Università degli Studi di Milano
Data Science for Economics

Understanding Employee Attrition

Emile Rahal

Statistical Learning Module

August 2023

Outline

1. **Abstract**
2. **Introduction**
    - Background
    - Research Questions
    - Significance of the Study
3. **Data Description**
    - Data Cleaning and Preprocessing
4. **Explanatory Data Analysis**
    - Numerical Variables
    - Categorical Variables
    - Correlation Analysis
5. **Preparing the data**
6. **Supervised Learning**
    - Logistic Regression
    - Random Forest
    - Decision Tree
    - k-NN
    - Models Comparison
        - Full Models
        - Subset Models (based on the feature importance)
    - Conclusion
7. **Unsupervised Learning**
    - PCA
    - k-Means Clustering
    - Interpretation
8. **Conclusion**
    - Summary of Findings
    - Future Research
9. **References**

1 Abstract

Employee attrition is a pressing challenge for organizations. This study investigates its factors and explores retention strategies. Using supervised and unsupervised machine learning, we identify critical attrition features and uncover behavior patterns.

Supervised models aid in feature selection, predicting attrition, and enhancing interpretability. Unsupervised techniques like PCA and K-Means clustering reveal employee segments, enabling specific retention strategies.

This report provides actionable HR insights that not only benefit organizations but also have a positive impact on employee well-being. By addressing attrition, organizations can create a better workplace, enhance job satisfaction, and improve overall employee experiences.

2 Introduction

Employee attrition poses significant challenges for organizations in today's competitive job market. Keeping skilled and motivated employees is important for maintaining productivity and reducing recruitment costs. We will try to understand the factors contributing to employee attrition and explore strategies to improve retention, with a focus on improving the work life of employees.

2.1 Background

Employee attrition, the phenomenon of employees leaving their jobs voluntarily, is a common concern for organizations across industries. In an era where talent is a valuable asset, the loss of skilled and experienced employees can have bad effects on a company's performance. Understanding the drivers behind attrition and implementing effective strategies are important for organizational success and for creating a better workplace for employees.

2.2 Research Questions

This study addresses the following key research questions:

- What are the main factors contributing to employee attrition?
- How can organizations keep employees in their jobs while enhancing the overall work environment?
- What insights can be gained from applying machine learning techniques to employee data?

2.3 Significance of the Study

By uncovering the factors influencing attrition and profiling employee clusters, companies can create a more supportive and satisfying work environment for employees. This report contributes to the knowledge of employee attrition, focusing on the importance of considering employees' well-being in addressing this critical problem in the corporate world.

3 Data Description

3.1 Overview

The **"Employee Attrition"** dataset, found on Kaggle, contains information related to employees in a company and aims to predict employee attrition, indicating whether an employee leaves the company or not. The dataset consists of two parts: a training set and a test set, totaling 25,491 records in the training set and 4,507 records in the test set.

3.1.1 Features

Feature	Type	Description
satisfaction_level	Numeric	Represent the employee's reported level of satisfaction, ranging from 0 to 10.
last_evaluation_rating	Numeric	Represent the employee's last evaluation rating scored on a scale from 0 to 10.
projects_worked_on	Numeric	Indicate the number of projects each employee has been involved in.
average_monthly_hours	Numeric	Represent the average number of monthly working hours for each employee.
time_spend_company	Numeric	Indicate the number of years an employee has spent with the company.
work_accident	Binary	Indicate whether an employee has experienced a work accident.
promotion_last_5years	Binary	Indicate whether an employee has been promoted in the last 5 years.
department	Categorical	Represent the department in which each employee works.
salary	Categorical	Represent the salary level of each employee, with categories such as "low," "medium," and potentially "high."

3.1.2 Target Variable

The target variable in this dataset is **Attrition**, which is a binary feature indicating whether an employee has left the company (1 for attrition) or is still employed (0 for no attrition).

3.2 Data Cleaning and Preprocessing

The dataset, originally split into training and test sets, was merged to increase the number of data points. No missing values were found in the dataset so no need for imputation.

3.2.1 Attrition Class Imbalance

The class distribution revealed a significant imbalance, with a notably larger number of employees in the "No Attrition" class (Class 0) compared to the "Attrition" class (Class 1). This class imbalance can pose challenges when developing predictive models, as machine learning algorithms may tend to favor the majority class, potentially leading to reduced predictive accuracy for the minority class.

3.2.2 Downsampling

To address the class imbalance, a downsampling approach was employed. This method involves reducing the size of the majority class (Class 0) to match the size of the minority class (Class 1). Achieving a balanced class distribution is beneficial for building machine learning models, as it helps prevent model bias and leads to more accurate predictions for both classes.

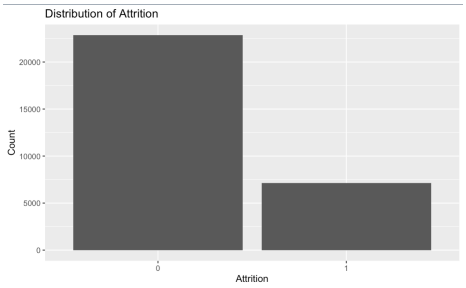


Figure 1: Before Downsampling

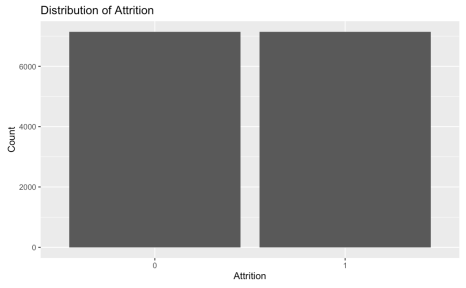


Figure 2: After Downsampling

4 Explanatory Data Analysis

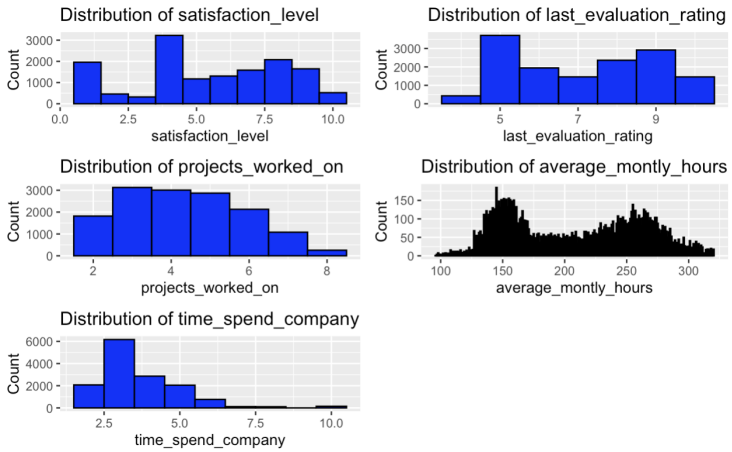
In this section, we explore the dataset to gain a deeper understanding of the variables and their relationships. Our exploration includes numerical variables, categorical variables, correlation analysis, and some specific factors potentially linked to attrition.

