International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I10P113

Volume 72 Issue 10, 130-139, October 2024

Original Article

Factors Influencing Voluntary Turnover Among Young College Graduates Using the XGBoost with Bagging Aggregation Algorithm: Findings from Nationwide Survey in South Korea

Haewon Byeon

AI Convergence College, Inje University, South Korea.
Inje University Medical Big Data Research Center.

Corresponding Author : bhwpuma@naver.com

Received: 11 June 2024 Revised: 05 October 2024 Accepted: 08 October 2024 Published: 25 October 2024

Abstract - *This study aims to analyze the factors influencing turnover among young professionals aged 20 to 30 and compare the predictive performance of various machine learning models using data from the 2019 Graduates Occupational Mobility Survey (n=11,605). The XGBoost with Bagging model was selected for its ability to handle complex interactions and large datasets. The dataset includes demographic information, job characteristics, job satisfaction scores, and other relevant variables. Data preprocessing involved handling missing values, one-hot encoding for categorical variables, and normalization. The model's performance was compared to KNN, Logistic Regression, SVM, and Bagging using metrics such as Area Under the Curve (AUC) and F1-score. The XGBoost with Bagging model demonstrated superior performance, with an AUC of 0.85 and an F1-score of 0.86, outperforming the other models. Key features influencing turnover intentions included permanent employment status, salary, job satisfaction, job security, and career advancement. These findings provide actionable insights for human resource management strategies aimed at reducing employee turnover. The study concludes that the XGBoost with Bagging model is a robust tool for predicting turnover intentions and recommends future research to integrate additional features and apply the model to different age groups and industries for further validation.*

Keywords - *Turnover intentions, XGBoost with Bagging, Machine learning models, Job satisfaction, Human resource management strategies.*

I. Introduction

The continuous development of technology has produced a dynamic South Korean labor market, marked by significant transformations in recent years. Despite the growth and evolution in various sectors, there remains a critical gap in understanding the turnover behavior of recent college graduates a demography increasingly essential to the workforce. Key elements of this evolution include the rise of knowledge workers, the expansion of competency-based human resource management [1], increased labor flexibility, the introduction of performance-based evaluation systems, and the balancing of work and family life [2]. These changes present both challenges and opportunities within the labor market. However, existing studies have largely focused on turnover pathways and intentions, often overlooking complex, non-linear interactions that machine learning models can elucidate. Particularly, the global economic crisis and growth without employment have led to increased unemployment rates and job instability [3], diminishing the concept of lifelong employment. As a result, the importance of facilitating the inflow and outflow of talent and the necessity of securing and retaining skilled personnel have become increasingly pronounced. This study addresses these issues by introducing a novel hybrid model, XGBoost with Bagging, aimed at offering a deeper analysis of turnover factors, thereby filling a significant gap in current research methodologies. As the proportion of knowledge workers in the South Korean labor market grows, their turnover significantly impacts both corporate and national economies [4]. The hiring and organizational adaptation of new employees are recognized as key means of securing a company's growth potential. However, turnover among organizational members, where employees move to different organizations after a certain period, poses a substantial challenge for human resource management [4]. Existing research methods often rely on regression analysis, which assumes linear relationships and fails to capture the nuanced interactions present in labor dynamics. Machine learning

models, by contrast, can process and analyze complex datasets, offering insights that traditional methods cannot. The initial entry into the labor market for young workers is often marked by unstable information acquisition and a higher likelihood of both voluntary and involuntary turnover, which can adversely affect their career development. Turnover among recent college graduates disrupts individual career development, increases recruitment and training costs, and leads to the loss of educational investments, causing frustration and job dissatisfaction among existing employees. By utilizing advanced machine learning techniques, this study not only predicts turnover behavior but also provides a comprehensive understanding of the underlying factors, such as job satisfaction and organizational culture, that influence these decisions. This issue is particularly critical for recent graduates who have the potential to become key human resources for corporate performance and national development. Studying their turnover is crucial for addressing youth unemployment [4, 5] and preventing the wastage of employment budgets. Despite its importance, there is a lack of comprehensive research on the turnover of recent college graduates.Thus, this research contributes to the literature by employing a machine learning framework to uncover complex patterns and interactions, offering novel insights into early career turnover behavior.

In contrast, machine learning models have recently been employed across various fields to address the limitations of traditional statistical models [4]. These models are powerful tools for learning patterns and making predictions from large datasets, making them particularly valuable in the labor market and human resource management contexts. Machine learning models can accurately detect and predict complex, non-linear relationships that are typical in real-world turnover behavior [6]. They can process a wide range of variables and large datasets simultaneously, allowing for a comprehensive analysis of multidimensional factors influencing turnover. Our study leverages these capabilities to offer a nuanced analysis, utilizing the 2019 Graduates Occupational Mobility Survey (GOMS) data to explore turnover phenomena among young graduates. For instance, factors such as job satisfaction, organizational culture, individual career goals, and economic conditions can all be integrated and analyzed through machine learning models [7, 8].

