IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Recommendation System to Mitigate Attrition Risks in IT organizations using Machine Learning

[1]Kishore Kumar K, [2]Selvakumar G, [3]Harini S, [4]Priyanka G

[1]Student, [2]Assistant Professor, [3]Student, [4]Student
[1]Artificial Intelligence and Data Science,
[1]KPR Institute of Engineering and Technology, Coimbatore, India

I. Abstract : In the ever-evolving landscape of organizational management, the conservation of human resources stands out as a paramount concern, elevating employee attrition to the forefront of strategic agendas. The complexities of attrition, arising from various causes, present challenges for HR managers and department leaders striving to proactively identify these signs. The widespread consequences of attrition, including disruptions in ongoing tasks, re-employment costs, and potential compromise of core technologies, underscore the urgency of our study. Our research introduces an innovative approach, proposing a predictive model that leverages machine learning algorithms to foresee employee attrition. This model assesses 35 variables, meticulously evaluating their impact on attrition and identifying significant contributors to the phenomenon. Variables such as environmental satisfaction, overtime work, and relationship satisfaction are highlighted as key contributors. Given the multifaceted repercussions of employee attrition, a comprehensive investigation into its causes and the development of an effective predictive framework become imperative. Our research delves into organizational factors influencing attrition, utilizing advanced machine learning algorithms for in-depth analysis. Departing from traditional methods, our approach adopts a holistic learning framework, deliberately omitting specific algorithmic mentions. The study reveals crucial contributors to attrition, including factors like monthly income, hourly rate, job level, and age. By pioneering this novel approach, organizations gain strategic insight to enhance factors contributing to attrition, fostering a more stable, resilient, and harmonious workforce environment. This innovative framework marks a paradigm shift in addressing employee attrition.

Keywords: Machine learning, Supervised Learning, Logistic Regression.

II. Introduction

In the dynamic realm of organizational management, the pervasive challenge of employee attrition necessitates strategic interventions to ensure sustainable workforce stability. Departures, whether voluntary resignations or organizational decisions, bring about far-reaching consequences, including disrupted tasks, re-employment and retraining costs, and potential knowledge leakage. The attrition rate, a pivotal metric, serves as a barometer for gauging an organization's trajectory, with a high rate indicative of frequent departures and associated setbacks.

Not all industries have equal employee turnover rates. In December 2023, the average employee turnover rate was 3.8%. Of this figure, approximately 2.5% of turnovers are due to employees quitting, while the remaining is attributed to redundancies and firings. The industry with the highest turnover rate is leisure and hospitality, with a rate of 5.8% in the last month of 2023. Other sectors with higher turnover rates are construction (4.2%), the retail trade (3.9%), and real estate (3.4%). Government jobs have some of the best retention rates, with an average number of total separations at just 1.5% in December 2023. The industries within retail with the highest levels of turnover aren't terribly surprising, with restaurants (17.2%), retail (16.2%), and sporting goods (14.8%) leading the way. The roles that people left the most include lower-level, often-seasonal jobs, like retail salesperson (19.3%), food service professional (17.6%) and hospitality professional (17.0%). Media and entertainment exhibit diverse turnover rates across industries, with newspapers, online media, and sports all closely tied. However, occupation-wise, animators and 3D artists face notably higher turnover rates, more than double the global average, reflecting their high demand. Marketing specialists experience the third-highest turnover. Interestingly, only 36% of job-leavers in this sector transition to another media/entertainment job, emphasizing a substantial sector exit trend.

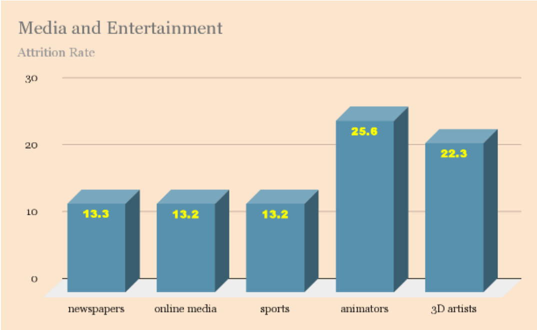

Fig 2.1: Bar graph showing the statistics about the attrition rate in Media & Entertainment Industry

The paper focuses on employing Logistic Regression as the primary machine learning technique to predict employee attrition. Logistic Regression offers a robust framework for modelling the probability of a binary outcome, making it suitable for predicting attrition events within organizations. By leveraging Logistic Regression, the study aims to provide accurate predictions of employee attrition, thereby enabling organizations to proactively address retention challenges and cultivate a stable workforce environment. In addition to predicting attrition, the study also identifies the top three reasons contributing to attrition within organizations. Through careful analysis of the dataset and feature importance measures derived from the Logistic Regression model, key factors influencing attrition rates are identified. These factors may include variables such as job satisfaction, work-life balance, career growth opportunities, compensation and benefits, organizational culture, leadership effectiveness, and workload distribution. By pinpointing the top three reasons for attrition, organizations gain actionable insights into areas that require immediate attention and intervention. Armed with this knowledge, decision-makers can implement targeted strategies to address underlying issues, mitigate attrition risks, and improve overall employee satisfaction and retention. This approach not only enhances organizational stability but also fosters a positive work environment conducive to employee engagement and productivity.

