Research on Employee Turnover Prediction Based on Machine Learning

Authors: Zhao Sanglin, Deng Hao, planet, Bingkun Yuan, Zhao Sanglin

Date: 2024-08-19T00:00:00+00:00

Abstract

This study investigates the problem of predicting employee turnover rates based on machine learning algorithms, aiming to help enterprises reduce talent attrition and lower operating costs through data analysis. With the advent of the knowledge economy era, employee turnover has become a critical factor affecting enterprise development. High turnover rates not only increase recruitment and training costs but may also lead to technology resource outflow and customer attrition, severely impacting enterprise competitiveness and profitability. This paper first introduces the multifaceted negative impacts of employee turnover on enterprises, including increased financial costs, disrupted business progress, and declining team morale. Subsequently, it elaborates on the necessity and feasibility of utilizing machine learning techniques for employee turnover prediction against the backdrop of artificial intelligence and big data. Through a review of relevant literature, this study selects three machine learning algorithms—Naive Bayes, K-Nearest Neighbors (KNN), and Random Forest—to construct employee turnover prediction models. In the data preprocessing stage, an employee turnover dataset from the Kaggle website is utilized, containing 1,470 records covering multiple dimensions such as employee basic information, position information, compensation and benefits, and job satisfaction. Through data cleaning and feature selection, irrelevant variables and variables with high multicollinearity were removed, retaining key factors influencing employee turnover. In the model construction and evaluation phase, this study establishes three predictive models—Naive Bayes, KNN, and Random Forest—and conducts applicability analysis and parameter optimization. By comparing the predictive performance of each model, this study ultimately selects the Random Forest model as the optimal predictive model, which demonstrates excellent performance across metrics including accuracy, recall, and F1-score. Based on the results of the optimal predictive model, this study further analyzes key factors influencing employee turnover, including compensation and benefits, job satisfaction, and promotion opportunities. Addressing these factors, this paper

proposes corresponding employee retention and management recommendations, such as improving compensation and benefits levels, enhancing work environments, and strengthening employee career development, to help enterprises reduce employee turnover rates and improve team stability. Finally, this study summarizes the main findings and contributions, identifies research limitations, and suggests future research directions. This research not only provides enterprises with an effective employee turnover prediction tool but also offers new ideas and methods for research in the field of human resource management.

Full Text

Title and Authors

Title: Machine Learning-Based Prediction of Employee Turnover

Authors: Zhao Sanglin[1], Deng Hao[1], You Xing[1], Yuan Bingkun[2]

Affiliations: [1] Hunan University of Finance and Economics, 410205 [2] Central South University, 410083

Abstract

This study examines the prediction of employee turnover using machine learning algorithms, aiming to help enterprises reduce talent loss and lower operational costs through data analysis. As the knowledge economy era unfolds, employee turnover has become a critical factor affecting organizational development. High turnover rates not only increase recruitment and training expenses but may also lead to the loss of technical resources and clients, severely impacting corporate competitiveness and profitability.

The paper first introduces the multifaceted negative impacts of employee turnover on enterprises, including increased financial costs, disrupted business progress, and declining team morale. Subsequently, it elaborates on the necessity and feasibility of utilizing machine learning techniques for turnover prediction against the backdrop of artificial intelligence and big data. Through a literature review, this study selects three machine learning algorithms—Naive Bayes, K-Nearest Neighbors (KNN), and Random Forest—to construct employee turnover prediction models.

During the data preprocessing phase, an employee turnover dataset from Kaggle is employed, containing 1,470 records across multiple dimensions including employee demographics, position information, compensation and benefits, and job satisfaction. Through data cleaning and feature selection, irrelevant variables and those with high multicollinearity were removed, retaining only the key factors influencing employee turnover.

In the model construction and evaluation phase, three prediction models—Naive Bayes, KNN, and Random Forest—were developed with applicability analysis

and parameter optimization. By comparing predictive performance across models, the Random Forest model was ultimately selected as the optimal predictor, demonstrating superior performance across accuracy, recall, and F1-score metrics.

Based on the optimal model's results, this study further analyzes the key factors influencing employee turnover, including compensation and benefits, job satisfaction, and promotion opportunities. Targeted recommendations for talent retention and management are proposed, such as improving compensation packages, enhancing work environments, and strengthening employee career development, to help organizations reduce turnover rates and improve team stability.

Finally, the paper summarizes the main findings and contributions, identifies research limitations, and suggests future research directions. This study not only provides enterprises with an effective employee turnover prediction tool but also offers new perspectives and methodologies for research in human resource management.

Keywords: Naive Bayes; KNN; Random Forest; Employee Turnover; Artificial Intelligence

---

Table of Contents

Chapter 1: Introduction
1.1 Introduction
1.2 Research Background
1.2 Problem Statement
1.4 Research Objectives
1.5 Research Questions
1.6 Significance of the Study
1.7 Structure of the Report

Chapter 2: Model Introduction
2.1 Naive Bayes (NB)
2.2 K-Nearest Neighbors (KNN)
2.3 Random Forest

Chapter 3: Methodology
3.1 Model Building Process
3.2 Data Source
3.3 Data Preprocessing
3.4 Performance Metrics
3.5 Permutation Feature Importance Method
3.6 Specific Tools

