



HR ATTRITION

Abstract

HR attrition significantly impacts the organisation's performance, cost, and team morale. [1]
This study uses Power BI to analyze an HR dataset by using the Binary Logistic model to identify which factors have more influence on attrition rates. The findings reveal that Overtime, YearSinceLastPromotion, DistanceFromHome are some features that influence the turnover. By implementing strategies to support employees can reduce the attrition and increase stability.

Irin Mary Thomas
irinthomas0@gmail.com

Keywords- Employee turnover, Retention strategies, Binary Logistic Regression

I. Introduction: -

Employee turnover is a major issue for many companies tend to lose their efficiency, performance, and overall workspace stability. [2]Understanding the factors influencing employee attrition is essential to derive strategic retention plans. This study focuses on identifying the key elements responsible for employee turnover like Overtime, DistanceFromHome, YearSinceLastPromtion

Higher turnover results in higher recruitment and training costs and hinders operational efficiency. By applying the Binary Logistic regression model this project aims to find the relationship among the features affecting the attrition and gives insights to the HR department to foster a more reliable and supportive work environment.

Ultimately, this research highlights the areas on which HR should focus to understand the main source of employee turnover and implement retention strategies.

II. Literature Review: -

1. Title: *HR Analytics: Employee Attrition Analysis using Random Forest* (Shobhanam Krishna and Sumati Sidharth [2])

Contextual summary: The author has presented an approach using the Random Forest algorithm on an IBM dataset which contains 35 features and around 1470 samples. The study focuses on the importance of understanding the factors that influence attrition such as job satisfaction, pay, and work environment. Attending to such issues is very crucial it reduces the cost associated with employee turnover and improves the workforce efficiency.

Existing Literature and Identified Gaps: The authors use methods like Decision Trees and Logistic regression to predict attrition in their paper. However certain gaps have been noticed while addressing imbalanced data and exploring advanced ensemble models. Although using SMOTE improved the sensitivity of the model the final validation results only showed small improvements.

Key Contributions and Intellectual Context: This research focuses on the strong performance of predicting attrition through the Random Forest method

and proves how to boost accuracy by feature selection and data balancing

Relevance: This report was chosen for its practicality in HR analytics as this emphasizes machine learning's role in fixing attrition problems.

2. Title: *Machine Learning in HR Analytics: A Comparative Study on the Predictive Accuracy of Attrition Models* (Md Shaik Amzad Basha et al.) [4]

Contextual Summary: The study mainly focuses on comparing various machine learning models like the Logistic Regression, SVM, Random Forest, and GBM, to foresee attrition. It highlights the significance of detecting the risk of employees leaving and ensuring that the model works well for this minority group.

Existing Literature and Identified Gaps: The article uses models like Naive Bayes and Decision Trees. Instead of ensemble models like GBM and XGBoost which are known for handling complex relationships, A key gap is the lack of focus on recall, which measures accuracy for predicting employees likely to leave.

Key Contributions and Intellectual Context: The research shows that advanced models, like SVM and GBM, outperform traditional ones in balancing accuracy and recall. The authors also stress the need to choose models based on organizational needs, as no single model fits all situations.

Relevance: This article stands out for its broad comparison of models and practical insights for organizations

III. Methodology-

Data Analytical Life Cycle Stages-

1. Discover: -

The data cover data of almost a decade till 2017. It is observed that the firm is divided into 6 different departments and 9 different job roles. There are almost 1470 number of employees, and the attrition rate is 16.1%. [5] Up until the year 2017, 237 employees have left the company.

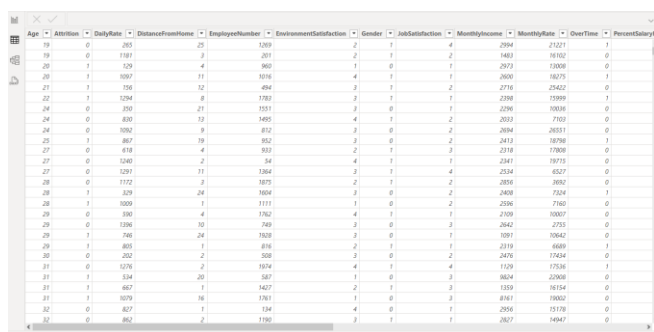
[illegible]

Figure 1-Initial dataset

2. Data Preparation

The dataset was collected from website data. World. Various data cleaning processes were involved, and duplicate tables were created to better analyze the prediction. The 'Initial dataset' was duplicated with the 'Final dataset' and this was used for the prediction of the new employee turnover.