Furthermore, Logistic Regression allows for the interpretation of model coefficients, providing insights into the relative importance of each independent variable in predicting attrition. By examining the coefficients associated with different predictor variables, organizations can prioritize interventions based on the factors that have the most significant impact on attrition rates. For example, if the coefficient for job satisfaction is found to be highly negative, indicating a strong inverse relationship with attrition, efforts to improve job satisfaction levels among employees may yield substantial reductions in attrition rates. Similarly, if the coefficient for work-life balance is positive and significant, it underscores the importance of implementing policies and programs that promote a healthy balance between work and personal life to enhance employee retention. In addition to identifying the top three reasons for attrition, the Logistic Regression model can also be leveraged to develop a predictive attrition risk score for individual employees. By combining the coefficients associated with each predictor variable and the corresponding values for individual employees, organizations can calculate a personalized attrition risk score for each employee. This score provides a quantitative measure of an employee's likelihood of attrition, allowing HR professionals and managers to prioritize retention efforts and tailor interventions based on the unique needs and circumstances of each employee. Armed with this predictive insight, organizations can implement targeted retention strategies, such as career development opportunities, mentoring programs, or compensation adjustments, to mitigate attrition risks and retain valuable talent within the organization.

2.1. Background Study

In the organizational management, employee attrition has emerged as a formidable challenge with consequences that extend beyond mere personnel changes. Traditional statistics-based analysis methods often fall short in providing highly accurate predictions, particularly when faced with complex or atypical patterns. Moreover, relying on researcher-designed hypotheses imposes limitations on the ability to analyse patterns not anticipated by the researcher. In response to these limitations, a growing body of research is turning to more sophisticated data science approaches, such as machine learning, to enhance the accuracy of predicting employee attrition. Machine learning is a process wherein machines learn from patterns and training data sets, enabling them to understand hidden patterns and establish relationships within the data. This methodology is crucial for efficient machine processing, mimicking the cognitive functions of the human brain.

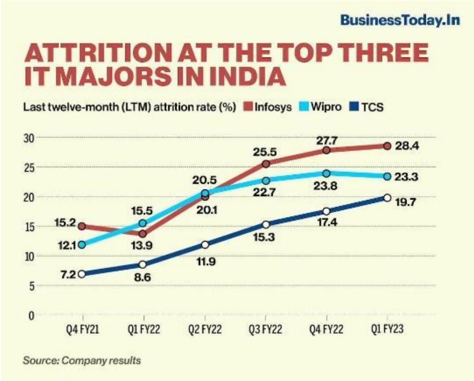

Fig 2.1.1: Comparison of Attrition in three major IT Companies.
source-url : https://www.businesstoday.in/latest/corporate/story

This image depicts the attrition rate of the 3 Major IT companies in India during the last twelve-months. Wipro has an attrition rate of 23.3% in 2023, Infosys has an attrition rate of 28.4% in 2023, TCS has an attrition rate of 19.7% in 2023. Out of all the three companies present TCS has the lowest attrition rate and Infosys has the highest attrition rate. TCS throughout the years had the lowest attrition rate compared to the others. Infosys after 2022 had the highest attrition rate. Wipro had the average attrition throughout the years. At the start of 2022, The attrition rate of Infosys was lower than Wipro but it gradually increased and overtook Wipro.

Several studies have contributed to the exploration of machine learning in the context of employee attrition. These studies utilized diverse datasets and applied various classification algorithms without explicitly mentioning algorithm names or accuracy scores. For instance, Chakraborty et al. (2021) employed Logistic Regression, Linear Discriminant Analysis, Ridge Classification, Lasso Classification, Decision Tree, and Random Forest to predict turnover probabilities. Fallucchi et al. (2020) presented an attrition prediction model using HR data with variables and samples, highlighting the superior performance of Logistic Regression. Similarly, Zhao et al. (2018) proposed a comprehensive predictive model utilizing different data sizes and complexity levels. Ganthi et al. (2022) focused on predicting employee attrition rates using classification algorithms and regularization techniques. The background of machine learning, characterized by its iterative learning from data and adaptability, aligns well with the contemporary need to overcome the limitations of traditional analysis methods. While machine learning is not a new concept, recent advancements have facilitated its application to vast datasets, enabling computers to autonomously learn and make decisions, a development that brings renewed enthusiasm to the field. This paper emphasizes the critical need for effective approaches to predict employee attrition, with a particular focus on machine learning as a powerful tool for extracting valuable insights and patterns from complex workforce dynamics.

3. Related Works

The reviewed literature offers a comprehensive exploration of various approaches employed in predicting employee attrition, reflecting the growing significance of this challenge in contemporary organizational settings. Piotr Płoński and colleagues [1] proposed analytic methods to enhance Human Resources (HR) management, utilizing machine learning to predict employee attrition. By leveraging a dataset of 1200 employees, they employed binary classification to predict attrition, highlighting the importance of such predictive analytics for large-scale HR management. Le Zhang and Graham Williams [2] underscored employee retention as a paramount challenge for companies, advocating for the recognition of behavioral patterns to better understand employees. Their approach involved using R for predictions, employing feature extraction methods such as word-to-vector and term frequency. Furthermore, they concluded that ensemble techniques, combining multiple models, can significantly enhance predictive performance, offering valuable insights for companies striving to improve employee retention strategies. Ashish Mishra et al. [3] emphasized the criticality of recruiting the right personnel for effective talent management, highlighting the initial steps in addressing attrition challenges. This underscores the significance of a holistic approach, starting from the recruitment phase, to mitigate attrition risks within organizations.

