**CS 590**

**DATA ANALYTICS IN BUSINESS USING R**

**ANALYSIS OF FACTORS AFFECTING EMPLOYEE RETENTION PROJECT**

**FINAL REPORT**

**Submission Date: 5th December 2023**

A graph with a line going up

Description automatically generated

**Group Members:**

1. **Pauline Korukundo**
2. **Martin Vivas**

**TABLE OF CONTENTS**

[**ABSTRACT ii**](#_heading=h.gjdgxs)

[**1**](#_heading=h.30j0zll) **INTRODUCTION 1**

[**2**](#_heading=h.1fob9te) **METHODOLOGY 1**

[*2.1*](#_heading=h.3znysh7) *EXPLORATORY DATA ANALYSIS 1*

[*2.2*](#_heading=h.1t3h5sf) *DESCRIPTIVE ANALYTICS 4*

[*2.3*](#_heading=h.4d34og8) *PREDICTIVE ANALYTICS 7*

[2.3.1 Classification of Departments with Highest Retention Rates 7](#_heading=h.2s8eyo1)

[2.3.2 Predicting Retention Rates 8](#_heading=h.17dp8vu)

[2.3.3 Employment Status 9](#_heading=h.3rdcrjn)

[**3**](#_heading=h.26in1rg) **RESULTS 9**

[*3.1*](#_heading=h.lnxbz9) *Retention Rate 9*

[**4**](#_heading=h.35nkun2) **DISCUSSION 9**

[*4.1*](#_heading=h.1ksv4uv) *Which Employees are Likely to Leave the Company 9*

[*4.2*](#_heading=h.44sinio) *Which Departments Have the Highest Retention Rates 10*

[*4.3*](#_heading=h.2jxsxqh) *Factors Influencing Retention 10*

[*4.4*](#_heading=h.z337ya) *Predicting Retention Rates 10*

[*4.5*](#_heading=h.3j2qqm3) *Classification of Employees with the Highest Retention Rates 10*

[*4.6*](#_heading=h.1y810tw) *Limitations 10*

[**5**](#_heading=h.4i7ojhp) **CONCLUSION 11**

[**6**](#_heading=h.2xcytpi) **REFERENCES 12**

# ABSTRACT

A skilled workforce ensures that a company operates smoothly (Bongale, Dharrao and Urolagin). To maintain this workforce, talent acquisition and management professionals would greatly benefit from understanding the factors that keep employees within an organization. While several techniques exist, data analytics techniques such as machine learning have undergone extensive research that can be leveraged to analyse and understand key factors influencing employee retention.

As such, this project set out to study these factors and provide tools that can be used to understand these factors, determine business units with high retention rates, predict retention rates and classify employees. With these insights, human resource experts can reinforce departments or job roles that are performing well, strengthen those at risk and enhance the overall work environment. The predictive model employed in determining retention rates for subsequent periods can be improved further and applied to a specific employee. As such, this knowledge can be used to foresee employees at risk of leaving the company and those that will stay with the company.

Prior hypothesis such as senior-level positions being associated with employee retention were confirmed by our analysis. Additionally, the younger and more energetic employees were found in operationally intensive job roles such as processed foods which they did not keep for long periods of time, while the older employees were retained for longer periods of time with the same company.

Furthermore, the younger employees aged 40 and below were prone to involuntary exits, indicative of a trend where individuals within this demographic seek diverse experiences and career paths.

This work therefore provides an account of the process followed to prepare the data, gain an understanding of the employee trends within the company, and derive insights from the trends and clusters.

# INTRODUCTION

The Analysis of Factors Affecting Employee Retention investigated employee data over ten years. The data used is the *MFGJOYearTerminationData* dataset from Kaggle, which contains 18 variables and 49,653 observations on employee data, including their demographics, how long they have worked, those that are still employed, and for terminated employees it shows the varied reasons for termination.

The dataset "Employee Attrition" was obtained from Kaggle under a Creative Commons license CCO: Public Domain, permitting its usage in our study. This fictional data facilitates researchers to explore the underlying factors that affect employee retention.

The dataset contains information related to employee demographics, length of services, job role and department, first employment data and whether they are still employed. There are 49,653 observations and 18 variables within the dataset were used to answer the following research questions identified at the start of the project:

1. Can we accurately predict which employees are likely to leave the organization?
2. What are the key factors contributing to employee retention within the organization?
3. Do specific departments or job titles experience higher retention rates?
4. Which employees are at the highest risk of leaving a company?
5. What actionable strategies can HR professionals implement to reduce attrition rates?

