Job Layoffs factors: Analysis & Prediction Model

Sumit Nalwade MS. Data Science ’23

Mentor: Dr. Christelle Scharff

Pace University, Seidenberg School of CSIS

Github

Abstract

In response to the escalating uncertainty in the job market and the prevalence of sudden layoffs across industries, the imperative for a robust layoff prediction model has become increasingly evident. Despite the scale of the problem, there is a notable lack of systematic study in this area. This study is attempting to close this significant gap by proposing a sophisticated categorization algorithm for forecasting staff layoffs.
This study describes and evaluates the use of two renowned classification models, Random Forest and XGBoost, in the context of layoff prediction. The inquiry includes a thorough performance analysis of both models, revealing their different strengths and drawbacks. Furthermore, the article digs into the identification of important variables critical for accurate layoff forecasts, applying model interpretation approaches such as Lime, SHAP, and Eli5. The models are evaluated on a dedicated dataset that includes employee variables such as age, gender, job engagement, and years of experience. To assess the efficiency of the models, performance indicators such as accuracy score, AUC-ROC, confusion matrix, and classification report are rigorously used. The observed findings highlight both models' excellent performance, with accuracy values over 90%.
This research not only addresses a significant void in current literature but also serves as a catalyst for further exploration within this untapped domain. The findings pave the way for advanced analyses, enabling the development of even more accurate models and fostering future research avenues. In essence, this study represents a foundational step towards enhancing our understanding of layoff prediction and propelling the field towards new realms of inquiry.

Research Question

Que 1- What are the key features influencing layoff predictions, and how can advanced interpretability techniques such as Lime, SHAP, and Eli5 contribute to a deeper understanding of model decision-making processes in the context of job layoffs?

Que 2- Can we predict the job layoff for an employee based on factors like Age, Sex, Skills, Experience , work preference , Job Title, location ?

Dataset:

We've compiled a dataset of 3,101 records from layoffs.fyi, a source aggregating global layoff data. Furthermore, our exclusive training dataset, which includes 3,067 items and 16 columns, is designed for predictive model training.

Key Points:
Source: layoffs.fyi
Records: 3,101 (Extracted), 3,067 (Training)
Purpose: Training predictive models on employee layoffs.
Features Example:

1. Age
2. Job Involvement
3. Performance rating
4. Years of experience
5. Hourly wages
6. Education
7. Gender

Methodology

❑ **Data Collection and Preprocessing:**

- ➢ Aggregated data from diverse sources.
- ➢ Merged datasets into a unified dataset.
- ➢ Applied data cleaning and manipulation techniques to ensure data quality.

❑ **Exploratory Data Analysis (EDA) with Visualization:**

- ➢ Conducted detailed data analysis using visualization libraries like seaborn and matplotlib .
- ➢ Explored patterns and insights from the data.

❑ **Feature Encoding:**

- ➢ Utilized Label Encoder library in Python to encode categorical variables.
- ➢ Assigned numerical values to labeled data for model compatibility.

❑ **Data Splitting:**

- ➢ Employed Train_test_split library to partition the dataset into an 80:20 ratio for training and testing, respectively.

❑ **Addressing Class Imbalance:**

- ➢ Applied Synthetic Minority Over-sampling Technique (SMOTE) on the training dataset to mitigate class imbalance, particularly addressing the overrepresentation of the 'No' class label.

❑ **Interpretable Model Implementation:**

- ➢ Implemented Local Interpretable Model-Agnostic Explanations (LIME) to explain individual predictions.
- ➢ Employed SHAP (Shapley Additive explanations) for global explanations of the models.

