**Objectives**

The overarching goal of this project was to employ the HR Analytics data set to uncover the determinants of employee attrition within a company. Through meticulous data processing, feature engineering, and the application of multiple predictive modeling techniques, the project aimed to:

***Predict Employee Attrition:*** Utilize advanced machine learning algorithms, including Logistic Regression, Random Forest, and XGBoost, to create robust models capable of predicting the likelihood of employee turnover. This objective was pursued with the intention of enabling preemptive interventions for at-risk employees.

***Identify Key Attrition Factors:*** Through exploratory data analysis, correlation matrix inspection, and the interpretation of model feature importances, pinpoint the most influential factors contributing to employee attrition. Special attention was given to variables such as OverTime, JobLevel, TotalWorkingYears, and StockOptionLevel, which were hypothesized and later confirmed to have significant impacts on attrition rates.

***Optimize Model Performance:*** Methodically tune hyperparameters for each model to optimize their performance. This included experimenting with different split ratios to balance training and testing data, regularizing logistic regression to prevent overfitting, and adjusting parameters for the Random Forest and XGBoost models to enhance their predictive accuracy and generalization to unseen data.

***Cross-Model Comparisons:*** Compare the effectiveness of different modeling approaches in predicting attrition, utilizing metrics such as accuracy, precision, recall, and the F1 score. This comparison was aimed at identifying the most suitable model or models for deployment in real-world HR analytics applications.

***Provide Actionable Insights:*** Translate the analytical findings into actionable recommendations for HR practices, including targeted retention strategies, tailored employee engagement programs, and personalized career development plans for employees identified as high risk for attrition.

***Future-Proofing HR Strategies:*** Leverage the insights gained from the analysis to future-proof HR strategies against evolving workforce dynamics. This involves suggesting agile, data-driven approaches to talent management that can adapt to changing employee expectations and organizational needs.

By fulfilling these objectives, the project contributes valuable insights into the mechanics of employee turnover, offering a data-driven foundation upon which companies can build more cohesive, enduring, and satisfying work environments for their employees.

**Data Preparation**

***Initial Assessment:*** The dataset was first examined for completeness, identifying any missing values or anomalies. It was found that the 'YearsWithCurrManager' column had some missing values, which was handled by replacing the null values with the value of ‘YearSinceLastPromotion’, this when not ideal provides a reasonable estimation for the field ‘YearsWithCurrManager’.

A screenshot of a computer

Description automatically generated

***Preliminary Feature Selection:*** The dataset includes a wide range of features, some of which are highly relevant to the analysis, while others may not contribute significantly to the outcomes of interest. Features were evaluated for their importance and relevance to the analysis goals. Unnecessary columns were identified and removed to streamline the dataset for analysis.

|  |  |
| --- | --- |
| **Feature** | **Reason for removal** |
| YearsInCurrentRole | Same as YearsSinceLastPromotion |
| Age | AgeGroup will be considered |
| MonthlyIncome | HourlyRate, StockOptionLevel is already considered |
| SalarySlab | HourlyRate, StockOptionLevel is already considered |
| DailyRate | HourlyRate, StockOptionLevel is already considered |
| MonthlyIncome | HourlyRate, StockOptionLevel is already considered |
| StandardHours | Constant for all entries |
| EmployeeCount | Constant for all entries |
| EmployeeNumber | Not significant for analysis |
| Over18 | Constant for all entries |
| EmpID | Not significant for analysis |

***Data Transformation and Encoding:*** Several columns in the dataset contain categorical data that need to be encoded into a numerical format suitable for machine learning algorithms. For instance, 'BusinessTravel', ‘AgeGroup’ are transformed using label encoding, considering their ordinal nature while 'OverTime', 'Attrition’, ‘Gender’, ‘Department’, ‘EducationField’, ‘JobRole’, ‘MaritalStatus’ were encoded one-hot encoding.

***Feature Engineering:*** 'TotalWorkingYears' was binned into categories such as '0-5', '6-10',... to simplify its distribution and enhance its predictive power regarding attrition.

***Normalization and Scaling:*** All Features were scaled to ensure that no single feature dominates the model due to its scale since models like logistic regression can be sensitive to the scale of the data MinMaxScaler() was used to ensure all datapoints was within 0-1.

***Correlation Matrix Analysis***

The correlation matrix served as an incipient tool for visualizing the linear relationships among the numerical variables within the HR dataset. Through the application of the corr() function to the processed data, a comprehensive matrix was generated, encapsulating the correlations between variables.

***Visualization and Interpretation***

The visualization of this matrix was achieved using a heatmap, facilitated by the seaborn library. This color-coded heatmap not only enhanced interpretability but also allowed for the immediate identification of variables that exhibit strong positive or negative correlations. Variables closely correlated with each other were highlighted, indicating potential predictors of attrition that warrant further examination.

A graph with red and blue squares

Description automatically generated

***Chi-squared Test Application***

Concurrently, Chi-squared tests were conducted to evaluate the association between categorical variables and employee attrition. This statistical test is crucial for identifying significant differences in the distribution of categorical variables, thereby shedding light on potential factors influencing attrition.

***Visualization and Interpretation***

The analysis involved creating contingency tables for categorical variables against 'Attrition\_encoded' and calculating the Chi-squared statistic and p-values. The findings were systematically organized, enabling a clear distinction between variables based on their level of significance in relation to attrition. The visualization of these results through a color-coded bar chart further facilitated an intuitive understanding of the statistical significance of these relationships.

A graph of a bar graph

Description automatically generated with medium confidence