HR Analytics: IBM Employee Attrition Analysis

Section C TEAM 53: Jamelia Gordon, Zixiao Xu, Xinyi Zhou, Ayush Pagaria

Business Understanding

The Business Problem

Employees leaving can lead to a talent drain, with loss of innovative ideas to competing rivals, which is particularly detrimental for the tech industry. Employee attrition can be problematic as it often reduces talent within the company and negatively impacts the operations of a company as the company allocated time, money, and resources in training employees. Therefore, if they leave, they will further need to re-allocate the same efforts into another new employee. It is therefore vital for companies to study employee attrition to identify and address problematic issues for their employees. For example, a high attrition rate could be from employees leaving due to a poor workplace culture. Only by investigating the reasons for this employee attrition can management make changes to improve the organization's work culture for other employees (Wooll, 2022).

International Business Machine (IBM)

IBM is a multinational technology company that provides services including consulting, software, infrastructure, and ecosystem support with over 345,000 employees globally.

For our analysis, we are considering a database that contains employee attrition at IBM, where we would review the factors that may contribute to employee attrition. Based on the data, we noticed that the attrition rate was approximately **16%** (247 observations out of 1470)

Use Scenario

IBM's Human Resources (HR) Department can use the results of this project to enhance their decision-making in implementing strategies to reduce employee attrition. In particular, this project will help them identify potential reasons for attrition in advance and help gain a deeper understanding of the root causes

driving attrition and how IBM can successfully deploy strategies to target selected employees that are most likely to leave.

Targeting Strategy

Causes
Retirement or Devoid of Life
Position Elimination/Termination
Pay Satisfaction
Lack of Career Development Opportunities
Poor Work-life Balance
Lack of Sense of Belonging
Poor Workplace Culture
Inadequate Benefits
Lack of Professional Development Opportunities

Deployment
Offer Employee Training Programs
Increase Employee Benefits
Review Pay Structure
Focus on Employee Well-Being
Contribute to Career Growth & Planning

Based on the information available to us, we further identified potential causes of employee attrition at IBM and identified potential deployment strategies to overcome such instances of attrition in the future. We believe introducing an employee allowance is the most appropriate deployment strategy as it tackles, to some extent, the majority of the causes mentioned above and is therefore, the most effective course of action.

The Core Business Question

Given the size of the firm, it is not possible to target all potential employees that may leave the firm as there are many factors for attrition, some beyond the control of IBM. Therefore, our deployment strategy has to target employees that are most likely to leave, and given the specialized employee benefits, they may choose not to leave the firm.

The Data Mining Solution

Our approach to analysis is rooted in supervised learning techniques, with the aim of addressing the issue of employee attrition within IBM. We intend to leverage a diverse set of machine learning models,

encompassing logistic regression, decision trees, random forests, and Lasso regression. Our primary objective is to predict employee job satisfaction and, in doing so, to explore effective strategies for proactively mitigating attrition, thereby retaining valuable talent within the organization.

These predictive models will play a pivotal role in identifying employees who exhibit a higher propensity to leave IBM. This early warning system, derived from the predictive capabilities of our models, will empower us to take preemptive measures to engage and retain these at-risk employees. Subsequent in-depth analyses will be undertaken to ascertain which specific employee benefits, policies, or interventions might have the most significant potential to reduce attrition rates.

Our approach is underpinned by the aspiration to extract actionable insights from the data. This data-driven strategy will enable IBM to not only gain a deeper understanding of the dynamics behind employee turnover but also to deploy tailored solutions aimed at enhancing job satisfaction and increasing employee retention. Ultimately, our goal is to fortify the organization's long-term stability by addressing the underlying concerns related to employee attrition.

Data Understanding

What is the Dataset?