- **Removing Values:** Certain columns were removed from the Final dataset to predict the attrition which would not affect the prediction, such as Business Travel, Education, Marital Status, Num Companies, Worked, Over18, Years In Current Role, Training Times Last Year, Stock Option Level, Standard Hours, Years With Current Manager, Total Working Years, Job Level, Job Involvement, Employee Count, Employee Number, Hourly Rate, Education Field, Job Role.
- **Replacing Values:** Certain values of the columns like Attrition, Overtime, and Gender initially had Yes and No inputs which were replaced by 0 and 1. Certain pivoted tables had null values which were later replaced with 0 for the accuracy in attrition prediction.
- **Splitting Columns:** Certain columns had categorical data which were converted to numerical (0/1) and were distributed into new columns. The occurrence of each value in the respective columns was represented with 1 and the remaining were initially null which were later replaced with 0 for better analysis. [6] Custom columns with only value 1 were added so that the values could be easily distributed into their respective columns while pivoting.



Age	Attrition	DailyRate	DistanceFromHome	EmployeeNumber	EnvironmentSatisfaction	Gender	JobSatisfaction	MonthlyIncome	MonthlyRate	OverTime	PercentSalaryHike
19	0	201	20	201	2	1	4	2094	21212	1	0
19	0	1181	3	201	2	1	4	1483	16162	0	0
20	1	129	4	960	1	0	1	2973	13008	0	0
20	1	1087	11	9196	4	1	1	3006	16175	1	0
21	1	156	12	494	3	1	2	2716	25422	0	0
22	1	1294	8	1783	3	1	1	2398	13999	1	0
24	0	390	27	1551	3	0	1	2396	16906	0	0
24	0	830	13	3495	4	1	2	3033	7103	0	0
24	0	1092	9	812	3	0	2	2694	26311	0	0
25	1	867	19	862	3	0	2	2413	16786	1	0
27	0	618	4	933	2	1	3	2318	17068	0	0
27	0	1240	2	54	4	1	1	2341	19711	0	0
27	0	1291	11	1384	3	1	4	2334	6227	0	0
28	0	1122	3	1875	2	1	2	2856	3862	0	0
28	1	329	24	3604	3	0	2	2409	7534	1	0
28	1	1000	1	1111	1	0	2	2596	7160	0	0
29	0	390	4	1762	4	1	1	2109	10007	0	0
29	0	1396	10	749	3	0	3	2642	2751	0	0
29	1	746	24	1030	3	0	1	1091	10642	0	0
29	1	803	1	816	2	1	1	2319	4689	1	0
29	0	202	2	530	3	0	2	2476	17424	0	0
31	0	1276	2	1074	4	1	4	1126	17536	1	0
31	1	534	20	587	1	0	3	8624	22008	0	0
31	1	467	1	3427	2	1	3	1339	16194	0	0
31	1	1039	16	1761	1	0	3	8161	16662	0	0
32	0	827	1	134	4	0	1	2556	15178	0	0
32	0	862	2	1789	3	1	1	2627	14447	0	0

Figure 2. Final Dataset

- **Train-Test Split:** - The data was split into two parts that are Train and Test where 20% of the data was set aside for testing and 80% was used for testing, this helps in evaluating the model's performance after the training. [7]

3. Model Planning: -

Choose Binary Logistic Regression as the dataset is about employee turnover and this model focuses on giving a binary output. It makes it easy to find the impacts that the factors have on the likelihood of an employee living or staying in the company. [8]

KNN model: KNN model is found to be effective in non-linear cases and involved high complexity but was less interpretable due to its "black box" nature. It has also been proven that it can narrow down its utility for making decisions based on the contribution made by the factors. [9]

Regularization Logistic Regression: It not only alters the coefficients of the regression model but also penalizes large values. It is not suited for a small dataset like the one used in this report. The regularization model also focuses on the predictive performance rather than the factors or the insights solely affecting the prediction. [10] Thus, the binary logistic regression model is used as the main aim of the analysis is to uncover actionable insights for the HR team.

