ISSN 2616-5775 Vol. 9, Issue 9: 8-14, DOI: 10.25236/AJCIS.2026.090902

LightGBM-MLP Weighted Fusion Model for AI-Job Salary Prediction

Yuman Zou

Lanzhou No.1 Senior High School, Lanzhou, Gansu, 730000, China
2209306498@qq.com

Abstract: *Aiming at the challenges including long-tailed salary distribution, difficulties in categorical-feature processing, time-series data leakage, and performance bottlenecks of single models in AI-job salary prediction, this paper constructs a complete data preprocessing pipeline and proposes a LightGBM-MLP weighted fusion prediction model. In the preprocessing stage, IQR-based outlier elimination, logarithmic transformation for skewed-distribution correction, categorical-feature encoding, and time-series dataset splitting are adopted to avoid temporal leakage. The LightGBM gradient-boosting tree and regularized optimized MLP sub-models are trained separately. Weight coefficients of 0.6 and 0.4 are assigned according to their performance on structured datasets for model fusion. Multiple baseline comparison experiments are carried out on the AI Job Market recruitment dataset. Experimental results show that the proposed fusion model achieves the highest coefficient of determination $R^2=0.4165$ and the lowest root-mean-square error $RMSE=23994.12$. Time-series 5-fold cross-validation demonstrates low standard deviations for all metrics and good generalization stability, which can provide predictive references for AI-job salaries.*

Keywords: *salary prediction; LightGBM-MLP; multi-layer perceptron; model fusion; time-series regression*

1. Introduction

With the continuous expansion of the artificial intelligence industry, recruitment demands for AI-related positions have grown rapidly. AI-job salary prediction is a data-mining regression task that uses historical recruitment information such as job experience level, company scale, and job posting date as input features, and predicts numerical salary values through machine learning and deep learning algorithms.

AI-job salary prediction has multiple practical values. For job seekers, prediction results help them form reasonable salary expectations and provide quantitative references for salary negotiation. For enterprise HR and recruitment departments, it supports benchmarking market salary levels and designing competitive compensation packages. For researchers in human resource studies, it facilitates the analysis of temporal salary trends in the AI industry and provides evidence for macro talent market evaluation.

Existing technical solutions for job-salary prediction can be divided into four categories. The first category includes traditional statistical regression methods represented by multiple linear regression. Such models are simple and interpretable, but they are insufficient for capturing complex nonlinear interactions among features. The second category consists of ensemble tree-based machine learning algorithms, such as Random Forest, GBDT, LightGBM, and CatBoost. These methods are robust to structured discrete categorical features and achieve stable performance in salary regression tasks. The third category includes deep learning approaches represented by the Multi-Layer Perceptron (MLP), which is good at mining subtle nonlinear correlations hidden in data. The fourth category is multi-model fusion, which combines outputs from different models to complement their respective strengths and further improve prediction performance.

Although existing approaches have achieved promising results, several practical difficulties remain in AI-job salary prediction. First, salary data generally follows a right-skewed long-tailed distribution. Extreme high-salary and low-salary outliers interfere with model training and reduce prediction performance. Second, many recruitment attributes are categorical discrete features that require appropriate encoding before being fed into models. Third, recruitment data is time-series data. If the

ISSN 2616-5775 Vol. 9, Issue 9: 8-14, DOI: 10.25236/AJCIS.2026.090902

training and test sets are split randomly, temporal leakage will occur, which may overestimate the generalization ability of the model. Fourth, single models have inherent limitations. Tree-based models handle structured discrete features well but have limited capability for complex nonlinear mining. Neural networks tend to overfit under few-feature and small-sample conditions. Therefore, it is difficult for a single model to balance prediction accuracy and generalization stability at the same time.

To address the above problems, this paper proposes a LightGBM-MLP weighted fusion method for AI-job salary prediction. A complete preprocessing workflow, including outlier filtering, logarithmic transformation, categorical encoding, and time-series splitting, is designed to handle the long-tailed distribution and temporal leakage. The LightGBM and optimized MLP sub-models are trained separately, and model fusion is performed by weighted averaging to combine their respective advantages. Comparative baseline experiments on a real-world AI recruitment dataset and time-series 5-fold cross-validation verify that the proposed fusion model achieves satisfactory prediction accuracy and generalization stability.

