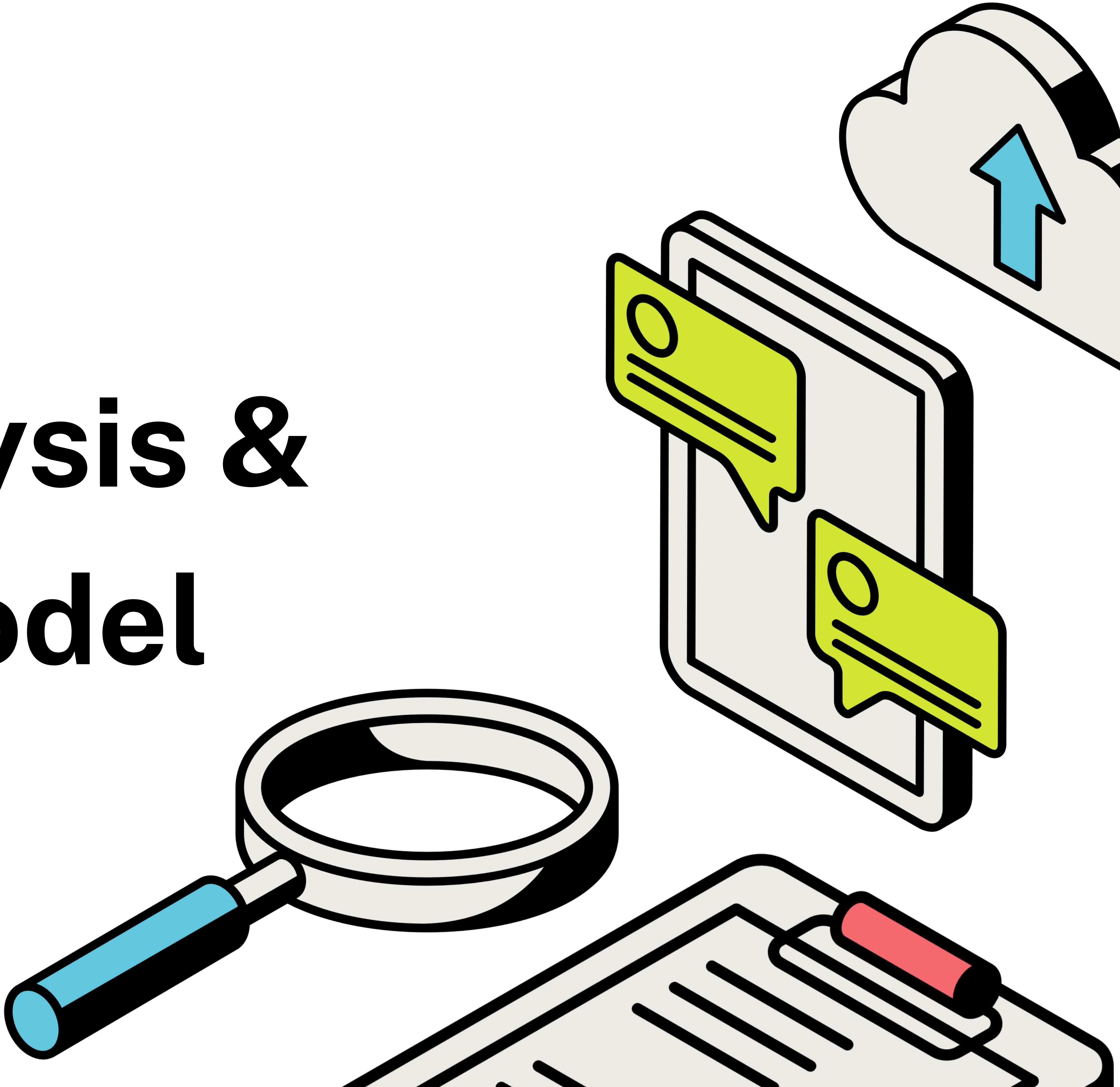
PACE UNIVERSITY

Job Layoffs factors: Analysis & Prediction Model



Date: December 13, 2023

Prepared by: Sumit Nalwade for CS-668

Email: Sumit.nalwade@pace.edu

Mentor: Dr. Christelle Scharff