4. Model Building

Primary Technique: The chosen technique is the binary logistic regression model for its interpretability and the ability to produce binary outcomes more efficiently. It helps in understanding the probability of the events being a success or failure (0/1). It also helps study the relationship between the given data set and the predicted outcome. [8]

Alternative techniques researched were KNN which, unlike logistic regression, did not provide coefficients, [9] and the second technique regularization logistic regression was used for high-dimensional datasets which made it less ideal for the HR Attrition dataset. [10]

5. Communication: -

The visuals that were created were

1. A line chart was created to know the attrition by the number of years an employee has been in the company
2. A Scatter chart to find the relation that the attrition has with the distance employees travel from home to the company.
3. A donut chart giving insights about the relationship between turnover and the employees who work overtime.
4. A Stacked column chart grouping the age which has high attrition chances.
5. A matrix to show the relationship between job satisfaction among the different departments.
6. A stacked bar chart to show which job role has the most attrition rates.

6. Measures Effectiveness/Apply Live: - This includes focusing on the privacy, consent, and handling of the employee's data responsibly. Hr analytics needs a

structural and ethical framework to promote data privacy, transparency, and trust among the employees as a key principle. Especially when managing sensitive employee data limited access to essential personnel should be implemented to identify and resolve ethical issues proactively. Frequent privacy assessments should be done to create transparent communication to keep updated information. [11]Moreover, the maintenance of privacy, consent, and data security throughout the process ensures that the data is used responsibly and there is no compromise on the trust and the privacy of the employee's information.

IV. RESULTS AND DISCUSSIONS

- **Model Overview:** This report helps to use the Binary Logistic Regression model as a primary supervised model to analyze HR Attrition. This model is popularly used for effective binary classification tasks which makes it easier to find the probability of the employee leaving the country based on the independent variables in the data set. It includes a series of linear input features which are transformed through the logistic functions which gave outputs in 0 and 1 representing the probability of attrition.
- **Key Variables Influencing the Predictions:** The model has certain key variables that significantly affect the attrition and those are: -
 - Overtime
 - YearSinceLastPromotion
 - DistanceFromHome
- **Data Transformation and Model Fitting:** Data transformation involves several steps that include pivoting tables that were categorical and were converted into numerical. Deleting and adding certain features to predict the data effectively. The dataset was also split into training (80%) and testing (20%) subsets to validate the performance of the model.
- **Evaluation Metrics:** The metrics and techniques that were used to find how well the model works are:
 - **Confusion Matrix:** In the confusion matrix that was calculated [12] it showed that there where the numbers were:

Metric	Sum of Value
False Negative	29
False Positive	10
True Negative	245
True Positive	10
Total	294

Figure 3. Confusion Matrix

- **Accuracy:** The accuracy rate of the binary logistic regression model implemented on the dataset [13] was 85% which shows that we were successful in predicting 85% correct predictions out of the total predictions. **Figure 4.** Accuracy Score(Appendix)
- **F1 Score:** The F1 score of 0.34 indicates room for improvement in identifying the positive cases of attrition. It also shows a moderate balance between precision and recall. [14]. **Figure 5.** F1Score(Appendix)
- **ODD Ratio:** The odd ratio derived from the logistic regression coefficient provides insights into the likelihood of attrition based on predictor values. [15]Variables that increase the odds of attrition are DistanceFromHome (1.23), Overtime (2.28), PerformanceRating(1.02), YearssinceLastPromotion (1.73), Sales(1.17), HumanResources (1.03). These are the variables that the HR should mainly focus on and implement retention measures to reduce attrition. **Figure 6.** Odd Ratio (Appendix)
- **Pseudo R-squared:** - The Pseudo R-squared value indicates the model fitness which showcases the variance of the model. The HR dataset with 0.23 pseudo-R-squared is found reasonable as the dataset has a lot of unpredictable and subjective factors which makes it challenging for any model to explain a large proportion of variance McFadden's pseudo-R-squared values between 0.2 and 0.4 are commonly viewed as indicating a good model fit in contexts like HR analytics, where perfect prediction is not expected. **Figure 7.** Pseudo-R-squared(Appendix)

