Employee Attrition: A Supervised Learning Approach to Predict Whether Employees Will Leave Their Current Company

Amirreza Dashti Genave

March 2025

Abstract

Many employees leave their current company, while managers often perceive this as an impulsive decision. However, numerous factors influence an employee's attrition and performance in their role. In this study, we aim to identify the most important factors affecting an employee's performance and predict whether they are at risk of attrition. We explore and visualize the attrition rate in relation to the available factors provided by the IBM HR Analytics dataset [1]. to gain a better understanding of the data. We then apply a supervised learning algorithm (SVM) to predict whether an employee is likely to leave their current company. Our SVM model achieved an accuracy of 78%, demonstrating its effectiveness in predicting employee attrition.

1 Introduction

What is employee attrition? This concept refers to the voluntary departure of employees from an organization due to various reasons, such as distance from the office, working atmosphere, working hours, or even personal life events. Aside from the reasons, employee attrition can significantly impact a company by reducing performance, decreasing productivity, and increasing recruitment costs. Managers attempt to mitigate attrition by addressing the key factors that contribute to it. Being able to predict attrition allows organizations to make informed decisions to retain valuable employees.

Although attrition may seem like an impulsive decision, it is influenced by several factors that can be predicted. The goal is to develop a predictive model that provides relatively accurate predictions regarding employee attrition.

The main aim of this work is to use statistical analysis to identify the factors most strongly correlated with attrition and to apply machine learning techniques to predict employee turnover. Among various supervised learning algorithms, we selected the Support Vector Machine (SVM) algorithm due to its ability to capture the nuanced effects of the factors involved in attrition and make efficient predictions. The details of the chosen model will be discussed in the corresponding section.

The IBM HR Analytics Employee Attrition & Performance dataset is a fictional dataset created by IBM scientists, containing more than 30 features, both numerical and categorical (e.g., age, education, distance from work, etc.). This well-structured dataset is ideal for performing prediction tasks and is labeled with "Yes" or "No" values, making it suitable for binary classification.

The main research questions addressed in this study are: What factors are most relevant to employee attrition, and can an SVM algorithm accurately predict attrition based on this dataset?

In this study, we first examine the data to gain meaningful insights and better understand the distributions and rates of attrition. Next, we perform a data preprocessing phase to retain the most important features. Then, we create the supervised learning model and conduct the training phase. Following this, we evaluate the performance of the model and interpret the results. Finally, we conclude the study and propose directions for future work.

2 Dataset Description

IBM HR Analytics Employee Attrition & Performance, is a fictional dataset created by IBM scientists containing 1470 samples and 35 columns including the label "Attrition". The columns include both numerical and categorical data which requires the use of encoding methods during preprocessing. Some key features may include:

- Age

- BusinessTravel
- DistanceFromHome
- Education
- JobRole

The target label has two unique values “Yes” or “No” indicating that this is a binary classification task.

Are the target labels evenly distributed? To answer this question, we must look at the plot that shows the class distribution. Figure 1 shows a significant imbalance: approximately 83% of the samples belong to the ”No” class, while only 16% correspond to the ”Yes” class. Since this is a challenge for the performance of our model, we will address this issue in the preprocessing phase.

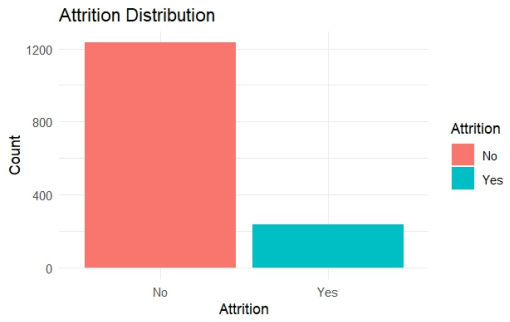

Figure 1: Class distribution

To find out more about the attrition rate with respect to other features such as job role, department, gender, stock option level, we used histogram charts to better understand these rates. These charts help us understand how the data is distributed across various attributes. Figures 2, 3, 4, 5, and 6 illustrate these distributions.