This approach not only enhances predictive accuracy but also provides HR professionals with actionable insights, enabling them to develop strategies that effectively address turnover issues. Therefore, utilizing machine learning for analysis offers a more precise and in-depth understanding of turnover behavior among recent college graduates. The goal of this study was to develop and utilize a novel hybrid model, XGBoost with Bagging aggregation, to analyze and predict the multifaceted and critical factors influencing turnover behavior among early-career employees aged 20 to 30. This innovative approach establishes a new benchmark in turnover research, highlighting the model's superiority in capturing complex interactions that traditional methods overlook. This research leveraged data from the 2019 Graduates Occupational Mobility Survey (GOMS) to conduct an in-depth analysis of turnover phenomena among early-career graduates, assess the impact of these factors on their career development, and provide significant insights for the formulation of human resource management strategies at both corporate and national levels. Additionally, the study evaluated the performance of the proposed XGBoost with Bagging model against several baseline models, including KNN, Logistic Regression, SVM, Bagging, and XGBoost, to establish a reliable and interpretable predictive tool for HR professionals. By offering a detailed comparative analysis, the study underscores the model's ability to outperform existing techniques, thus contributing to more effective human resource management.

2. Methodology

2.1. Research Overview

The goal of this study is to develop and utilize a novel hybrid model, XGBoost with Bagging aggregation, to analyze and predict the multifaceted and critical factors influencing turnover behavior among early-career employees aged 20 to 30. This model addresses the limitations of traditional statistical methods by capturing complex, non-linear interactions within the data, providing a comprehensive analysis of turnover phenomena. XGBoost, a powerful variant of the Gradient Boosting algorithm, is known for its high predictive performance and overfitting prevention capabilities. Bagging involves independently training multiple models and aggregating their predictions to improve performance. By integrating these techniques, the study aims to enhance predictive accuracy and offer robust insights into turnover behavior. The systematic diagram of this study is presented in Figure 1.

2.2. Data Collection and Preprocessing

This research utilized data from the 2019 Graduates Occupational Mobility Survey (GOMS) to analyze factors influencing turnover among young workers. The dataset provides a rich source of information, capturing a wide range of variables that impact labor market transitions. The data preprocessing steps were meticulously designed to prepare the dataset for analysis:

- Handling Missing Values: Missing values were either removed or replaced with mean or median values.
- Categorical Data Encoding: Categorical variables were converted into numerical variables using one-hot encoding.
- Normalization: All numerical variables were scaled using Min-Max normalization or Z-normalization.
- Data Splitting: The data was split into training and validation sets in an 80:20 ratio to evaluate model performance.

