Università degli studi di Milano

Data Science for Economics

STATISTICAL LEARNING PROJECT

Employee Attrition: Investigating Why Workers Quit

Student: Ketrin Hristova

Registration number: 13209A

Academic Year: 2022/2023

Contents

1. Abstract

Employee attrition is a critical challenge facing organizations across industries. High turnover can lead to decreased productivity, loss of valuable knowledge, and increased hiring costs. The report seeks to uncover the factors driving attrition at IBM by applying a diverse set of statistical learning techniques to a rich HR dataset. The analysis began with extensive exploratory data analysis to understand correlations in the data. Supervised learning models, including logistic regression, LASSO, random forest, decision trees, and support vector machines, were trained to predict employee attrition, with LASSO being the top performer. For unsupervised learning, principal component analysis identified underlying structures related to employee experience, performance, and job attributes. Multiple correspondence analyses revealed associations between departments, job roles, and attrition. Finally, t-SNE visualization and density-based clustering segmented employees into distinct groups based on their characteristics and attrition tendencies. The project provides meaningful insights into the drivers of attrition at IBM. The predictive models can support data-driven retention initiatives by identifying high-risk employees.

2. The Dataset

The dataset this analysis uses comes from IBM HR Analytics and represents a rich set of employee characteristics and outcomes. The dataset, which is hosted on Kaggle (https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset), contains 35 variables capturing demographics, compensation, tenure, satisfaction, and performance for 1470 employees. The target variable is a binary indicator of whether employees left the company during the observation period. Key variables include:

- Demographics: age, gender, marital status, distance from home
- Compensation: hourly/monthly rates, income
- Tenure: years in company, role, with manager
- Satisfaction: job, environment, relationship
- Performance: rating, salary hike
- Outcome: attrition

A challenge with the data is the imbalance between classes for the attrition variable. Only 16.1% of employees left the company, while 83.9% remained. This skew could impair model performance, an issue that should be addressed through techniques like oversampling. Overall, the richness of the data provides a strong foundation for investigating the factors associated with employee retention.

3. Explanatory Analysis

3.1 Introduction

The IBM employee dataset contains a rich set of variables capturing demographics, compensation, tenure, satisfaction, performance, and attrition for 1470 employees. Initial data quality assessment identified an imbalance issue, with only 16.1% attrition cases. Additional exploration was undertaken to improve data quality and integrity further.

3.2 Correlation Between Variables

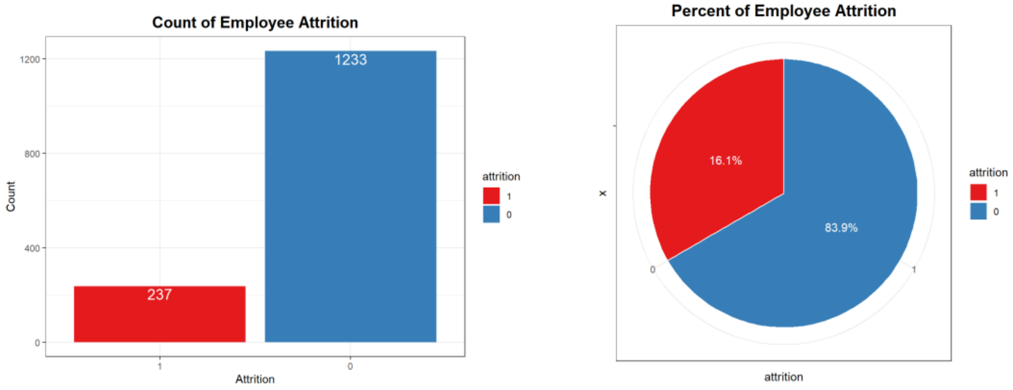


Figure 1

Upon computing the turnover rate, denoting the percentage of employees departing from the company within a specified timeframe, the visual representations revealed that 1,233 employees (83.9%) opted to remain within the organization, whereas 237 employees (16.1%) chose to depart. Notably, the dataset exhibits an imbalance, as a larger proportion of individuals opted to retain their association with the organization, while a comparatively smaller subset decided to leave.