2. Related Work

Job-salary prediction is a classic regression task in human resource data mining. In recent five years, scholars from different countries have conducted extensive research on traditional machine learning, deep learning, and multi-model fusion. This chapter reviews existing studies, analyzes their limitations, and introduces the research motivation of this paper.

2.1. Salary Prediction Based on Traditional Machine Learning

Traditional machine learning methods are common baselines for salary prediction tasks, mainly including statistical regression and ensemble tree models. Sofianopoulos et al. [1] compared seven regression algorithms, including Random Forest, GBDT, SVR, and KNN, for employee salary prediction. The experimental results show that ensemble tree methods outperform kernel-based methods and k-neighbor algorithms on mixed-type human resource datasets, but individual tree models have performance limitations. Ali et al. [2] applied XGBoost to salary prediction and verified its effectiveness in human resource regression. However, random dataset splitting was used without considering the temporal nature of recruitment data, which may introduce time-series leakage. Li and Zhang [3] constructed a hybrid model based on Random Forest and AdaBoost-KNN using crawled big-data recruitment data. The experimental results show that model fusion can improve prediction accuracy, but the stacking structure increases computational overhead. Novák and Kučera [4] studied global job salary forecasting using LightGBM and CatBoost. They pointed out that outlier removal and logarithmic transformation can significantly improve the stability of tree models in salary regression tasks.

2.2. Salary Prediction Based on Deep Learning

Deep learning uses strong nonlinear fitting ability to discover complex feature relationships, and some studies combine natural language processing to process job description texts. Shahin [5] used TF-IDF to vectorize job titles and applied MLP for salary prediction. This method performs well on datasets with job description texts, but neural networks are prone to overfitting when only a limited number of categorical features are available. Sun et al. [6] proposed an interpretable salary prediction algorithm based on collective utility marginal contribution learning to model the combined effects of skill sets. This algorithm has advantages under high-dimensional skill inputs, but its benefits are limited when the number of available features is small. Chen and Zhang [7] constructed a CNN-Attention-BiGRU hybrid deep network for salary prediction. The error decreases significantly when sufficient samples are available, but the generalization ability drops sharply on small datasets.

2.3. Salary Prediction Based on Ensemble and Fusion

Ensemble fusion reduces the inherent bias of single models by combining outputs, and it is an effective way to improve prediction performance. Mobin et al. [8] used the Super-Learner ensemble framework and fused multiple gradient boosting models for AI sector salary prediction. This method improves bias calibration, but it requires many sub-models and consumes high computational resources. Wang [9] selected the four best-performing models on the validation set and used hard voting ensemble for AI-job salary prediction to optimize the MAE metric. However, hard voting lacks a

ISSN 2616-5775 Vol. 9, Issue 9: 8-14, DOI: 10.25236/AJCIS.2026.090902

refined weight assignment strategy. Nidal and Ahmad [10] compared hard voting and weighted average fusion strategies on a recruitment salary dataset. The experimental results show that a properly designed weighted average can further improve R^2 compared with hard voting.

2.4. Review of Existing Research

To sum up, tree-based models are suitable for structured discrete feature processing, while deep learning models are good at capturing nonlinear relationships. However, existing studies still have some limitations. Some studies split time-ordered recruitment datasets randomly and ignore temporal leakage. Many fusion methods require a large number of sub-models and high computational resources. Lightweight weighted fusion methods for few-feature time-series AI-job scenarios are still insufficient. Therefore, this paper proposes a lightweight LightGBM-MLP weighted fusion prediction method. Strict time-series dataset splitting is performed to avoid temporal leakage, and salary prediction performance is improved under limited input features.

3. Methodology

This paper proposes a LightGBM-MLP weighted fusion prediction model. The overall workflow includes data preprocessing, LightGBM sub-model, optimized MLP sub-model, and weighted fusion output. Raw recruitment data are preprocessed and then split chronologically into a training set and a test set. The preprocessed data are fed into LightGBM and the optimized MLP for independent training. Both sub-models output salary predictions in the logarithmic space. The weighted sum is calculated to obtain the fusion output. Finally, inverse logarithmic transformation is applied to restore the real salary scale for metric evaluation.

3.1. Data Preprocessing

The raw salary dataset follows a right-skewed long-tailed distribution and contains many categorical features and time-series fields. Preprocessing includes four steps.

1) IQR outlier elimination. The 25th percentile Q_1 and the 75th percentile Q_3 of salary are calculated, and the interquartile range $IQR=Q_3-Q_1$. Samples with salaries in the interval $[Q_1-1.5\times IQR, Q_3+1.5\times IQR]$ are retained. Extreme high-salary and low-salary outliers are removed to reduce their interference with model training.

