Employee Attrition Prediction

IBM HR ANALYTICS EMPLOYEE ATTRITION & PERFORMANCE

Namrata Gulati | 9/5/24

Contents

Introduction:

Employee attrition, or the phenomenon of employees leaving the organization, can have a significant impact on the business. Some of its implications include decreased productivity, loss of institutional knowledge, and increased costs associated with hiring and training new employees.

This project will let us interpret the factors that contribute to employee attrition and accordingly develop strategies to mitigate it. By the end of this report, one can predict employee attrition using ML techniques. By analyzing the features of the given dataset, I have built a predictive model that can identify employees who are at risk of leaving the organization.

The primary **objectives** of this project are as follows:

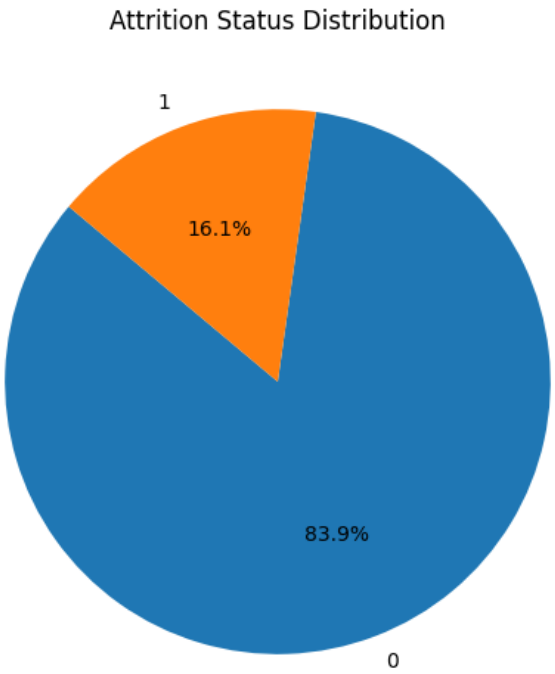
- **Data Exploration**: To know about the distribution of features, identify any patterns and understand the prevalence of attrition.
- **Model Development**: Using machine learning algorithms I'll develop predictive models to classify employees as either likely to churn or likely to stay.
- **Evaluation and Optimization**: We will evaluate the performance of each model using appropriate metrics such as accuracy, precision, recall, and F1-score. This will be followed by optimizing model parameters to improve predictive accuracy.
- **Insights and Recommendations**: Finally, based on the findings, certain insights and recommendations have been included to better understand the drivers of employee attrition and devise effective retention strategies.

Dataset Analysis:

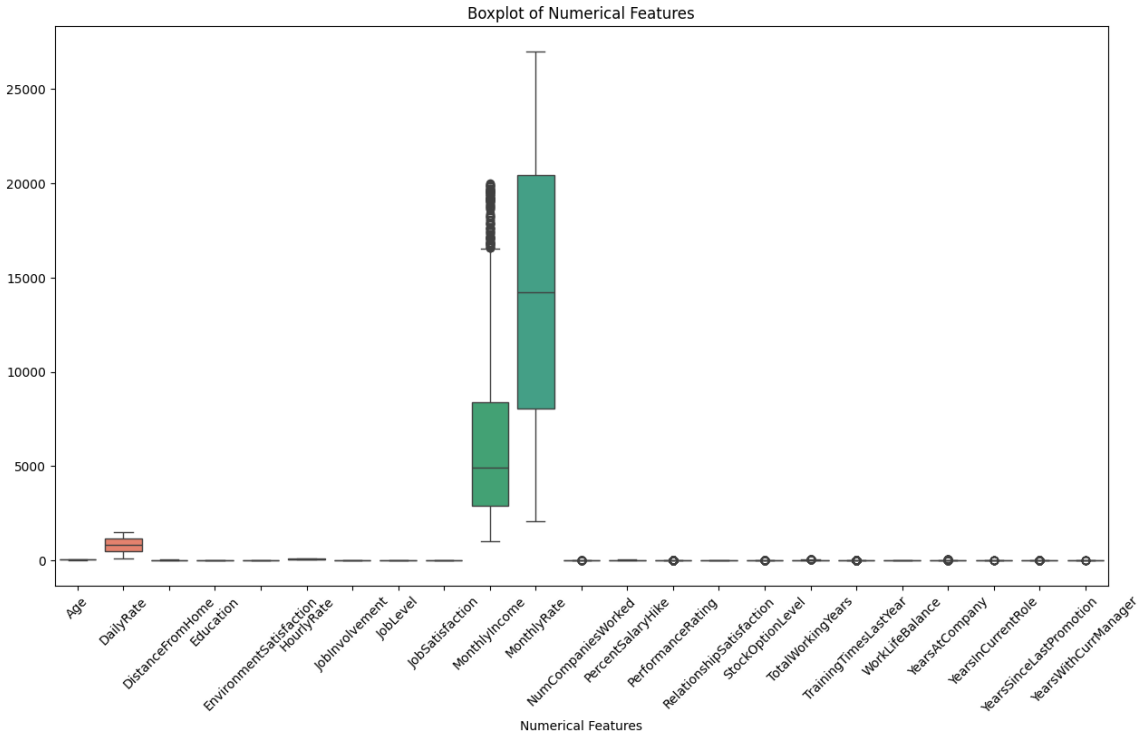
This project focuses on IBM HR Analytics Employee Attrition & Performance dataset that has been downloaded from Kaggle.

