well the model explains variance in the target variable.

Classification reports include precision, recall, F1-score, and support for each job role.

Confusion Matrices:

Visual representations of true vs. predicted labels to identify misclassification trends.

Large-Scale Role Analysis:

Synthetic random data was created to analyze role distribution over predictions.

Probabilities for each job role were calculated using the Voting Classifier in identifying how many different types of roles there would be in mixed cases.

Comparison of the Models:
Individual models were compared to the ensemble model, including precision, error rates, and robustness.

The Voting Classifier always performed more accurately than individual models since it was appropriate for career advising.

This process resulted in a very accurate and interpretable job prediction system. The Voting Classifier was the most reliable solution because it combined the strengths of multiple models. The system is very suitable for deployment in career counseling platforms that provide real-time, user-friendly, and scalable recommendations for a variety of job roles.

III. RESULTS

The performance of the job prediction system was tested across multiple machine learning models and a Voting Classifier, which is an ensemble model, using several metrics. Below is a detailed summary of the results:

1. Individual Model Performance

Random Forest Classifier:

Accuracy: 87.5%

Mean Absolute Error (MAE): 0.1

Mean Squared Error (MSE): 0.23

R^2 Score: 0.85

Key Features:

The model captured well the critical predictors such as programming skills, AI/ML expertise, and communication skills.

Confusion Matrix:

It had very low misclassifications, especially in the technical roles of "Database Administrator" and "Cyber Security Specialist."

K-Nearest Neighbors (KNN)

Accuracy: 84.3%

MAE: 0.15

MSE: 0.28

R^2 Score: 0.81

Observations:

KNN could not classify some of the overlapping feature distributions that happen in similar roles, for instance, "Networking Engineer" and "Software Developer."

SVM

Accuracy: 86.8%

MAE: 0.13

MSE: 0.25

R^2 Score: 0.83

Strengths:

The model performed pretty well with high-dimensional data and maintained a very balanced classification on all the roles.

Weakness:

Computationally costly compared with other models.

Decision Tree Classifier

Accuracy: 82.1%

MAE: 0.18

MSE: 0.31

R^2 Score: 0.78

Observation:

It was an interpretable model that had a tendency to overfit; thus, its generalization performance reduced on test set.

2. Voting Classifier (Ensemble Model)

Accuracy: 89.3%

This is the most accurate of all models, thanks to the complementary strengths of Random Forest, KNN, SVM, and Decision Tree.

MAE: 0.11

Minimum error rate among all the models; the higher, the better for predicting the right output.

MSE: 0.21

It means there are very few bad predictions with higher margins.

R^2 Score: 0.88

It means that it can explain about 88 percent of variance of the target's prediction regarding job role.

Confusion Matrix:

Classification accuracy is similar in both technical and non-technical roles. Very few instances of misclassifications are available. "AI ML Specialist" and "Graphics Designer" have distinct separations between these two classes.

3. Role- Specific Insights

High Prediction Accuracy:

Predictions such as "Cyber Security Specialist," "Software Developer," and "AI ML Specialist" can be done with maximum confidence. There are high definitions in terms of feature importance.

Moderate Prediction Accuracy:

For instance, "Project Manager" and "Business Analyst" were incorrectly categorized at some point due to similarity between their characteristics (such as communication, management, etc.).

Low Prediction Accuracy:

Customer service executive is the role for which maximum confusion was seen probably because of technical theme of the dataset.

4. Visualization and Comparison

Confusion Matrix:

Pointing out the strength of each model and confusion regarding the mistakes.

Voting Classifier was the best balanced for the model prediction of all the roles.

Model Accuracy Comparison:

A bar chart was used for comparison of accuracy across all models, and the Voting Classifier performed the best.

5. Role Distribution Analysis

Synthetic random input data:

The system generated a probability distribution over all job roles.

Key takeaways:

"The Ai ML Specialist" had the highest average likelihood because of high representation of technical skills in the dataset.

Graphics Designer and "Customer Service Executive" had a lower probability due to niche-like characteristics.

Conclusion

The Voting Classifier performed better than the individual models, providing strong and reliable career predictions.

The prediction accuracy was high for technical roles, while the non-technical roles showed minor classification challenges.

The system is scalable and can generate high-confidence predictions for diverse user inputs, making it a valuable tool for career guidance.