2) Logarithmic transformation of the target variable. A logarithmic transformation is applied to the salary variable to convert the right-skewed distribution into an approximately normal distribution and reduce the regression bias caused by the long-tailed effect.

$$y' = \ln(y + 1) \tag{1}$$

where y is the original salary, and y' is the transformed salary. In the prediction stage, the inverse transformation $y = exp(y') - 1$ is used to restore the real salary value.

3) Ordinal encoding of categorical features. Two categorical features, experience level and company size, are encoded into ordinal numerical values.

Experience level mapping: Entry → 1, Mid → 2, Senior → 3, Executive → 4.

Company size mapping: Startup → 1, SMB → 2, Enterprise → 3.

Missing values are uniformly filled with 0.

4) Time-series dataset splitting. The dataset is split according to the job posting year using the 0.8 quantile. Samples posted earlier are used as the training set, and samples posted later are used as the test set. The dataset is not shuffled to avoid time-series leakage. Finally, three input features are used: encoded experience level, encoded company size and job-posting year.

3.2. LightGBM Sub-model

LightGBM is a histogram-based gradient boosting tree algorithm and has good adaptability to regression tasks on structured discrete data. The hyperparameters are set as follows: maximum tree

ISSN 2616-5775 Vol. 9, Issue 9: 8-14, DOI: 10.25236/AJCIS.2026.090902

depth is 4, learning rate is 0.08, L2 regularization coefficient is 0.2, and number of iterations is 120. A fixed random seed is used to ensure experimental reproducibility. This sub-model is good at mining piecewise regularity in discrete features and has strong stability on structured recruitment data, so it is assigned a higher weight in fusion.

3.3. Optimized MLP Sub-model

The Multi-Layer Perceptron (MLP) is used to mine subtle nonlinear relationships in the data. The hidden layers are set to 64, 32, and 16, with ReLU activation functions. The initial learning rate is 0.001, L2 regularization coefficient is 0.001, maximum number of iterations is 600, and the random seed is fixed. L2 regularization is introduced to reduce the overfitting risk of neural networks under few-feature conditions and to supplement nonlinear patterns that are difficult for tree models to capture.

3.4. LightGBM-MLP Weighted Fusion

Both sub-models output predictions in the logarithmic space, and model fusion is implemented by weighted averaging.

$$\hat{y}_{fusion}=0.6\times\hat{y}_{lgb}+0.4\times\hat{y}_{mlp} \tag{2}$$

where $\hat{y}_{fusion}$ is the fused output of LightGBM-MLP, $\hat{y}_{lgb}$ is the prediction of the LightGBM sub-model, and $\hat{y}_{mlp}$ is the prediction of the MLP sub-model.

Weight assignment is based on model performance. LightGBM has stronger stability on structured recruitment data and is responsible for the main prediction task, so its weight is set to 0.6. MLP is used to supplement the mining of subtle nonlinear relationships, so its weight is set to 0.4. The fusion output is then restored to the original salary scale through inverse logarithmic transformation for metric calculation. This fusion strategy combines the advantages of tree models for structured feature processing and the nonlinear fitting ability of neural networks, reduces the inherent bias of single models, and improves overall prediction performance.

4. Experiments and Results

4.1. Experimental Settings

4.1.1. Experimental Dataset

The AI Job Market recruitment dataset is used in the experiments. It contains 20 fields, including job ID, job title, company size, industry, country, remote type, experience level, working years, education, skill tags, salary, posting month, posting year, recruitment urgency, and number of openings. The prediction target is job salary. After IQR outlier filtering, clean samples are obtained. The dataset is split chronologically into a training set and a test set. The model uses three input features: encoded experience level, encoded company size, and job-posting year.

4.1.2. Baseline Methods

To verify the effectiveness of the proposed LightGBM-MLP fusion model, five baseline models are selected for comparison: GBDT baseline, Random Forest, optimized MLP, single LightGBM, and single CatBoost.

4.1.3. Evaluation Metrics

Four regression evaluation metrics are used to measure model performance: R^2, RMSE, MAE, and MAPE.