- **Co-efficient:** -The co-efficient is calculated to know the effect that particular feature had on the target variable (Predicted Attrition) OverTime(0.85), YearSinceLastPromotion (0.56), and DistanceFromHome are the top three significant factors influencing the model and the target. Whereas JobSatisfaction, YearsAtCompany, and MonthlyIncome features have relatively negative coefficients, suggesting they negatively affect the target outcome. **Figure 8.** Co-efficient bar chart(Appendix)

The graphs and visualizations are prepared to show the metrics and trends related to the HR Attrition dataset that we have, like:

1. **Job Satisfaction Table by Department:**

A table has been created to represent job satisfaction scores (1-4) across different departments. This table helps HR understand how the job satisfaction impacts the employee turnover.

Job Satisfaction among the Departments				
Department	1	2	3	4
Human Resources	3	5	5	1
Research & Development	25	22	47	43
Sales	6	9	12	10

Figure 9. Job Satisfaction among the departments table

2. **Age Bar Chart:** This bar chart categorizes different ages into groups to identify which group experiences the highest turnover which allows the HR to implement tailored retention techniques based on the age demographics.

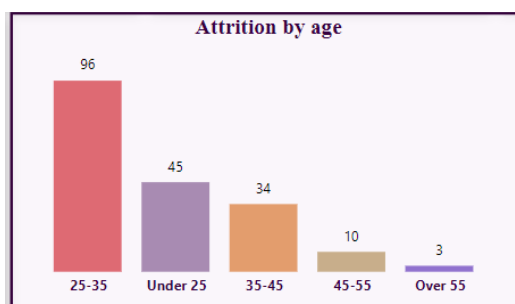


Figure 10. Attrition by age

3. **Key Matrix Cards:** 3 matrix cards show the total attrition count (188), overall model accuracy(84.83%) and total employee count(1470). This gives a brief view of the dataset by the number of employees and also shows the performance of the model by attrition count and the accuracy of the model.

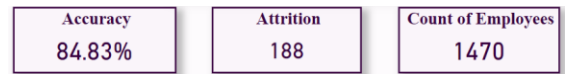


Figure 11. Matrix Card for Accuracy, Attrition and Total Employees

4. **Line Chart for Attrition by Years at Company:** This line chart illustrates the attrition with the number of years the employee has worked for the employer. The spike in the chart shows the year in which employees leave the company most often. This helps the HR team to make improvements in the initiatives particularly in those years to help retain the employees and reduce the attrition rate

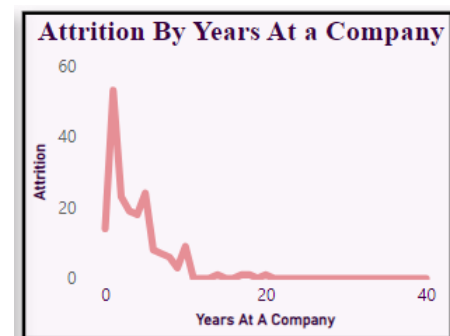


Figure 12. Line Chart for attrition by years at company

5. **Attrition By Job Role Bar Chart:** The horizontal bar charts show how the job role influences attrition. The HR can observe through this chart that the “Research Scientists and the “Laboratory Technicians” tend to leave the organization more often than others and can focus on implementing role-specific retention methods.

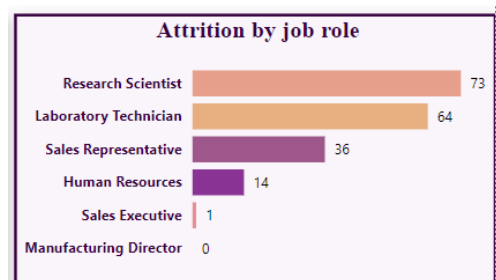


Figure 13.Bar Chart for Attrition by Job Role

6. **Overtime Donut Chart:** The donut chart fragments show the overtime status(1/0) which

helps to reveal how the employees working overtime affect the attrition. It helps HR to assess overtime is an important factor in increasing the HR rate.