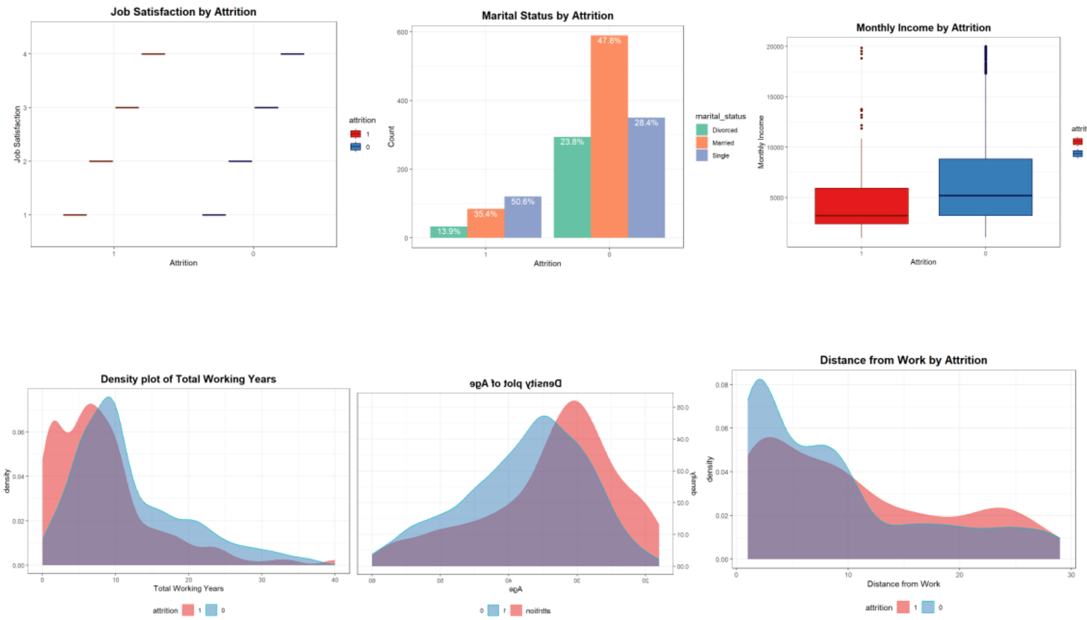


Figure 2

The plots above show us the correlation between some independent variables and attrition variables.

-The density plot of Age shows that the majority of employees are between 28 and 36 years. So it seems that those who left the organization were relatively younger.

-The Marital Status plot indicates that the majority of those who left was relatively single.

-From the Monthly Income we can see that to a large majority of those who left had a relatively lower monthly income.

-From the Job Satisfaction plot it seems to a large majority of those who left had a relatively lower job satisfaction.

-The density plot of Total Working Years we see that a large majority of those who left had a relatively shorter working years in the organization.

-The Distance from Work plot indicates that a large majority of those who left had a relatively lower distance from work.

3.3 Correlation Analysis and Detection

The next step involved examining correlations among numerical features, emphasizing the importance of assessing potential relationships between numeric predictors in the dataset.

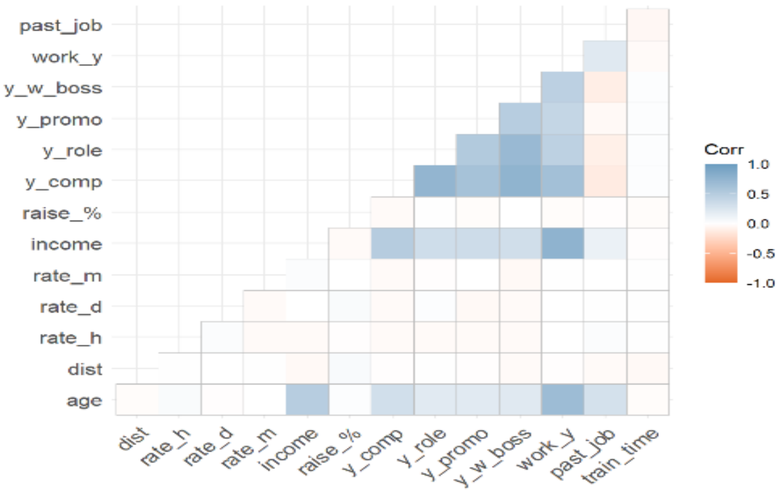


Figure 3

The graphical representation unveils variables that pose concerns, including distance from home, hourly rate, daily rate, monthly rate, percent salary hike, and training times last year. (figure 3)

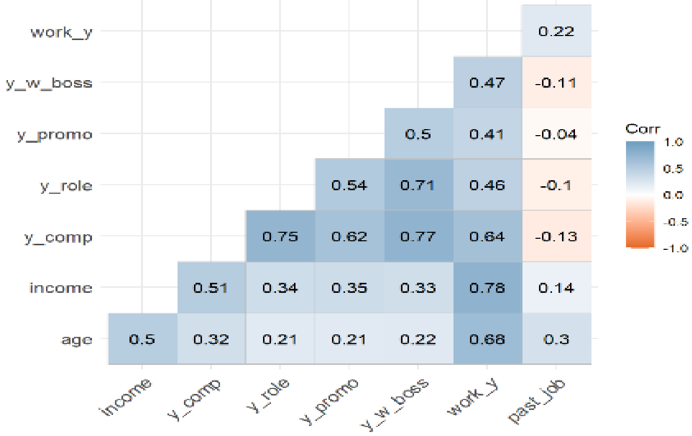


Figure 4

The graphical representation above shows the correlation without the randomly generated variables, indicating a strong positive correlation between four pairs of features. (figure 4)

3.4 Outliers

In the context of this project, the identification of outliers has been integrated with the application of Cook's distance as an essential analytical tool. Cook's distance is a statistical metric employed in regression analysis to assess the impact of individual data points on the model's estimated coefficients. It quantifies the influence of each observation by measuring the change in predicted values when that observation is excluded from the analysis. The implemented action involves the exclusion of outlier rows due to their tendency to introduce undesired and noteworthy associations in the dataset, as shown in Figure 5.