Moving beyond individual studies, the literature reveals a broader exploration of methodologies and techniques applied in predicting employee attrition. The Performance Assessment of Data Balancing Techniques [11] addresses the challenges posed by imbalanced datasets, offering insights into the effectiveness of various data balancing techniques. This research experimentally demonstrated that data balancing enhances the performance of applied classifiers, laying the groundwork for improving predictive models in attrition scenarios. Neural network techniques, explored by [12], showcased their effectiveness in solving learning problems, particularly when dealing with vast datasets. The study highlighted the advantages of neural networks, such as high-speed processing and parallelism with big data, making them instrumental in addressing real-world problems. The performance comparison and challenges identified contribute to a better understanding of the capabilities and limitations of neural network techniques in the context of employee attrition prediction.The study by [13] delved into machine learning-based classification algorithms for predicting employee attrition rates, applying techniques such as K-Nearest Neighbors, extreme gradient boosting, Ada Boosting, Decision Tree, neural networks, and Random Forest. By utilizing HR employee data, the research demonstrated the applicability of these techniques, achieving an accuracy score of 88%, providing organizations with valuable insights into attrition prediction. The automated prediction of employee attrition proposed by [15] further showcased the practical application of various machine learning models, including Ad boost Model, Random Forest Regressor, Decision Tree, Logistic Regression, and Gradient

Boosting Classifiers. Their goal of accurate detection of employee attrition aligns with organizational objectives to enhance employee satisfaction. In summary, the reviewed literature provides a rich tapestry of methodologies, ranging from machine learning algorithms to data balancing techniques and neural network applications, all aimed at predicting and understanding employee attrition. These studies collectively contribute to the evolving landscape of employee management strategies, offering valuable insights for organizations seeking to mitigate attrition risks and foster a more stable workforce environment.

4. Methodology

4.1Logistic Regression

Logistic Regression is a statistical method used for binary classification tasks, where the goal is to predict the probability that a given input belongs to one of two possible outcomes or classes. Despite its name, logistic regression is a classification algorithm, not a regression algorithm. The logistic regression model estimates the probability of the binary outcome using a logistic function, also known as the sigmoid function. The logistic function ensures that the predicted probabilities lie between 0 and 1, making it suitable for binary classification tasks. If the predicted probability is greater than or equal to 0.5, the input is classified as belonging to class 1; otherwise, it is classified as belonging to class 0.The logistic function is defined as:

P(Y=1) = 11+e-(0 +1x)

In the above equation, P(Y = 1) is the probability of the dependent variable Y being 1

e is the base of the natural algorithm

0 is the intercept

β1, β2, …. , βn are the coefficients of the independent features X1,X2,....,Xn respectively. The logistic function 11 + e-z ensures that the predicted values are between 0 and 1, making it suitable for binary classification problems. The logistic regression model estimates the coefficients β1, β2 ,...,βn based on the training data to make predictions on the data.

4.2Variable Selection

Variable selection is a critical step in building machine learning (ML) models, as it involves choosing a subset of features (variables) that are most relevant for predicting the target variable. The goal is to improve model performance, reduce overfitting, and enhance interpretability. The required variables for this model are selected using various visualization & data manipulation techniques. Every categorical and numerical attribute were visualized with the attrition attribute and the attributes which are correlated with it were selected for further processing.

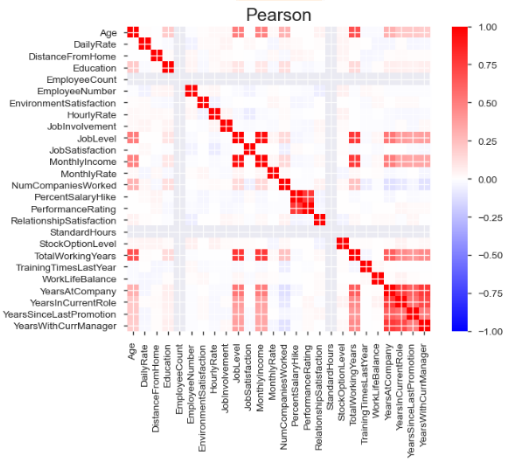

Fig 4.2.1: Heatmap depicting the correlation between different attributes
Source: Google colab - Data visualization on IBM HR Dataset

Heatmap showing the correlation between different attributes present in the dataset. This heatmap is used to find the relation between two attributes. This heatmap is one of the factors that can be used to identify the required attributes to create a model. A heatmap is a graphical representation of data where values are depicted using color gradients. In the context of the attrition prediction project, a heatmap could be utilized to visually showcase the correlation between various attributes and the attrition rate. Each attribute's impact could be represented by different shades or colours, allowing users to quickly identify the most influential factors. This visual tool enhances the interpretability of data, aiding both users and organizations in understanding the relationships within the dataset. By presenting complex information in an intuitive and accessible format, the heatmap becomes a valuable component in the decision-making process, guiding efforts to address attrition effectively.

