SEP-787: Machine Learning – Classification Model

Final Project

On

IBM HR Analytics Employee Attrition & Performance

By

Abhishek Gambhir (400546232)

Snigdha Pandey (400547068)

1. Introduction

Objective: The project aims to analyze and predict employee attrition using decision tree-based models[1]. Employee attrition is the voluntary or involuntary reduction of staff members, which can have negative impacts on the organization's performance and reputation.

Significance: Understanding attrition is crucial for maintaining a productive workforce and economic stability. By identifying the factors that influence attrition and the employees who are most likely to leave, HR managers can design effective retention strategies and improve employee satisfaction and loyalty.

2. Problem Statement

i. **Identification of Challenge:** The project addresses the pressing challenge of employee attrition, seeking to employ predictive modelling techniques on the IBM HR Analytics dataset.

ii. **Defined Goal:** The primary objective is to develop a robust and accurate predictive model for employee attrition, with a specific focus on leveraging Decision Tree-based models.

iii. **Data Exploration Phase:** The project initiates a comprehensive Exploratory Data Analysis (EDA) using the Kaggle-sourced employee attrition dataset. This phase aims to summarize critical features, identify patterns, and analyze factors influencing employee attrition, laying the groundwork for subsequent decision-making and feature engineering.

iv. **Formulation of Central Question:** To tackle the challenge of employee attrition within the IBM HR Analytics dataset, it is essential to compare the efficacy of different models best suited for binary classification problem such as Logistic Regression, SVM, Random Forest, and Gradient Boosting. By doing so, we can better understand which model works best and how it can help resolve the issue. This investigation will serve as a crucial step in guiding subsequent phases and enable us to assess these models directly and comparatively for optimal results.

v. **Model Creation and Evaluation:** Given the binary classification nature of the problem, the project involves the creation and evaluation of Logistic Regression, SVM, Random Forest, and Gradient Boosting models. Selecting the most suitable model hinges on rigorous evaluation metrics such as accuracy, precision, recall, and roc-auc scores. Hyper parameters will be optimized through grid search and cross-validation, ensuring the chosen model aligns with the unique characteristics of the dataset and enhances predictive accuracy for employee attrition.

In summary, this project aims to address the pressing challenge of employee attrition by developing a robust and accurate predictive model. The project comprises a comprehensive data exploration phase, formulating a central question, and creating and evaluating predictive models using rigorous evaluation metrics to enhance predictive accuracy for employee attrition.

3. Dataset Overview

i. **Source:** The dataset originates from IBM and is sourced from Kaggle (Employee attrition via Ensemble tree-based methods | Kaggle).

ii. **Size:** The dataset comprises 1470 records and 35 columns, making it a substantial dataset.

iii. **Target Variable:** 'Attrition' is the target variable.

iv. **Categorical Columns:** Out of the 35 columns, 8 are categorical, including attributes such as 'Attrition,' 'BusinessTravel,' 'Department,' 'EducationField,' 'Gender,' 'JobRole,' 'MaritalStatus,' 'Over18,' and 'OverTime.'

v. **Non-Categorical Columns:** The dataset features 26 non-categorical columns, including essential parameters such as 'Age,' 'DailyRate,' 'DistanceFromHome,' 'Education,' 'EmployeeNumber,' 'JobInvolvement,' 'JobLevel,' 'MonthlyIncome,' and others.

vi. **Data Integrity:** No null and duplicate values found in dataset.

vii. **Constant Columns:** Three columns, namely 'EmployeeCount,' 'Over18,' and 'StandardHours,' exhibit constant values across all records.

4. Approach

Our approach involves the following stages:

1. **Libraries and Dataset**:
   a. Imported all necessary Python libraries such as numpy, pandas, seaborn etc.
   b. Imported the dataset csv file named 'WA_Fn-UseC_-HR-Employee-Attrition'.
   c. Printed the first few records to check if the dataset had been successfully loaded.
2. **Exploratory Data Analysis:**
   a. Checked the number of records and columns in the dataset. Also known as the shape of the dataset.
   b. Checked for categorical and non-categorical columns, as well as the number of columns in each category.
   c. Plotted bar graph for categorical features where the X-axis represents the unique value in the column, the Y-axis represents the count, and the bar represents the relation with attrition rate.
   d. Insights received through the visualization of these graph:
      i. **Education:** There doesn't appear to be a strong correlation between education level and attrition rate. Attrition is present across all education levels, but it does not significantly vary as education level changes.
      ii. **Job Role:** Certain job roles, like Sales Representative and Laboratory Technician, show a higher attrition rate compared to others like Manager or Research Director. This could indicate that roles with possibly lower entry barriers or lower levels of specialization might have higher turnover.
      iii. **Marital Status:** Single employees tend to have a higher attrition rate compared to married or divorced employees. This might be due to different life stages or commitments.