1. The given dataset has 1470 rows and 31 columns.
2. The target variable, "Attrition," indicates whether an employee has left the organization (Yes) or is still employed (No).
3. There are 26 numerical features and 9 categorical features.
4. Numerical Features: These include attributes like age, daily rate, monthly income, years at the company, etc., which are represented by numerical values.
5. Categorical Features: Attributes such as business travel frequency, department, job role, marital status, etc., are categorical in nature.
6. There are 0 missing values in the entire dataset.
7. describe() has been used to find out the count, mean and other mathematical operations.
8. Using tabulate library of python, a table detailing the unique values of each feature has been created.
9. Attrition Status Distribution:

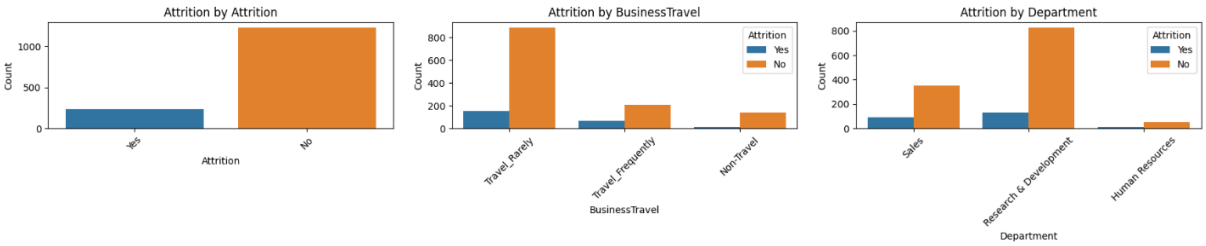
   Using value_counts(), attrition as Yes or No has been depicted in the form of a pie chart. 83.9% have not left the company while 16.1% represents the employees who have left the organization. Another noteworthy thing is that the target column is imbalanced. This issue will be handled at a later stage.



10. Box Plot of Numerical Features
    This has been curated to analyze the data of numerical features and understand their range, median and most importantly, outliers. The 'MonthlyIncome' feature has outliers that will be removed through log transformation. For other features, the outliers don't exist and all values exist below 2500.