ATTRIBUTES USED: Attrition, BusinessTravel, Education, EducationField, Gender, JobLevel, MaritalStatus, OverTime, PerformanceRating, StockOptionLevel, TrainingTimesLastYear, Total_Satisfaction, Age, Department, DistanceFromHome, JobRole, MonthlyIncome, NumCompaniesWorked, TotalWorkingYears.

5.Implementation

Machine learning is about creating algorithms that learn from data. Through this process, machines can make predictions or decisions without explicit programming. It's categorized into supervised, unsupervised, semi-supervised, and reinforcement learning. In supervised learning, algorithms are trained on labelled data to predict outcomes. Unsupervised learning discovers patterns in unlabelled data. Semi-supervised learning uses both labelled and unlabelled data. Reinforcement learning teaches agents to make decisions through trial and error. Common techniques include decision trees, neural networks, SVMs, and k-nearest neighbours. Data preprocessing involves cleaning, normalization, and feature engineering. Models are trained using optimization algorithms like gradient descent. Evaluation metrics like accuracy, precision, recall, and F1 score assess model performance.

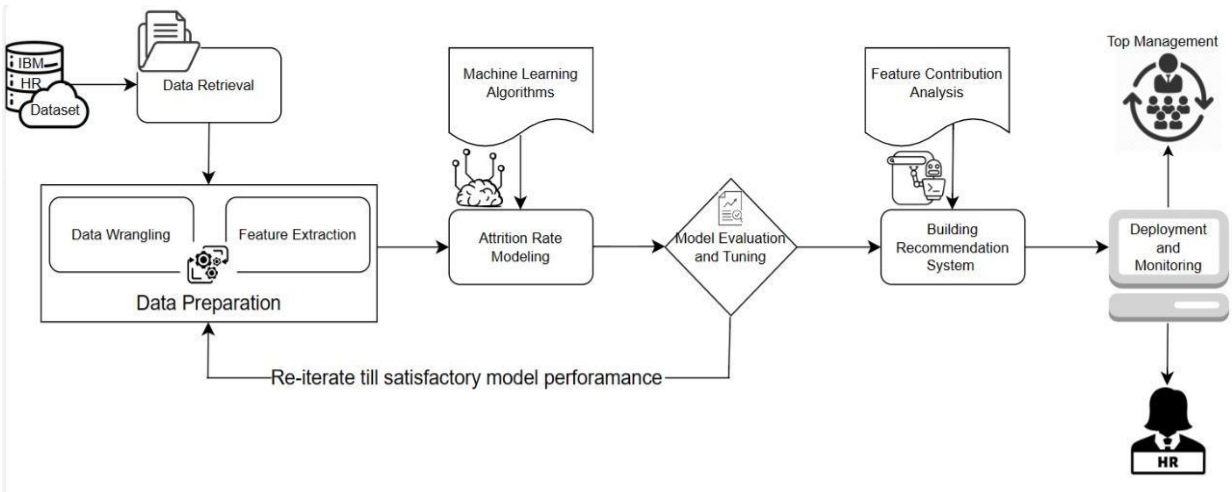

Fig 5.1.1 System Architecture

The Figure 5.1.1 represents the workflow of our project. The data - IBM HR dataset was collected. Then the collected data undergoes data preparation. The data preparation consists of data wrangling and feature extraction. Then an ML model is built to find the attrition rate percentage of an employee. The developed model undergoes model evaluation and tuning. Then using feature contribution analysis, a recommendation system for mitigating attrition rate is built. Then the recommendation system will be deployed and monitored. The recommendation system will be used by HR and the top management in an organization.

Overfitting occurs when a model learns noise instead of patterns. Cross-validation helps in estimating a model's performance. Hyperparameter tuning optimizes model performance. Model deployment involves deploying trained models into production. Deep learning, a subset of ML, uses neural networks with multiple layers. TensorFlow and PyTorch are popular deep learning frameworks. MLaaS (Machine Learning as a Service) provides ML capabilities via cloud platforms. Ethical considerations include bias, fairness, and privacy in ML applications. Continuous learning is essential to stay updated with new techniques and advancements. Machine Learning Algorithm Logistic Regression (LR) is applied in this research study. The RFC is our proposed approach for predicting employee attrition.

5.1. Data Collection

The dataset utilized in this study originates from the 'IBM HR Analytics Employee Attrition & Performance data,' accessible through Kaggle's Competitions platform and developed by IBM's team of data scientists. This dataset proves to be apt for exploring employee attrition prediction due to its comprehensive coverage of factors closely associated with attrition trends. Noteworthy examples include analyses on the relationship between job roles and attrition rates based on distance from home, as well as comparisons of average monthly income across different education levels and their respective attrition rates. Employing this dataset as our foundation, our study focuses on training a predictive model tailored to forecast employee attrition using machine learning methodologies. Python 3.7 serves as the primary tool for training our predictive model. The dataset encompasses a total of 1470 records and encompasses 31 variables, with one serving as the dependent variable (employee attrition) and remaining 30 as independent variables. These independent variables are broadly categorized into three main groups: personal attributes, job-related factors, and career-related aspects.

