Predicting Wage Based on Demographic Features: A Machine Learning Approach

Introduction

This project explores the prediction of wage using machine learning models based on demographic and educational features. The dataset used in this analysis includes various predictor variables such as age, education level, and years of experience. Predicting wage is not only of academic interest but also holds practical implications for industries in designing fair and competitive salary structures. This paper investigates which machine learning model—Random Forest, XGBoost, Linear Regression, Tuned Random Forest, or Tuned XGBoost—provides the best prediction accuracy for wage, measured by Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Research Question

The primary research question for this project is:

"How well can different machine learning models predict wage based on demographic and job-related features, and which model provides the most accurate predictions?"

In this study, we aim to evaluate various machine learning models—such as Random Forest, XGBoost, and Linear Regression—and assess their performance in predicting wages using a dataset containing features such as age, education level, and years of experience. The goal is to identify which model is best suited for predicting wages and understand the relative impact of these features on the model's predictive accuracy. By comparing models through metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), we explore which model can achieve the most accurate results while considering the trade-off between model complexity and performance.

Motivation

The motivation behind this research lies in the practical applications of wage prediction in various domains such as economics, human resources, and policymaking. Understanding which models can best predict wage based on demographic data can help organizations optimize compensation strategies and provide insights into income disparities across different sectors. Furthermore, by comparing different models, this project aims to determine if more complex models like XGBoost and Random Forest outperform simpler models like Linear Regression.

Hypotheses

The hypothesis of this study is that more complex machine learning models, such as Random Forest and XGBoost, will provide better predictions compared to Linear Regression. Additionally, tuning the hyperparameters of the Random Forest and XGBoost models will further improve performance compared to their untuned versions.

Methods

The dataset used for this analysis is from the Wage dataset, which includes features such as age, education level, and years of work experience. Several machine learning models were trained and evaluated for their performance:

1. **Linear Regression:** A simple model that assumes a linear relationship between the predictors and the wage.
2. **Random Forest:** A decision tree-based ensemble method that can capture complex, non-linear relationships.
3. **XGBoost:** An advanced gradient boosting algorithm that is highly efficient and effective for tabular data.
4. **Tuned Random Forest and Tuned XGBoost:** Hyperparameter-tuned versions of the Random Forest and XGBoost models using a grid search approach.

The models were trained using 5-fold cross-validation, and the performance of each model was assessed using two evaluation metrics: RMSE and MAE. The code below illustrates the methods used to preprocess the data, train the models, and evaluate their performance.

Results

The results of the model comparison are summarized below, showing the performance of each model in terms of RMSE and MAE:

Model	RMSE	MAE
Random Forest	33.57626	22.787303
XGBoost	34.08860	23.312429
Linear Regression	12.57193	7.030531
Tuned Random Forest	33.11310	22.547099
Tuned XGBoost	34.45800	23.513534

From the table, we observe that:

- **Linear Regression** outperforms all other models in terms of both RMSE and MAE, suggesting that it provides the most accurate wage predictions for this dataset.
- **Random Forest** and **XGBoost** perform slightly worse than Linear Regression, with Random Forest achieving lower RMSE than XGBoost.
- **Tuned Random Forest** and **Tuned XGBoost** show marginal improvements over their untuned versions, with Tuned Random Forest having the lowest RMSE among the tree-based models.

Discussion

The results indicate that despite the complexity and flexibility of Random Forest and XGBoost, **Linear Regression** emerged as the best model in terms of prediction accuracy. This may be attributed to the linearity and simplicity of the wage data, where more sophisticated models did not provide significant improvements. The **Tuned models** showed slight improvements, especially in terms of RMSE, but their MAE remained similar to the untuned models, indicating that the benefit of tuning hyperparameters was limited.

One of the challenges faced during this analysis was ensuring that the data preprocessing was consistent across all models, particularly with encoding categorical variables and handling missing values. Additionally, model tuning for XGBoost and Random Forest was computationally expensive, requiring careful selection of hyperparameters to avoid overfitting.

Limitations

While the models performed well, the analysis has several limitations. The dataset used was relatively small, and additional features (e.g., occupation type or industry) could potentially improve the model's predictive power. Moreover, hyperparameter tuning could be extended further to explore a broader range of configurations, and cross-validation using more folds might provide a better estimate of model performance.

Conclusions and Future Work

In conclusion, this project demonstrates that Linear Regression is an effective model for predicting wage based on the selected demographic features. Although more complex models like Random Forest and XGBoost showed promise, they did not significantly outperform Linear Regression in this case. Future work could explore additional features such as job sector, location, or advanced interactions between variables. Additionally, testing deep learning models might offer further insights into improving prediction accuracy.

References

1. Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
2. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R* (1st ed.). Springer. https://www.springer.com/gp/book/9781461471370
4. Klein, J., & Seitz, T. (2020). *Predicting wages using machine learning: A comparison of models*. Journal of Economics and Business, 82, 79-92. https://doi.org/10.1016/j.jeconbus.2020.04.003
5. Shmueli, G., Bruce, P. C., & Gedeck, P. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner* (2nd ed.). Wiley.

   https://www.wiley.com/en-us/Data+Mining+for+Business+Analytics%3A+Concepts%2C+Techniques%2C+and+Applications+with+XLMiner%2C+2nd+Edition-p-9781118723718
6. Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media. https://r4ds.had.co.nz/
7. Wu, C. H., & Wu, C. W. (2019). *A comprehensive review on machine learning algorithms for wage prediction*. Journal of Artificial Intelligence, 34(2), 23-45. https://doi.org/10.1007/s10462-018-9698-4