9th North American Conference on Industrial Engineering and Operations Management
Washington D.C., USA, June 4-6, 2024
Publisher: IEOM Society International, USA DOI: 10.46254/NA09.20240026
Published: June 4, 2024

Human Resources Analytics: Determining Possible Turnovers with Feature Engineering Approach

Eren Darici, Kira K. Hamelink and Ilgin Acar
Department of Industrial and Entrepreneurial Engineering and Engineering Management
Western Michigan University
Kalamazoo, MI, USA
eren.darc@wmich, kira.k.hamelink@wmich.edu, ilgin.acar@wmich.edu

Abstract

In human resources operations, it is crucial to anticipate potential employee turnovers to implement effective planning strategies, like employee recruitment and training planning. In addition to exploring the feasibility of predicting employee turnover based on non-job-specific features, this study identifies the most important candidate characteristics for predicting potential turnover. These features encompass demographic and educational data and are applicable across various industries rather than being confined to a specific industry. The dataset used is characterized by its imbalanced nature where class proportions are skewed, necessitating the inclusion of every data point for comprehensive analysis. A comparative investigation was conducted, evaluating different sampling methodologies to handle imbalanced data including up-sampling and down-sampling, alongside various classification algorithms such as ensemble learning techniques and Support Vector Machines (SVM). As the result of this study, the significance of each feature was determined through the application of the most effective model, namely Random Forest, which achieved an accuracy rate of 87.3% and an area-under-curve score (AUC) of 87.3%, which exceeded previous studies using the same dataset. These metrics not only indicate the model's ability to correctly classify employees into potential turnover and non-turnover categories with high accuracy, but also highlight its capability to minimize false positives and false negatives, which is crucial for decision-making processes.

Keywords

Human Resources Management, Human Resources Analytics, Employee Turnover Prediction, Feature Engineering, Classification Algorithms

1. Introduction

In one of the primary operations of Human Resources Management Planning (HRMP), a sub-sect of Human Resources Management (HRM), the effective fulfillment of an organization's requirements is accomplished by determining appropriate staffing levels and necessary skills for hired staff (Mouzala, 2022). Training newly hired employees is a proven way to both motivate employees (Ozkeser, 2019) and improve their job performance. However, the administration of training for new hires comes with a monetary and time cost; according to the Association of Talent Development's (ATD) 2016 report (ATD Releases 2016 State of the Industry Report, 2015), the average cost of training is 84 dollars per hour as of 2015, and the average required training time is 33.5 hours per employee. According to the same report, organizations around the world spent an average of 1,252 dollars for the training and development of each new employee hired.

Although the definition of turnover may vary among different organizations, employee turnover is generally measured as the number of employees who leave the organization within some specified timeframe. In literature, attrition and turnover are often used interchangeably for employees leaving jobs at an organization. However, attrition refers to employees leaving a position for natural or voluntary reasons, such as death or retirement, whereas the turnover metric focuses on employees leaving the organization for work at a new organization, or for personal reasons (*Turnover vs. Attrition*, 2023). Turnover and attrition involve significant amounts of costs in different levels of human resources operations from recruiting to training. When a turnover occurs, replacement of the lost employee comes with replacement costs that include advertising, candidate interviewing and selection processes, and training after recruitment (Mitchell et al., 2001).

According to the February 2024 report by the Bureau of Labor Statistics (Job Openings and Labor Turnover - February 2024, 2024), 3.5 million instances of employees quitting were recorded between February 2022 and February 2024. Figure 1 shows the downward trend of the percentage of hires and separations recorded between February 2022 and February 2024. As shown in the figure, the employee-leaving rate has decreased by roughly 0.5% over the two-year study period. Despite the general decrease in employee quitting rates, the total amount of money spent on employee training continues to increase over time when the increasing cost of training and 3.8% turnover rate are considered.