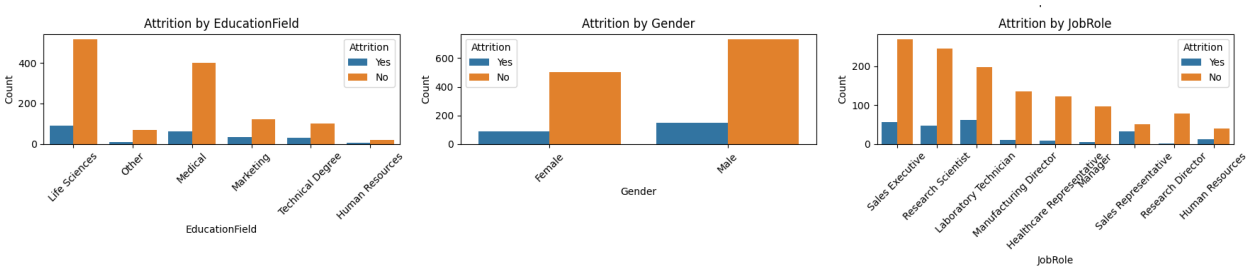


11. Relationship between 'Attrition' and categorical features
    This relationship has been mapped using bar plots made with a "for" loop that iterates over all features.

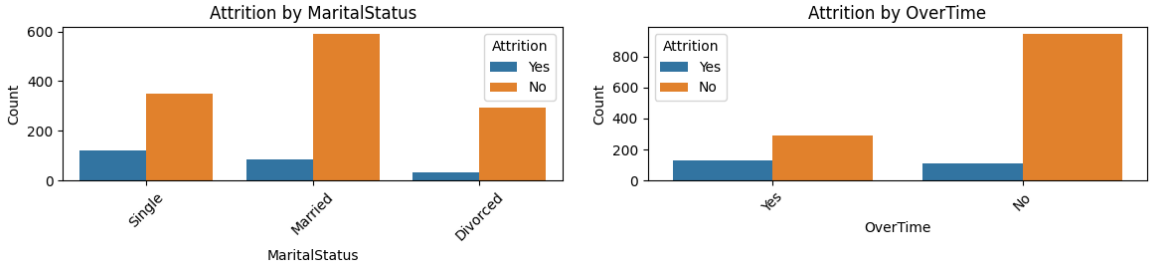


- Attrition by BusinessTravel states that people who travel frequently may engage in organization switch depending on their location.
- The Attrition by Department chart suggests that R&D Department provides a sense of stability to employees. This comparatively lower

attrition rate suggests greater job satisfaction, a positive work environment, and rewarding career opportunities within the R&D team. With respect to the Yes and No division, sales department depicts comparatively high attrition owing to high-pressure sales targets, job stress, or dissatisfaction with compensation structures.



- Attrition by EducationField shows that Life Sciences and Medical fields demonstrate the lowest attrition rates. These fields often offer specialized roles that require specific expertise, leading to higher job satisfaction and lower turnover. Also, HR professionals may seek opportunities in other organizations to gain broader experience or pursue career advancement. Technical roles may involve continuous learning and skill development to keep pace with industry trends. Additionally, employees with technical backgrounds may have diverse career opportunities in various industries, leading to higher mobility and turnover.
- Attrition by Gender depicts a higher attrition rates among males compared to females. This can be due to several factors like unequal opportunities for career advancement, pay disparities, etc. Males may be more inclined to seek new opportunities, career changes, or job transitions compared to females, leading to higher turnover.
- Attrition by JobRole suggests that employees in leadership or specialized roles are more likely to remain with the company over time. Possible reasons for this lower attrition include greater job satisfaction, competitive compensation, opportunities for career advancement, etc. Sales Representatives, Laboratory Technicians, Human Resources Personnel may be more prone to turnover due to factors such as job dissatisfaction, limited growth opportunities, lower salaries, or higher workloads.



12. Correlation check with numerical features

    Five Features namely 'MonthlyIncome', 'TotalWorkingYears', 'JobLevel', 'YearsInCurrentRole', 'Age' have been used to check how they are related to the target variable i.e. Attrition. These numerical features have been selected based on top correlation.

    KDE plots provide a smooth estimate of the distribution of each variable, making it easier to identify patterns or differences. Due to presence of hue parameter, **Blue dots** represent employees who have not experienced attrition ('No' in the 'Attrition' column). **Orange dots** represent employees who have experienced attrition ('Yes' in the 'Attrition' column). With respect to Attrition, more blue dots indicate a class imbalance in the dataset, which means that the model may have more examples of “No” class than the “Yes”, potentially leading to biased predictions.

