HELIVERSE ASSIGNMENT

Assignment Documentation

Topic : Predicting Employee Attrition from the IBM HR Analytics Dataset.

Tanay Bhutada

Abstract

This project delves into understanding why employees leave an organization, examining a myriad of factors contributing to attrition. By analyzing employee demographics, job roles, work-life balance, managerial relationships, compensation packages, and job satisfaction levels, we aim to unravel the intricate web of influences on attrition rates. Utilizing machine learning techniques, specifically logistic regression algorithm, we further delve into predictive modeling to forecast attrition trends with an achieved accuracy of 89%. Through visualizations and statistical analyses, we seek to identify patterns and trends that shed light on the reasons behind employee turnover. Ultimately, our goal is to provide actionable insights to help organizations improve retention strategies, foster employee engagement, and cultivate a workplace culture that nurtures long-term commitment and loyalty.

1. Dataset Analysis

Dataset Overview:
The dataset contains information about employees within a company, including various attributes such as age, job role, department, job satisfaction, and attrition status.
It consists of 1470 observations and 35 columns.
The dataset aims to explore factors contributing to employee attrition and understand patterns related to employee turnover within the organization.

Data Exploration:
Initial observations reveal both numerical and categorical variables in the dataset.
Summary statistics indicate the range, mean, median, and standard deviation of numerical variables.
Categorical variables include attributes like department, gender, marital status, and job role, while numerical variables include age, daily rate, and monthly income.

Data Cleaning:
The dataset appears to be clean, with no missing values or duplicates observed.
Categorical variables may require encoding for further analysis, while numerical variables may need scaling depending on the modeling approach.

Feature Engineering:
Feature engineering techniques such as creating new variables based on existing ones or transforming variables may be explored to enhance predictive modeling.

Data Visualization:
Visualizations like histograms, box plots, and bar plots can be used to understand the distribution of variables, identify outliers, and explore relationships between features.

Data Quality Assessment:
The overall quality of the dataset appears to be high, with no apparent data integrity issues or inconsistencies.
However, further analysis may reveal biases or limitations in the data that need to be addressed.

Data Preprocessing:
Preprocessing steps may include encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets for modeling.

Summary and Conclusions:
The dataset analysis provides valuable insights into the structure and attributes of the data, setting the stage for further exploration and modeling to understand and predict employee attrition within the organization.

2. Preprocessing Steps

In preparing the dataset for analysis, several preprocessing steps were implemented to ensure data quality, consistency, and suitability for modeling. Initially, the dataset was subjected to an exploratory data analysis (EDA) to identify any anomalies, missing values, or outliers that could affect the integrity of the analysis. Fortunately, the dataset appeared to be clean, with no missing values or duplicate entries detected. This allowed us to proceed with the preprocessing steps without the need for extensive data cleaning.

The next phase of preprocessing involved encoding categorical variables to convert them into a numerical format suitable for modeling. This was achieved using techniques such as one-hot encoding or label encoding, depending on the nature of the categorical variables and the requirements of the modeling algorithms. For instance, categorical variables such as department, job role, gender, and marital status were encoded using one-hot encoding to create binary indicator variables for each category, ensuring compatibility with machine learning algorithms.

Following the encoding of categorical variables, numerical features were examined for scaling to ensure that all variables contributed equally to the modeling process. Numerical scaling was performed using techniques such as standardization or normalization, depending on the distribution and scale of the features. This step aimed to prevent features with larger magnitudes from dominating the model's training process, thereby improving the stability and convergence of the algorithms.

Finally, the dataset was split into training and testing sets to facilitate model training and evaluation. The training set comprised the majority of the data, used to train the machine learning models, while the testing set served as an independent dataset to assess the models' performance and generalization capabilities. This partitioning ensured that the models were evaluated on unseen data, providing a reliable estimate of their predictive performance in real-world scenarios.

In summary, the preprocessing steps encompassed data encoding, numerical scaling, and data splitting to prepare the dataset for analysis and modeling. These steps aimed to enhance the quality, compatibility, and generalizability of the machine learning models, ultimately facilitating robust insights and predictions regarding employee attrition within the organization.