4.1.4. Validation Strategy

Time-series 5-fold cross-validation is used to evaluate model stability instead of ordinary random K-fold cross-validation. Recruitment data is time-series data, and random shuffling will cause temporal leakage. Time-series 5-fold cross-validation strictly splits the dataset according to time, ensuring that the training set is always earlier than the validation set. The mean and standard deviation of each metric are reported to measure the generalization ability of the model across different time windows.

ISSN 2616-5775 Vol. 9, Issue 9: 8-14, DOI: 10.25236/AJCIS.2026.090902

4.2. Experimental Results

4.2.1. Performance Comparison of Different Methods

Table 1. Performance comparison of different models on the test set

Model	R^2	RMSE	MAE	MAPE (%)
GBDT Baseline	0.4145	24035.31	19635.45	18.18
Random Forest	0.4140	24045.40	19668.24	18.24
Optimized MLP	0.4113	24100.23	19896.34	18.84
LightGBM	0.4128	24069.87	19694.13	18.26
CatBoost	0.4153	24019.15	19662.27	18.26
LightGBM-MLP Fusion (Ours)	0.4165	23994.12	19713.98	18.42

Table 1 shows the experimental results. Several observations can be made as follows.

First, the proposed LightGBM-MLP weighted fusion model achieves the highest R^2=0.4165 and the lowest RMSE=23994.12 among all compared models. This result indicates that the weighted fusion strategy effectively combines the advantages of tree models and neural networks and improves overall fitting performance.

Second, CatBoost achieves the best comprehensive performance among the single models, with R^2=0.4153. The optimized MLP performs worse than the tree models on the few-feature structured dataset, which shows that neural networks are prone to overfitting under limited features.

Third, the MAPE values of all models are distributed between 18.18% and 18.84%. The fusion model has a slightly higher MAPE than some baseline models, but the gap is small. Since R^2 and RMSE are global fitting metrics and have higher priority, the overall performance of the proposed fusion model is better.

Fourth, the overall prediction error is within a reasonable range. This study uses only three input features, and the prediction accuracy can be further improved by introducing additional features such as industry, region, and skills.

The comparison results of R^2 and MAPE are shown in Figure 1 and Figure 2, respectively.

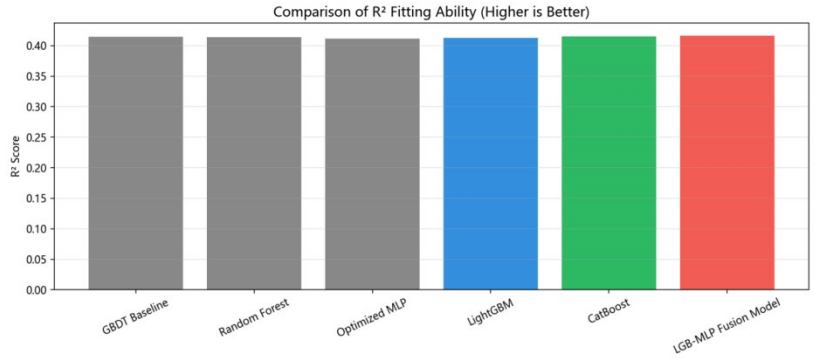


Figure 1. Comparison of R^2 fitting capacity for each model

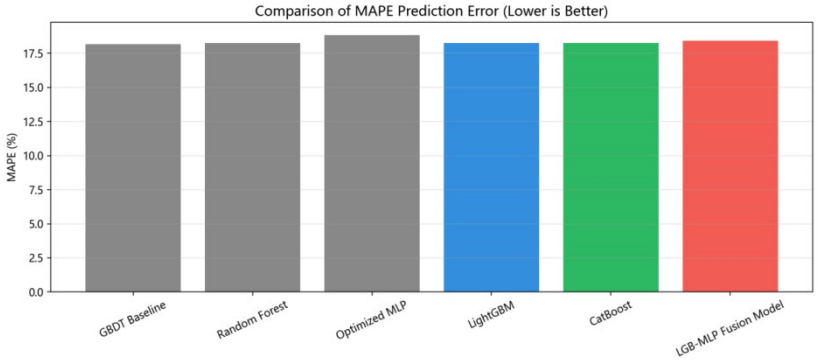


Figure 2. Comparison of MAPE prediction error for each model

ISSN 2616-5775 Vol. 9, Issue 9: 8-14, DOI: 10.25236/AJCIS.2026.090902

It can be seen from Figure 1 that the proposed fusion model has the highest R^2 value among all models. Figure 2 shows that the MAPE of the proposed model is at a medium level, which is consistent with the analysis in Table 1.