    YearsinCurrentRole vs MonthlyIncome & TotalWorkingYears suggest an extremely weak correlation as these points are spread out.

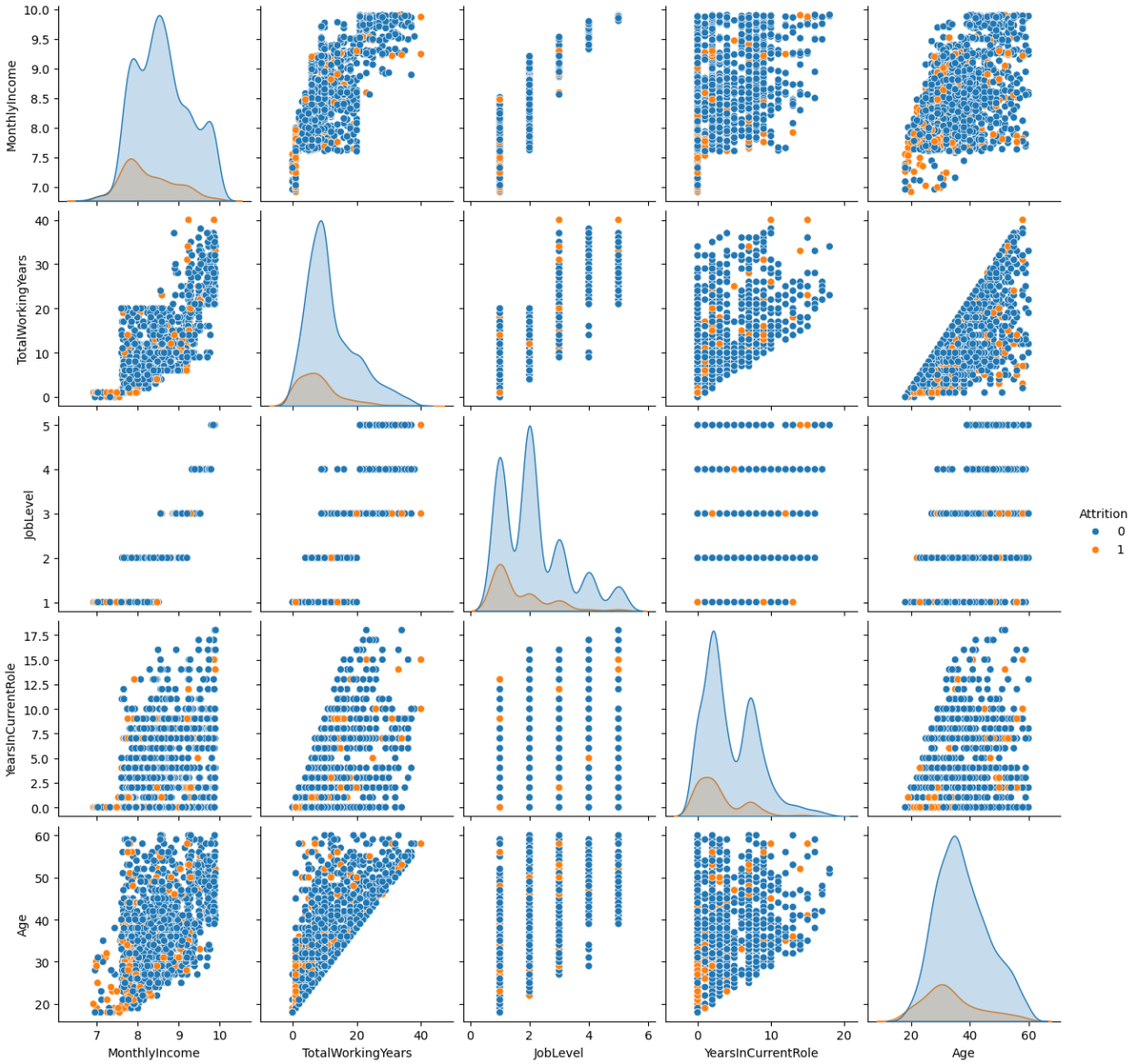
    Peaks in Density Plots along diagonal imply regions where more data points exist i.e. a peak corresponds to the mode of the distribution. So, for example, JobLevels has 5 peaks corresponding to the 5 different classes.

