Introduction to Data Science
DS501

Case Study 4

IBM HR Analytics – Business Consulting

Submitted By:

Chithramvel Sanarpalayam Selvamuthukumar
Sreeram Marimuthu
Nitya Phani Oruganty Santosh

Index:

1. Abstract
2. Introduction
3. The Business Part
    a. Your business problem to solve:
    b. Why the problem is important to solve?
    c. What is your idea to solve the problem?
    d. What differences you could make with your data science approach?
    e. Why do you believe the idea deserves the financial resources of your company?
    f. Business Proposition
    g. Importance & Motivations

4. Data Gathering
    a. Dataset Source and Relevance
    b. Features Exploration
    c. Data Pre-processing for Analysis
    d. Balancing the Dataset

5. Conjectures
    a. Conjecture 1
    b. Conjecture 2
    c. Conjecture 3
    d. Conjecture 4

6. Exploratory Data Analysis (EDA)
    a. Conjecture Evaluation
    b. New Findings

7. Machine Learning Models
    a. Decision Tree
    b. RandomForest
    c. XGBoost
    d. Logistic Regression

8. Model Integration with Business Proposition

9. The Story of the Group

10. Acknowledgments

11. Summary and Conclusion

1. Abstract:

This case study explores the application of data science methodologies to address the prevalent business challenge of employee attrition. One of the biggest assets for any large company is their employees. When organizations put so much effort and invest substantially in acquiring the best talents and keeping their workforce happy, it is important to keep track and keep their retention rates stable. Employees leaving frequently despite everything seemingly going well can often lead to companies wasting time and effort in the wrong areas and also lose productivity in training new hires. The study employs machine learning models, including Logistic Regression and XGBoost, to develop predictive tools for proactively managing and mitigating attrition risks. The results demonstrate the potential of data-driven approaches to enhance organizational decision-making, improve workforce stability, and ultimately contribute to the long-term success of the company.

2. Introduction:

In the contemporary business landscape, employee attrition poses a significant challenge to organizations across industries. The loss of skilled and experienced personnel not only impacts productivity but also incurs substantial costs associated with recruitment and training. This case study delves into the proactive use of data science to address the intricate problem of employee attrition.

The selected dataset, derived from HR records, forms the basis for our exploratory data analysis (EDA). Using the IBM HR Analytics dataset, we can perform exploratory analysis to identify the key factors that lead to employees leaving a company and also those that improve the likelihood of retaining them. Using this information and keeping track of certain metrics possibly through monthly employee surveys, we can then build a machine learning model that can automatically predict certain employees that are likely to start looking for work elsewhere or leave the company, and suggest the specific areas to focus HR involvement in to increase the chances of retaining them. An extra feature could be the detection of high value employees (not necessarily the highest paid) and the cost-benefit of retaining them.

The results presented herein contribute to the growing body of knowledge on leveraging data science to address critical business challenges, particularly in the realm of human resource management.

3. The Business Part:

3. a. Your business problem to solve:

Chosen Market: Business Consulting

One of the biggest assets for any large company is their employees. When organizations put so much effort and invest substantially in acquiring the best talents and keeping their workforce happy, it is important to keep track and keep their retention rates stable.

Employees leaving frequently despite everything seemingly going well can often lead to companies wasting time and effort in the wrong areas and also lose productivity in training new hires. Therefore, there needs to be a data driven approach that can accurately determine the factors leading to employee attrition and develop a system that can help HR departments predict employee attrition in advance, and direct their involvement towards specific aspects that will increase the chances of retaining them. A model that provides this functionality will be a valuable service that could potentially be offered by a business consulting firm.

The business problem we aim to address is employee attrition within our company. Employee attrition can have significant negative impacts on productivity, morale, and overall organizational performance. Understanding the factors contributing to attrition and developing strategies to mitigate it is crucial for maintaining a healthy and stable workforce.

3. b. Why the problem is important to solve?

Employee attrition is a pervasive issue in many industries, leading to increased recruitment costs, loss of institutional knowledge, and potential disruptions in team dynamics. By identifying and understanding the factors that contribute to attrition, we can proactively implement measures to retain valuable employees, enhance job satisfaction, and ultimately improve the overall organizational climate.

3. c. What is your idea to solve the problem?

