Comparative Analysis of XGBoost and Random Forest in Predicting the Success of Trainees in the Technical Intern Training Program (TITP)

Syaban Maulana, Nenden Siti Fatonah, Gerry Firmansyah and Agung Mulyo Widodo

March 1, 2025

Comparative Analysis of XGBoost and Random Forest in Predicting the Success of Trainees in the Technical Intern Training Program (TITP)

Syaban Maulana[1,] Nenden Siti Fatonah[2*,] Gerry Firmansyah[3,] , [4] and Agung Mulyo Widodo[*]
[1,2,3,4]University Esa Unggul
syaban@student.esaunggul.ac.id, nenden.siti@esaunggul.ac.id, gerry@esaunggul.ac.id, agung.mulyo@esaunggul.ac.id

Abstract

Japan's population is projected to decline from 125 million (2020) to 88 million (2065), prompting the government to open opportunities for foreign workers through the Technical Intern Training Program (TITP). This research aims to analyze success factors of TITP interns using XGBoost and Random Forest models with SMOTE method to handle imbalanced data. Data was sourced from labor sending companies comprising 784 samples with the following distribution: 57 cases of pre-training dropouts, 67 cases of training dropouts, 52 cases of internship dropouts, 16 cases of runaways, and 592 cases of successful internship completion. Results show Random Forest slightly outperforming XGBoost with balanced accuracy of 0.32 compared to 0.27, though both achieved macro F1-scores of approximately 0.28-0.32. Feature importance analysis revealed age, test scores, and health factors as key predictors of internship success. The main challenge was extreme class imbalance with minority classes such as runaways (only 16 samples or 2% of the total). While the models performed well for the majority class, improvements are needed for minority class detection.

Keywords: Technical Intern Training Program, XGBoost, Random Forest, SMOTE, multiclass classification, Imbalanced Data.

[*] Created the first draft of this document

1 Introduction

Japan faces a significant demographic challenge, with projections indicating a population decline from 125 million in 2020 to 88 million by 2065 and a substantial reduction in working-age individuals. To address this labor shortage, Japan implemented the Technical Intern Training Program (TITP), which has faced challenges including a high rate of intern runaways.

Machine learning techniques offer potential solutions for improving intern selection processes. This research compares XGBoost and Random Forest algorithms to predict TITP intern success using a dataset of 784 samples with five outcome categories ranging from pre-training dropouts to successful completion. Our models address the significant class imbalance using the SMOTE method.

Results show Random Forest slightly outperforming XGBoost (balanced accuracy of 0.32 vs. 0.27), with both models achieving macro F1-scores of 0.28-0.32. Feature importance analysis identified internship duration, demographics, and health indicators as key success predictors. These findings provide an empirical foundation for enhancing intern selection and monitoring processes, potentially reducing program failures and their associated costs.

2 Related Works

In the study titled "Application of Random Forest Algorithm on Credit Risk Analysis," Kurniawan et al. (2024) explore the challenges of credit risk, which arises from the inability of borrowers to repay loans, potentially leading to significant losses for lenders. The authors emphasize the importance of conducting objective and accurate analyses of borrowers using data collected during the credit application process, which includes personal information, monthly income, and employment history. By employing the Random Forest algorithm, the research identifies key features that significantly influence loan approval decisions. The study reveals that important factors include repeat user status, mobile likelihood, monthly salary, industry, and years of service. The model demonstrates a commendable recall score of 0.9091, indicating its effectiveness in credit risk analysis. This research contributes to the growing body of literature on machine learning applications in financial decision-making, highlighting the potential of Random Forest in enhancing the accuracy of credit assessments (Kurniawan et al., 2024).

In the paper titled "Credit Scoring: Does XGBoost Outperform Logistic Regression? A Test on Italian SMEs," Zedda (2024) investigates the effectiveness of XGBoost compared to traditional logistic regression in predicting loan defaults among Italian small and medium-sized enterprises (SMEs). The study analyzes a dataset of 35,535 cases across seven business sectors, utilizing 28 banking variables and 55 balance sheet ratios. The findings indicate that while both models demonstrate similar capabilities in selecting good borrowers, the performance varies significantly across different sectors. Notably, the choice of cutoff settings plays a crucial role in the models' effectiveness, with XGBoost showing slightly better results in certain scenarios. The research highlights the importance of balancing the risks of rejecting creditworthy applicants against the costs of lending to potentially defaulting borrowers, ultimately suggesting that machine learning methods like XGBoost can enhance credit scoring processes in banking (Zedda, 2024).

