Enhanced Salary Prediction System Using Random Forest and Categorical Encoding

Authors: **Faiz Azam (BAI-22S-010), Abdullah Khalid (BAI-22S-003)**

Submitted to: **Syed Mohammad Hassan**

Department of Artificial Intelligence

Sindh Madressatul Islam University (SMIU)

Abstract: Using important employment characteristics such job title, experience level, employment type, geography, and company size, this study offers an improved machine learning-based wage prediction method. Although salary prediction has been researched before, this project adds a useful prediction module, strong preprocessing, and efficient handling of outliers. To increase the accuracy of our USD salary predictions, we used a Random Forest Regressor. Additionally, outlier filtering was employed to cut down on noise and label encoding was utilized for categorical variables. Our contribution is to improve the prediction pipeline and make it more replicable and easy to utilize for deployment in the future.

Index Terms: Machine Learning, Salary Prediction, Regression, Random Forest, Label Encoding, Feature Engineering

I. Introduction Predicting salaries has grown in importance for employers, job boards, and job searchers. For this, conventional statistical techniques have been employed, but new advances in machine learning provide greater flexibility and performance. In order to improve prediction accuracy, this work expands on previous methods by adding improvements to the data preprocessing and modeling phases.

II. Literature Review Prior research has used simple decision tree models and linear regression to forecast compensation based on a few job characteristics. Nevertheless, these models have drawbacks include low generalization, sensitivity to outliers, and an inability to detect non-linear patterns. Our study expands on this framework by utilizing Random Forest, an ensemble learning method that is more scalable and resilient. Furthermore, categorical feature transformation was optimized for memory efficiency using label encoding rather than conventional one-hot encoding.

III. Dataset Overview The dataset used (salaries.csv) contains real-world job postings with columns such as:

- experience level
- employment type
- job title
- salary in USD
- employee residence
- company size

After cleaning, the dataset consisted of approximately N rows (excluding rows with missing values and extreme salary outliers).

IV. Methodology

A. Data Preprocessing

- Null values were eliminated.

  Salary outliers (more than $400,000) were eliminated.

  All categorical variables—job_title, experience_level, company_size, employee_residence, and employment_type—were subjected to label encoding.

B. Model Selection

- Random Forest Regressor was selected for its capability to handle non-linearity and large datasets.
- Data was split using an 80:20 ratio for training and testing.

C. Model Evaluation

- R^2 Score and RMSE were used as evaluation metrics.
- The model achieved an R^2 score of ~0.75 and RMSE around ~$21,000, indicating good predictive power.

V. Key Contributions & Advancements

To increase model accuracy, sophisticated outlier removal was applied to the wage column.

reduced dimensionality by using label encoding rather than one-hot.

created a unique prediction cell that accepts user-specified input for estimating salaries.

performed better than linear models in comparable previous initiatives.

VI. Results and Visualizations

Job titles with the highest average wages were displayed in bar plots.

Predicted and actual salaries were compared using scatter plots.

Job_title_enc and experience_level_enc were the most significant features, according to feature importances.

VI. Results and Visualizations

The highest average salary job titles were displayed in bar plots.

Scatter plots were used to compare expected and actual salaries.

The most significant features, according to feature importances, were job_title_enc and experience_level_enc.

VII. Limitations and Future Work

Predictions may be skewed by bias brought on by the regional concentration of jobs (such as in the US).

Non-tree-based models can be used to test one-hot encoding and feature scaling.

With Streamlit, a graphical user interface (GUI) or web application can be created for real-world use.

VIII. Conclusion This study shows how to use Random Forest to estimate salaries in a scalable and useful way. The system attains competitive accuracy with the application of robust regression algorithms and the improvement of preprocessing processes. Future research can incorporate real-time data from APIs like Glassdoor or LinkedIn, or expand this model into a full-stack web application.

References

[1] Scikit-learn documentation, https://scikit-learn.org

[2] Kaggle Datasets, https://www.kaggle.com

[3] J. Brownlee, "Machine Learning Mastery with Python", 2019.