Our approach involves leveraging data science methodologies to analyse and predict employee attrition. By examining various features such as job satisfaction, work-life balance, and career development opportunities, we aim to identify patterns and trends that correlate with attrition. This will enable us to develop targeted interventions and retention strategies to address specific challenges faced by employees.

3. d. What differences you could make with your data science approach?

Traditional methods of addressing attrition may lack specificity and effectiveness. Data science allows us to delve deep into the underlying causes of attrition by analysing a myriad of factors simultaneously. The insights gained from this approach enable us to tailor retention strategies to individual or departmental needs, leading to more precise and impactful interventions.

3. e. Why do you believe the idea deserves the financial resources of your company?

Investing in data science for attrition prediction and prevention is a strategic decision that can yield substantial returns on investment. By retaining skilled and experienced employees, the company can save on recruitment and training costs. Moreover, maintaining a stable workforce positively influences the company's reputation, customer satisfaction, and overall competitiveness in the market.

3. f. Business Proposition:

Our business proposition centres around the proactive management of employee attrition to enhance workforce stability. By leveraging data science methodologies, we aim to identify key factors influencing attrition within our organization. The proposition extends to the development of predictive models that provide actionable insights for strategic decision-making. Our goal is to empower the company with the ability to implement targeted interventions, ultimately fostering a stable and satisfied workforce.

3. g. Importance & Motivations:

This topic holds paramount importance to us as it directly impacts the overall success and resilience of our organization. A stable and satisfied workforce contributes to increased productivity, improved organizational culture, and cost savings associated with recruitment and training. The motivations behind this analysis stem from a genuine interest in understanding the intricate dynamics of employee attrition, debunking assumptions, and providing tangible solutions for effective workforce management.

4. Data Gathering:

4. a. Dataset Source and Relevance:

The initial phase of our study involves sourcing a comprehensive dataset directly from the company's HR records. This proprietary dataset serves as a rich repository of employee-related information, ensuring that our analyses are contextually relevant to the organization. By utilizing internal records, we gain access to specific nuances and intricacies that may not be captured in publicly available datasets. This approach enhances the authenticity and applicability of our findings.

4. b. Features Exploration:

Upon acquiring the dataset, an in-depth exploration of its features was conducted. Key variables such as 'BusinessTravel,' 'MaritalStatus,' 'JobRole,' 'Department,' 'EducationField,' and 'Attrition' were identified. Understanding the nature of these features is paramount for formulating hypotheses and conducting analyses that align with the intricacies of the company's workforce.

4. c. Data Pre-processing for Analysis:

To facilitate effective analysis, certain preprocessing steps were undertaken. Categorical variables were transformed into numerical representations. For instance, 'BusinessTravel' categories like 'Travel_Rarely' or 'Travel_Frequently' were mapped to numerical values. Additionally, columns deemed irrelevant for our analysis, such as 'Over18,' 'EmployeeCount,' and 'StandardHours,' were removed to streamline the dataset.

4.d. Balancing the Dataset:

A subsample (Attrition 350 (0): 237 (1)) of the original dataset was taken to make it more balanced. The data selected for Attrition (0) were randomly selected.

5. Conjectures

5.a Conjecture 1

“EnvironmentSatisfaction”, “RelationshipSatisfaction”, “JobSatisfaction” and “WorkLifeBalance” scores are inversely proportional to attrition rates.

5.b Conjecture 2

Employees from specific education fields will have a higher attrition rate due to the nature of work.

5.c Conjecture 3

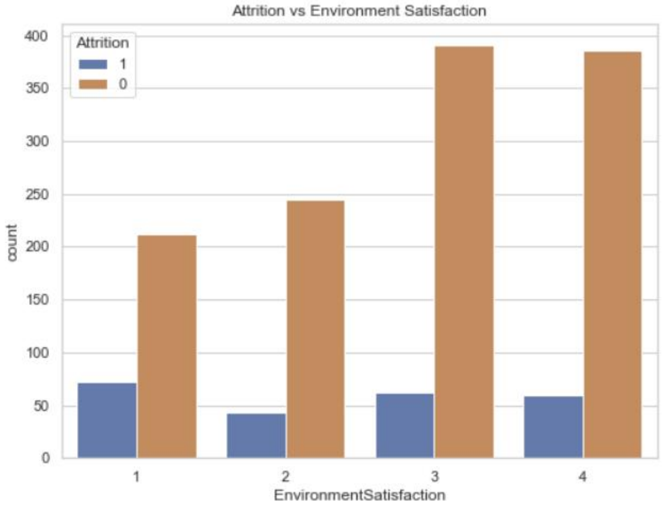
More qualified (highest educational level) employees are more likely to leave.