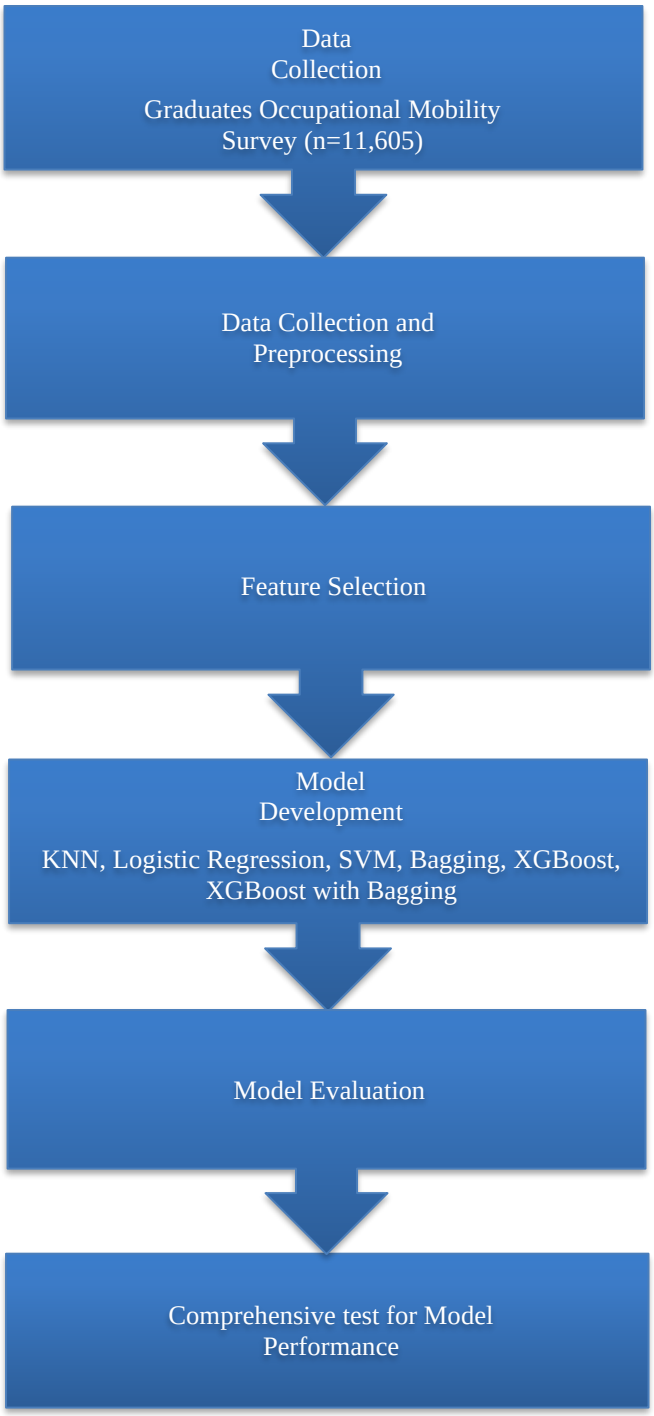

Fig. 1 Systematic diagram of this study

2.3. Feature Selection and Model Development

This study developed an XGBoost with Bagging model by combining the strengths of both XGBoost and Bagging algorithms. XGBoost builds strong predictive models by sequentially training multiple weak learners (typically decision trees) to correct the errors of previous models. Bagging enhances performance by training multiple models independently and aggregating their predictions.

2.4. XGBoost Algorithm

XGBoost (eXtreme Gradient Boosting) is an improved version of the Gradient Boosting algorithm. The objective function of XGBoost is defined as:

$$\left[\text{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}i) + \sum k = 1^K \Omega(f_k)\right]$$

Where (l) is the loss function, (\Omega) is the regularization term, and (\Theta) represents the set of model parameters. The loss function measures the difference between the predicted values ((\hat{y}_i)) and the actual values ((y_i)), while the regularization term controls the model complexity to prevent overfitting. XGBoost uses the Second-Order Taylor Expansion to approximate the objective function and perform optimization:

$$\left[\text{Obj}(\Theta) \approx \sum_{i=1}^{n}\left[l\left(y_i, \widehat{y_i^{(t)}}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t)\right]$$

Where (g_i) and (h_i) are the first and second-order derivatives of the loss function, respectively. The learning process of XGBoost involves updating the prediction function (F_t(x)) at each step (t) by adding a new decision tree (h_t(x)):

$$[F_t(x) = F_{t-1}(x) + \eta h_t(x)]$$

Here, (\eta) is the learning rate, and (h_t(x)) is the new decision tree that minimizes the loss:

$$\left[h_t(x) = \arg\min_h \sum_{i=1}^{n} l\left(y_i, F_{t-1}(x_i) + h(x_i)\right)\right]$$

2.5. Bagging Algorithm

Bagging (Bootstrap Aggregating) improves model performance by reducing variance and preventing overfitting. The key steps in Bagging are:

- Data Sampling: Generate multiple bootstrap samples from the original dataset with replacement.
- Model Training: Train independent models on each bootstrap sample.
- Prediction Aggregation: Aggregate the predictions from all models by averaging or majority voting.

2.6. XGBoost with Bagging Model

The model proposed in this study combines XGBoost and Bagging, two powerful ensemble learning techniques, to create a more robust and accurate predictive algorithm. XGBoost, which stands for eXtreme Gradient Boosting, is an enhancement of the Gradient Boosting Machine (GBM) algorithm. It is widely used due to its strong predictive performance, fast execution speed, and overfitting prevention capabilities. On the other hand, Bagging, short for Bootstrap Aggregating, reduces variance and prevents overfitting by training multiple independent models in parallel and aggregating their results. The XGBoost with Bagging model trains multiple decision trees using bootstrap samples (i.e.,

samples drawn with replacement from the original data) in each iteration of XGBoost. The predictions of these trees are then averaged to produce the final prediction. This process can be mathematically expressed as follows:

$$\left[F(x) = \sum_{i=1}^{N} f_i(x)\right]$$

Where (F(x)) is the final prediction model, (N) is the number of decision trees generated from the bootstrap samples, and (f_i(x)) is the prediction function of the (i)-th decision tree. Each (f_i(x)) is calculated as follows:

$$[f_i(x) = wq(x)]$$

Where (w) represents the weight of the leaf node and (q(x)) is the function that determines which leaf node a particular data point (x) belongs to within the tree.The learning process of XGBoost aims to minimize the loss function (L), which the following equation can represent:

$$\left[L(\phi) = \sum_i l(y_i, \hat{y}i) + \sum k\,\Omega(f_k)\right]$$

Here, (l(y_i, \hat{y}_i)) is the function measuring the loss between the actual value (y_i) and the predicted value (\hat{y}_i), and (\Omega(f_k)) is the regularization term that controls the complexity of the (k)-th tree. The XGBoost with Bagging model applies Bagging to the basic structure of XGBoost by independently training models on each bootstrap sample and averaging their predictions to produce the final output. This approach helps prevent overfitting and enhances the generalization capability of the model. By doing so, the XGBoost with Bagging model allows each tree to learn different aspects of the data, thereby contributing to the overall stability and accuracy of the model. Additionally, this model maintains fast training speeds through parallel processing and provides more robust predictions due to the diversity achieved through Bagging. The XGBoost with Bagging model combines the strengths of both methods. The process is as follows:

- Bootstrap Sampling: Generate (B) bootstrap samples from the original dataset.
- XGBoost Model Training: Train an XGBoost model on each bootstrap sample. Here, (\Theta_b) represents the parameters of the XGBoost model trained on the (b)-th bootstrap sample.
- Prediction Aggregation: Aggregate the predictions from all XGBoost models by averaging:

$$\left[\hat{y} = \frac{1}{B}\sum_{b=1}^{B} \widehat{y_b}\right]$$

Where (\hat{y}_b) is the prediction from the (b)-th XGBoost model. This approach maintains the high predictive performance of individual XGBoost models while leveraging Bagging to enhance stability and prevent overfitting. The combined model provides accurate predictions of turnover among young workers.

2.7. Model Evaluation

The performance of the models was evaluated using AUC (Area Under the Curve) and F1-score. AUC measures the area under the ROC curve, indicating the model's classification performance. F1-score is the harmonic mean of precision and recall, which is useful for imbalanced datasets. Each model's performance was assessed through cross-validation, and the XGBoost with Bagging model was found to outperform the other models.

Through this methodology, we developed a turnover prediction model for young workers and identified the best model through comprehensive performance comparisons.

2.8. Comprehensive Test for Model Performance

This study aims to establish a comprehensive benchmark for evaluating the performance of the XGBoost with the Bagging model by implementing and assessing several other models: k-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Bagging, and XGBoost. The selection of these models is intended to create a comparative framework to assess the performance of the XGBoost with the Bagging model thoroughly.

2.9. k-Nearest Neighbors (KNN)

The KNN model is a simple, non-parametric method used for classification and regression. In this study, KNN was configured with various values of (k) to find the optimal number of neighbors. Cross-validation was performed to determine the best (k) value, which was found to be 7.

$$\left[\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i\right]$$

2.10. Logistic Regression

Logistic Regression is a widely used statistical method for binary classification problems. In this study, both L1 and L2 regularization techniques were employed to prevent overfitting. The optimal regularization parameters were determined through cross-validation.

$$\left[P(Y = 1 \mid x) = \frac{1}{1 + e^{-(\alpha + \beta^T x)}}\right]$$

Where(α) is the intercept, (β) is the coefficient vector, and (x) is the predictor variable.

2.11. Support Vector Machine (SVM)

The SVM model was used with the Radial Basis Function (RBF) kernel to capture non-linear relationships. A grid search was conducted to find the optimal values for the kernel parameters (C) and (γ).

$$\left[\min_{\mathbf{w},\xi}\left(\frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{n}\xi_i\right)\right]$$

subject to:

$$[y_i(\mathbf{w}\cdot\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0]$$

2.12. Bagging

Bagging, or Bootstrap Aggregating, is an ensemble method that improves the stability and accuracy of machine learning models. In this study, Bagging was implemented using decision trees as base learners. A total of 100 decision trees were aggregated to make the final predictions.

$$\left[\hat{Y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x)\right]$$

Where (T) is the total number of trees, and$\left(f_t(x)\right)$is the prediction made by the (t)-th tree.

2.13. XGBoost with Bagging

The XGBoost with Bagging model combines the strengths of XGBoost and Bagging. In this approach, XGBoost serves as the base learner, and Bagging is applied to enhance stability and reduce overfitting. A total of 50 XGBoost models were trained on bootstrap samples, and their predictions were aggregated.

2.14. Data Source

In this study, panel data from the Korea Employment Information Service's '2019 Graduates Occupational Mobility Survey (GOMS)' was utilized. The "College Graduates' Occupational Movement Path Survey" is a national statistical survey that investigates various aspects of graduates' transitions from education to the labor market. This includes school education, employment preparation, work experience, and vocational training among graduates from 2-year, 4-year, and education colleges. The primary objective of this survey is to provide reliable information on the supply and demand of manpower between the education sector and the labor market and to offer foundational data for policy-making aimed at alleviating discrepancies through long-term follow-up on the career development and job mobility paths of college graduates. The survey population comprises graduates from 2-year colleges, 4-year universities, and education colleges. Each year, a sample of 18,000 graduates from the previous year is selected using a complex sampling method.

GOMS data encompass a wide range of variables that can influence entry and settlement in the labor market. These variables include educational courses, job search activities, work experience, vocational training, certifications, personal information, and the family background of college graduates. This dataset is extensively used for analyzing the transition performance from the education market to the labor market, job mismatch analysis, and evaluating the return on investment in education.