4.1 Numerical Variables

We begin by examining the distribution of numerical variables in the dataset. These variables include:

- **Satisfaction Level:** The average satisfaction level among employees is approximately 5.52/10, indicating a moderate level of satisfaction.
- **Last Evaluation Rating:** The average last evaluation rating is about 7.17/10, indicating relatively high evaluation ratings.
- **Projects Worked On:** On average, employees have worked on approximately 4.32 projects, suggesting they are involved in multiple projects.
- **Average Monthly Hours:** The average number of average monthly hours worked by employees is roughly 208.08 hours, indicating a moderate workload.
- **Time Spend Company:** On average, employees have spent about 3.64 years with the company, indicating a relatively short to moderate tenure.

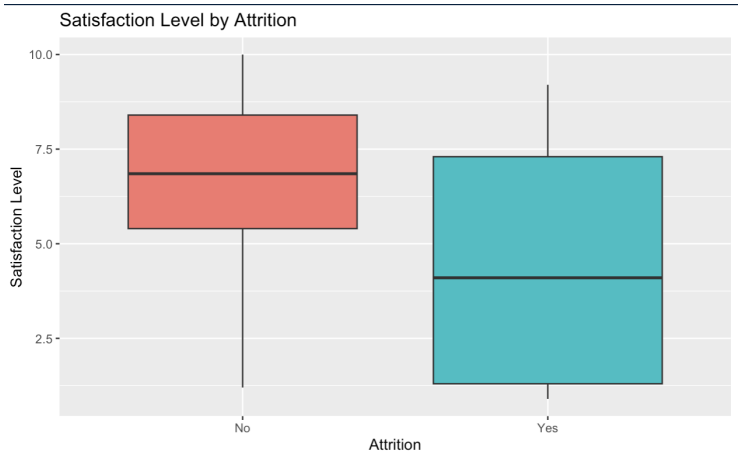
We present histograms to visualize the distribution of these variables.



We have included a selection of plots that we believe are particularly influential in understanding employee attrition within the organization. These visualizations have been chosen for their ability to shed light on the factors that are related to attrition and provide valuable insights into the dataset.

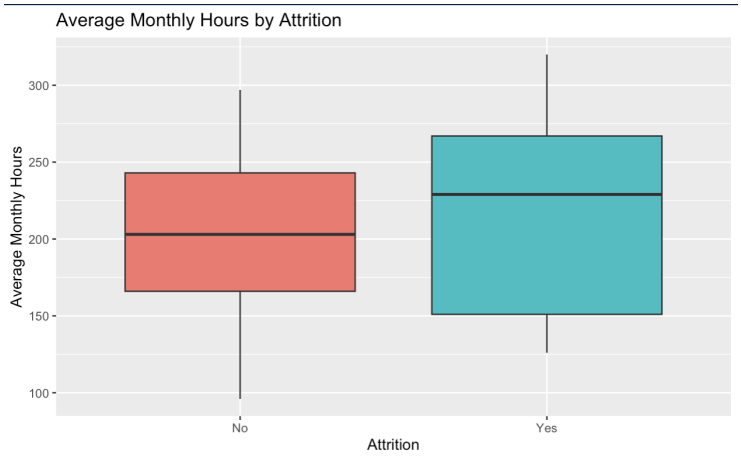
4.1.1 Satisfaction Level by Attrition

We explore how satisfaction level relates to attrition. Employees who left the company (attrition 1) had significantly lower satisfaction levels than those who stayed.



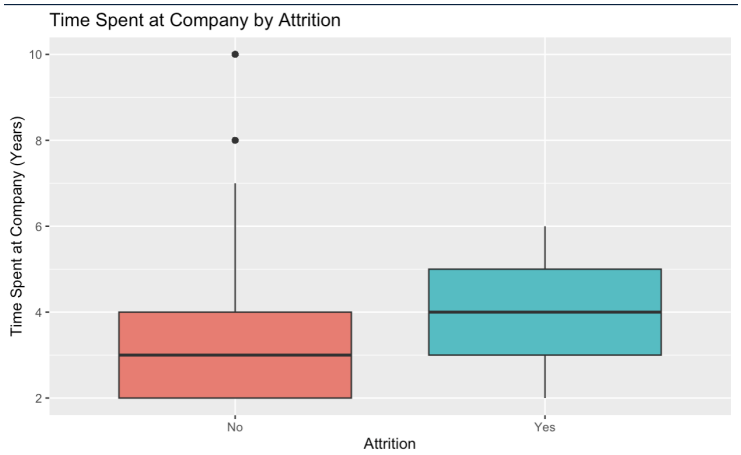
4.1.2 Average Monthly Hours by Attrition

Examining average monthly hours worked, we find that employees who left the company (attrition 1) worked more hours on average compared to those who stayed.



4.1.3 Time Spend Company by Attrition

When considering the time spent at the company, employees who left the company (attrition 1) had spent slightly more time than those who stayed.