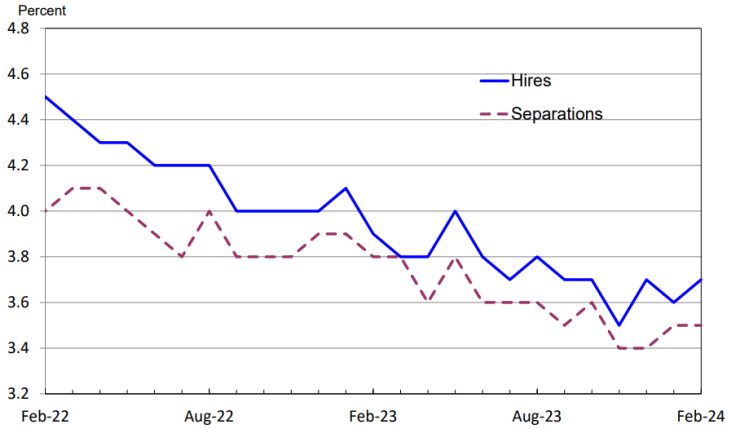

Figure 1. “Hires and total separations rates, seasonally adjusted, February 2022 - February 2024” (Job Openings and Labor Turnover - February 2024, 2023)

Besides incurring a monetary loss, high turnover rates negatively affect the overall public appearance of an organization and its working environment (Lee et al., 2016). Long-term effects of high turnover rates result in reduction in the number of entry-level workers, forcing the affected organization to use temporary workers for short periods (Al-Suraihi et al., 2021).

When prioritizing employee retention, the identification of characteristics which indicate that potential new hires are likely to leave an organization within the first few years of hiring can aid HRM professionals in choosing long-term employees from a candidate pool. These characteristics, referred to in machine learning as “features,” can be divided into two kinds: (1) numerical and (2) categorical. Numerical features are continuous values which can be measured on a given scale, such as height in centimeters. Categorical features are discrete values that can be grouped into categories, such as zip codes. In machine learning (ML), feature engineering is a data preprocessing step that involves transforming raw data into meaningful inputs (Hastie et al., 2009) for a more accurate decision-making outcome.

In addition to exploring the feasibility of predicting employee turnovers based on employee characteristics with a comparative approach to classification algorithms, this study aims to determine the most important employee candidate characteristics which indicate that the candidate is likely to leave their position soon after hiring. Characteristics identified in this study can then be used to aid in HRMP hiring decision-making processes across industries.

2. Literature Review

As stated in the introduction section, selecting candidate employees for hire and their subsequent training are primary operations of HRM. In Ozkeser's 2019 study of the impacts of training as a function of HRM on employee motivation, it is shown that "perceived training opportunities" positively affect employee motivation. Employee motivation, or lack thereof, has been proven to decrease job satisfaction and increase the rate of employee turnover (Al-Suraihi et al., 2021).

In HRM, the integration of ML techniques has revolutionized traditional practices from applicant screening to planning fields, offering new opportunities for operational efficiency, and ML in employee turnover and employee attrition are widely studied topics in literature. This section will focus on recent literature which uses the same or similar data used in this study under varying contexts. Marvin et al. (2021) used the same dataset (HR Analytics, n.d.) that is used in this study to predict turnovers to prevent potential costs associated with replacing lost workforce. They used different sampling techniques to deal with imbalanced data, however, they only up-sampled the dataset and only tested using accuracy, which is an unsuitable metric for imbalanced classification (Jeni et al., 2013). The findings of the research were consistent with the results reported by V Papineni et al. (2021), in which Marvin et al (2021) achieved a training accuracy of 99.1% and a testing accuracy of 84.6%. In contrast, V Papineni et al. (2021) reported 100% training accuracy, which indicates overfitting, where the model performs nearly perfectly on the training data but demonstrates a lack of performance on the testing data. Another study using the same data was conducted by Conlon (2021), where feature importance was mentioned but noted a lack of accuracy, with an accuracy rate of 78.5% and an area-under-curve (AUC) score of 80.5%. Jain and Nayyar, (2018) used a different dataset (IBM HR Analytics Employee Attrition & Performance, n.d.) to predict employee attrition. Using a feature engineering approach, they created three new features as a better representation of the existing ones. The resulting accuracy was 89.1% using an XGBoost (Chen and Guestrin, 2016) model. In this study, however, the imbalanced nature of the dataset was left unaddressed, and no other metrics were considered to evaluate the developed models.

