PREDICTING HUMAN ATTRITION USING MACHINE LEARNING

BDA 650 CAPSTONE PROJECT REPORT

Abhishek Gundala
Sajjiynie Suraweera
Academic Advisor: Prof Nikanor Volkov, Ph.D, CVA,MAFF

Introduction

In the context of today's rapidly evolving workplace, employee retention has become a crucial concern for organizations. For a company like IBM, with its global footprint and diverse workforce, this challenge is particularly significant. Understanding the factors influencing employee turnover, or attrition, is imperative for maintaining the company's intellectual capital, operational efficiency, and overall competitive advantage. The business question specifically seeks to uncover:

Factors Leading to Attrition: What are the primary reasons why employees choose to leave IBM? This includes both tangible factors like compensation, work-life balance, and career growth, as well as intangible aspects such as workplace culture and employee engagement.

Differences Across Departments: Are there department-specific trends in attrition? For instance, do certain departments have consistently higher attrition rates, and what drives this disparity?

Employee Demographics and Attrition: How do employee demographics (e.g., age, tenure, job level, and location) influence attrition rates? Understanding this helps in identifying high-risk groups and tailoring retention strategies accordingly.

Impact of Leadership and Management Practices: How does leadership style and management affect attrition rates? This involves evaluating how factors like communication, feedback, and leadership development influence employee decisions.

External Influences: What external factors, such as industry trends or economic conditions, impact employee turnover at IBM? This provides context for attrition beyond internal factors.

Effectiveness of Current Retention Strategies: How effective are IBM's current

strategies in retaining employees? This involves evaluating existing programs, policies, and initiatives intended to improve retention.

Background

Employee attrition has been a persistent challenge across various industries, and IBM is no exception. Given the rapid pace of technological advancements and competitive job market, retaining skilled employees has become increasingly difficult. Historically, IBM has experienced fluctuations in employee numbers, which have impacted productivity and morale. With a workforce known for its technical expertise and innovation, IBM relies heavily on its employees' knowledge and skills. High attrition can lead to a loss of intellectual capital, increased hiring and training costs, and disruption in project continuity.

Importance

Understanding and addressing employee attrition is critical for several reasons:

Cost Efficiency: Hiring and training new employees are costly endeavors. Retaining current employees saves the company significant resources.

Knowledge Retention: Experienced employees possess valuable institutional knowledge. Their departure can lead to a knowledge gap, affecting productivity and innovation.

Employee Morale: High attrition rates can negatively impact the morale of remaining employees, potentially leading to further turnover.

Business Continuity: Retaining key personnel ensures smooth business operations and project continuity.

Talent Management: By understanding attrition patterns, IBM can improve its talent management strategies, ensuring the right people are in the right roles.

Reputation and Competitiveness: A company known for high employee turnover may struggle to attract top talent, impacting its competitive edge.

By addressing the root causes of attrition, IBM can enhance employee satisfaction, foster a positive work environment, and build a more resilient and committed workforce. The insights derived from this analysis will guide strategic decision-making, helping IBM to implement policies and practices that will support a strong and stable workforce.

Literature Review

Organizational Policies and Employee Satisfaction:

The intertwining of organizational policies and employee satisfaction plays a pivotal role in influencing attrition. ***Merceline Anitha et al. (2023)*** stress the importance of organizational strategies that are closely aligned with employee needs to effectively reduce turnover rates. Similarly, ***Kesavan and Dhivya (2022)*** identify salary, work-life balance, and job dissatisfaction as primary drivers of employee departure, suggesting

enhancements in these areas as critical to improving retention.

Demographic and Socio-Economic Influences:

The study by ***Kesavan and Dhivya (2022)*** also brings to light the significant impact of demographic factors, such as age, gender, and marital status, on attrition rates, underscoring the necessity of personalized retention strategies. ***Norsuhada Mansor, Nor Samsiah Sani, and Mohd Aliff (2021)*** further elaborate on the influence of demographic and socio-economic factors, advocating for a nuanced approach to addressing these variables to mitigate turnover effectively.

Leadership Styles and Work Environment:

The leadership style and the work environment significantly dictate employee turnover. ***Negi (2013)*** discusses how leadership that fosters support and engagement is essential for minimizing turnover, highlighting the critical role of leadership in retention strategies.

The Role of HR Analytics and Machine Learning

Predictive Modeling: The potential of machine learning in predicting employee attrition is exemplified by **Fallucchi et al. (2020)**, who demonstrate the effectiveness of *Gaussian Naïve Bayes classifiers*. This highlights the utility of machine learning techniques in identifying at-risk employees. The study by ***Norsuhada Mansor et al. (2021)*** further explores various algorithms, finding that an *optimized SVM model* yields the highest accuracy, emphasizing the importance of selecting appropriate models based on the data's characteristics.

Integrating HR Analytics

Drawing from "Quantifiably Better," ***VanWieren***'s DATA-INSIGHT-ACTION cycle and the ITEM model advocate for a structured approach to utilizing HR analytics in addressing attrition. By enhancing data quality, as outlined through the Seven C's of data quality, predictive models can become more reliable and actionable

Department-Specific Attrition Insights

Departmental Analysis: Insights from LinkedIn (***Lewis & Soroñgon, 2022***) reveal disparities in turnover rates across departments, with HR roles experiencing the highest rates. This departmental lens emphasizes the necessity for tailored retention strategies, informed by a nuanced understanding of department-specific challenges and factors contributing to attrition.

Data Dictionary

Category	Feature Name	Type of Data	Data Description
Employee Attributes	Age	Continuous,Numeric	The age of the individual employee
	Over18	Categorical, Nominal	(1 = Yes, 2 = No)
	Gender	Categorical, Nominal	Female, Male
	JobRole	Categorical, Nominal	Job Role (1=HR Rep, 2=HR, 3=Lab Technician, 4=Manager, 5=Managing Director, 6=Research Director, 7=Research Scientist, 8=Sales Executive, 9=Sales Representative)
	Education	Categorical, Numeric	Level of education attained (1 = 'Below Collage', 2 = 'College', 3 = 'Bachelor', 4 = 'Master', 5 = 'Doctor')
	OverTime	Categorical, Nominal	Overtime (1 = No, 2 = Yes)
	Department	Nominal	Employee department (HR, R&D, Sales)
	MaritalStatus	Categorical, Nominal	Marital Status (1 = Divorced, 2 = Married, 3 = Single)
	BusinessTravel	Categorical, Nominal	Business travel frequency (No Travel, Travel Frequently, Travel Rarely)
	EducationField	Nominal	Field of education (HR, Life Sciences, Marketing, Medical Sciences, Others, Technical)
	YearsAtCompany	Continuous, Numeric	Total Number of Years at the Company
	TotalWorkinYears	Continuous, Numeric	Total Years Worked
	DistanceFromHome	Continuous, Numeric	The distance from work to home
	TrainingTimesLastYear	Continuous, Numeric	Training sessions last year
	NumCompaniesWorked	Continuous, Numeric	No. of Companies Worked At Previously
Supervisor Survey	JobLevel	Categorical, Numeric	Level of Job (1 to 5)
	PerformanceRating	Categorical, Numeric	Performance Rating
	YearsInCurrentRole	Continuous, Numeric	Years in Current Role
	YearsSinceLastPromotion	Continuous, Numeric	Last Promotion
	YearsWithCurrManager	Continuous, Numeric	Years Spent with Current Manager
Compensation & Benefits	DailyRate	Continuous, Numeric	Salary Level
	HourlyRate	Continuous, Numeric	Hourly Salary
	MonthlyRate	Continuous, Numeric	Monthly Rate
	StockOptionLevel	Categorical, Numeric	Stock Options
	PercentSalaryHike	Continuous, Numeric	Percentage Increase in Salary
Employee Survey	JobSatisfaction	Categorical, Numeric	Satisfaction with the job (1='Low', 2='Medium', 3='High', 4='Very High')
	JobInvolvement	Categorical, Numeric	Job Involvement (1 = 'Low', 2 = 'Medium', 3 ='High', 4 = 'Very High')
	WorkLifeBalance	Categorical, Numeric	Time Spent between Work and Outside (1 'Bad', 4 'Best')
	RelationshipSatisfaction	Categorical, Numeric	Relations Satisfaction (1 = 'Low', 4 'Very High')
	EnvironmentSatisfaction	Categorical, Numeric	Employee satisfaction with the environment (1 = 'Low', 2 = 'Medium', 3 = 'High', 4 = 'Very High')
Operational Data	EmployeeNumber	Continuous, Numeric	Employee ID
	StandardHours	Continuous, Numeric	Standard Hours
	EmployeeCount	Continuous, Numeric	Count of instance

Importance of Classification framework & Data Dictionary

Structured Data Management: It organizes and categorizes large datasets into logical groups, making it easier to manage, analyze, and draw insights from the data. This is particularly useful when dealing with thousands of records and numerous attributes.

Insightful Analysis: It helps break down complex data into digestible sections, allowing analysts to focus on specific areas like employee attributes, supervisor feedback, and compensation details. This specialization enables targeted analysis that can reveal trends and insights within each category.

Improved Decision-Making: By categorizing data, companies can quickly identify problem areas or strengths in their workforce. For example, understanding how compensation impacts retention can inform salary structures, while insight into employee satisfaction can guide management practices.

Consistency and Standardization: A framework ensures consistent data usage and interpretation across the organization. This standardization is vital when comparing data over time or across different departments.

Enhanced Communication: It provides a common language for discussing data among stakeholders, making it easier to communicate findings and recommendations. For instance, referring to specific sections like "Supervisor Survey" or "Compensation & Benefits" streamlines discussions.

Regulatory Compliance: In industries where data privacy is a concern, a classification framework ensures sensitive data is handled appropriately, aligning with data protection regulations.

Predictive Modeling: Classification frameworks are essential for predictive analytics, allowing data scientists to build accurate models by selecting relevant features from well-organized data categories

Data Preprocessing

1. Handling Missing Values

Steps Taken: Missing values were identified using the sapply function, which helps in pinpointing columns that contain missing data. Once identified, appropriate steps such as imputation or removal were applied.

Rationale: Handling missing data is crucial for ensuring the dataset's completeness and robustness. Unaddressed missing values can lead to biased results and errors during analysis.

2. Managing Data Types and Conversion

Step 1: Identification and Conversion of Numeric Variables

The process begins by identifying a set of variables that should be numeric. These variables cover various aspects of employee characteristics and workplace metrics such as age, different rates (daily, hourly, monthly), education levels, job satisfaction levels, and tenure-related metrics.

To ensure accuracy, our script first verifies which of these intended numeric variables are present in the dataset. This preventive measure avoids errors during the conversion process by only adjusting variables that exist in the dataset.

After verifying their presence, the script converts these variables from their current format (which could be character strings or factors) to numeric. This is done individually for each variable to ensure that all specified numeric data is correctly interpreted as numerical values for subsequent analysis.

Step 2: Identification and Conversion of Categorical Variables

Following the handling of numeric variables, the script addresses categorical variables. A list of variables intended to be categorical is defined, including job roles, departments,

and demographic information such as gender and marital status.

The script checks for the presence of these variables within the dataset and then converts the existing ones to factors. This conversion is essential for proper handling of categorical data in R, facilitating correct encoding in statistical models.

Step 3: Verification of Data Conversion

To conclude, the script includes a step to print the structure of the dataset after conversion. This allows for a thorough examination of the data types within the dataset, ensuring each variable has been converted as planned. This output serves as a validation step, confirming that the dataset's structure has been successfully modified.

Through these conversions, the script not only standardizes the dataset but also readies it for robust and error-free statistical analysis.

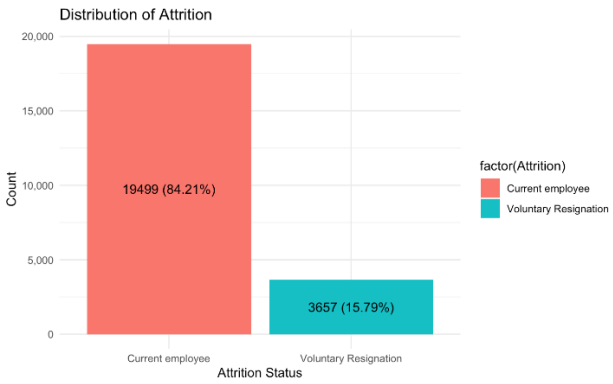
Distribution of Attrition Among Employees

In our analysis of employee attrition, we explored the proportion of current employees compared to those who have voluntarily resigned. The visualization presented in the figure highlights the distribution of attrition within the organization, providing a clear, quantitative insight into employee turnover.

The bar chart illustrates that a significant majority of the workforce, 84.21% (19,499 employees), are currently employed within the company. In contrast, 15.79% (3,657 employees) have chosen to voluntarily resign. This distribution is crucial as it underscores the stability of the employee base and also flags potential areas where employee retention strategies could be applied.

By analyzing these figures, we can gauge the effectiveness of existing employee engagement and retention strategies. The relatively lower percentage of voluntary resignations suggests that while the company maintains a strong retention rate, there remains an opportunity to further reduce turnover through targeted interventions.

This data informs our understanding of the organization's dynamics and can guide HR policies and practices aimed at enhancing job satisfaction and employee loyalty, ultimately contributing to a more stable and productive workforce.



Quick Overview :

Our analysis of the dataset has primarily focused on understanding the factors that influence employee attrition. Through comprehensive data profiling, univariate and bivariate distribution analysis, correlation study, and Principal Component Analysis (PCA), we have gained insights into the underlying dynamics of employee turnover.

Univariate Distribution Analysis

In the univariate distribution analysis, we examined key variables such as the number of companies worked, percent salary hike, relationship satisfaction, and work-life balance among others. The box plots for each variable by attrition status show varying distributions, indicating how different factors are distributed among those who remain with the company versus those who resign voluntarily.

Key observations from this analysis reveal:

Longer tenured employees are less likely to resign, as indicated by the higher years at the company, years in the current role, and years with the current manager for current employees compared to those who resigned.

Employees who have worked with fewer companies tend to stay longer, suggesting that frequent job switchers are more likely to resign.

Factors like stock option level and total working years also show significant differences between the groups, with current employees generally benefiting more from company stock options and having longer working tenures.

Correlation Analysis

The correlation analysis highlights the relationships between various attributes. Notably, attributes like age, total working years, job level, and monthly income show negative correlations with attrition, indicating that as these increase, the likelihood of attrition decreases. Conversely, variables with less significant correlation scores, such as distance from home and number of companies worked, suggest a weaker relationship with attrition.

Principal Component Analysis

PCA was conducted to reduce the dimensionality of our dataset while capturing the maximum variance within the data. This analysis helped in identifying the principal components that account for most of the variance, thus summarizing the data efficiently without much loss of information. The components related to job level, total working years, and age were found to be particularly influential.

Bivariate Distribution Analysis

The bivariate distribution analysis, particularly through box plots segmented by attrition status, provided a deeper understanding of how variables differ between current employees and those who have resigned. This analysis is crucial for pinpointing specific areas that could be targeted to improve employee retention strategies.

Initial Variable Selection for Data Analysis

In our approach to analyzing the dataset, an essential first step involved refining the dataset by selecting only the most relevant variables. This process, often referred to as feature selection, is crucial in simplifying the data analysis, improving model accuracy, and reducing computational complexity.

Rationale for Variable Exclusion

The subset was created by excluding specific variables that were deemed less relevant or redundant for the primary analysis goals. The criteria for exclusion included:

Descriptive Variables: Variables such as Education Desc, EnvironmentSatisfaction Desc, JobInvolvement Desc, JobSatisfaction Desc, RelationshipSatisfaction Desc, and WorkLifeBalance Desc were excluded. These are descriptive counterparts to other numerical or categorical variables that already capture the necessary information in a more analytically useful format.

Identifiers and Indexes: Unique identifiers like EmployeeNumber, Unique Emp ID, Application ID, and Index do not contribute to analytical models as they are specific to individuals and do not hold any intrinsic analytical value.

Constant or Near-Constant Variables: StandardHours was removed as it is a constant (or near-constant) across all employees, providing no variance useful for analysis.

Variables with Limited Impact on Objective: Attrition and Employee Source were excluded in this initial phase to focus on underlying patterns and correlations without the influence of the outcome variable or recruitment channels, which are planned to be analyzed separately.

Implementation

The variable selection was implemented in R using a subset creation approach. This method involves explicitly listing the variables to be excluded, ensuring a clear and manageable dataset moving forward. This step helps in maintaining focus on the variables that provide the most value for subsequent explorative and predictive analyses.

Benefits of This Approach

By streamlining the dataset in this manner, we ensure that our analysis is not only faster but also more robust, avoiding the pitfalls of overfitting or undue influence from less relevant variables. This strategic exclusion of variables allows us to concentrate our computational resources and analytical efforts on the data that are most likely to yield meaningful insights and actionable results.

Correlation Matrix Heatmap Analysis

Overview

We conducted a correlation analysis on selected variables from our dataset to identify relationships that could influence subsequent modeling and hypothesis testing.

Methodology

Using Pearson's correlation coefficient, we analyzed the linear relationships between pairs of numeric variables. Missing values were handled using pairwise complete observations to maximize data use without bias.

Visualization Technique

A heatmap was created using R's ggplot2 package, providing a color-coded representation of the correlation coefficients:

Red indicates strong positive correlations.

Blue indicates strong negative correlations.

White shows no correlation.

The symmetrical layout on the x and y axes allows for quick visual assessment of variable interconnections, highlighting potential collinearity or independent influences.

Key Observations

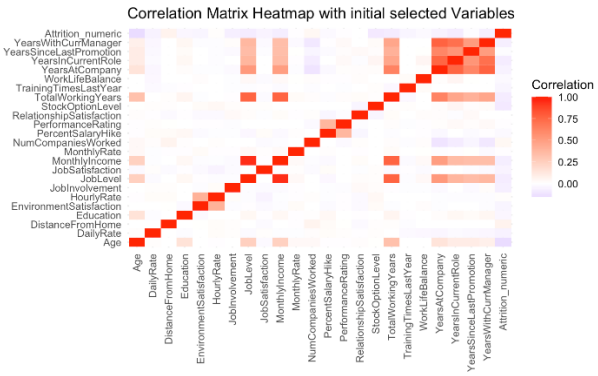
Employee Tenure and Level: Variables such as TotalWorkingYears, YearsAtCompany, and JobLevel are strongly positively correlated, suggesting a typical advancement in career as employees gain tenure.

Less Correlated Factors: Variables like DistanceFromHome and Attrition_numeric show weaker correlations, indicating their influences might stem from a broader set of external factors not directly related to job-specific metrics.

Implications for Further Analysis

The correlations are crucial for refining predictive models and generating hypotheses. High correlation areas might require adjustments in regression models to manage multicollinearity, while less correlated variables offer unique insights into their impacts on outcomes like employee satisfaction and retention.

This analysis not only enhances our understanding of the dataset but also guides our further investigative efforts, ensuring robust and actionable data-driven decision-making.



Converting Variables for Analytical Suitability

In our preprocessing steps, a crucial modification involved converting Attrition_numeric from numeric to a categorical format (factor). This change is vital because Attrition_numeric, despite being represented numerically, inherently describes a category (attrition status) and should be treated as such in any statistical modeling to ensure correct interpretation and analysis.

Multicollinearity Assessment

Multicollinearity within a dataset can severely impact the performance and interpretation of a model. To address this, we performed a multicollinearity check. This involves calculating a correlation matrix for all numeric variables in the dataset, using pairwise complete observations to handle missing data without introducing biases.

We then identified pairs of variables with high correlation, defined by an absolute correlation coefficient greater than 0.7 but not equal to 1 (to exclude self-correlations). High correlations can indicate redundant information, which may not only skew the model results but also affect the stability of the model coefficients.

Key Findings and Actions

The analysis highlighted several pairs of highly correlated variables:

MonthlyIncome and JobLevel exhibited a very high correlation coefficient of approximately 0.943, suggesting that as job levels increase, monthly income tends to increase correspondingly.