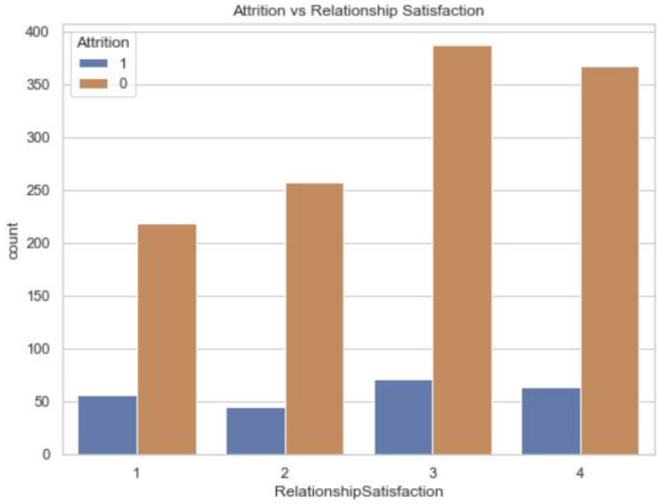
5.d Conjecture 4

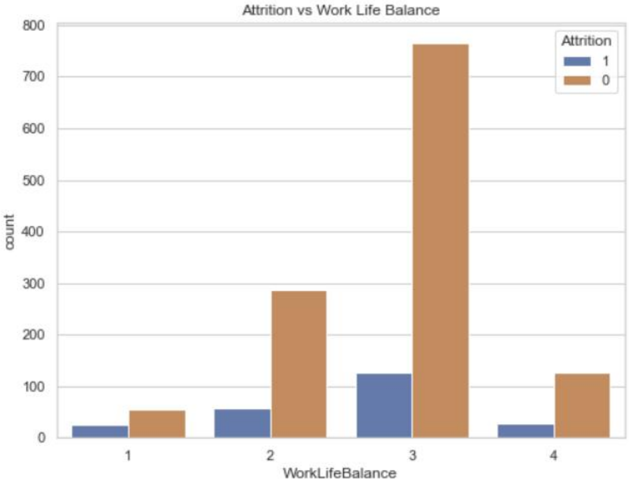
More experienced (years worked in total and in current role) employees are more likely to leave.

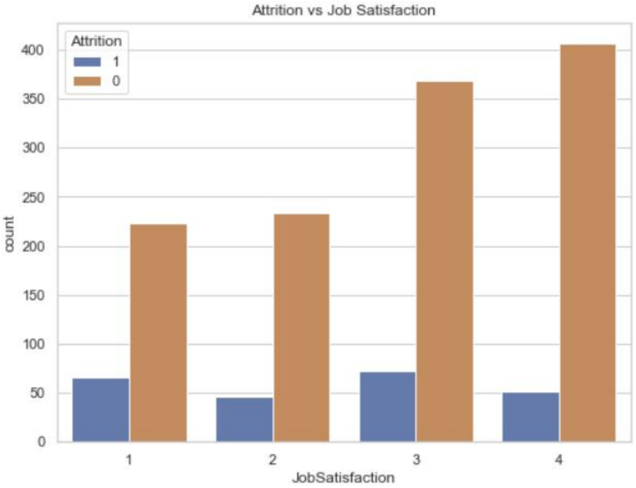
6. Exploratory Data Analysis (EDA)

6.a Conjecture Evaluation


Attrition vs Environment Satisfaction
Attrition
1
0
count
400
350
300
250
200
150
100
50
0
1
2
3
4
EnvironmentSatisfaction

