IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151162

# FUTURE-AWARE JOB MARKET DEMAND FORECASTING AND RESUME MATCHING USING MACHINE LEARNING

**Shri Nivetha R[1], Kavya S Prasad[2], Jeyaathri V[3], Dr. G. Paavai Anand[4]**

Student, Department of Computer Science and Engineering,

SRM Institute of Science and Technology, Vadapalani, Chennai, India[1-3]

Mentor, Department of Computer Science and Engineering,

SRM Institute of Science and Technology, Vadapalani, Chennai, India[4]

**Abstract:** This paper presents a future-aware job market analytics and resume matching system that combines monthly job-posting demand forecasting with skill-trend analysis and candidate skill-coverage assessment. A global Random Forest regression model is trained using lagged demand, rolling statistics, growth indicators, seasonal features, forecast horizon, series type and job-domain information. A chronological hold-out design is used to evaluate twelve-month-ahead predictions against naïve and seasonal-naïve baselines. Forecasted skill demand is then combined with recent role-specific skill relevance to construct a weighted target skill profile. Resumes are parsed for recognised skills and assessed using a transparent weighted coverage score, while TF-IDF cosine similarity is reported separately as a supplementary text-alignment measure. On the supplied artificial dataset, the model achieved MAE 3.07, RMSE 4.31 and $R^2$ 0.908, outperforming the pooled baselines. These findings demonstrate the feasibility of integrating forecasting and explainable skill-gap analysis, but do not establish real labour-market performance.

**Keywords:** Machine Learning, Random Forest Regression, Job Market Forecasting, Skill Trend Analysis, Resume Parsing, TF-IDF, Skill Gap Analysis.

## I. INTRODUCTION

The labour market changes as organisations adopt new technologies, business models and operating practices. Skills that are frequently requested at one point in time may become less prominent, while emerging tools and competencies may gain relevance. Consequently, a resume assessment based only on the current wording of job descriptions can provide an incomplete view of a candidate's preparedness for future roles. A useful career-support system should therefore consider both the skills associated with a target occupation and the direction in which demand for those skills may be moving.

This work develops a Future-Aware Job Market Prediction and Resume Matching System. It integrates two connected functions: (i) analytics and forecasting of monthly job-posting demand at market, domain, role and skill levels; and (ii) resume analysis that compares a candidate's evidenced skills with a forecast-weighted profile for a selected role. The system is implemented as a web application with a React frontend and a FastAPI backend. Pandas is used for data preparation, while scikit-learn provides the machine-learning pipeline.

The principal contribution is an interpretable workflow that links demand forecasts to a transparent skill-coverage score. Instead of treating resume similarity as a proxy for future readiness, the system separates keyword-based skill evidence, forecast-weighted coverage and document-level text similarity. This separation makes the score easier to explain and reduces the risk of presenting a text-similarity value as a hiring probability.

The study has an important boundary: the supplied dataset contains artificial records generated for software demonstration. The reported metrics therefore evaluate the implementation on that dataset only. They must not be interpreted as estimates of actual hiring demand or as evidence that the system can predict the real labour market.

## II. LITERATURE REVIEW

Random Forest is an ensemble learning method that constructs multiple decision trees using randomised training samples and feature subsets, and combines their outputs to improve predictive stability. It is suitable for tabular data containing nonlinear relationships and mixed feature types. In this project, the estimator is used as a regression model to predict normalised future posting counts.

Time-series forecasting commonly relies on information available at the forecast origin, including lagged observations, seasonal values, rolling summaries and trend indicators. Chronological evaluation is essential because random shuffling

 This work is licensed under a Creative Commons Attribution 4.0 International License 

DOI: 10.17148/IJARCCE.2026.151162

can allow future observations to influence model training and produce overly optimistic estimates. Forecasting practice also recommends comparison with simple benchmarks such as naïve and seasonal-naïve methods.

For document analysis, TF-IDF represents text using term-frequency and inverse-document-frequency weighting. It reduces the influence of terms that occur widely and gives greater weight to terms that are relatively distinctive within a corpus. Cosine similarity can then compare vector representations of two documents. These techniques are computationally efficient and useful for a supplementary resume-to-role text comparison, although they do not fully capture semantic equivalence or skill proficiency.