4.2 Categorical Variables

4.2.1 Salary by Attrition

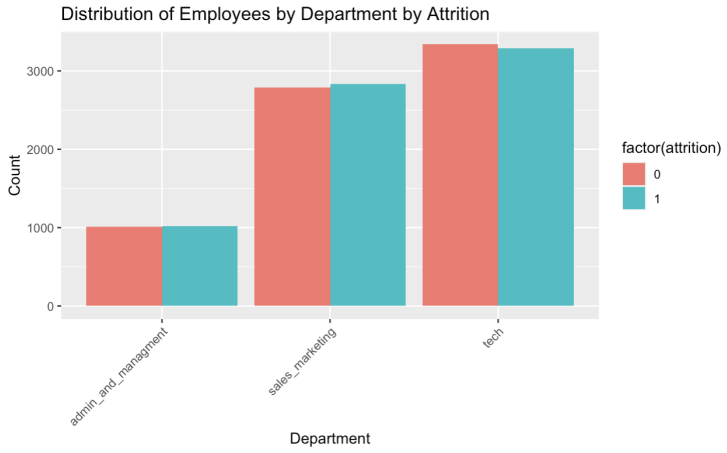
We also investigate the relationship between salary levels and attrition. It appears that employees with low and medium salaries have a higher attrition rate compared to those with high salaries.

4.2.2 Department by Attrition

To simplify the analysis and reduce the number of categories, we have mapped the departments into broader categories. After mapping, the specific department names were grouped into the following broader categories:

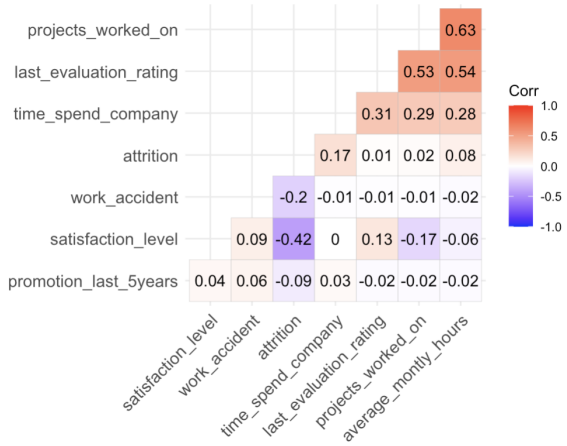
- **Sales and Marketing**: This category includes employees from departments like "product mng," "sales," and "marketing." These employees are primarily involved in sales and marketing activities.
- **Technical**: The "tech" category comprises employees from departments such as "IT," "technical," and "RandD." These employees are primarily engaged in technical and research-oriented roles.
- **Administration and Management**: This category covers employees from departments like "hr," "management," and "accounting." These employees are often involved in administrative and managerial functions.

While there are variations in the total number of employees across these broader department categories, attrition rates appear to be fairly uniform. The attrition rates are quite similar across these redefined department categories.



4.3 Correlation Analysis

Finally, we conduct a correlation analysis to identify relationships between variables.



Employee "satisfaction level" appears to be the most critical factor influencing attrition, with lower satisfaction levels associated with higher attrition rates. Other factors such as "evaluation ratings", "projects worked on", and "average monthly hours" have weaker associations with attrition. Factors like "work accidents", "time spent at the company", and recent "promotions" also play a role but to a lesser extent. It's important to note that correlation does not imply causation and additional analyses may be needed to understand the underlying causes of attrition.

Variable	GVIF
satisfaction_level	1.122206
last_evaluation_rating	1.688310
projects_worked_on	1.929474
average_montly_hours	1.855082
time_spend_company	1.146753
work_accident	1.012281
promotion_last_5years	1.019240
department	1.019516
salary	1.029708

To ensure that our analysis is not affected by multicollinearity, we employed the Variance Inflation Factor (VIF). While some variables show relatively high correlations, the VIF analysis reveals that there is no severe multicollinearity among the features (VIF values near 1). This indicates that although certain variables may have strong relationships, they do not create problematic issues in the analysis.

5 Preparing the Data

5.1 Categorical Encoding

In the preprocessing phase, we prepared categorical variables for machine learning algorithms. For the "Salary" feature, we employed ordinal encoding, categorizing levels as "low," "medium," and "high," ensuring that the order of categories is preserved. In contrast, the "Department" variable was one-hot encoded and transformed into binary columns for each department category. Importantly, we removed one of the dummy columns to avoid the dummy trap, which can arise from a perfect correlation between dummy variables, potentially affecting model stability and interpretability.

5.2 Numerical Scaling

Additionally, we addressed numerical variables through a process known as standardization or "z-score scaling." This technique makes numerical variables more directly comparable and can enhance the performance of specific machine-learning algorithms. By scaling the features to have a mean of 0 and a standard deviation of 1, we ensure that no single feature dominates the learning process, resulting in a more balanced and effective model.

5.3 Splitting the Data

To evaluate our supervised learning models effectively, we divided the dataset into training (80%) and testing (20%) sets. This partitioning strategy allows us to train our models on one subset and assess their performance on an independent test set, ensuring the reliability of our model evaluations.

6 Supervised Learning

In this section, we enter the world of supervised learning to construct predictive models for employee attrition and to understand the factors that influence attrition in an organization. The models considered in this analysis include Logistic Regression, Random Forest, Decision Tree, and k-Nearest Neighbors (k-NN). We will evaluate their performance and conclude their effectiveness for attrition prediction.

6.1 Logistic Regression

We initiate our supervised learning journey with Logistic Regression, a straightforward yet powerful technique for binary classification. This model will serve as our baseline for attrition prediction.