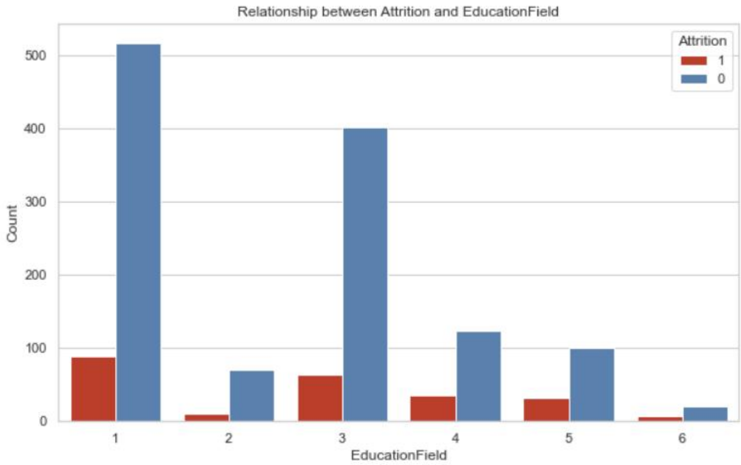

Attrition vs Relationship Satisfaction
Attrition
1
0
count
400
350
300
250
200
150
100
50
0
1
2
3
4
RelationshipSatisfaction

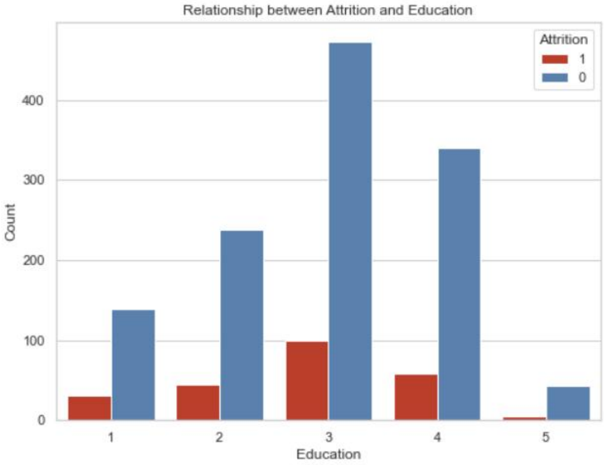


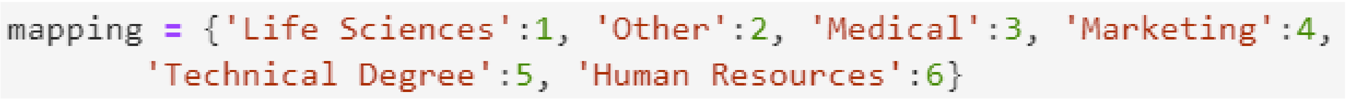


Conjecture 1 Update

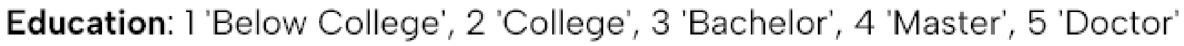
“**EnvironmentSatisfaction**”, “**RelationshipSatisfaction**”, “**JobSatisfaction**” “**WorkLifeBalance**” scores are in fact not of particular concern. The tree based models will give a better idea about the specific combination of these metrics that lead to attrition.







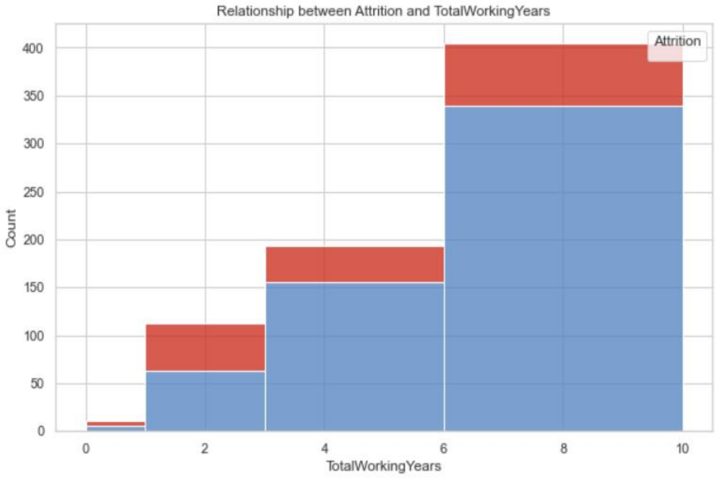
```
mapping = {'Life Sciences':1, 'Other':2, 'Medical':3, 'Marketing':4,
       'Technical Degree':5, 'Human Resources':6}
```