    MonthlyIncome vs. Age clearly shows that people in the lower age group are more prone to switching the organization due to desire of better salary and other factors. However, as age increases, people desire for stability thus we can observe less attrition there. As employees age and gain more experience, they may prioritize stability and career growth over immediate financial gains, leading to lower attrition rates.

    TotalWorkingYears vs. Age shows that there is a cluster of younger employees with relatively high total working years, indicating early career mobility or job-hopping behavior.

    JobLevel vs. Age seems to be constant for all job levels.

YearsInCurrentRole vs. Age: The scatter plot suggests that younger employees are more likely to change roles frequently, possibly seeking career advancement or exploring different job opportunities. Conversely, older employees tend to stay in their current roles for longer durations, indicating a greater sense of stability and job satisfaction.



Preprocessing Steps:

1. Python Libraries like matplotlib, scipy, wordcloud, pandas, sklearn, mlxtend and others.
2. 'EmployeeCount', 'EmployeeNumber', 'StandardHours', 'Over18' columns have been dropped initially.
3. For the purpose of transformation, a copy of original dataframe has been created as df1.
4. The 'Attrition' column, which has already been encoded as 0 & 1, is excluded from the numerical features that will undergo transformation.
5. A StandardScaler object is initialized to scale the numerical features. A list of LabelEncoder objects is created one for each categorical feature. LabelEncoder is used to convert categorical labels into numerical values.
6. A ColumnTransformer named preprocessor, which applies transformations to the numerical and categorical features separately, is used. For numerical features, it applies the numerical_transformer, which is the StandardScaler object created earlier. For categorical features, it uses the 'passthrough' option, which means it leaves them unchanged because they have already been encoded using LabelEncoder.
7. This results in 30 columns (31ST column is Attrition). With Label Encoding, original columns stay intact. All columns have their dtype as "**float64".**
8. This is the coding that is being followed:

Mapping for feature 'BusinessTravel':
Non-Travel: 0
Travel_Frequently: 1
Travel_Rarely: 2
Mapping for feature 'Department':
Human Resources: 0
Research & Development: 1
Sales: 2
Mapping for feature 'EducationField':
Human Resources: 0

Life Sciences: 1
Marketing: 2
Medical: 3
Other: 4
Technical Degree: 5
Mapping for feature 'Gender':
Female: 0
Male: 1
Mapping for feature 'JobRole':
Healthcare Representative: 0
Human Resources: 1
Laboratory Technician: 2
Manager: 3
Manufacturing Director: 4
Research Director: 5
Research Scientist: 6
Sales Executive: 7
Sales Representative: 8
Mapping for feature 'MaritalStatus':
Divorced: 0
Married: 1
Single: 2
Mapping for feature 'OverTime':
No: 0
Yes: 1

9. Correlation after normalization is calculated with threshold as 0.7 to check which features are highly correlated and will contribute more in model development. These include:
    - 'MonthlyIncome' and 'JobLevel'
    - 'JobSatisfaction' and 'TotalWorkingYears'
    - 'MonthlyIncome' and 'TotalWorkingYears'
    - 'YearsAtCompany' and 'YearsInCurrentRole',
    - 'YearsWithCurrManager' and 'YearsAtCompany'
    - 'YearsWithCurrManager' and 'YearsInCurrentRole'
10. Handling Imbalanced Data (of target variable) using SMOTE (Synthetic Minority Over-sampling Technique)

    For Attrition target column:

    0 1233

    1 237

This depicts that the data is biased towards class 0 i.e. Attrition as NO. SMOTE (Synthetic Minority Over-sampling Technique) is being used here to address the issue of class imbalance in the target variable 'Attrition'. SMOTE is applied to the feature matrix X and the target vector y to generate synthetic samples for the minority class ('Attrition' = 1). After applying SMOTE, the class proportions are printed to show that the class distribution is now balanced, which can help improve the performance of machine learning models trained on this data.