Github: https://github.com/switchtosumit/job_layoffs_prediction_model

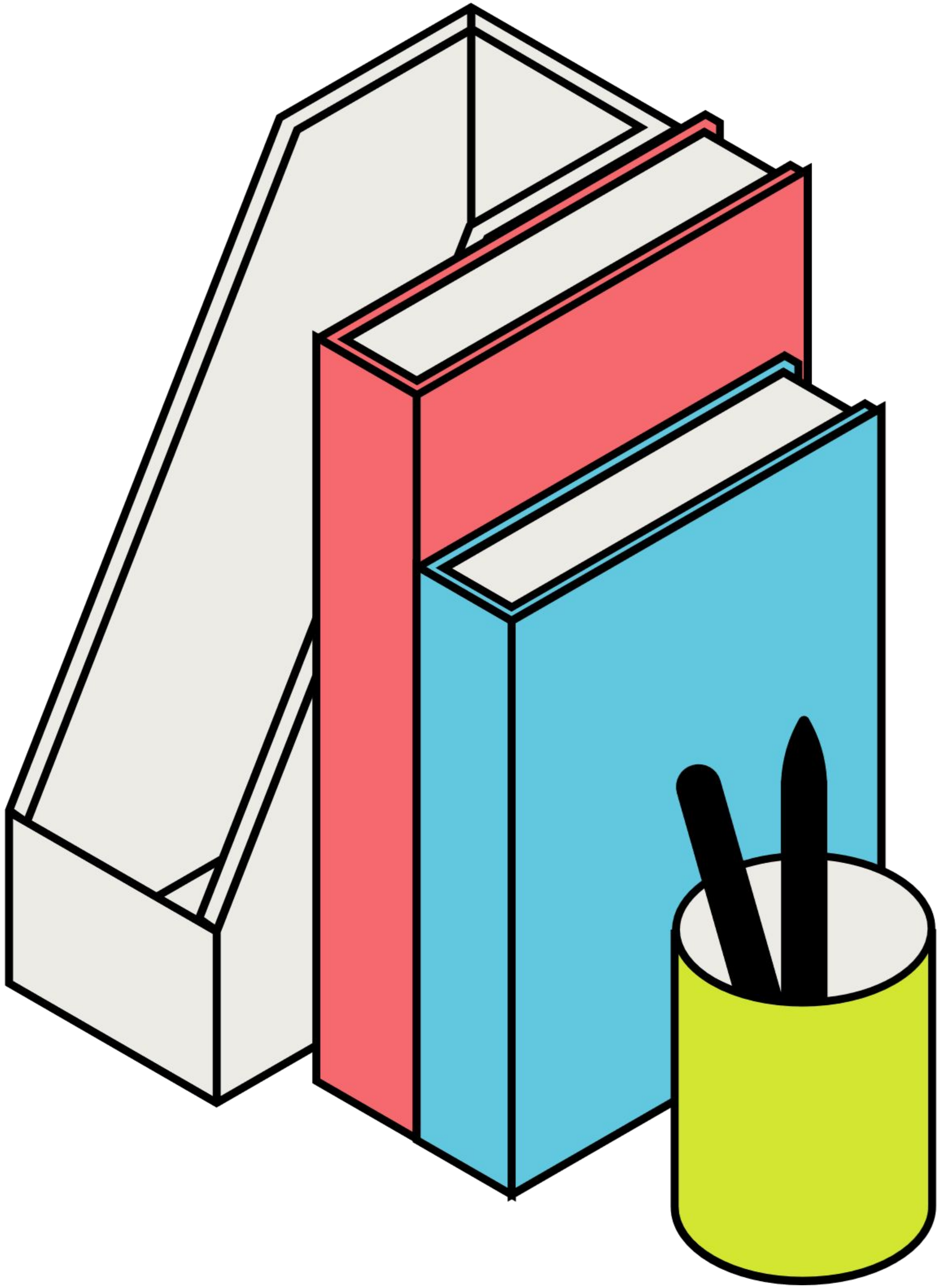
Outline

- Project Proposal
- Literature Review
- Dataset
- EDA & Methodology
- Modeling
- Findings & Results
- Limitations
- Conclusion & Future works