In existing literature, many approaches to determine turnover and attrition have been studied on different datasets. However, most of these studies focus only on accuracy as a metric, neglecting the significance of False Positives (FP) (misclassifying an employee as a potential turnover) and False Negatives (FN) (misclassifying a potential turnover as a staying employee), which are crucial for planning purposes. Moreover, imbalanced data handling and preprocessing techniques are often overlooked or superficially addressed.

3. Data Collection

The dataset used in this study is a publicly available dataset found on the Kaggle platform (HR Analytics, n.d.). The context of the dataset is that a company in the Big Data and Data Science field wishes to hire data scientists after a series of courses conducted by the company. To reduce the monetary and time costs associated with hiring new employees, as well as to improve the quality of training that will be given to the candidates, the company decides to attempt to predict potential future turnover. The dataset contains 13 features detailing information about each candidate, such as the city where they are located, their highest level of education, and the candidate's major discipline. A full list of characteristics for each candidate can be found in appendix X. The collected data only consists of current credentials, demographics, and experience data that are not industry-specific; the data used in this study can be collected by companies in different industries to test potential applications.

Figure 2 shows the distribution of those candidates which are not looking for a job change, represented as a target value of 0.0, and those who are looking for a job change, represented as a target value of 1.0.

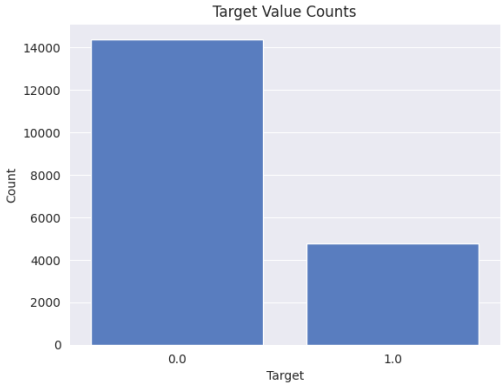

Figure 2. Proportions of target classes

As can be seen from Figure 2, the data is highly imbalanced where nearly 75% of the 19,158 candidates are not looking for a job change while only 25% of the candidates are looking for a job change. In general, an imbalanced dataset is likely to produce unreliable results unless the issue is properly addressed (Jeni et al., 2013).

4. Methods

The problem addressed in this study is defined as a binary classification problem; the dataset contains only one target variable with two categories: (0) candidates not looking for a job change, and (1) candidates looking for a job change. As stated in the data collection section, the dataset is highly imbalanced which necessitates the use of every data point, along with different metrics to evaluate the models for FPs and FNs.

The first step in data preprocessing is to clean the data based on feature relationships and data types. Data cleaning is a measure typically taken to prepare data for effective processing and includes dealing with any missing data in the dataset. Figure 3 shows a frequency diagram representing the frequencies of missing features, created using the "missingno" module for Python (Bilogur, 2016). In this figure, blank white lines in each column represent a missing feature value for that candidate.