The logistic regression model is fitted to predict employee attrition based on all the features. The estimated coefficients and their significance levels are presented below:

```
Call:
glm(formula = attrition ~ ., family = binomial, data = train_data)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -0.37943    0.06657  -5.700 1.20e-08 ***
satisfaction_level          -1.12669    0.02738 -41.144  < 2e-16 ***
last_evaluation_rating       0.14281    0.02920   4.890 1.01e-06 ***
projects_worked_on          -0.50847    0.03279 -15.508  < 2e-16 ***
average_montly_hours         0.26301    0.03125   8.417  < 2e-16 ***
time_spend_company           0.52232    0.02519  20.733  < 2e-16 ***
work_accident1              -1.57005    0.08427 -18.631  < 2e-16 ***
promotion_last_5years1      -1.26953    0.21663  -5.860 4.62e-09 ***
salary.L                    -1.47663    0.08210 -17.987  < 2e-16 ***
salary.Q                    -0.39441    0.05467  -7.214 5.44e-13 ***
department_sales_marketing1  0.02147    0.06915   0.310    0.756
department_tech1            -0.01403    0.06792  -0.207    0.836
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The coefficients represent the estimated impact of each variable on the likelihood of attrition. Significance levels are denoted as *** for highly significant. Here are some key observations:

- **Satisfaction Level:** A decrease in satisfaction level is associated with a significantly higher likelihood of attrition.
- **Time Spend in the Company:** More time spent in the company is associated with a significantly higher likelihood of attrition.
- **Work Accident:** Having a work accident is associated with a significantly lower likelihood of attrition.
- **Promotion in the Last 5 Years:** Having received a promotion in the last 5 years is associated with a significantly lower likelihood of attrition.
- **Salary Level:** Lower salary levels are associated with a significantly higher likelihood of attrition. This suggests that employees with lower salaries are more likely to leave the company.

These coefficients provide valuable insights into the factors influencing employee attrition in the organization. We will further evaluate the model's performance in subsequent sections.

6.2 Random Forest

Random Forest is a powerful ensemble learning technique used for classification tasks. In our analysis, we employed Random Forest to predict employee attrition based on various features. The model was built using 100 decision trees and considered a random subset of three variables at each split.

Type of Random Forest	Classification
Number of Trees	100
Number of Variables Tried at Each Split	3

The Random Forest model exhibited promising performance. The out-of-bag (OOB) estimate of the error rate, a measure of the model's accuracy, was approximately 1.69%. This low error rate suggests that the Random Forest model is effective at classifying employees into attrition and non-attrition categories.

The confusion matrix, presented below, provides further insights into the model's performance:

	Attrition (1)	**No Attrition (0)**
Attrition (1)	5570	144
No Attrition (0)	49	5665
Class Error	0.025	0.009

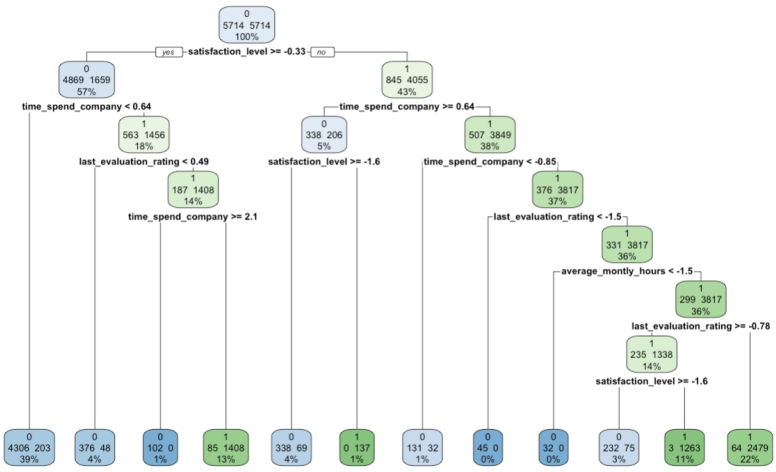
From the confusion matrix, we can observe that the model achieved a relatively low-class error rate for both attrition and no attrition cases. This suggests that the Random Forest model is effective at correctly classifying employees who leave (attrition = 1) and those who stay (attrition = 0).

Overall, the Random Forest model demonstrates strong predictive capabilities for identifying employee attrition. In the subsequent sections, we will compare the performance of this model with other supervised learning algorithms to determine the most suitable approach for addressing attrition in our organization.

6.3 Decision Tree

Decision Trees offer transparency in model decision-making. We explore this technique to gain insights into the hierarchy of factors influencing attrition.

A Decision Tree model was constructed to predict employee attrition based on all the features. The tree was developed using recursive partitioning, where the data was split into subgroups based on the most influential features.

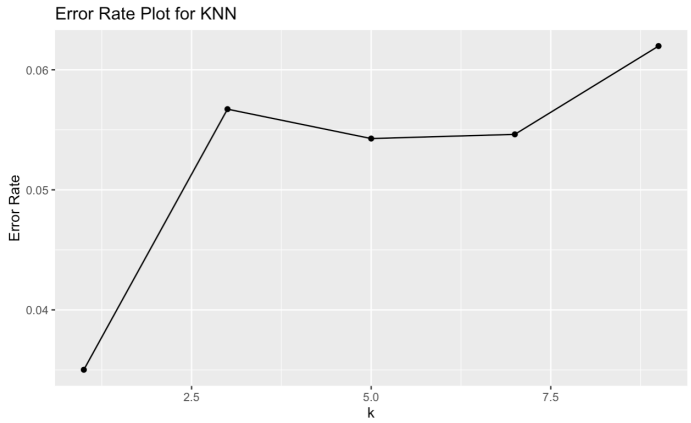


The Decision Tree provides a clear and interpretable framework for understanding the factors influencing employee attrition within the organization. Each branch of the tree represents a set of conditions that lead to a prediction of attrition or non-attrition. In the subsequent sections, we will compare the performance of this model with other supervised learning algorithms and conclude the most effective approach to address attrition.