The proposed system combines these established techniques in an application-oriented workflow. Its distinction is not a claim of a new forecasting algorithm; rather, it is the integration of multi-level demand forecasting, relative skill-trend classification, forecast-weighted role profiles and explainable resume skill coverage within one interface.

## III. METHODS AND MATERIALS

### A. System Architecture

The application follows a client-server architecture. The frontend is implemented using React and TypeScript, with pages for the overview dashboard, dataset exploration, job-demand forecasts, skill trends, model performance and resume analysis. The backend is implemented using FastAPI and exposes REST endpoints for dataset metadata, analytics, forecasts, skill profiles, resume parsing and scoring. Pandas performs data loading and aggregation; scikit-learn implements preprocessing and Random Forest regression; Joblib persists the trained model.

The dataset is read from a user-configured CSV path. The loader validates required columns, dates, salary values and mandatory fields, creates monthly timestamps, identifies geographic regions and expands pipe-separated skill lists into posting-skill events. The original CSV is retained for dataset exploration and is not modified by the application.

### B. Dataset Description and Provenance

The project documentation describes a dataset containing 12,037 records and 14 required fields: job identifier, posting date, job title, job domain, location, experience level, industry, salary minimum and maximum, salary currency, employment type, required skills, preferred skills and job description. Skill lists are represented as pipe-separated values. The documented date coverage is January 2019 to December 2026, corresponding to 96 monthly observations.

Crucially, the project README states that these records were generated artificially by a one-off script and were not collected from real job advertisements or an external source. The dataset is therefore a synthetic demonstration dataset. All quantitative findings in this paper are restricted to this dataset and cannot support claims about actual employers, salaries, occupational demand or future employment conditions. The stated end date also extends beyond the current preparation date (30 October 2026); the authors should verify the file's date range and explicitly document how future-dated synthetic records were created before submission.

The forecasting target is monthly job-posting count for a market, domain, role or skill series. Salary and free-text description fields are not forecasting targets. Series with insufficient history, too few active months or very low recent average volume are excluded from forecasting.

### C. Data Preprocessing and Feature Engineering

Posting dates are converted to monthly periods, and each series is reindexed over the complete monthly calendar with missing months represented as zero counts. Skill demand is calculated from the number of postings that mention a skill as required or preferred. Role, domain and location filters are applied before aggregation.

For a series y and forecast origin t, the model uses information available up to and including t. Numerical features include normalised lags at 0, 1, 2, 5 and 11 months; three- and six-month rolling means; three- and twelve-month rolling standard deviations; a short-term slope; twelve-month growth; log-transformed demand level; and the value from the corresponding month in the previous year. Forecast horizon and sine/cosine encodings of the target calendar month are also included. Categorical features identify the series type and job domain. Missing numerical values are imputed using the median, and categorical variables are one-hot encoded with unknown categories ignored.

### D. Random Forest Forecasting Model

A single global RandomForestRegressor is trained using a direct multi-horizon strategy. Each training row represents a series, an origin t and a horizon h, where h ranges from 1 to 12. The target is the future posting count normalised by the mean count over the preceding twelve months:

 This work is licensed under a Creative Commons Attribution 4.0 International License 



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151162

$$z(t,h) = y(t+h) / L(t), \quad L(t) = (1/12)\ \Sigma[j=0 \text{ to } 11]\ y(t-j)$$

The predicted ratio is multiplied by L(t) to recover the predicted posting count, with negative predictions clipped to zero. Normalisation allows a shared model to learn across series with different demand magnitudes. The implementation uses 80 trees, maximum depth 14, minimum leaf size 30, maximum feature fraction 0.6, row subsampling of 0.25 and random_state 42.

The evaluation split is chronological. The last twelve months are held out, and training examples are restricted so that their target months fall within the training period. The model is evaluated from the final training origin over the twelve held-out months. Series eligibility is determined using the training window only. After evaluation, the same model configuration is refitted on all available months for operational forecasts. This separation is intended to prevent test-period targets from entering training features or training labels.