Chapter 4: Results and Discussion
4.1 Exploratory Data Analysis Results

4.2 Machine Learning Performance
4.3 Permutation Feature Importance
4.4 Employee Turnover Dashboard

Chapter 5: Conclusion
5.1 Summary
5.2 Limitations
5.3 Future Work

References

Appendix: Python Programming for Model Building

Chapter 1: Introduction

1.1 Introduction

Talent is one of the most critical resources for enterprises and a key factor in their thriving development. Employee turnover can be simply understood as the voluntary departure of company staff, which is reactive from the company's perspective (Lee & Mitchell, 2017). Employee mobility has increasingly attracted organizational attention, so it is no surprise that it has drawn the interest of numerous scholars (Hom et al., 2017). Employee attrition represents more than just the loss of personnel; it generates many negative impacts on human resources, finances, operations, and other aspects (Barvey & Pathak, 2018). For instance, vacancies may hinder certain business operations and affect project timelines. Additionally, the time costs of recruitment, expenses for training new employees, and the time required for newcomers to adapt to their roles constitute significant hidden costs for companies. More importantly, departing employees may take away important clients or critical technologies, forcing the company to suffer substantial losses. In summary, high employee turnover rates have become one of the most crucial operational costs for enterprises. Against this backdrop, taking effective measures to reduce employee attrition and thereby lower operational costs has become one of the most important tasks for corporate HR departments.

With the accelerated development of artificial intelligence and next-generation information technologies, data has become a strategic asset for most organizations across various industries, including those related to business processes. All types of organizations benefit from adopting new technologies (Zhang & Lu, 2021), and the collection, management, and analysis of data bring numerous advantages in terms of efficiency and competitive edge. Indeed, analyzing large volumes of data can improve decision-making processes, achieve predetermined corporate goals, and enhance business competitiveness (Duan & Dwivedi, 2019). Within organizations, the adoption of AI influences corporate decision-making activities across several domains (Varian, 2018). For example, Chief Marketing Officers track detailed shopping patterns and preferences to predict and inform

consumer behavior. Chief Financial Officers use real-time, forward-looking integrated analytics to better understand different business lines. Now, Chief Human Resource Officers are beginning to deploy predictive talent models that can more effectively and rapidly identify, recruit, develop, and retain the right people (de Romrée et al., 2016).

In recent years, increasing attention has focused on human resources because employee quality and skills constitute a company's growth factor and genuine competitive advantage (Varian, 2018). In 21st-century information science and massive data analysis, AI—after being applied to sales and marketing—is now beginning to guide corporate decisions about employees, with the goal of establishing human resource management decisions on objective data analysis rather than subjective considerations (Gupta & Jain, 2018). People analytics will help organizations and their HR managers reduce attrition by changing the way talent is attracted and retained (Kemparaju, 2023).

This study applies machine algorithms to the human resources department, using machine learning to analyze and predict the primary factors influencing employee turnover rates. In this way, employees with resignation tendencies can be identified in advance, and their potential reasons for leaving can be analyzed, thereby helping companies and their HR managers improve employee job satisfaction through measures such as salary increases and improved work environments, ultimately reducing employee turnover.

1.2 Research Background

With the advent of the knowledge economy era, human resources have become an important factor driving human civilization progress and supporting economic and social development (Bai et al., 2020). Talent elements are becoming increasingly active, and their mobility is becoming more widespread globally. In China, according to the "2022 Turnover and Salary Adjustment Research Report" released by 51job, the overall employee turnover rate in 2021 was 18.8%, an increase of 4% compared to 2020. The voluntary turnover rate was 14.1%, up 6% from 2020. This indicates that employee attrition in enterprises is becoming increasingly severe (Zhou, 2022). Moreover, the report shows significant variation in turnover rates across industries, with catering, hospitality, and tourism experiencing the highest rate at 16.9%. On one hand, catering, hospitality, and tourism are labor-intensive industries with younger workforces and relatively lower stability. On the other hand, due to COVID-19, customer traffic dropped sharply, which had a tremendous impact on the operations of related companies, leading to significantly reduced employee performance income and stronger intentions to resign.

For enterprises, normal employee mobility can optimize internal personnel structure. However, many companies currently suffer from excessively high turnover rates, which severely hinders their growth and development. Evidence suggests that turnover is triggered by dissatisfaction with supervisors, company manage-

ment policies, corporate culture, work environment, promotion opportunities, and compensation (Griffeth & Gaertner, 2000). If companies can uncover the deep-rooted causes of employee attrition and implement improvement measures, they can effectively reduce employee turnover.

As the concept and application of big data permeate all industries, enterprises have gradually begun using machine learning algorithms to analyze their internal data. Machine learning is a collective term for a class of algorithms whose essence is pattern recognition. These algorithms are trained on large amounts of data to 挖掘出一些潜规则来进行预测 (Shitole & Priyadarshini, 2021). The machine learning process is like finding a function where sample data is input into the function and the desired result is output, but this function is very complex and difficult to express. Furthermore, machine learning does not only pursue good performance on the training set; its most important goal is to enable the learned function to achieve more accurate predictions when facing new samples. This capability is called generalization ability (Mahadevkar et al., 2022). Indeed, by using machine learning algorithms to establish enterprise employee turnover prediction models and identify employees with potential to leave, companies can prepare in advance and take corresponding measures, achieving a transformation from passively accepting employee turnover to actively intervening. When enterprises predict a high probability of employee resignation, they can communicate and confirm with employees adequately. If due to corporate factors, the enterprise should pay attention and can adjust relevant systems and policies to avoid further attrition. If due to other personal reasons, the company can also formulate recruitment plans as soon as possible to find talent for corresponding positions in a timely manner, avoiding adverse impacts caused by sudden employee departures. Additionally, using machine learning algorithms can identify important factors causing employee turnover, which companies can analyze to discuss how to reduce turnover rates.

