A Case Study on

Predictive Analytics for Employee Retention: Leveraging Machine Learning to Understand and Mitigate Attrition Risks

By

HIMANSHU KOTKAR [10]

HSNC UNIVERSITY, MUMBAI

School of Applied Sciences
Department of Data Science and Business Analytics

April 2024

Abstract

Employee attrition within the HR department of any company can be a significant challenge, impacting organizational stability and workforce management. In this case study, we delved deep into the reasons behind high attrition rates, aiming to uncover underlying factors and develop effective strategies to retain valuable talent. Leveraging advanced predictive modeling techniques, we meticulously analyzed various aspects such as employee demographics, job satisfaction levels, and work-life balance to understand their influence on attrition risk.

Our findings shed light on the key drivers of employee turnover within the HR department, providing invaluable insights for HR management to address these issues proactively. Armed with a comprehensive understanding of the factors contributing to attrition, organizations can implement targeted retention strategies to foster a more engaging work environment and enhance employee satisfaction. By harnessing the power of data-driven insights, companies can cultivate a culture of retention and empower their HR teams to effectively manage talent attrition, ultimately leading to greater organizational success and employee satisfaction.

Keywords

1. Employee Attrition: The phenomenon of employees leaving a company voluntarily or involuntarily, often measured by the rate at which employees depart over a specific period.

2. Predictive Analytics: The practice of extracting insights from data to predict future outcomes or behaviors, often employing statistical algorithms and machine learning techniques.

3. Employee Retention: The strategies and practices employed by organizations to retain their employees and reduce turnover rates, encompassing various initiatives such as career development, benefits, and work-life balance.

4. Machine Learning: A branch of artificial intelligence focused on developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed.

Introduction

Employee attrition is a significant concern for organizations across various industries, impacting workforce stability and organizational effectiveness. In this case study, we focus on understanding and addressing the issue of employee attrition within the Human Resources (HR) department of companies.

The purpose of this study is to delve into the factors contributing to high attrition rates in HR departments and to develop effective strategies to mitigate attrition risk. By examining the demographics, job satisfaction levels, and work-life balance of HR employees, we aim to uncover insights that can inform HR management practices and improve retention efforts. This case study seeks to provide actionable recommendations for HR departments to enhance employee satisfaction, reduce turnover, and foster a more engaging work environment within the HR function.

Through a comprehensive analysis of attrition trends and potential drivers, this study aims to contribute to the development of effective retention strategies tailored to the unique needs of HR professionals.

This project originates from the possibility of enhancing employee contentment, cutting down expenses, elevating organizational efficiency, and fostering a favorable workplace atmosphere. It represents a chance to leverage data and analytics to enact substantial improvements that are advantageous to both employees and the entire organization.

Objective/Problem Definition

The objective of this case study is to identify the factors contributing to employee attrition within the HR department of companies and to develop a predictive model to forecast attrition risk.

The inputs to this problem include various demographic and job-related features of HR employees, such as age, gender, job level, job satisfaction, work-life balance, and performance ratings.

The output of the predictive model is a probability score indicating the likelihood of an HR employee leaving the company within a specified timeframe. By accurately predicting attrition risk, HR departments can proactively implement retention strategies to retain valuable talent and maintain workforce stability.

This problem definition aims to guide the development of predictive modeling techniques that can provide actionable insights to HR managers, enabling them to make informed decisions and take proactive measures to reduce attrition rates within the HR department.

Algorithm Definition

1. Logistic Regression:

A statistical method used for binary classification, where the algorithm models the probability of an event occurring as a function of input features. It predicts the likelihood of an outcome (e.g., yes/no, 1/0) based on linear relationships between input variables and applies a sigmoid function to transform the output into probabilities.

2. K-Nearest Neighbors (KNN):

A non-parametric classification algorithm that assigns a class label to an input data point based on the majority class of its k nearest neighbors in the feature space. It makes predictions by calculating distances between data points and finding the most common class among its nearest neighbors.

3. Decision Tree:

A supervised learning algorithm that constructs a tree-like structure to make decisions by partitioning the feature space into segments based on the values of input features. It recursively splits the data into subsets, with each split maximizing the homogeneity of the target variable within the subsets.