The research, guided by the works of Bongale, Dharrao, and Urolagin, embarks on a comprehensive analysis to provide valuable tools for talent acquisition and management professionals. The objective is not only to understand the determinants of employee retention but also to develop predictive models capable of forecasting retention rates, classifying employees, and identifying business units with commendable retention rates.

Therefore, the project seeks to find key factors that drive employee retention such that human resource or talent managers can amplify these factors for the efficiency of their organizations.

# METHODOLOGY

## EXPLORATORY DATA ANALYSIS

This section discusses and thoroughly examines the employee data in preparation for further analysis. The data tidying principles were followed to facilitate visualization, manipulation, and analysis. This involved removal of unnecessary and duplicated data, defining new variables, changing variable data types.

There were no missing values in the data; however, some variables were removed since they were not necessary to answer the project's research questions. These variables include gender short (M or F) and gender full (Male, Female) reduced to one variable, birth date removed since the age was already computed, termination date and record date since length of service and STATUS YEAR were also given.

Other variables were renamed or introduced to the data to derive key insights on employee retention. An extensive exploratory data analysis was then carried out to understand the data further to get quick insights on the distribution and relationships between the different variables.

1. **Age and Gender**

A quick glance at the age, length of service, departments and job titles shows the youngest employees being 19 while the oldest are 65, and a median age of 42. The active employee age range has a bimodal distribution. However, most active employees in their twenties have been working for less than 10 years. Those over 40 have been working for longer years i.e. 15 to 26 years.

A graph of a number of employees below

Description automatically generated

*Figure 1: Departments with Active Employees*

The company has slightly more females than males. The youngest workforce has a mixture of males and females in the customer service and produce departments. The customer service department has the highest number of employees under 40 years.

The oldest employees are female and working in the Diary and Meats departments with tenures of up to 13 years. Generally, this group of employees works in Produce, Meats and Bakery and very few in store Management.

1. **Job Role and Functional Unit**

*Figure 1* shows that employees working at the Head Office business unit have longer tenures, with the average being 19 years, while those at the Stores business unit work for an average of 10 years and at most 15 years.

A diagram of a business unit

Description automatically generated

*Figure 2: Length of Service vs. Business Unit*

The Meats department has the highest number of employees above the age 40, followed by the Produce department. Since age is a major factor in employee retention, these departments will most likely keep most of their employees.

1. **Employee Terminations**

The terminated employees in their 40s were laid off, while those that resigned were mostly in their 30s with some outliers in their 60s, and others went into retirement at a mean age of 65.

A diagram of a number of different colored squares

Description automatically generated with medium confidence

*Figure 3: Termination Reasons*

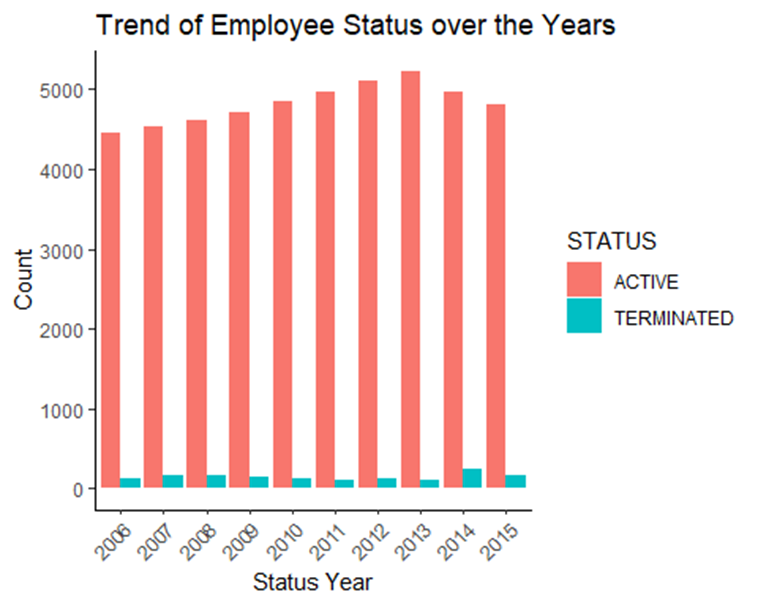
The table below shows a slight rise in active employees in 2013, with a subsequent return to normal levels by 2015.