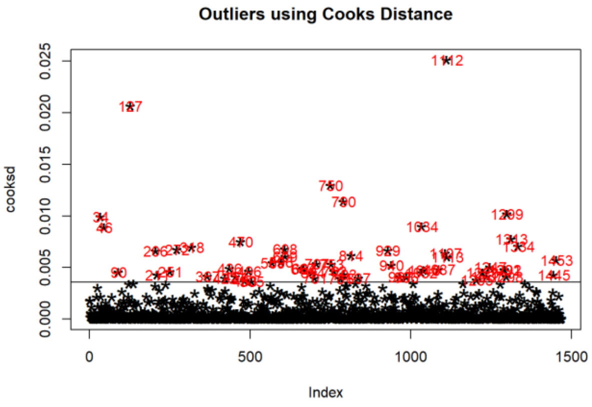


Figure 5

4. Supervised Learning

Supervised learning is a form of machine learning where a defined dataset, known as the training dataset, is employed to make predictions. With a clean dataset and initial insights established, supervised machine-learning models were trained to predict employee attrition. The following section details the models' approaches, parameters, and performance.

4.1 Logistic Regression

Logistic regression is an appropriate baseline classifier for a binary target like employee attrition. An initial model was trained using all remaining predictors. Its main goal is to figure out how the input data changes (independent variables) into forecasts for a result (dependent variable). In this situation, the model guesses if a certain piece of data will be classed as "1." Similar to linear regression, logistic regression helps us make predictions. In the first use of logistic regression, by using all possible predictors, the model showed very good results. In detail, it reached an accuracy score of 0.858 and a ROC_AUC score of 0.791. These numbers give us a clear understanding of how well the model can guess results and its overall power to separate things.

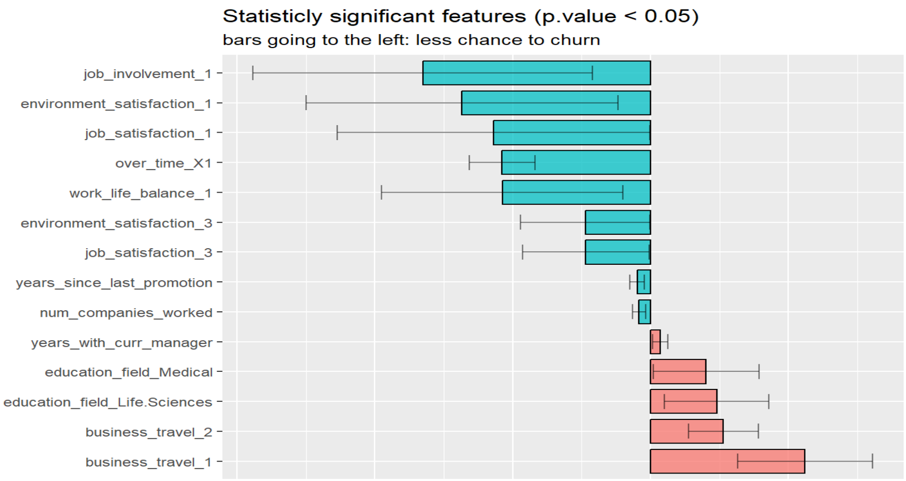


Figure 6

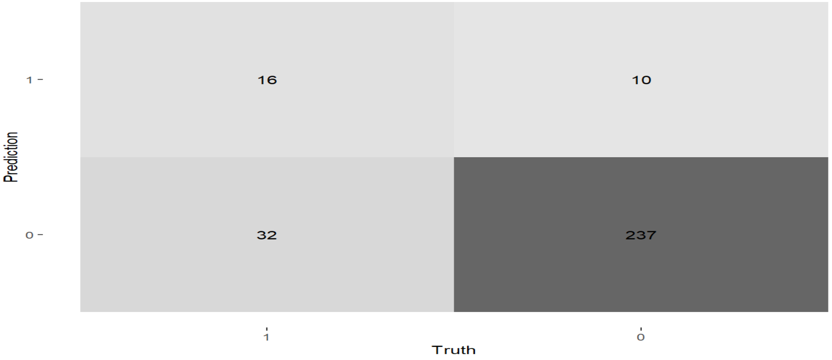


Figure 7

4.2 Lasso Reggresion

Lasso regression is a widely used method in statistical modeling and machine learning for estimating variable connections and making projections. It aims to strike a balance between model simplicity and accuracy, incorporating a shrinkage approach. Pulling data values towards a central point is what is referred to Shrinkage. The lasso procedure fosters the development of simple and sparse models and is particularly effective in cases of high correlation among variables or when automating aspects of model selection, such as parameter elimination.