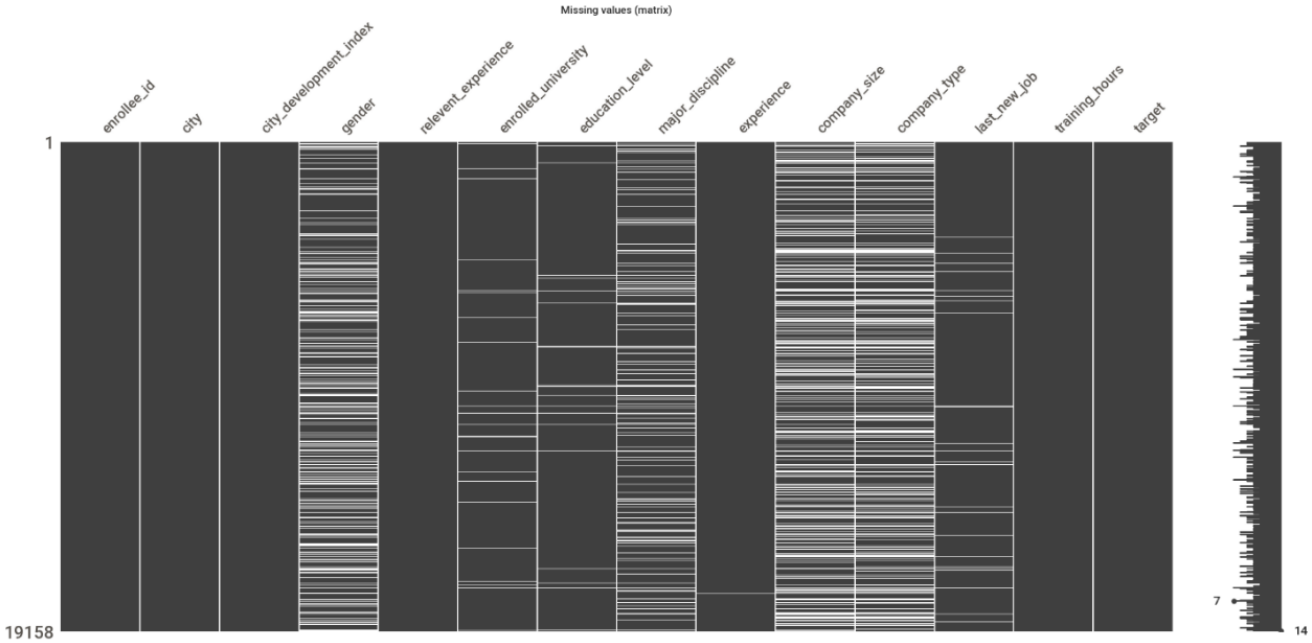

Figure 3. Missing data frequency for each feature

Based on the diagram, which shows a significant number of missing entries for several categories, including gender, company_size, and company_type, it is decided to drop those observations having more than four missing features. After this operation, the dataset is left with 18,280 observations.

In Figure 4, a dendrogram showing the relationships of missing data was created using the "missingno". In the dendrogram, the x-axis labels represent features of the dataset, and the y-axis represents the distance or dissimilarity. The lines connecting these features indicate hierarchical clustering based on similarity in their missing data patterns, meaning that they might have a relationship based on their values; horizontal connecting lines lower on the y-axis (with a higher value) represent dissimilar missing feature patterns, while those features with higher connecting lines share more similar missing feature patterns.