TotalWorkingYears and JobLevel, as well as TotalWorkingYears and MonthlyIncome, also showed substantial correlations, indicating that higher job levels and income are associated with longer working tenures.

YearsInCurrentRole and YearsAtCompany, along with YearsWithCurrManager and YearsAtCompany, were highly correlated, implying that longer tenures at the company are often associated with longer durations in the current role and under the same management.

Implications for Model Development

Based on these findings, we need to consider strategies to mitigate the effects of multicollinearity, potentially through variable selection or dimensionality reduction techniques such as Principal Component Analysis (PCA) before proceeding with predictive modeling. This step is critical to enhance model accuracy and ensure robustness in our predictive insights.

Positive Correlation Between Monthly Income and Working Years

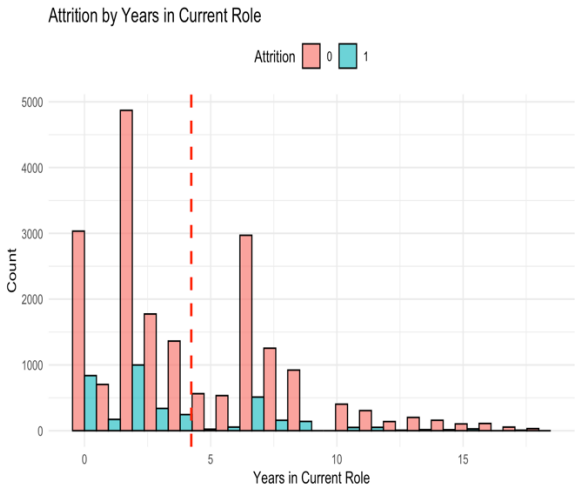
The scatter plot illustrating the relationship between TotalWorkingYears and MonthlyIncome clearly demonstrates a positive correlation. As the number of working years increases, there is a notable rise in monthly income. This trend is smoothed and highlighted by a loess curve, underscoring the progressive increase in income with increased tenure. This visualization not only confirms the expected economic progression but also helps in understanding employee compensation growth over their career span.

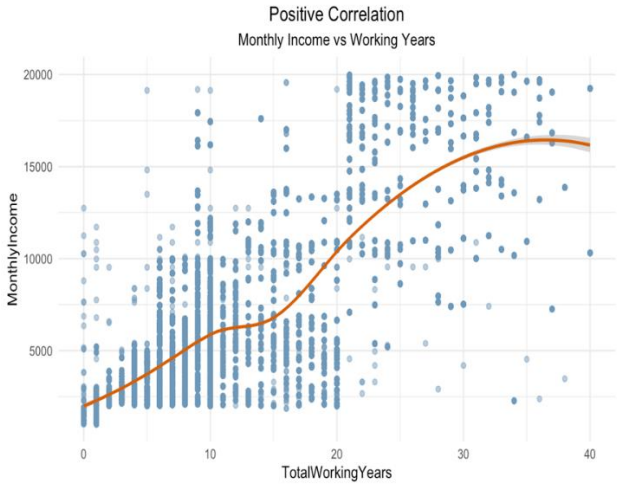
Attrition Trends by Years in Current Role

The histogram for YearsInCurrentRole against Attrition_numeric provides insightful details into employee retention. Most notably, attrition is higher among employees with fewer years in their current roles, with a significant drop as tenure increases. A dashed line representing the average years in current roles delineates the point where attrition rates start to decline significantly. This graph effectively showcases the critical period in employee roles where retention efforts could be most impactful.

Impact of Training on Attrition

Similarly, the histogram comparing TrainingTimesLastYear with Attrition_numeric reveals patterns in training impact on employee turnover. Employees with fewer training sessions last year tend to have higher attrition rates. The average training times are marked by a dashed line, beyond which attrition rates noticeably decrease. This suggests that increased training opportunities could correlate with higher employee retention, emphasizing the potential benefits of investing in employee development programs.





Analysis of Age Distribution by Attrition Status

Overview

To better understand the impact of age on employee retention, we conducted an analysis segmenting the workforce into age groups and examined the attrition rates within each group. This approach helps to

identify demographic trends in attrition and informs targeted retention strategies.

Method

The dataset was enhanced by categorizing employees into age groups (20-30, 30-40, 40-50, and 50-60 years). For each group, we calculated the total count and percentage of employees who stayed with the company versus those who left. This allowed us to observe attrition patterns across different age segments.

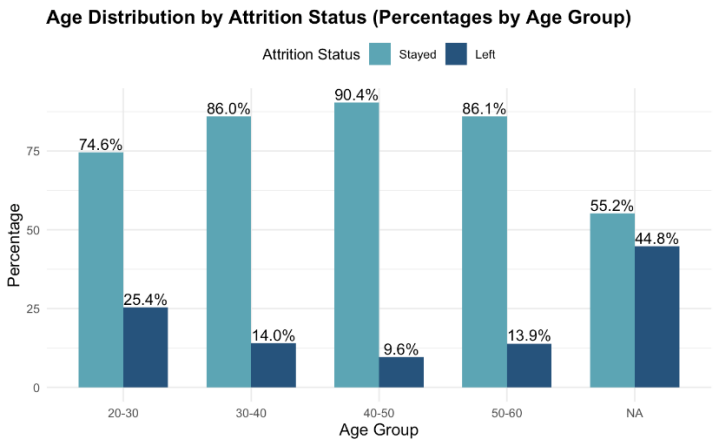
Visualization

The results were visualized using a bar graph, where each age group was represented with percentages indicating the proportion of employees who stayed or left. The color-coded bars (teal for stayed, dark blue for left) make it easy to distinguish between the two attrition statuses at a glance.

Key Findings

The youngest age group (20-30 years) showed a relatively high attrition rate, with 25.4% of individuals in this group leaving the company.

Attrition rates decrease progressively in older age groups, with the lowest rates observed in the 40-50 years group, where only 9.6% left the company.



Analysis of Monthly Income Variation with Age

Overview

To further explore the dynamics within our workforce, we analyzed the correlation between age and monthly income, hypothesizing that income would generally increase with age due to accumulated experience and career progression.

Visualization and Methodology

We utilized a scatter plot to visually represent each employee's age against their corresponding monthly income, overlaying a loess curve to highlight the overall trend.

The plot, rendered using R's ggplot2, displays age on the x-axis and monthly income on the y-axis, with individual data points plotted to show the distribution and density across different age groups.

Key Observations

The scatter plot reveals a clear positive correlation between age and monthly income, as indicated by the upward trajectory of the loess curve. This suggests that, on average, employees earn higher incomes as they age, which could be attributed to several factors such as greater job responsibilities, advancements in position, or accumulation of skills and experience.

Analysis

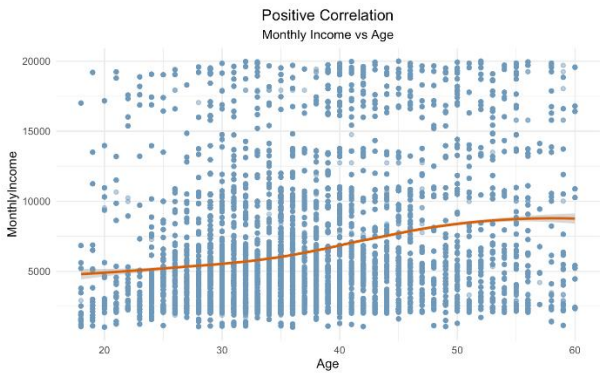
Initial Career Phase (20s to early 30s): Income levels are generally lower but increase steadily as individuals progress through their early career stages.

Mid-Career Phase (mid-30s to early 50s): There is a more pronounced increase in income, likely reflecting career stability, advancement to managerial or senior roles, and peak productivity periods.

Late Career Phase (late 50s and beyond): The income growth rate tends to plateau, which might reflect a transition towards retirement or a shift in work roles, such as reduced responsibilities or part-time positions.

Implications

This analysis not only validates the positive correlation between age and income but also highlights critical stages in an employee's career where targeted interventions could maximize income growth and job satisfaction. Understanding these patterns helps tailor compensation packages and career development programs that align with the age and career stage of the workforce, thereby aiding in retention and job satisfaction.



Dataset Preparation for Predictive Modeling

Dataset Splitting

To ensure the robustness and generalizability of our predictive model, the

dataset was split into training and testing sets. Using a stratified sampling method, approximately 70% of the data was allocated to the training set and 30% to the testing set. This approach maintains a similar distribution of the target variable, Attrition_numeric, across both sets, which is crucial for unbiased model training and validation.

Balancing the Dataset

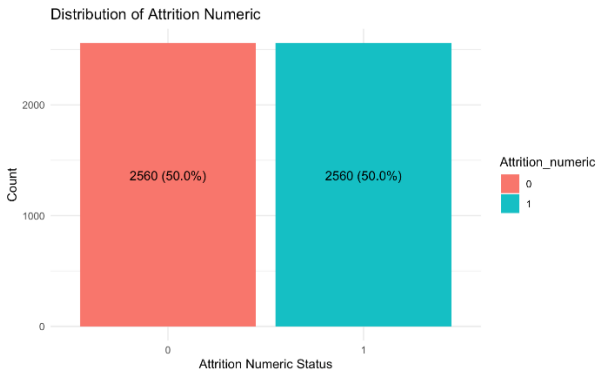
The initial examination of the Attrition_numeric distribution revealed an equal split between the classes (50% stayed, 50% left), as visualized in the bar graph. However, to further enhance model accuracy and prevent overfitting due to any underlying imbalance not visible in the overall dataset, we applied a downsampling technique to the training data. This method adjusts the class distribution to ensure that both outcomes are equally represented, thereby allowing the model to learn generalized patterns without bias toward the majority class.

Implementation

Class Balancing: The downSample function from the caret package was employed on the training dataset to equalize the number of instances in each class of Attrition_numeric. This ensures that our predictive modeling does not skew towards the more prevalent class.

Validation: After balancing, the new distribution was verified to confirm that the classes are now equally represented, enhancing the reliability of the model's predictive power on unseen data.

Data Integrity: Post-balancing, the attribute was renamed back to Attrition_numeric from Class to maintain consistency with the original dataset framework.



Comprehensive Analysis of Logistic Regression Model on Employee Attrition

Overview

Our analysis employed a logistic regression model to predict employee attrition based on a range of predictor variables. The model was developed using a balanced dataset to ensure accuracy and generalizability of the results. This section describes the model's formulation, training, testing, and the interpretation of its output.

Model Formulation

The logistic regression model was constructed to understand the impact of various factors on the likelihood of employee attrition. The formula for the model is as follows:

Attrition_numeric ~ Department + EducationField + Age +

BusinessTravel + DailyRate + DistanceFromHome + Education +

EnvironmentSatisfaction + Gender + HourlyRate + JobInvolvement +

JobSatisfaction + MaritalStatus + MonthlyRate + NumCompaniesWorked +

OverTime + PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +

StockOptionLevel + TrainingTimesLastYear + WorkLifeBalance +

YearsSinceLastPromotion + YearsWithCurrManager +

MonthlyIncome * JobLevel + TotalWorkingYears * JobLevel +

YearsInCurrentRole * YearsAtCompany

This formula includes:

Main Effects: Basic demographic and job-related variables such as department, education field, age, business travel frequency, daily rate, distance from home, education level, environment satisfaction, and several others.

Interaction Effects: Interaction terms like Monthly Income * Job Level and Total Working Years * Job Level are included to capture the combined effects of these variables, which are not evident when considered separately.

These variables were selected based on their potential relevance to employee satisfaction and retention as indicated by exploratory data analysis.

Data Preparation for Modeling

Data Splitting: The dataset was split into a 70:30 ratio for training and testing to validate the model effectively. This helps in understanding how the model performs on unseen data, an important step to evaluate the generalizability of the model.

Class Balancing: Given the critical nature of balanced classes in classification models, we employed an upsampling technique on the training data. This approach adjusts the dataset so that the classes of the target variable (attrition) have equal representation, which is crucial for avoiding bias toward the majority class.

Model Training and Testing

Training: The model was trained on the balanced training dataset using the glm() function in R with a binomial family, suitable for binary classification tasks like ours.

Testing: Predictions were made on the test dataset to assess the model's performance. The predicted probabilities were converted to class outputs (0 or 1) based on a threshold of 0.5.

Model Evaluation

Output Summary: The summary of the logistic regression provides coefficients for each predictor, which represent the log odds of attrition given a one-unit change in the predictor, holding other variables constant. Significant predictors were identified based on their p-values.

Model Accuracy: The accuracy of the model was calculated as the proportion of correct predictions (both true positives and true negatives) over all predictions made. A confusion matrix was also generated to visualize the performance of the model across the two classes.

Interpretation of Key Results

Significant Predictors: Variables like Age, Business Travel, Daily Rate, Environment Satisfaction, and OverTime showed significant effects. For instance, increases in Age and Environment Satisfaction are associated with decreases in the likelihood of attrition, underscoring their importance in employee retention.

Interaction Effects: The interaction terms provided insights into complex relationships, such as how the effect of Monthly Income on attrition varies by Job Level. Although not all interaction terms were significant, their inclusion is justified by the need to explore these nuanced relationships.

Model Fit: The AIC (Akaike Information Criterion) and the residual deviance indicate the goodness of fit of the model. A lower AIC and a residual deviance close to the degrees of freedom suggest a well-fitting model.

Conclusion

This logistic regression analysis has successfully highlighted various factors influencing employee attrition and confirmed the importance of considering both direct and interaction effects. The model's accuracy and the insights gained from significant predictors provide a robust framework for strategic human resource planning aimed at reducing employee turnover. Future interventions can be tailored based on these findings to address specific areas contributing to attrition, thereby enhancing employee retention strategies.

Decision Tree Model Analysis Using Cross-Validation

Overview

Our decision tree model, developed to predict employee attrition, was rigorously validated using cross-validation techniques to ensure its stability and reliability. The process involved setting up specific controls for cross-validation, adjusting the data, and analyzing the results.

Model Setup and Validation

Cross-Validation Settings:

Method: We implemented k-fold cross-validation, a robust method for assessing model performance. Initially, we set up a 5-fold cross-validation and later increased to a 10-fold configuration to refine our assessment.

Reproducibility: A fixed seed (set.seed(123)) ensured that our results are reproducible, providing consistent outputs across different runs.

Class Probability Storage: Class probabilities were saved during the cross-validation process, aiding in detailed performance metrics such as ROC and AUC calculations.

Data Preparation:

Factor Levels Adjustment: The levels of the Attrition_numeric factor were adjusted to be valid R variable names using the make.names function. This adjustment ensured compatibility with R modeling functions and avoided potential errors in the modeling process.

Model Training

Decision Tree Training:

Formula Setup: The model used Attrition_numeric as the dependent variable, predicted from all other variables in the dataset.

Model Method: We used the rpart method for recursive partitioning, a popular approach for building decision trees.

Performance Metric: Model performance was evaluated based on accuracy, ensuring the model effectively distinguishes between employees who left and those who stayed.

Results and Model Selection

Cross-Validation Outcomes:

The model was evaluated at different complexity parameter (cp) values to control tree growth. The optimal cp was chosen based on the highest accuracy observed during cross-validation:

Best CP Value: 0.01978022, yielding an accuracy **of 65.85%,** with a kappa **of 0.3169963.**

Model Performance:

Accuracy: The final model achieved an accuracy of 65.85% across the 10 folds, significantly better than the No Information Rate of 50%, indicating that the model has good predictive power.

Confusion Matrix: Detailed predictions versus actuals were tabulated:

9504 true negatives (correctly predicted non-attrition),

8473 true positives (correctly predicted attrition),

4146 false negatives,

5177 false positives.

Interpretation of Key Results

Model Efficacy: With a sensitivity of 69.63% and specificity of 62.07%, the model is relatively strong at identifying true positives but also somewhat likely to misclassify negatives.

Balanced Accuracy: The balanced accuracy of 65.85% mirrors the standard accuracy, indicating consistent performance across both classes.

Statistical Significance: The p-value (< 2.2e-16) for accuracy being greater than the no information rate confirms that the model predictions are significantly better than random guessing.

Conclusion

This decision tree model, validated through extensive cross-validation, provides a robust predictive tool for understanding employee attrition. Its performance metrics suggest it can reliably identify factors contributing to attrition, supporting HR strategies in effectively managing and reducing turnover. Adjustments and improvements can be further explored by tweaking the complexity parameter and considering additional or fewer predictors in the model.

Department-Specific Analysis: Logistic Regression Model for the Sales Department

Overview

To address unique characteristics within different organizational segments, we conducted a department-specific analysis focusing on the Sales department. This tailored approach helps in understanding attrition dynamics specific to sales personnel, enabling targeted retention strategies.

Data Preparation and Model Setup

Subset Creation and Cleaning:

A subset of data was created specifically for the Sales department.

The Department column was removed post-filtering since all records were from Sales, rendering this variable redundant.

Single-level factors, which can potentially skew the model, were removed to simplify the dataset and focus on meaningful variables.

Data Splitting and Balancing:

The Sales dataset was randomly split into training (75%) and testing (25%) sets, ensuring a representative sample for both model training and validation.

The training set was balanced using a downsampling technique to equalize the class distribution of the Attrition_numeric target variable, enhancing model fairness and reducing bias.

Logistic Regression Modeling

Model Training:

A logistic regression model was fitted using the balanced training data. The model included a variety of predictors ranging from demographic variables like age and gender to job-related factors like job involvement and monthly income.

Interaction terms and polynomial expansions were not included in this initial model to focus on direct relationships.

Model Performance and Evaluation:

Key Predictors: Variables such as BusinessTravel, DistanceFromHome, and OverTime were significant, indicating their strong influence on attrition decisions within the Sales department.

Model Fit and Statistics: The model showed a good fit, with an AIC of 1970.6 and a significant reduction in residual deviance, indicating that the model captured substantial information about the data.

Predictive Performance on Testing Data

Accuracy and Predictive Power:

The model achieved an accuracy of approximately 70.3% on the testing set, which is considerably higher than random chance.

The confusion matrix for the testing data showed that the model was more effective at identifying true negatives (employees who stayed) but also made substantial correct predictions for true positives (employees who left).

Interpretation and Implications

The high significance of BusinessTravel and OverTime suggests that sales employees who frequently travel or work overtime are at a higher risk of attrition, possibly due to work-life balance issues.

The negative coefficient for JobInvolvement and positive coefficient for TrainingTimesLastYear suggest that enhancing job involvement and providing adequate training could mitigate attrition risks.

Department-Specific Policies:

Based on the model's insights, HR strategies could be developed to improve job satisfaction and retention among sales employees. This might include reviewing travel policies, adjusting overtime requirements, and enhancing training programs.

Conclusion

This department-specific logistic regression model provides valuable insights into attrition drivers within the Sales department. By understanding specific factors that influence attrition, tailored interventions can be designed to enhance employee satisfaction and retention, directly contributing to the department's stability and success.

Department-Specific Analysis: Logistic Regression Model for the Research & Development Department

Overview

In our ongoing effort to understand and manage employee attrition across different segments of the organization, we focused on the Research & Development (R&D) department. This analysis aims to uncover specific patterns and predictors of attrition within this group, facilitating the development of targeted retention strategies.

Data Preparation and Model Setup

Subset Creation and Cleaning:

We extracted a subset of data pertaining exclusively to the Research & Development department.

The 'Department' column was removed after filtering because it was redundant within the subset.

Single-level factors were eliminated to avoid skewing the model, ensuring a focus on relevant and impactful variables.

Data Splitting and Balancing:

The R&D dataset was divided into training (75%) and testing (25%) sets, with random sampling to maintain representativeness.

To combat imbalance in the training dataset, a downsampling technique was applied to equalize the distribution of the Attrition_numeric target variable. This enhances the model's fairness and accuracy by reducing bias.