Therefore, in this study, we will discuss and establish three machine learning models using Naive Bayes, KNN, and Random Forest respectively to classify and predict employee turnover. At the conclusion of this research, we will select an optimal prediction model based on relevant metrics and analyze the important factors causing employee attrition.

1.2 Problem Statement

Employee turnover is divided into internal and external mobility. Internal mobility refers to employees taking on new tasks within the same company, while external mobility refers to employees leaving the company to work for another organization. This study explores the concept of external turnover.

Key factors for business success include formulating correct business strategies, maintaining streamlined and efficient organizational structures, and possessing excellent management teams (Hani, 2021). As corporate management awareness improves, companies gradually realize that employees are core resources and that

talent plays a critical role in company development. Therefore, companies invest substantial time and money in introducing and cultivating talent. The exact cost of employee turnover is difficult to predict and varies based on factors such as role, tenure, and skills. When employees leave prematurely or too frequently, companies increase human capital costs, and work efficiency and quality decline. More seriously, it may lead to the outflow of technical resources or leakage of business operations. Excessive attrition may give outsiders the impression of internal management problems, resulting in deteriorating corporate reputation (Lee & Li, 2018).

The negative impact of turnover on enterprises is that the loss of some personnel negatively affects the emotions and work attitudes of other current employees (Mahadi et al., 2020). An individual's performance is directly influenced by their conditions and living environment. Armstrong (2020) considers turnover a disruptive phenomenon that brings costly consequences to organizations. When an employee voluntarily resigns, their decision may cause greater negative impact on the company than when an underperforming employee is dismissed. After an employee resigns, remaining employees may need to undertake more work tasks and responsibilities due to the company's failure to recruit suitable personnel in a timely manner, increasing their work pressure and generating dissatisfaction. Additionally, it may cause remaining employees to question the company's promotion and development opportunities, thereby generating ideas of following suit and resigning. This creates a vicious cycle that further exacerbates attrition. Furthermore, employee turnover reduces job satisfaction among employees in handling their work (Mahadi et al., 2020). When job satisfaction is low, employees lack work enthusiasm, affecting communication and cooperation between teams, leading to decreased work efficiency. When communicating with customers, employees may fail to meet customer needs well due to their negative attitude, bringing poor experiences to customers and thereby damaging company interests. Employees with higher job satisfaction tend to fulfill work responsibilities, actively communicate and assist colleagues, and have lower turnover intentions (Lin & Huang, 2021).

According to Ooi and Teoh (2021), employee turnover can also be classified based on employees' willingness to leave the company. This type is called voluntary and involuntary turnover. The voluntary turnover rate is actually the total number of employees leaving the organization divided by the total number of employees, typically measured over a one-year period (Hausknecht & Trevor, 2017). Voluntary turnover refers to situations where employees decide to leave the organization according to their own will, while involuntary turnover refers to situations where employees leave the organization reluctantly and are forced to resign due to various factors including poor work performance. People decide to voluntarily leave an organization for countless reasons, including low compensation, high work stress, poor performance evaluations, lack of job satisfaction, insufficient career development opportunities, lack of organizational commitment, lack of autonomy, and unfair labor practices (Lee et al., 1987).

Therefore, the main problem our research addresses is identifying the most influential factors among those mentioned above to reduce employee turnover rates and minimize the substantial losses that talent attrition brings to enterprises. To effectively solve this problem, this study establishes three machine learning prediction models for classification and prediction and selects an optimal prediction model based on final results. The data for these algorithm studies was obtained through the Kaggle website. This study will use Microsoft Power BI to construct a dashboard, transforming data into easily understandable and analyzable images and achieving data visualization.

1.4 Research Objectives

(1) Establish machine learning models for employee turnover prediction: Based on preprocessed available data, use multiple machine learning methods to build prediction models, conduct applicability analysis and parameter optimization, evaluate model effectiveness, and select the best model for predicting employee turnover rates.

(2) Reveal factors influencing employee turnover: Based on actual enterprise employee turnover data and redundancy among variables, use correlation analysis methods to examine relationships between variables and independence between each variable and the response variable. Through feature engineering and feature selection, identify the most important factors influencing employee turnover phenomena and use them as important early warning indicators for turnover.

(3) Employee retention and management recommendations: Provide recommendations based on influencing factors of employee attrition and model prediction results, implementing defensive measures before employees leave or at the first signs of turnover intention to reduce talent loss risks and minimize corporate losses.

1.5 Research Questions

The research questions for this study are as follows:

(1) Which model is best for predicting employee turnover after comparative evaluation?

(2) In the selected prediction model, which factors are most influential in predicting the presence of employee turnover?

(3) How can employee turnover rates be reduced based on the influencing factors and model prediction results?

1.6 Significance of the Study

Employee attrition can severely impact a company's organizational structure, human resource management, and loss of core technologies. In severe cases, it

may even lead to the company's demise. Employee turnover costs have become one of the company's major operational expenses. This phenomenon occurs frequently and has attracted increasing attention from corporate managers and social research scholars (Barvey & Pathak, 2018). Studying this topic can help us understand factors affecting employee turnover and enable rational allocation of human resources based on these factors. It helps corporate managers pay more attention to employees' work attitudes, psychological states, and turnover intentions, identifying employees' willingness to leave as early as possible. Spending sufficient time communicating with employees to understand problems in their actual work and reasons for leaving, using some incentive measures to encourage employees to continue working, and reducing corporate losses. This provides enterprises with theoretical methods and prediction models regarding employee management and career strategies, improves human resource management efficiency, enhances employee work stability, reduces talent loss risks, helps companies lower recruitment and management costs, and promotes sustained and healthy enterprise development.