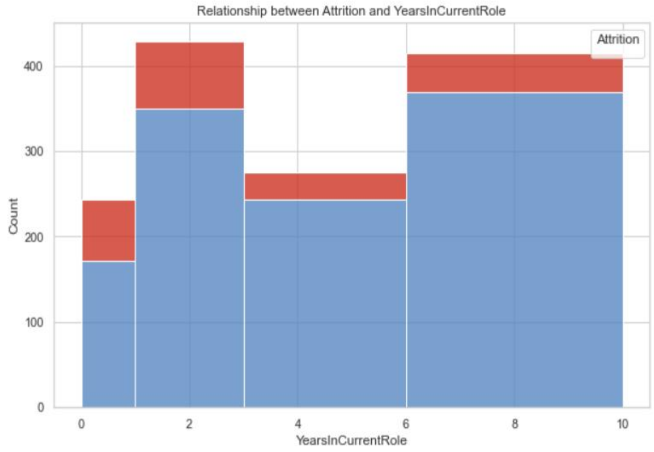


Education: 1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor'

Conjecture 2 and 3 Update

Professionals in the life science field and employees whose highest education level is a bachelor's degree are more likely to leave.





Conjecture 4 Update

The total work experience of an employee is directly proportional to attrition likelihood, and employees who have worked in their current role for 2-3 years are particularly likely to jump companies frequently.

6.b New Findings

Overtime worked (boolean), **business travel** (frequency) and **distance from home** (miles) were particularly strong factors that increase the likelihood of an employee leaving the company.

7. Machine Learning Models:

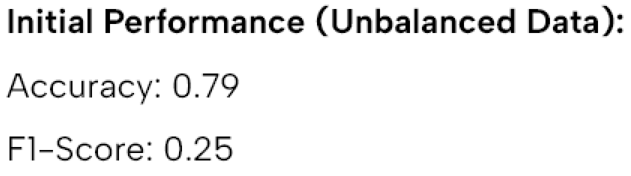
7. a. Decision Tree:

Model Overview:

Tree based methods were our first choice, given the suitability for a classification task such as this but also because the final tree/model will be **human-readable**, so that a HR professional can read and adapt the structure if required on a case-by-case basis.

Performance:

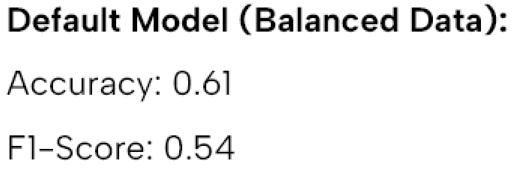
Accuracy and F1-Score were taken, due to the imbalanced of target variable data in the dataset.



Initial Performance (Unbalanced Data):

Accuracy: 0.79

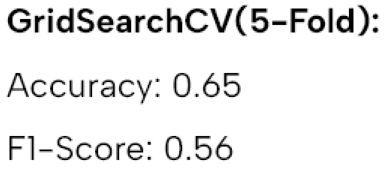
F1-Score: 0.25



Default Model (Balanced Data):

Accuracy: 0.61

F1-Score: 0.54



GridSearchCV(5-Fold):

Accuracy: 0.65

F1-Score: 0.56

Hyperparameter tuned Decision Tree exhibits improved classification compared to the initial model, but still overfits the data and gives a very low F1-Score, yielding unsatisfactory performance.

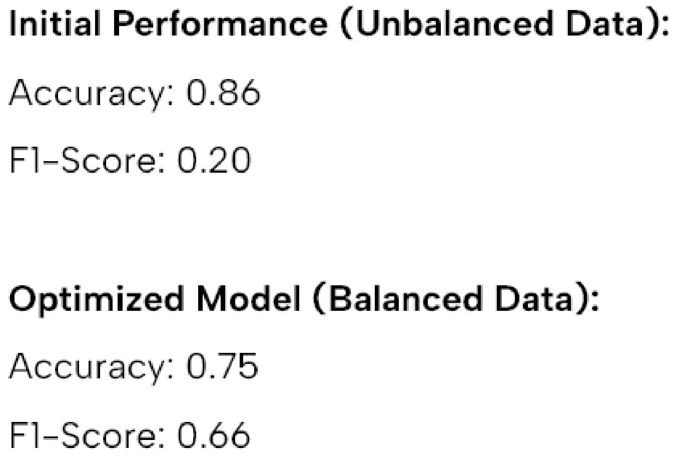
7. b. RandomForest:

Model Overview:

Having seen that even the optimized Decision Tree model overfits the data, the Random Forest model is the next choice as **bagging** the predictions will help mitigate the overfitting of using one tree alone.