6.4 k-NN (k-Nearest Neighbors)

K-NN is a versatile and intuitive classification method that makes predictions based on the majority class of its nearest neighbors. However, the crucial decision in K-NN lies in selecting the number of neighbors, denoted as "k." Several values of "k" were explored to evaluate model performance.

To visualize the selection of "k," an error-rate vs. "k" plot was generated to illustrate the error rate's behavior as "k" varies. This plot helped inform the decision to choose "k=5" as the optimal parameter value.



When "k" is set to 1, the error rate is approximately 0.0350. This indicates that the K-NN model with only one neighbor closely fits the training data but may be sensitive to noise or outliers.

As "k" is increased to 3, 5, 7, and 9, the error rate also increases. This suggests that the model's performance becomes less sensitive to individual data points with larger "k" values. However, it's crucial to strike a balance, as an excessively high "k" value might lead to underfitting.

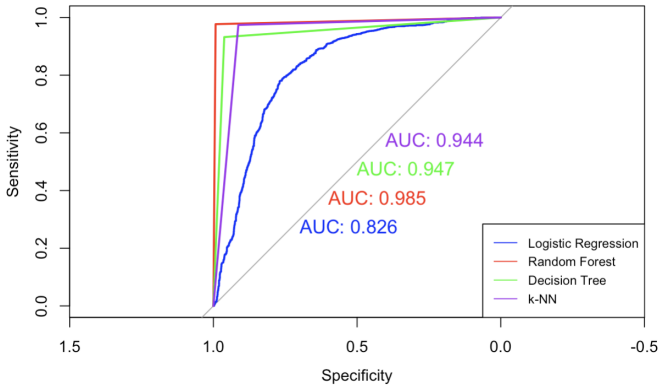
After careful evaluation, a "k" value of 5 was selected. The 5-NN model performs well on both the training and test datasets, achieving high accuracy and balanced sensitivity and specificity. This indicates that the model generalizes effectively to new, unseen data, making it a robust classifier for predicting attrition.

6.5 Models Comparison

In this subsection, we comprehensively compare the performance of the models employed in the previous sections. We assess their accuracy, precision, recall, and other relevant metrics to determine which model(s) best suit our attrition prediction needs. We also included a feature importance plot to show the contribution of the variables in the models.

6.5.1 ROC-AUC

The Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) is a critical evaluation metric for binary classification models. It provides a comprehensive measure of a model's ability to discriminate between the positive and negative classes across different probability thresholds. A higher ROC-AUC score signifies better model performance, with values closer to 1 indicating excellent discrimination ability.



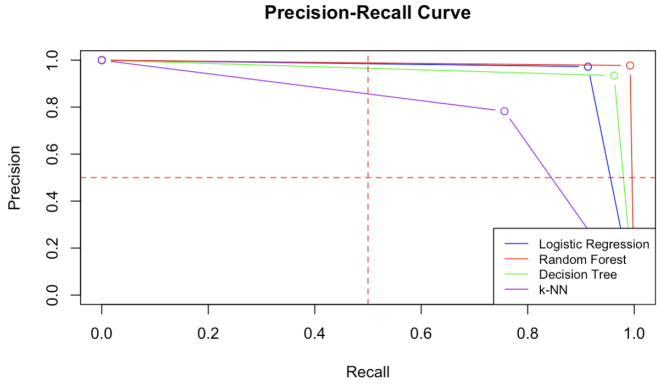
- **Logistic Regression:** An AUC score of 0.826 indicates moderate discriminative power.
- **Random Forest):** An AUC score of 0.985 suggests very high discriminative power.
- **Decision Tree:** An AUC score of 0.947 indicates good discriminative ability.
- **k-Nearest Neighbors:** An AUC score of 0.944 shows strong discriminative power.

In summary, the Random Forest model stands out with the highest score, indicating superior performance in distinguishing employees who leave from those who stay.

6.5.2 Precision-Recall Analysis

In addition to ROC-AUC analysis, we evaluate the models using Precision-Recall curves.

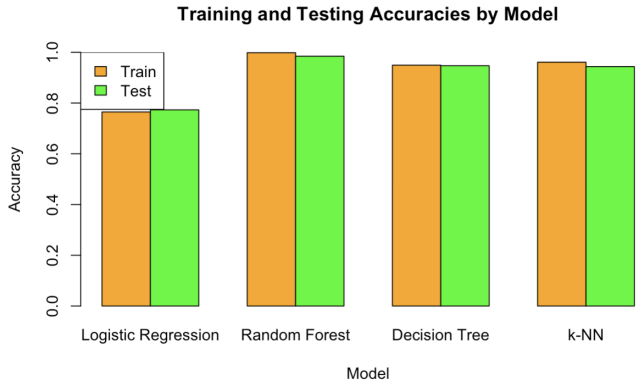
This analysis focuses on the trade-off between precision (the ability of the model to correctly identify true positives) and recall (the ability of the model to find all relevant instances). A model with high precision ensures that when it predicts an employee will leave, while a high recall indicates that the model is effective at identifying most of the employees who leave the company.



From our results, we observe that the Random Forest model exhibits exceptional precision and recall scores, indicating its strong predictive power. On the other hand, the Decision Tree and k-NN models also perform well, striking a balance between precision and recall. However, the Logistic Regression model, while respectable, shows slightly lower precision and recall values compared to the other models.

6.5.3 Accuracy

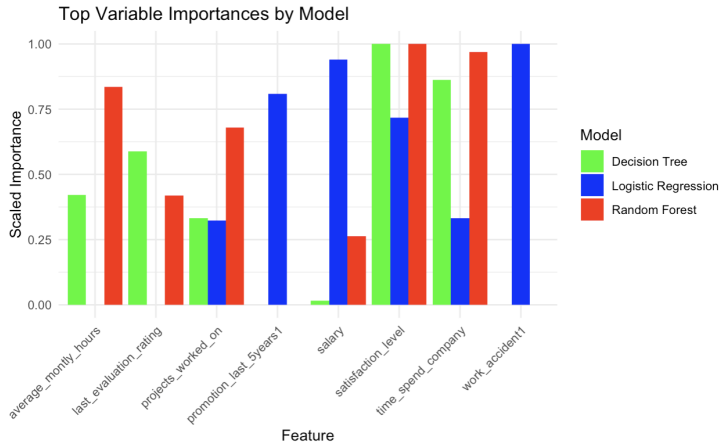
This metric provides valuable insights into the models' overall classification performance. It measures the overall correctness of the model's predictions.



From our results, we observe that the Random Forest model achieves remarkable accuracy value on both the training and test datasets. The Decision Tree and k-NN models also exhibit strong performance. While the Logistic Regression model's accuracy is respectable, they are slightly lower compared to the other models.

6.5.4 Feature Importance

We evaluated the importance of the features in predicting employee attrition for the models. Understanding feature importance helps us identify which factors play a crucial role in predicting attrition. It's important to note that k-NN does not provide explicit feature importance scores because it relies on proximity to neighbors for classification and does not assign weights or ranks to individual features. Therefore, we could not include it in the plot.