Analysing the demographic composition within the dataset reveals a diverse distribution across different age groups. The dataset exhibits varying proportions of individuals across age brackets, including 1.6% in their 10s, 25.8% in their 20s, 42.5% in their 30s, 21.1% in their 40s, 8.7% in their 50s, and a negligible 0.2% comprising those aged 60 and above. Gender distribution within the dataset showcases a nearly balanced ratio, with 49.2% identified as male and 50.7% as female. Further examination of educational backgrounds within the dataset reveals a diverse distribution across different attainment levels. Notably, educational achievements are segmented into categories, with 13% indicating below college level, 21% at the college level, 43% holding bachelor's degrees, 21% possessing master's degrees, and a minor 2% holding doctorate degrees. These detailed insights into the composition of independent variables provide a comprehensive understanding of the dataset's demographics and educational backgrounds, forming the basis for our subsequent analyses and model training endeavours.

Weights Assigned

S.No	Parameter	Weight	Remark
1	Education	0.15	Moderate Positive Impact
2	Job Level	0.2	Strong Positive Impact
3	Number of companies worked	0.3	Very Strong Positive Impact
4	Age	0.2	Moderate Positive Impact
5	Department	-0.1	Weak Negative Impact
6	Monthly Income	-0.05	Strong Negative Impact
7	Years at company	0.2	Strong Positive Impact
8	Percent Salary Hike	-0.1	Moderate Negative Impact
9	Training times last year	0.2	Strong Positive Impact
10	Years With Current Manager	0.1	Moderate Positive Impact

In addition to the fundamental attributes like education level, job level, age, and distance from home, we can consider several other pertinent factors to enhance the predictive accuracy of our employee attrition model. These include the employee's social network within the workplace, measured by the number of friends or connections they have; their punctuality and adherence to office hours, captured through arrival and departure times; the duration of their presence in the office and breaks taken during work hours, reflecting their engagement and work habits. Moreover, behavioural aspects such as workplace demeanour, facial expressions denoting emotions like happiness or sadness, and overall happiness index can provide valuable insights into employee satisfaction and potential attrition risk. Additionally, factors like career advancement opportunities, diversity and inclusion, and perceived organizational support can further enrich the predictive model. By incorporating a diverse range of attributes, both fundamental and nuanced, we can develop a comprehensive employee attrition model that accurately anticipates workforce dynamics and empowers organizations to implement proactive retention strategies. The project's webpage facilitates a user-friendly experience by collecting 12 crucial attributes to predict attrition rates. Users input their data, and the system promptly calculates the percentage of attrition. Notably, the platform goes beyond mere predictions, offering insightful recommendations tailored to the individual's circumstances. For instance, if distance from home emerges as the primary influencing factor, proactive measures like arranging temporary housing or accommodation are suggested. Additionally, the system highlights the top three influential attributes, providing valuable insights to help organizations address potential attrition issues with targeted interventions and personalized solutions. This holistic approach empowers users with actionable information to mitigate attrition and foster a more engaged workforce.

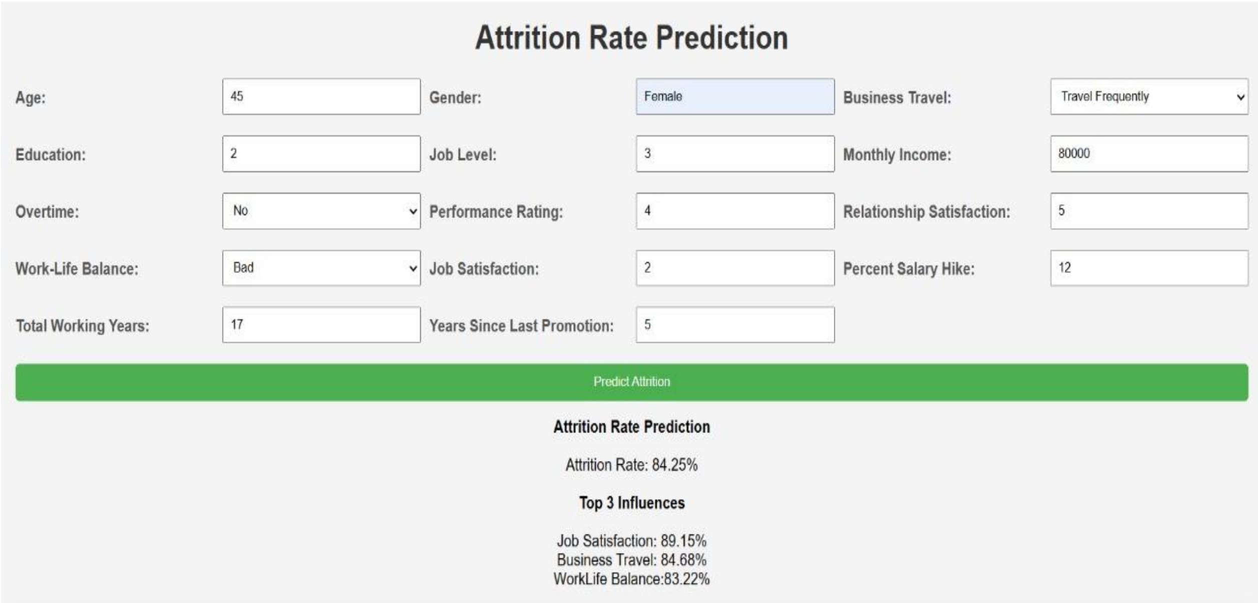

Fig 5.1.2 Web Page

Figure 5.1.2 presents the web page interface of our attrition prediction model, showcasing a comprehensive set of attributes essential for accurate forecasting. These attributes include age, education level, overtime work, work-life balance, total working years, gender, job level, performance rating, job satisfaction, years since last promotion, business travel frequency, monthly income, relationship satisfaction, and percent salary hike. Through the model's analysis, users receive the percentage of attrition rate, offering a clear understanding of the likelihood of employee turnover within the organization. Additionally, the model highlights the top three factors influencing attrition, providing valuable insights for strategic intervention and targeted retention efforts.