3 Research Methods

3.1 Dataset and Preprocessing

The dataset utilized in this research was sourced from the Technical Intern Training Program (TITP) participants at Sending Organization or 送り出し機関, comprising a total of 784 records collected over the period from 2019 to 2023. This dataset includes 37 variables that encompass various aspects of the interns' information, ranging from demographic data to their readiness for the program. Key demographic variables include birthplace, date of birth, gender, marital status, nationality, religion, and address. Additionally, physical condition variables such as height, weight, and blood type are included, along with health-related information. The preprocessing phase involved several critical steps to ensure data quality. Missing values in numerical data were filled with the mean, while categorical data were filled with the mode. Label encoding was applied to convert categorical variables into numerical format, facilitating their use in machine learning models. The dataset was then split into training and testing sets, with 80% of the data allocated for training and 20% for testing, ensuring a robust evaluation of the models. Furthermore, feature scaling was performed using StandardScaler to normalize the range of numerical features, and the Synthetic Minority Over-sampling Technique (SMOTE) was employed to address class imbalance in the dataset, thereby enhancing the model's predictive performance.

The target variable "current_status" was classified into five categories:

1. Class 0: Pre-training dropout (57 samples)
2. Class 1: Training dropout (67 samples)
3. Class 2: Internship dropout (52 samples)
4. Class 3: Runaway cases (16 samples)
5. Class 4: Successfully completed (592 samples)

3.2 Machine Learning Model: XGBoost

XGBoost (Extreme Gradient Boosting) was implemented as one of the primary models in this research due to its proven effectiveness in handling imbalanced classification tasks. The algorithm employs a gradient boosting framework that builds an ensemble of weak prediction models, typically decision trees, in a sequential manner. In our implementation, XGBoost was configured with specific hyperparameters to address the five-class classification problem. The model architecture utilized a multi:softprob objective function with a learning rate of 0.1 and a maximum tree depth of 4. Additional parameters included n_estimators set to 100, subsample and colsample_bytree both at 0.8, and min_child_weight of 3 to control model complexity. The implementation leveraged XGBoost's built-in regularization capabilities (L1 and L2) and automatic handling of missing values. The training process included cross-validation to validate hyperparameter selection before final model training on the SMOTE-processed data.

3.3 Machine Learning Model: Random Forest

Random Forest was employed as the second primary model in this research, offering a different ensemble learning approach to predict TITP intern outcomes. This algorithm constructs multiple decision trees during training and outputs the class that is the mode of the individual trees' predictions. In our implementation, the Random Forest model was configured with 100 estimators and a maximum depth of 4 to balance complexity and generalization ability. The model utilized bootstrap aggregating (bagging), training each tree on a random subset of the data with replacement. For the five-class classification task, we configured the algorithm with balanced class weights to account for the class imbalance present in the original data. The number of features considered for splitting at each node was set to the square root of the total number of features, following standard practice. The minimum samples required to split an internal node was set to 2, and the criterion used for measuring split quality was Gini impurity. Like XGBoost, the Random Forest model was trained on the SMOTE-processed data to address class imbalance issues.

3.4 Optimization

The optimization phase involved a three-stage approach to enhance model performance on the highly imbalanced TITP dataset. First, I implemented resampling techniques, with SMOTE (Synthetic Minority Over-sampling Technique) selected as the optimal method after comparative evaluation. SMOTE generated synthetic samples for minority classes by interpolating between existing instances, effectively balancing the training data distribution while preserving feature relationships. The second stage focused on model-level optimization, where I maintained the default configurations of both XGBoost and Random Forest models but trained them on the SMOTE-processed data. This approach allowed me to assess the impact of data rebalancing on model performance while keeping algorithmic parameters consistent. Finally, I conducted evaluation metric optimization, maintaining default probability thresholds (0.5) for class prediction despite the initial consideration of threshold adjustments. This decision was based on preliminary tests showing minimal performance improvements with threshold modifications. Throughout the optimization process, I prioritized balanced accuracy and macro F1-score as the primary evaluation metrics to ensure fair assessment across all classes, particularly the minority classes representing various dropout scenarios.

4 Results and Discussion

The implementation of XGBoost and Random Forest models on the TITP dataset yielded several significant findings. Random Forest demonstrated slightly superior performance with a balanced accuracy of 0.32 compared to XGBoost's 0.27, while both achieved comparable macro F1-scores (0.32 and 0.28 respectively). Both models excelled at predicting successful program completions (Class 4) with precision above 0.93, but struggled with minority classes, particularly runaways (Class 3) where precision and recall remained below 0.30.

Feature importance analysis revealed that internship duration was the dominant predictor (importance score 0.295), followed by demographic factors including ID, BMI, and place of birth. This suggests that administrative and physical attributes significantly influence program outcomes. The stark importance difference between the top feature and others indicates that commitment duration serves as a critical success indicator in TITP internships.

The class imbalance in the dataset (592 successful completions versus only 16 runaways) presented a substantial challenge. While SMOTE improved overall performance, models still exhibited classification bias toward the majority class. This demonstrates the inherent difficulty in predicting rare events like internship abandonment, despite resampling efforts.