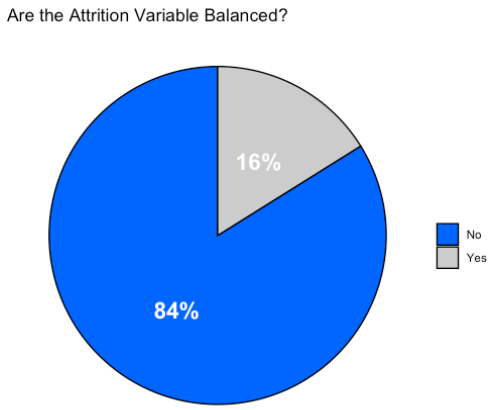
The provided dataset contains information about 1470 employees' attribution and performance in IBM, including personal details (such as age, gender, marital status, education, commute), job-related information (such as job role, department, workload, job satisfaction, salary level, work experience), and company-related data (such as work location, business travel). 35 different variables are collected. The data unit is recorded by a single employee and related to their personal information and work performance. It was collected from Kaggle website, and the data period is 2018. These data can be used to address various issues related to employee attrition and organizational management.

Potential Bias

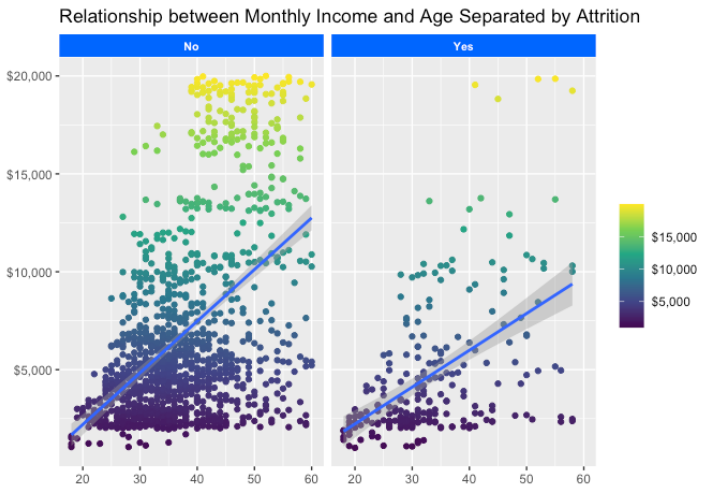
The dataset is entirely fictional and was created by IBM data scientists. While it may not accurately reflect real-world behaviors and factors that lead to employee turnover, it still provides valuable insights into these aspects. Importantly, this dataset does not exhibit a clear optimism bias because it is not influenced by employee concerns about retaliation. Additionally, the use of synthetic data can protect employee privacy during the development phase of the model, thereby reducing potential privacy violations.

Exploratory Data Analysis

Pie Chart on Target Variable: The pie chart representing our target variable 'Attrition' provides an overview of its distribution. It shows that 84% of employees have chosen to stay with the company, whereas 16% have left. This observation highlights the imbalance within our dataset, a critical factor guiding us in selecting the most effective strategy for constructing our predictive model.



Attrition by Monthly Income and Age: The scatter plots are categorized by those who remained with the company and those who departed. A notable observation emerges when comparing the two plots: a significant proportion of employees who left the company have lower monthly incomes, irrespective of their ages. However, both plots reveal a positive association between monthly income and age. It is noteworthy that five employees chose to leave despite having high monthly incomes of around $20000 and being above the age of 40.



Data Preparation

Data Cleaning

We first looked for rows or columns containing missing values or N/A values, and luckily there is no missing values in this dataset. Secondly, depending on the specific requirements for our modeling, unnecessary variables like 'Over18', 'EmployeeCount', and 'StandardHour' have been removed because they might not contribute significantly to the IBM attrition analysis. Thus, in data cleaning process, they are removed to simplify the dataset.

Lastly, dummy variables are assigned to better fit in our data analysis in future regression, classification, and other modeling techniques. To be specific, categorial data like 'Gender', 'Marital Status' , whether to take business travel, work over time or not and Attrition or not, altogether 6 key variables are assigned into binary dummy variables (True = 1 and False = 0). And for marital status, we coded single equals to 1, married equals 2 and divorced equals to 3. As is shown in the table below.