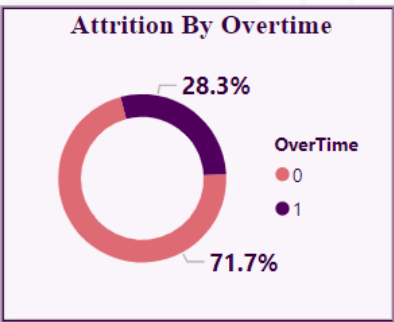


Figure 14.Donut Chart of Overtime

7. Scattered Chart:The chart shows relationship between the age and the distance affecting the attrition. The younger employees who tend live farther have the most chances to have attrition(1)

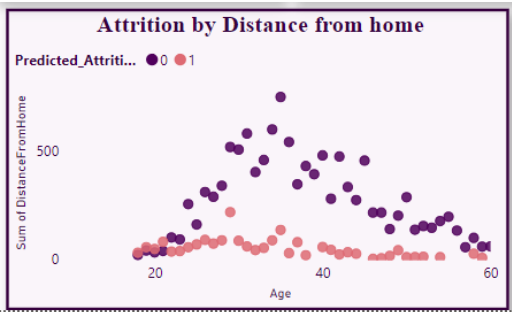


Figure 15.Scattered plot attrition according to the age and distance from home

V. PEER REVIEW

1:- The peer review helped me to know how other models work and how they can be implemented in real-world data. **Figure 15.** Peer Review Given 1,**Figure 16:** -Peer Review Given 2(Appendix)

2:- As stated, I worked on the Literature review and provided the sources for the research. Additionally, I included the actionable solutions in the conclusion section which enhanced my report quality. The need for more information about Binary logistic regression was also fulfilled in the Model Building section. The Power BI Visuals were also explained in detail in the Results and the discussion section These helped me improve my report quality and made it easy and proper for anyone to understand.**Figure 17:** - Peer Review Given 1,

Figure 18: - Peer Review Received 1(Appendix)

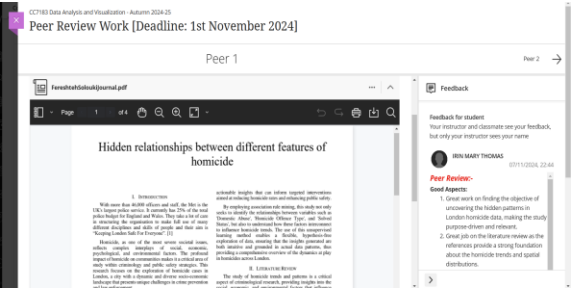


Figure 16. Peer Review Given 2

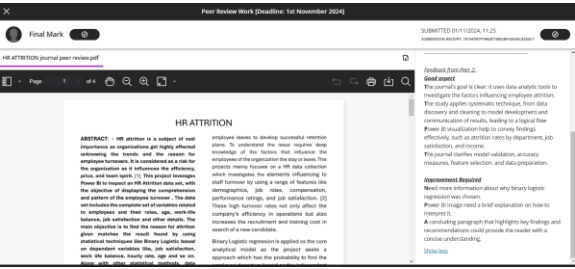


Figure 19: - Peer Review Received 2

VI. CONCLUSIONS AND RECOMMENDATIONS

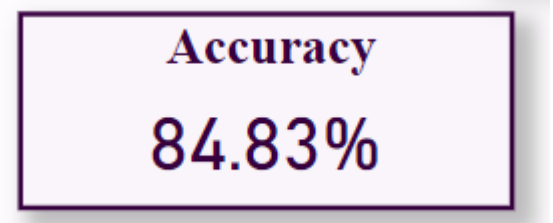
The insights derived from the binary logistic regression model were significant predictors like Overtime, Year Since Last Promotion and Distance from home helped in achieving the initial objective of finding key features impacting attrition The unsupervised model provided a broader perspective through Power BI visuals like the relationship between the job satisfaction among the departments which helps uncover hidden relationships which will help in taking retention measures. [16] For future work advanced machine learning techniques could be used to provide ongoing insights into employee satisfaction and early warning signs of potential turnover. Additionally, implementing cross-validation techniques could increase the model's reliability, ensuring that the insights derived are both actionable and robust [17] Finally, this project helps to address the immediate HR concerns and lay a groundwork for future research and model enhancement in employee retention strategies.