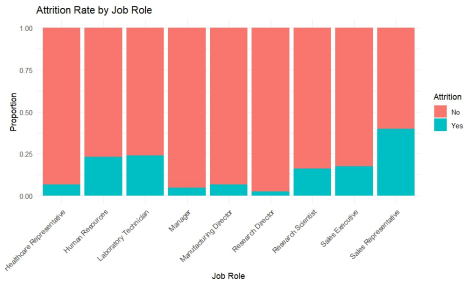

Figure 2: Attrition rate by job role

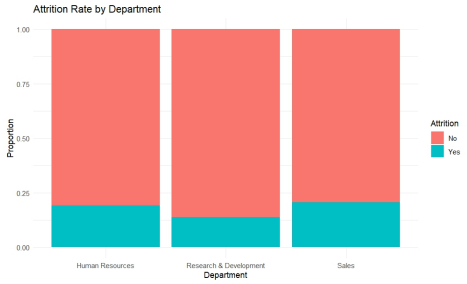

Figure 3: Attrition rate by department

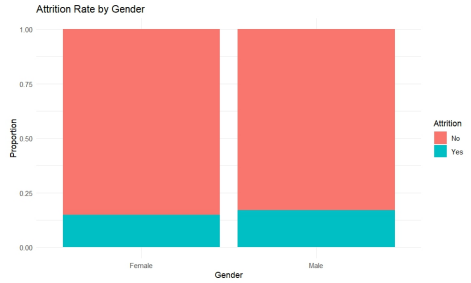

Figure 4: Attrition rate by gender

Figure 5: Attrition rate by stock level option

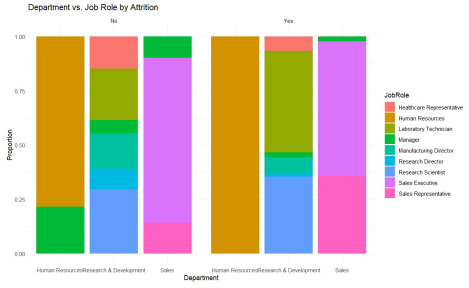

Figure 6: Department VS. job role by attrition

Additionally, the distribution of years at company and age with respect to attrition are provided in figures 7 and 8 respectively. Figure 7 highlights that there is a significant concentration of attrition within the first few years of employment. From figure 8, we can understand that younger employees

(particularly those in their 20s and early 30s) are more likely to leave.

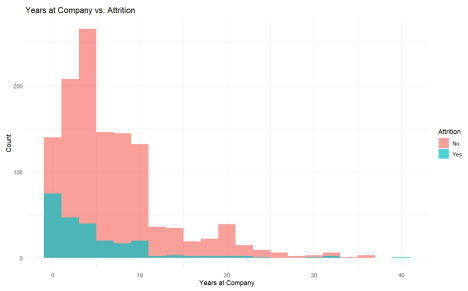

Figure 7: Years at Company vs. Attrition

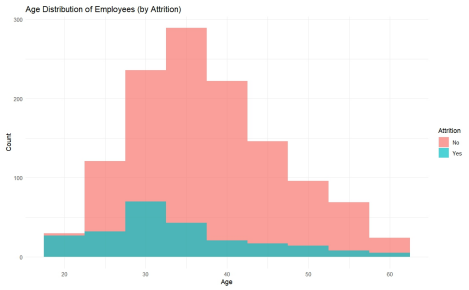

Figure 8: Age Distribution of Employees (by Attrition)

Figure 9 shows that employees who did not leave (Attrition = No) tend to have a higher median monthly income compared to those who left (Attrition = Yes). Lower-income employees are more likely to leave the company.