Logistic Regression Modeling

Model Training:

A logistic regression model was developed using the balanced training dataset. The model incorporated various predictors from demographic details (such as age and gender) to job-related factors (including job involvement and monthly income).

The initial modeling phase focused on direct relationships without including interaction terms or polynomial expansions.

Model Performance and Evaluation:

Significant Predictors: Factors such as BusinessTravel, OverTime, and JobInvolvement were highly significant, highlighting their strong influence on attrition within the R&D department.

Model Fit and Statistics: The logistic regression demonstrated an effective fit, with an AIC of 4026.3 and a significant reduction in residual deviance. This indicates that the model captured substantial information about the attrition dynamics in the department.

Predictive Performance on Testing Data

Accuracy and Predictive Power:

The model achieved an accuracy of approximately 66.58%, which, while modest, offers insights above random guessing.

The confusion matrix indicated the model's effectiveness in predicting both true negatives (employees who stayed) and true positives (employees who left), although improvements are desirable, particularly in reducing false negatives.

Interpretation and Implications

Insights for Retention Strategies:

The significance of BusinessTravel and OverTime suggests that R&D employees frequently involved in travel or overtime are more likely to leave, potentially due to strains on work-life balance.

Factors like JobSatisfaction and JobInvolvement being significant indicate that increasing engagement and satisfaction at work could potentially reduce attrition rates.

Department-Specific Policies:

Based on the findings, HR strategies could include reviewing work assignment policies, especially those relating to overtime and travel. Enhancing job satisfaction through various engagement activities and

recognizing employee efforts could also play a crucial role in retention.

Conclusion

This tailored logistic regression analysis for the R&D department provides crucial insights into the specific factors influencing employee attrition. Understanding these can help in designing effective interventions aimed at improving job satisfaction and retention, thereby enhancing the stability and success of the department.

Department-Specific Analysis: Logistic Regression Model for the Human Resources Department

Overview

This analysis is part of a broader initiative to understand employee attrition across various organizational departments, with this section focusing on the Human Resources (HR) department. The objective is to identify key factors that influence attrition specifically within HR, enabling the creation of more effective employee retention strategies tailored to this department.

Data Preparation and Model Setup

Subset Creation and Cleaning:

A subset of data was specifically extracted for the Human Resources department.

Post-extraction, the 'Department' column was removed as it became redundant within this focused dataset.

We removed single-level factors to ensure that the remaining variables were relevant and had sufficient variability to contribute meaningfully to the model.

Data Splitting and Balancing:

We split the HR dataset into training (75%) and testing (25%) subsets using a random sampling method to maintain representativeness.

The training set was balanced using downsampling to equalize the distribution of the Attrition_numeric variable. This step helps reduce bias in the model by ensuring a fair representation of both attrition outcomes.

Logistic Regression Modeling

Model Training:

A logistic regression model was employed, utilizing the balanced training data. The model incorporated various predictors, encompassing both demographic and job-related factors.

The direct relationships were the focus in this initial modeling phase, avoiding the incorporation of interaction terms or polynomial expansions to maintain model simplicity and interpretability.

Model Performance and Evaluation:

Significant Predictors: Key variables such as Age, BusinessTravel, DistanceFromHome, and MonthlyIncome proved significant, indicating their substantial impact on HR employee attrition decisions.

Model Fit and Statistics: The model achieved an effective fit, as shown by an AIC of 517.35 and a notable reduction in residual deviance, suggesting that the model was successful in capturing the dynamics of attrition within HR.

Predictive Performance on Testing Data

Accuracy and Predictive Power:

The logistic regression model achieved an accuracy of approximately 69.95% on the testing data, which is significantly better than chance.

The confusion matrix revealed the model's ability to correctly predict both true negatives and positives, although there remains room for improvement, especially in decreasing false negatives.

Interpretation and Implications

Insights for Retention Strategies:

The importance of BusinessTravel and MonthlyIncome suggests that concerns related to travel demands and compensation are crucial in the attrition rates among HR employees. Addressing these could help reduce turnover.

The negative impact of Age and TrainingTimesLastYear on attrition suggests that more experienced HR employees or those receiving more training are less likely to leave, highlighting areas where retention efforts could be concentrated.

Department-Specific Policies:

Insights derived from the model can guide HR policy adjustments, particularly around travel requirements, compensation structures, and training opportunities. Enhancing job satisfaction and engagement through targeted initiatives could be beneficial.

Conclusion

The focused logistic regression analysis for the Human Resources department provides valuable insights into specific factors that influence attrition. These findings can assist in crafting targeted interventions that improve job satisfaction and retention rates, thereby fostering a more stable and effective HR department.

Department-Specific Analysis: Decision Tree Models for Different Departments

Sales Department

Overview

In our comprehensive study to analyze employee attrition, a decision tree model was developed specifically for the Sales department. This model aims to identify and understand the critical factors leading to employee turnover within this sector, providing insights to formulate effective retention strategies.

Model Training and Evaluation

Data Splitting and Balancing: The data was split into training (70%) and testing (30%) sets. To address class imbalance in the training dataset, a downsampling technique was used.

Decision Tree Model: The decision tree was trained to capture complex patterns and interactions among the predictors with respect to attrition.

Model Evaluation: The model achieved an accuracy of 74.97%. The confusion matrix indicated a sensitivity of 77.37% and a specificity of 65.48%, reflecting a moderate ability to distinguish between employees who would leave or stay.

Implications for Retention Strategies

The decision tree model identified key predictors impacting attrition in sales. These insights can guide the development of targeted interventions aimed at addressing specific factors contributing to turnover.

Enhanced training programs, improved work conditions, and tailored incentive structures may prove effective in retaining sales personnel.

Human Resources Department

Overview

A decision tree model was also implemented for the Human Resources department to pinpoint the underlying causes of employee departures and assist in crafting tailored HR policies to enhance retention.

Model Training and Evaluation

Data Handling: Similar to the Sales department, the data for HR was divided and balanced appropriately to ensure a representative training sample.

Decision Tree Model: This model provided a structured analysis of attrition factors within HR, utilizing a diverse set of predictors.

Model Evaluation: The model's accuracy stood at 77.66%. It demonstrated a high specificity of 95.74%, indicating strong performance in identifying employees who are likely to stay.

Implications for Retention Strategies

The findings suggest that specific factors in HR, such as job satisfaction and work-life balance, are critical. Initiatives to enhance these areas could lead to significant improvements in employee retention.

Adjustments to HR policies, focusing on professional development and employee engagement, could mitigate attrition risks effectively.

Research & Development Department

Overview

For the Research & Development department, a decision tree model was constructed to analyze attrition rates and provide actionable insights to reduce turnover.

Model Training and Evaluation

Data Preparation: The R&D data underwent similar processing with an emphasis on balancing the training set to improve model reliability.

Decision Tree Model: The model explored various factors affecting attrition within R&D, aiming to uncover complex patterns.

Model Evaluation: Achieving an accuracy of 71.9%, the model showed balanced accuracy with a sensitivity of 72.74% and specificity of 66.97%.

Implications for Retention Strategies

The decision tree results emphasize the importance of addressing specific work-related factors such as project assignments, career progression opportunities, and recognition programs to retain R&D employees.

Enhancing career development paths and creating a more engaging work environment could significantly lower attrition rates in this department.

Conclusion

The decision tree analyses for the Sales, Human Resources, and Research & Development departments have provided crucial insights into department-specific attrition factors. These insights will guide the creation of targeted retention strategies, tailored to address the unique challenges and needs of each department, ultimately fostering a more stable and engaged workforce across the organization.

Comprehensive Comparison of Logistic Regression and Decision Tree Models

Overview

In our extensive evaluation of modeling techniques to analyze employee attrition, we compared the performance of two popular predictive models: Logistic Regression and Decision Trees. This

comparison aims to determine which model is more effective in capturing the complexities of employee turnover, thereby providing a robust foundation for developing targeted retention strategies.

Model Performance Metrics

The comparison focused on five key metrics:

Accuracy: Measures the overall correctness of the model.

Kappa: Evaluates the agreement of prediction with the actual classes, adjusted for chance.

Precision: Reflects the model's accuracy in predicting positive classes.

Recall (Sensitivity): Indicates the model's ability to identify actual positives.

F1 Score: Harmonic mean of precision and recall, providing a balance between the two.

Evaluation Results

Accuracy: Both models showed similar accuracy, with Logistic Regression slightly outperforming the Decision Tree model.

Kappa: The Decision Tree demonstrated a slightly better Kappa score, indicating a better agreement than the Logistic Regression under conditions where positive cases are less frequent.

Precision: Logistic Regression achieved higher precision, suggesting it is more reliable when predicting positive cases of attrition.

Recall: Both models exhibited nearly identical recall, indicating a similar ability to identify all relevant cases of attrition.

F1 Score: The F1 scores were equivalent, suggesting that both models provide a balanced approach to precision and recall.

Plot Analysis

The bar plot visually represents the comparison of the two models across these metrics, underscoring their strengths and weaknesses. Logistic Regression tends to have higher precision, which is crucial for minimizing false positives in employee attrition prediction. On the other hand, the Decision Tree's slightly better Kappa score suggests it may handle random chances more effectively, a significant aspect when dealing with unbalanced datasets.

Conclusion

The choice between Logistic Regression and Decision Trees may depend on specific organizational needs and the nature of the dataset. Logistic Regression's higher precision makes it suitable for scenarios

where the cost of false positives is high. In contrast, Decision Trees might be preferred for their ease of interpretation and robustness against overfitting with appropriate tuning. The detailed metric comparison provides a clear direction for selecting the model that best fits the strategic goals of attrition analysis and retention planning, helping to enhance overall organizational resilience.

Full Logistic Regression and Decision Tree model comparison

Recommendations for Reducing Employee Attrition at IBM

Overview

In response to the challenges posed by employee attrition at IBM, we have employed predictive modeling and targeted interventions as a strategic approach to mitigate turnover and enhance employee retention. This section outlines a comprehensive set of recommendations designed to address the specific factors contributing to employee attrition across different departments, leveraging insights gained from our analysis to propose actionable and cost-effective strategies.

Economic Rationale and Cost-Benefit Analysis

Baseline Attrition Costs

Initial Attrition Impact: With a total of 3,657 employees leaving, the attrition has imposed a significant cost of approximately $14,628,000 on the organization, considering an average cost of $4,000 per employee. This cost encompasses recruitment, training, and lost productivity.

Predictive Model Insights and Financial Implications

Logistic Regression Model: Our logistic regression model, which achieved a sensitivity of 65.36%, indicates that through targeted interventions, we could potentially retain up to 2,390 employees who might have otherwise left the company.

Projected Savings: By applying the model and implementing recommended strategies, the potential savings are estimated at $9,560,000. Even after accounting for the remaining attrition post-intervention, the

projected attrition cost would reduce to $5,068,000.

Cost of Implementation: The estimated cost for implementing these retention strategies is about $1,500,000.

Net Savings

After considering the cost of retention strategies, the net savings are projected to be approximately $8,060,000. This substantial figure highlights the financial viability of our strategic recommendations, providing a strong incentive for their adoption.

Recommendations

Company-Wide Strategies

Work-Life Balance: Implement flexible work arrangements to foster a healthier work-life balance, enhancing employee satisfaction and reducing turnover.

Business Travel: Modify travel policies to reduce the frequency and necessity of travel, easing the related stress and contributing to better overall well-being.

Department-Specific Strategies

Sales:

Age and Career Development: Provide age-appropriate career development opportunities tailored to enhance skill sets and job satisfaction.

Overtime Practices: Review and adjust overtime policies to ensure they are fair and do not contribute to employee burnout.

Compensation: Regularly assess and adjust compensation packages to ensure competitiveness and equity.

Human Resources (HR):

Distance from Home: Offer flexible working conditions or relocation assistance to HR employees who face long commutes.

Professional Development: Expand access to training and development programs, reinforcing the HR department's role in fostering a supportive work environment.

Research & Development (R&D):

Environment Satisfaction: Initiate programs that enhance workplace satisfaction, such as recognition schemes and career advancement opportunities.

Education and Training: Provide targeted training and continuous skill development to keep pace with technological advancements and industry demands.

Implementation and Expected Impact

The implementation of these recommendations is expected to significantly reduce attrition rates, thus retaining valuable human capital and reducing recurrent costs associated with hiring and training new employees. By investing in these strategic areas, IBM can expect to see not only cost savings but also

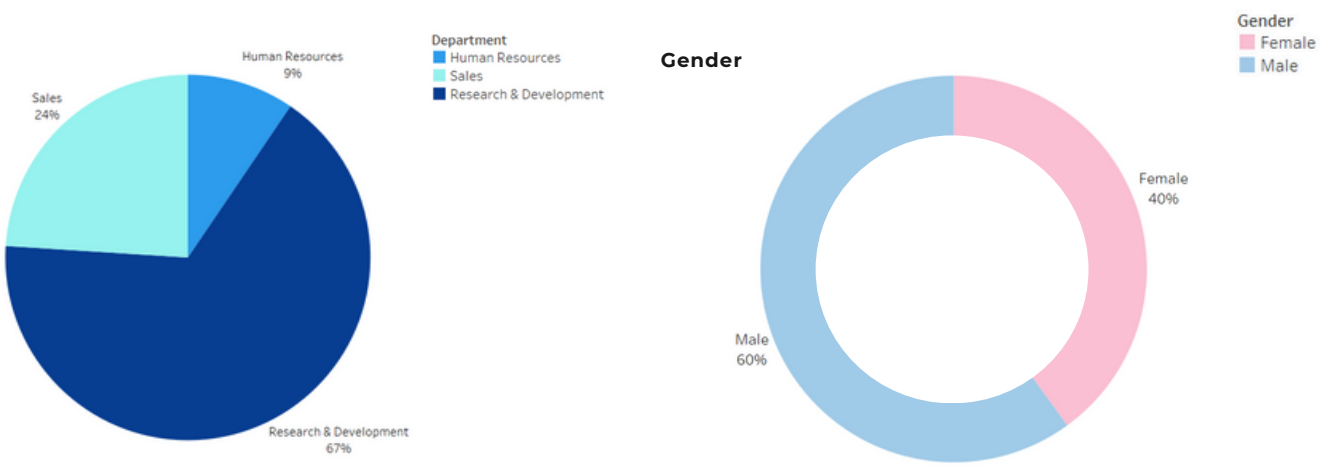
an enhancement in employee engagement and productivity, leading to sustained organizational growth and stability. These initiatives will ensure that IBM remains a competitive and desirable place to work, attracting and retaining top talent in an increasingly competitive market.

References

1. Anitha, B. M. (2023). *A Study On Employee Attrition And Retention With Reference To Evron Impex*. Anvesak.
2. Kesavan, L., & Dhivya, S. (2022). *A Study On Causes Of Employee Attrition*. Journal of Pharmaceutical Negative Results.
3. Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). *Predicting Employee Attrition Using Machine Learning Techniques*. Computers.
4. Lewis, G., & Soroñgon, J. (2022). *Industries with the Highest (and Lowest) Turnover Rates*. LinkedIn.
5. Mansor, N., Sani, N. S., & Aliff, M. (2021). *Machine Learning for Predicting Employee Attrition*. International Journal of Advanced Computer Science and Applications (IJACSA), 12(11).
6. Negi, G. (2013). *Employee Attrition: Inevitable Yet Manageable*. International Monthly Refereed Journal of Research In Management & Technology.
7. VanWieren S, (2017), *Quantifiably Better: Delivering HR Analytics from Start to Finish,* 1st Edition, ISBN 9781634622219

APPENDIX

Percentage of Employees in each Department



% of Employees in Each Dept.

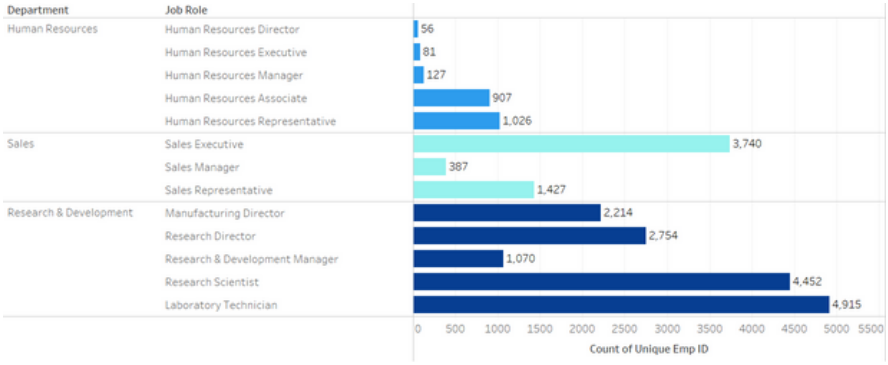
The 67% proportion in R&D underscores IBM's strong emphasis on innovation and development of new technologies, services, and solutions, which is central to maintaining its competitive edge in the tech industry.

The 24% in Sales reflects a robust sales force, which is crucial for a company like IBM that operates in a highly competitive global market and needs a strong sales team to market and sell its diverse array of products and services.

The 9% in Human Resources is relatively small compared to the operational sectors, which is typical as HR's role is generally to support the larger workforce without needing to be a large department itself

The entire workforce consists of approximately 60% males and 40% females.

Job Roles Segmented by Departments



Research & Development (R&D) is the largest department, with the highest count of unique employee IDs, especially in roles such as Laboratory Technician, Research Scientist, and Research Director. This suggests a significant focus on R&D activities within the organization.

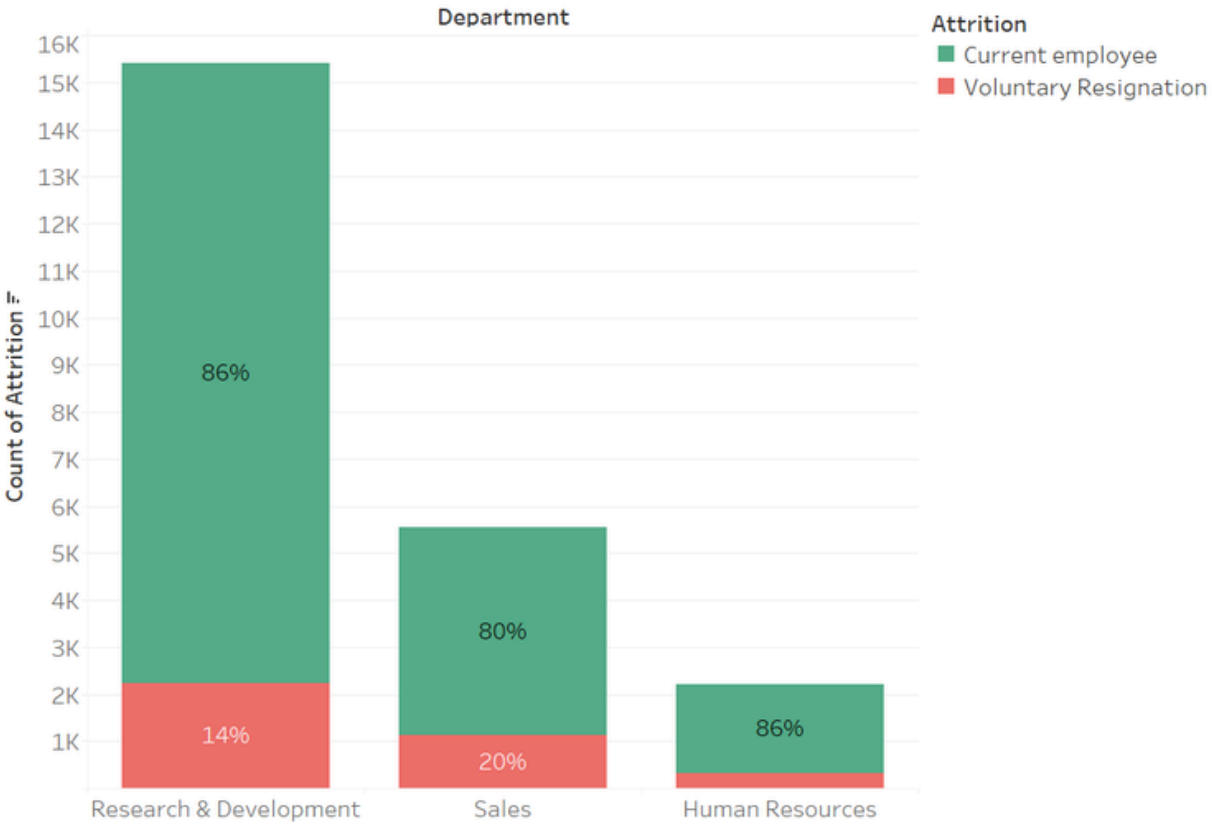
Sales is the second-largest department by employee count, with the role of Sales Executive being the most common, followed by Sales Representative and Sales Manager. This implies a robust sales operation with multiple levels of sales roles.