Performance:

Accuracy and F1-Score were taken, due to the imbalanced of target variable data in the dataset.



Initial Performance (Unbalanced Data):

Accuracy: 0.86

F1-Score: 0.20

Optimized Model (Balanced Data):

Accuracy: 0.75

F1-Score: 0.66

Even with optimum hyperparameters and balanced data, Random Forest performs better than Decision Tree but some overfitting still persists. Better accuracy and precision/recall required.

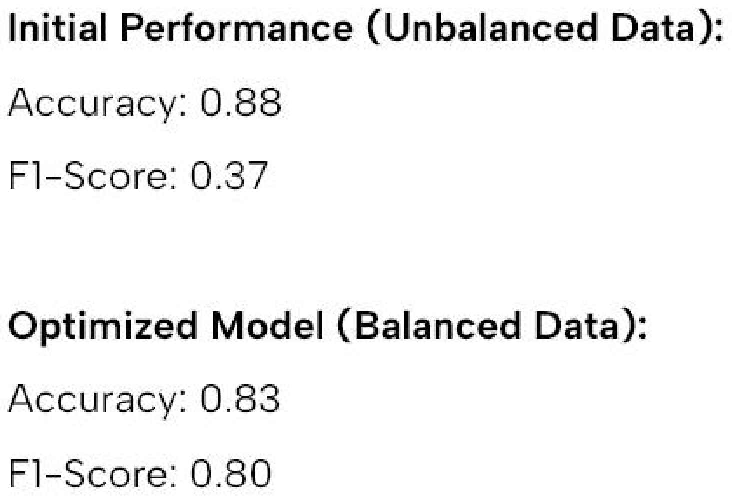
7. c XGBoost

Model Overview:

Gradient boosting trees (XGBoost) were chosen to maintain interpretability and also give **further room for optimization** through more hyperparameters.

Performance:

Accuracy and F1-Score were taken, due to the imbalanced of target variable data in the dataset.



Initial Performance (Unbalanced Data):

Accuracy: 0.88

F1-Score: 0.37

Optimized Model (Balanced Data):

Accuracy: 0.83

F1-Score: 0.80

The optimized XGBoost model yields the highest test accuracy and F1-Score, and shows the biggest improvement in F1-Score when using balanced data

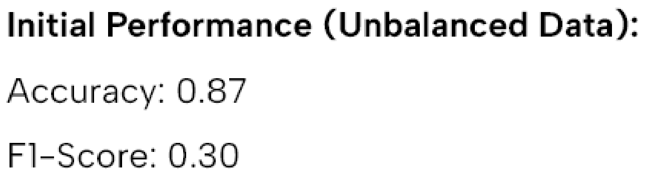
7. d Logistic Regression

Model Overview:

The immediate choice for a classification problem, especially for one that has linear relationships, but being a non-tree based model makes it less interpretable.

Performance:

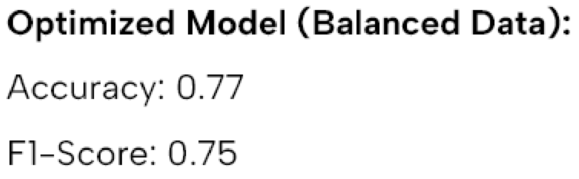
Accuracy and F1-Score were taken, due to the imbalanced of target variable data in the dataset.



Initial Performance (Unbalanced Data):

Accuracy: 0.87

F1-Score: 0.30



Optimized Model (Balanced Data):

Accuracy: 0.77

F1-Score: 0.75

The optimized Logistic Regression model gives a good test accuracy and F1-Score, but is lower than XGBoost.

8. Model Integration with Business Proposition:

The tree based models align with our business proposition by serving as tools for proactive attrition management as well as being easy to interpret. These models, trained on the insights gained from data analysis, offer the ability to predict attrition risks, allowing the company to implement targeted strategies for employee retention. The models serve as a tangible manifestation of our business proposition, translating data-driven insights into practical solutions.