4. Random Forest:

An ensemble learning method that builds multiple decision trees during training and combines their predictions through a voting mechanism to improve accuracy and robustness. It randomly selects subsets of features and data points to train each tree, reducing overfitting and enhancing generalization performance.

Research Methodology

- Load the Dataset: The HR Attrition Dataset is loaded using the pd.read_csv() function. The head() and info() methods are used to display the first few rows and get information about the dataset, respectively.

- Understanding the Dataset: Basic Information about the dataset is generated, numerical and categorical attributes are enlisted.

- Data Cleaning: Any missing values in the dataset are dropped using the dropna() method.

- Data Preprocessing:

    - The target variable 'Attrition' is mapped to binary values (1 for 'Yes' and 0 for 'No').
    - Selected features are extracted from the dataset and one-hot encoded using the get_dummies() function.
    - Upsampled the target variable “Attrition”.As the data is very much biased towards “No”, biases data will provide biased result so for achieving better accuracy of model.

- Splitting the Dataset: The dataset is split into training and testing sets using the train_test_split() method from scikit-learn.

- Implementing Machine Learning Algorithms: Logistic Regression, K Nearest Neighbour, Decision Tree, and Random Forest classifiers are initialized and trained using the training data.

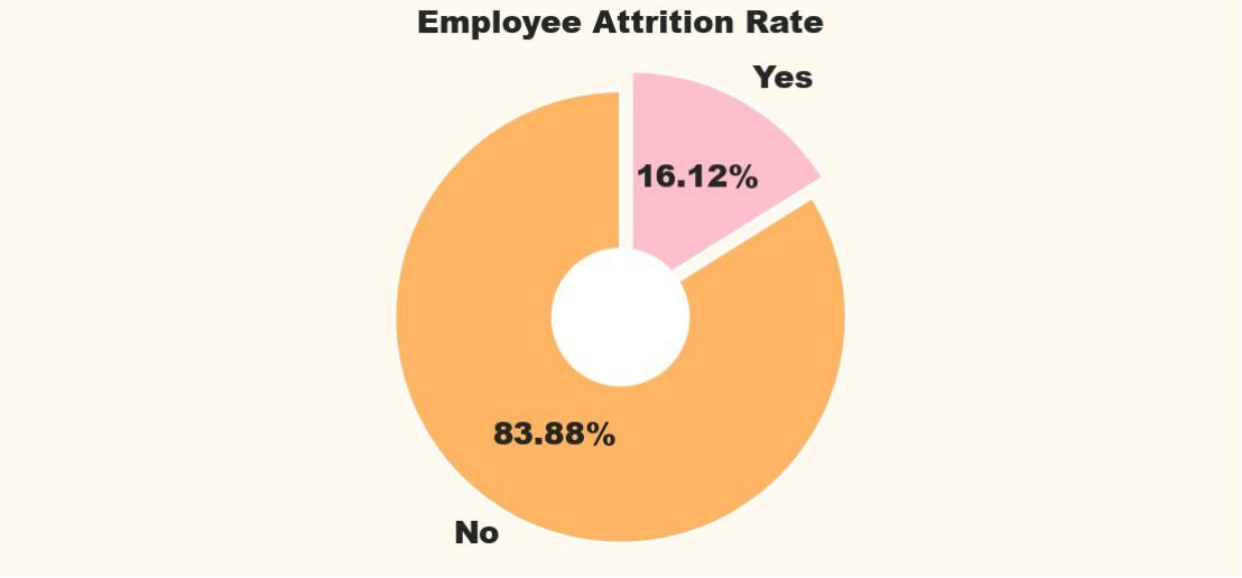
- Model Evaluation: The accuracy score and confusion matrix are computed to evaluate the performance of each algorithm on the testing data.

- Results: The results, including the accuracy and confusion matrix, are printed for each algorithm.

- Model Performance Comparison: The hvPlot library is used to visualize the ROC curve diagram comparing the performance of all models used.

Results

1] DISTRIBUTION OF TARGET VARIABLE



Distribution of target variable "ATTRITION"

Observation:

1. The employee attrition rate of this organization is 16.12%.
2. According to online resources and domain experts, it says that the attrition rate 4% to 6% is normal in organization.