In this study, we analyzed data from 11,605 participants who are currently employed, aged 20 to 39, and have a college degree or higher. All subjects in this study were graduates from August 2018 to February 2019. Part-time jobs were excluded from the definition of workers. Table 1 summarizes the main variables included in the GOMS panel data and their characteristics.

Table 1. The main variables included in the GOMS data

Variable Category	Variable Name
Educational Background	University Name
	Major
Job Search	Job Search Duration
Job Experience	Work Duration
Vocational Training	Training Completion
Certification	Certification Ownership
Personal Information	Age
Household Background	Household Income
Employment Status	Determination of Employment Status
Current Job	Job Retention at Initial Survey
	Industry and Occupation
	Employment Status and Type
	Changes from the Initial Survey
Job Changer/New Hire	Entry Time
	Job Search Activities
	Industry and Occupation
	Employment Status and Type
Common	Promotion Experience, Timing, Frequency
	Fringe Benefits and Social Insurance
	Non-wage Employment and Startup Preparation
	Job Type and Number of Employees
	Company Location
	Working Hours and Income
	Satisfaction and Job Level
	Foreign Language Utilization and Proficiency
	Concurrent Jobs
	Job Change Preparation and Desired Job
Job Search	Job Search Duration and Efforts
	Desired Employer and Working Conditions
	Difficulties and Stress in Job Search
Non-economic Activity	Last Job Search Timing
	Methods of Solving Economic Problems
Previous Job	Duties at Time of Resignation
	Employment Status and Type at Time of Resignation
	Working Hours and Income at Time of Resignation
	Reason for Leaving

Experienced Job	Work Duration
	Job Search
	Industry and Occupation
	Job Type and Number of Employees
	Company Location
	Satisfaction and Job Level
	Employment Status and Type
	Working Hours and Income
	Reason for Leaving
School Experience	Further Education after College Graduation
	Graduation Status of School Attended at Last Survey
	Current Enrollment Status
Job Preparation	Job-related Education and Training
	Certification
	Exam Preparation
Personal Details	Personal Details, Family Education and Occupation, Income

3. Results

3.1. Model Performance

This section details the performance of the XGBoost with the Bagging aggregation model for predicting turnover behavior among early-career employees aged 20 to 30. The model's effectiveness was compared to several baseline models: KNN, Logistic Regression, SVM, Bagging, and XGBoost. Evaluation metrics such as Area Under the Curve (AUC) and F1-score were used to assess the model's predictive capabilities thoroughly.

3.2. Initial Model Performance

Initially, all models were trained and validated using their default hyperparameters on the entire dataset to establish a performance benchmark. Table 2 and Figure 2 summarize the results.

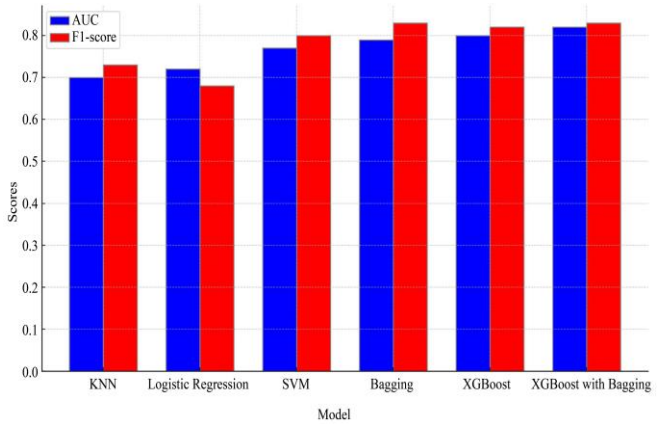

Fig. 2 Pre-tuning performance of models with default parameters

Table 2. Performance of models with default settings before hyperparameter tuning

Model	AUC	F1-score
KNN	0.70	0.73
Logistic Regression	0.72	0.68
SVM	0.77	0.80
Bagging	0.79	0.83
XGBoost	0.80	0.82
XGBoost with Bagging	0.82	0.83

3.3. Hyperparameter Tuning

To enhance the performance of the models, a grid search with cross-validation was employed for hyperparameter tuning. For the XGBoost with Bagging model, the key hyperparameters included the number of trees, learning rate, maximum depth of each tree, and the minimum number of samples required to split an internal node. The specific ranges for hyperparameter tuning were:

- Number of Trees (n_estimators): 50 to 500
- Learning Rate (learning_rate): 0.001 to 0.2 (log scale)
- Maximum Depth (max_depth): 3 to 10
- Minimum Samples per Split (min_samples_split): 2 to 20
- Subsample: 0.6 to 1.0
- Colsample_bytree: 0.6 to 1.0
- The optimal hyperparameters identified for the XGBoost with the Bagging model were:
- Number of Trees: 300
- Learning Rate: 0.05
- Maximum Depth: 7
- Minimum Samples per Split: 10
- Subsample: 0.8
- Colsample_bytree: 0.8

3.4. Model Performance Post-Hyperparameter Tuning

After hyperparameter tuning, the models were reevaluated. The results are shown in Table 3 and Figure 3. The optimized models' performance illustrates the superior performance of the XGBoost with the Bagging model.