### E. Forecast Evaluation

Predictions are evaluated using mean absolute error (MAE), root mean squared error (RMSE) and coefficient of determination ($R^2$). Let $y_i$ be the observed count and $\hat{y}_i$ the predicted count for N test observations:

$$MAE = (1/N)\ \Sigma\ |y_i - \hat{y}_i|$$

$$RMSE = \sqrt{[(1/N)\ \Sigma\ (y_i - \hat{y}_i)^2]}$$

$$R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$$

The Random Forest is compared with a naïve forecast, which repeats the latest observed value, and a seasonal-naïve forecast, which repeats the value from twelve months earlier. The implementation also reports results by series type and forecast horizon. Prediction intervals are estimated from empirical quantiles of held-out model errors, grouped by demand level and horizon where sufficient residual observations are available.

### F. Skill Trend Classification

For each skill, current demand is the average monthly number of postings mentioning that skill during the latest twelve months. Historical growth compares the latest twelve-month total with the preceding twelve-month total. Forecast growth compares the average predicted demand over the next twelve months with current demand. To avoid labelling a skill as growing merely because the entire market is growing, skill growth is adjusted relative to the corresponding market or domain growth.

When a valid forecast is available and average current demand is at least three postings per month, the system blends relative forecast growth and relative historical growth. Forecast growth receives a weight of 0.6 and historical growth a weight of 0.4. A blended growth value of at least 0.08 is classified as growing; a value at or below −0.08 is classified as declining; intermediate values are stable. A skill is marked emerging when it is classified as growing and blended growth is at least 0.20. Skills with insufficient volume or unavailable forecasts are not forced into a trend category.

### G. Future Skill Profile and Resume Matching

For a selected role, the system derives skill relevance from postings in the latest 24-month window. Required skill mentions receive a weight of 1.0 and preferred mentions receive a weight of 0.5. Relevance is calculated relative to the number of postings in the selected role. If fewer than 30 postings are available for that role, the system falls back to its broader job domain and records this scope change.

Each skill's relevance is multiplied by a demand factor based on forecast growth relative to the market. The factor is clipped to the range 0.5 to 2.0. The resulting raw weights are normalised to sum to one, and the highest-ranked 20 skills form the target profile. This design combines evidence of current role relevance with a forward-looking demand signal.

The resume parser extracts recognised skills from PDF or DOCX files using keyword and alias normalisation. The future-readiness score is defined as:

$$Score = 100 \times (\Sigma \text{ weights of evidenced target skills}) / (\Sigma \text{ weights of all target skills})$$

The score is bounded between 0 and 100 and is accompanied by a skill-by-skill breakdown. TF-IDF cosine similarity between the resume text and a weighted target-profile text is calculated separately and displayed as resume similarity; it does not modify the coverage score. Recommendations are rule-based and identify relevant skills that are not evidenced in the resume. The score is not a probability of being hired and does not measure depth of competence.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026

DOI: 10.17148/IJARCCE.2026.151162

## IV. RESULTS AND DISCUSSION

### A. Forecasting Performance

The project's executed evaluation reports pooled results across 285 eligible series and twelve held-out months, giving 3,420 predictions. The training dataset for the evaluation model contained 183,786 rows. Table I reproduces the measured results documented by the project.

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| Random Forest Regressor | 3.07 | 4.31 | 0.908 |
| Seasonal-naïve baseline | 3.85 | 5.30 | 0.861 |
| Naïve baseline | 4.29 | 6.00 | 0.821 |

**TABLE I. POOLED HELD-OUT FORECASTING PERFORMANCE**

On the pooled evaluation, the Random Forest achieved lower MAE and RMSE and higher $R^2$ than both baselines. The project reports that the Random Forest achieved a lower per-series MAE than the best baseline for 59% of evaluated series. This indicates that the pooled improvement does not apply uniformly to every series.