2] STATISTICAL DESCRIPTION OF VARIABLES

```
In [8]: # Descriptive Analysis of numerical attributes
        df.describe().T
```

Out[8]:

	count	mean	std	min	25%	50%	75%	max
Age	1470.0	36.923810	9.135373	18.0	30.00	36.0	43.00	60.0
DailyRate	1470.0	802.485714	403.509100	102.0	465.00	802.0	1157.00	1499.0
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.00	7.0	14.00	29.0
EmployeeCount	1470.0	1.000000	0.000000	1.0	1.00	1.0	1.00	1.0
EmployeeNumber	1470.0	1024.865306	602.024335	1.0	491.25	1020.5	1555.75	2068.0
HourlyRate	1470.0	65.891156	20.329428	30.0	48.00	66.0	83.75	100.0
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.00	4919.0	8379.00	19999.0
MonthlyRate	1470.0	14313.103401	7117.786044	2094.0	8047.00	14235.5	20461.50	26999.0
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.00	2.0	4.00	9.0
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.00	14.0	18.00	25.0
StandardHours	1470.0	80.000000	0.000000	80.0	80.00	80.0	80.00	80.0
StockOptionLevel	1470.0	0.793878	0.852077	0.0	0.00	1.0	1.00	3.0
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.00	10.0	15.00	40.0
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.00	3.0	3.00	6.0
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.00	5.0	9.00	40.0
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.00	3.0	7.00	18.0
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.00	1.0	3.00	15.0
YearsWithCurrManager	1470.0	4.123129	3.568136	0.0	2.00	3.0	7.00	17.0

```
In [9]: # Descriptive Analysis of categorical attributes
        df.describe(include='O').T
```

Out[9]:

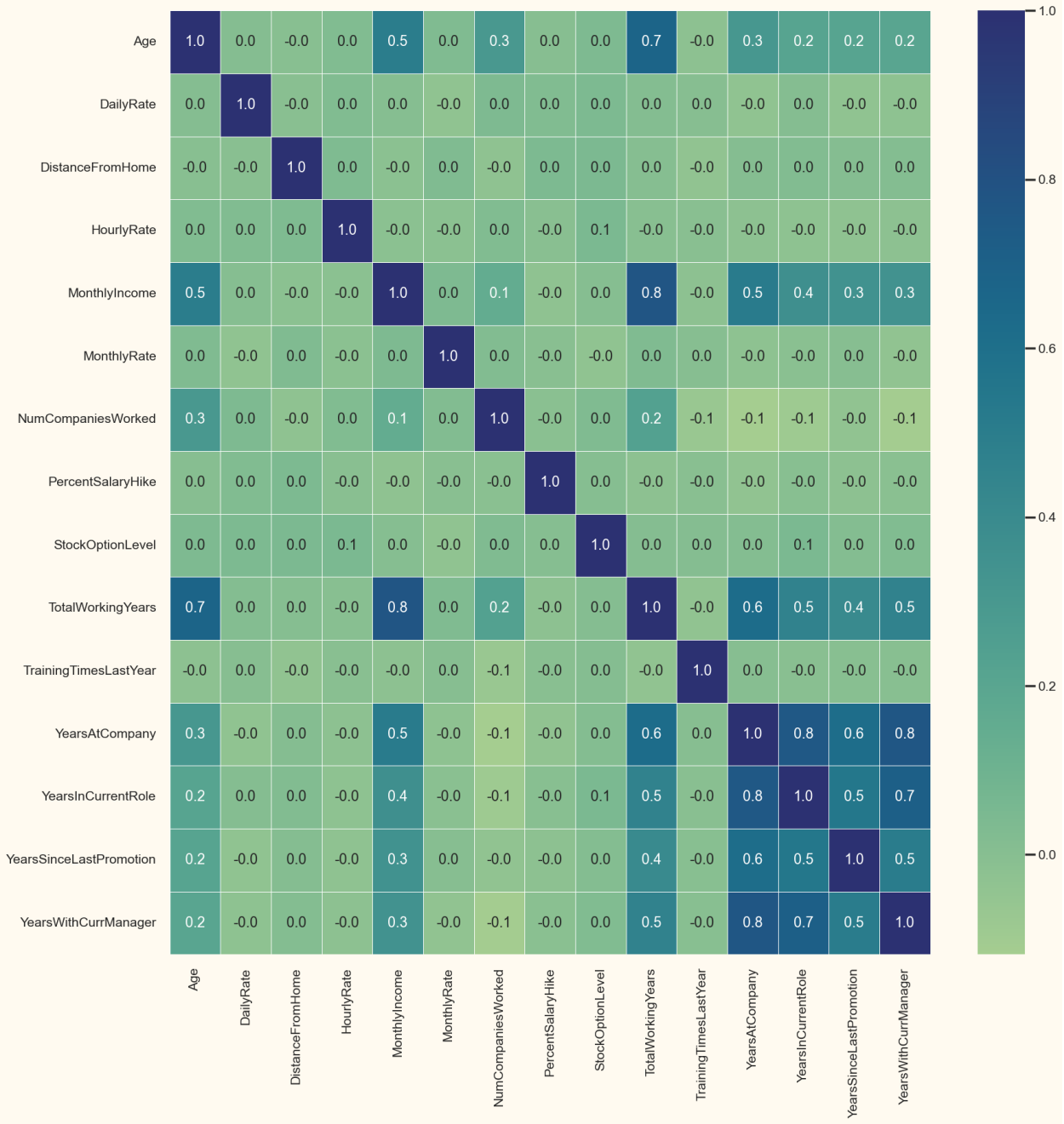
	count	unique	top	freq
Attrition	1470	2	No	1233
BusinessTravel	1470	3	Travel_Rarely	1043
Department	1470	3	Research & Development	961
Education	1470	5	Bachelor	572
EducationField	1470	6	Life Sciences	606
EnvironmentSatisfaction	1470	4	High	453
Gender	1470	2	Male	882
JobInvolvement	1470	4	High	868
JobLevel	1470	5	Entry Level	543
JobRole	1470	9	Sales Executive	326
JobSatisfaction	1470	4	Very High	459
MaritalStatus	1470	3	Married	673
Over18	1470	1	Y	1470
OverTime	1470	2	No	1054
PerformanceRating	1470	2	Excellent	1244
RelationshipSatisfaction	1470	4	High	459
WorkLifeBalance	1470	4	Better	893