❑ **Performance Evaluation:**

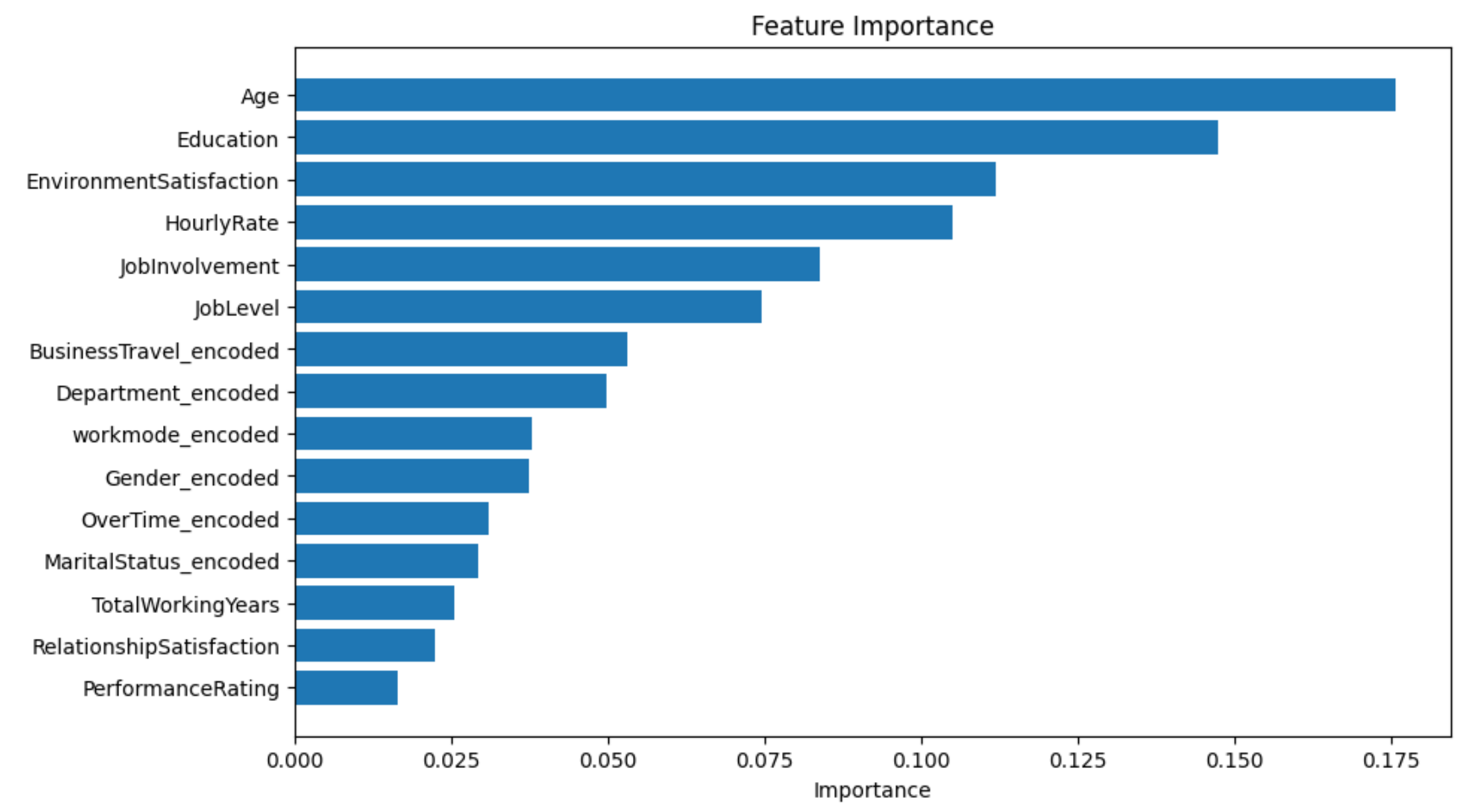
- ➢ Evaluated model performances using diverse metrics:
  - Accuracy_score
  - Confusion matrix
  - AUC-ROC
  - Classification report

Note: These steps were executed to ensure robust model training, interpretability, and comprehensive evaluation of the classification models.