In the above analysis, the Attrition Rate is determined to be 84.25%, with key contributing factors identified. Job Satisfaction stands out prominently with a high percentage of 89.15%, indicating a significant influence on attrition. Business Travel contributes substantially with a percentage of 84.68%, highlighting the importance of considering travel-related factors in addressing attrition. Moreover, WorkLife Balance, though slightly lower at 83.22%, remains a notable factor influencing attrition rates. It's crucial to note that a positive work environment and employee satisfaction play a pivotal role in retention. By understanding these influencing factors, organizations can tailor interventions to create a more conducive and satisfying work environment for their employees."

Logistic Function (Sigmoid)

The dependent variable Y as the likelihood of employee attrition
where.,
$Y = 1$ indicates attrition,
$Y = 0$ indicates no attrition.
The independent variable $(X_1, X_2, ..., X_n)$ are various attributes related to employees such as Age, Gender, Business Travel frequency, Job Satisfaction, etc. The logistic regression equation becomes:

$$\mathrm{P(Attrition = 1)} = \frac{1}{1+e^{-(\beta_0 + \beta_1.Age + \beta_2.Gender + \cdots + \beta_n.BusinessTravel)}}$$

In the above equation β1, β2, ..., βn are the coefficients estimated by the logistic regression model. The attributes like Age, Gender, Business Travel, etc., contribute to the logistic function, determining the probability of an employee experiencing attrition. The logistic regression model calculates the likelihood of an employee leaving the company based on their individual characteristics, providing insights into factors influencing attrition.

In the domain of workforce analytics, the logistic regression model applied to predict employee attrition yields probabilities denoting the likelihood of attrition occurrence. This probability, symbolized as $P(\text{Attrition} = 1)$ is subsequently converted into a percentage scale to enhance interpretability. The derived metric, referred to as the "Attrition Percentage," encapsulates the model's predictive confidence in the prospect of attrition for a given employee. Expressed formally as:

$$\text{Attrition Percentage} = \text{P(Attrition} = 1)\times 100$$

By way of illustration, if the model assigns a probability of 0.75 to an employee, the corresponding attrition percentage becomes 0.75 * 100 = 75% This percentage signifies the model's level of conviction in predicting attrition, with higher percentages denoting heightened confidence in the likelihood of attrition for the specified individual. This transformation of probabilities into percentages serves to render the model's predictions more accessible and facilitates effective communication of attrition predictions within the organizational context.

$$\text{Logit(Attrition)}=\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \cdots + \beta_n \cdot X_n$$

Logit(Attrition): The linear combination of coefficients (β)and feature values (X).
Logit(Attrition) represents the log-odds of the event "Attrition."

$$\text{P(Attrition)} = \frac{1}{1+e^{Logit(attrition)}}$$

In the above equation the terms represent,
P(Event) : The probability of the event occurring
e : Euler's number(approximately 2.71828)

1. Logistic Regression Model: It combines the features $(X_1, X_2, \ldots, X_n)$ with their respective coefficients (β1,β2,...,βn) to calculate the log-odds the log-odds.
2. Log-Odds to Probability: The logistic (sigmoid) function transforms the log-odds into probabilities between 0 and 1. The probability represents the likelihood of the event occurring.

For a specific employee, you can use the coefficients and feature values to calculate the log-odds. Then apply the logistic function to obtain the predicted probability.

$$\text{Predicted Probability} = \frac{1}{1+e^{Log-Odds}}$$

This probability represents the likelihood of the event "Attrition" occurring. Positive coefficients increase the log-odds (and thus, the odds), while negative coefficients decrease them. The sigmoid function ensures that the predicted probability falls within the range of 0 to 1.

5.2 Feature Contribution Analysis

In the context of employee attrition analysis using the IBM retention dataset, understanding feature contributions goes beyond merely identifying influential factors. It involves a deeper comprehension of the coefficients assigned to each attribute in the logistic regression model. These coefficients represent the quantitative impact of each feature on the log-odds of attrition.

A positive coefficient indicates that an increase in the corresponding attribute raises the odds of attrition, while a negative coefficient suggests the opposite effect. The magnitude of the coefficient reflects the strength and direction of this influence. For example, a higher positive coefficient for 'overtime' signifies a more substantial impact on the likelihood of attrition for employees working overtime.

$$\text{Contribution } (Feature_i) = \beta_i * \text{(Valuenew - Valuereference)} \qquad (5.2.1)$$

In the equation 5.2.1,
Contribution $(Feature_i)$ represents the contribution of the i-th feature to the predicted log-odds.

β_i is the coefficient associated with the i-th feature

Valuenew is the new value of the feature

Valuereference is the reference value of the feature (typically 0 or the mean value).

Moreover, to delve deeper into understanding the contribution of each feature to the prediction for individual employees, we employed SHAP (SHapley Additive exPlanations). SHAP values provide a unified measure of feature importance and directionality in the prediction task.