Against the above research background, taking enterprise employee attrition as the research object and using machine learning algorithms to establish employee attrition prediction models has certain practical significance. This study selects the employee turnover dataset from Kaggle, constructs machine learning models to predict employee turnover, analyzes the important factors causing employee attrition, so that enterprise HR departments can track and manage employee work dynamics in a timely manner, and proposes solutions and recommendations for important influencing factors of losses based on turnover results. This study uses multiple machine learning algorithms for research and analysis, elaborating in detail the process from data analysis to method selection, model establishment, and model analysis and comparison, providing a feasible solution for employee turnover prediction with certain research significance.

Additionally, other companies can analyze employee attrition issues according to their actual situations, strengthen human resource management, and improve team stability.

1.7 Structure of the Report

Chapter 1, the introduction, discusses the significance of establishing machine learning models to predict employee turnover against the background of continuously rising enterprise employee turnover rates, and finally explains the research content and structure of this study.

Chapter 2 is the literature review and theoretical introduction. It first reviews literature on employee turnover and introduces the machine learning methods used in this paper, including Naive Bayes, KNN, and Random Forest.

Chapter 3 covers data preprocessing and exploratory analysis. It first introduces the data source, explains the meaning of variables in the dataset and preprocesses the data, then conducts exploratory data analysis, and finally uses

machine learning methods to model the analysis.

Chapter 4 presents the findings and analysis based on the prediction models.

Chapter 5 summarizes all findings, research limitations, and future plans.

Chapter 6 lists all references used in this study. Finally, the appendix of this research will be presented in Chapter 7.

Chapter 2: Model Introduction

2.1 Naive Bayes (NB)

The Naive Bayes classifier is a typical probabilistic machine learning method used for classification applications due to its simplicity and effectiveness. It is based on Bayes' theorem and makes a "naive" assumption that features are independent given the class label (Kim et al., 2020). It assumes that all variables contribute to classification and are interrelated. This assumption is known as class-conditional independence (Friedman et al., 1997). This is an unrealistic assumption for most datasets, but it leads to a simple prediction framework that yields surprisingly good results in many practical situations (Hetal, 2012). Furthermore, it is very easy to construct and does not require any complex iterative parameter estimation schemes. This means it can be easily applied to huge datasets. It is easy to interpret, so users unfamiliar with classifier technology can understand why it makes classifications (Wu et al., 2008).

2.2 K-Nearest Neighbors (KNN)

The KNN algorithm is an instance-based learning method that can be used to solve classification, regression, and search problems. It classifies objects based on the closest training examples in the feature space (Boateng et al., 2020). In the KNN algorithm, the classification of a new test feature vector is determined by the classes of its k nearest neighbors. The algorithm assumes that similar items exist nearby. In other words, similar items are close to each other. The KNN algorithm depends on whether this assumption is true enough to be useful.

KNN is effective for large numbers of training samples. However, for this algorithm, the value of parameter k (the number of nearest neighbors) and the type of distance to use must be determined. Computation time can be long because distances from each query instance to all training samples need to be calculated, and computation time slows down significantly as the number of instances or predictive independent variables increases (Statnikov et al., 2008). The Euclidean distance metric is used to implement the KNN algorithm for locating nearest neighbors (Pan et al., 2004).

The Euclidean distance metric d(x, y) between two points x and y is calculated using Equation (2.1):

$$d(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

where N is the number of features.

2.3 Random Forest

Random Forest classification is a popular machine learning method used to develop prediction models in many research settings. Typically in predictive modeling, the goal is to reduce the number of variables needed for prediction to lessen the burden of data collection and improve efficiency (Speiser et al., 2019). In the Random Forest setting, many classification and regression trees are constructed using randomly selected training datasets and random subsets of predictor variables for modeling outcomes. Results from each tree are aggregated to give predictions for each observation. Therefore, compared to single decision tree models, Random Forest generally provides higher accuracy while maintaining some beneficial qualities of tree models (e.g., the ability to interpret relationships between predictors and outcomes) (Speiser, Durkalski, & Lee, 2015).

The Random Forest flowchart is shown below:

[Figure 2: see original paper]-1 Random Forest Algorithm Flowchart

---

Chapter 3: Methodology

In this study, the process of building employee turnover prediction models consists of several components, including data preprocessing, machine learning modeling, and data visualization.

3.2 Data Source

The data in this paper comes from the Kaggle website, linked to https://www.kaggle.com/datasets/thedesvastator/employee-attraction-and-factors. The dataset contains a total of 1,470 data rows, including one target variable and 33 independent variables, with no missing values and high completeness.

Among these, attrition rate is the target variable indicating employee turnover, where “Yes” means the employee has left and “No” means the employee has not left. Independent variables include employees’ basic identity information, position information, compensation and benefits, job satisfaction, and related variables. EmployeeNumber is an irrelevant variable, as all samples have identical values for Over18, StandardHours, and EmployeeNumber, while the significance of DailyRate, HourlyRate, and MonthlyRate is unclear. Therefore, these variables are removed. Secondly, due to the large number of variables and

high multicollinearity between some variables, such as JobLevel and MonthlyIncome, the following variables were removed to facilitate analysis and observation: DistanceFromHome, EducationField, JobInvolvement, JobLevel, NumCompaniesWorked, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager. The dataset after removing these variables is shown in Table .1.