9. The Story of the Group

Our team initiated this data exploration endeavor with a shared curiosity, seeking to unravel latent patterns within our company's HR dataset. The primary motivation was to gain insights into the intricate dynamics surrounding employee attrition and to validate our preconceptions about the factors influencing workforce stability. The allure of this analysis lay in the prospect of devising strategies that could potentially maximize employee retention and, consequently, contribute to the overall success of the organization.

Navigating through the extensive dataset required each team member to assume a pivotal role in the data collection, cleaning, and analysis phases.

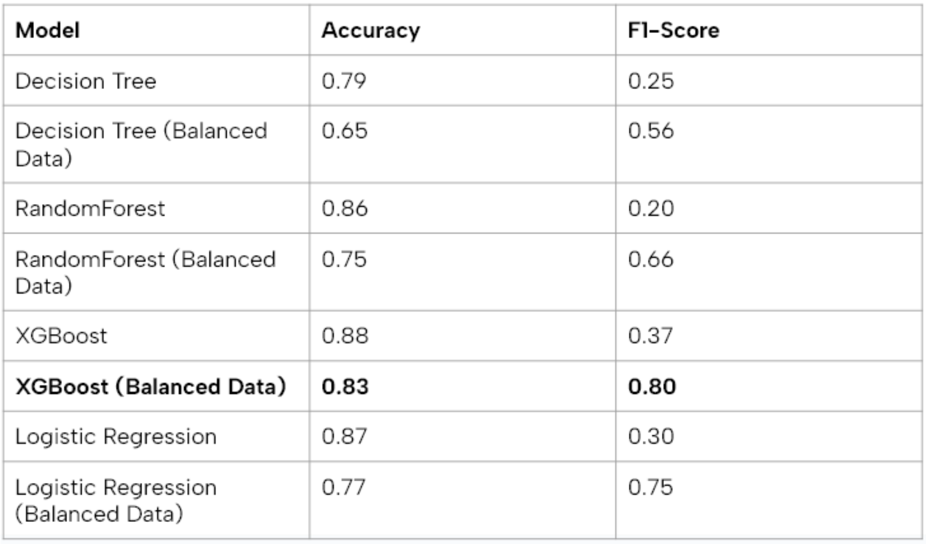
Throughout the journey, we confronted challenges inherent in working with real-world data, including imbalances and nuances specific to our models. It was the collective dedication, diverse skills, and collaborative spirit of our team that ensured the project's success. The analysis not only corrected certain assumptions about employee attrition but also generated new actionable insights derived from a data-driven perspective.

10. Acknowledgments

We express our gratitude to the contributors of the HR dataset, the data science prrofessionals at IBM, without whom this analysis would not have been possible. Their commitment to providing a comprehensive and insightful dataset has enabled our team to delve into the complexities of employee attrition within our organization. We also extend our appreciation to our team members whose collaborative efforts in data collection, cleaning, and analysis have been instrumental in the success of this project. Their dedication and expertise have contributed to the depth and accuracy of our findings, enhancing the overall quality of our data-driven insights.

11. Summary and Conclusion:

A machine learning model that can automatically predict employees that are likely to leave the company, and suggest the specific areas to focus HR involvement in to increase the chances of retaining them was successfully built and optimized. XGBoost performed the best on the balanced data that we subsampled and cleaned.



Model	Accuracy	F1-Score
Decision Tree	0.79	0.25
Decision Tree (Balanced Data)	0.65	0.56
RandomForest	0.86	0.20
RandomForest (Balanced Data)	0.75	0.66
XGBoost	0.88	0.37
XGBoost (Balanced Data)	**0.83**	**0.80**
Logistic Regression	0.87	0.30
Logistic Regression (Balanced Data)	0.77	0.75

The key factors that lead to employees leaving a company and also those that improve the likelihood of retaining them were identified. It was found that:

Employees being satisfied with their company’s work environment, relationship with their manager or their work-life balance are not actually determining factors in their leaving of the company.

1. Professionals in the life science field and employees whose highest education level is a bachelor’s degree are more likely to leave.
2. The total work experience of an employee is directly proportional to attrition likelihood, and employees who have worked in their current role for 2-3 years are particularly likely to jump companies frequently.
3. **Overtime** worked (boolean), **business travel** (frequency) and **distance from home** (miles) were particularly strong factors that increase the likelihood of an employee leaving the company.