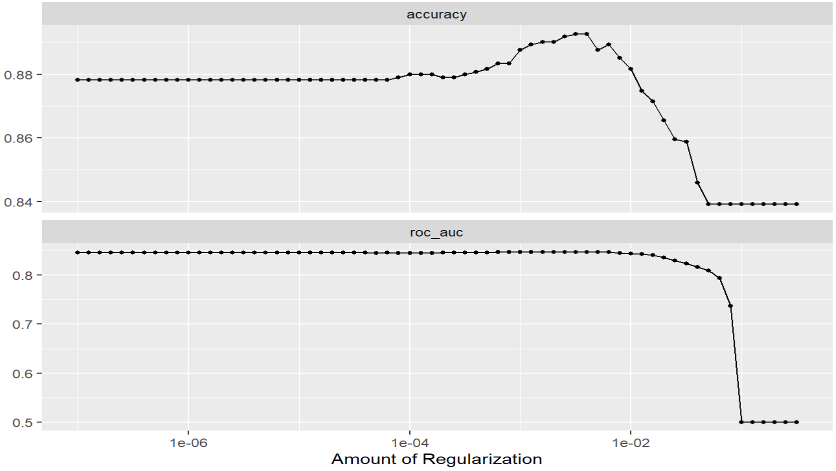


Figure 8

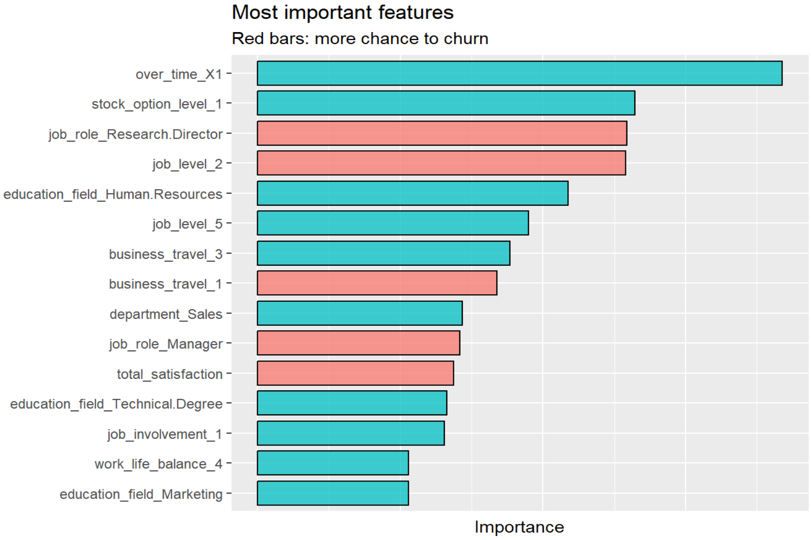


Figure 9

Applying Lasso yielded an accuracy score of 0.851 and an AUC score of 0.791. Performance has improved relative to preceding models, as depicted in the figures (Figures 8 and 9).

4.3 Random Forest

Assembling multiple decision trees with random forests can improve predictive performance. A grid search found ideal parameters of 500 estimators, a maximum depth of 5, and a minimum leaf size of 50. With these hyperparameters, the random forest classifier achieved an accuracy of 0.851 and AUC of 0.75. The random forest model combines the predictions of the estimators to produce a more accurate prediction.

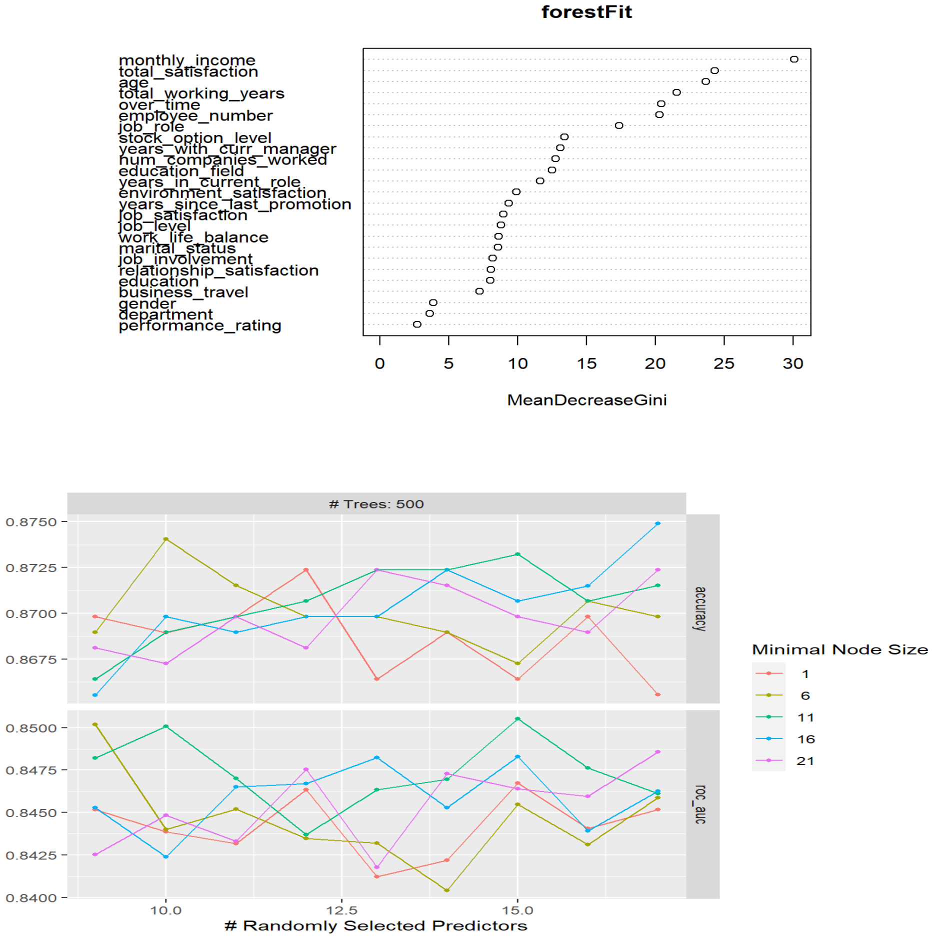


Figure 10

Anticipated attrition outcomes were derived from the testing set. As delineated in Figure 10, the Random Forest model exhibited commendable performance.