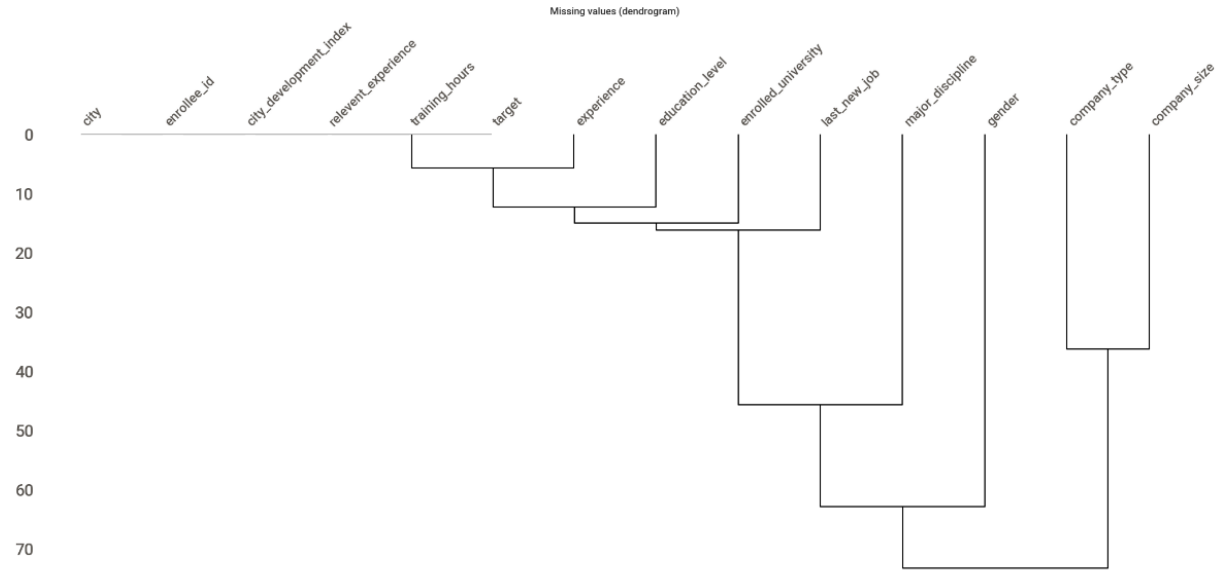

Figure 4. Missing data relationships with dendrogram

Based on the dendrogram and further investigation through the dataset, it is observed that it is possible to fill major discipline features based on their relationships with the other features. When achieved education level, major discipline, and enrolled university features were compared together, any missing major discipline field was filled with appropriate values. For this feature specifically, a new named category was also created for unknown disciplines. The rest of the categorical features were filled with their mode values to preserve their original distributions where missing data was minimal. However, if there was a great amount of missing data, observations with missing values were removed from the dataset altogether. Lastly, binning was applied for all numerical features, besides city development index and training hours, to turn them into categorical features, and label encoding was applied to all categorical features.

For feature selection, a Chi-squared test at the significance level of 0.05 (α=0.05) was conducted between all categorical variables and target variable. Table 1 lists all features found to have a statistically significant impact on each candidate's classification.

Table 1. Significant columns chi-squared test p-values (α=0.05)

Feature
relevant_experience
enrolled_university
education_level
major_discipline
experience
company_size
company_type
last_new_job

In addition to the significant columns, the training hours feature was selected for use in the modeling phase to capture the effect of training hours. Similarly, the gender feature was selected to capture any differences related to gender, and the city development index feature was selected as a better representation of the city feature. However, all company related features were dropped due to substantial missing data.

After feature selection, the dataset was resampled into two different copies: (1) majority class down-sampled to the minority class, and (2) minority class up-sampled to the majority class. All sampling procedures were conducted using bootstrapping (Pedregosa et al., 2011). Then, the data was split into a training dataset and a testing dataset with the respective proportions of 70% and 30%.

For model selection, k-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), and Random Forests (RF) were used. Each model was trained twice, on down-sampled and up-sampled datasets, using cross-validation (cv=5) and grid search with AUC scoring for hyperparameter tuning.

5. Results and Discussion

After running five models with both down-sampled and up-sampled data, results were calculated using scikit-learn's metrics library (Pedregosa et al., 2011). For each run, the accuracy and AUC metric were recorded. Results of the model on the test data are given below in Table 2.

Table 2. Model results of test data
*: calculated using different functions

	Down-sampled		Up-sampled	
Model	**Accuracy**	**AUC**	**Accuracy**	**AUC**
KNN	56.80%	56.80%	82.59%	82.59%
LR	67.04%	67.04%	67.72%	67.72%
SVM	61.24%	61.24%	62.20%	62.20%
DT	69.54%	69.54%	84.35%	84.35%
RF	67.53%	67.53%	87.33%	87.38 & 94%*

As seen from Table 2, the model which resulted in the highest accuracy is a Random Forest (RF) trained on up-sampled data with the maximum depth of 200 along width Gini impurity as criterion.

The result of the model is 87.33% accuracy, surpassing previous studies(Conlon, 2021; Marvin et al., 2021; V Papineni et al., 2021). using the same dataset. Even though it falls a little behind of one previous study (Kyalkond et al., 2022) in terms of accuracy, the other metrics used in that study were not provided for comparison for the model performance on FP and FN. Using RF on up-sampled data, results for the AUC score varied when using different ways to calculate the score. The first score, 87.38%, was recorded with "roc_auc_curve" function from scikit-learn with the predictions on the test set. The second score, 94%, was recorded with the "plot_roc_curve" function from the previous versions of the scikit-learn that is available on different online platforms (Kaggle: Your Home for Data Science, n.d.) using a Grid Search Classifier, that is used for training on the test set. The difference between the two results can be explained by over-fitting for each cross-validation (Cawley and Talbot, 2010).