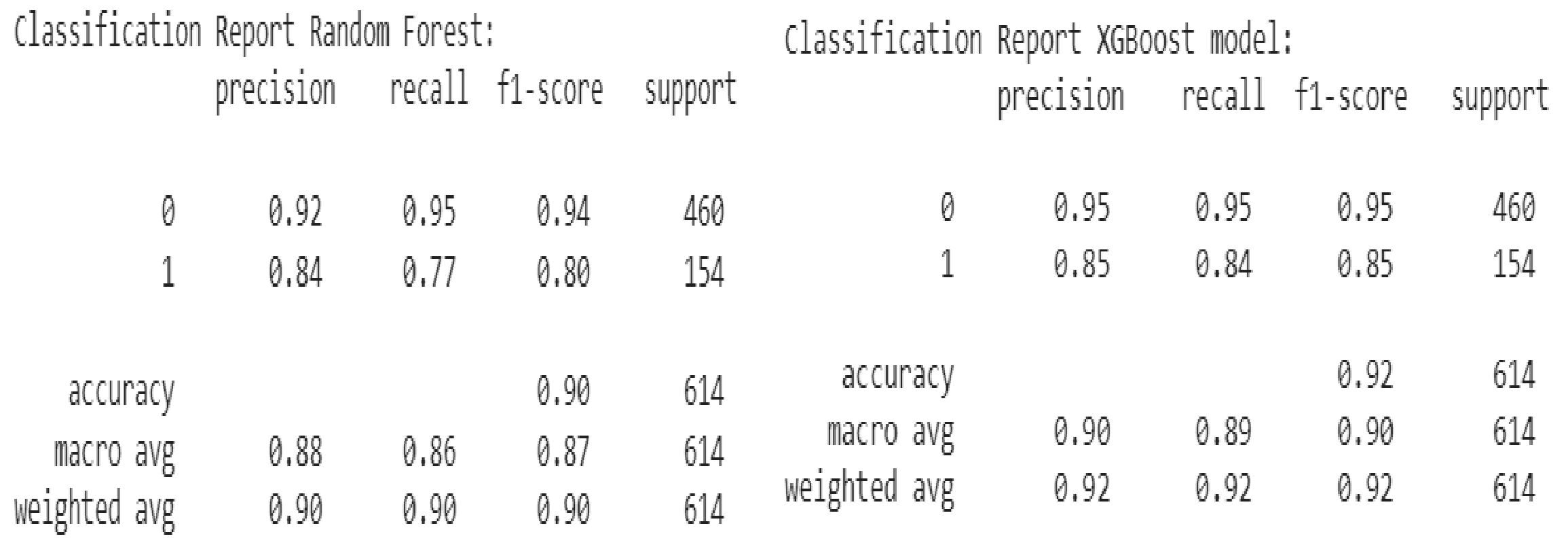
Results

The model has revealed the importance of several factors impacting employee layoffs. Here are some noteworthy conclusions from our most recent job layoff data analysis:

- The highest number of layoffs occurred in 2023 over the past four years.
- The United States experienced the highest number of layoffs compared to any other country.
- Sectors most susceptible to layoffs in the United States: Finance, Real Estate, and Marketing.
- Top three companies with the most significant layoffs: Google, Meta, and Salesforce.
- Employees aged 30-35 and younger individuals (19-22 years) face a higher risk of layoffs, the latter possibly due to limited experience.
- The Random Forest model identified Age as the most crucial factor in predicting layoffs.