iv. **OverTime:** Employees who work overtime show a significantly higher attrition rate. This could point towards work-life balance being a crucial factor in employee retention.

v. **WorkLifeBalance:** Interestingly, employees who rated their work-life balance as low (1) show a higher attrition rate. This aligns with the observation regarding overtime and suggests that maintaining a healthy work-life balance is vital for employee retention.

e. Plotted a histogram using seaborn to draw a relation between Age i.e. a non-categorical column against Attrition (Target variable). The age distribution in relation to attrition reveals:

i. **Younger Employees and Attrition:** There's a noticeable trend where younger employees (especially in their late 20s to early 30s) show higher attrition rates compared to older age groups. This could be attributed to younger employees being in the early stages of their careers and more likely to change jobs for better opportunities or career advancement.

ii. **Decrease in Attrition with Age:** As age increases, the attrition rate seems to decrease. Employees in their mid-30s and onwards show lower attrition rates. This might be due to increased job stability, higher positions, or more responsibilities both in their personal and professional lives.

iii. **Peaks and Troughs:** Certain age groups, like mid-20s and early 30s, have distinct peaks in attrition, suggesting that these age ranges might be critical points for employee turnover.

f. Plotted box plot for Monthly Income and Total Working Years by attrition status. The box plots for Monthly Income and Total Working Years in relation to attrition reveal the following insights:

i. **Monthly Income:** There's a noticeable difference in the distribution of monthly income between employees who left (Attrition = Yes) and those who stayed (Attrition = No). Employees who left generally had lower median monthly incomes compared to those who stayed. This could suggest that

compensation is a significant factor in an employee's decision to stay with or leave a company. The wider range and higher outliers in the monthly income for employees who stayed might indicate more opportunities for financial growth within the company, which could contribute to higher retention.

ii. **Total Working Years:** Employees who left the company tend to have fewer total working years compared to those who stayed. This aligns with the earlier observation about younger employees being more likely to leave. The more extensive range of working years among employees who stayed suggests a correlation between longer tenure at the company and lower attrition rates.

g. Created a temporary new data frame consisting of columns without any constant value and filtered further to get columns with numerical variables to create a correlation matrix instance and use it to create a heat map that provides insight into how different numerical variables in the dataset are related.

i. **High Correlation Pairs:** Some variables show a high degree of correlation. For example, TotalWorkingYears is highly correlated with Age, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager, and YearsSinceLastPromotion. This is expected as these variables are all related to the length of an employee's career and tenure at the company.

ii. **Moderate Correlations:** There are also moderate correlations between some variables, such as between MonthlyIncome and TotalWorkingYears, indicating that employees with more years of working experience tend to have higher salaries.

iii. **Low Correlation with Attrition:** Interestingly, most variables have a relatively low correlation with Attrition, suggesting that no single factor strongly predicts whether an employee will leave. However, Age, TotalWorkingYears, and YearsWithCurrManager have a somewhat more negative correlation with attrition, implying that more experienced and older employees, or those with a longer tenure with their current manager, are less likely to leave.

iv. **Job Level and Income:** JobLevel is strongly correlated with MonthlyIncome, which is expected as higher job positions usually come with higher pay.

3. Data Cleaning

a. After gathering insights from the data, the next step is to clean the data and make it ready for the model.

b. The first step involves checking for any missing values and duplicate records. We discovered that our data had no missing values or duplicate records that could have hampered our model prediction.

c. The next step includes outlier detection and, if they exist, seeing their impact on our data and handling them. For outlier detecting we choose IQR method because it help you find unusual patterns in data sets with a binary target variable. This method is strong against outliers and doesn't require any assumptions about the distribution of the data. It is effective for a quick initial assessment, providing insights into deviations from the norm and helping to understand factors that influence the binary target variable.

d. After applying the Interquartile Range (IQR) method we found that the 10 columns have the outliers and to identify potential outliers in each column, we undertook a further step to confirm these findings by analyzing strong correlations of these columns with others. This correlation analysis aimed to determine whether all data points, including those initially flagged as outliers, exhibited consistent trends with related variables. If the data points, regardless of their extremity, followed the same trend as the majority of the data, it was concluded that these were not true outliers, but rather valid, albeit extreme, scenarios within the dataset.