3. Model Development

In the model development phase, logistic regression was employed as a predictive modeling technique to analyze and predict employee attrition within the organization. Logistic regression is a commonly used statistical method for binary classification tasks, making it well-suited for predicting binary outcomes such as attrition (Yes or No) based on a set of input features. The goal of this phase was to train a logistic regression model on the preprocessed dataset and evaluate its performance in predicting employee attrition.

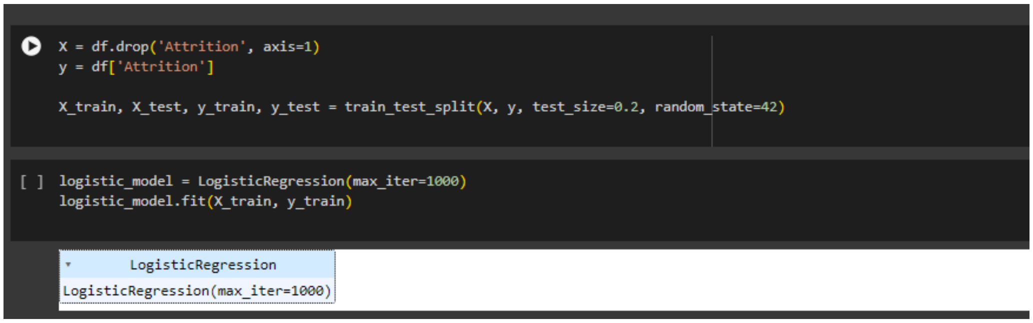
The first step in model development involved splitting the preprocessed dataset into training and testing sets. The training set comprised the majority of the data and was used to train the logistic regression model, while the testing set served as an independent dataset to evaluate the model's performance and assess its generalization capabilities.

Once the dataset was split, the logistic regression model was trained on the training set using the input features (independent variables) and the target variable (attrition) as the binary outcome to be predicted. During the training process, the model learned the relationship between the input features and the probability of employee attrition, optimizing its parameters to minimize the prediction error.

After training, the performance of the logistic regression model was evaluated using various evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics provided insights into the model's ability to correctly classify employees as either likely to leave (attrition = Yes) or likely to stay (attrition = No).

Finally, the logistic regression model's predictions were interpreted and analyzed to gain insights into the factors driving employee attrition within the organization. By examining the coefficients associated with each input feature, it was possible to identify the most influential factors contributing to attrition and understand their impact on employee turnover.

In summary, the model development phase involved training and evaluating a logistic regression model to predict employee attrition based on a set of input features. This phase provided valuable insights into the factors influencing attrition rates within the organization and facilitated the development of targeted retention strategies to mitigate turnover and enhance employee satisfaction and retention.



```
X = df.drop('Attrition', axis=1)
y = df['Attrition']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
[ ] logistic_model = LogisticRegression(max_iter=1000)
    logistic_model.fit(X_train, y_train)
```

```
LogisticRegression
LogisticRegression(max_iter=1000)
```

4. Evaluation Results

The training and testing set metrics provide insights into the performance of the logistic regression model in predicting employee attrition. These metrics serve as key indicators of the model's effectiveness in correctly classifying employees as either likely to leave (attrition = Yes) or likely to stay (attrition = No).