*Table 1: Trends in Employee Status over the Years*

|  |  |  |  |
| --- | --- | --- | --- |
| **SN** | **Year** | **Active** | **Terminated** |
|  | 2006 | 4445 | 134 |
|  | 2007 | 4521 | 162 |
|  | 2008 | 4603 | 164 |
|  | 2009 | 4710 | 142 |
|  | 2010 | 4840 | 123 |
|  | 2011 | 4972 | 110 |
|  | 2012 | 5101 | 130 |
|  | 2013 | 5215 | 105 |
|  | 2014 | 4962 | 253 |
|  | 2015 | 4799 | 162 |

We further examined the reasons for employee termination (layoff, resignation, and retirement). To gain deeper insights into the trends observed in 2013 and 2014, we created a stacked bar plot showcasing the distribution of the reasons for termination across these years.

The number of active employees increased gradually with a slight drop in 2013. The company recorded the highest terminations during this year.



*Figure 4: Trend in Employee Status over the Years*

## DESCRIPTIVE ANALYTICS

The departments were combined into Operational Units to get a better understanding of their retention and termination rates.

A graph of different types of data

Description automatically generated with medium confidence

*Figure 5: Trends in Retention Rates by Functional Units*

On the other hand, the operations unit started off with a high termination rate which dropped to zero towards 2015. The termination rate of the HR unit is also rising. Management has not lost many of its employees.

A graph of different types of data

Description automatically generated with medium confidence

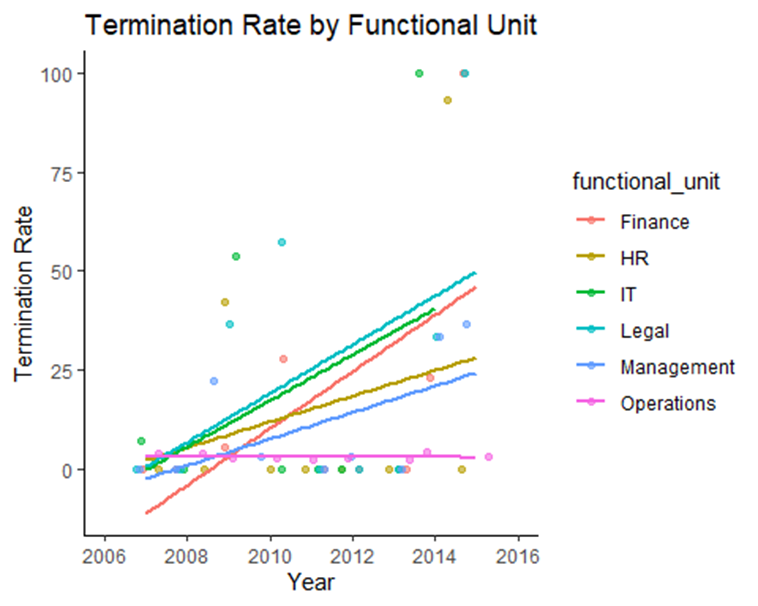
*Figure 6: Trends in Termination Rates across Functional Units*

A linear model shows different behavior of these functional units. The Operations Functional Unit has had a stable retention rate over the years while that of the Finance department is gradually dropping. In terms of departments, Produce, Meats and Diary have the highest retention rates.