Statistical description of all variables

Observation :

1. The Minimum Age is 18 which conveys that all employees are Adult. So there's no need of Over18 Attribute for our analysis.
2. The Stanard Deviation value of EmployeeCount and StandardHours is 0.00 which conveys that All values present in this attribute are same.
3. Attribute EmployeeNumber represents a unique value to each of the employees, which will not provide any meaningful inisghts.
4. Since this Attribute will not provide any meaningful insights in our analysis we can simply drop these attributes

3] HEATMAP OF ALL VARIABLES



Heatmap (Showing correlations of all variables)

Analysis

- Converting target variable into numerical format

```
In [51]: # Convert Target Variable into numerical form.
         label = LabelEncoder()
         df["Attrition"] = label.fit_transform(df.Attrition)
```

- Up sampling the target variable “Attrition”

```
In [55]: from sklearn.utils import resample
         #seperate the case of yes attrition and yes-subscribe
         emp_attr_no = df[df.Attrition == 0]
         emp_attr_yes = df[df.Attrition == 1]

         #Upsample the yes-subs cases
         df_minority_downsampled = resample(emp_attr_yes,replace =True, n_samples = 600)

         #Combine majority class with upsampled minority class
         new_df = pd.concat([emp_attr_no, df_minority_downsampled])
```

- Convert Categorical Variable into numerical form

```
In [52]: # Convert Categorical Variable into numerical form.
         num_df = pd.get_dummies(new_df,drop_first=True)
         num_df.head(5)
```

Out[52]:

	Age	Attrition	DailyRate	DistanceFromHome	HourlyRate	MonthlyIncome	MonthlyRate	NumCompaniesWorked	PercentSalaryHike	StockOptionLevel	...	Ma
1	49	0	279	8	61	5130	24907	1	23	1	...	
3	33	0	1392	3	56	2909	23159	1	11	0	...	
4	27	0	591	2	40	3468	16632	9	12	1	...	
5	32	0	1005	2	79	3068	11864	0	13	0	...	
6	59	0	1324	3	81	2670	9964	4	20	3	...	