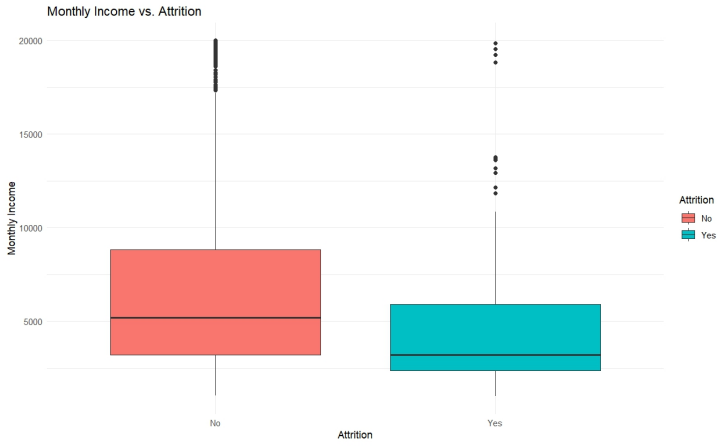

Figure 9: Monthly Income vs. Attrition

Given these observations, the next step is to prepare the data for modeling by addressing class imbalance and handling both numerical and categorical features in the preprocessing phase.

3 Data Preprocessing

IBM attrition dataset needs several key steps to be prepared to be ready for being fed into the model.

1. **Counting missing values and improper columns:** This dataset has zero missing values. Therefore, no handling of missing data is needed. However, some columns with constant values or columns having zero standard deviation must be deleted. Because they don't contribute to the learning phase, and they can produce issues for the algorithm.

2. **Converting the label values to numeric values:** This task can be done by mapping the "Yes" values to 1 and "No" values to 0. In this way, the learning algorithm can interpret and use this column as the target.

3. **Converting categorical features into Factors:** : In R programming language, factor is an approach to handle categorical columns. A factor stores the unique values of categorical columns and uses ordinal numbers for each unique value. This approach ensures storage efficiency, accuracy, and correct interpretation of the data.

4. **Feature selection:**Feature selection is adopted to reduce the redundant or irrelevant features, speed up the training phase, and simplify the model. To achieve this, we calculate the correlation matrix of the columns. Figure 10, demonstrates the correlation matrix, which shows that a certain number of columns are highly correlated to the target column. Furthermore, highly correlated features may increase the model error by introducing multicollinearity. Therefore, we introduce a threshold equal to 0.75 to identify highly correlated pairs of features and retain the feature that is more correlated to the target column.

5. **Generate dummy variables using one-hot encoding:** This step has been taken to convert categorical variables into a numerical format that machine learning models can understand. This approach ensures that no ordinal relationship is introduced between categories.

6. **Handling class imbalance:** It is undeniable that the labels are extremely imbalanced. A good approach to mitigate this issue is to

use SMOTE (Synthetic Minority Oversampling Technique) algorithm. SMOTE uses K-nearest neighbor (KNN) to identify k neighbors of a sample from the minority class, and creates a new sample. Figure 11 shows that the class distribution has become even after applying SMOTE algorithm on the dataset.