4.4 Decision Tree and Pruned Decision Tree

A decision tree is constructed through the iterative partitioning of input data into subsets contingent upon the values of individual attributes. This recursive process mimics a series of inquiries, each dichotomizing the data based on responses. Formulating these inquiries involves the judicious selection of attributes and threshold values, thereby delineating data subsets.

Subsequently, a pruned decision tree model was implemented. Pruning, a prevalent machine learning technique, entails reducing decision tree dimensions by excising segments that contribute marginally to the prediction of target variables. This selective removal is undertaken with the explicit aim of streamlining the model architecture and mitigating overfitting. (figure 11)

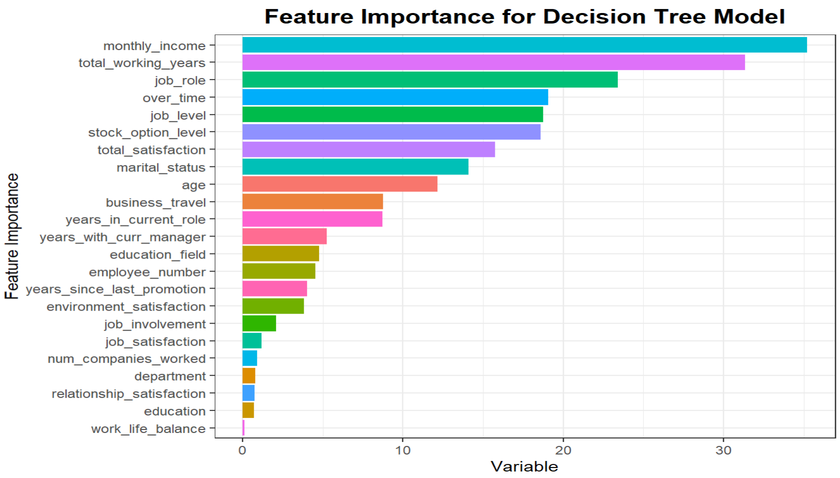


Figure 11

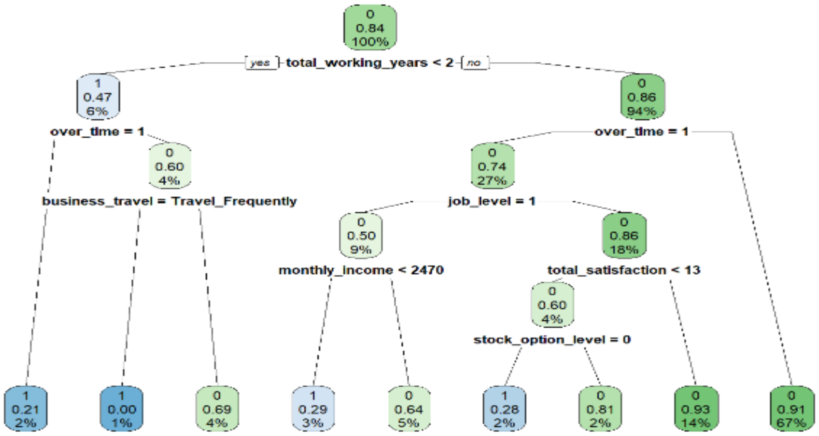


Figure 12

Utilizing the pruned Decision Tree resulted in an accuracy score of 0.833 and an AUC score of 0.565, which is marginally lower than others.

4.5 Support Vector Machine

A Radial Basis Function Support Vector Machine (RBF SVM) was employed as an alternative approach, achieving an accuracy of 0.8606. However, it exhibited a comparatively lower Area Under the Curve (AUC) of 0.573. The disparity between the high accuracy and the suboptimal AUC implies a tendency to misclassify numerous negative cases. As depicted in Figure 13, the SVM displayed inconsistency in its performance metrics.

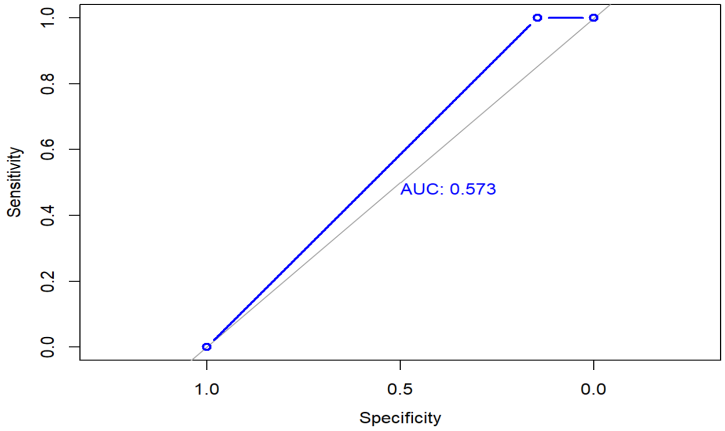


Figure 13

In summary, the predictive modeling surfaced LASSO as this dataset's best-supervised technique for predicting employee attrition. The regularization helps avoid overfitting and improve generalizability.

5. Unsupervised Learning

Unsupervised learning approaches were used in addition to the supervised models to find patterns and groups among personnel. Beyond anticipating results, these strategies assisted in segmenting personnel and enriching comprehension.