This methodology led us to the conclusion that there were no true outliers in the dataset. The data points at the extreme ends, though initially appearing as outliers through the IQR method, were in fact valid scenarios when viewed in the context of their correlation with other variables. This approach ensures that our data analysis is robust and accounts for the natural variability inherent in real-world data, rather than discarding potentially significant information as outliers too hastily. Then For

these columns we calculated the mean and median to see the data distribution in these columns. We observed that for every column there was not much difference between the mean and median value except '**YearsSinceLastPromotion**' where the mean value is over twice as significant as the median value. The median indicates that most employees were promoted just one year ago, whereas the mean suggests that most were promoted over two years ago. Therefore, the outliers should be removed to correct this discrepancy.

e. Therefore, in the next step, we eliminated 107 records from the dataset, rechecked the mean and median, and found that it was better than before and had not even impacted any other column. Thus, the insight gained from the above analysis are :

i. **Keep Contextually Relevant Outliers:** For fields like MonthlyIncome, TotalWorkingYears, YearsAtCompany, where high values might realistically exist, we should keep the outliers as they could represent actual scenarios, such as high-income executives or long-tenured employees.

ii. **Assess Skewed Distributions:** For PerformanceRating and TrainingTimesLastYear, the high number of outliers suggests a skewed distribution. This could be realistic (e.g., most employees have a standard performance rating, with a few exceptional cases). We will keep these outliers as they represent important variations.

iii. **Review Outliers with Suspicious Counts:** In a real-world scenario, some employees might indeed have worked at a significantly higher number of companies compared to their peers. This could be due to a variety of reasons, such as career choices, industry norms, or personal circumstances. In such cases, these outliers represent genuine data points that provide insights into certain employee segments

iv. **Variables with Few Outliers:** For variables like YearsInCurrentRole and YearsWithCurrManager, where the number of outliers is relatively low, it could be reasonable to keep them unless they represent impossible scenarios.

Given these considerations, the recommended approach would be to retain the outliers in this dataset except for 'YearsSinceLastPromotion'.

4. Feature Engineering

a. Converted categorical variables to numerical using Label Encoding.

b. **Addressing Multicollinearity:** We excluded 'MonthlyIncome' from our analysis owing to its significant correlation with 'JobLevel'. Likewise, the columns 'YearsAtCompany', 'YearsInCurrentRole', 'YearsWithCurrManager', and 'YearsSinceLastPromotion' were dropped due to their substantial correlations with 'TotalWorkingYears'. This action was taken to mitigate the effects of multicollinearity. Multicollinearity, which occurs when independent variables in a regression model are highly correlated, can adversely affect the model's performance. It can lead to unreliable and unstable estimates of regression coefficients, making it challenging to ascertain the effect of each independent variable.

c. **Handle Class Imbalance**: Initially, the dataset showed a significant imbalance in the class distribution with about 83.88% of the data falling into the 'No Attrition' category (0.0) and only 16.12% in the 'Yes Attrition' category (1.0). Such an imbalance can lead to a biased model that overly favors the majority class. In this case, a predictive model might become overly proficient at identifying 'No Attrition' cases while failing to accurately detect the more critical 'Yes Attrition' cases. To address this, we adopted the SMOTE oversampling approach, after which classes were equally balanced.

5. Model Implementation

a. Random Forest

1. Comparison between initial and best hyper-parameters.

Initial Parameter Grid	Best Parameter Selected
n_estimators: [50, 80, 100, 200]	n_estimators: 50
max_depth: [3, 5, 10, 15, 20, 30]	max_depth: 20
min_samples_split: [3, 5, 8]	min_samples_split: 3
min_samples_leaf: [1, 2, 3, 5, 10]	min_samples_leaf: 1

2. The top 5 features for Random Forest Model – JobLevel, StockOptionLevel, OverTime, MaritalStatus and JobInvolement.

b. Gradient Boosting

1. Comparison between initial and best hyper-parameters.

Initial Parameter Grid	Best Parameter Selected
n_estimators: [50, 80, 100, 200]	n_estimators: 200
learning_rate: [0.01, 0.05, 0.1]	learning_rate: 0.1
max_depth: [3, 5, 10, 15, 20, 30]	max_depth: 10

2. The top 5 features for Gradient Boosting Model – JobLevel, StockOptionLevel, OverTime, MaritalStatus and EmployeeNumber.

c. Light Gradient Boosting

1. Comparison between initial and best hyper-parameters.

Initial Parameter Grid	Best Parameter Selected
num_leaves: [31, 50]	num_leaves: 31
learning_rate: [0.01, 0.05, 0.1]	learning_rate: 0.05
n_estimators: [50, 80, 100, 200]	n_estimators: 100