References

- [1 P. Quinn, "Understanding Employee Attrition: Causes, Impacts, and Solutions," Training Outlook, 2024. [Online]. Available: <https://trainingoutlook.com/employee-attrition/>. [Accessed 19 October 2024].
- [2 S. a. S. S. Krishna, "HR Analytics: Employee Attrition Analysis using Random Forest," *HR Analytics: Employee Attrition Analysis using Random Forest*, vol. 18, no. 10.23940/ijpe.22.04.p5.275281, p. 7, 2022.
- [3 F. Fallucchi, M. Coladangelo, R. Guiliano and E. W. D. Loca, "settingsOrder Article Reprints," *MDPI*, 2020.
- [4 M. S. A. Basha, O. Varikunta, A. U. Devi and S. Raja, "Machine Learning in HR Analytics: A Comparative Study on the Predictive Accuracy of Attrition Models," *A Comparative Study on the Predictive Accuracy of Attrition Models*, no. 10.1109/DICCT61038.2024.10533064, p. 6, 2024.
- [5 J. Pathak, "HR Attribution Data," 2018. [Online]. Available: <https://data.world/juhipathak7/hr-attrition-data>. [Accessed 20 11 2024].
- [6 J. C. Wu, "RPods by RStudio," *HarvardX Data Science Capstone 2 – Predicting Employee Attrition*, 2022.
- [7 A. Preeth, "Employee Attrition Prediction – A Comprehensive Guide," *Analytical Vidya*, 2024.
- [8 G. f. greeks, "Advantages and Disadvantages of Logistic Regressio," *GeeksForGeeks*, 8 2024. [Online]. Available: <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/>. [Accessed 1 11 2024].
- [9 A. Kumar, "KNN vs Logistic Regression: Differences, Examples," *Analytics Yogi*, 02 12 2023. [Online]. Available: <https://vitalflux.com/knn-vs-logistic-regression-differences-examples/#:~:text=Non%2DLinear%20Relationships%3A%20KNN%20can,size%20of%20the%20data%20grows..> [Accessed 18 10 2024].
- [1 A. Akalin, "Logistic regression and regularization," in *Computational Genomics with R*, Berlin, Germany, Chapman & Hall/CRC, 2020.
- [1 I. o. Data, "Exploring Data Privacy and Ethical Considerations in HR Analytics and Training," Institute of Data, 04 12 2023. [Online]. Available: <https://www.institutedata.com/us/blog/exploring-data-privacy-and-ethical-considerations-in-hr-analytics-and-training/>. [Accessed 5 11 2024].
- [1 A. Navlani, "Understanding Logistic Regression in Python," datacamp, 11 8 2024. [Online]. Available: <https://www.datacamp.com/tutorial/understanding-logistic-regression-python>. [Accessed 1 11 2024].
- [1 L. Su, "Logistic Regression, Accuracy, and Cross-3] Validation," Medium, 14 5 2019. [Online]. Available: https://medium.com/@lily_su/logistic-regression-accuracy-cross-validation-58d9eb58d6e6. [Accessed 02 11 2024].
- [1 R. Kundu, "F1 Score in Machine Learning: Intro & 4] Calculation," V7, 16 12 2022. [Online]. Available: <https://www.v7labs.com/blog/f1-score-guide>. [Accessed 7 11 2024].
- [1 J. Frost, "Odds Ratio: Formula, Calculating & Interpreting," 5] Statistics By Jim, 2024. [Online]. Available: <https://statisticsbyjim.com/probability/odds-ratio/>. [Accessed 11 11 2024].
- [1 A. Preet, "Employee Attrition Prediction – A Comprehensive 6] Guide," *Analytical Vidya*, 11 11 2024. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/>. [Accessed 22 11 2024].
- [1 scikit-learn developers, "Cross-validation: evaluating 7] estimator performance," scikit learn, 2007-2024. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html. [Accessed 20 11 2024].