The SHAP (SHapley Additive exPlanations) value for a feature in the context of our analysis with the employee attrition dataset represents the impact of that feature on the model's output for a specific employee compared to the average model output. Mathematically, it is defined as:

$$\text{SHAP } (Feature_i) = \Sigma \text{ [(Model Prediction(x') - E[Model Prediction]) * } \phi\text{i(s')]}$$

Here,
SHAP $(Feature_i)$ represents the SHAP value for the i-th feature

Model Prediction(x') is the model's prediction for the instance x', which in our case would be the predicted probability of attrition for a specific employee.

E[Model Prediction] is the expected model prediction over all instances, which can be interpreted as the average predicted probability of attrition across the entire dataset.

$\phi i(s')$ is the Shapley value for the i-th feature in the instance s', which quantifies the contribution of the feature to the difference between the model prediction for the specific instance x' and the expected model prediction.

In our analysis, we converted the raw SHAP values into percentages for meaningful insights. By doing so, we provided a more interpretable representation of feature contributions, allowing stakeholders to grasp the relative importance of each feature in predicting attrition.

Analysing feature contributions, along with SHAP values, offers actionable insights into the specific nature of the relationship between attributes and attrition. Features with larger absolute coefficients and SHAP values contribute more significantly to the predictive power of the model. Therefore, a detailed examination of these coefficients and SHAP values not only reveals influential attributes but also aids in prioritizing intervention strategies.

In summary, feature contribution analysis, coupled with SHAP values and a nuanced understanding of coefficients, enables organizations to adopt a strategic approach in addressing attrition challenges. By fostering a data-driven perspective on employee retention, organizations can implement targeted intervention strategies and mitigate risks associated with influential attributes.

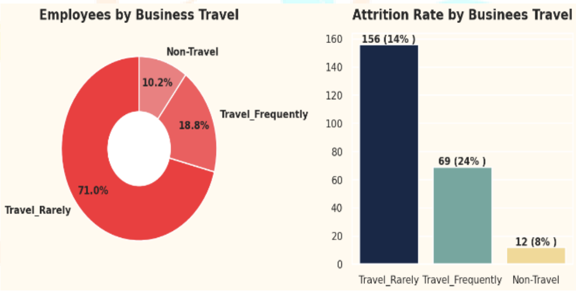

Fig: 5.2.1-Visualization of Employees by Business Travel

6. Results and Discussion

The examination of employee attrition prediction utilizing machine learning methods offered profound insights into the multifaceted dynamics of employee turnover within organizational settings. Leveraging the expansive dataset provided by the IBM HR Analytics Employee Attrition & Performance data, our study embarked on a journey to deploy a Logistic Regression algorithm to construct predictive models customized to forecast employee attrition accurately. Also the model gives the top three factors for their attrition. Our findings illuminated the significant predictors influencing attrition rates within organizational contexts. Notably, variables such as environmental satisfaction, overtime work, and relationship satisfaction emerged as pivotal contributors to the propensity of employees to leave their roles. Through the sophisticated analysis facilitated by machine learning techniques, we unraveled intricate relationships between these variables and attrition trends, offering actionable insights for organizational decision-makers. The demographic analysis of the dataset provided intriguing insights into the composition of the workforce across various age groups, genders, and educational backgrounds. The results of the predictive model reveal valuable insights . Logistic Regression emerged as the top-performing model with an accuracy of 87%. This unexpected outcome highlights the robustness of Logistic Regression, known for its simplicity and interpretability. Despite its linear nature, Logistic Regression demonstrated competitive accuracy, suggesting that even simpler models can yield reliable predictions in attrition forecasting tasks. The Precision, Recall, and F1-score metrics further validate Logistic Regression's performance, with it consistently demonstrating strong performance across all metrics. This underscores the versatility and efficacy of Logistic Regression in accurately identifying employees likely to attrite.
The predictive models were trained and evaluated using the dataset on employee attrition. The performance metrics were measured in terms of accuracy, precision, recall, and F1-score, with the results summarized in the table below:

Performance Metrics	Logistic Regression
Accuracy	87%
Precision	0.85
F1 Score	0.88
Recall	0.83

Table 6.1 - Comparison of ML algorithms based on different performance metrics

Accuracy is a fundamental metric derived from the confusion matrix that provides an overall measure of how well a classification model is performing. The accuracy of the model is 87%. Precision is a metric derived from the confusion matrix that focuses on the accuracy of the positive predictions made by a classification model. The precision of our model is 0.85. The F1 score is a metric derived from the confusion matrix that provides a balance between precision and recall. The F1 Score for our model is 0.88. Recall, also known as Sensitivity or True Positive Rate, is a metric derived from the confusion matrix that measures the ability of a classification model to capture all positive instances. The Recall of our model is 0.83.

Overall, these findings emphasize the importance of considering simpler models like Logistic Regression alongside more complex ones, as they can provide valuable insights and reliable predictions for organizations seeking to mitigate the effects of employee attrition. As the field of data science continues to evolve, the application of machine learning in HR management holds promise for revolutionizing organizational performance and enhancing employee satisfaction.