Human Resources has fewer employees compared to the other departments, which is typical for many organizations as the HR function often has a supporting role rather than being directly involved in the primary business operations like R&D and sales.

The job roles within Human Resources include Human Resource Associate and HR Representative, HR Manager, and HR Executive, with Human Resource Representative having the highest count. This could indicate a dedicated HR function for Human Resource benefits or similar employee welfare programs.

Specific roles such as Human Resources Director and Human Resources Manager have relatively low counts, which is consistent with organizational structures where there are fewer higher-level positions compared to entry-level or mid-level roles.

Attribution Rate by Department



Research & Development has the highest number of employees and the lowest percentage of voluntary resignations (14%), suggesting a relatively stable workforce within this department.

Sales has a higher attrition rate with 20% having voluntarily resigned, which might reflect the challenging nature of sales roles or market conditions affecting the sales department.

Human Resources shows an attrition rate that appears to be similar to Research & Development with 14% resignations, but the overall number of employees is much lower, consistent with typical company structures where HR departments are smaller.

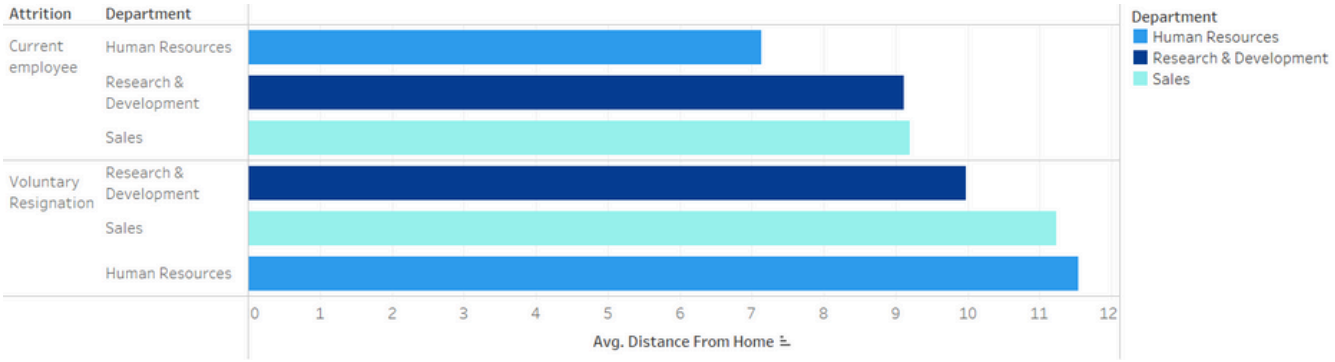
Average Number of Companies employees have worked at before

Department	Job Role	
Human Resources	Human Resources Associate	2
	Human Resources Director	1
	Human Resources Executive	5
	Human Resources Manager	2
	Human Resources Representative	2
Research & Development	Laboratory Technician	2
	Manufacturing Director	2
	Research & Development Manager	2
	Research Director	3
	Research Scientist	3
Sales	Sales Executive	4
	Sales Manager	2
	Sales Representative	4

Avg. Num Co..
1 5

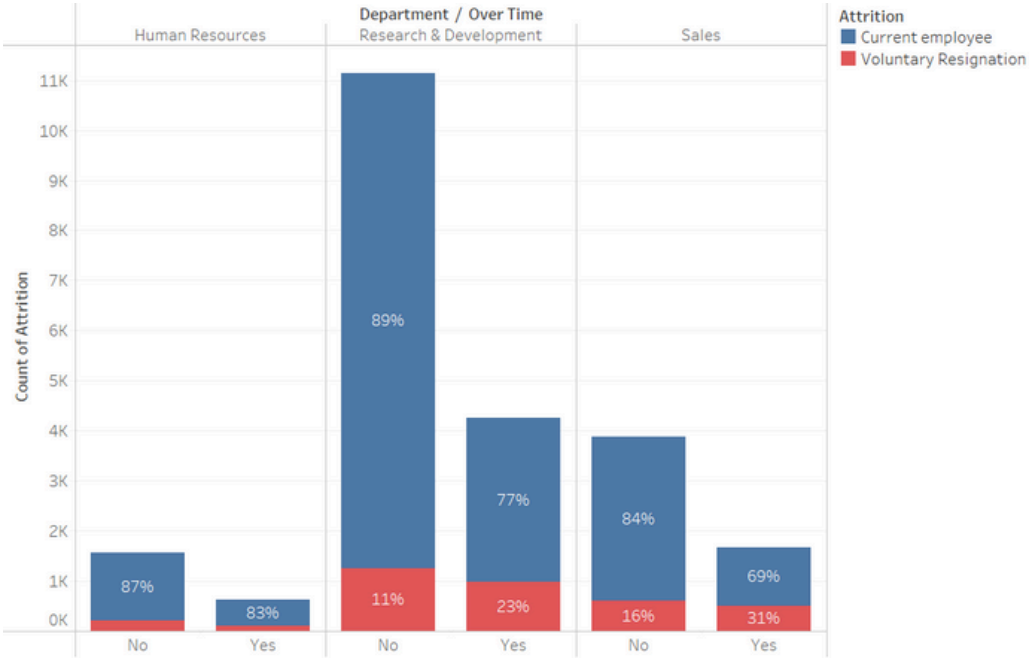
Majority of the work fore has worked at least in 2 companies before joining IBM.

Average Distance from Home and Attrition



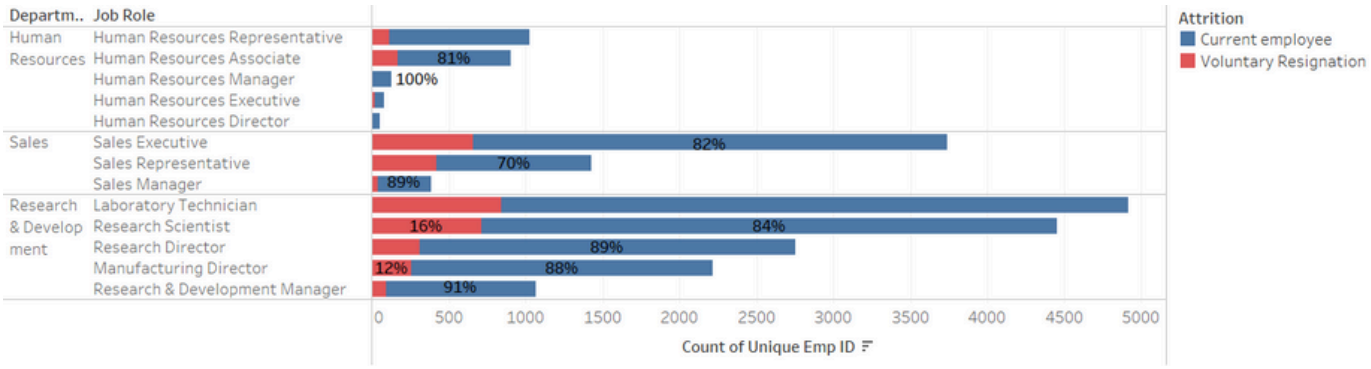
There seems to be a pattern where resigned employees lived farther from work than current employees, suggesting that a longer commute might be a factor in an employee's decision to leave the company.
Human Resources has the most considerable difference in average commuting distance between current employees and those who have resigned, which might indicate that commute distance is a particularly significant factor in turnover for this department.
All departments show higher average commute distances for resigned employees, which supports the idea that commute could be a universal factor in employee turnover across the company.

Department OverTime by Attrition



In Research & Development (R&D), the highest percentage of attrition is for those who worked overtime (23% attritlion compared to 77% current employees).
The Human Resources department shows the least amount of attrition across the overtime groups.
Sales demonstrates significant attrition among employees who worked overtime, with 31% of those employees leaving.

Attribution Rate by Job Role



The HR Manager role has a 100% retention rate, with no voluntary resignations.
The Human Resource Representative and HR roles show some level of attrition, with a larger proportion of current employees than those who have resigned.

The R&D roles display varying levels of attrition. The Laboratory Technician role shows a significant proportion of resignations (17%).
Other roles in R&D, such as Manufacturing Director, R&D Manager, and Research Director, have higher retention rates with lower percentages of voluntary resignations.

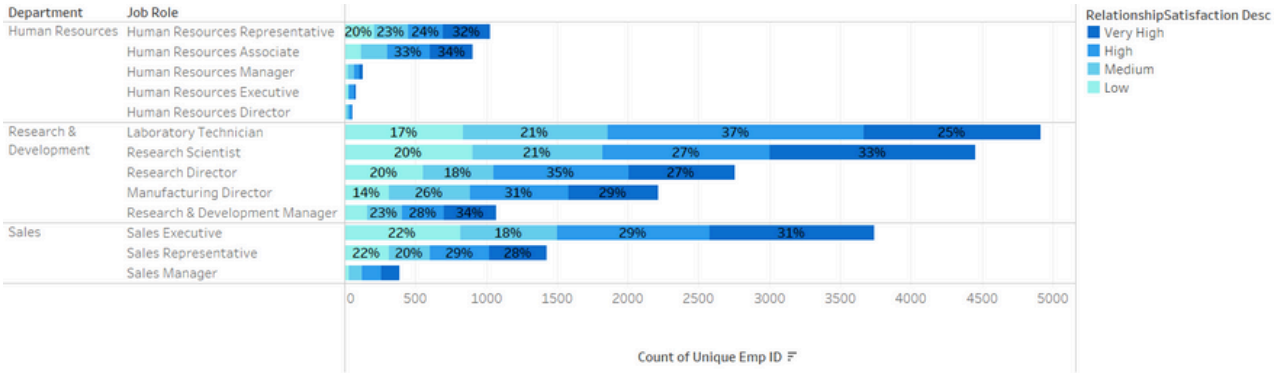
The Sales Representative role shows the highest rate of voluntary resignation (30%) in comparison to retention within this department.
Sales Executives and Sales Managers have a majority of current employees with lower attrition rates (18% and 11% respectively).

Across all departments, the majority of individuals in each role are current employees, with the retention rate being notably high for several roles.
The bar lengths represent the total number of employees for each role, with Research Scientist having the highest number and HR Director having the lowest count among the roles listed.

Attrition vs. Retention:

While most job roles have higher retention, certain roles such as Sales Representative and Laboratory Technician have relatively higher voluntary resignation rates.

Relationship Satisfaction by Job Role

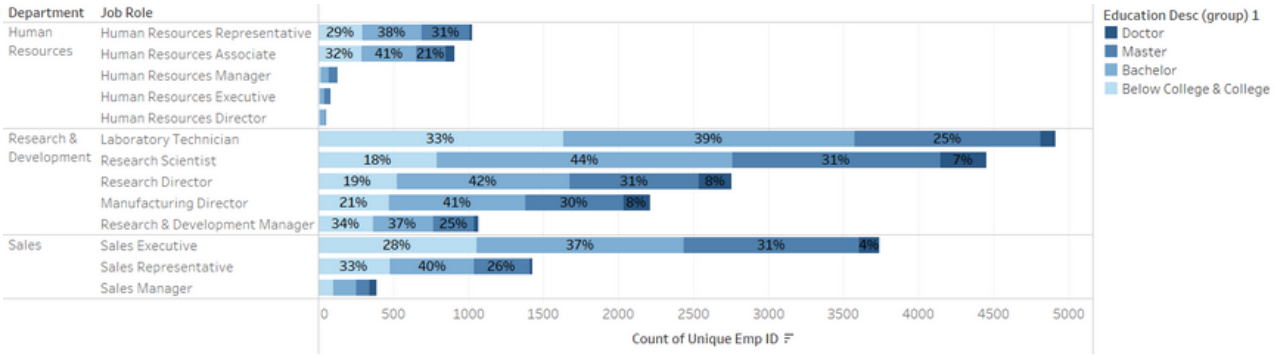


The bars indicate a mixture of relationship satisfaction levels across various roles. For example, roles like 'Research & Development Manager' and 'Sales Representative' have a notable spread across all four categories of relationship satisfaction.
Significant High Satisfaction:

Many roles have a high percentage of employees reporting 'High' relationship satisfaction, which could suggest positive interpersonal and professional relationships within these roles.
Concern Areas:

There are roles with a substantial percentage of 'Low' relationship satisfaction, which might be an area for HR to investigate to improve team dynamics or address specific issues.

Education Qualifications for Job Roles



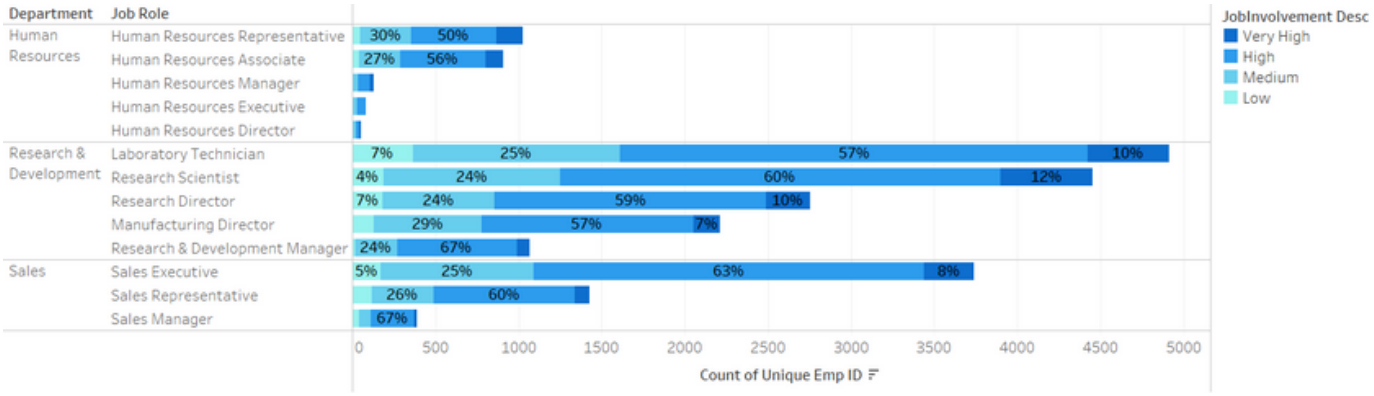
In job roles like 'Human Resources Representative,' there is a high proportion of employees with a 'Bachelor' level of education.
'Human Resources Manager' shows that all employees in this role have at least a 'College' education, with a substantial portion holding a 'Master's degree.

The 'Laboratory Technician' role shows a significant percentage of employees with 'Bachelor' and 'Below College' education, while 'Research Scientist' and 'Research Director' roles have a higher proportion of employees with 'Doctor' level education, which is expected in roles that involve advanced research.

For the 'Sales Representative' role, there's a relatively large segment of employees with 'Below College' education.
'Sales Executives' and 'Sales Managers' primarily have 'Bachelor' and 'Master' degrees.
Highly Educated Workforce:

IBM appears to have a highly educated workforce, with most roles filled by individuals holding at least a 'Bachelor' degree.
Advanced roles like 'Manufacturing Director' and 'Research & Development Manager' are mostly filled by individuals with 'Master' degrees, reflecting the need for advanced education in managerial positions that may require both technical and leadership skills.

Job Involvement by Job Role

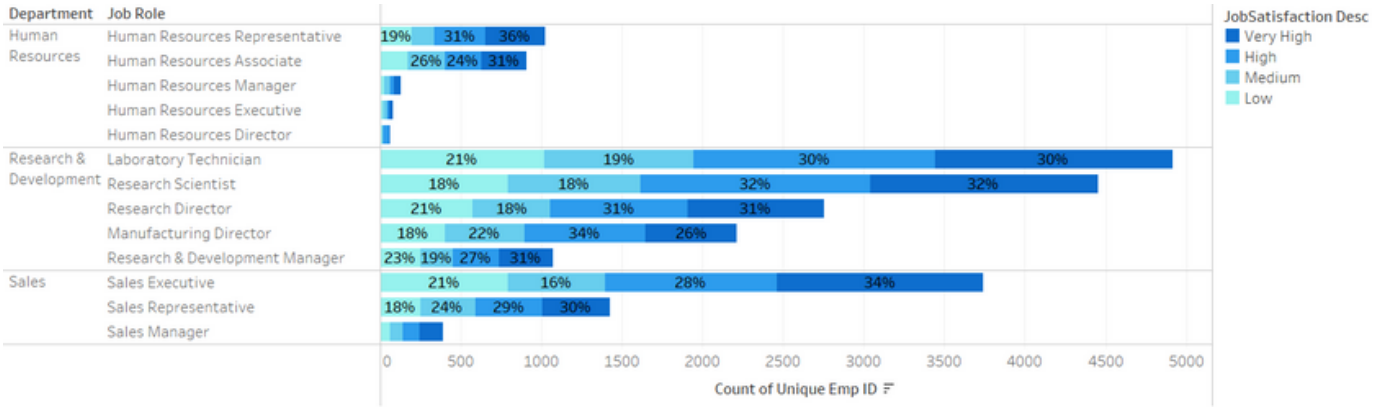


Human Resources Department: Job roles within HR seem to have varying levels of involvement, with significant portions displaying "High" and "Very High" job involvement.

Research & Development (R&D): It's a key department with many roles showing "High" to "Very High" job involvement, which is typical for roles that are closely tied to the core business of innovation at IBM.

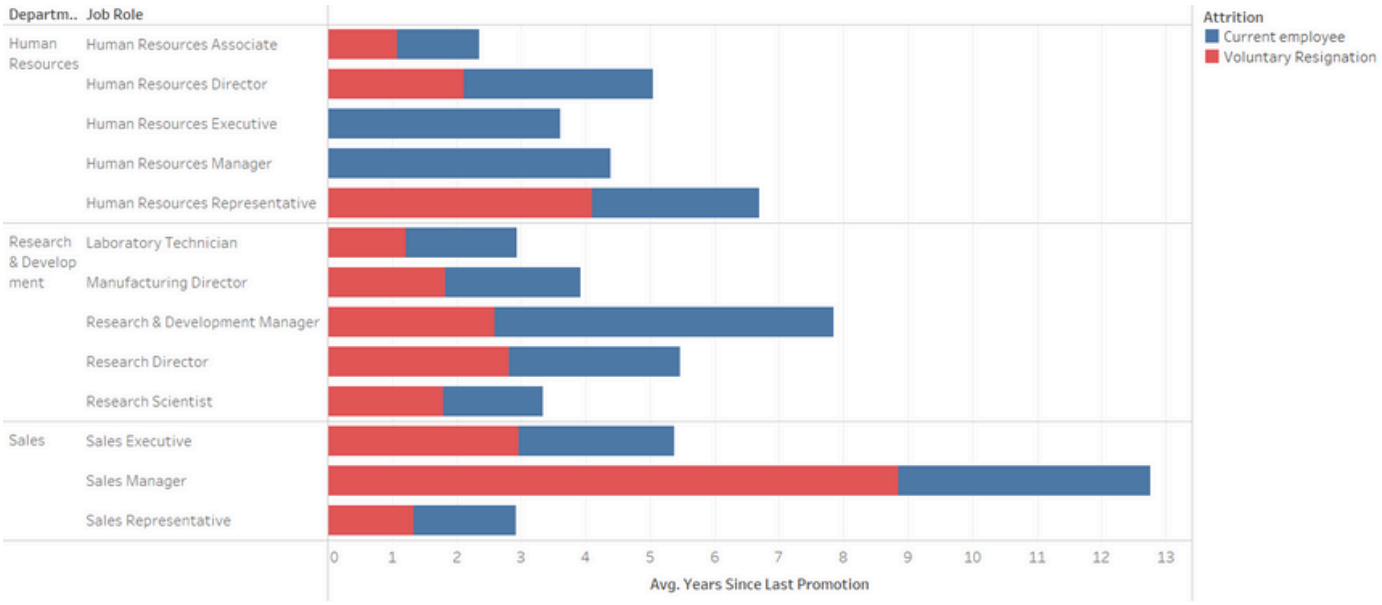
Sales Department: The Sales roles, particularly "Sales Executive" and "Sales Manager," exhibit a significant degree of "High" and "Very High" involvement, which may reflect the demanding nature of sales roles that often have performance tied closely to measurable outcomes like sales targets.