5 rows × 61 columns

After performing all these steps, the data is ready for model building. As all the variables are now in numerical form and the target variables is also balanced. The model won’t provide any bias.

DATA MODELING

1] Logistic Regression:

```
TESTING RESULTS:
================================

CONFUSION MATRIX:
[[278  34]
 [ 35 112]]

ACCURACY SCORE:
0.8497

CLASSIFICATION REPORT:
                    0           1  accuracy   macro avg  weighted avg
precision    0.888179    0.767123  0.849673    0.827651      0.849409
recall       0.891026    0.761905  0.849673    0.826465      0.849673
f1-score     0.889600    0.764505  0.849673    0.827053      0.849537
support    312.000000  147.000000  0.849673  459.000000    459.000000
```

2] KNN:

```
TESTING RESULTS:
================================

CONFUSION MATRIX:
[[256  56]
 [ 54  93]]

ACCURACY SCORE:
0.7603

CLASSIFICATION REPORT:
                    0           1  accuracy   macro avg  weighted avg
precision    0.825806    0.624161  0.760349    0.724984      0.761227
recall       0.820513    0.632653  0.760349    0.726583      0.760349
f1-score     0.823151    0.628378  0.760349    0.725765      0.760773
support    312.000000  147.000000  0.760349  459.000000    459.000000
```

3] Decision Trees:

```
TESTING RESULTS:
================================

CONFUSION MATRIX:
[[265  47]
 [ 12 135]]

ACCURACY SCORE:
0.8715

CLASSIFICATION REPORT:
                    0           1  accuracy   macro avg  weighted avg
precision    0.956679    0.741758   0.87146    0.849218      0.887848
recall       0.849359    0.918367   0.87146    0.883863      0.871460
f1-score     0.899830    0.820669   0.87146    0.860249      0.874478
support    312.000000  147.000000   0.87146  459.000000    459.000000
```

4] Random Forest:

```
TESTING RESULTS:
================================

CONFUSION MATRIX:
[[309   3]
 [ 14 133]]

ACCURACY SCORE:
0.9630

CLASSIFICATION REPORT:
                    0           1  accuracy   macro avg  weighted avg
precision    0.956656    0.977941  0.962963    0.967299      0.963473
recall       0.990385    0.904762  0.962963    0.947573      0.962963
f1-score     0.973228    0.939929  0.962963    0.956579      0.962564
support    312.000000  147.000000  0.962963  459.000000    459.000000
```

Conclusion

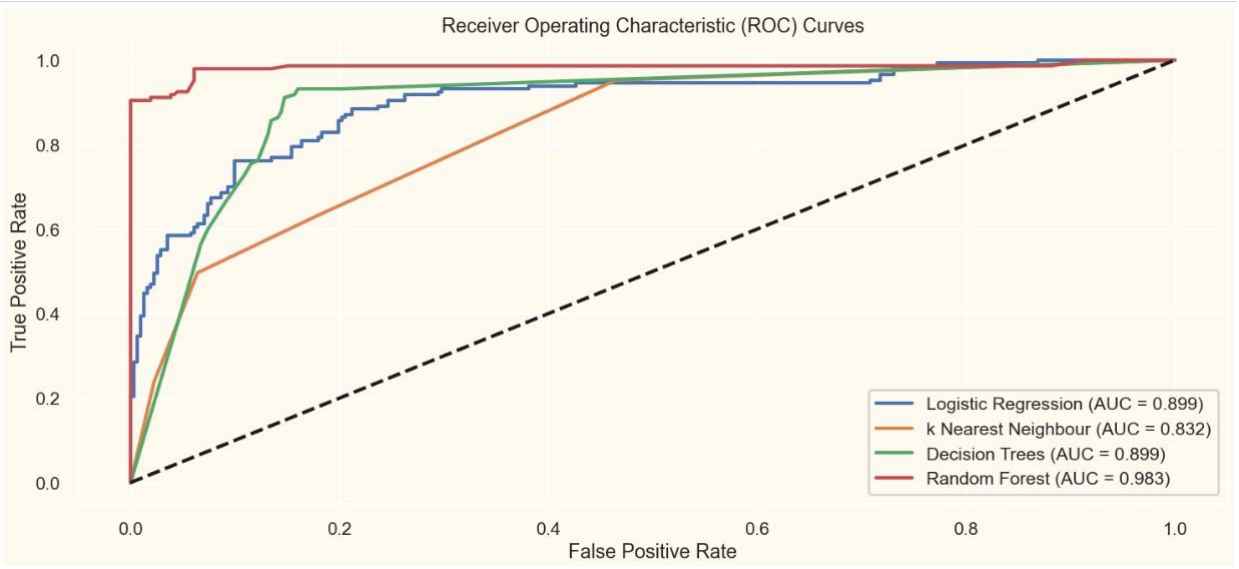
1. **Logistic Regression:**
- Accuracy Score: **0.8497**
- The model achieved a decent accuracy of around 85%.
- Precision, recall, and F1-score for both classes (0 and 1) are also reasonable.
- The model performed consistently well across different evaluation metrics.