IV. CONCLUSION

This paper discusses and compares a number of classification models, namely Random Forest, KNN, SVM, and Decision Tree, while emphasizing strengths and weaknesses for a voting classifier ensemble learning method that is particularly strong. The remainder of this paper summarizes these findings:

1.Effectiveness of ensemble learning

The Voting Classifier was more accurate, being 89.3%, because it combined the strengths of Random Forest, KNN, SVM, and Decision Trees. That is a clear indication that a variety of different classifiers are better at getting good reliability as well as accuracy in finding the right careers to play into.

2.Role of Feature Engineering

Feature scaling and transformation of key technical skills, soft skills, and personality traits enable the models to differentiate between the job roles better. This therefore explains the role of preprocessing and feature scaling in model performance.

3.Advantages of Random Forest

Feature importance of Random Forest can handle complex interaction, and in this case of the ensemble, it did an excellent job if taken probabilistic output.

4.Non-Technical Role Prediction Challenges

It sometimes does not come out clearly with slight inaccuracies while forecasting overlapping roles such as "Business Analyst" and "Project Manager." This once again reflects the fact that it lacks data collection granularity along with the choice of features of non-technical roles.

5.Scalability and Real Applications

It can easily be merged with real-world career counseling and recruitment platforms because of the scalability of the system. It's modularity can adapt for an increase in datasets, real-time inputs, and evolving job markets.

6.Cold Start Scenarios

Cold start scenarios where there is limited amount of data performed well with the models Random Forest and Voting Classifier, thus lessening reliance on long histories of users' interaction. It makes the system very useful for career prediction in the case of new or sparse datasets.

7.Hybrid Models

Adding an NLP flavor to machine learning predictions can be taken further by, for example, examining resumes or open-ended responses for more refined predictions. Feedback loops can also be used to improve adaptability over time.

Conclusion:

This is a framework for a scalable, very accurate, and quite practical prediction of careers. Thus, it presents an open scope for professional development and recruitment strategy research

REFERENCES

Job Role Prediction System
IEEE Conference Publication, 2022.
This paper discusses the development of a machine learning application that predicts candidate eligibility for specific job roles based on a set of features.
IEEE Xplore

An Ensemble Method for Job Recommender Systems
ResearchGate, 2016.
This study presents an ensemble approach to job recommendation, aiming to predict job postings relevant to users by combining multiple models to enhance accuracy.
ResearchGate

Ensemble Learning in Recommender Systems
ACM Digital Library, 2014.
This paper proposes a technique that utilizes multimodal user interactions to generate more accurate recommendation lists, optimized through ensemble learning methods.
ACM Digital Library

Presentation of a Recommender System with Ensemble Learning
arXiv preprint, 2020.
This research employs group classification and ensemble learning techniques to increase prediction accuracy in recommender systems, addressing challenges in user needs analysis.
arXiv

Job Prediction: From Deep Neural Network Models to Applications
arXiv preprint, 2019.
This study focuses on job prediction using various deep neural network models, including TextCNN and Bi-GRU-CNN, and proposes an ensemble model that achieved a high F1 score, illustrating the effectiveness of combining different models.
arXiv

Let's Predict Who Will Move to a New Job
arXiv preprint, 2023.
This paper discusses the use of machine learning algorithms, such as Random Forest and XGBoost, to predict whether an individual will seek new employment, highlighting the application of SMOTE to address class imbalance.
arXiv

A Machine Learning-Based Job Forecasting and Trend Analysis System
IEEE Xplore, 2023.
This study implements a prediction system for job trends, enabling stakeholders to understand and align with evolving job market demands through machine learning techniques.
IEEE Xplore

Hire Prediction System Using Machine Learning
International Journal of Research in Advent Technology, 2019.
This paper presents a computational framework for automatically predicting a candidate's hireability by evaluating various machine learning methods on collected applicant data.
IJRAT

Student Career Prediction Using Algorithms of Machine Learning
SSRN Electronic Journal, 2023.
This article provides a career prediction model using machine learning to assist students in choosing appropriate career paths, emphasizing the application of specific algorithms.
SSRN

Ensemble Learning Based Employment Recommendation Under Interaction Sparsity for College Students (EERIS)
ACM Digital Library, 2023.
This paper proposes a novel model called EERIS, which utilizes ensemble learning to provide employment recommendations for college students, effectively addressing interaction sparsity issues.
ACM Digital Library