The results indicate that the XGBoost with Bagging model outperforms other models in terms of AUC and maintains a balanced performance in terms of F1-score. The improvement in AUC from 0.82 to 0.85 demonstrates the effectiveness of the hyperparameter tuning process.

Table 3. Performance of models with optimized hyperparameters

Model	AUC	F1-score
KNN	0.72	0.75
Logistic Regression	0.74	0.70
SVM	0.79	0.82
Bagging	0.81	0.85
XGBoost	0.83	0.84
GBoost with Bagging	0.85	0.86

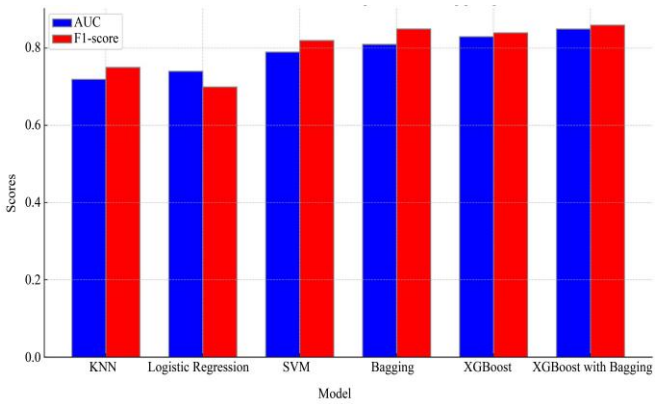

Fig. 3 Performance of Optimally Tuned Models

3.5. Comparative Analysis

The comparative analysis highlights the superior performance of the XGBoost with the Bagging model. The model achieved the highest AUC and demonstrated a strong balance between recall and precision, making it a robust tool for predicting turnover behavior. Figure 4 summarizes the comparative performance of the models.

Table 4. Top 10 important features for the XGBoost with Bagging model

Feature	Importance
Permanent Employment (types of employment)	0.220
Salary (monthly average)	0.195
Job Satisfaction	0.170
Job Security	0.150
Career Advancement (individual's potential for growth)	0.120
Skill Level of Work (Match between current job skill level and one's skill level)	0.100
Personal Development	0.085
Relationship with one's colleagues	0.065
Industry Type	0.050
Company Size	0.045

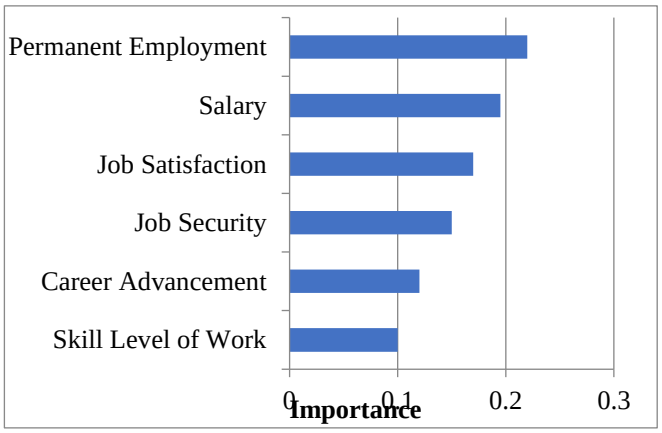

Fig. 4 Main 6 features of highest importance for the XGBoost with Bagging model

3.6. Feature Importance

To understand the factors influencing turnover intentions, feature importance was analyzed for the XGBoost with Bagging model. Table 4 lists the top 10 most important features. The analysis reveals that permanent employment status, salary, job satisfaction, job security, and career advancement are the most significant predictors of turnover intentions.

4. Discussion

This study utilized data from the Employment and Labor Panel Survey to analyze the factors influencing turnover among young workers aged 20 to 30 and to compare the predictive performance of various machine learning models. The results indicated that the XGBoost with Bagging model exhibited the highest performance. The findings that the XGBoost with Bagging model exhibited the highest performance in analyzing factors influencing turnover among young workers underscore the strengths of advanced machine learning models in tackling complex datasets and relationships. This advantage extends broadly across numerous applications, as highlighted in several studies [9-15]. For instance, the implementation of XGBoost feature importance could improve LightGBM's performance in bankruptcy prediction, emphasizing the adaptability and effectiveness of complex machine learning techniques over traditional models [16].

Additionally, machine learning techniques allowed for the extraction of valuable insights from complex economic data, overcoming challenges faced with traditional statistical methods [17]. Also, in the industrial field, a deep learning framework specialized for anomaly detection showcased superiority over conventional techniques in computing efficiency and detection accuracy, demonstrating its effectiveness with large-scale data [18]. Furthermore, the use of machine learning models in predicting the performances of metal-organic frameworks for trace CH3I capture illustrated their advanced capabilities in handling complex patterns and large data volumes [19].