2. **K-Nearest Neighbors (KNN):**
- Accuracy Score: **0.7603**
- The model's accuracy is lower compared to logistic regression.
- Precision, recall, and F1-score for class 1 are notably lower than for class 0, indicating potential issues with class imbalance or misclassification.
- The model may benefit from further tuning or feature engineering to improve performance.

3. **Decision Tree:**
- Accuracy Score: **0.8715**
- The decision tree model shows improved accuracy compared to logistic regression and KNN.
- It achieves high precision and recall for both classes, suggesting effective discrimination between positive and negative instances.
- The model demonstrates robust performance and could be considered a viable option for classification tasks.

4. **Random Forest:**
- Accuracy Score: **0.9630**
- The random forest model exhibits the highest accuracy among the tested algorithms.
- It achieves exceptional precision, recall, and F1-score for both classes, indicating strong predictive performance.
- Random forest outperforms other algorithms and appears well-suited for the classification task.

ROC Curve: An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.



- The graph is well-structured and displays multiple lines of varying colors, each representing a different machine learning model.
- The models include Logistic Regression, KNN, Decision Tree and Random Forest.
- Each model is also associated with an Area Under Curve (AUC) value, indicating their performance. The graph presents the True Positive Rate on the y-axis, ranging from 0.0 to 1.0, and the False Positive Rate on the x-axis, also ranging from 0.0 to 1.0.
- Here, model like Random Forest and Logistic regression have better performance comparing with KNN and Decision Tree.

In conclusion, the random forest model shows the highest performance in terms of accuracy and classification metrics, followed by the decision tree. Logistic regression performs reasonably well, while KNN lags slightly in accuracy and class-specific metrics. Therefore, the random forest model seems to be the most valid and effective for this classification task.

Limitations

Limited Data Availability: Small datasets may contain insufficient samples to adequately represent the variability and complexity of attrition patterns within the organization. This limitation restricts the diversity of employee profiles and attrition scenarios captured in the data, potentially leading to biased or unreliable model predictions.

Data Quality: The accuracy and reliability of the predictive models heavily rely on the quality of the data used for training. As the dataset contained an imbalance in the target variable, It can negatively impact the performance of the models and the validity of the predictions.

Feature Engineering: Domain knowledge enables the identification of relevant predictors of attrition beyond the available dataset. HR professionals can provide valuable insights into organizational factors, job-related variables, and employee behaviors that may influence attrition risk. This knowledge guides the selection and creation of meaningful features for improved model performance.

References

[1] Dattatreya, G. R., and Kanal, L. N. (1985). Decision trees in pattern recognition. University of Maryland. Computer Science.

[2] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. San Francisco CA: Morgan Kaufmann.

[3] Neapolitan, R. E. (1989). Probabilistic reasoning in expert systems: theory and algorithms. Wiley.

[4] Hoffman M and Tadelis S 2018 People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis National Bureau of Economic Research

[5] https://github.com/aastha985/Employee_Attrition_Prediction

[7] N. Mansor, N. S. Sani, and M. Aliff, “Machine Learning for Predicting Employee Attrition,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, pp. 435–445, 2021, doi: 10.14569/ IJACSA.2021.0121149