Model performance varied notably across different target classes. For pre-training dropouts (Class 0), both models achieved reasonable precision (0.22) but limited recall. For training dropouts (Class 1), performance was similar with precision and recall around 0.15. The models performed better for internship dropouts (Class 2), with Random Forest achieving 0.29 F1-score compared to XGBoost's 0.09.

These findings highlight the potential of machine learning approaches in predicting TITP outcomes while acknowledging their limitations with extremely imbalanced classes. The models provide valuable insights for sending organizations to identify at-risk candidates and implement targeted interventions. Future research should explore more sophisticated techniques for handling extreme class imbalance and incorporate temporal features to further improve predictive performance.

5 Conclusion

This study demonstrates the effectiveness of machine learning approaches in predicting Technical Intern Training Program outcomes, with Random Forest slightly outperforming XGBoost in handling the inherently imbalanced dataset. The analysis identified internship duration, demographic factors, and health indicators as the most significant predictors of success, providing an empirical foundation for improved selection processes. Practically, these findings enable sending organizations to implement more robust screening protocols focused on the identified key factors, potentially reducing program failures and their associated economic and social costs. The predictive models developed can serve as decision support tools during selection, while the feature importance analysis offers guidance for designing targeted pre-departure orientation programs addressing the most critical risk factors. For future research, I recommend exploring more sophisticated techniques specifically designed for extreme class imbalance, investigating the temporal aspects of program dropouts, and expanding the dataset to include post-arrival variables such as workplace environment and supervisor relationships. Additionally, implementing these models in a production environment with continuous feedback loops would further validate their real-world effectiveness and facilitate ongoing refinement of the predictive algorithms.

References

Kurniawana, R. (2024). Application of Random Forest algorithm on credit risk analysis. 9th International Conference on Computer Science and Computational Intelligence.

Meng, D., Xu, J., & Zhao, J. (2021, December). Analysis and prediction of hand, foot and mouth disease incidence in China using Random Forest and XGBoost. PLoS ONE.

Mika E., Aisyah Savira, N., & Febi Triyanti. (2022). Fenomena childfree di Jepang dalam perspektif teori feminisme eksistensialis.

Murnawan, S., Lestari, S., Samihardjo, R., & Dewi, D. A. (2024, September). Sustainable educational data mining studies: Identifying key factors and techniques for predicting student academic performance. Journal of Applied Data Sciences.

National Institute of Population and Social Security Research. (2019). Population and social security in Japan.

Ng, M., & Hernandez, M. (2024, November). Predicting solar curtailment using XGBoost and Random Forests. NHSJS Reports.

Nguyen, H. Q., Nguyen, D. D. K., Le, T. D., Mai, A., & Huynh, K. T. (2023, September). Career path prediction using XGBoost model and students' academic results. Journal of Innovation and Sustainable Development.

Nie, M., Xiong, Z., Zhong, R., Deng, W., & Yang, G. (2020, April). Career choice prediction based on campus big data—Mining the potential behavior of college students. Applied Sciences.

Putri Elsy. (2018). Fenomena tenaga kerja asing di Jepang dewasa ini.

Wahyuningsih, T., Iriani, A., Purnomo, H. D., & Sembiring, I. (2024, March). Predicting students' success level in an examination using advanced linear regression and extreme gradient boosting. Computer Science and Information Technologies.

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023, September). Explainable prediction of loan default based on machine learning models. Data Science and Management.

Hakkal, S., & Ait Lahcen, A. (2024, December). XGBoost to enhance learner performance prediction. Computers and Education: Artificial Intelligence.

Hossen, M. K., & Uddin, M. S. (2023, August). Attention monitoring of students during online classes using XGBoost classifier. Computers and Education: Artificial Intelligence.

Alzakari, S. A., Abdel Menaem, A., Omer, N., Abozeid, A., Hussein, L. F., Mohamed Abass, I., Rami, A., & Elhadad, A. (2024, October). Enhanced heart disease prediction in remote healthcare monitoring using IoT-enabled cloud-based XGBoost and Bi-LSTM. Alexandria Engineering Journal.

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. (2023, September). Explainable prediction of loan default based on machine learning models. Data Science and Management.

Kozina, A., Kuźmiński, Ł., Nadolny, M., Miałkowska, K., Tutak, P., Janus, J., Płotnicki, F., Walaszczyk, E., Rot, A., Dziembek, D., & Król, R. (2023, October). The default of leasing contracts prediction using machine learning. KES.

Nie, M., Xiong, Z., Zhong, R., Deng, W., & Yang, G. (2020, April). Career choice prediction based on campus big data—Mining the potential behavior of college students. Applied Sciences.

Brooks, W. (2024). The dynamics of demand: The Japanese TITP and SSW programs in an era of change.

National Institute of Population and Social Security Research. (2019). Population and social security in Japan.