Table 3. Random Forest on up-sampled – classification report

Class	Precision	Recall	F1-Score	Support
0	0.93	0.81	0.87	4047
1	0.83	0.94	0.88	0.88
Accuracy				
Macro Avg	0.88	0.87	0.87	8094
Weighted Avg	0.88	0.87	0.87	8094

In the classification report (Table 3), the model's performance is evaluated based on its ability to distinguish between candidates who are not actively seeking a job (class 0) and candidates who are actively seeking a job (class 1). Out of a total of 4047 instances classified as not looking for a job (class 0), the model correctly identified 81% (true positives, TP), while 19% were incorrectly classified as candidates looking for a job (false negatives, FN). This suggests instances where candidates were incorrectly labeled as not seeking employment when searching for a job. The precision for class 0 is 0.93, indicating that the proportion of correctly identified instances among all instances predicted as not looking for a job change, while the recall is 0.81, representing the model's ability to correctly identify individuals who are not seeking re-employment out of all actual instances of not seeking re-employment. Conversely, for candidates actively seeking a job change (class 1), out of 4047 instances, the model correctly classified 93% (true positives, TP), while 7% were incorrectly classified as not seeking a job change (false positives, FP). This indicates instances where individuals were wrongly labeled as re-employment seekers when they were not. The precision for class 1 is 0.83, and the recall is 0.93. These metrics collectively provide a comprehensive assessment of the model's effectiveness in accurately classifying individuals based on their job-seeking status, with an overall accuracy of 87% (Figure 5).

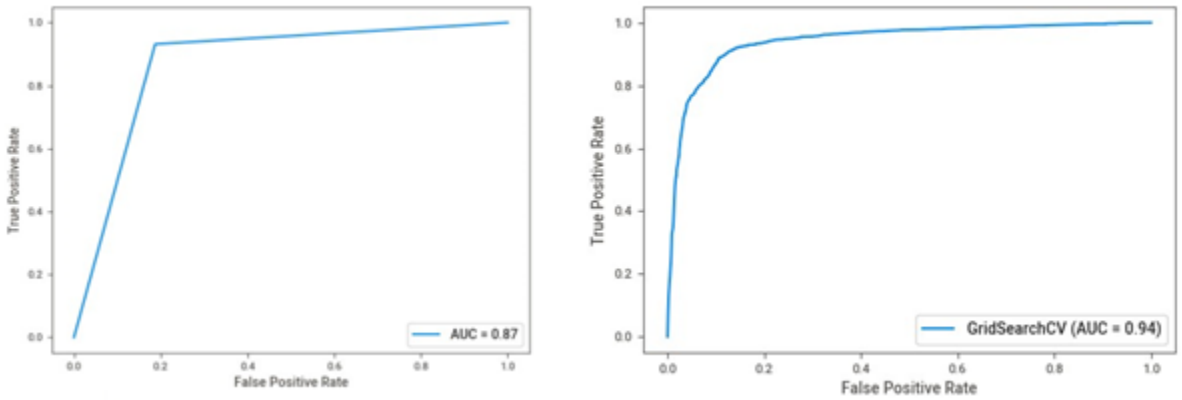

Figure 5. Receiver operating characteristic (ROC) curve with and without cross-validation over-fit

As can be seen from Figure 6, feature importances from Random Forest Classifier on up-sampled data were extracted. According to the results, the most important features were the training hours, city development index and experience, respectively.

