IBM HR Analytics Employee Attrition & Performance Dataset Analysis and Modeling

1. Introduction:

Employee attrition, or the rate at which employees leave a company, is a critical concern for organisations. High attrition rates can lead to increased recruitment costs, loss of productivity, and decreased morale among remaining employees. I aimed to create various models to predict employee attrition and then after that found out the major contributing factors to employees leaving.

2. Dataset Analysis:

The dataset contains information on various attributes of employees, including demographics, job satisfaction, and performance metrics. There are 35 columns out of which 9 are having categorical variables while the rest have integers.

3. Data Preprocessing:

As part of preprocessing I checked whether there are any NULL values but there are no NULL values in the complete dataset. Then I checked for duplicate rows to avoid model getting biassed toward a specific kind of data, none were found to be duplicate.

As a next step I performed Label encoding for the categorical variables which had binary values(like ‘Yes’ or ‘No’, ‘Male’ or ‘Female’ etc.) and One Hot Encoding for the categorical variables which had multiple classes('BusinessTravel', 'Department', 'EducationField', 'JobRole', 'MaritalStatus').

Next I scaled the columns with integers and the Dataset is ready for training.

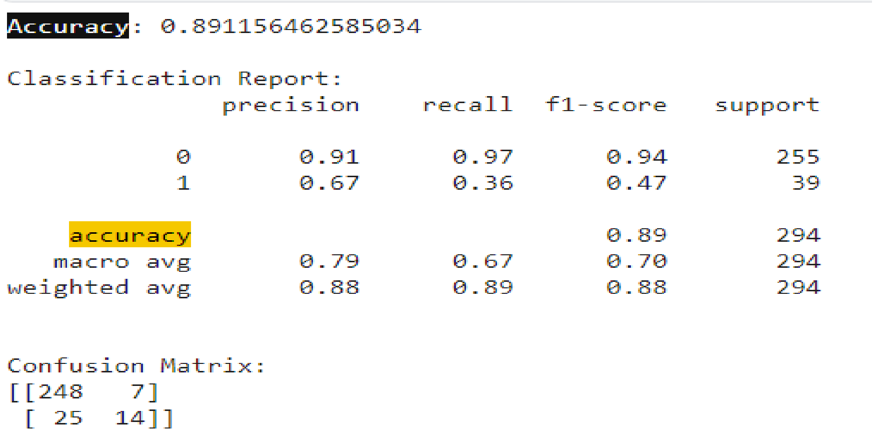
4. Model Development:

Several classification models, including **Logistic Regression, Random Forest, XGBoost, Gradient Boosting, and SVM,** were trained to predict employee attrition. Each model was evaluated based on the metrics in the classification report among others.

5. Model Evaluation:

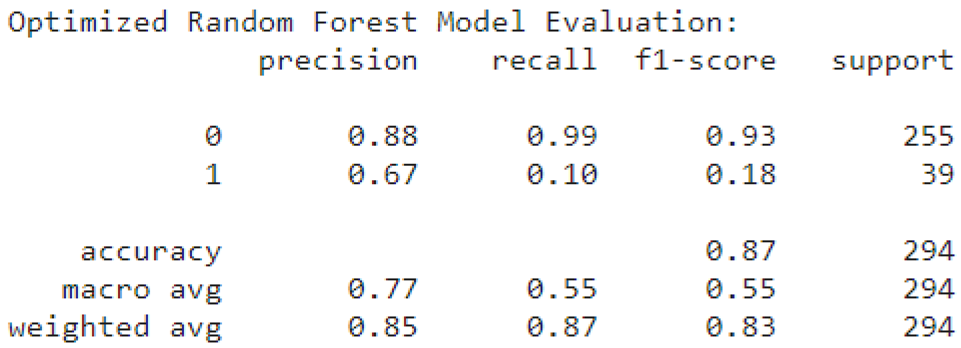
Evaluation results show the performance of each model in terms of accuracy, precision, recall, and F1-score:

- Logistic Regression:



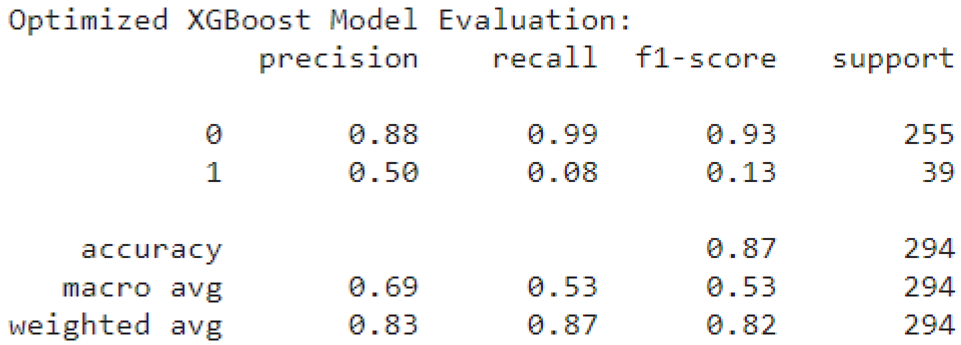
```
Accuracy: 0.891156462585034

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94       255
           1       0.67      0.36      0.47        39

    accuracy                           0.89       294
   macro avg       0.79      0.67      0.70       294
weighted avg       0.88      0.89      0.88       294


Confusion Matrix:
[[248   7]
 [ 25  14]]
```