Table 3.1 Description of Variables in the Dataset

Dataset Variable	Description Meaning
EnvironmentSatisfaction	1-5 (higher value = greater satisfaction)
Age	18-60
BusinessTravel	Travel_{Rarely}/R&D/Human Resources/Sales
Education	1-5 (higher value = higher education level)
Gender	Male/Female
JobRole	Sales Executive/Research Scientist/Laboratory Technician/Manufacturing Director/Medical Representative/Sales Representative/Research Director
JobSatisfaction	1-4 (higher value = greater satisfaction)
MaritalStatus	Single/Married/Divorced
MonthlyIncome	1009-19999
OverTime	Yes/No
PercentSalaryHike	11-25
RelationshipSatisfaction	1-4 (higher value = greater satisfaction)
StockOptionLevel	Integer
TotalWorkingYears	Integer
WorkLifeBalance	1-4 (higher value = better work-life balance)

3.3 Data Preprocessing

Data preprocessing refers to a series of processes including cleaning, integrating, transforming, and discretizing raw data, which is considered an essential step. Its purpose is to reduce data complexity and facilitate better data analysis and modeling (Ramírez-Gallego et al., 2017). This study employs several data preprocessing techniques to prepare data for machine learning algorithms, as described in detail in Chapter 2.

3.3.1 Exploratory Data Analysis Exploratory data analysis typically employs data visualization methods to analyze complex datasets and summarize

their main characteristics. In this study, descriptive analysis and heatmap correlation matrices were used to explore the data before building machine learning models. Conducting descriptive statistical analysis on the dataset before modeling can reveal basic information about employees in the dataset and provide certain guidance for feature selection and classification prediction models. Descriptive analysis is a method used to objectively describe the nature and magnitude of variable characteristics (Kemp et al., 2018). This analysis can be implemented using the describe function in the Pandas library available in the Python programming language. Heatmap matrices can be implemented using the corr function in the Pandas library.

3.3.2 Feature Selection Feature selection eliminates irrelevant or redundant features, thereby reducing the number of features, finding the optimal feature subset, improving model accuracy, and reducing runtime. In this study, EmployeeNumber is an irrelevant variable, as all samples have identical values for Over18, EmployeeNumber, and StandardHours, while the significance of DailyRate, HourlyRate, and MonthlyRate is unclear. Moreover, due to the large number of variables and high multicollinearity between some variables, such as JobLevel and MonthlyIncome, the following variables were removed to facilitate analysis and observation: DistanceFromHome, EducationField, JobInvolvement, JobLevel, NumCompaniesWorked, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager. All the above variables were removed from the dataset using the drop function in the Pandas library.

3.3.3 Data Normalization Data normalization is a method for standardizing different variables or data feature directions, also known as feature scaling. Normalizing data allows features of different dimensions to be compared together, which can greatly improve model accuracy. Data normalization processing is a typical approach to data standardization, including min-max normalization and z-score normalization. The processed data has a mean of 0 and a standard deviation of 1. The transformation formula is as follows:

$$z = \frac{x - \mu}{\sigma} \quad (3.1)$$

where is the mean of the original data and σ is the standard deviation of the original data. Z-score standardization is relatively simple to calculate and is suitable for situations where the maximum and minimum values of the dataset are unknown, there are outliers in the data, or the data distribution is highly discrete.

Min-max normalization is a method for linear transformation of original data, where the processed result values are mapped to [0, 1]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.2)$$

Min-max normalization converts both positive and inverse indicators into positive indicators with consistent direction, facilitating comparison. It is suitable for simply transforming data mapping to a certain interval for comparison and observing data distribution.

In this study, because some numerical variables have large differences in value ranges, they can easily cause incomparability between variables. Therefore, this study adopted Z-score standardization before model building. In this method, each numerical input variable will be individually assigned a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation (Brownlee, 2020). This technique can be implemented using the StandardScaler function in the Scikit-Learn library provided by Python (Brownlee, 2020).

3.3.4 Transforming Categorical Variables Feature encoding is the corresponding encoding of various special feature values in categorical variables and is also a quantification process. The reason data needs to be encoded is that numerical data can be recognized and calculated by machine models. It mainly includes label encoding and one-hot encoding. Label encoding, as the name suggests, uses label encoding to convert unordered and ordered categorical data into numerical data. Label encoding solves the problem of categorical encoding, and encoding values can be customized as needed, with the values themselves having no specific meaning. In this study, I will use label encoding for 7 variables.

3.3.5 Data Sampling The most commonly used method for handling imbalanced data is oversampling. Its basic principle is to directly increase the number of minority classes, improve the accuracy of minority class classification, thereby changing the data distribution and reducing this imbalance.

(1) The most convenient method for oversampling is to directly duplicate sample data from the minority class, but this method simply repeats data and increases quantity without changing any information, thus potentially causing overfitting. For this situation, generally adding newly synthesized data or random Gaussian noise is used to reduce the aforementioned problem, such as the SMOTE algorithm.