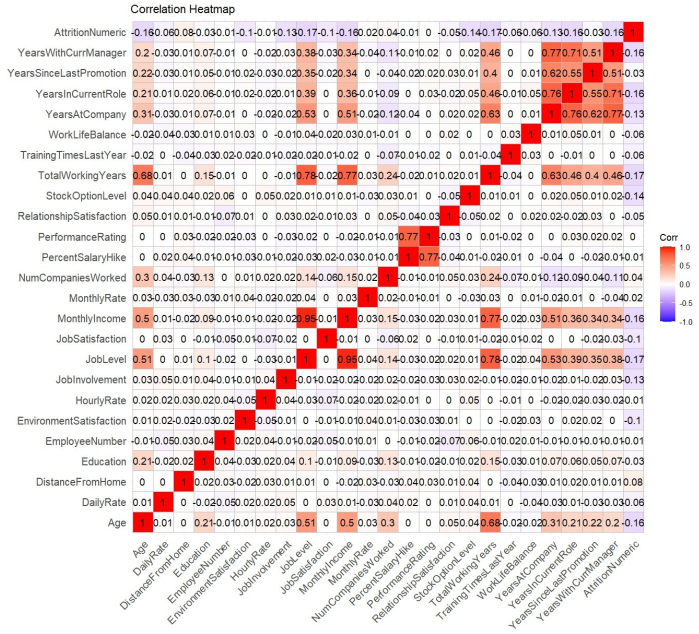

Figure 10: Correlation matrix

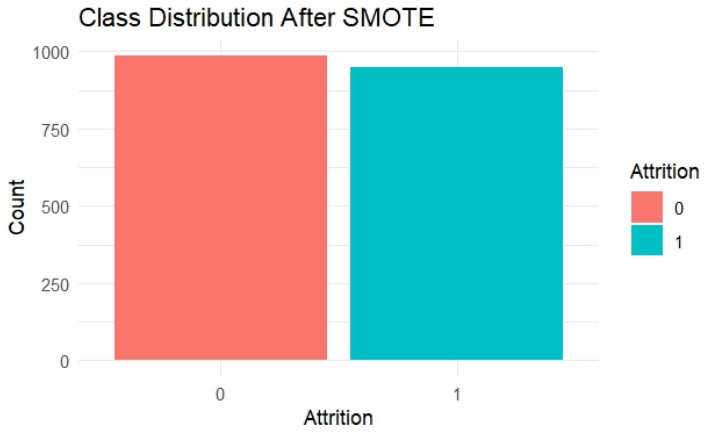

Figure 11: Class distribution after SMOTE

4 Methodology

4.1 What is SVM?

Support Vector Machines (SVM) are a type of supervised learning algorithm used for classification and regression tasks. SVM finds the optimal hyperplane that separates data points into different classes. It works by maximizing the distance between the nearest points of each class, known as support vectors. The goal in linear SVM is to find the best linear boundary that separates the "Yes" and "No" labels.

The decision boundary in SVM is defined by the following equation:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b \tag{1}$$

where:

- $\mathbf{w}$ is the weight vector,
- $\mathbf{x}$ is the feature vector,
- b is the bias term.

SVM tries to maximize the margin between the support vectors and the hyperplane, effectively finding the optimal boundary between the classes.

4.2 Why SVM?

Support Vector Machines are well-suited for binary classification tasks, making them an ideal choice for our employee attrition dataset. We chose SVM because it can handle high-dimensional data efficiently and is particularly effective for datasets with a moderate number of samples, such as our dataset.

4.3 Cross-Validation

A 5-fold cross-validation technique was applied to the data to reduce the risk of overfitting and to achieve more reliable performance metrics. K-fold cross-validation divides the dataset into k subsets (or folds), training the model on $k-1$ folds and testing it on the remaining fold. By using this technique, we can ensure that the model generalizes well to unseen data.

Through cross-validation, we tuned a hyperparameter known as C, which is a regularization term used in the optimization objective of the SVM algorithm. The tuned parameter obtained for C is 0.01.

This approach ensures that we build a robust and reliable model capable of making accurate predictions on new, unseen data.

5 Results and Evaluation

The results of the SVM model demonstrate its performance in predicting employee attrition, with key metrics such as accuracy, Kappa, Specificity, balanced accuracy, and the confusion matrix.

The model was trained on 1,937 samples with 39 predictors and two target classes. The accuracy of the model is 78.16%, with a 95% confidence interval of (72.98%, 82.75%). This shows that the model performs well in predicting employee attrition. The Kappa statistic, which measures the agreement between the predicted and actual values beyond what would be expected by chance (random guessing), is 0.4044. This indicates moderate agreement. A value of 0 for Kappa means the model is no different from random guessing.