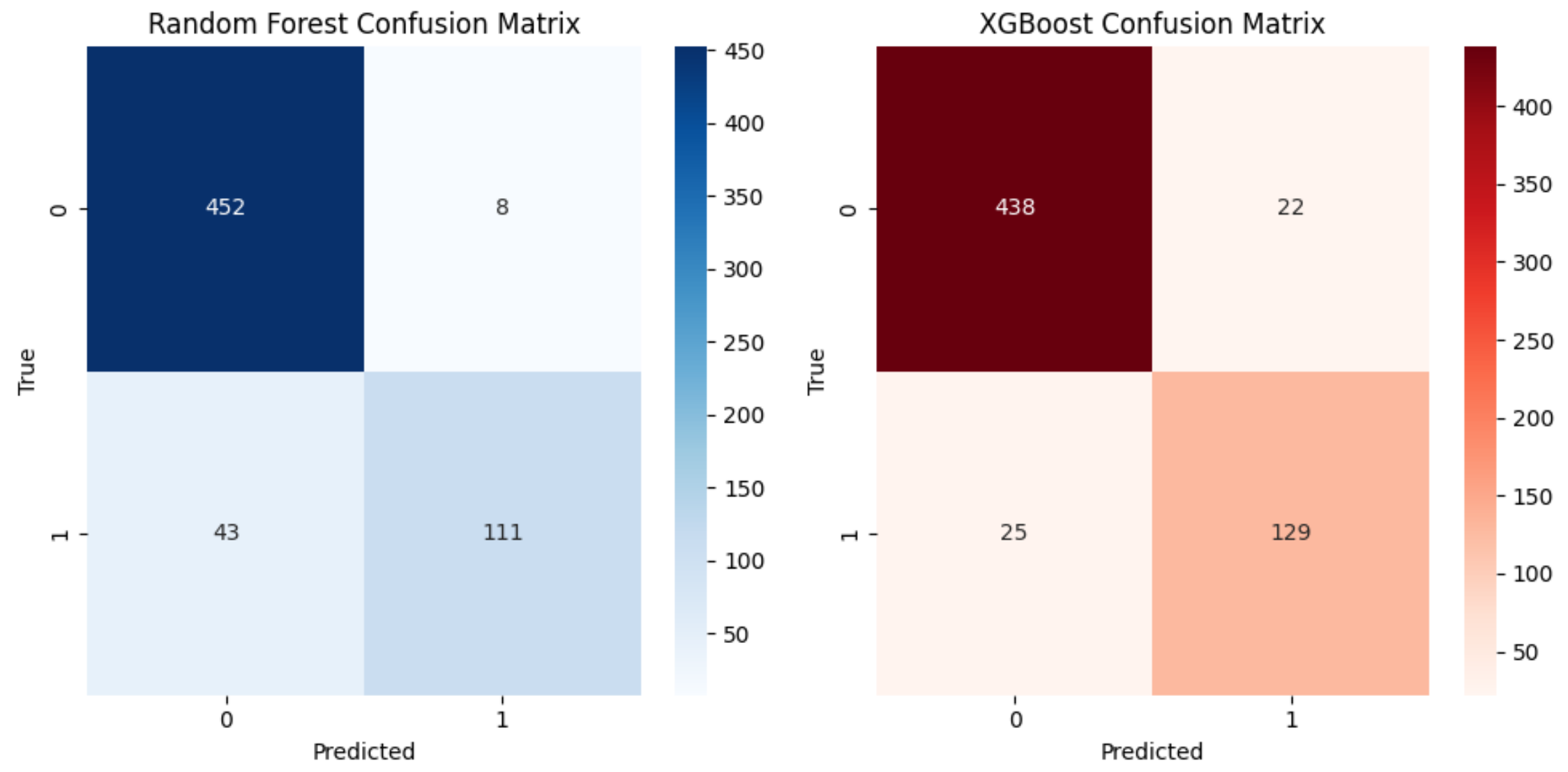
- We achieved an impressive F1 score of 92 & 90 for the XGBoost and Random Forest models respectively. Notably, the XGBoost model demonstrated higher precision scores across both labels compared to the Random Forest model. Below are the detailed classification reports for the Random Forest and XGBoost models:



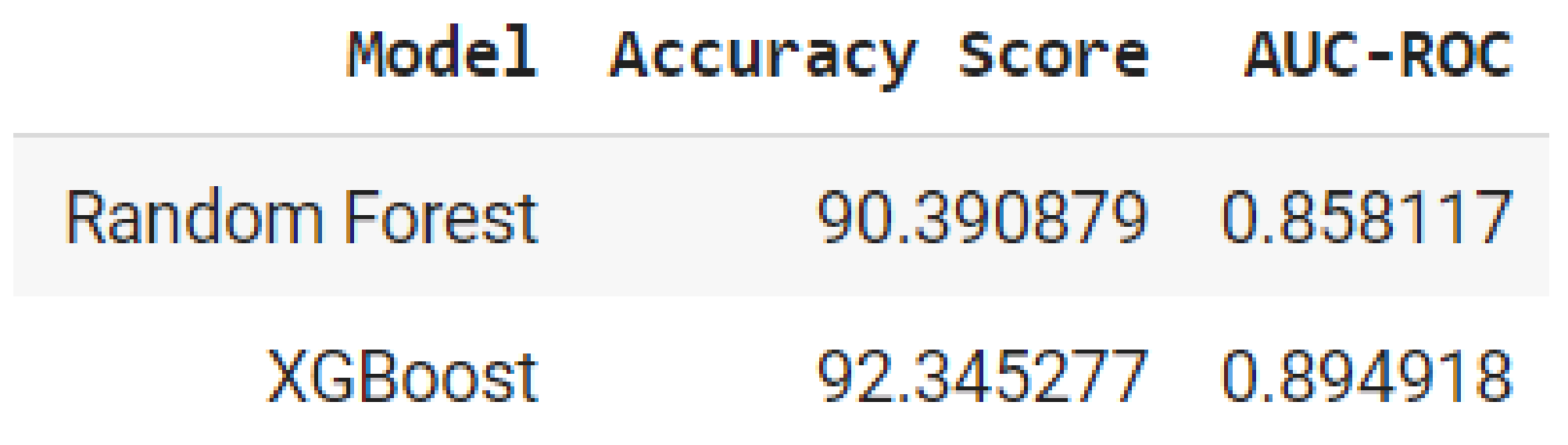
```
Classification Report Random Forest:
              precision    recall  f1-score   support

           0       0.92      0.95      0.94       460
           1       0.84      0.77      0.80       154

    accuracy                           0.90       614
   macro avg       0.88      0.86      0.87       614
weighted avg       0.90      0.90      0.90       614
```

```
Classification Report XGBoost model:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95       460
           1       0.85      0.84      0.85       154

    accuracy                           0.92       614
   macro avg       0.90      0.89      0.90       614
weighted avg       0.92      0.92      0.92       614
```

- The XGBoost model outperforms the Random Forest model in terms of predicting True Positive classes. Additionally, false negative predictions have been substantially decreased by 50%. This is the corresponding confusion matrix for each of the two models:



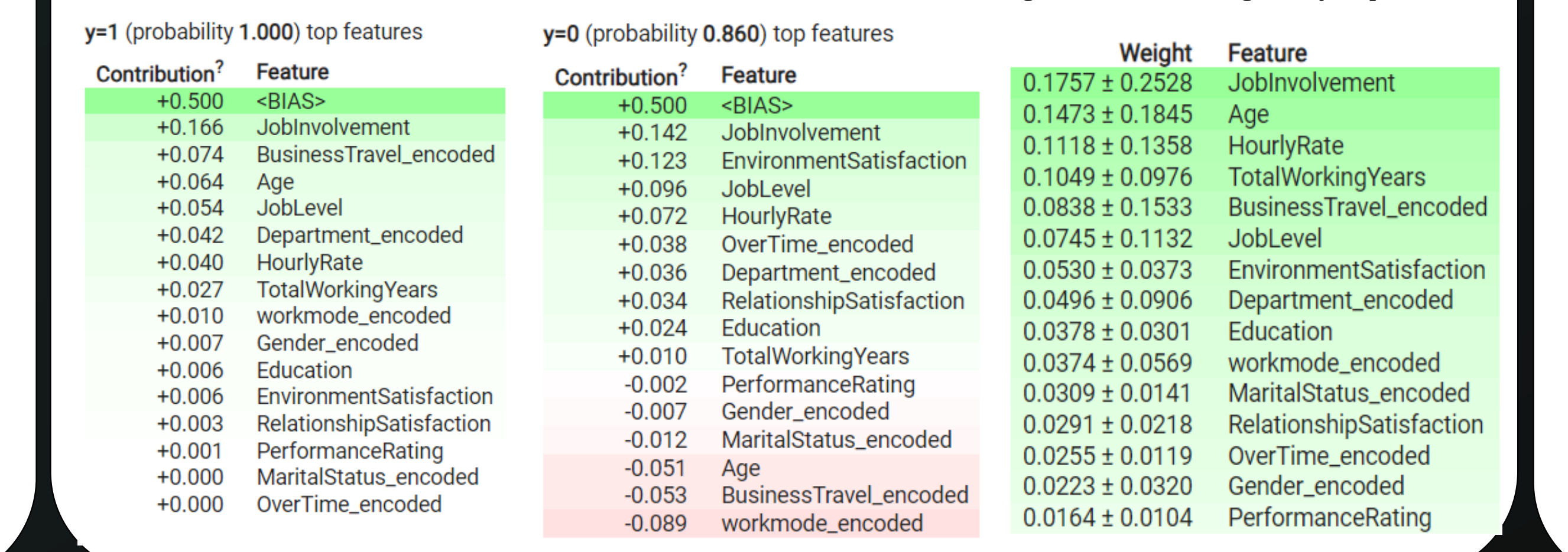
- Here is a comparison of the XGBoost and Random forest models' accuracy and AUC_ROC scores. Comparing the XGBoost model to Random Forest, it is more accurate.