Gender		**OverTime**		**MaritialStatus**		**BusinessTravel**		**Attrition (Target)**	
Male	1	Yes	1	Single	1	Travel_Rarely	1	Yes	1
Female	0	No	0	Married	2	Travel_Frequently	2	No	0
				Divorced	3	Non-Travel	3		

Modelling

Mathematical Framework

Uncertainty: Employee exits or Employee stays

Deployment: Target employee with employee benefits

Goal: $E[Profit|X, D] = P(Employee\ stays \mid \text{X, D}) * \text{V(X, D)}$

Model Selection

We've selected four distinct modeling approaches: Logistic regression, Classification tree, Random Forest, and LASSO because of the regression nature of our target variable Attrition. Each of these methodologies has its own set of advantages and limitations when applied to the context of understanding and predicting employee attrition.

1. Logistic regression models the relationship between employee characteristics and the likelihood of attrition. It is ideally suited for binary classification, making it a valuable tool for forecasting whether an employee will stay or leave, but it may not capture complex and non-linear patterns frequently observed in the dataset because it assumes linear relationship.
2. Classification trees divide data into homogeneous groups based on predictor variables. It can capture complex and non-linear relationships and interactions among predictors, but it can lead to overfitting, which may result in a model that is too complex.
3. Random forest is an ensemble of multiple decision trees to enhance predictive accuracy and reduce overfitting, and it is more effective in capturing complex and non-linear attrition patterns.
4. LASSO builds on logistic regression and improves overfitting in models by selecting features, and it helps reducing multicollinearity issues common in employee turnover datasets.

Solving the Business Problem and Improving Profitability

Preventative Measures: By leveraging historical data, the models predict which employees are at risk of leaving soon. This proactive approach allows the company to take preventive measures to retain valuable talent.

Cost Reduction: By reducing attrition, the company can lower recruitment and onboarding costs, which can be substantial. This contributes to cost savings and operational efficiency.

Cost Efficiency: Retained talent tends to be more productive over time as they become more familiar with the organization, its processes, and their roles. By reducing attrition, the company contributes to its bottom line in the long run.

Evaluation

We evaluated each model through a comprehensive four-step analysis, assessing accuracy and performance using in-sample accuracy, out-of-sample accuracy, False Negative Rate (FNR), and Area Under the Curve (AUC).

In-Sample Accuracy

Comparing the accuracy of each model, Random Forest showed the highest accuracy at 0.999, followed by Decision Tree at 0.896 and Logistic Regression at 0.880. However, given the potential overfitting risk of Random Forest, further scrutiny was focused on out-of-sample performance.

Out-of-Sample Accuracy

Visualizing out-of-sample accuracy revealed that the Logistic model performed the best at 0.859, with Lasso Regression ranking second and Random Forest third.

False Negative Rate (FNR)

The False Negative Rate (FNR) is a critical metric in classification model evaluation. It quantifies the proportion of instances that are incorrectly predicted as negative (i.e., the model predicts the employee

will stay at the company) among all the actual positive instances. Among the models we assessed, Lasso Regression exhibited the lowest FNR at 0.00824, followed by Random Forest at 0.011, Logistic Regression at 0.0357, and Decision Tree at 0.0934. These low FNR values across all four models signify their strong performance.

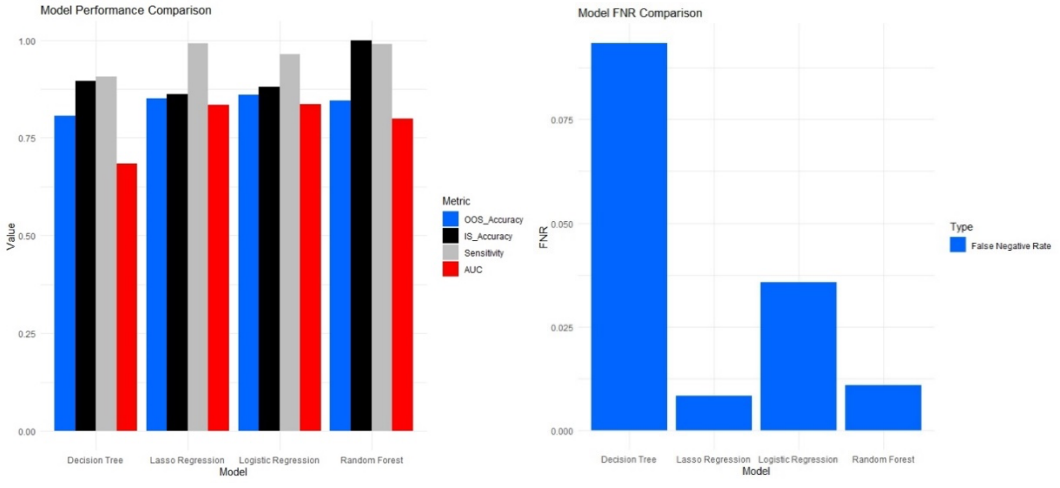
Area Under the Curve (AUC)