Job Satisfaction by Job Role



Overall, most roles have a significant portion of employees with "High" or "Very High" job satisfaction.
There are noticeable percentages of "Low" job satisfaction across various roles, which could be a focal point for HR interventions.

avg years since last promoted

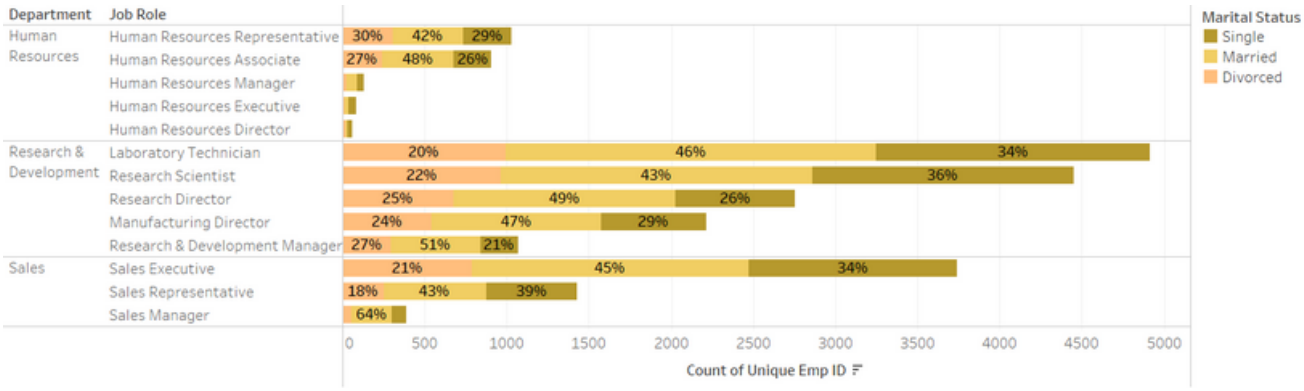


Departm..	Attrition	Job Role	Years Since Last Promotion (bin) 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Human Resources	Current employee	Human Resources Associate	46%	27%	11%	4%	4%	4%	1%	3%	0%	0%		0%				
		Human Resources Director	9%	36%	19%		17%			2%	17%							
		Human Resources Executive	26%	24%	13%			13%		11%						13%		
		Human Resources Manager	20%	13%	10%	6%	13%			20%			6%	6%	6%			
		Human Resources Representative	35%	29%	9%	3%	5%	1%	2%	4%	0%	1%	1%	4%	3%	1%	1%	1%
		Total	37%	27%	10%	4%	5%	3%	1%	5%	1%	0%	1%	2%	2%	1%	0%	0%
	Voluntary Resignation	Human Resources Associate	40%	33%	14%	9%	2%			1%								
		Human Resources Director			89%	11%												
		Human Resources Executive	100%															
		Human Resources Representative	18%	26%		7%	17%			15%	3%		3%	7%				3%
		Total	34%	27%	10%	8%	8%			6%	1%		1%	3%				1%
	Total		37%	27%	10%	4%	6%	2%	1%	5%	1%	0%	1%	2%	1%	1%	0%	1%
Research & Development	Current employee	Laboratory Technician	44%	23%	11%	5%	3%	3%	2%	4%	1%	1%	0%	0%	1%			0%
		Manufacturing Director	39%	27%	6%	6%	4%	5%	2%	6%	1%	0%	0%	1%	0%		2%	1%
		Research & Development Manager	19%	15%	9%	2%	14%	2%	5%	5%	1%	3%	1%	10%	2%	3%	5%	5%
		Research Director	34%	25%	10%	4%	5%	5%	3%	5%	2%	1%	0%	1%	1%	2%	1%	1%
		Research Scientist	42%	29%	14%	2%	2%	1%	2%	4%	1%	1%	1%	1%	0%	0%		0%
		Total	39%	25%	11%	4%	4%	3%	2%	5%	1%	1%	0%	1%	1%	1%	1%	1%
	Voluntary Resignation	Laboratory Technician	53%	29%	8%	1%	1%	1%	0%	2%		1%	0%					2%
		Manufacturing Director	46%	22%	3%	0%	4%	13%	3%	7%								
		Research & Development Manager	44%	26%						21%		8%						1%
		Research Director	44%	17%	7%	5%	4%	3%	5%		3%	5%		3%	1%	3%		3%
		Research Scientist	37%	22%	21%	5%	4%	0%	1%	7%	2%	1%						
		Total	45%	24%	11%	3%	3%	2%	2%	5%	1%	2%	0%	0%	0%	0%		1%
	Total		40%	25%	11%	4%	4%	3%	2%	5%	1%	1%	0%	1%	1%	1%	1%	1%
Sales	Current employee	Sales Executive	39%	22%	10%	2%	4%	4%	2%	8%	2%	2%		3%	0%	1%		1%
		Sales Manager	31%	12%	10%	7%	9%	2%	6%	5%	2%	3%		3%	1%	0%	1%	7%
		Sales Representative	47%	22%	15%	3%	3%	4%	2%	1%			0%	1%	1%	1%	1%	0%
		Total	40%	21%	11%	3%	4%	4%	2%	6%	1%	2%	0%	2%	0%	1%	0%	1%
	Voluntary Resignation	Sales Executive	32%	19%	12%	5%	7%	0%	4%	11%	1%	2%	1%	2%	1%	1%	1%	
		Sales Manager	2%	22%					20%					20%			17%	20%
		Sales Representative	55%	21%	11%	1%	2%	4%		3%	1%		2%					1%
		Total	39%	20%	11%	3%	4%	2%	3%	8%	1%	1%	1%	2%	0%	1%	1%	1%
	Total		40%	21%	11%	3%	4%	3%	3%	7%	1%	2%	0%	2%	0%	1%	1%	1%

% of Total Distinct.. 0% 100%

Both Visualizations sjow that attrition is higher among employees who haven't been promoted for a significant number of years, particularly in sales roles and certain roles in the Research & Development department.

Marital Status by Job Role

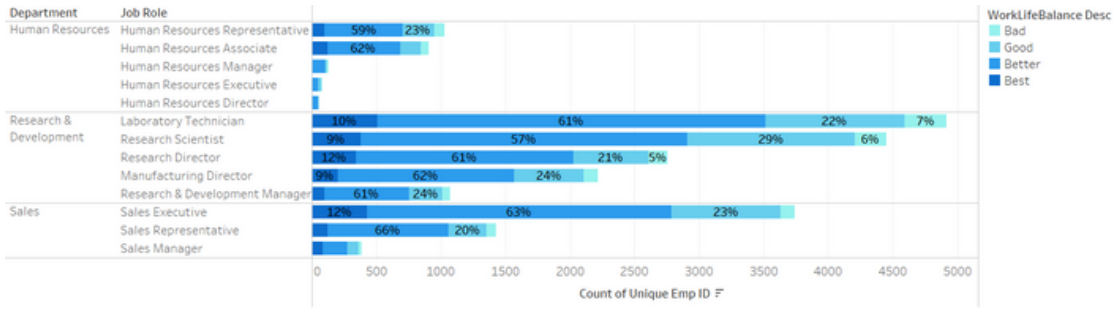


A significant portion of 'Human Resource Representative' and 'Human Resources Associate' roles are held by single individuals.
The job roles within HR seem to have a variety of marital statuses with no single status overwhelmingly dominant.

'Research & Development Manager' roles have a high proportion of married individuals.
'Laboratory Technician' and 'Manufacturing Director' roles also have a substantial number of married employees, followed by single and then divorced statuses.

'Sales Representative' roles have a relatively balanced distribution between single and married statuses, with singles slightly leading.
'Sales Executives' and 'Sales Managers' also display a mixture of marital statuses, but with a notable proportion being married.

Work Life Balance by Job Role



The majority of employees across all job roles seem to have a 'Better' or 'Best' work-life balance, with 'Better' being the most common.

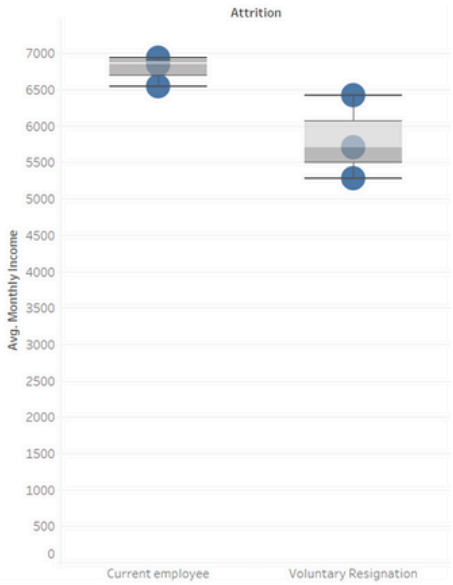
The Research & Development department has a significant number of employees, particularly in roles such as Research Scientist, Manufacturing Director, and Research & Development Manager, with the majority reporting 'Better' work-life balance.

Human Resources roles tend to have a good spread across all four work-life balance categories, with a noticeable portion of Human Resources Associates and Representatives falling into the 'Best' category.

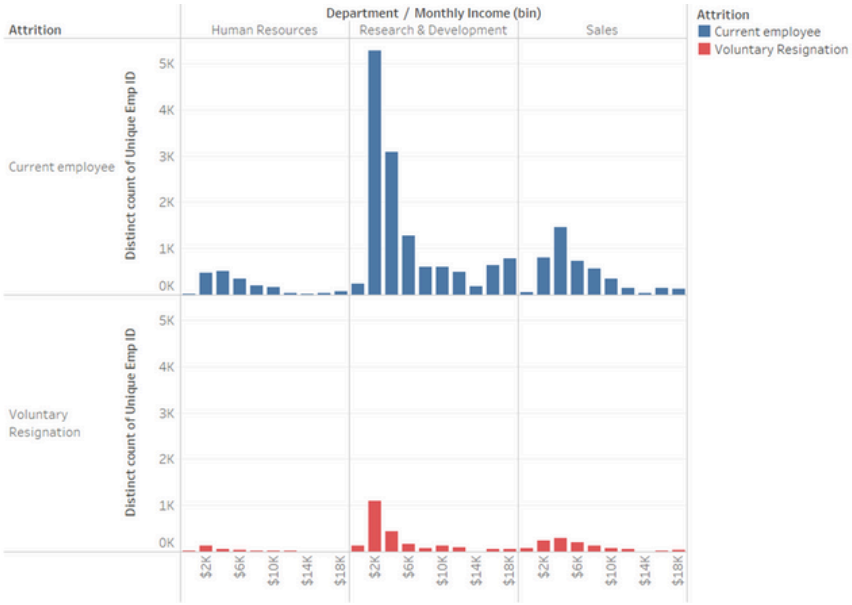
In the Sales department, Sales Representatives and Sales Executives have the highest counts, with most Sales Representatives reporting 'Better' work-life balance and a notable percentage of Sales Executives reporting 'Best' work-life balance.

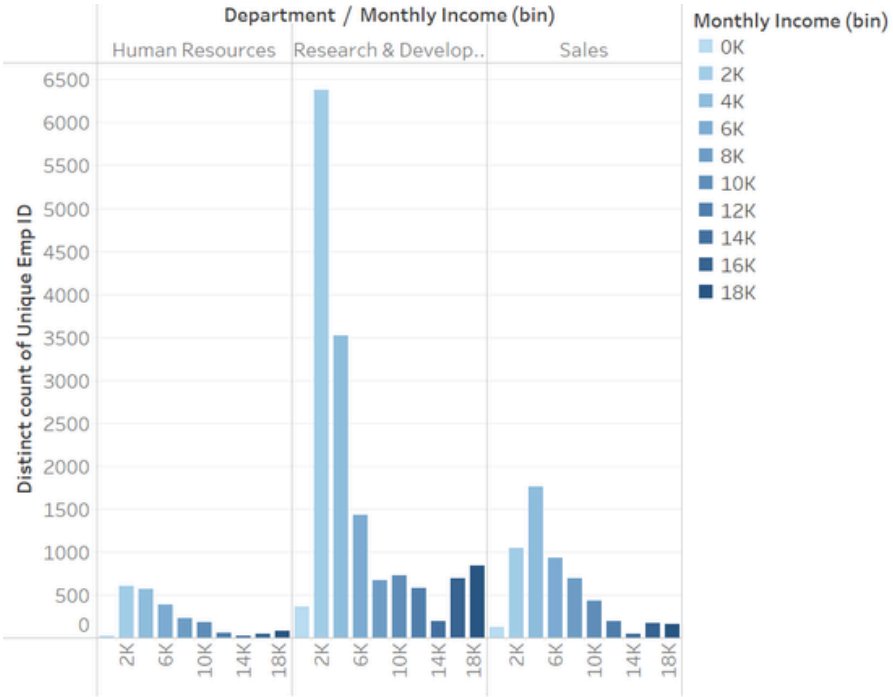
Sales Managers show the highest proportion of 'Bad' work-life balance in comparison to other roles visualized in the chart.

Avg Monthly Income / Attrition



Monthly Income Distribution - Dept. wise





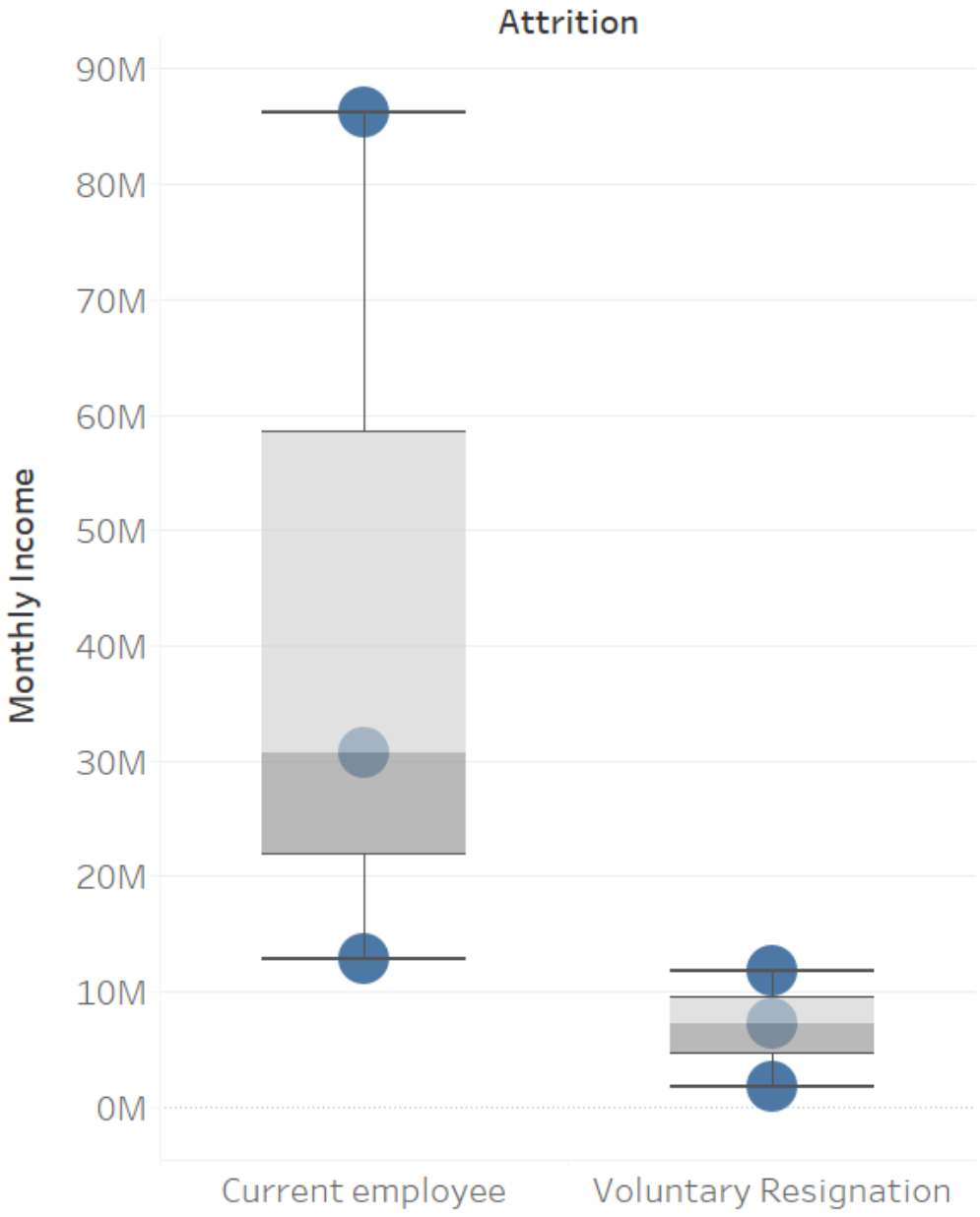
The majority of employees, especially in the Research & Development department, appear to be concentrated in a lower income bracket (possibly around 0-2K). There's a significant number of employees in this income range, suggesting it might be the starting or average salary for many positions in that department.

There are notable disparities between departments. For example, the Human Resources department seems to have a relatively even distribution across income brackets up to 10K, but few in the higher brackets. On the other hand, Sales shows fewer employees in the lower brackets but has representations up to the 18K bracket, which might indicate higher earning potential in Sales roles, possibly due to commissions or bonuses.

The higher income brackets (14K and above) are less populated across all departments, which is typical in many organizations where a smaller proportion of the workforce occupies senior or highly specialized roles that command higher salaries.

The Sales department has employees across a wide range of income brackets, including the highest ones represented in the chart. This could suggest a larger variety of roles within Sales, from entry-level to high-ranking positions, or it might reflect the success of sales personnel in achieving higher earnings through performance incentives.