A graph of different colored lines

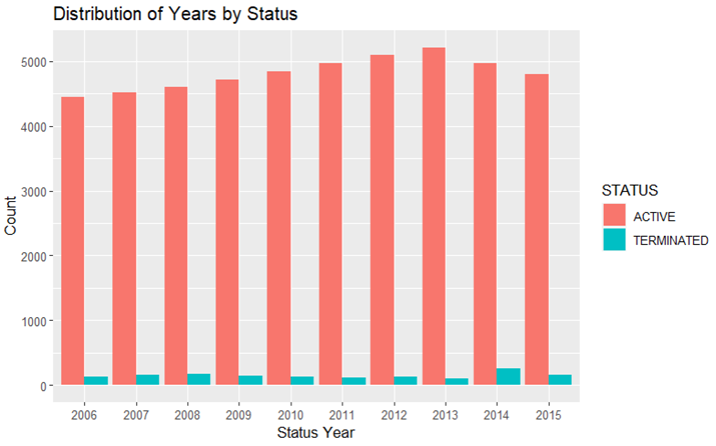
Description automatically generated

*Figure 7: Functional Unit Retention Rates over the Years*



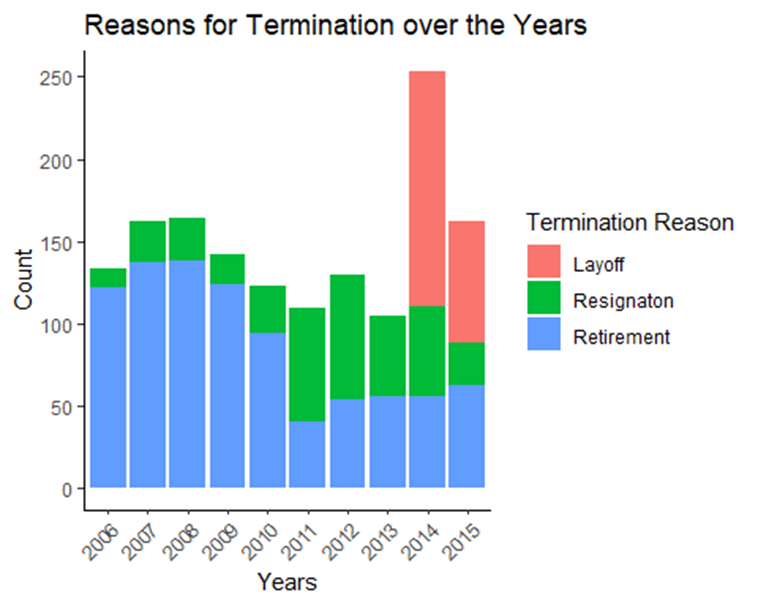
*Figure 8: Functional Unit Termination Rates over the Years*

To investigate whether our company has experienced higher employee attrition in recent years compared to earlier ones, we initially generated a bar plot representing the distribution of employee status over the years. The bar plot in Fig 9. that follows shows a visualization of this trend.



*Figure 9: Employee Status over the Years*

We further delved into this analysis by examining the reasons for employee termination (layoff, resignation, and retirement). To gain deeper insights into the trends observed in 2013 and 2014, we created a stacked bar plot showcasing the distribution of the reasons for termination across these years. Please find the stacked bar plot attached for reference.



*Figure 10: Termination Reasons over the Years*

Our analysis revealed intriguing patterns. The number of retirements gradually declined since 2009, reaching its lowest point in 2011. We can also see a high number of layoffs in 2014 and a small number of layoffs in 2015. After seeing this information, we would think that the number of active employees in the year 2016 will be similar to 2015, around 4800.

To confirm this, we created a time series ETS model was created, based on the year and STATUS. We printed the forecast values obtained from the model with the prediction values.



We can say our guess was correct based on the information printed. The point forecast is 4799, which could be rounded to 4800. We also forecast with 80% certainty that in 2016 the number of active employees will be between 4613 and 4985. With 95% certainty, we can forecast that the value is going to be between 4515 and 5083.

## PREDICTIVE ANALYTICS

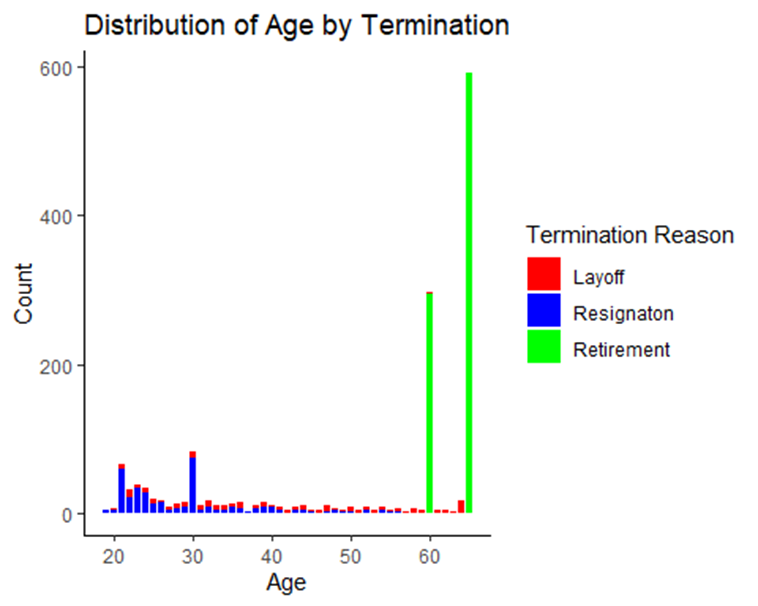
### Factors influencing the most to determine Active or Terminated employees