AUC values were calculated to assess model performance in classifying positive and negative instances across various thresholds. Lasso (0.835) and Logistic (0.837) demonstrated superior performance among the four models.

Considering these evaluations, we prioritize resources toward retaining employees who may be considering leaving to minimize false negatives, which can have significant costs. Therefore, we place particular emphasis on the False Negative Rate (FNR) as a critical metric to test model effectiveness, and we have found that **LASSO stands out as the top-performing model**.

Model Performances:

	In-sample Accuracy	OOS Accuracy	FNR	AUC
Logistic	0.880	0.859	0.0357	0.837
Decision Tree	0.896	0.807	0.0934	0.684
Random Forest	0.999	0.846	0.0110	0.798
LASSO	0.862	0.850	0.00824	0.835



Cost-Benefit Analysis

	Employee plans on Leaving	Employee Stays
Give Employee Benefit	**0.25**	**-0.05**
Don't Give Employee Benefit	**-0.2**	**0**

We view the likelihood of not extending an allowance to a potential "exiter" is particularly costly relative to the cost of wasting resources to target an employee who had no intention of leaving. On the other hand, we believe our efforts to retain an employee provides substantial benefits to the company even when we take into account the possibility that an employee, even after receiving benefits, decides to exit.

$E[Profit] = 0.25p - 0.2p - 0.05(1-p) \leftrightarrow p > 0.5$

Therefore, targeting becomes profitable and feasible when the probability of an employee leaving exceeds 50%.

The Lasso Regression model is the most suitable for our purposes, and by utilizing the respective confusion matrix, we estimated an expected profit of 0.05. While this positive expected profit is relatively marginal, it holds viable financial benefits for the firm. This suggests that the model is indeed viable for deployment in predicting attrition at IBM. However, the firm should be cautious about allocating excessive resources based on this return. Careful consideration is needed to strike the right balance between the potential benefits and resource allocation when implementing the model.

Deployment

In building our models we considered the following strategies:

1. **Directed approach:** Choose a suitable deployment based on the specific needs of each employee.

Disadvantages: Challenging to implement at scale

2. **One size fits all approach:** Choose 1 deployment strategy that caters to the gross majority of employees.

Disadvantages: May not address the underlying cause for a particular employee's desire to exit

Justification: Overall, we consider implementing an employee allowance to be the most suitable deployment strategy. It addresses many of the common attrition factors and allows us to get the most value for our investment, considering the limited profit margin. While it may not justify extensive resource allocation, the profit margin is substantial enough to warrant attention. Focusing our resources across the entire employee base proves to be the most effective approach.

Implementation

The data mining model will act as a predictive tool to identify employees at risk of leaving soon. This will assist the HR department in spotting potential departures and initiating initial discussions to assess the situation. With this insight, HR can make informed decisions when recommending the allowance package.

Risk and Cautions

Employee Segmentation: The firm needs to be wary of categorizing employees when considering the implementation of allowance packages. It is essential to use objective, job-related criteria for employee recommendations to avoid discrimination. HR professionals play a critical role in ensuring a fair recommendation process, guaranteeing equitable treatment for all employees. This approach allows for effective employee segmentation while upholding principles of fairness and objectivity throughout the process.

Human Bias: Human beings are naturally inclined to make decisions influenced by preconceived judgments or unconscious preferences for certain subgroups. Therefore, it's crucial to conduct regular reviews of these selections, identify potential biases, and provide ongoing training for our staff to be able to recognize and address biases in decision-making strategies.