7. Conclusion

In conclusion, this study presents a comprehensive exploration of employee attrition within organizational management, highlighting the critical need for strategic interventions to ensure workforce stability and organizational resilience. This study offers a comprehensive evaluation of machine learning algorithms for predicting employee attrition, with a focus on Logistic Regression model. Logistic Regression emerged as the top-performing model, providing individualized predictions indicating the likelihood of an employee leaving the company, such as percentage of attrition rate for a specific employee. This personalized approach demonstrates the model's robustness and versatility. The results underscore the importance of considering simpler models alongside more complex ones, as Logistic Regression's simplicity and interpretability make it a valuable tool for accurately identifying employees likely to attrite. These insights can empower organizations to proactively implement retention strategies, fostering a stable and productive workforce environment while minimizing the disruptive effects of attrition. Moving forward, further research could explore additional variables and feature engineering techniques to enhance the predictive capabilities of these models, ultimately contributing to more effective attrition management strategies in organizational settings. In the era of artificial intelligence and data science, the nuanced understanding of employee attrition is paramount, given its profound impact on organizational culture and financial sustainability. With millions of employees departing monthly, proactive measures become imperative for organizational management. This paper contributes a tailored approach, focusing on specific machine learning technique, to mitigate attrition risks and foster resilient organizational cultures. Through a thorough review of related literature, encompassing various predictive methodologies and techniques, this study consolidates insights and best practices. Logistic Regression stands out for its robust predictive capabilities in the context of employee attrition. In essence, this research emphasizes the importance of proactive attrition management strategies, informed by sophisticated data analysis and machine learning techniques. By embracing innovative approaches, organizations can navigate attrition dynamics more effectively, ensuring workforce stability and maximizing organizational productivity.

8. References

[1] Abiodun, O.I.; Jantan, A.; Omolara, A.E.; Dada, K.V.; Mohamed, N.A.E.; Arshad, H. State-of-the-art in artificial neural network applications: A survey. Heliyon 2018, 4, e00938. [CrossRef]

[2] Aggarwal, S.; Singh, M.; Chauhan, S.; Sharma, M.; Jain, D. Employee Attrition Prediction Using Machine Learning Comparative Study. Smart Innov. Syst. Technol. 2022, 265, 453–466. [CrossRef]

[3] Aljedani, N.; Alotaibi, R.; Taileb, M. HMATC: Hierarchical multi-label Arabic text classification model using machine learning. Egypt. Inform. J. 2021, 22, 225–237. [CrossRef]

[4] Ashish Mishra (Data Scientist, Experfy), "Using Machine Learning to Predict and explain Employee Attrition".

[5] Costa, André & Veloso, Adriano. (2015). Employee Analytics through Sentiment Analysis. 10.13140/RG.2.1.1623.3688.

[6] Dr. Jonathan Erhardt, "Artificial Intelligence: Opportunities and Risks," Policy paper by the Effective Altruism Foundation.

[7] Ganthi, L.S.; Nallapaneni, Y.; Perumalsamy, D.; Mahalingam, K. Employee Attrition Prediction Using Machine Learning Algorithms. Lect. Notes Netw. Syst. 2022, 288, 577–596. [CrossRef]

[8] Jadhav, A.; Mostafa, S.M.; Elmannai, H.; Karim, F.K. An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task. Appl. Sci. 2022, 12, 3928. [CrossRef]

[9] Jia, X.; Cao, Y.; O'Connor, D.; Zhu, J.; Tsang, D.C.W.; Zou, B.; Hou, D. Mapping soil pollution by using drone image recognition and machine learning at an arsenic-contaminated agricultural field. Environ. Pollut. 2021, 270, 116281. [CrossRef] [PubMed]

[10] Khare, Rupesh, Dimple Kaloya, and Gauri Gupta. "Employee Attrition Risk Assessment using Logistic Regression Analysis," from 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence.

[11] Lao, Randy. "Predicting Employee Kernelover," Kaggle.

[12] Mishra, Ashish (Data Scientist, Experfy), "Using Machine Learning to Predict and explain Employee Attrition".

[13] Piotr Płoński (MLJAR), "Human-first Machine Learning Platform," Human Resource Analytics Predict Employee Attrition.K. Elissa, "Title of paper if known," unpublished.

[14] Pyke, Sandra W., and Peter M. Sheridan. "Logistic Regression Analysis of Graduate Student Retention," from The Canadian Journal of Higher Education, Vol. XXIII-2, 1993.

[15] Raza, Ali, Kashif Munir, Mubarak Almutairi, Faizan Younas, and Mian Muhammad Sadiq Fareed. 2022. "Predicting Employee Attrition Using Machine Learning Approaches" Applied Sciences 12, no. 13: 6424. https://doi.org/10.3390/app12136424

[16] Ramchandra, Reshma N., and C. Rajabhushanam. "Machine learning algorithms performance evaluation in traffic flow prediction. Mater. Today Proc. 2022, 51, 1046–1050. [CrossRef]

[17] Tsai, I.-J.; Shen, W.-C.; Lee, C.-L.; Wang, H.-D.; Lin, C.-Y.; Tsai, I.-J.; Shen, W.-C.; Lee, C.-L.; Wang, H.-D.; Lin, C.-Y.; et al. Machine Learning in Prediction of Bladder Cancer on Clinical Laboratory Data. Diagnostics 2022, 12, 203. [CrossRef] [PubMed]

[18] Zhang, Le and Graham Williams (Data Scientist, Microsoft), "Employee Retention with R based Data Science Accelerator".