5.1. Principal Component Analysis

Principal component analysis condensed the correlated numeric variables into a smaller set of underlying factors. Optimal dimensionality reduction was achieved, retaining four components that explained 78% of the variance. The components represented employee experience at IBM, general work experience, IBM performance, and job involvement. This technique highlighted key employee attributes and relationships in the data, as shown below.

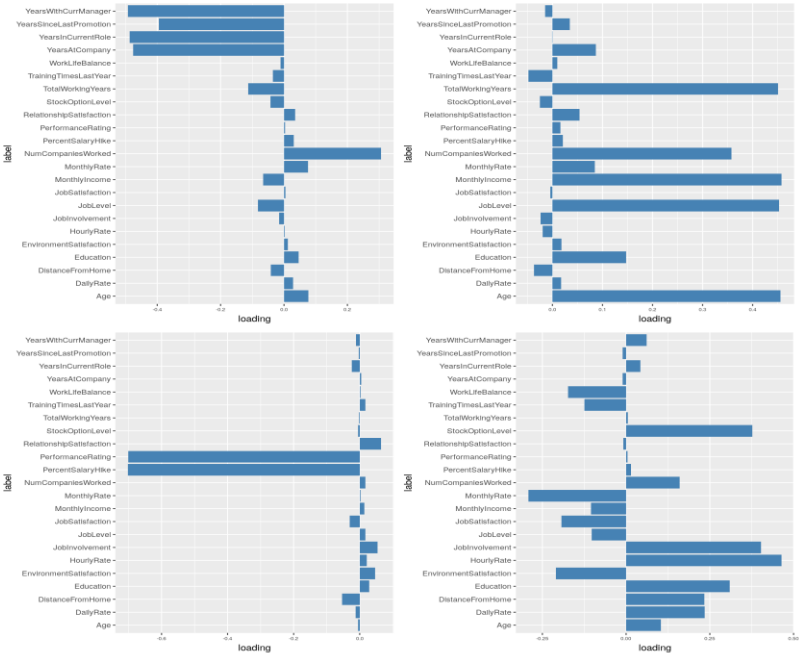


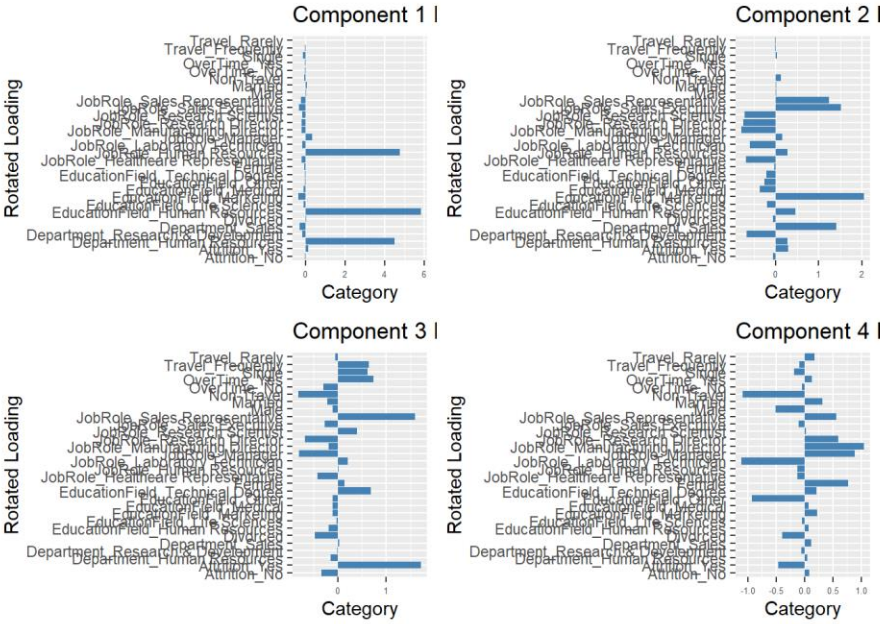
Figure 14

The four rotated main components of the continuous variables are interpreted as follows:

- The first component appears to indicate 'experience at IBM.' This component has strong negative loadings due to years with current management and years at the firm, whereas a handful of companies have significant positive loadings. As a result, people with high ratings on this component are relatively new to IBM.
- The second component represents 'workforce experience,' with significant positive loadings for age, job level, number of businesses worked for, and total working years.
- The third component indicates 'performance at IBM,' with significant negative loadings for performance rating and percent salary increase. Employees scoring high on this characteristic will likely perform poorly at IBM and earn less money.
- 'Job Involvement' might be represented by the fourth component. I picked this interpretation because of the high-loading for-work engagement, stock option level, and hourly rate. However, the monthly rate's significant negative loading is perplexing: how can monthly and hourly rates have opposing loadings?

5.2 Multiple Corresponding Analysis

MCA (Multiple Correspondence Analysis) is a data analytical technique tailored for nominal categorical data, aiming to elucidate and portray latent structures within a given dataset and is achieved by representing data points in a low-dimensional Euclidean space. Consequently, MCA functions as the definite data counterpart to principal component analysis. The MCA method is applied to the Burt Matrix of nominal variables in the IBM dataset. Four principal components are estimated for interpretative purposes, and a varimax rotation is employed to enhance clarity in the analysis.