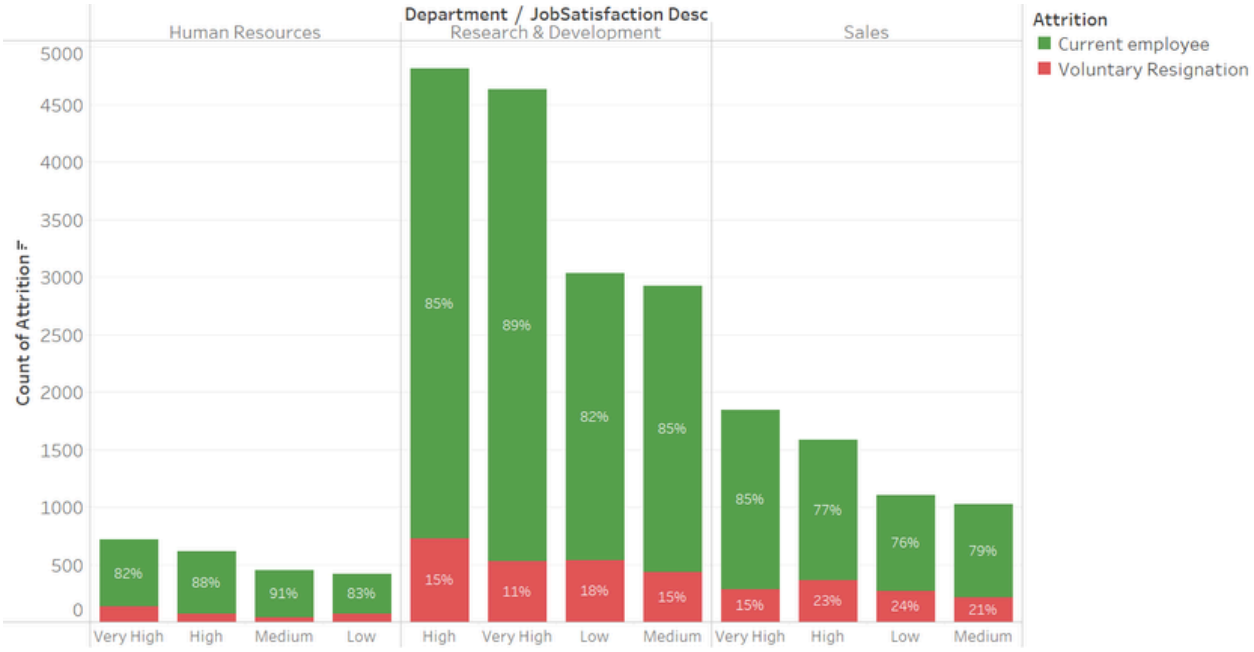
Total sum paid out by IBM in relation to Attrition



The 'Monthly Income' represent the total sum paid out to all current employees versus all employees who have resigned voluntarily in a month.

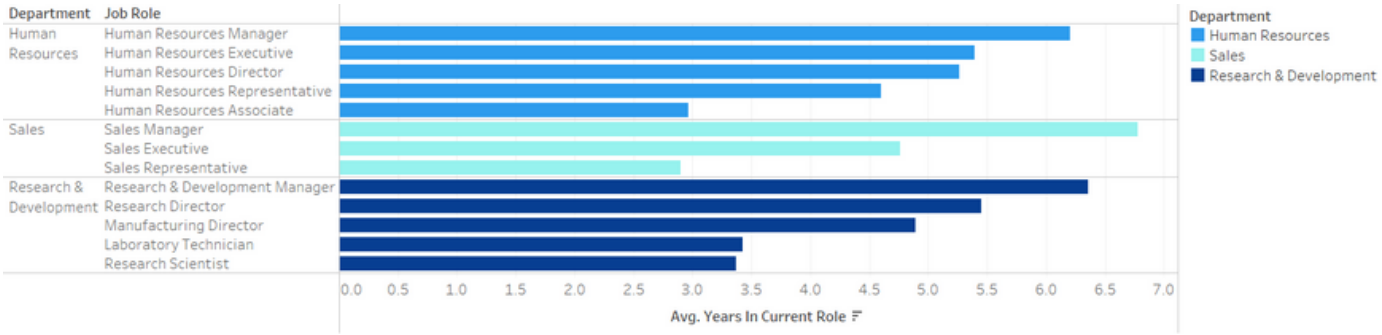
The higher median for current employees would then indicate that the company's total payroll cost for current employees is greater than for those who have resigned.
The wider range (IQR) for current employees suggests a more significant variation in total monthly payouts across different times, possibly due to fluctuations in workforce size, wage adjustments, or bonus distributions.
Outliers on the higher end for current employees might reflect months with exceptionally high payroll costs, possibly due to yearly bonuses or an increase in the workforce.

Total sum paid out by IBM in relation to Attrition

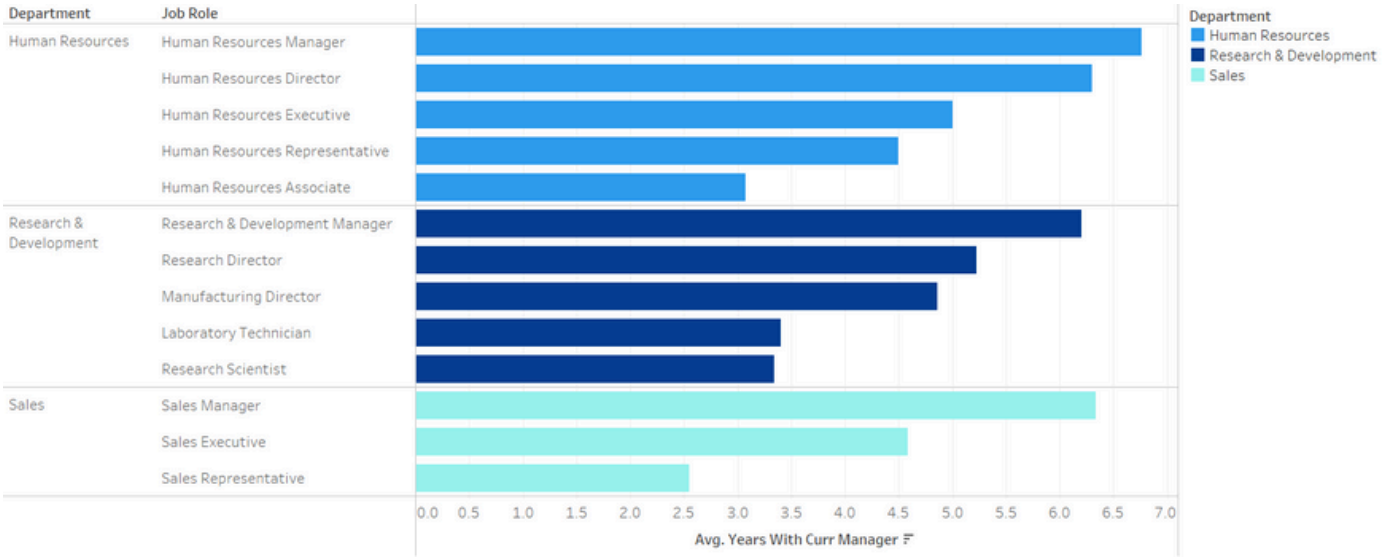


In all three departments, the majority of employees have 'Very High' or 'High' job satisfaction levels.
The Research & Development department has the highest count of employees, with most employees reporting 'Very High' job satisfaction.
Attrition seems higher among employees who have 'Low' job satisfaction, as can be seen by the proportion of red in those categories. This is most pronounced in the Sales department.
The Human Resources department has the lowest overall attrition, and a higher percentage of employees with 'Very High' satisfaction have resigned voluntarily compared to other satisfaction levels.
In the Sales department, employees with 'Medium' job satisfaction have the highest voluntary resignation rate, followed closely by those with 'Low' satisfaction.
It is interesting to note that in the Research & Development department, a small percentage of employees with 'Very High' satisfaction have also resigned voluntarily, indicating that factors other than job satisfaction might influence the decision to leave.

Average Number of Years in Current Role

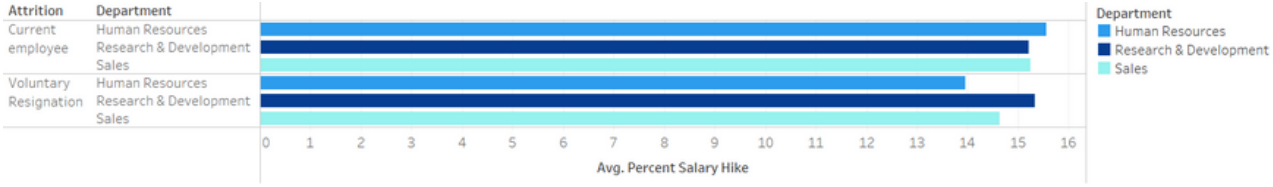


Average Yearswith Current Manager



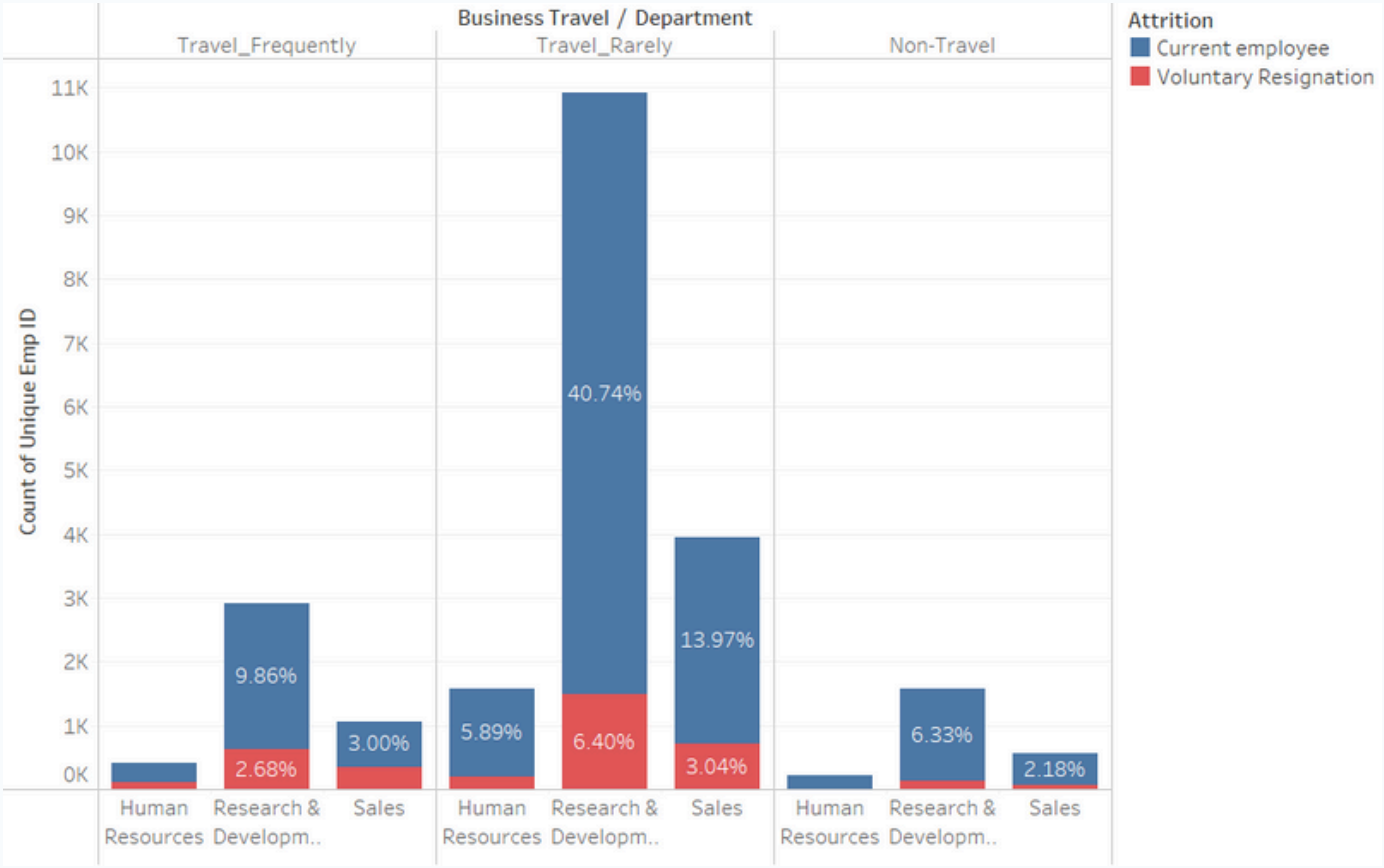
On average the Managers across all 3 departments have been in the current role for more than 6 years. Employees currently in lower level position tend to have been in this role for 3 years or more.

Avg Percentage Salary HIke



Employees who resigned voluntarily appear to have a higher average percentage salary hike compared to current employees. This could suggest that salary hikes are not enough to retain employees who already have intentions to resign.

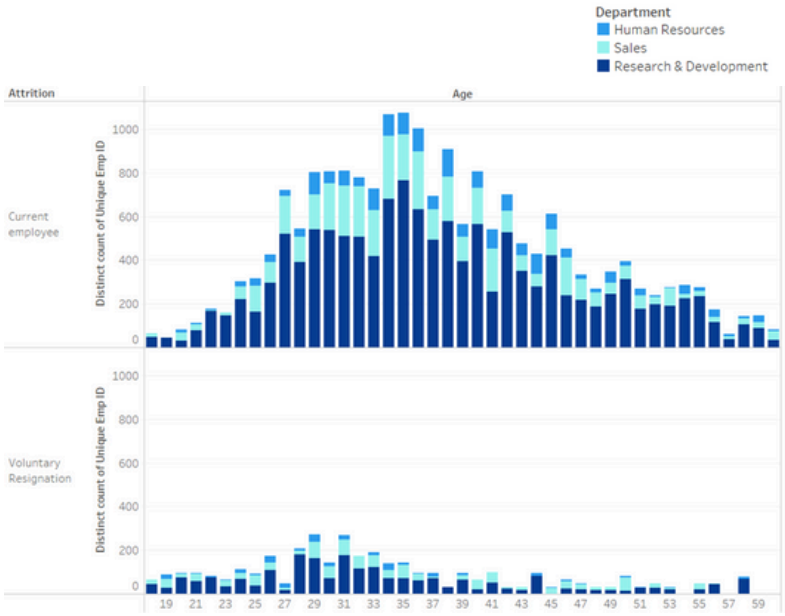
Business Travel

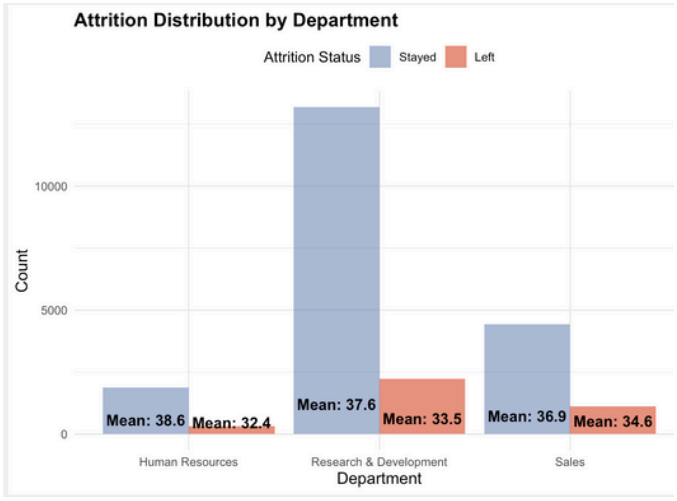


A higher rate of attrition (red bars) is visible among employees who travel frequently or rarely compared to non-travelers.

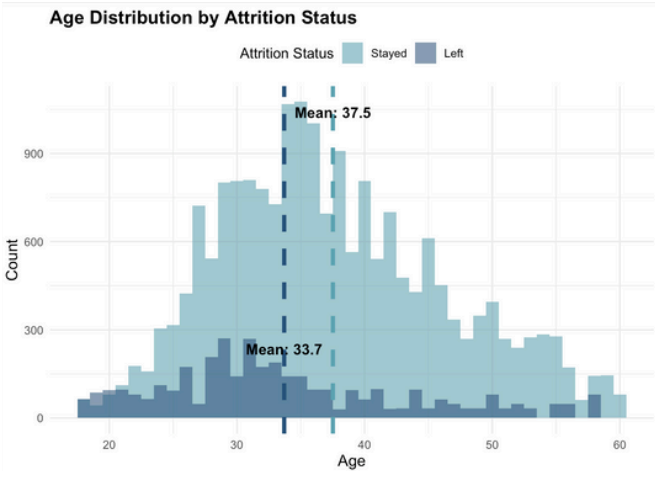
The Research and Development department has a significant portion of employees who travel rarely, and this group shows a noticeable amount of attrition.
The Sales department has a significant proportion of employees who do not travel, but the attrition rate here is moderate.

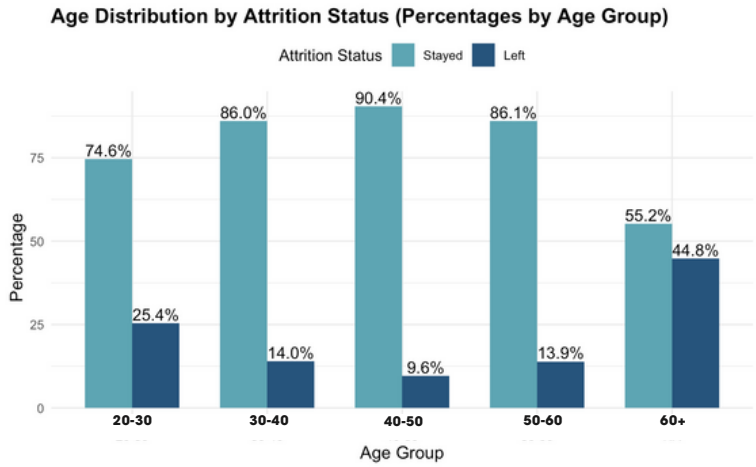
Frequent travel might contribute to employee burnout, which could explain the higher attrition rates among those who travel often.
Providing flexible travel policies or opportunities for employees who travel frequently could help reduce attrition.





The graph indicates that R&D department has a higher concentration of employees working within age range 27 – 43. Sales has a more even distribution compared to the 2 other departments.