Project Proposal

1	**Targeted Problem - Understanding the Underlying Causes of Workforce Reduction.** With the current wave of labor reductions across several industries, retaining one's job has become increasingly difficult. Many individuals are abruptly removed from their employment, frequently without a full understanding of the causes underlying these choices. As a result, there is an essential need to undertake a comprehensive data analysis to identify the primary factors leading to these workforce cutbacks. Such knowledge is critical for building effective mitigation methods.
2	**Research Question** -Can we predict the risk of job layoff for an employee based on factors like Age, Sex, Skills, Experience , work preference , Job Title, location ?
3	**Dataset** - Below site has live data update post-covid layoffs for multiple industries . I need to download these data and create my own dataset to use it for this project. https://layoffs.fyi/ I am also thinking of creating pseudo dataset for laid-off employees that include data such as skills, experience etc.
4	**Motivation** - As an aspiring data science student Analyze tech industry data to enhance career stability and minimize layoff risks while refining machine learning and analytical skills.

Literature Review



Mining Organizational Networks for Layoff Prediction Model Construction

Huo-Tsan Chang, Hui-Ju Wu , I-Hsien Ting

Article Link: https://ieeexplore.ieee.org/document/5231802

Goal: This paper aims to introduce the importance of the application of Data Mining and Social Network Analysis to predict layoffs through an empirical study.

Dataset: left employees' database of a semiconductor company at the Hsinchu Science Park Taiwan

Methodology: Social Network Analysis to analyze the relationship of the laid-off employee with a manager. Use of Data mining techniques to build a predictive model for layoffs in a company.

Result: Study finds out that the high compensation, high positions, high grades, and high education level are on the dangerous list of layoff.

Don't Fire Me, a Kernel Autoregressive Hybrid Model for Optimal Layoff Plan

Zhiling Luo, Ying Li , Ruisheng Fu, Jianwei Yin

Article Link: https://ieeexplore.ieee.org/document/7584978

Goal: This paper aims to build a model for an optimal layoff plan.

Dataset: A real dataset from the workflow system employed in the government of Hangzhou City in China, which results in 9750969 logs from 2050 activities and 16295 employees in two years.

Methodology: Regressing the activity throughput by the stuff number and inferring process throughput by the maximum flow or minimum cut algorithm on the Directed Acyclic Graph of the process. In the regressing step, a kernel autoregressive hybrid model is proposed. After that, an augmenting path-based algorithm is introduced to make an optimal layoff plan.