2. The top 5 features for Light Gradient Boosting Model – JobSatisfaction, Age EnvironmentSatisfacion, DistanceFromHome and NumCompaniesWoeked.

d. XGBoost

1. Comparison between initial and best hyper-parameters.

Initial Parameter Grid	Best Parameter Selected
n_estimators: [50, 80, 100, 200]	n_estimators: 50

learning_rate: [0.01, 0.05, 0.1]	learning_rate: 0.1
max_depth: [3, 5, 10, 15, 20, 30]	max_depth: 10

2. The top 5 features for XGBoosting Model JobLevel, StockOptionLevel, OverTime, MaritalStatus, and Department.

e. Logistic Regression

1. Comparison between initial and best hyper-parameters.

Initial Parameter Grid	Best Parameter Selected
C: [0.001, 0.01, 0.1, 1, 10, 100]	C: 1
penalty: ['l1', 'l2']	penalty: l2

2. The top 5 features for Logistic Regression Model JobLevel, StockOptionLevel, OverTime, MaritalStatus, Department.

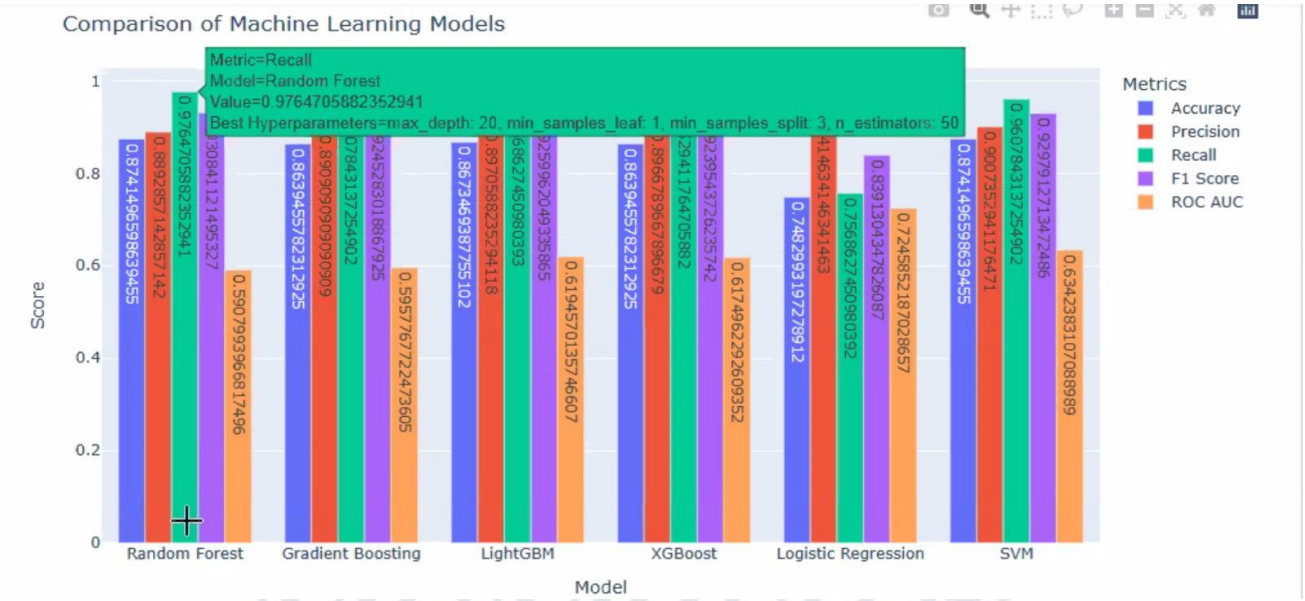
f. SVM

1. Comparison between initial and best hyper-parameters.

Initial Parameter Grid	Best Parameter Selected
svm_C: [0.1, 1, 10]	svm_C: 10
svm_kernel: ['linear', 'rbf']	svm_kernel: rbf
svm_gamma: ['scale', 'auto']	svm_gamma: scale

2. The top 5 features for SVM Model - Department , OverTime , RelationshipSatisfaction, WorkLifeBalance and JobInvolvement.

6. Model Selection



A high recall model efficiently identifies a more significant proportion of positive attrition cases. This feature is essential for proactive human resource management, as it allows timely intervention strategies to retain talent, plan for potential staff shortages, and maintain organizational knowledge and stability. High recall empowers employers to anticipate and act on potential employee turnover before it occurs, thus making them better equipped. **In summary, with its highest recall, the Random Forest model becomes a powerful tool in proactively managing employee attrition**.