We aimed to determine the key factors influencing employee retention and termination. Utilizing a random forest model, we considered various factors, finding the most important ones. Upon training and testing the model, we proceeded to assess the relative importance of each factor, and the resulting analysis culminated in the generation of an informative plot. Refer to the attached plot for a visual representation.



*Figure 10: Key Factors using random forest*

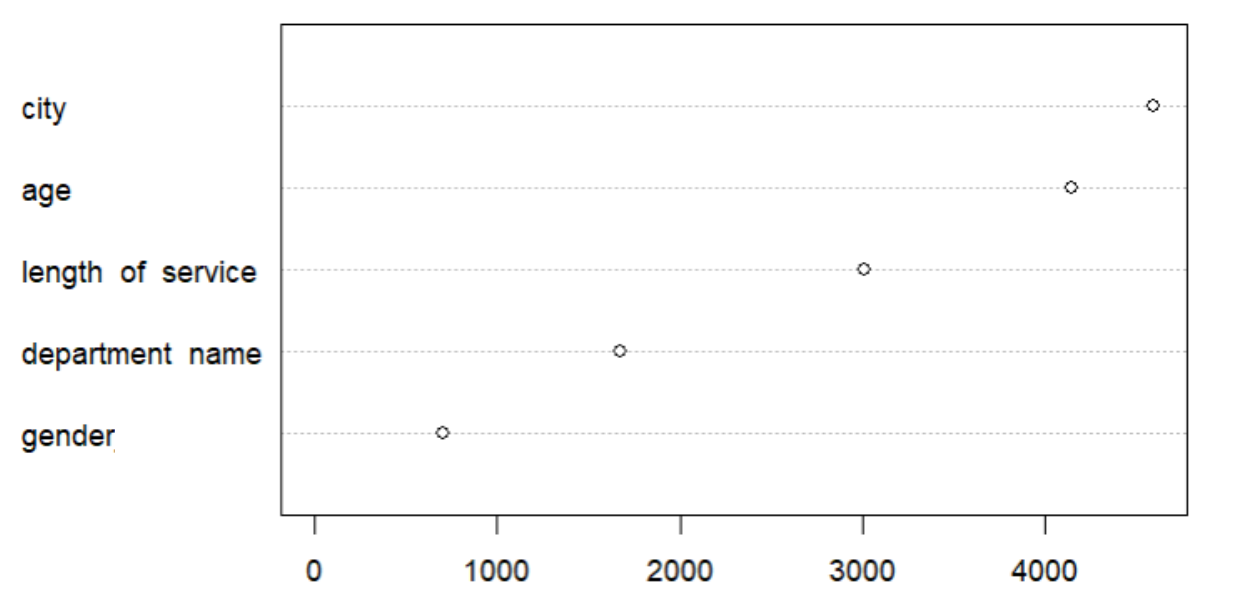
The analysis revealed that age stands out as the most crucial factor. To understand why age holds such significance, we conducted further investigation as depicted in Fig.11—a plot analyzing the reasons for terminations based on age.



*Fig.11 Analysis of Termination Reasons by Age*

Our examination revealed that for employees older than 60 years, retirement emerges as the predominant termination reason. Given that most employees retire after the age of 60, with every employee retiring by age 65, predicting retirement becomes straightforward. To refine our focus on termination by layoff and resignation, we excluded employees aged 60 and older.

After training and testing the model under these refined parameters, we reassessed the importance of each factor, resulting in a comprehensive plot represented in Fig.12.