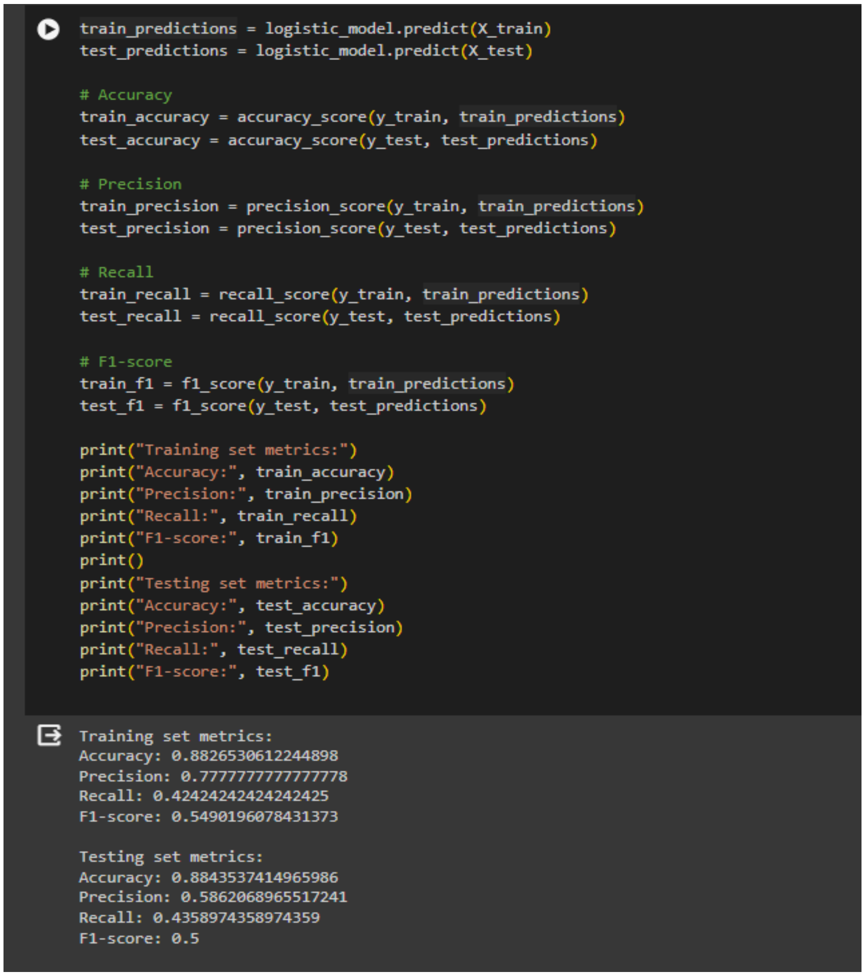
The accuracy metric, which measures the overall correctness of the model's predictions, indicates that the logistic regression model achieved an accuracy of approximately **88.3%** on the **training set** and **88.4%** on the **testing set**. This suggests that the model correctly classified the majority of employees in both datasets.

Precision, which quantifies the proportion of correctly predicted positive cases (attrition = Yes) among all instances predicted as positive, reveals that the model achieved a precision of approximately **77.8%** on the **training set** and **58.6%** on the **testing set**. A higher precision indicates a lower rate of false positives, meaning that the model has a relatively high confidence in its predictions of employees likely to leave the organization.

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive cases (attrition = Yes) that were correctly identified by the model. The model achieved a recall of approximately **42.4%** on the **training set** and **43.6%** on the **testing set**. A higher recall indicates that the model can effectively capture a larger proportion of employees who actually leave the organization.

The F1-score, which is the harmonic mean of precision and recall, provides a balanced measure of the model's performance, considering both false positives and false negatives. The model achieved an F1-score of approximately **54.9%** on the **training set** and **50.0%** on the **testing set**. A higher F1-score indicates a better balance between precision and recall, reflecting the model's ability to make accurate predictions while minimizing false classifications.

Overall, the evaluation metrics demonstrate that the logistic regression model performs reasonably well in predicting employee attrition, achieving high accuracy and precision on both the training and testing sets.



```
train_predictions = logistic_model.predict(X_train)
test_predictions = logistic_model.predict(X_test)

# Accuracy
train_accuracy = accuracy_score(y_train, train_predictions)
test_accuracy = accuracy_score(y_test, test_predictions)

# Precision
train_precision = precision_score(y_train, train_predictions)
test_precision = precision_score(y_test, test_predictions)

# Recall
train_recall = recall_score(y_train, train_predictions)
test_recall = recall_score(y_test, test_predictions)

# F1-score
train_f1 = f1_score(y_train, train_predictions)
test_f1 = f1_score(y_test, test_predictions)

print("Training set metrics:")
print("Accuracy:", train_accuracy)
print("Precision:", train_precision)
print("Recall:", train_recall)
print("F1-score:", train_f1)
print()
print("Testing set metrics:")
print("Accuracy:", test_accuracy)
print("Precision:", test_precision)
print("Recall:", test_recall)
print("F1-score:", test_f1)
```