Specificity (the true negative rate for class '1', attrition) is 0.7660, indicating the model correctly classifies 76.6% of the attrition cases. The Balanced Accuracy, which accounts for both sensitivity and specificity, is 0.7753, indicating a balanced performance between correctly predicting both classes.

The confusion matrix is presented below:

Prediction / Reference	0	1
0	195	10
1	51	37

This matrix shows how the model performs compared to the original labels. It indicates that the model correctly predicts 195 samples of the "0" class and incorrectly predicts 51 samples as class "1". It also correctly predicts 37 samples of the "1" class out of 47, which indicates moderate performance for this class.

Overall, the SVM model demonstrates strong performance in predicting employee attrition, with high accuracy and balanced accuracy.

5.1 Feature Importance

In this section, we present the feature importance derived from the Support Vector Machine (SVM) model, which helps us understand the contribution of each feature to the prediction of employee attrition.

The top five most important features contributing to the SVM decision boundary are:

- Job Role: Research Director
- Job Role: Sales Representative
- Environment Satisfaction
- Job Involvement
- Over Time (Yes)

6 Conclusion

Conclusion: In this work, we implemented a support vector machine (SVM) supervised learning model to predict employee attrition using several features from the IBM HR Analytics dataset. The data preprocessing included several steps such as removing the irrelevant and constant columns, one-hot encoding, and handling the class imbalance using SMOTE algorithm. We further pruned the features by analyzing the feature correlations using the

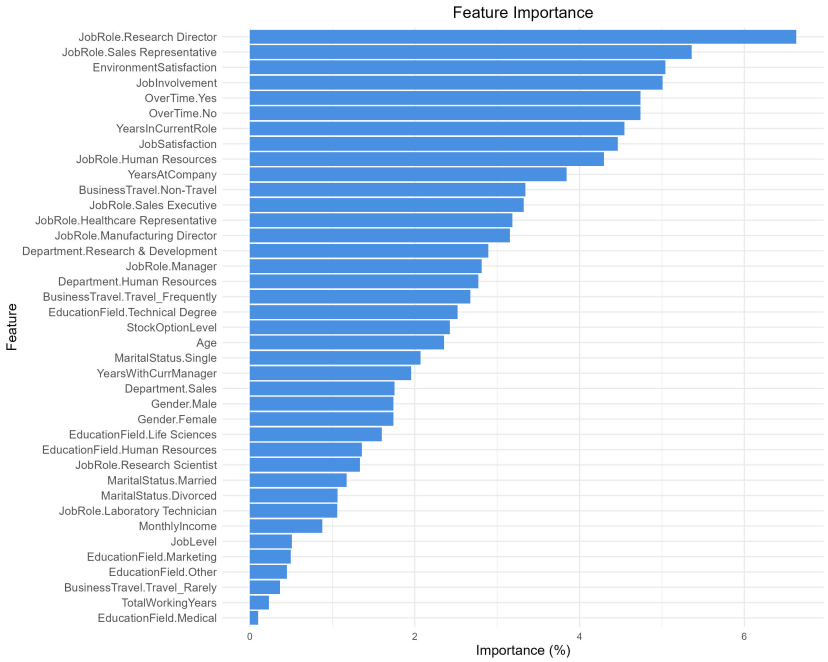

Figure 12: Feature importance plot derived from the SVM model.

correlation matrix. Through the cross-validation, the hyperparameter C was tuned, as it is the regularization term in the optimization function of the SVM algorithm. The results shown in the evaluation section show that while the model obtains a high balanced accuracy of 77%, there is still room for improvement in the prediction performance on the class "1". However the model's performance is satisfactory, it could be enhanced by adding more samples to the dataset, or experimenting other kernel functions. Overall, the current model provides a solid foundation for predicting the employees attrition, and the mentioned improvements can help optimizing the performance.

Code Availability

The code used for this analysis is available on GitHub: GitHub Repository.

References

[1] A. M S, T. Deshpande, and I. Data Scientists, "Ibm hr analytics employee attrition & performance," 2023.