- The first component is categories that indicate an employee works for the HR department.
- The second component is categories that indicate an employee works in the sales department.
- Interestingly, the third component seems relevant to employee attrition. Employee Attrition has strong loadings on this component. In addition, sales representatives have strong loadings on this component, which may indicate this job role has higher attrition than other roles.
- The fourth component is a mixed bag. It represents categories belonging predominantly to higher level positions: high negative weights on non-travel and lab technician and high positive weights on higher up job roles (Director, Manager).

5.3 Gower’s Distance

A critical exploratory analysis involved identifying coherent employee groupings based on multidimensional similarities. Gower distance was computed between all employees to quantify dissimilarity across diverse data types.

Gower’s Distance can be used to measure how different two records are. The records may contain a combination of logical, categorical, numerical, or text data. The distance is always a number between 0 (identical) and 1 (maximally dissimilar).

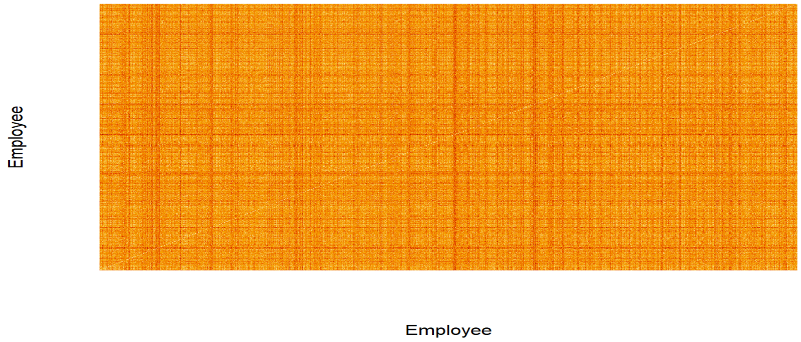


Figure 15

5.4 t-SNE Visualization and Density Clustering

T-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensionality reduction technique well-suited for visualizing high-dimensional data. Using conditional probabilities, it models similarities between data points in high and low dimensions. T-SNE minimizes

Kullback-Leibler divergence between the distributions to project data into two or three dimensions, facilitating visualization of intrinsic structure.

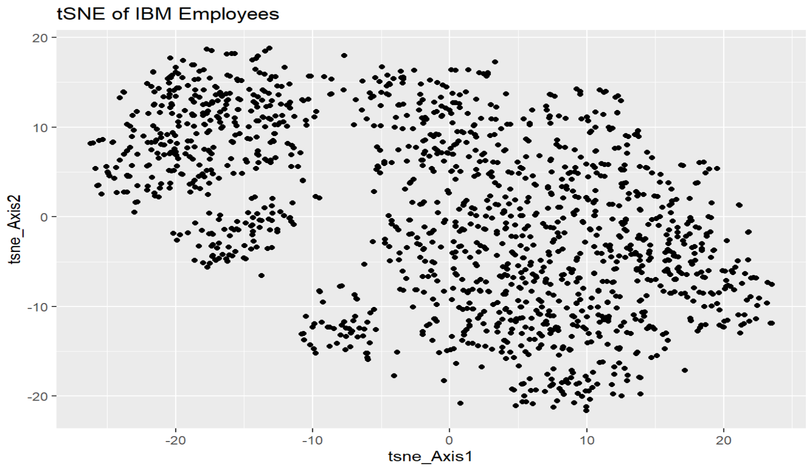


Figure 16

t-SNE projection using the default parameters, as seen, did not delineate the clustering structure. (figure 16)

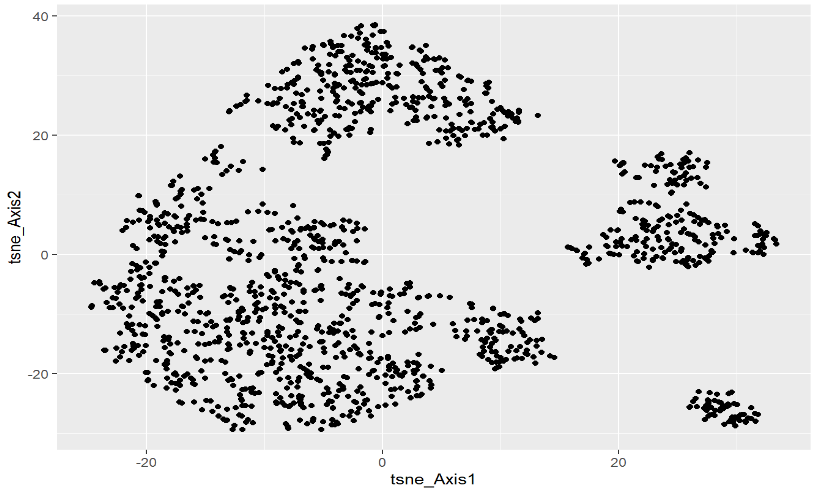


Figure 17

t-SNE projection using Euclidean distances provided enhanced visualization of four distinct clusters, as seen above. (figure 17)