### B. Performance by Series Type

| Series type | MAE | $R^2$ | Interpretation |
|---|---|---|---|
| Skills | 3.07 | 0.87 | Strongest reported fit among the listed groups. |
| Roles | 2.11 | 0.64 | Lower absolute error; moderate explanatory fit. |
| Domains | 4.82 | 0.55 | Forecast performance varies with domain scale and history. |
| Whole market | 12.33 | −0.33 | The aggregate series was difficult to forecast and did not outperform a simple fit in $R^2$ terms. |

**TABLE II. REPORTED PERFORMANCE BY SERIES TYPE**

The negative $R^2$ for the whole-market series is an important qualification. It shows that a pooled model can perform well across many series while remaining weak for a particular aggregate series. Moreover, pooled $R^2$ is influenced by large-volume series, whereas small-series metrics can be unstable when calculated over only twelve test observations. The results should therefore be read alongside per-series and by-type metrics rather than as a single universal accuracy claim.

### C. Application-Level Outputs

The system exposes observed historical demand, model-generated forecasts, skill trend classifications, future skill profiles and resume coverage results through separate interface pages. Dataset Explorer supports search, sorting, categorical and numerical filtering, pagination, summary statistics and CSV export. The model-performance page displays model status, evaluation metrics, feature importance and error by forecast horizon, and provides a retraining action.

The resume analysis pipeline produces a target-role profile, a weighted coverage score, a list of evidenced and missing target skills, supplementary TF-IDF similarity and rule-based recommendations. This combination is intended to make the assessment more actionable than a single opaque similarity score. However, keyword extraction can miss synonyms and contextual evidence, and the system cannot infer proficiency from a skill mention alone.

### D. Software Verification

According to the project README, 47 automated backend tests passed and the frontend production build completed successfully in the recorded verification environment. A headless Chromium run completed 18 interface checks covering navigation, dataset search and filtering, pagination, CSV downloads, model and forecast pages, skill trends, resume upload and scoring, with no reported console errors. These results demonstrate software-level verification in the documented environment; they do not constitute an independent validation of forecasting accuracy or a user study. Windows execution and browsers other than Chromium were not reported as tested.



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 11, November 2026
DOI: 10.17148/IJARCCE.2026.151162

### E. Threats to Validity and Limitations

The largest limitation is dataset provenance: the records are artificial. Synthetic data may encode assumptions and regularities that are easier to learn than real hiring behaviour. Accordingly, the measured accuracy cannot be generalised to external labour markets. The documented date range should also be audited because it includes months later than the preparation date of this manuscript.

Random Forest regression is not inherently designed to extrapolate arbitrary long-term trends beyond patterns represented in its training data. Sparse role or skill series may have wide forecast uncertainty, and the whole-market result illustrates that one global model is not uniformly effective. Role-level skill growth may use a domain-level forecast when the role-specific series is not independently forecastable. Resume parsing is keyword-based, does not support scanned image-only documents, and does not estimate skill proficiency. The future-readiness score is a project-defined decision-support indicator rather than a validated hiring measure.

## V. CONCLUSION

This paper presented a future-aware job market analytics and resume matching system that links monthly demand forecasting with skill-trend analysis and explainable resume skill coverage. A global Random Forest model uses lagged, rolling, trend, seasonal and categorical features to forecast posting counts over a twelve-month horizon. The forecast is incorporated into a role-specific skill profile, which is compared with skills extracted from a candidate's resume through a weighted coverage score. TF-IDF cosine similarity is retained as a separate text-alignment indicator.

On the supplied artificial dataset, the Random Forest achieved pooled MAE of 3.07, RMSE of 4.31 and $R^2$ of 0.908, outperforming the two reported baselines on pooled metrics. The weaker whole-market result and the 59% per-series baseline-win rate show why disaggregated evaluation is necessary. The application provides a practical and transparent prototype, but real-world validity remains untested. Future work should replace the artificial records with a documented real dataset, use rolling-origin backtesting across multiple periods, compare additional forecasting methods, incorporate semantic skill matching, estimate proficiency, and conduct user-centred evaluation.

## REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[2] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021. Available: https://otexts.com/fpp3/
[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.
[4] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press, 2008. Available: https://nlp.stanford.edu/IR-book/
[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[6] Project documentation, "Future-Aware Job Market Prediction and Resume Matching System (v2)," README.md and backend source code, supplied project archive, 2026. (Implementation and dataset-provenance documentation.)

 This work is licensed under a Creative Commons Attribution 4.0 International License 