Model	Accuracy Score	AUC-ROC
Random Forest	90.390879	0.858117
XGBoost	92.345277	0.894918

- Eli5 elucidates the influential factors behind predictions made by the random forest model. It identifies the feature "<BIAS>" as the predominant contributor to both "1" and "0" labels. Additionally, Eli5 emphasizes the substantial positive influence of the "Job Involvement" feature on the model's label predictions.

Weight of each feature globally in prediction



y=1 (probability **1.000**) top features

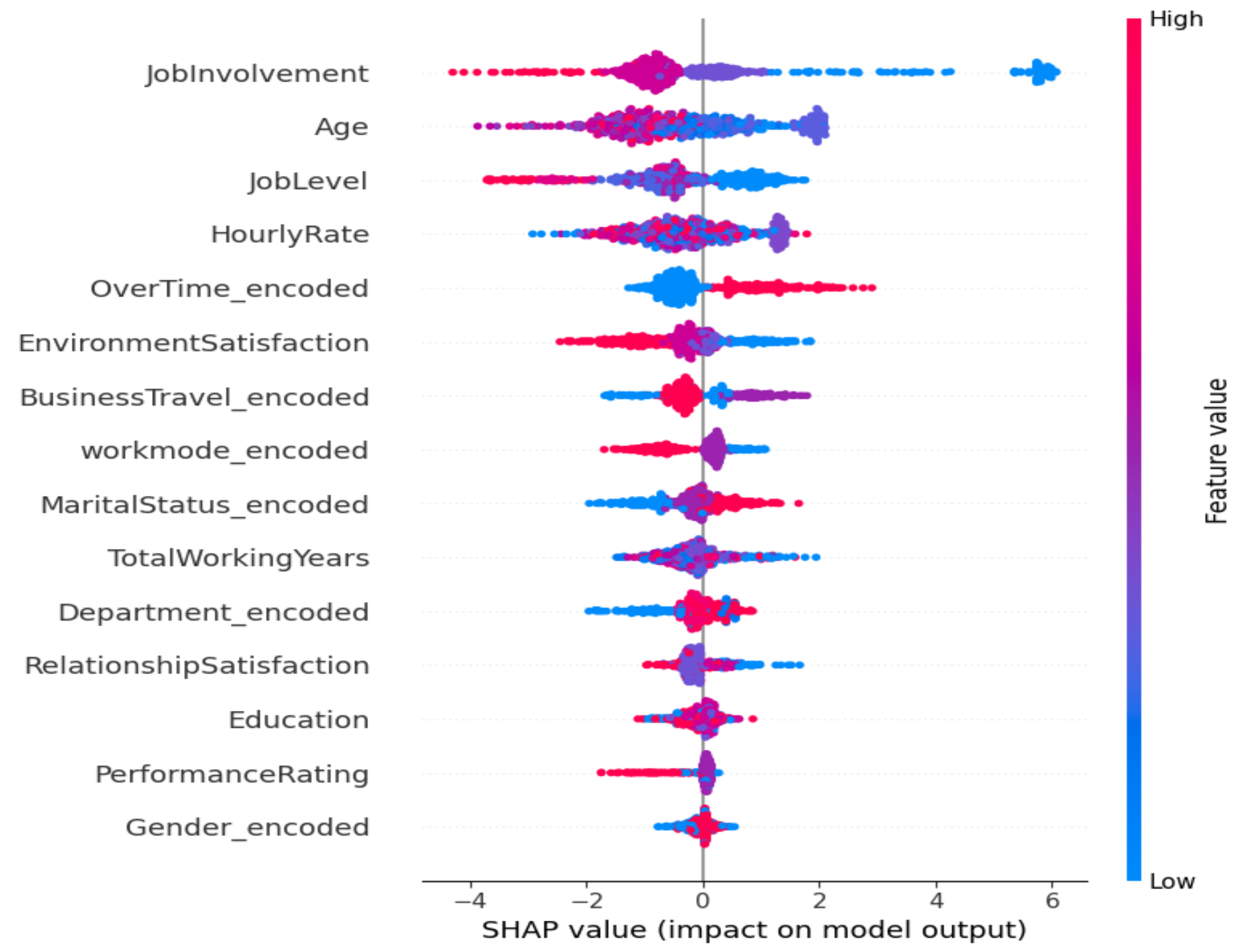
Contribution?	Feature
+0.500	<BIAS>
+0.166	JobInvolvement
+0.074	BusinessTravel_encoded
+0.064	Age
+0.054	JobLevel
+0.042	Department_encoded
+0.040	HourlyRate
+0.027	TotalWorkingYears
+0.010	workmode_encoded
+0.007	Gender_encoded
+0.006	Education
+0.006	EnvironmentSatisfaction
+0.003	RelationshipSatisfaction
+0.001	PerformanceRating
+0.000	MaritalStatus_encoded
+0.000	OverTime_encoded