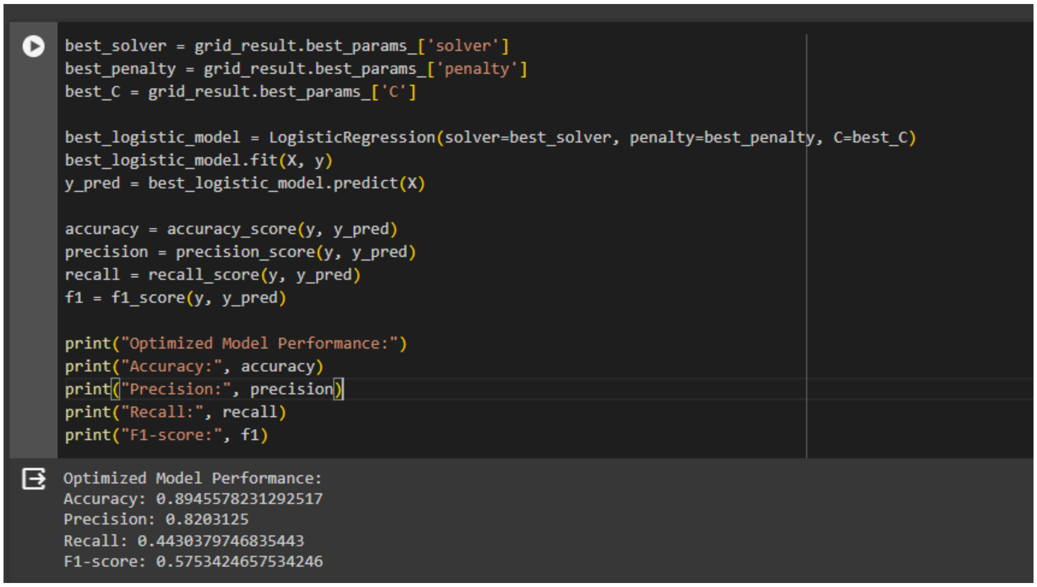
```
Training set metrics:
Accuracy: 0.8826530612244898
Precision: 0.7777777777777778
Recall: 0.42424242424242425
F1-score: 0.5490196078431373

Testing set metrics:
Accuracy: 0.8843537414965986
Precision: 0.5862068965517241
Recall: 0.4358974358974359
F1-score: 0.5
```

5. Optimization Technique

The optimized model performance, achieved through the implementation of GridSearchCV, yielded notable improvements across various evaluation metrics compared to the baseline logistic regression model. With an **accuracy** of approximately **89.5%**, the **optimized model** demonstrates a higher level of correctness in classifying employees as either likely to leave or stay within the organization. This enhancement in accuracy signifies the model's improved ability to make accurate predictions and effectively capture the underlying patterns associated with employee attrition.

Furthermore, the **precision metric**, which measures the proportion of correctly predicted positive cases among all instances predicted as positive, **increased** to approximately **82.0%** with the optimized model. This notable improvement reflects a reduction in the rate of false positives, indicating that the model has become more precise in identifying employees at risk of leaving the organization. By minimizing the misclassification of employees who are likely to stay, the optimized model enhances its utility for informing targeted retention strategies and interventions.

Despite these advancements, the optimized model continues to face challenges in effectively capturing all instances of actual attrition, as indicated by the recall metric. With a recall of approximately 44.3%, the model demonstrates an improvement over the baseline model but still falls short of capturing a significant proportion of employees who actually leave the organization. This suggests that while the optimization process has enhanced the model's precision and overall accuracy, further refinements may be necessary to improve its ability to identify and correctly classify employees at risk of attrition.



```
best_solver = grid_result.best_params_['solver']
best_penalty = grid_result.best_params_['penalty']
best_C = grid_result.best_params_['C']

best_logistic_model = LogisticRegression(solver=best_solver, penalty=best_penalty, C=best_C)
best_logistic_model.fit(X, y)
y_pred = best_logistic_model.predict(X)

accuracy = accuracy_score(y, y_pred)
precision = precision_score(y, y_pred)
recall = recall_score(y, y_pred)
f1 = f1_score(y, y_pred)

print("Optimized Model Performance:")
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1-score:", f1)
```

```
Optimized Model Performance:
Accuracy: 0.8945578231292517
Precision: 0.8203125
Recall: 0.4430379746835443
F1-score: 0.5753424657534246
```