4.2.2. 5-Fold Cross-Validation Results

Table 2. 5-fold cross-validation results of the proposed model

Metric	Mean Value	Standard Deviation
R^2	0.4221	0.0159
RMSE	23922.39	262.34
MAE	19667.58	308.86
MAPE (%)	18.26	0.26

Time-series 5-fold cross-validation results are shown in Table 2. The average R^2 is 0.4221, which is slightly higher than the test-set result. The standard deviations of all metrics are relatively low. In particular, the standard deviation of MAPE is only 0.26%, which indicates that the model has small performance fluctuations across different time windows and good generalization stability. The experimental results are not accidental due to a specific dataset split.

4.3. Comprehensive Analysis of Experimental Results

The experimental results on both the test set and the cross-validation set show that the proposed lightweight weighted fusion model can achieve better performance than individual baseline models when using only three input features. The MLP performs worse than the tree models, which reflects that neural networks are prone to overfitting on few-feature time-series structured datasets. Although the MAPE of the fusion model increases slightly, the global metrics R^2 and RMSE are improved. The limitation of this study is that the number of input features is relatively small. In future work, additional information such as job skills, region, and industry can be introduced to further improve prediction performance.

5. Conclusions

This paper focuses on practical problems in AI-job salary prediction, including long-tailed salary distribution, categorical feature processing, time-series leakage, and performance bottlenecks of single models. A complete data preprocessing workflow is designed, and a LightGBM-MLP weighted fusion salary prediction method is proposed. The LightGBM gradient boosting tree and the optimized MLP sub-model are trained separately, and their outputs are fused with weights of 0.6 and 0.4. Comparative experiments are conducted on a real-world AI recruitment dataset using GBDT, Random Forest, MLP, LightGBM, and CatBoost as baselines. The experimental results show that the proposed fusion model achieves the best R^2 and RMSE. Time-series 5-fold cross-validation also verifies its good generalization stability. Due to the limitation of few input features in this study, future work will introduce more features, such as job skills and regional information, to improve prediction accuracy. In addition, interpretability modules can be added to quantitatively analyze the contribution of each feature to salary.

References

[1] Sofianopoulos, S., & Mavridis, P., "Comparative analysis of machine learning models for employee salary prediction," Proceedings of the 12th International Conference on Information Systems and Technologies, Athens, Greece, 2021.

[2] Ali, M., & Gupta, S., "Machine learning approaches to salary prediction in human resource payroll systems," Journal of Computing and Social Science, 5(2):45-54, 2022.

[3] Li, Y., & Zhang, H., "A model for predicting salaries in big data roles: An integration of Random Forest and AdaBoost-KNN models," ACM Conference on Big Data and Artificial Intelligence, New York, USA, 189-194, 2023.

[4] Novák, J., & Kučera, T., "Explainable AI and robust forecasting of global salary trends: Addressing data drift and unseen categories by tree-based models," Mathematical Modeling and Computing, 12(3):993-1004, 2025.

[5] Shahin, N., "Job market cheat codes: Prototyping salary prediction and job grouping with synthetic job listings," arXiv preprint arXiv:2506.15879, 2025.

ISSN 2616-5775 Vol. 9, Issue 9: 8-14, DOI: 10.25236/AJCIS.2026.090902

[6] Sun, Y., Zhuang, F., Zhu, H., et al., "An interpretable salary prediction algorithm based on collective utility marginal contribution learning," Journal of Computer Research and Development, 61(5):1276-1289, 2024.
[7] Chen, L., & Zhang, Q., "Salary management data analysis and annual salary prediction based on artificial intelligence algorithms," 2024 7th International Conference on Data-Driven Intelligent Systems, New York, USA, 218-223, 2024.
[8] Mobin, M., Zhang, T. Y., & Hu, B., "Interpretable salary prediction in the AI sector using ensemble learning with uncertainty quantification," International Conference on Data Science and Human Resources, Piscataway, USA: IEEE, 2024.
[9] Wang, H., "Predicting AI position salaries: Model fusion via weighted averaging and feature correlation analysis," ACM International Conference on Artificial Intelligence and Employment, New York, USA, 211-216, 2024.
[10] Nidal, S., & Ahmad, M., "Ensemble regression strategies for job-posting salary prediction," 2025 International Conference on Applied AI and Data Analytics, Amman, Jordan: IEEE, 2025.