Younger employees (20-30) have the highest attrition rates, indicating a potential need for targeted retention strategies. Conversely, the lowest attrition rates are seen in the middle age brackets, which signifies that employees who have progressed in their careers tend to stay longer. The sharp rise in attrition in the 60+ category likely correlates with employees reaching retirement age

Propensity Score Matching

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3886  347
##          1 1963  750
##
##                Accuracy : 0.6674
##                  95% CI : (0.6562, 0.6785)
##     No Information Rate : 0.8421
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2178
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6644
##             Specificity : 0.6837
##          Pos Pred Value : 0.9180
##          Neg Pred Value : 0.2764
##              Prevalence : 0.8421
##          Detection Rate : 0.5595
##    Detection Prevalence : 0.6094
##       Balanced Accuracy : 0.6740
##
##        'Positive' Class : 0
##
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## Department                2.132924  2        1.208492
## EducationField            2.535214  6        1.080607
## Age                       1.082284  1        1.040329
## BusinessTravel            1.044822  2        1.011022
## DistanceFromHome          1.230991  1        1.109500
## Education                 1.035204  1        1.017450
## EnvironmentSatisfaction   1.085291  1        1.041773
## Gender                    1.026220  1        1.013025
## JobInvolvement            1.014915  1        1.007430
## JobSatisfaction           1.021082  1        1.010486
## MaritalStatus             1.734611  2        1.147626
## MonthlyIncome             1.474362  1        1.214233
## NumCompaniesWorked        1.139318  1        1.067388
## OverTime                  1.034080  1        1.016897
## PercentSalaryHike         1.212110  1        1.100959
## PerformanceRating         1.191801  1        1.091696
## RelationshipSatisfaction  1.017034  1        1.008481
## StockOptionLevel          1.693762  1        1.301446
## TrainingTimesLastYear     1.027627  1        1.013719
## WorkLifeBalance           1.016379  1        1.008156
## YearsAtCompany            2.043510  1        1.429514
## YearsSinceLastPromotion   1.657952  1        1.287615
```

LOGISTIC REGRESSION: FULL MODEL

OverTime	p-value < 2e-16 ***
Age	p-value < 2e-16 ***
BusinessTravelTravel_Frequently	p-value < 2e-16 ***
BusinessTravelTravel_Rarely	p-value = 9.50e-10 ***
YearsInCurrentRole	p-value = 3.19e-08 ***
DailyRate	p-value = 4.69e-08 ***
DistanceFromHome	p-value = 4.67e-09 ***
EnvironmentSatisfaction	p-value = 1.42e-08 ***
MaritalStatusSingle	p-value = 1.34e-08 ***
JobInvolvement	p-value = 4.31e-10 ***
JobSatisfaction	p-value = 2.13e-05 ***
YearsSinceLastPromotion	p-value = 1.03e-06 ***
TrainingTimesLastYear	p-value = 7.32e-05 ***
PercentSalaryHike	p-value = 9.52e-05 ***
DepartmentSales	p-value = 0.002599 **

```
          Reference
Prediction    0    1
         0 3888  321
         1 1961  776
```

Accuracy : 0.6715
Sensitivity : 0.6647
Specificity : 0.7074

No Information Rate : 0.8421

LOGISTIC REGRESSION: REDUCED MODEL

OverTimeYes	p-value < 2e-16 ***
Age	p-value < 2e-16 ***
BusinessTravelTravel_Frequently	p-value < 2e-16 ***
BusinessTravelTravel_Rarely	p-value < 2e-16 ***
JobInvolvement	p-value < 2e-16 ***
EnvironmentSatisfaction	p-value < 2e-16 ***
MaritalStatusSingle	p-value < 2e-16 ***
DistanceFromHome	p-value < 2e-16 ***
JobSatisfaction	p-value < 2e-16 ***
TrainingTimesLastYear	p-value < 2e-16 ***
WorkLifeBalance	p-value = 6.79e-13 ***
MonthlyIncome	p-value = 2.60e-13 ***
PercentSalaryHike	p-value = 5.00e-13 ***
StockOptionLevel	p-value = 1.28e-09 ***
YearsSinceLastPromotion	p-value = 1.64e-08 ***

```
          Reference
Prediction    0    1
         0 3886  347
         1 1963  750
```

Accuracy : 0.6674
Sensitivity : 0.6644
Specificity : 0.6837

No Information Rate : 0.8421

LOGISTIC REGRESSION: WITH INTERACTION TERMS

OverTimeYes	p-value <2e-16 ***
Age	p-value < 2e-16 ***
BusinessTravelTravel_Frequently	p-value < 2e-16 ***
DailyRate	p-value = 1.10e-06 ***
EnvironmentSatisfaction	p-value = 1.80e-06 ***
YearsInCurrentRole	p-value = 1.88e-06 ***
JobInvolvement	p-value = 5.82e-07 ***
JobSatisfaction	p-value = 0.000820 ***
DistanceFromHome	p-value = 0.002589 **
PercentSalaryHike	p-value = 0.004372 **

Non of the variables with interaction terms resulted as significant *

```
          Actual
Predicted    0    1
        0  502  236
        1  266  532
```

Accuracy : 0.6732
Sensitivity : 0.6536
Specificity : 0.6927

No Information Rate : 0.5

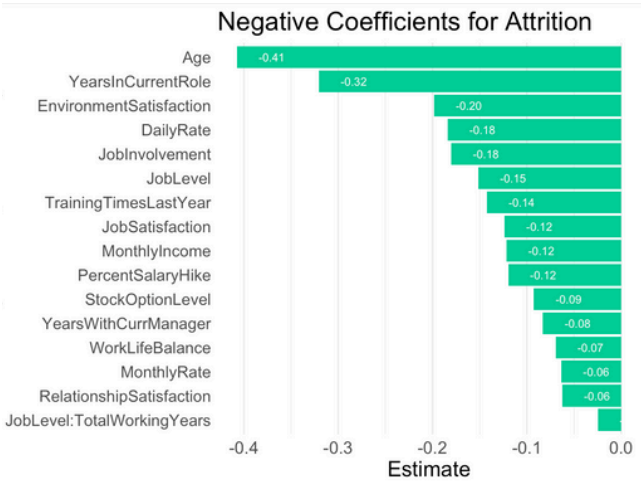
Variable Interaction for Logistic (Final model)

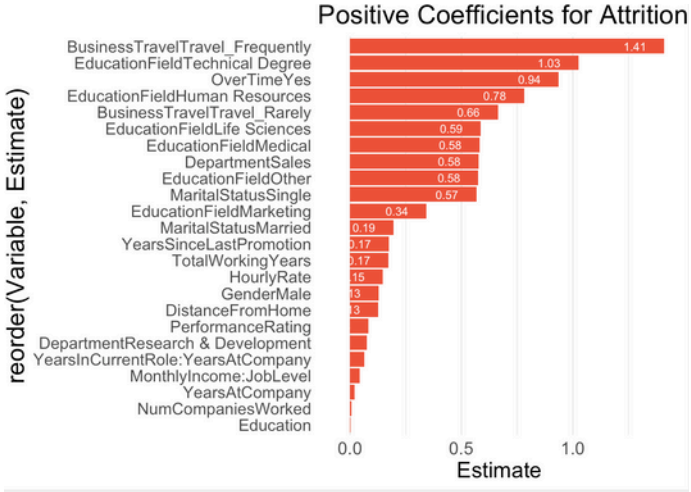
```
##
## Call:
## glm(formula = Attrition_numeric ~ Department + EducationField +
##     Age + BusinessTravel + DailyRate + DistanceFromHome + Education +
##     EnvironmentSatisfaction + Gender + HourlyRate + JobInvolvement +
##     JobSatisfaction + MaritalStatus + MonthlyRate + NumCompaniesWorked +
##     OverTime + PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +
##     StockOptionLevel + TrainingTimesLastYear + WorkLifeBalance +
##     YearsSinceLastPromotion + YearsWithCurrManager + MonthlyIncome *
##     JobLevel + TotalWorkingYears * JobLevel + YearsInCurrentRole *
##     YearsAtCompany, family = binomial(), data = balanceddata_interaction)
##
```

Conf Mat for Logistic Model with interaction

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 502 236
##          1 266 532
##
##                Accuracy : 0.6732
##                  95% CI : (0.6491, 0.6966)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.3464
##
##  Mcnemar's Test P-Value : 0.1956
##
##             Sensitivity : 0.6536
##             Specificity : 0.6927
##          Pos Pred Value : 0.6802
##          Neg Pred Value : 0.6667
##              Prevalence : 0.5000
##          Detection Rate : 0.3268
##    Detection Prevalence : 0.4805
##       Balanced Accuracy : 0.6732
##
##        'Positive' Class : 0
##
```

Coefficents for Logistic final model





Sales Subset

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2298.5  on 1657  degrees of freedom
Residual deviance: 1920.6  on 1633  degrees of freedom
AIC: 1970.6

Number of Fisher Scoring iterations: 4

         Actual
Predicted   0   1
        0 791 106
        1 306 185
[1] "Accuracy: 0.703170028818444"
```

```
##
## Call:
## glm(formula = Attrition_numeric ~ ., family = binomial(), data = train_data_balanced)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.910016   0.278052  -6.869 6.45e-12 ***
## EducationFieldMarketing        0.137648   0.126109   1.091 0.275055
## EducationFieldOther            0.286175   0.269705   1.061 0.288659
## Age                           -0.202272   0.059618  -3.393 0.000692 ***
## BusinessTravelTravel_Frequently 1.623971   0.245235   6.622 3.54e-11 ***
## BusinessTravelTravel_Rarely    0.790824   0.224385   3.524 0.000424 ***
## DistanceFromHome               0.378346   0.058956   6.417 1.39e-10 ***
## Education                     -0.187577   0.056760  -3.305 0.000951 ***
## EnvironmentSatisfaction       -0.193451   0.059834  -3.233 0.001224 **
## GenderMale                     0.096305   0.115787   0.832 0.405551
## JobInvolvement                -0.292967   0.055331  -5.295 1.19e-07 ***
## JobSatisfaction               -0.169771   0.055551  -3.056 0.002242 **
## MaritalStatusMarried          -0.178187   0.161806  -1.101 0.270795
## MaritalStatusSingle            0.775332   0.202421   3.830 0.000128 ***
## MonthlyIncome                 -0.114129   0.083884  -1.361 0.173651
## NumCompaniesWorked             0.091511   0.050069   1.828 0.067596 .
## OverTimeYes                    1.125056   0.123449   9.114  < 2e-16 ***
## PercentSalaryHike             -0.280459   0.064031  -4.380 1.19e-05 ***
## PerformanceRating              0.002623   0.066168   0.040 0.968381
## RelationshipSatisfaction      -0.109068   0.054675  -1.995 0.046060 *
## StockOptionLevel               0.136253   0.075686   1.800 0.071824 .
## TrainingTimesLastYear         -0.021633   0.056995  -0.380 0.704272
## WorkLifeBalance               -0.127593   0.058292  -2.189 0.028608 *
## YearsAtCompany                -0.205858   0.084336  -2.441 0.014649 *
## YearsSinceLastPromotion        0.370418   0.070870   5.227 1.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2298.5  on 1657  degrees of freedom
## Residual deviance: 1920.6  on 1633  degrees of freedom
## AIC: 1970.6
##
## Number of Fisher Scoring iterations: 4
```

Research and Development Subset

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2182  187
         1 1100  382

               Accuracy : 0.6658
                 95% CI : (0.6507, 0.6807)
    No Information Rate : 0.8522
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2021

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6648
            Specificity : 0.6714
         Pos Pred Value : 0.9211
         Neg Pred Value : 0.2578
             Prevalence : 0.8522
         Detection Rate : 0.5666
   Detection Prevalence : 0.6152
      Balanced Accuracy : 0.6681

       'Positive' Class : 0
```

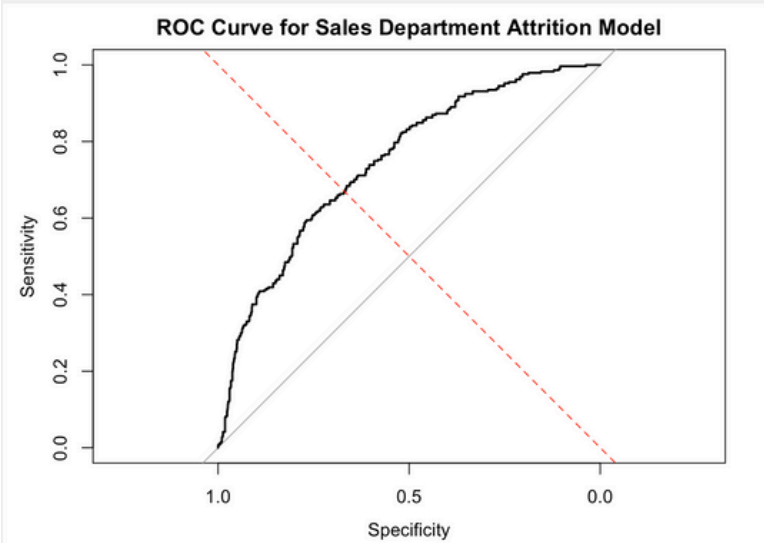
```
##
## Call:
## glm(formula = Attrition_numeric ~ ., family = binomial(), data = train_data_balanced)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.709372   0.554885  -4.883 1.05e-06 ***
## EducationFieldLife Sciences    0.900856   0.535904   1.681 0.092762 .
## EducationFieldMedical          0.909116   0.533733   1.703 0.088509 .
## EducationFieldOther            0.651873   0.558317   1.168 0.242981
## EducationFieldTechnical Degree 1.265953   0.543632   2.329 0.019875 *
## Age                           -0.438291   0.038596 -11.356  < 2e-16 ***
## BusinessTravelTravel_Frequently 1.237428   0.168963   7.324 2.41e-13 ***
## BusinessTravelTravel_Rarely    0.711752   0.154845   4.597 4.30e-06 ***
## DistanceFromHome               0.084496   0.043180   1.957 0.050369 .
## Education                     -0.030262   0.039970  -0.757 0.448972
## EnvironmentSatisfaction       -0.072961   0.039429  -1.850 0.064250 .
## GenderMale                     0.290066   0.079730   3.638 0.000275 ***
## JobInvolvement                -0.206738   0.038491  -5.371 7.83e-08 ***
## JobSatisfaction               -0.218284   0.039038  -5.592 2.25e-08 ***
## MaritalStatusMarried           0.234966   0.110855   2.120 0.034041 *
## MaritalStatusSingle            0.755564   0.139087   5.432 5.56e-08 ***
## MonthlyIncome                 -0.130092   0.046426  -2.802 0.005077 **
## NumCompaniesWorked            -0.003873   0.044468  -0.087 0.930600
## OverTimeYes                    0.870294   0.082148  10.594  < 2e-16 ***
## PercentSalaryHike              0.004104   0.041492   0.099 0.921218
## PerformanceRating              0.007081   0.041934   0.169 0.865909
## RelationshipSatisfaction      -0.041873   0.038481  -1.088 0.276526
## StockOptionLevel              -0.101840   0.050779  -2.006 0.044904 *
## TrainingTimesLastYear         -0.177414   0.039166  -4.530 5.91e-06 ***
## WorkLifeBalance               -0.078901   0.038435  -2.053 0.040087 *
## YearsAtCompany                -0.044004   0.055136  -0.798 0.424813
## YearsSinceLastPromotion       -0.020841   0.051443  -0.405 0.685389
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4585.9  on 3307  degrees of freedom
## Residual deviance: 3972.3  on 3281  degrees of freedom
## AIC: 4026.3
##
```

HR Subset

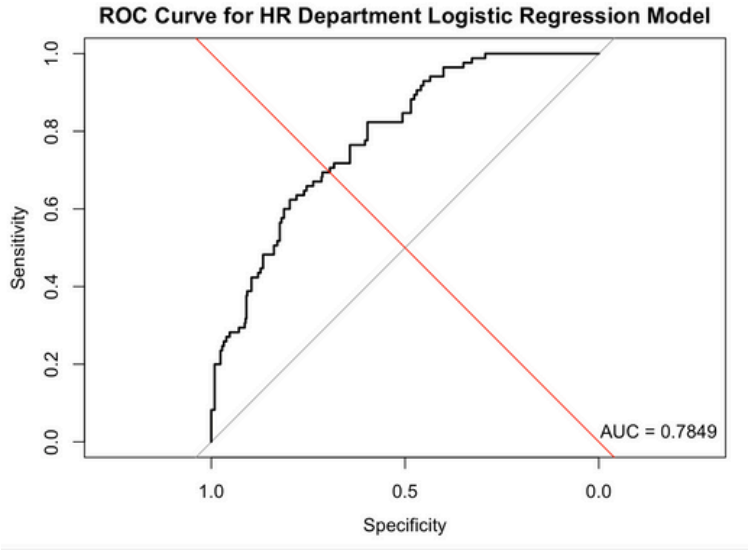
```
               Accuracy : 0.6995
                 95% CI : (0.6592, 0.7376)
    No Information Rate : 0.8452
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2557

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7004
            Specificity : 0.6941
         Pos Pred Value : 0.9259
         Neg Pred Value : 0.2980
             Prevalence : 0.8452
         Detection Rate : 0.5920
   Detection Prevalence : 0.6393
      Balanced Accuracy : 0.6973

       'Positive' Class : 0
```

```
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -0.66261    0.75415  -0.879 0.379609
## EducationFieldLife Sciences    -0.73767    0.40108  -1.839 0.065880 .
## EducationFieldMarketing        -0.72497    0.66125  -1.096 0.272916
## EducationFieldMedical           0.34875    0.46547   0.749 0.453713
## EducationFieldOther             0.44066    0.54978   0.802 0.422822
## EducationFieldTechnical Degree -0.11687    0.66201  -0.177 0.859876
## Age                            -0.78064    0.14690  -5.314 1.07e-07 ***
## BusinessTravelTravel_Frequently  1.34156    0.64306   2.086 0.036960 *
## BusinessTravelTravel_Rarely     1.01639    0.60731   1.674 0.094210 .
## DistanceFromHome                0.30631    0.14073   2.177 0.029508 *
## Education                       0.47500    0.13697   3.468 0.000525 ***
## EnvironmentSatisfaction        -0.26257    0.11520  -2.279 0.022659 *
## GenderMale                     -0.25698    0.26559  -0.968 0.333250
## JobInvolvement                 -0.16369    0.13484  -1.214 0.224775
## JobSatisfaction                 0.01171    0.12541   0.093 0.925614
## MaritalStatusMarried           -0.48239    0.33553  -1.438 0.150522
## MaritalStatusSingle            -0.52188    0.45337  -1.151 0.249687
## MonthlyIncome                  -0.50460    0.17893  -2.820 0.004800 **
## NumCompaniesWorked             -0.20397    0.12112  -1.684 0.092183 .
## OverTimeYes                     0.34287    0.26691   1.285 0.198930
## PercentSalaryHike              -0.43075    0.14070  -3.062 0.002202 **
## PerformanceRating               0.24800    0.14072   1.762 0.078005 .
## RelationshipSatisfaction       -0.20449    0.13423  -1.523 0.127656
## StockOptionLevel               -0.25464    0.15249  -1.670 0.094933 .
## TrainingTimesLastYear          -0.60331    0.15039  -4.012 6.03e-05 ***
## WorkLifeBalance                 0.07025    0.13359   0.526 0.598949
## YearsAtCompany                  0.31387    0.16130   1.946 0.051673 .
## YearsSinceLastPromotion         0.25176    0.16435   1.532 0.125557
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

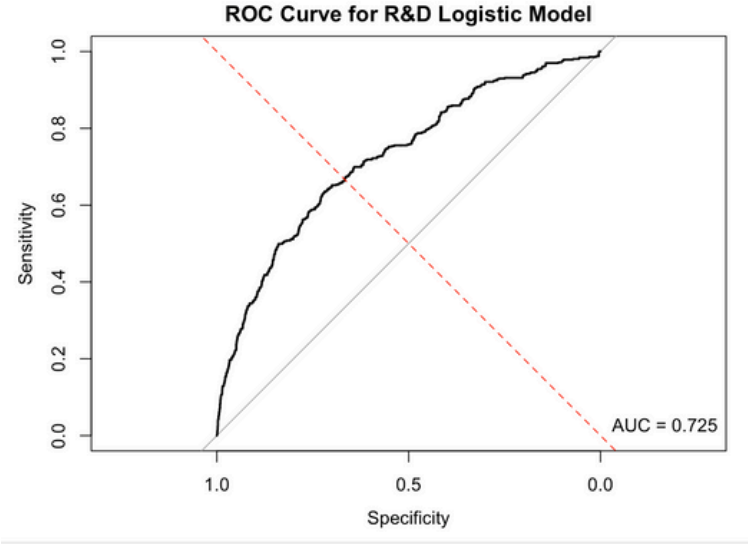
Sales ROC



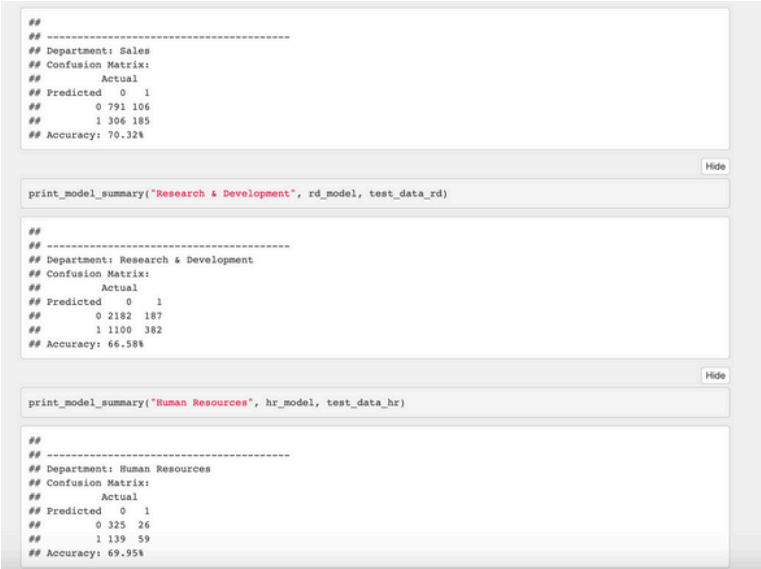
HR Subset ROC



R&D ROC



Comparison of Subsets (HR, Sales and R&D)



```
##
## ----------------------------------------
## Department: Sales
## Confusion Matrix:
##          Actual
## Predicted   0   1
##         0 791 106
##         1 306 185
## Accuracy: 70.32%
```

Hide