VII. APPENDIX

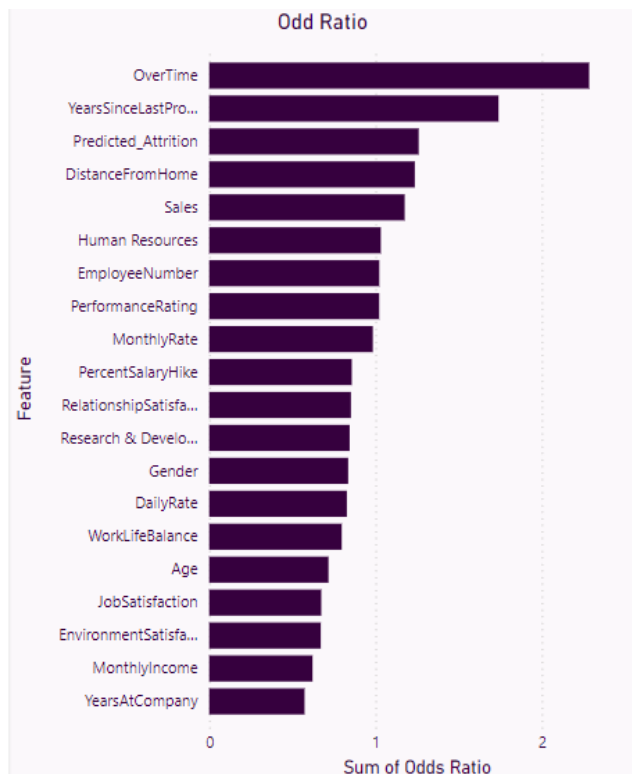
1. Figure 4. Accuracy Score



2. Figure 5. F1Score

Metric	Sum of Value
F1 Score	0.34
Total	0.34

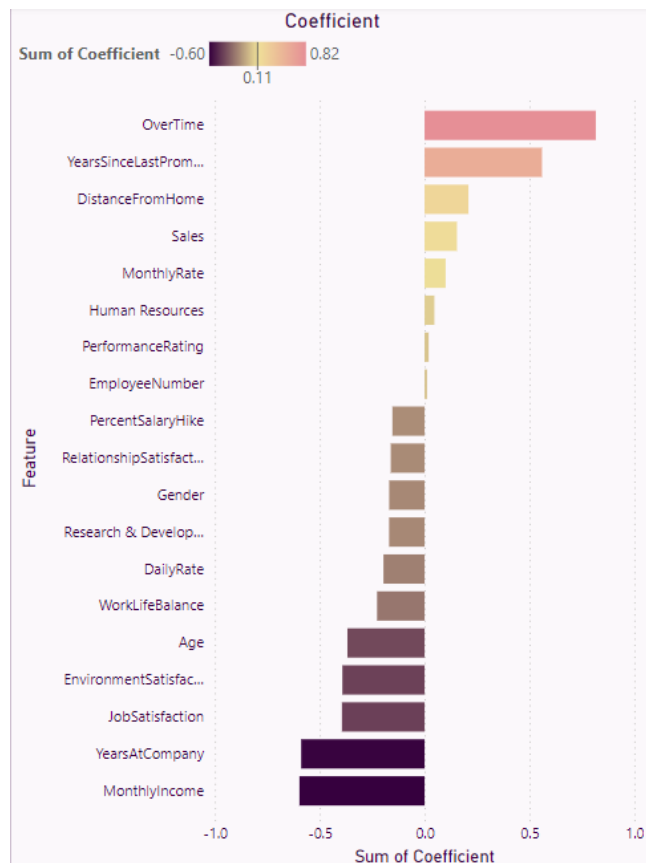
3. Figure 6. Odd Ratio



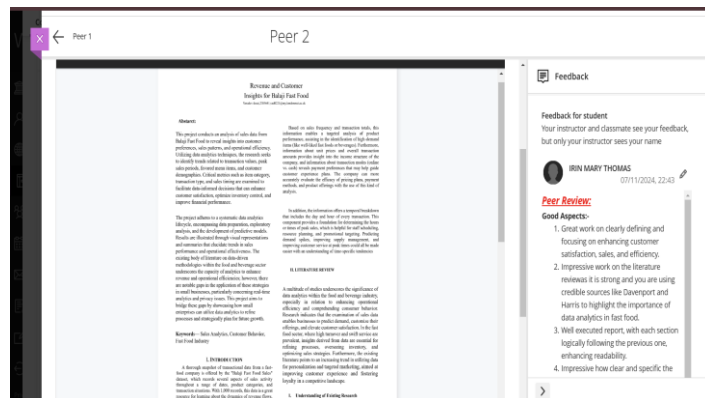
4. Figure 7. Pseudo-R-squared

Metric	Sum of Value
Pseudo R-squared	0.23
Total	0.23

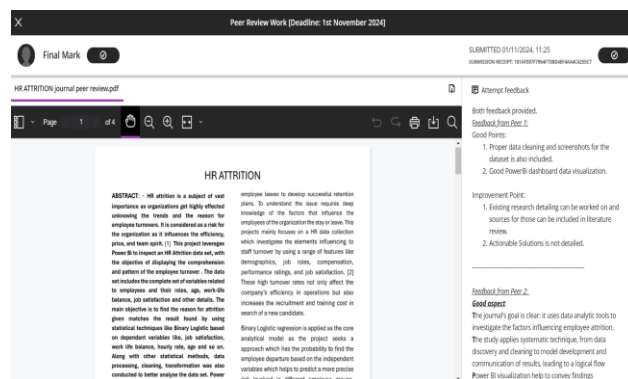
5. Figure 8. Co-efficient bar chart