(2) **SMOTE Algorithm**

SMOTE oversampling technology is a new technique different from traditional oversampling algorithms. Unlike oversampling that duplicates data, this method creates new data through synthesis, increasing the amount of data in minority samples, changing the data distribution to some extent while avoiding excessive samples. From the characteristics of new samples synthesized by SMOTE

technology, it can solve the overfitting problem that traditional oversampling is prone to to a certain extent. Its basic principle is shown in Figure [Figure 3: see original paper].1.

Figure 3.1 Basic Principle of SMOTE Algorithm

In simple terms, SMOTE technology randomly inserts a new sample data point between each minority class data point and its neighbor, repeating until the data volume of each class reaches a balanced state, then training the new dataset. Its advantages are: helping to break the overfitting problem caused by simple duplication and the problem of not increasing information volume in minority classes, thereby improving the learning efficiency of classifiers.

Since the dataset used in this paper only has 1,470 samples, and minority samples (attrition samples) only account for 16.12%, which is extremely imbalanced data, this paper will use the SMOTE algorithm to artificially generate new departing employee numbers to make the imbalanced samples become balanced, thereby achieving the result of expanding sample quantity without losing sample information.

3.3.6 Data Splitting Data splitting is a process in data science or machine learning where data is divided into two or more subsets. Typically, through two-part splitting, one part is used for evaluation or testing data, and the other part is used for training the model. Table 3.2 shows the dataset sizes used for training and testing in this study.

Table 3.2 Training and Testing Dataset Sizes

Dataset Attribute	Partition Percentage
Training Set	80%
Testing Set	20%

3.4 Performance Metrics

The main model evaluation metrics in classification problems include confusion matrix, accuracy, precision, recall, F1-score, etc. The confusion matrix is shown in Figure 3.2, where TP (True Positive) represents true positive cases; FN (False Negative) represents false negative cases; FP (False Positive) represents false positive cases; TN (True Negative) represents true negative cases. This matrix is commonly used to describe generalization errors of prediction results in binary classification problems.

The formulas for calculating accuracy score, precision, recall, and F1-score are shown in Equations (3.3), (3.4), (3.5), and (3.6) respectively.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.5)$$

$$\text{F1-Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3.6)$$

Figure 3.2 Binary Confusion Matrix

3.5 Permutation Feature Importance Method

After training a model, in addition to being interested in the model's predictions, we often want to know which features are more important in the model and which features have the greatest impact on prediction results. Permutation importance is a method for measuring feature importance. In this study, the permutation feature importance method will be used to calculate the importance of variables in predicting employee turnover, based on the proportional reduction in model scores when variables are randomly shuffled (Assar et al., 2021).

3.6 Specific Tools

This study uses Jupyter Notebook from Anaconda Navigator and Microsoft Power BI for modeling and data visualization respectively. Jupyter Notebook is an open-source web application that allows users to create and share documents containing live code, equations, visualizations, and narrative text. This notebook supports the Python programming language, which is necessary for building prediction models in this study. Python is a high-level interpreted general-purpose dynamic programming language that emphasizes code readability (Özgür et al., 2017). This language consists of large and comprehensive standard libraries such as Pandas, Scikit-Learn, Seaborn, and Matplotlib for model building (Jasim et al., 2019). For data visualization, Microsoft Power BI is an easy-to-use tool that can analyze and present data in an interactive and highly visual manner (Aspin, 2016).

Chapter 4: Results and Discussion

4.1 Exploratory Data Analysis Results

In this study, descriptive statistics were generated to observe and explore the characteristics of all variables, including numerical and categorical variables present in the dataset. These characteristics include count, unique values, top values, frequency, mean, standard deviation, minimum and maximum values,

and 25%/50%/75% percentiles. The results of the descriptive statistics are shown in Table .1. The results in Table 4.1 clearly indicate that no null values exist in the dataset (no missing values), as the count for all variables matches the total number of observations in the dataset (1,470), demonstrating that all variables consist of values or labels in each observation describing respective variables.

Additionally, a heatmap correlation matrix was generated to explore monotonic associations between two variables in this study, with results shown in Figure [Figure 4: see original paper].1. According to Figure 4.1, all values showing relationships between variables range from -0.66 to 0.77, indicating direct relationships. The value between variables "MonthlyIncome" and "TotalWorkingYears" is 0.77, which is the highest compared to other relationships in the correlation matrix. This indicates that the direct relationship between these two variables is the strongest compared to other relationships.

Table 4.1: Dataset Descriptive Statistics

Figure [Figure 4: see original paper].2 Heatmap Correlation Matrix

4.2 Machine Learning Performance

The performance of machine learning algorithms in predicting employee turnover presence was evaluated through confusion matrices, accuracy scores, precision, recall, and F1-score. By comparing these performance metrics, this study selected a golden prediction model that can accurately predict employee turnover. The confusion matrix results and performance metrics of the prediction models are shown in the following sections.

4.2.1 Confusion Matrix Comparison The confusion matrix results among KNN, Naive Bayes, and Random Forest models are shown in Figure 4.2. Based on these results, KNN accurately classified 159 employees who did not leave the company and 36 employees who did leave. However, the algorithm misclassified 13 employees who actually left but were classified as not leaving. Additionally, the algorithm misclassified 86 employees who actually did not leave but were classified as leaving.

Naive Bayes accurately classified 179 employees who did not leave and 32 employees who left. However, the algorithm misclassified 17 employees who actually did not leave but were classified as leaving. Additionally, the algorithm misclassified 66 employees who actually left but were classified as not leaving.

Random Forest accurately classified 235 employees who did not leave and 16 employees who left. However, the algorithm misclassified 33 employees who actually did not leave but were classified as leaving. Additionally, the algorithm misclassified 10 employees who actually left but were classified as not leaving.