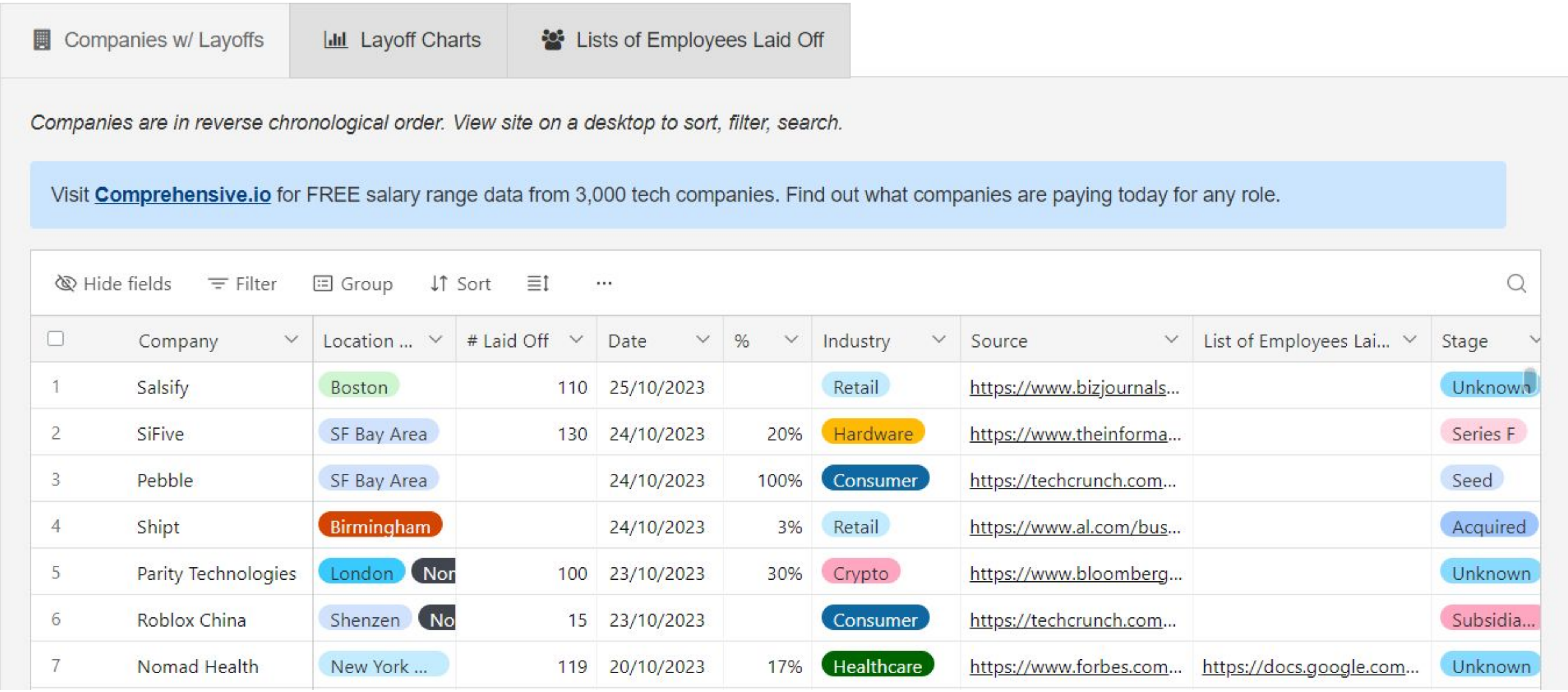
Result: The model can provide an optimal layoff plan with the least throughput loss, given the historical data. Since this study is unable to answer Who should be fired? , future work will focus on determining personal efficiency and firing those with the least efficient.

Dataset

Dataset Link: https://layoffs.fyi/ ,
https://www.aihr.com/blog/hr-data-sets-people-analytics/

Description: The above site data includes companies name ,location, employee list, roles etc for laid-off employees. It has data in csv, airtables and coda.io that i am going to combine and use for this project.

Size: 3101 rows ,9 columns.



	Company	Location ...	# Laid Off	Date	%	Industry	Source	List of Employees Lai...	Stage
1	Salsify	Boston	110	25/10/2023		Retail	https://www.bizjournals...		Unknown
2	SiFive	SF Bay Area	130	24/10/2023	20%	Hardware	https://www.theinforma...		Series F
3	Pebble	SF Bay Area		24/10/2023	100%	Consumer	https://techcrunch.com...		Seed
4	Shipt	Birmingham		24/10/2023	3%	Retail	https://www.al.com/bus...		Acquired
5	Parity Technologies	London Nor	100	23/10/2023	30%	Crypto	https://www.bloomberg...		Unknown
6	Roblox China	Shenzen No	15	23/10/2023		Consumer	https://techcrunch.com...		Subsidia...
7	Nomad Health	New York ...	119	20/10/2023	17%	Healthcare	https://www.forbes.com...	https://docs.google.com...	Unknown