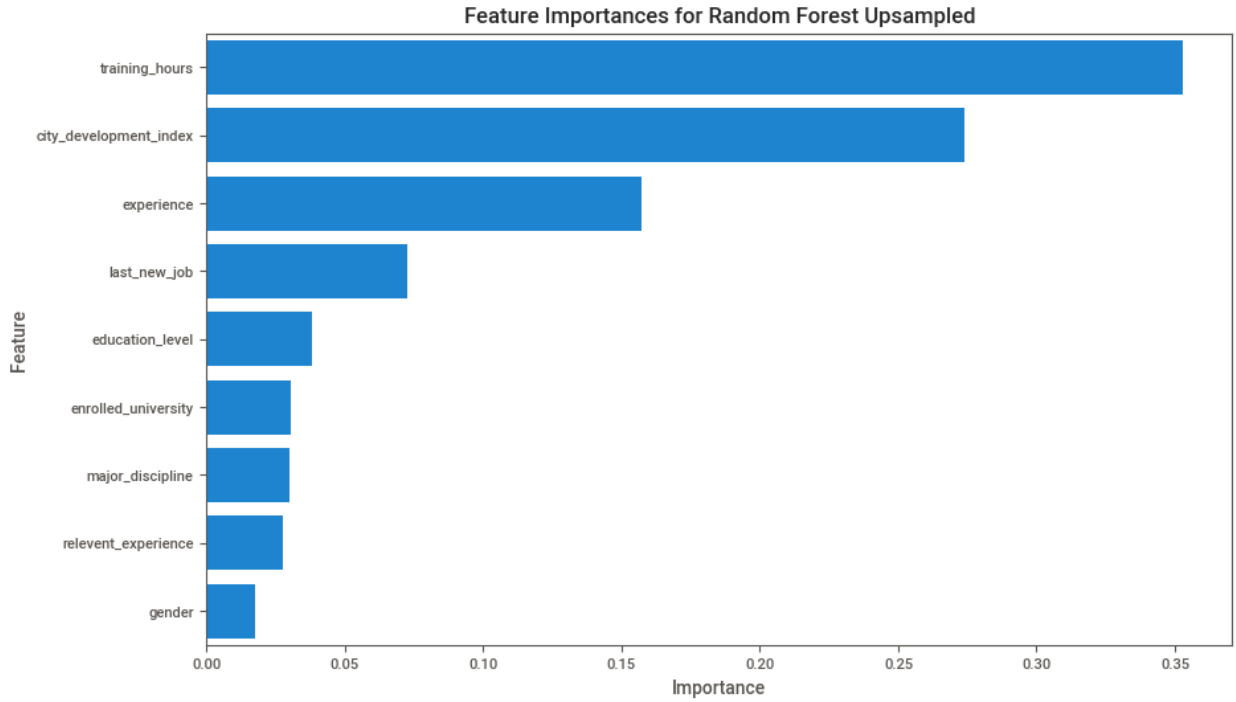

Figure 6. Extracted feature importances

6. Conclusions and Future Work

In conclusion, this study explores the integration of machine learning (ML) and feature engineering methods for predicting employee turnovers, specifically in the context of effective training planning and recruitment purposes in Human Resources Management (HRM). The model developed in this study achieves these results by surpassing the accuracy and other metrics achieved by previous studies using the same or similar dataset.

The data and features used in this study are not related to a specific job, making them applicable to different industries. Through comparative analysis and experimentation with various classification algorithms, it was found that a Random Forest (RF) trained on up-sampled data yielded the best performance, achieving an accuracy of 87.33% and showcasing promising results in terms of precision, recall, and F1-score. However, it is acknowledged that further improvements can be made**,** including further hyperparameter optimization with techniques such as grid search or random search, and different approaches to address the issue of imbalanced data like Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) could be explored.

Overall, this study contributes to the advancement of predictive analytics in HRMP, offering valuable insights and methodologies to enhance efficient planning. Future work of this study could include validation of extracted feature importances through real-life scenarios and case studies across different industries.

Appendices

Appendix A

Features and their descriptions in the dataset used in this study are shown in Table 4 below.

Table 4. Features and descriptions (HR Analytics, n.d.)

Feature	Description
enrollee_id	Unique ID for candidate
city	City code
city_development_index	Development index of the city (scaled)
gender	Gender of candidate
relevant_experience	Relevant experience of candidate
enrolled_university	Type of University course enrolled, if any
education_level	Education level of candidate
major_discipline	Education major discipline of candidate
experience	Candidate's total experience in years
company_size	Number of employees in current employer's company
company_type	Type of current employer
last_new_job	Difference in years between previous job and current job
training_hours	Training hours completed
target	0 – Not looking for job change, 1 – Looking for a job change

References

Al-Suraihi W.A., Samikon S.A., Al-Suraihi A.-H.A. and Ibrahim I., Employee Turnover: Causes, Importance and Retention Strategies, *European Journal of Business and Management Research,* vol. 6, no. 3, pp. 1-10, 2021.

ATD Releases 2016 State of the Industry Report, Available: https://www.td.org/insights/atd-releases-2016-state-of-the-industry-report, Accessed on May 21, 2024.

Bilogur A. ResidentMario/missingno [Python]. Available: https://github.com/ResidentMario/missingno 2024. (Original work published 2016), Accessed on May 21, 2024.

Cawley G.C. and Talbot N.L.C., On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, *The Journal of Machine Learning Research*, vol. 11, pp. 2079-2107, 2010.

Chen T. and Guestrin C., XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.