- **Satisfaction Level:** Across all three models, namely Logistic Regression, Random Forest, and Decision Tree, the "Satisfaction Level" consistently emerges as the most important feature. This uniformity in importance shows the effect of employee satisfaction level in predicting and influencing attrition.

- **Time Spent at the Company:** In both the Random Forest and Decision Tree models, "Time Spent at the Company" ranks as the second most important feature. This highlights the significance of the number of years an employee has spent with the company as a robust predictor of attrition.

- **Salary:** Interestingly, "Salary" assumes varying importance across the models. In the Logistic Regression model, it holds the position of the second most important feature. However, in the Random Forest and Decision Tree models, it carries a lower importance score. This divergence could be attributed to differences in how these models handle categorical variables or nonlinear relationships.

- **Other Features:** Features such as "Last Evaluation Rating," "Average Monthly Hours," and "Projects Worked On" exhibit varying levels of importance across the models. Generally, these features are considered less important than satisfaction level and time spent at the company.

- **Promotion in the Last 5 Years:** Notably, the feature "Promotion in the Last 5 Years" demonstrates relatively higher importance in the Logistic Regression model but does not hold the same prominence in the Random Forest and Decision Tree models. This suggests that the influence of promotions on attrition may not be consistent across all modeling techniques.

This comprehensive feature analysis enhances our understanding of the key factors influencing attrition prediction and will help us make a subset selection for the features.

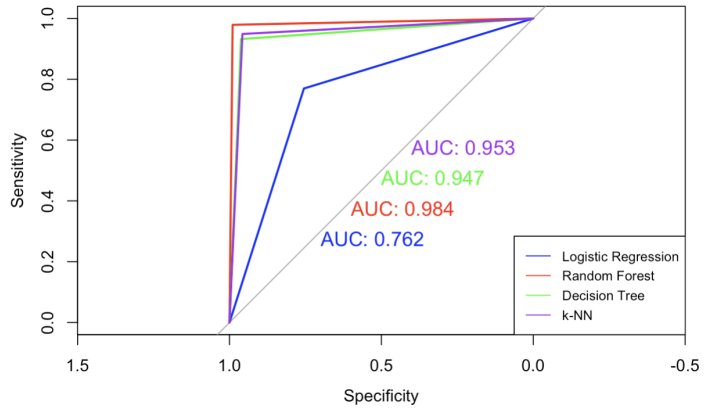
6.6 Subset Feature Selection and Model Assessment

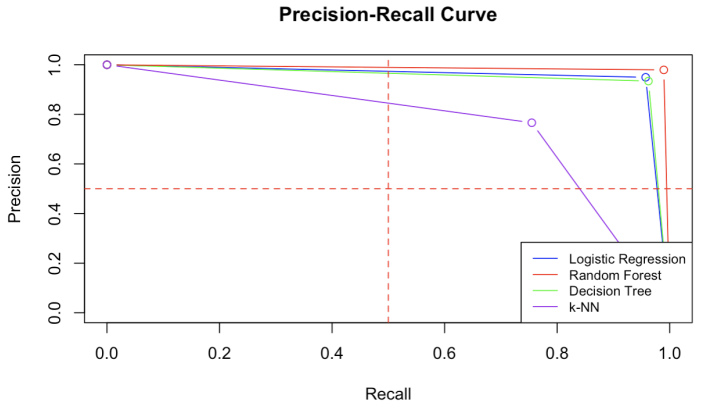
In the pursuit of constructing an accurate model for predicting employee attrition, we employed feature selection techniques. These techniques involve identifying and selecting a subset of the most important features, as determined by our earlier analysis. The rationale behind feature subset selection is to enhance model interpretability, reduce computational complexity, and potentially improve model generalization.

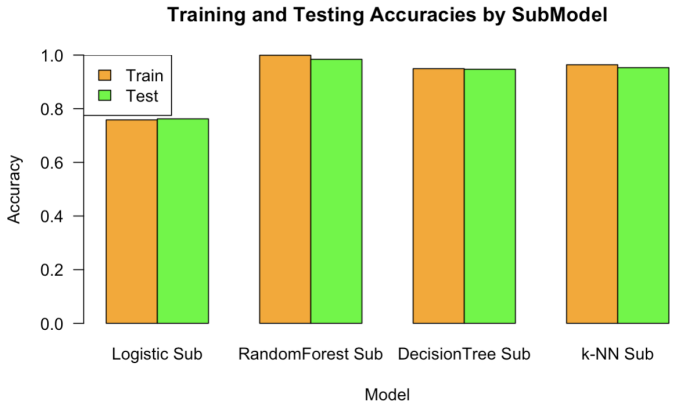
To facilitate this process, we initially determined the most influential features (as mentioned in the earlier section), with a particular focus on "Satisfaction Level," "Time Spent at the Company," "Salary," and "Promotion in the Last 5 Years." These features consistently exhibited high importance across our models—Logistic Regression, Random Forest, and Decision Tree.

Our objective was twofold: first, to evaluate the performance of the chosen features in predicting attrition, and second, to compare the effectiveness of the algorithms on this refined model.

By exploring different model variations with this subset of features, we aimed to strike a balance between predictive accuracy and model simplicity. The following plots show the outcomes of this feature selection and model assessment process, shedding light on the submodels' performance.







In transitioning from full models to their respective subset models, we observed somehow consistent patterns in performance across different algorithms.

-The Random Forest Subset maintained its predictive strength with a reduced feature set, preserving high AUC, precision, recall, and accuracy.

-The k-NN Subset also demonstrated adaptability to feature reduction, with improved AUC and precision while maintaining good recall.

-Decision Tree Subset retained consistent AUC, precision, and recall, indicating its resilience to feature reduction.

-The Logistic Regression Subset saw a notable drop in AUC but exhibited a slight increase in precision.

These insights into our subset models' performance metrics show that the transition from full to subset models not only optimizes performance but also enhances efficiency.

6.7 Conclusion

In our exploration of attrition prediction models in supervised learning, we've explored the world of feature selection and subset models. Our findings show that reducing features can make models easier to understand while maintaining their predictive abilities. Remarkably, Random Forest kept performing well even with fewer features. k-NN also adapted to fewer features, and Decision Tree remained strong.

Additionally, we consistently found 'Job satisfaction,' 'Last evaluations,' 'Time Spent at the company,' and 'Salary' to be crucial attrition predictors in all models. These insights emphasize the influence of these factors on employee attrition.