Class Proportions after SMOTE:

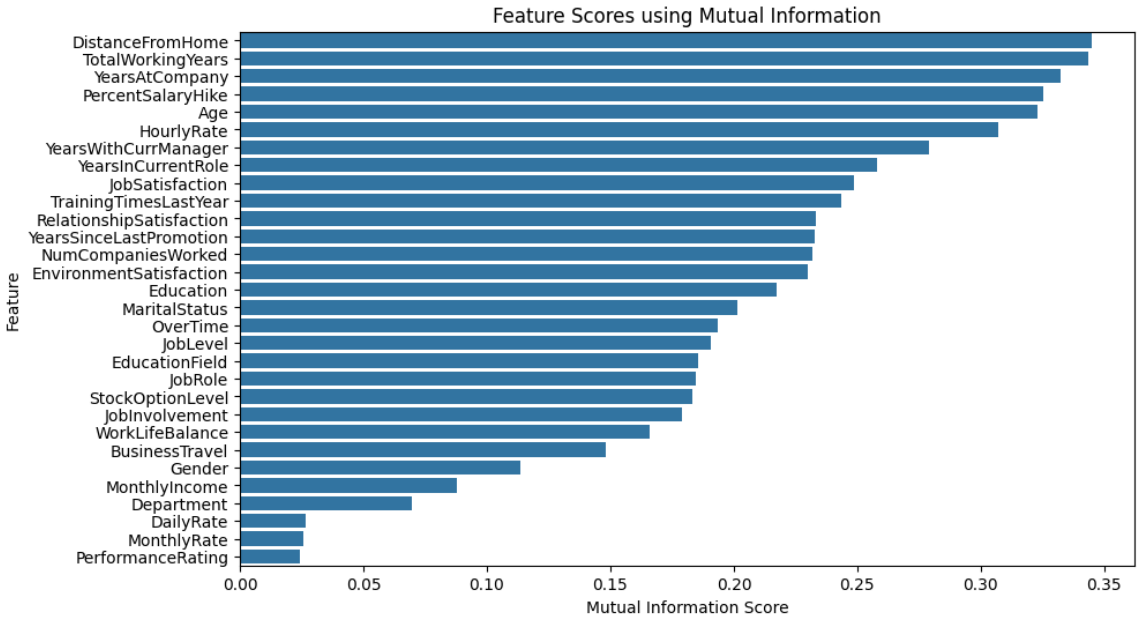
Attrition

1 0.5

0 0.5

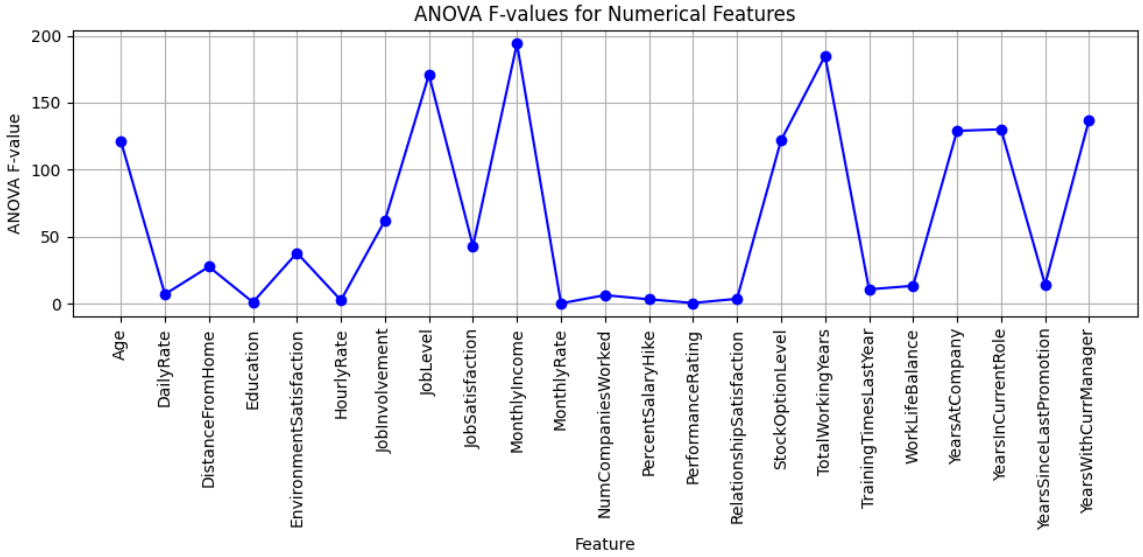
11. Feature Scores using Mutual Information