y=0 (probability **0.860**) top features

Contribution?	Feature
+0.500	<BIAS>
+0.142	JobInvolvement
+0.123	EnvironmentSatisfaction
+0.096	JobLevel
+0.072	HourlyRate
+0.038	OverTime_encoded
+0.036	Department_encoded
+0.034	RelationshipSatisfaction
+0.024	Education
+0.010	TotalWorkingYears
-0.002	PerformanceRating
-0.007	Gender_encoded
-0.012	MaritalStatus_encoded
-0.051	Age
-0.053	BusinessTravel_encoded
-0.089	workmode_encoded

Weight	Feature
0.1757 ± 0.2528	JobInvolvement
0.1473 ± 0.1845	Age
0.1118 ± 0.1358	HourlyRate
0.1049 ± 0.0976	TotalWorkingYears
0.0838 ± 0.1533	BusinessTravel_encoded
0.0745 ± 0.1132	JobLevel
0.0530 ± 0.0373	EnvironmentSatisfaction
0.0496 ± 0.0906	Department_encoded
0.0378 ± 0.0301	Education
0.0374 ± 0.0569	workmode_encoded
0.0309 ± 0.0141	MaritalStatus_encoded
0.0291 ± 0.0218	RelationshipSatisfaction
0.0255 ± 0.0119	OverTime_encoded
0.0223 ± 0.0320	Gender_encoded
0.0164 ± 0.0104	PerformanceRating

- Interpreting SHAP Summary for XGBoost Model
  - ○ **Key Contributors:**
    - ✓ The SHAP summary plot highlights that "Job Involvement," "Age," "Job Level," and "Overtime" are the most influential features shaping the XGBoost model's predictions.
  - ○ **Direction of Impact:**
    - ✓ Specifically, the positive impact of "Job Involvement" indicates a favorable contribution towards predicting class label 1.



Limitations

1. The abundance of data poses a challenge, restricting our ability to train the model on real-time data, potentially impacting its responsiveness to dynamic changes.

2. Insufficient literature and research work in this domain further compound the challenge, limiting the availability of established methodologies and benchmarks for effective model development.

Conclusions & Future work

Age, Experience, and Work engagement all have a major influence on job layoffs; having less experience and less job engagement increases the risk. Future efforts involve enhanced real-time data collection, exploring untapped research avenues, and incorporating factors like employee-manager relationships into advanced models for more precise predictions. These findings pave the way for a comprehensive understanding and effective anticipation of job layoffs.

Bibliography

1. Berson, S. Smith and K. Thearling, Building Data Mining Applications for CRM, New York, NY:McGraw-Hill, 2000.

2. M. Kudo and J. Skalansky, "Comparison of Algorithms That Select Features for Pattern Classifiers", Pattern Recognition, vol. 33, no. 1, pp. 25-41, 2000.

3. Z. Luo, Y. Li, R. Fu and J. Yin, "Don't Fire Me, a Kernel Autoregressive Hybrid Model for Optimal Layoff Plan," 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, 2016, pp. 470-477, doi: 10.1109/BigDataCongress.2016.72.