6. Figure 17. Peer Review Given 1



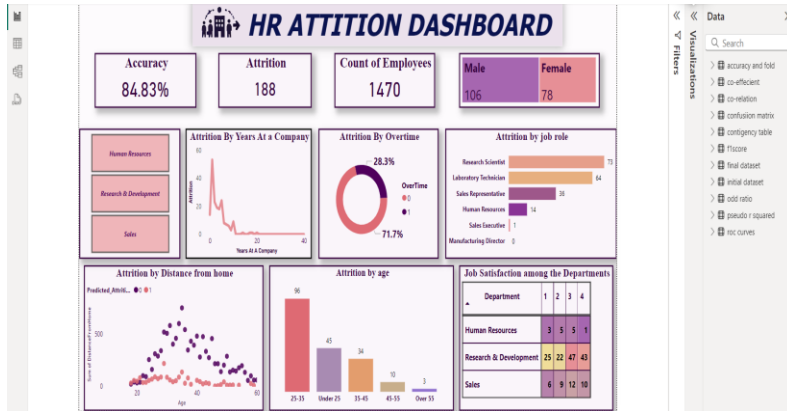
7. Figure 18. Peer Review Received 1



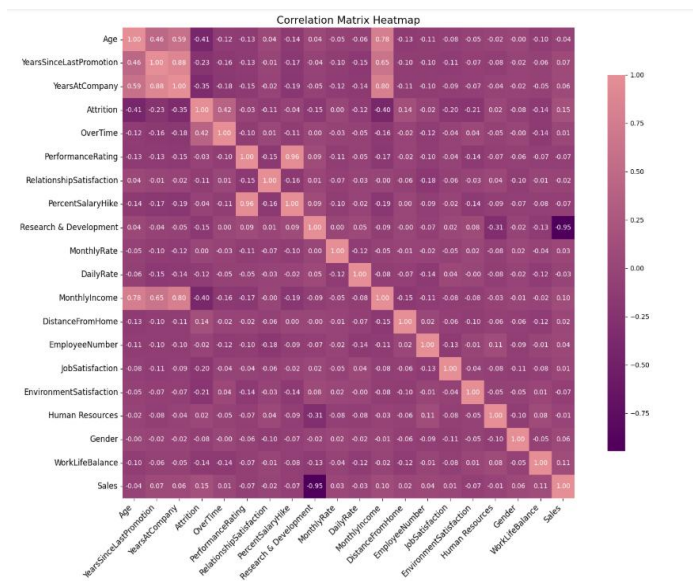
8. Figure 20: Power BI Dashboard-

Link:-

https://app.powerbi.com/groups/me/report/s/5da343c9-ac1c-466c-8f49-ee28f1fd20e0?ctid=3d1cee9c-8bf0-4375-b5b9-5c0315d1187e&pbi_source=linkShare



9. Figure 21- Correlation Heatmap



10. Figure 22- ROC Curve