Mutual Information measures the dependency between two variables, indicating how much information about one variable can be obtained by observing the other variable. In this case, it helps determine the relevance of each feature to the target variable- Attrition. These scores have been arranged in descending order with the help of the bar graph. The ones with score ~ 0 will not contribute much to the prediction task.



12. Anova test for numerical variables

For each numerical feature, the ANOVA F-value is calculated using the f_oneway function. This function performs a one-way ANOVA test to determine whether there are significant differences in the means of the feature across different levels of the target variable (attrition in this case). Features with higher F-values indicate larger differences in means across different levels of attrition, suggesting they may have more influence on attrition. Peaks in the line chart relate to the numerical features where the differences in means across different levels of attrition are particularly pronounced. Conversely, dips (where F value ~ 0) suggest features where the differences are less significant. Thus, the less significant features can be easily dropped.



13. Thus 'Gender', 'BusinessTravel', 'EducationField', 'Department','JobInvolvement','WorkLifeBalance', 'NumCompaniesWorked','MaritalStatus', 'OverTime', 'JobRole', 'DailyRate', 'HourlyRate', 'PerformanceRating', 'MonthlyRate' have been dropped keeping in mind both Mutual information and Anova F value scores.

Model Development & Evaluation:

1. For predicting employee attrition, six different models were chosen: **Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, XGBoost, and K-Nearest Neighbors (KNN).**
2. Evaluation Metrics: Each model's performance was assessed using several evaluation metrics:
    a. Cross-validation Score: The mean accuracy obtained through cross-validation.
    b. Accuracy: The proportion of correctly predicted instances.
    c. Precision: The ratio of correctly predicted positive observations to the total predicted positives.
    d. Recall: The ratio of correctly predicted positive observations to the all observations in actual class.
    e. F1 Score: The weighted average of precision and recall, providing a single metric to assess a model's performance.
3. Result:
    a. Random Forest and XGBoost achieved the highest accuracy, precision, recall, and F1 score among the models evaluated, indicating their effectiveness in predicting employee attrition.
    b. Decision Tree and Logistic Regression also performed reasonably well, with competitive scores across all metrics.
    c. SVM and KNN exhibited lower accuracy and precision compared to other models, suggesting they may not be the best choices for this task.
4. Thus, Random Forest is my preferred choices for predicting employee attrition due to its performance across multiple evaluation metrics.

Optimization Techniques:

a. Hyperparameter tuning
b. Bagging & AdaBoost Classifier ensemble methods
    - Further hyperparameter tuning and ensemble methods, specifically Bagging and AdaBoost Classifier have been utilized to improve the predictive performance of a Random Forest base classifier for predicting employee attrition.
    - To perform search over specified parameter values and cross-validate each combination using 5-fold cross-validation, grid search cv has been used here.
    - BaggingClassifier is instantiated with the best Random Forest classifier as the base estimator and 10 estimators.
    - AdaBoostClassifier is instantiated with the best Random Forest classifier as the base estimator, 50 estimators, and a learning rate of 1.0.
    - The following results have been obtained:
      Bagging Classifier Accuracy: 0.8724696356275303
      AdaBoost Classifier Accuracy: 0.9089068825910931

    Thus, Ensemble method: AdaBoost has further enhanced predictive accuracy by combining multiple base estimators.

c. Feed Forward Neural Network with KerasClassifier
    Hyper parameter tuning using GridSearchCV with a KerasClassifier built using the Functional API has been covered next. KerasClassifier works for Binary classification tasks like predicting employee attrition.
    Structure of FFNN:
        - The first layer defined in the build_model function is the input layer, which receives the features of the input data.
        - Two hidden layers are defined with ReLU activation functions ('relu'). These layers perform transformations on the input data.
        - The final layer is the output layer, which produces the model's predictions. It consists of a single neuron with a sigmoid activation function ('sigmoid'),

which outputs values in the range [0, 1] representing the probability of the positive class in a binary classification task.