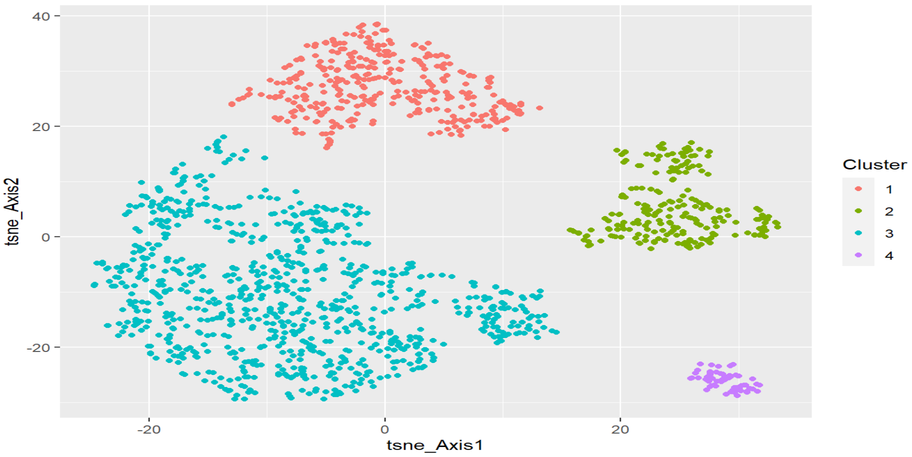


Figure 18

Clusters are grouped by color in the t-SNE solution.

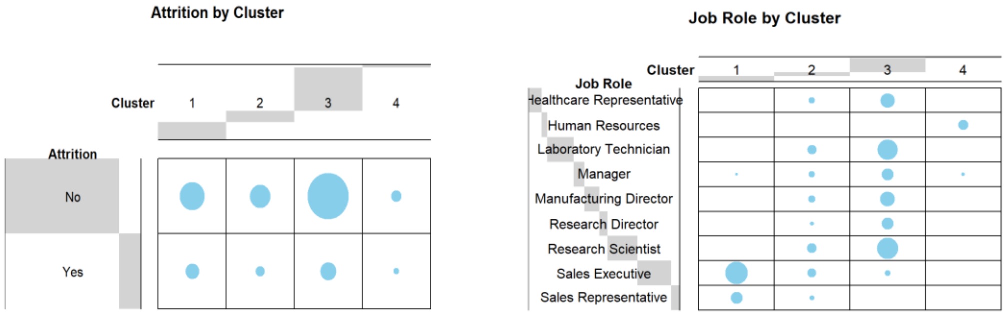
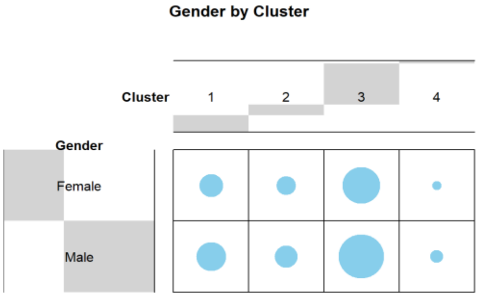


Figure 19



Examining cluster composition by job role provided insight into the characteristics of the four employee segments identified through unsupervised learning.

- Cluster 1 comprised sales roles, namely sales executives and representatives. This cluster exhibited the highest attrition rate at 30%, indicating turnover issues among sales employees.
- Clusters 2 and 3 showed comparable distributions of job functions. However, attrition rates diverged, with Cluster 2 at 18% and Cluster 3 at 14%. Further analysis of cluster features revealed Cluster 3 scored higher across satisfaction, performance, income, and education dimensions.

- Cluster 4 uniquely consisted of human resources employees. Though a smaller segment, it demonstrated an elevated attrition rate of 26%.

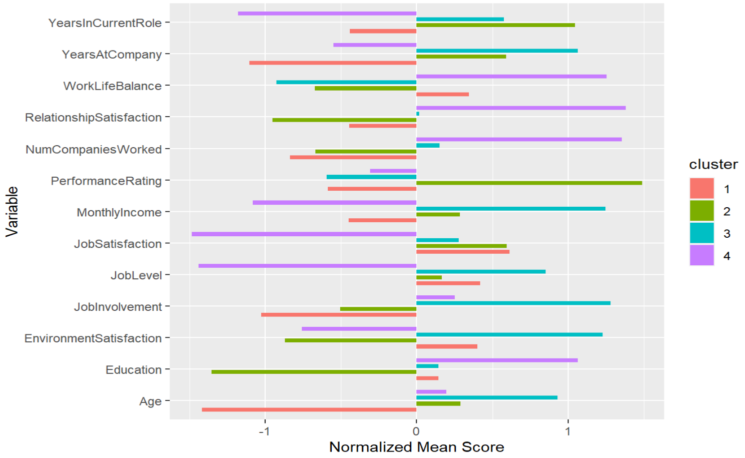


Figure 20

In summary, the data-driven clustering revealed heterogeneity in workforce dynamics across groups. Sales roles experienced pronounced attrition, while more satisfied and skilled employees showed more excellent retention. Targeted interventions could be designed based on the nuances of each employee segment.

6. Conclusion

This study underscores the instrumental role of numerical analysis in comprehending the factors influencing employee attrition. Both methodologies, LASSO and self-sufficient techniques have yielded novel insights into the dynamics within IBM's workforce. The significant findings establish correlations between job cessation and variables such as occupational position, remuneration, and job satisfaction. Predictive models indicate a likelihood of widespread job exits; however, closer examination reveals distinct worker cohorts predisposed to separation from the company. These insights, derived from empirical data, serve as a foundation for informed decision-making and the formulation of effective strategies for employee retention, emphasizing a positive organizational culture.