Our study not only highlights the balance between model simplicity and power but also underscores the significance of these key factors. These insights guide HR professionals and management to focus on employee engagement, work-life balance, compensation, and career growth to reduce attrition and improve the employee experience.

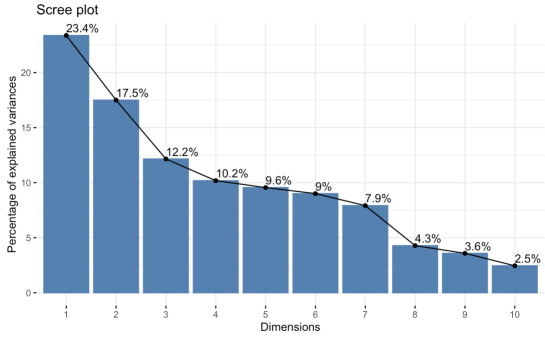
7 Unsupervised Learning

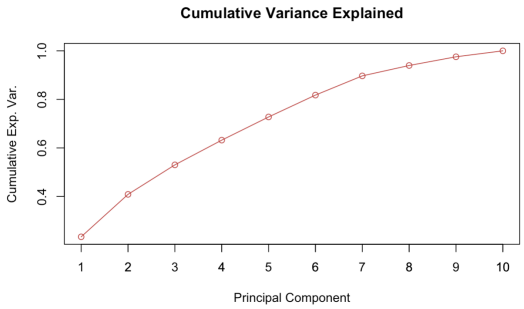
In our exploration of employee attrition, we venture into the realm of unsupervised learning techniques. This analytical journey involves uncovering hidden patterns and employee profiles within the dataset, offering a better perspective on the underlying factors contributing to attrition. Through Principal Component Analysis (PCA) and K-Means Clustering, we shed light on distinct dimensions of employee behavior and unveil unique profiles.

7.1 Principal Component Analysis (PCA)

In our quest to explore the complex dynamics of employee attrition, Principal Component Analysis (PCA) emerges as a powerful tool. This method allows us to show patterns and dependencies among various employee attributes into a manageable set of principal components. By reducing the dimensionality of our data, PCA unveils the underlying structure that drives employee behaviors and engagement.

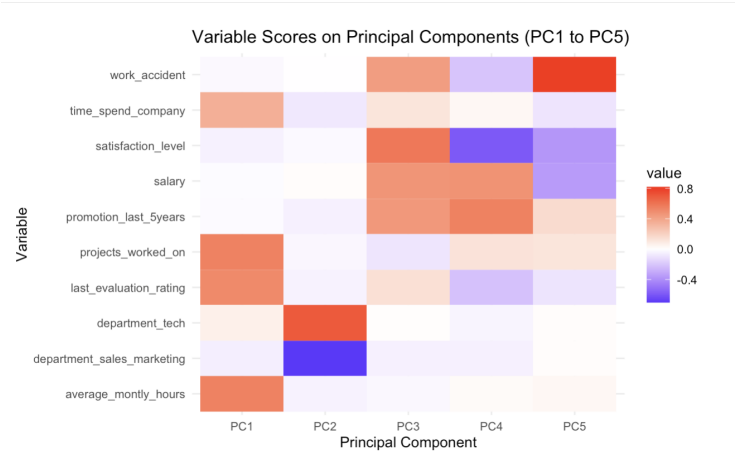
After applying the PCA on the features, we noticed that the first five components explain approximately 73% of the variance in our dataset as shown in the scree and the cumulative variance explained plots below.





7.1.1 PCA Scores

The PCA scores for each feature provide insights into their relationships with each other. The following heatmap illustrates these relationships:



7.1.2 Interpretation of scores

Our interpretation of the principal components scores for the features is as follows:

PC1: It represents employee work engagement, where the values indicate commitment through involvement in projects, higher average monthly hours, and positive evaluation ratings.

PC2: It signifies departmental variation, particularly between the "sales/marketing" and "tech" departments, highlighting differences in employee characteristics or behaviors based on their department.

PC3: It relates to job satisfaction, promotions in the last 5 years, and higher salaries, with employees having accidents possibly having better scores on this component.

PC4: It reflects career progression influenced by promotions and salary, with some influence from work accidents on satisfaction.

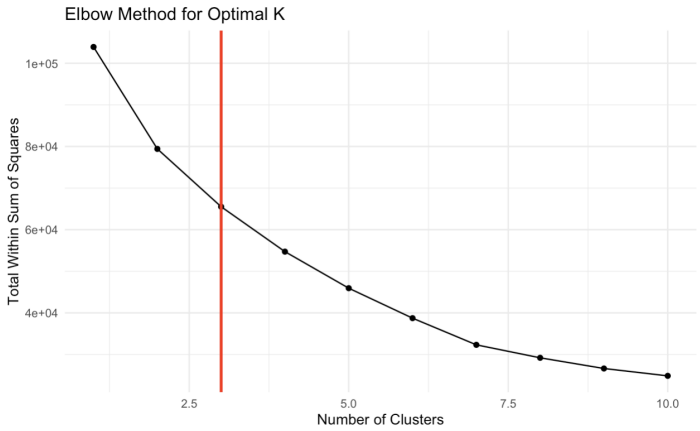
PC5: It indicates a trade-off between work safety (more accidents) and lower job satisfaction levels, suggesting a delicate balance between these factors in employee profiles.

We identified five key principal components, each shedding light on different aspects of employee behavior in an organization

7.2 Clustering Analysis

In this section, we employ the k-means clustering technique on the first 5 PCs to group employees based on their profiles, aiming to uncover meaningful patterns and provide insights for employee behavior.

To determine the optimal number of clusters, we utilize the elbow method. This technique helps us identify a suitable value for K, the number of clusters, by analyzing the within-cluster sum of squares (WSS) as K varies. The point at which the rate of decrease in WSS starts to slow down is typically considered the "elbow" or the optimal number of clusters.



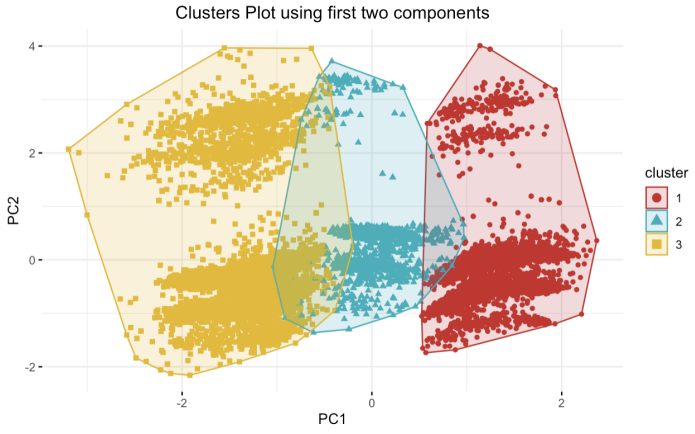
we observe that the elbow point occurs at $K = 3$. Therefore, we proceed with K-Means clustering with $K = 3$ as the optimal number of clusters.

7.2.1 Cluster Visualization in Principal Components

Now, we visualize the clusters in the principal component space (PC1 and PC2) to gain insights into the distribution of employees within each cluster.

We can see below the distribution of the clustered employees within the first two principal components. This visualization allows us to observe the separation of clusters in this reduced-dimensional space. As we can see the 3 clusters distributed on the PC1 indicate the level of work commitment so we can understand how it changes among the clusters, and we notice that in each cluster we have 2 divided groups that show the departmental variation of the employees, in simpler words, we can see that employees are not only separated by their level of work commitment (PC1) but also by their department (PC2).

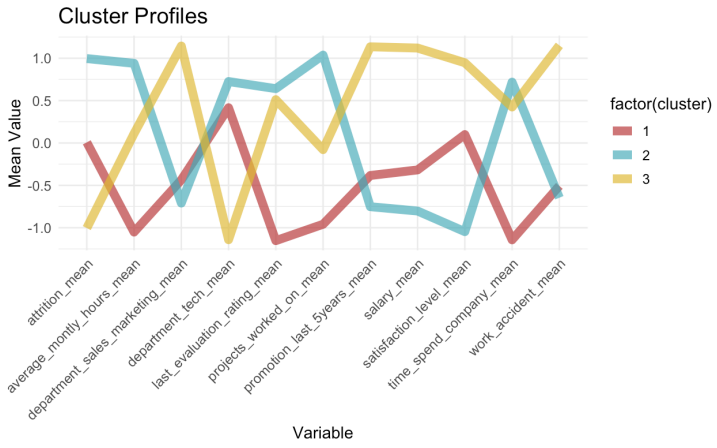