Through confusion matrix comparison of these three algorithms, Random Forest outperforms Naive Bayes and KNN in predicting employee turnover because

the sum of true positives and true negatives for Random Forest is higher than for Naive Bayes and KNN. Therefore, Random Forest can accurately classify employees based on the provided independent variables.

Figure 4.2: Confusion Matrices for KNN, Naive Bayes, and Random Forest Models

4.2.2 Performance Metrics Comparison Table 4.2 shows the overall performance results of the KNN, Naive Bayes, and Random Forest models in terms of accuracy scores, precision, recall, and F1-score. According to the accuracy scores in Table 4.2, Random Forest outperforms Naive Bayes and KNN because the algorithm can predict the maximum number of true positives and true negatives compared to Naive Bayes and KNN.

Table 4.2: Accuracy Scores, Precision, Recall, and F1-Score Results for Models

Model	Accuracy Score	Precision	Recall	F1-Score
KNN	...	...	...	...
Naive Bayes	...	...	...	...
Random Forest	...	...	...	...

Furthermore, Random Forest also outperforms Naive Bayes and KNN in terms of precision and F1-score. This is because the precision and F1 values for Random Forest are 0.96 and 0.92 respectively, which are higher than those for Naive Bayes and KNN. Therefore, when these two metrics are a concern in predicting employee turnover, Random Forest can minimize false positives and false negatives.

4.2.3 Selection of Optimal Model Based on the performance metrics results, Random Forest was selected as the best model in this study. This is because all performance metrics for Random Forest are high, far exceeding those of Naive Bayes and KNN. Therefore, these results indicate that Random Forest can accurately classify employees into corresponding categories based on the provided independent variables when predicting employee turnover. Additionally, the algorithm can minimize the costs or impacts caused by false positives and false negatives because it can reduce misclassification problems in the employee turnover prediction process.

4.3 Permutation Feature Importance

In this study, the average scores derived from permutation feature importance for each variable in the golden model (Random Forest) are shown in descending order in Figure 4.6. According to Figure 4.6, the top three most important variables for predicting employee turnover are “OverTime,” “Age,” and “MonthlyIncome,” as the importance scores for these variables are 0.068, 0.037, and 0.033

respectively, which are higher than other variables. Among these four variables, "OverTime" is the most important variable in the employee turnover prediction process because this variable has the highest average importance score compared to other variables.

Figure 4.3: Importance Score Results for Each Variable

4.4 Employee Turnover Dashboard

A dashboard is a commonly used tool for data visualization. Figure 4.4 shows the dashboard constructed for this study to visually present all variables related to employee turnover graphically.

Figure 4.4: Employee Turnover Dashboard

According to the dashboard in Figure 4.1, the employee sample used in this study includes 83.86% of employees (1,233 people) who did not leave the company, while the remaining 16.12% of employees (237 people) decided to leave. This dashboard shows some important information regarding overtime and age. The first insight is that employees who work overtime tend to decide to leave the company. Secondly, compared to people aged 40-60, those aged 20-40 are more inclined to leave the company.

Chapter 5: Conclusion

5.1 Summary

In summary, human resource analytics is a technique used in this study to predict employee turnover presence by establishing machine learning algorithms. This study utilizes a human resource analytics dataset consisting of 1,470 rows of data and 17 variables as a sample to build prediction models. Additionally, descriptive analysis and heatmaps were used to explore patterns and relationships between variables.

Before the model building process, data preprocessing was performed to transform the data into a form suitable for the modeling process. As a result of this process, the following variables have been removed: "EmployeeNumber," "Over18," "StandardHours," "EmployeeNumber," "DailyRate," "HourlyRate," "MonthlyRate," "DistanceFromHome," "JobInvolvement," "NumCompaniesWorked," "PercentSalaryHike," "PerformanceRating," "TrainingTimesLastYear," "YearsAtCompany," "YearsSinceLastPromotion," "JobLevel," "YearsWithCurrManager," "EducationField," and "YearsInCurrentRole." All categorical data has been converted to numerical data, but this data has still been scaled to a standard range. Additionally, the dataset was split into training and testing datasets according to the Pareto principle, with 80% used for the training dataset and the remaining 20% for the testing dataset. This study used three machine learning algorithms—K-Nearest Neighbors,

Naive Bayes, and Random Forest—to predict employee turnover presence. The research results achieved all research objectives and answered all research questions posed by this study.

For research question (1), "Based on variables in the dataset, which model is most suitable for predicting employee turnover presence?" this study concludes that the Random Forest model is the best model for predicting employee turnover. This is because all performance metrics (accuracy score, precision, and F1-score) derived from Random Forest far exceed those of Naive Bayes and KNN. Therefore, the results indicate that the Random Forest model can accurately predict employee turnover presence based on the provided variables.

For research question (2), "In the selected prediction model, which factors are most influential in predicting employee turnover presence?" this study demonstrates that the top three variables—"OverTime," "Age," and "MonthlyIncome"—are the most important variables for predicting employee turnover because these variables scored higher than other variables in the permutation feature importance process. Additionally, the results show that "OverTime" has the greatest impact on employee turnover prediction because this variable has the highest score compared to other variables in this study.