Conlon S., Why Do Data Scientists Want to Change Jobs: Using Machine Learning Techniques to Analyze Employees' Intentions in Switching Jobs, *INTERNATIONAL JOURNAL OF MANAGEMENT & INFORMATION TECHNOLOGY,* vol. 16, pp. 59-71, 2021.

Hastie T. Tibshirani R. and Friedman J.H., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer, 2009.

HR Analytics: Job Change of Data Scientists. Available: https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists. Accessed on April 6 2024.

IBM HR Analytics Employee Attrition & Performance, Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset, Accessed on April 6, 2024.

Jain R. and Nayyar A., Predicting Employee Attrition using XGBoost Machine Learning Approach, *Proceedings of the 2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, pp. 113-120, 2018.

Jeni L.A., Cohn J.F., and De La Torre F., Facing Imbalanced Data—Recommendations for the Use of Performance Metrics, *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 245-251, 2013.

Job Openings and Labor Turnover—February 2024, Available: https://www.bls.gov/news.release/pdf/jolts.pdf, Accessed on February 4, 2024.

Kaggle: Your Home for Data Science. Available: https://www.kaggle.com/. Accessed on April 6, 2024.

Kyalkond S.A., Manikanta Sanjay V., Manoj Athreya H., Aithal S.S., Rajashekar V. and Kushal B.H., Data Scientist Job Change Prediction Using Machine Learning Classification Techniques, P. Karuppusamy, F.P. García Márquez and T.N. Nguyen (Eds.), *Ubiquitous Intelligent Systems*, pp. 211-219, Springer Nature, 2022.

Lee B., Seo D., Lee J.-T., Lee A.-R., Jeon H.-N. and Han D.-U., Impact of work environment and work-related stress on turnover intention in physical therapists, *Journal of Physical Therapy Science*, vol. 28, no. 8, pp. 2358-2361, 2016.

Marvin G., Jackson M., and Alam Md.G.R., A Machine Learning Approach for Employee Retention Prediction, *Proceedings of the 2021 IEEE Region 10 Symposium (TENSYMP)*, pp. 1-8, 2021.

Mitchell T.R., Holtom B.C., Lee T.W. and Graske T., How to keep your best employees: Developing an effective retention policy / Executive commentary, *The Academy of Management Executive*, vol. 15, no. 4, pp. 96-109, 2001.

Ozkeser B., Impact of training on employee motivation in human resources management, *Procedia Computer Science*, vol. 158, pp. 802-810, 2019.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. and Duchesnay É., Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825-2830, 2011.

Turnover vs. Attrition: Definitions, Differences and Tips. Indeed Career Guide, Available: https://www.indeed.com/career-advice/career-development/turnover-vs-attrition, Accessed on February 3, 2023.

V Papineni S. L., A. Reddy M., Yarlagadda S., Yarlagadda S. and Akkineni H., An Extensive Analytical Approach on Human Resources using Random Forest Algorithm*, International Journal of Engineering Trends and Technology*, vol. 69, no. 5, pp. 119-127, 2021.

Mouzala K., Human resources operations management and its application to standards., B.S. Thesis, Sch. Of Eng., University of The Aegean, Chios, 2022.

Biographies

Eren Darici is an Industrial Engineer and current master's student in Industrial Engineering. He earned his bachelor's degree in industrial engineering in 2023 from Eskisehir Technical University, Türkiye. Following this, he commenced his Master of Science program in Industrial Engineering at Western Michigan University, USA, in the same year. His research interests include optimization, scheduling, and applied-machine learning.

Kira Hamelink received her bachelor's degree in computer science and master's degree in engineering management in 2021 and 2023, respectively, from Western Michigan University, USA. She is pursuing her Ph.D. in Industrial Engineering at the same institution while serving as a part-time computer science instructor at Kalamazoo Valley Community College, Michigan. Her research interests include optimization, applied operations research, artificial intelligence and machine learning, data analytics, and engineering education.

Ilgin Acar earned her Ph.D. in Industrial Engineering from Western Michigan University, USA. Currently, she is an Assistant Professor in the Department of Industrial and Entrepreneurial Engineering and Engineering Management at Western Michigan University, USA. Her research interests include linear optimization, the application of optimization and data analytics in both healthcare and logistics domains.