These findings suggest that future research should explore the integration of machine learning models with other advanced analytical techniques to enhance predictive performance further and provide deeper insights into turnover behavior and other complex phenomena. For instance, job satisfaction has been identified as a significant behavioral indicator of turnover intention among young professionals in China [20]. Similarly, other predictors, such as career facilitation and skill utilization, have been recognized for their impact on turnover intentions [21].

Additionally, factors such as organizational commitment, job satisfaction, work environment [22-24], and promotion and compensation have been found to

influence turnover intentions. Interestingly, the role of job satisfaction and social support has also been highlighted. Job satisfaction and career advancement have been seen as factors that encourage employees to stay, thereby reducing turnover intentions [25, 26].

Moreover, salary reference points—including minimum requirements, status quo, and goal levels—along with pay satisfaction and estimated peer salaries have been examined for their effects on turnover intention [26]. This diversity in predictors across studies underscores the complexity of the turnover intention phenomenon and highlights the need for organizations to approach it from multiple angles. By considering both individual and organizational factors, organizations can more effectively manage and mitigate turnover intentions.

The findings of this study can assist HR departments and corporate training managers in more effectively utilizing turnover prediction models to enhance human resource management strategies. For instance, turnover prediction models can help in the early identification of employees with a high likelihood of leaving the organization. This enables the provision of tailored training programs for these individuals or the implementation of measures to improve overall job satisfaction within the organization. The study has several limitations.

First, the data scope and generalizability are constrained as the study utilizes data from the Korea Employment Information Service's 2022 GOMS. While this dataset provides comprehensive insights into the labor market transitions of young graduates in Korea, the findings may not be generalizable to other countries or regions with different economic conditions, labor market structures, or cultural contexts. Second, the cross-sectional nature of the data captures information at a single point in time, limiting the ability to infer causal relationships between the identified factors and turnover intentions. Longitudinal data, which tracks individuals over time, would provide a more robust framework for understanding the dynamic nature of career development and turnover behavior.

Third, although the GOMS dataset includes a wide range of variables, it may not capture all the factors influencing turnover intentions. For instance, variables such as organizational culture, leadership style, and external economic conditions could also play significant roles but were not included in the analysis.Lastly, while the XGBoost with Bagging model offers high predictive accuracy, its complexity can make it challenging to interpret the results. Decision-makers may find it difficult to derive actionable insights from complex models. Future research should consider using diverse datasets from multiple regions, incorporating additional relevant variables, and exploring the use of simpler models or techniques that enhance interpretability without significantly compromising predictive performance.

5. Conclusion

This research employed data from the Employment and Labor Panel Survey to investigate the factors influencing turnover among young professionals aged 20 to 30 and to evaluate the predictive capabilities of various machine learning models. The study found that the XGBoost with Bagging model outperformed other models, underscoring the effectiveness of advanced machine learning techniques in processing large datasets and complex interactions. Consistent with prior studies, the results emphasized the significance of factors such as permanent employment, salary, job satisfaction, job security, and career advancement in predicting turnover intentions. These findings hold valuable implications for HR departments and corporate training managers.

By leveraging turnover prediction models, organizations can identify at-risk employees early and implement targeted interventions, such as customized training programs or initiatives to enhance job satisfaction. This comprehensive approach to managing turnover intentions highlights the importance of addressing both individual and organizational factors. Future research should explore the combination of machine learning models with other advanced analytical methods to improve predictive accuracy and offer deeper insights into turnover behavior and other complex issues.

Acknowledgment

The Basic Science Research Program supports this research through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF- RS-2023-00237287, NRF-2021S1A5A8062526) and local government-university cooperation-based regional innovation projects (2021RIS-003).

References

[1] Pei Xu, and Ke Zhang, "Research on Human Resource Management from the Perspective of Competency," *Proceedings of the 2018 International Conference on Management, Economics, Education and Social Sciences*, vol. 236, pp. 7-9, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Hyounju Kang, "*Work-Life Balance in South Korea: Experiences of the Highly Educated and Married Female Korean Employees with Flexible Workplace Arrangements*," Doctor of Philosophy, Electronic Theses, Texas A&M University, pp. 1-248, 2016. [Google Scholar] [Publisher Link]

[3] Gary S. Fields, "The Employment Problem in Korea," *Journal of the Korean Economy*, vol. 1, no. 2, pp. 207-227, 2000. [Google Scholar] [Publisher Link]