For research question (3), "Based on the influencing factors and model prediction results, what can be done to reduce employee turnover?" since overtime, age, and salary are the three most important factors affecting employee turnover, improvements should be made in these three areas. Three recommendations are proposed regarding overtime. First, workflows should be optimized to ensure tasks are appropriately distributed to reduce unnecessary overtime. Second, managers should allocate tasks reasonably to ensure team members' workloads are balanced and avoid excessive overtime for individual members. Finally, companies can implement flexible work systems, such as flexible working hours or remote work, to help employees better balance work and life and reduce the need for overtime.

Regarding age, first, personalized development opportunities and training programs should be provided for employees across different age groups to meet their career development needs and enhance their career development motivation. Second, companies should understand the needs and concerns of employees in different age groups and provide age-appropriate benefits and treatment, including health insurance, training programs, and flexible work arrangements.

For compensation, first, market research should be conducted to ensure salary levels are competitive and can attract and retain talent. Second, provide a performance evaluation and reward system that motivates employees in a fair and transparent manner, making them feel their work is valued and rewarded. Finally, salary structures can be evaluated regularly and adjusted based on employee contributions and growth.

5.2 Limitations

One limitation encountered in this study is that integer encoding was used as a technique to convert nominal data for "Department" into numerical data. Generally, nominal data should be handled using one-hot encoding, where each label is encoded into a Boolean value and stored in a newly added column to provide equal importance for each label and reduce the risk of model misdirection (Brownlee, 2020; Wang & Zhi, 2021). However, this method caused a problem in this study because the number of columns increased according to the number of unique observations for the "Department" variable, which resulted in scores for each variable that could not clearly indicate the importance score for the "Department" variable. Therefore, in this study, this method became a mapping technique that mapped each nominal data to a specific value.

The second limitation is that the number of independent variables in this dataset was too large, and deleting some of them before conducting the research led to minor errors between analysis and reality.

5.3 Future Work

In future research, primary sources of employee data will be used to build machine learning algorithms to predict employee turnover. This is because original data is more reliable and accurate than secondary data, which can be obtained by distributing questionnaires or interviewing the target population of the study. Additionally, the scope of the dataset needs to be expanded to include small and medium-sized enterprises. Furthermore, future research can explore other machine learning algorithms, such as Decision Trees, Support Vector Machines, and Logistic Regression, to predict employee turnover based on collected data.

References

[1] Ongori, H. (2007). A review of the literature on employee turnover. *African Journal of Business Management*, 1(2).

[2] Armstrong, M., & Taylor, S. (2020). *Armstrong's handbook of human resource management practice* (15th ed.). Kogan Page. https://books.google.at/books?id=g7zEDwAAQBAJ

[3] Barvey, A., Kapila, J., & Pathak, K. (2018). Proactive intervention to downtrend employee attrition using artificial intelligence techniques. *arXiv* [stat.ML]. http://arxiv.org/abs/1807.04081

[4] Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing*, 08(04), 341–357. https://doi.org/10.4236/jdaip.2020.84020

[5] Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective

naïve Bayes algorithm. *Knowledge-Based Systems*, 192(105361), 105361. https://doi.org/10.1016/j.knosys.2019.105361

[6] de Romrée, H., Fecheyr-Lippens, B., & Schaninger, B. (2016, July 22). People analytics reveals three things HR may be getting wrong. *McKinsey.com*; McKinsey & Company. https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/people-analytics-reveals-three-things-hr-may-be-getting-wrong

[7] Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. *International Journal of Information Management*, 48, 63-71.

[8] Angelo, P. (2008). Machine learning. *WO2008053161* [P]. 2008-05-08.

[9] Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management*, 26(3), 463-488.

[10] Hani, E. H. (2021). The effect of key business success factors on start-up performance. *Network Intelligence Studies*, IX(18 (2/2021)), 117–129. https://doaj.org/article/978305af399045479625c88f7f7798fe

[11] Bhavsar, H., & Ganatra, A. (2012). A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(4), 74-81.

[12] Hom, P. W., Lee, T. W., Shaw, J. D., et al. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology*, 102(3), 530-545.

[13] Kaur, B. (n.d.). Antecedents of turnover intentions: A literature review. Ripublication.com. Retrieved June 20, 2023, from https://www.ripublication.com/gjmbs_{spl}/gjmbsv3n10_{

[14] Kemparaju, N. (2023). Employee retention analysis using machine learning. *Zenodo*. https://doi.org/10.5281/ZENODO.7697798

[15] Kim, H.-C., Park, J.-H., Kim, D.-W., & Lee, J. (2020). Multilabel naïve Bayes classification considering label dependence. *Pattern Recognition Letters*, 136, 279–285. https://doi.org/10.1016/j.patrec.2020.06.021

[16] Lee, T. W., & Mitchell, T. R. (2017). Control turnover by understanding its causes. In *The Blackwell Handbook of Principles of Organizational Behaviour* (pp. 93–107). Blackwell Publishing Ltd.

[17] Lee, T. W., Hom, P. W., Eberly, M., et al. (2018). Managing employee retention and turnover with 21st century ideas. *Organizational Dynamics*, 47(2), 88-98.

[18] Lee, T. W., Mowday, R. T. (1987). Voluntarily leaving an organization: An empirical investigation of Steers and Mowday's model of turnover. *The Academy*

of Management Journal, 30(4), 721-743.

Appendix: Python Programming for Model Building

The prediction models used in this study were built using the Python programming language. The complete input code and corresponding output are shown below.

Data Preprocessing

```
# Code implementation will be presented here with corresponding outputs
```

Note: Figure translations are in progress. See original paper for figures.