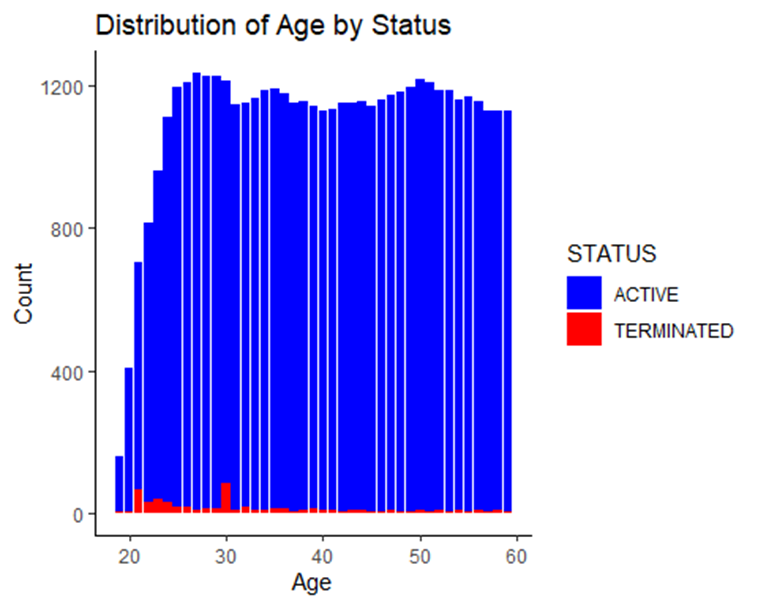


*Fig.12: Key Factors using random forest for employees under 60 years of age*

We found out that "city\_name" emerged as the most crucial factor, surpassing "age" in importance. Despite excluding retirements from the analysis, age continues to play a significant role in predicting employee terminations.

In addition, we created another model that utilized the "region" column, a summary of "city\_name" grouping cities into fewer regions, and the "functional\_unit" column, summarizing "department\_name" by grouping departments into fewer units, but they notably exhibited lower importance, and the models had lower accuracy.

Further investigation into the importance of "city\_name" reveals that certain cities exhibit significantly higher counts of both ACTIVE and TERMINATED employees. Moreover, some cities, despite having fewer employees, demonstrate termination above average percentages.



*Fig.13: Distribution of Age by Status*

In our examination of the distribution of ACTIVE and TERMINATED employees based on "age," the second most influential factor, we uncover interesting patterns. Employees in their early 20s and at the age of 30 are identified as more susceptible to termination, even in the presence of a smaller workforce within this age range.

As we delve deeper into predicting individual terminations, we encounter challenges. The model struggles to accurately forecast the termination of individual employees. However, when applied at a group level, the model proves more effective in predicting overall termination and retention rates. While this may not provide granular insights into specific employees, it remains a valuable tool for identifying factors influencing employee terminations and determining which groups of employees are more likely to face termination.



### Classification of Departments with Highest Retention Rates

#### Model Selection

The model selection process involved grouping departments with retention rates above 65% as High and those below as having Low Retention Rates. These groups were then added to the data and the data prepared for analysis.

Three approaches were followed, one being randomly analyzing the accuracy of different K Nearest Neighbors in classifying employees into these groups. Accuracies of 93 - 94% were observed for K-10, 30 and 50. With this method, the best K-values was 30.

*Table 2: Accuracies of Different K-values*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **K-value** | K3 | K5 | K10 | K30 | K50 |
| **Accuracy** | 0.7827160 | 0.7814815 | 0.7882716 | 0.8024691 | 0.8067901 |

To ascertain which of these is the best, another step was taken using cross validation. Different K-values were evaluated over k-folds and predictions of employee groupings computed. The best K-value was selected as K=80 based on the cross validation which is then used to classify the employees into groups with the highest and lowest retention rates

*Table 3: Confusion Matrix for CV*

|  |  |  |
| --- | --- | --- |
| **Predicted** | **Actual** | |
| High | Low |
| 2520 | 1260 |

The model predicts both classes accurately with an accuracy of 100%. However, these results were further confirmed by applying the AUC\_ROC Curve “Area Under the Curve” (AUC) of the “Receiver Operating Characteristic” (ROC).

The area under the curve was 0.88 which is close to zero thereby indicating the KNN model’s ability to correctly distinguish between and predict both classes. Therefore, we can conclude that the predictions, very high accuracies, and sensitivity is not a random behavior of the model.