For example, employees in one cluster might have a high work commitment, but within that cluster, there could be a subgroup from the "sales/marketing" department showing different characteristics compared to the "tech" department. This indicates that the department to which an employee belongs plays a role in shaping their profile or behavior.



7.2.2 Cluster Profiles

These clusters represent distinct groups of employees who share similar characteristics and behaviors. Understanding their profiles is essential for organizations aiming to tailor their strategies for employee engagement, retention, and development.

We will explore the key attributes, such as job satisfaction, evaluations, workload, tenure, salary, and attrition rates, within each cluster. By gaining insights into these profiles, organizations can make informed decisions to enhance employee satisfaction, avoid attrition risks, and offer a more productive and harmonious workplace.



Cluster 1 *"New Talent"*

This cluster represents employees with moderate levels of satisfaction, low evaluation ratings, and lower workloads. They are relatively new to the company and have medium salaries and attrition rates. This cluster might represent a mix of junior or new employees who haven't yet reached high workloads or evaluation ratings.

Advices:

-Invest in training and development programs to help these employees reach their potential.

-Monitor their progress and offer clear pathways for career advancement within the organization.

Cluster 2 *"High Potential, High Attrition"*

This cluster represents a group of employees with high workloads and evaluation ratings but low job satisfaction. They relatively spent a long time in the company, but their low salaries and high attrition rates indicate a risk of quitting. These employees might be dissatisfied due to the salary factor.

Advices:

-Conduct regular surveys or feedback sessions to understand the specific causes of dissatisfaction.

-Consider salary adjustments or performance-based incentives to increase job satisfaction.

-Provide opportunities for career progression to keep valuable employees.

Cluster 3 *"Loyal High Performers"*

This cluster represents highly engaged and satisfied employees who are dedicated to their work. They spent a lot of time in the company, with high salaries, indicating loyalty and compensation. The low attrition rate suggests that these employees are likely to stay with the company.

Advices:

-Continue to recognize and reward these high-performing, loyal employees.

-Offer opportunities for career growth and advancement within the company to keep them engaged.

-Leverage their experience and expertise to mentor and train newer employees.

8 Conclusion

8.1 Summary of Findings:

In our comprehensive analysis of employee attrition, we explored various supervised and unsupervised learning techniques to gain insights into the factors influencing attrition within the organization. We found that employee satisfaction, evaluations, and tenure are consistently vital factors affecting attrition. The balance between model interpretability and predictive strength highlighted the importance of these key features.

Through supervised learning models, we identified Random Forest as resilient to feature reduction, maintaining strong predictive performance even with fewer features. K-nearest neighbors (K-NN) demonstrated adaptability to feature reduction, and Decision Tree displayed robustness as well.

Unsupervised learning, specifically K-means clustering, allowed us to group employees into distinct clusters based on their characteristics. We identified three key clusters: "New Talent," "High Potential, High Attrition," and "Loyal High Performers." These clusters provided actionable insights into addressing attrition through tailored HR strategies.

8.2 Future Research:

In future statistical learning projects, we can delve deeper into the attrition prediction model. Investigating advanced modeling techniques, such as deep learning, may lead to improved predictive accuracy. Moreover, fine-tuning hyperparameters and conducting feature engineering can further enhance model performance.

Additionally, exploring alternative unsupervised learning approaches, like hierarchical clustering, can provide different perspectives on employee segmentation to address the broader goal of creating a better workplace and enhancing employee satisfaction.

By combining statistical learning with a comprehensive approach to employee well-being, future research can provide comprehensive solutions for organizations, ultimately leading to improved employee retention, job satisfaction, and organizational success.

9 Refrences

Dataset: Employee Attrition Dataset on Kaggle
Code: GitHub Repository