- Also, along with Adam optimizer and binary cross-entropy loss, accuracy is used as the evaluation metric.

The output is:

Best: 1.000000 using {'batch_size': 32, 'epochs': 10}

Thus, we obtain a 100% accuracy with a Keras Classifier on this FFNN. FFNN turns out to be the most suitable for predicting employee attrition.

Insights and Recommendations:

A. INSIGHTS

- Attrition by BusinessTravel suggests that frequent travelers may be more prone to changing organizations based on their location, potentially impacting turnover rates.
- The Attrition by Department chart reveals that the Research and Development (R&D) Department fosters stability, with lower turnover rates attributed to positive work environments and rewarding career opportunities. Conversely, the Sales Department experiences higher attrition due to factors like high-pressure sales targets.
- Attrition by EducationField highlights variations in turnover rates across different fields, with Life Sciences and Medical fields showing the lowest attrition. Conversely, HR and technical roles may see higher turnover due to career advancement or skill development opportunities elsewhere.
- Attrition by Gender indicates higher turnover among males, potentially influenced by unequal career opportunities or differing aspirations.
- Attrition by JobRole suggests that leadership or specialized roles have lower turnover, while roles like Sales Representatives or Laboratory Technicians may experience higher attrition due to job dissatisfaction or limited growth opportunities.
- Correlation Check with Numerical Features explores relationships between key factors like MonthlyIncome and Age, revealing potential trends such as higher attrition among younger employees seeking better opportunities. Also, Attrition vs Numerical columns display less correlation. This means that certain features are not dependent on target variable.

B. CHALLENGES ENCOUNTERED

- Data Quality: Ensuring the accuracy and completeness of the data used for analysis posed a challenge, particularly when dealing with large datasets or missing values. Robust data cleaning and preprocessing techniques were employed to address these challenges.
- Interpretation Complexity: Analyzing complex relationships and patterns within the data required careful consideration and domain knowledge. Understanding the nuances of employee behavior, organizational dynamics, and external factors influencing attrition was essential for deriving meaningful insights.
- Addressing Bias: Avoiding bias in data analysis and interpretation was crucial to ensure the fairness and reliability of the findings. Steps were taken to mitigate bias in data collection, preprocessing, and analysis to enhance the validity of the results.

- Model Interpretability: While machine learning models offer predictive power, interpreting the underlying factors driving predictions can be challenging. Ensuring model interpretability is essential for understanding the reasons behind attrition and deriving actionable insights for organizational decision-making.

C. RECOMMENDATIONS FOR REDUCING EMPLOYEE ATTRITION

- Enhance Employee Engagement: Implement initiatives to improve employee engagement, satisfaction, and overall well-being. Conduct regular employee surveys, feedback sessions, and performance evaluations to understand and address underlying concerns.
- Provide Growth Opportunities: Offer opportunities for career development, training, and advancement to employees at all levels. Establish clear career paths, mentorship programs, and skill development initiatives to encourage professional growth and retention.
- Promote Work-Life Balance: Foster a supportive work environment that prioritizes employee well-being and work-life balance. Implement flexible work arrangements, wellness programs, and stress management initiatives to support employee health and reduce burnout.
- Recognize and Reward Performance: Recognize and reward employee contributions and achievements to reinforce positive behavior and performance. Implement performance-based incentives, bonuses, and recognition programs to motivate employees and increase job satisfaction.
- Enhance Organizational Culture: Cultivate a positive organizational culture that values diversity, inclusion, and collaboration. Foster open communication, trust, and transparency to build strong relationships between employees and leadership.
- Monitor and Evaluate Interventions: Continuously monitor employee turnover rates, satisfaction levels, and retention metrics to evaluate the effectiveness of implemented interventions. Adjust strategies based on feedback and evolving organizational needs to maintain a supportive and engaging work environment.