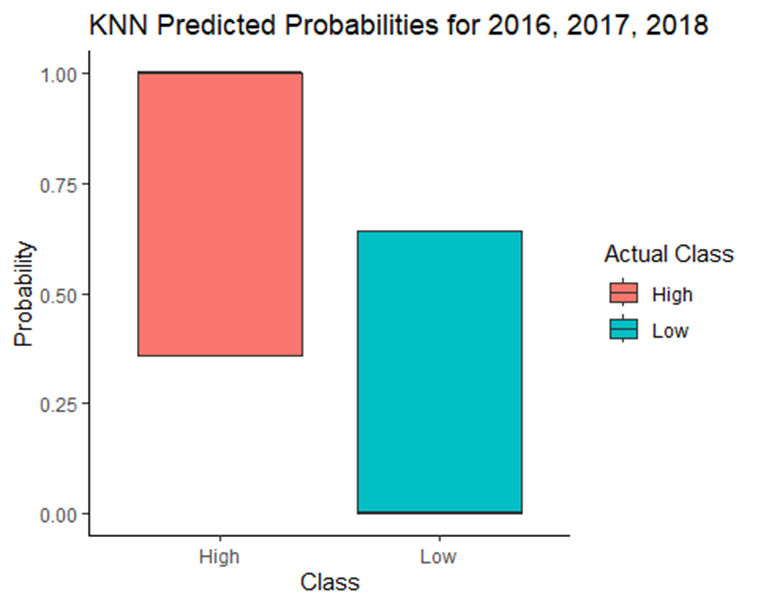
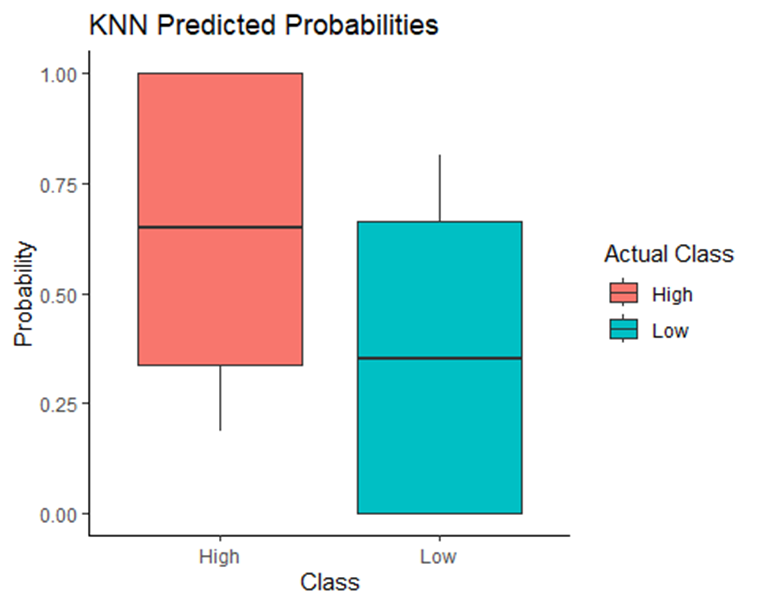
A graph with a blue line

Description automatically generated

*Figure 9: AUC\_ROC Curve*

#### Employee Classification into Group of Highest and Lowest Retention Rates

The KNN model was therefore used to classify the employees into groups of highest and lowest retention rates. Additionally, the model was used to predict these classes for the subsequent year for which no data was provided.



*Figure 10: Classification of Employees with High and Low Retention Rates (right) Prediction for 2016-2018*

### Predicting Retention Rates

The prediction of retention rate followed a rigorous model selection process using the AICc (Akaike Information Criterion with a correction for small sample sizes) evaluation. Criterion selected the model with age, employment period and the year under consideration as predictors of retention rate with the lowest AICc of 607.57. The model with the functional unit/ department was eliminated. Therefore, this Logistic Regression model was developed and used to forecast the retention rate over time. The prediction shows a reduction in the Retention Rates over the 10-year period.

A graph with a line going up

Description automatically generated

*Figure 11: Predicted Retention Rates over Time*

The model can therefore be used to predict the Retention Rate for a specific year. To get an idea for 2016 where no existing data on the employees was available, a hypothesis of expected employee age and employment was made. The model was then given this data to predict the retention rates for that year.

# RESULTS

## Retention Rate

The logistic regression model predicts 49.43% as the retention rate for the subsequent year 2016 which is low across the entire company. This indicates the need for immediate attention to counter this low retention.

The prediction shows that all departments across the company will have a high retention rate. However, without test data, it is hard to verify these results.

Additionally, the model accurately classifies the employees with the highest and lowest retention rates for the subsequent years.