[4] Sookyung Park, Sungmin Lee, and Jongphil Bae, "The Relationships between Job Burnout, Supervision and Turnover Intention of Early-Career Social Workers in South Korea," *Korean Journal of Social Welfare Research*, vol. 65, pp. 135-164, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Lihe Ma, "Employee Turnover Prediction Based on Machine Learning Model," *2022 5th Asia Conference on Machine Learning and Computing*, Bangkok, Thailand, pp. 22-27, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Cecil Mlatsheni, *The Youth Labour Market in South Africa*, The Oxford Handbook of the South African Economy, pp. 690-706, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Markus Atef, Doaa Elzanfaly, and Shimaa Ouf, "Early Prediction of Employee Turnover Using Machine Learning Algorithms," *International Journal of Electronics and Communication Engineering Studies*, vol. 13, no. 2, pp. 135-144, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Chenyu Liao, "Employee Turnover Prediction Using Machine Learning Models," *International Conference on Mechatronics Engineering and Artificial Intelligence*, Changsha, China, vol. 12596, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Joseph S. Harrison et al., "Using Supervised Machine Learning to Scale Human-Coded Data: A Method and Dataset in the Board Leadership Context," *Strategic Management Journal*, vol. 44, no. 7, pp. 1780-1802, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz, "A Comparative Analysis of Gradient Boosting Algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937-1967, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Malak Abdullah, Doaa Obeidat, and Heba Nammas, "Using Ensemble Machine Learning Algorithms to Predict a Scrabble Player's Rating," *2023 14th International Conference on Information and Communication Systems*, Irbid, Jordan, pp. 1-6, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Keyou S. Mao et al., "Identifying Chemically Similar Multiphase Nanoprecipitates in Compositionally Complex Non-Equilibrium Oxides via Machine Learning," *Communications Materials*, vol. 3, pp. 1-13, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] S. Keerthika et al., "Enhancing Soil Moisture Prediction through Machine Learning for Sustainable Resource Management," *2023 7th International Conference on Electronics, Communication and Aerospace Technology*, Coimbatore, India, pp. 1175-1179, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] N. Savitha, and K. Ravikumar, "Machine Learning Techniques for Agriculture Crop Recommendations Based on Productivity: A Survey," *International Journal of Science and Research*, vol. 12, no. 10, pp. 111-116, 2023. [CrossRef] [Publisher Link]

[15] Felipe Pérez de los Cobos et al., "First Large-Scale Peach Gene Coexpression Network: A New Tool for Predicting Gene Function," *Cold Spring Harbor Laboratory*, pp. 1-23, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Risma Moulidya Syafei, and Devi Ajeng Efrilianda, "Machine Learning Model Using Extreme Gradient Boosting (XGBoost) Feature Importance and Light Gradient Boosting Machine (LightGBM) to Improve Accurate Prediction of Bankruptcy," *Recursive Journal of Informatics*, vol. 1, no. 2, pp. 64-72, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Zhekai Liu, "Review on the Influence of Machine Learning Methods and Data Science on the Economics," *Applied and Computational Engineering*, vol. 22, pp. 137-141, 2023. [CrossRef] [Publisher Link]

[18] R. Anuradha et al., "Deep Learning for Anomaly Detection in Large-Scale Industrial Data," *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering*, Gautam Buddha Nagar, India, pp. 1551-1556, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Xiaoyu Wu et al., "Mapping the Porous and Chemical Structure-Function Relationships of Trace CH_3I Capture by Metal-Organic Frameworks Using Machine Learning," *ACS Applied Materials & Interfaces*, vol. 14, no. 41, pp. 47209-47221, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Nailin Bu, Carol A. McKeen, and Wenguo Shen, "Behavioural Indicators of Turnover Intention: The Case of Young Professionals in China," *International Journal of Human Resource Management*, vol. 22, no. 16, pp. 3338-3356, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[21] Thomas M. Mitchell, and Benjamin Schneider, "*Work and Career Considerations in Understanding Employee Turnover Intentions and Turnover: Development of the Turnover Diagnostic*," Psychology, no. 84-2, pp. 1-44, 1984. [Google Scholar] [Publisher Link]

[22] Eric G. Lambert et al., "A Test of a Turnover Intent Model," *Administration in Social Work*, vol. 36, no. 1, pp. 67-84, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[23] Yoon Jik Cho, and Gregory B. Lewis, "Turnover Intention and Turnover Behavior: Implications for Retaining Federal Employees," *Review of Public Personnel Administration*, vol. 32, no. 1, pp. 4-23, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[24] Melisa Erdilek Karabay et al., "Analyzing The Effect of Antecedents of Turnover Intention According to Generations," *European Proceedings of Social and Behavioural Sciences*, vol. 54, pp. 578-589, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jinuk Oh, and Nita Chhinzer, "Is Turnover Contagious? The Impact of Transformational Leadership and Collective Turnover on Employee Turnover Decisions," *Leadership & Organization Development Journal*, vol. 42, no. 7, pp. 1089-1103, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Guanxing Xiong, X.T. Wang, and Aimei Li, "Leave or Stay as a Risky Choice: Effects of Salary Reference Points and Anchors on Turnover Intention," *Frontiers in Psychology*, vol. 9, pp. 1-10, 2018. [CrossRef] [Google Scholar] [Publisher Link]