```
print_model_summary("Research & Development", rd_model, test_data_rd)
```

```
##
## ----------------------------------------
## Department: Research & Development
## Confusion Matrix:
##          Actual
## Predicted    0    1
##         0 2182  187
##         1 1100  382
## Accuracy: 66.58%
```

Hide

```
print_model_summary("Human Resources", hr_model, test_data_hr)
```

```
##
## ----------------------------------------
## Department: Human Resources
## Confusion Matrix:
##          Actual
## Predicted   0   1
##         0 325  26
##         1 139  59
## Accuracy: 69.95%
```

Department Wise Regressions

Department wise (Human Resources, Sales and Research and Development) Logistic Regressions

Hide

```
# Split the subset data into three separate data frames, one for each department
hr_subset <- subset(ab_dataset_reduced, Department == "Human Resources")
sales_subset <- subset(ab_dataset_reduced, Department == "Sales")
rd_subset <- subset(ab_dataset_reduced, Department == "Research & Development")

# Function to check factor levels in each subset, revised to exclude non-factors and return only factors
check_factor_levels <- function(data) {
  lapply(data[, sapply(data, is.factor)], function(x) length(levels(x)))
}

# Apply to each department
hr_levels <- check_factor_levels(hr_subset)
sales_levels <- check_factor_levels(sales_subset)
rd_levels <- check_factor_levels(rd_subset)
```

Department wise (Human Resources, Sales and Research and Development) Logistic Regressions

Hide

```
# Split the subset data into three separate data frames, one for each department
hr_subset <- subset(ab_dataset_reduced, Department == "Human Resources")
sales_subset <- subset(ab_dataset_reduced, Department == "Sales")
rd_subset <- subset(ab_dataset_reduced, Department == "Research & Development")

# Function to check factor levels in each subset, revised to exclude non-factors and return only factors
check_factor_levels <- function(data) {
  lapply(data[, sapply(data, is.factor)], function(x) length(levels(x)))
}

# Apply to each department
hr_levels <- check_factor_levels(hr_subset)
sales_levels <- check_factor_levels(sales_subset)
rd_levels <- check_factor_levels(rd_subset)
```

LOGISTIC REGRESSION: RESULTS FROM DEPARTMENT WISE ANALYSIS

- Similar for all 3 Departments
- Similar for Sales and R&D
- Similar for Sales & HR
- Similar for R&D and HR

SALES	P Value	Research & Development	P Value	HR	P Value
OverTimeYes	p< 2e-16 ***	Age	p < 2e-16 ***	TrainingTimesLastYear	p = 6.03e-05 ***
BusinessTravelTravel_Frequently	p = 3.54e-11 ***	OverTimeYes	p < 2e-16 ***	Age	p = 1.07e-07 ***
DistanceFromHome	p = 1.39e-10 ***	BusinessTravelTravel_Frequently	p = 2.41e-13 ***	MonthlyIncome	p = 0.004800 **
YearsSinceLastPromotion	p = 1.73e-07 ***	JobSatisfaction	p = 2.25e-08 ***	PercentSalaryHike	p = 0.002202 **
JobInvolvement	p = 1.19e-07 ***	MaritalStatusSingle	p = 5.56e-08 ***	Education	p = 0.000525 ***
PercentSalaryHike	p = 1.19e-05 ***	JobInvolvement	p = 7.83e-08 ***	EnvironmentSatisfaction	p = 0.022659 *
BusinessTravelTravel_Rarely	p = 0.000424 ***	TrainingTimesLastYear	p = 5.91e-06 ***	DistanceFromHome	p = 0.029508 *
EnvironmentSatisfaction	p = 0.001224 **	BusinessTravelTravel_Rarely	p = 4.30e-06 ***	BusinessTravelTravel_Frequently	p = 0.036960 *
JobSatisfaction	p = 0.002242 **	MonthlyIncome	p = 0.005077 **	PerformanceRating	p = 0.078005 .
MaritalStatusSingle	p = 0.000128 ***	GenderMale	p = 0.000275 ***	BusinessTravelTravel_Rarely	p = 0.094210 .
Age	p = 0.000692 ***	MaritalStatusMarried	p = 0.034041 *	StockOptionLevel	p = 0.094933 .
Education	p = 0.000951 ***	WorkLifeBalance	p = 0.040087 *	NumCompaniesWorked	p = 0.092183 .
YearsAtCompany	p = 0.014649 *	StockOptionLevel	p = 0.044904 *	YearsAtCompany	p = 0.051673 .
RelationshipSatisfaction	p = 0.046060 *	EducationFieldTechnical Degree	p = 0.019875 *	EducationFieldLife Sciences .	p = 0.065880
WorkLifeBalance	p = 0.028608 *	EnvironmentSatisfaction	p = 0.064250 .	YearsSinceLastPromotion	p = 0.125557

```
--------------------
Department: Sales
Confusion Matrix:
          Actual
Predicted   0   1
        0 791 106
        1 306 185
Accuracy: 70.32%
```

```
---------------------------------
Department: Research & Development
Confusion Matrix:
          Actual
Predicted    0    1
        0 2182  187
        1 1100  382
Accuracy: 66.58%
```

```
---------------------------
Department: Human Resources
Confusion Matrix:
          Actual
Predicted   0   1
        0 325  26
        1 139  59
Accuracy: 69.95%
```

10 fold cross validation for DT model

10 Fold Cross validation

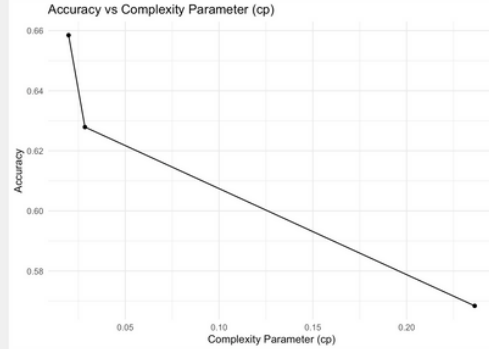
```r
# Set up cross-validation settings
train_control <- trainControl(
  method = "cv",          # Use k-fold cross-validation
  number = 10,            # Number of folds in k-fold cross-validation
  savePredictions = "final",
  classProbs = TRUE       # Store class probabilities
)

# Train the model with cross-validation
set.seed(123)  # For reproducibility
decision_tree_model_cv <- train(
  Attrition_numeric ~ .,   # Formula
  data = trainingdata_balanced_ab,  # Training data
  method = "rpart",        # Training method (decision tree)
  trControl = train_control,  # Control object
  metric = "Accuracy"      # Performance metric
)

# Print the model summary
print(decision_tree_model_cv)
```

```
## CART
##
## 27300 samples
##    22 predictor
##     2 classes: 'X0', 'X1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 24570, 24570, 24570, 24570, 2457
## Resampling results across tuning parameters:
##
##   cp          Accuracy   Kappa
##   0.01978022  0.6584982  0.3169963
##   0.02849817  0.6279121  0.2558242
##   0.23626374  0.5683883  0.1367766
##
## Accuracy was used to select the optimal model using the l
## The final value used for the model was cp = 0.01978022.
```

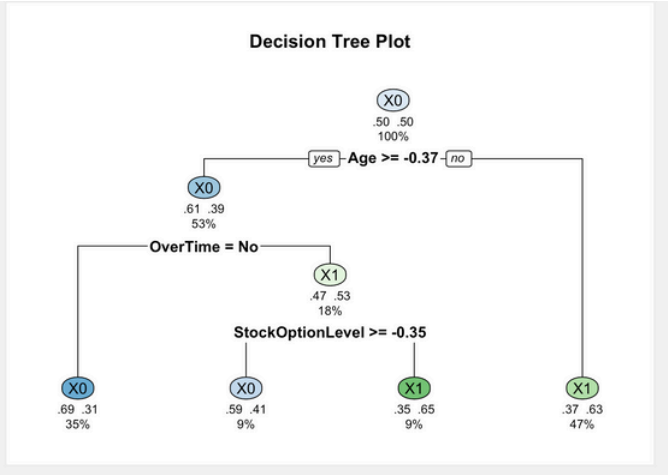
Final Decision Tree model

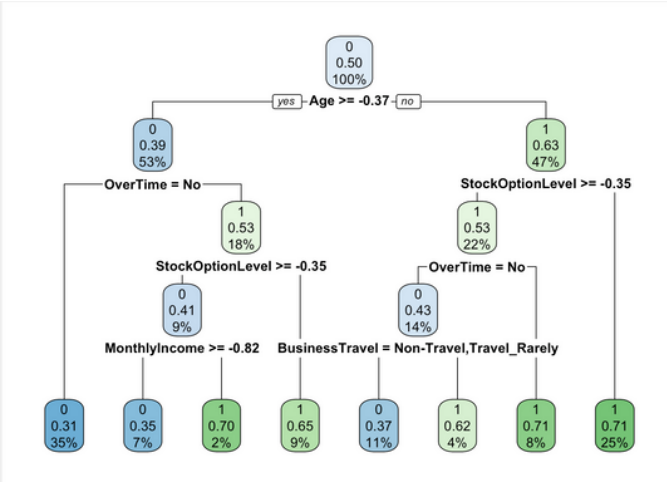
```
# Confusion Matrix and Statistics
#
#           Reference
# Prediction   X0   X1
#         X0 9504 5177
#         X1 4146 8473
#
#                Accuracy : 0.6585
#                  95% CI : (0.6528, 0.6641)
#     No Information Rate : 0.5
#     P-Value [Acc > NIR] : < 2.2e-16
#
#                   Kappa : 0.317
#
#  Mcnemar's Test P-Value : < 2.2e-16
#
#             Sensitivity : 0.6963
#             Specificity : 0.6207
#          Pos Pred Value : 0.6474
#          Neg Pred Value : 0.6714
#              Prevalence : 0.5000
#          Detection Rate : 0.3481
#    Detection Prevalence : 0.5378
#       Balanced Accuracy : 0.6585
#
#        'Positive' Class : X0
#
```

Decision Tree Accuracy vs CP

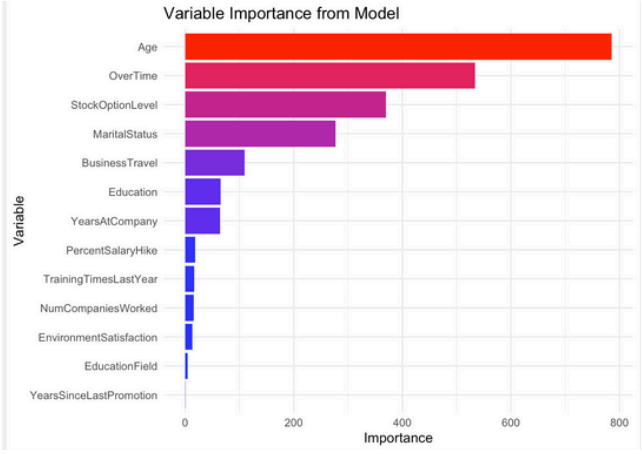


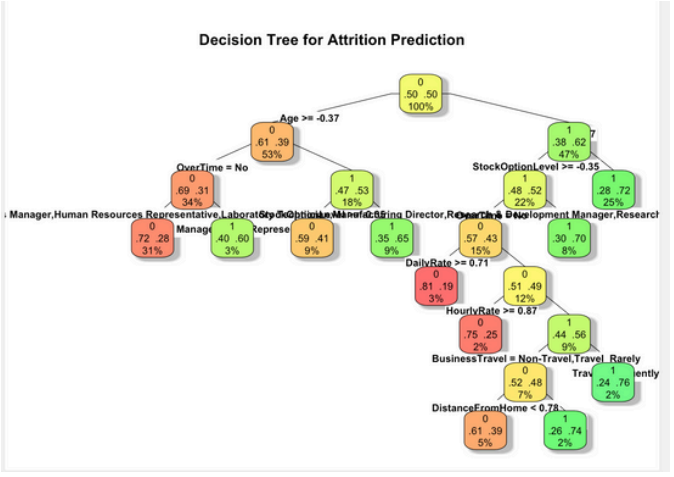
DT Models





DT model Variable importance





DT for Subsets

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 421   4
##          1 143  90
##
##                Accuracy : 0.7766
##                  95% CI : (0.7428, 0.8079)
##     No Information Rate : 0.8571
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.4355
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7465
##             Specificity : 0.9574
##          Pos Pred Value : 0.9906
##          Neg Pred Value : 0.3863
##              Prevalence : 0.8571
##          Detection Rate : 0.6398
##    Detection Prevalence : 0.6459
##       Balanced Accuracy : 0.8520
##
##        'Positive' Class : 0
##
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1029  116
##          1  301  220
##
##                Accuracy : 0.7497
##                  95% CI : (0.7282, 0.7703)
##     No Information Rate : 0.7983
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3553
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7737
##             Specificity : 0.6548
##          Pos Pred Value : 0.8987
##          Neg Pred Value : 0.4223
##              Prevalence : 0.7983
##          Detection Rate : 0.6176
##    Detection Prevalence : 0.6873
##       Balanced Accuracy : 0.7142
##
##        'Positive' Class : 0
##
```

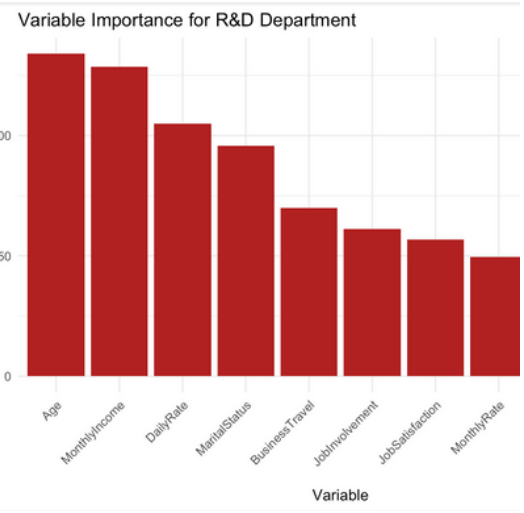
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2876  220
##          1 1078  446
##
##                Accuracy : 0.719
##                  95% CI : (0.7058, 0.732)
##     No Information Rate : 0.8558
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2585
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7274
##             Specificity : 0.6697
##          Pos Pred Value : 0.9289
##          Neg Pred Value : 0.2927
##              Prevalence : 0.8558
##          Detection Rate : 0.6225
##    Detection Prevalence : 0.6701
##       Balanced Accuracy : 0.6985
##
##        'Positive' Class : 0
##
```

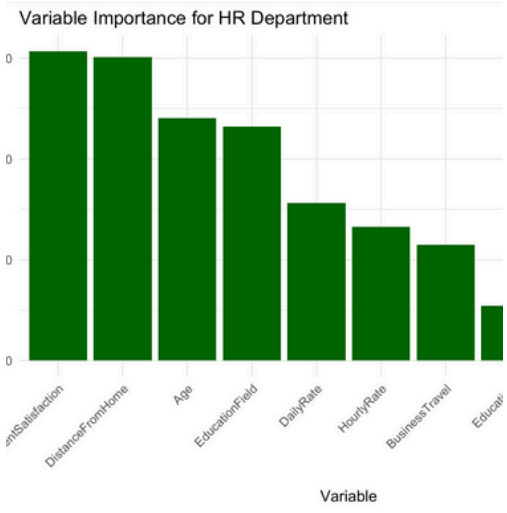
Sales, HR and R&D subsets variable importance

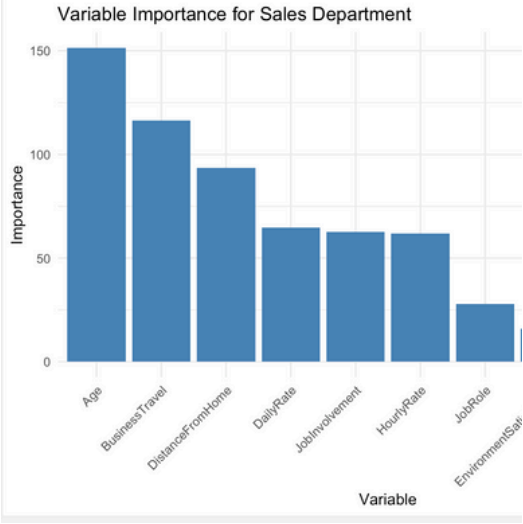
	Overall
Age	62.367579
BusinessTravel	22.913872
DistanceFromHome	69.957259
Education	16.652117
EducationField	42.574613
EnvironmentSatisfaction	63.398729
JobInvolvement	9.319177
JobSatisfaction	12.407505
MaritalStatus	3.758586
MonthlyIncome	27.340278
NumCompaniesWorked	2.403565
PercentSalaryHike	22.899110
PerformanceRating	3.953970
RelationshipSatisfaction	15.531891
StockOptionLevel	29.644549
TrainingTimesLastYear	43.370503
YearsAtCompany	29.429738
YearsSinceLastPromotion	3.508567
Department	0.000000
Gender	0.000000
OverTime	0.000000
WorkLifeBalance	0.000000

	Overall
Age	160.693968
BusinessTravel	69.933187
DistanceFromHome	6.121094
Education	7.423855
JobInvolvement	52.371749
JobSatisfaction	34.568477
MaritalStatus	95.858599
MonthlyIncome	128.497173
NumCompaniesWorked	8.517866
OverTime	185.272132
PercentSalaryHike	19.806168
StockOptionLevel	107.883099
TrainingTimesLastYear	45.172356
YearsAtCompany	26.572917
YearsSinceLastPromotion	10.655467
Department	0.000000
EducationField	0.000000
EnvironmentSatisfaction	0.000000
Gender	0.000000
PerformanceRating	0.000000
RelationshipSatisfaction	0.000000
WorkLifeBalance	0.000000

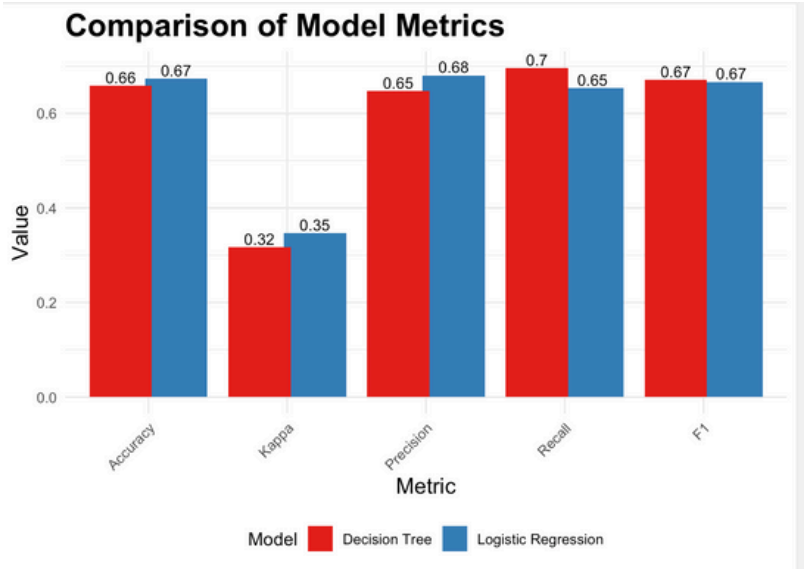
	Overall
Age	158.917796
BusinessTravel	107.007484
DistanceFromHome	83.735931
Education	26.841194
EnvironmentSatisfaction	25.082053
JobInvolvement	65.167166
JobSatisfaction	44.118585
MaritalStatus	52.295666
MonthlyIncome	95.513794
NumCompaniesWorked	24.978251
OverTime	36.146428
PercentSalaryHike	35.233779
PerformanceRating	2.892857
StockOptionLevel	15.059482
TrainingTimesLastYear	30.083053
YearsAtCompany	59.115339
YearsSinceLastPromotion	16.635839
Department	0.000000
EducationField	0.000000
Gender	0.000000
RelationshipSatisfaction	0.000000
WorkLifeBalance	0.000000







Final Models comparison



ROC for final DT and Logit Models