# DISCUSSION

## Which Employees are Likely to Leave the Company

Employees in their forties and below are more likely to leave the company involuntarily. Within this age bracket, employees often find themselves in a phase of exploration, grappling with decisions related to ideal workplace, residence, and a broader exploration of life's diverse options. Consequently, this demographic might display lower levels of workplace productivity or seek varied experiences as part of their journey to discern personal preferences. However, metrics such as employee performance or indicators of performance could reinforce this claim.

## Which Departments Have the Highest Retention Rates

Employees in the Finance, HR and IT functional units are at risk of leaving the company since these units have high termination rates.

Since the Operations Functional Unit (departments of Produce, Meats and Diary) and Management have had stable retention rates over the years, non-statistical factors in these units can be examined further for application to departments at risk.

## Factors Influencing Retention

Different models were used to understand the biggest factors affecting retention. These include:

1. Application of random forest model revealed age as the most significant factor in determining employment status.
2. Further analysis of retention rate, which is a statistical factor, revealed factors of age, period of employment, and department as predictors in a liner regression model.

## Predicting Retention Rates

The retention rates over the 10-year period generally reduced across the entire company. If this data is monitored closely, it is possible to apply mitigating factors to realize improvements.

With this model, in place, an estimate of employee age, and current employment duration was made and used to predict retention rate of employees that fit that description. The model predicted a low retention rate of 49.3% for the subsequent year 2016. To get this prediction, it must be noted that the departments with a Retention rate of 100% were eliminated from the dataset since there were many, causing an imbalance in the data.

While Logistic Regression is a powerful tool that can be used to determine which employees are likely to remain with an organization, it is limited when the variables in the data are not based on a linear relationship.

## Classification of Employees with the Highest Retention Rates

The applied KNN model shows that all departments across the company will have a high retention rate and has a high probability of predicting an employee in this class with the given dataset.

For the subsequent periods where synthetic data was used, the model performs well in predicting these two employee segments. However, without test data for the synthetic data used for the 2016 – 2018 period, it is hard to verify these results.

## Limitations

A key challenge was the low statistical data that could be used with models such as linear regression to make satisfactory predictions. Moreso, parameters such as salary, financial background, family dynamics such as number of children, and work hours can improve the analysis and determination of employee retention.

Additionally, it must be noted that the data was imbalanced with more active employees compared to terminated employees. This has a potential impact on the quality of the predictions and classifications made. Therefore, these models can be further improved with a more balanced dataset or with techniques such as bootstrap resampling to create a generalized model that will perform well on different data.

The insights from this work are applicable to the employees with similar factors as those present in the MFG dataset. While employees are risk can be determined using these models, individuals with similar characteristics should not be discriminated against during recruitment, but rather supported to foster their retention.

# CONCLUSION

By identifying the underlying factors contributing to retention and deriving insights, HR professionals can therefore take proactive measures such as creating personalized development plans or retention strategies for a more stable and productive work environment. This is important because high retention rates can lead to significant cost savings and increased productivity. (Taylor, El-Rayes and Smith).

Descriptive analytics as well as predictive models were successfully applied in this work such as linear regression and K-means classification models to provide insights on employee retention. The project showcases the power of data analytics and predictive modeling in predicting retention rates and the factors driving these rates.

The findings show an employee’s age, and their employment period can be used to predict the retention rate for a given department or functional unit. The predictive models can be refined further and applied to an individual employee basis.

The work also uncovers patterns in younger employees working in specific roles and tending to have shorter tenures while the older employees tend to work for extended periods. Moreso, the analysis affirms prior hypothesis such as the correlation between senior-level positions and higher employment retention.

The derived insights and clusters can be adopted for strategic decision-making in talent acquisition, retention, and overall human resource management. The fusion of business analytics and workforce dynamics underscores the potential for organizations to proactively profile workforce strategies, ensuring sustained success and adaptability in a dynamic business landscape.

# REFERENCES

1. Bongale, A. M., D. Dharrao and S. Urolagin. "Exploratory Data Analysis and Classification of Employee Retention based on Logistic Regression Model." *9th International Conference on Advanced Computing and Communication Systems (ICACCS)* 2023 : 1929-1933.
2. Taylor, S., N. El-Rayes and M. Smith. "An Explicative and Predictive Study of Employee Attrition Using Tree-based Models." *Proceedings of the 53rd Hawaii International Conference on System Sciences*. 2020.