- Random Forest:



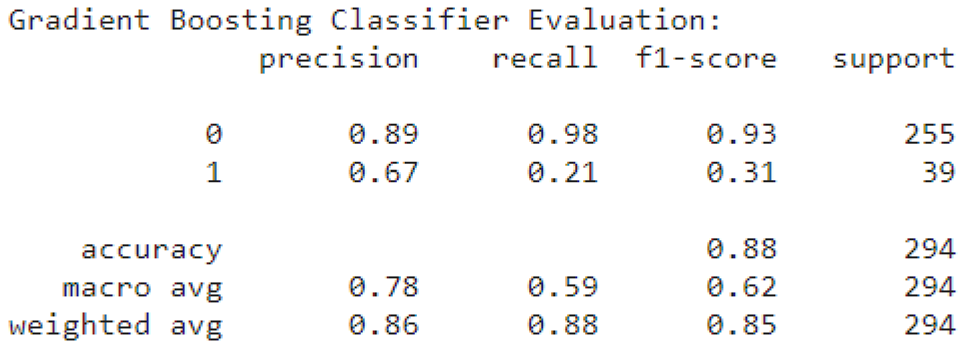
```
Optimized Random Forest Model Evaluation:
              precision    recall  f1-score   support

           0       0.88      0.99      0.93       255
           1       0.67      0.10      0.18        39

    accuracy                           0.87       294
   macro avg       0.77      0.55      0.55       294
weighted avg       0.85      0.87      0.83       294
```

- XGBoost:



```
Optimized XGBoost Model Evaluation:
              precision    recall  f1-score   support

           0       0.88      0.99      0.93       255
           1       0.50      0.08      0.13        39

    accuracy                           0.87       294
   macro avg       0.69      0.53      0.53       294
weighted avg       0.83      0.87      0.82       294
```

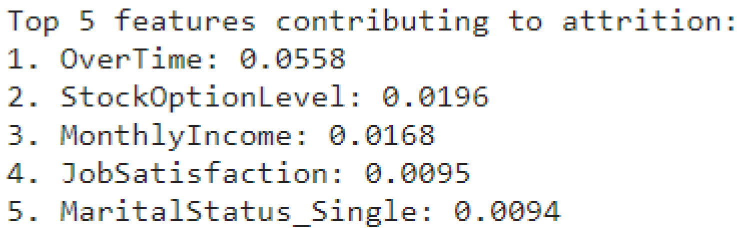
- Gradient Boosting:



```
Gradient Boosting Classifier Evaluation:
              precision    recall  f1-score   support

           0       0.89      0.98      0.93       255
           1       0.67      0.21      0.31        39

    accuracy                           0.88       294
   macro avg       0.78      0.59      0.62       294
weighted avg       0.86      0.88      0.85       294
```

6. Model Optimization:
Hyperparameter tuning, feature selection, and ensemble methods were explored for optimising model performance. Methods such as GridSearchCV for hyperparameter tuning and SelectFromModel for feature selection are employed to improve model accuracy and robustness.

7. Factors contributing the most to Attrition:
Based on the Features Importances derived from Permutation Importance calculated after training Random Forest Classifier, top 5 factors contributing to the attrition were observed.



```
Top 5 features contributing to attrition:
1. OverTime: 0.0558
2. StockOptionLevel: 0.0196
3. MonthlyIncome: 0.0168
4. JobSatisfaction: 0.0095
5. MaritalStatus_Single: 0.0094
```

This means that to reduce attrition the features OverTime, StockOptionLevel, MonthlyIncome, JobSatisfaction, MaritalStatus_Single will have to be shifted more in the favour of the Employee, where applicable.

8. Conclusion:

In conclusion, Logistic Regression was most accurate and precise in predicting Employee Attrition with 89% accuracy and 0.88 weighted avg precision, logistic regression also has the highest weighted average recall-> 0.89 and f1-score->0.88. This analysis sheds light on the factors influencing employee attrition, by implementing effective retention strategies, organisations can foster a positive work environment and enhance employee satisfaction and productivity.