EDA & Methodology

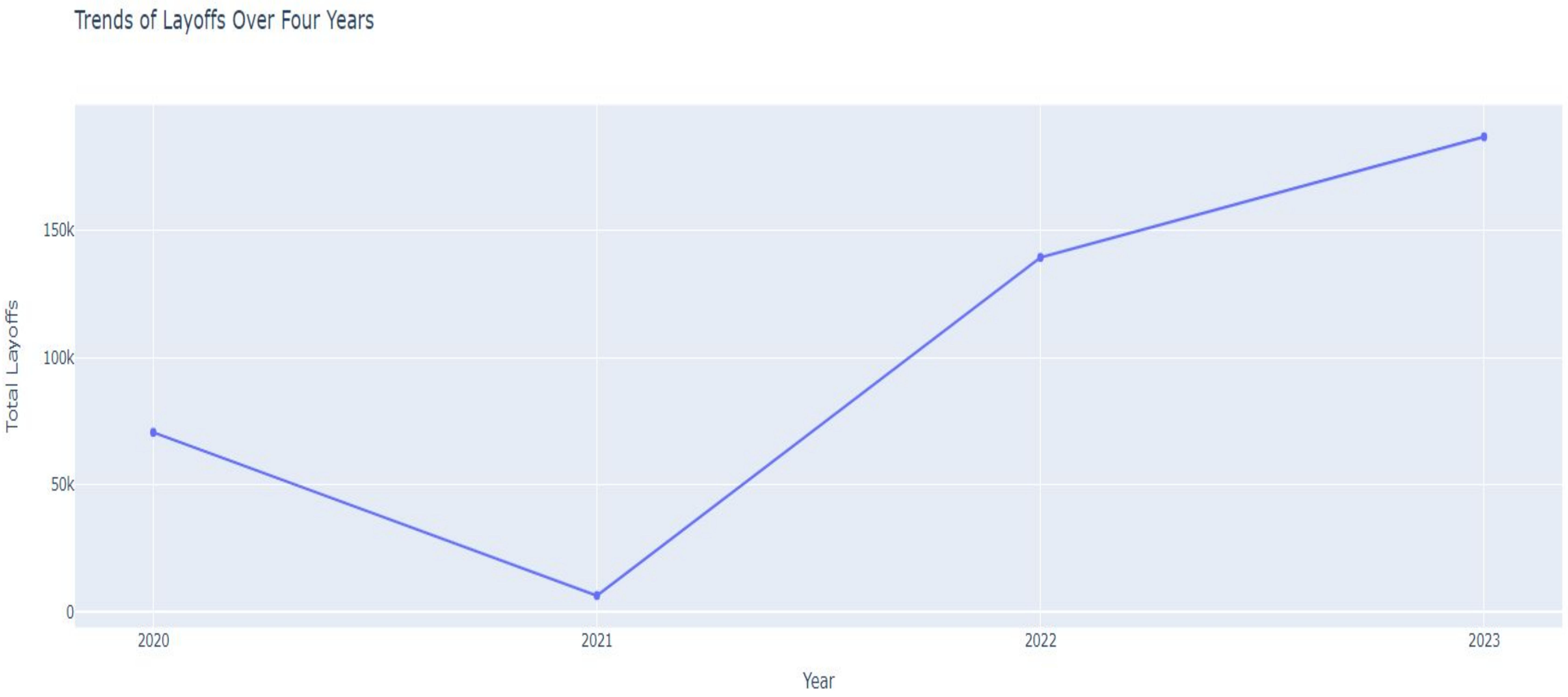
In the context of this project, an extensive Exploratory Data Analysis (EDA) was conducted to meticulously visualize and analyze the trends in layoffs across various dimensions, including countries, industries, and companies over the four-year duration. Subsequently, the acquired insights and findings have been presented in the following slides.

Methodology

- Dataframe creation
- Data cleaning includes dropping null values, check data types.
- Creating barplots, pie chart, scatter plot for insight visualization.
- Split the dataset into training and testing set.
- Handle class imbalance by SMOTE technique
- Train Random forest and XGBoost classification model to predict employee Layoff label.
- Evaluate performance of both the models using various metrics like Accuracy score, F1-score, confusion matrix.
- Interpret model prediction by using Eli5, LIME and SHAP.

What have been the trends in layoffs over the past four years?

- The chart indicates an upward trend in layoffs post-2021.
- The lowest point in the number of layoffs was observed in the year 2021.

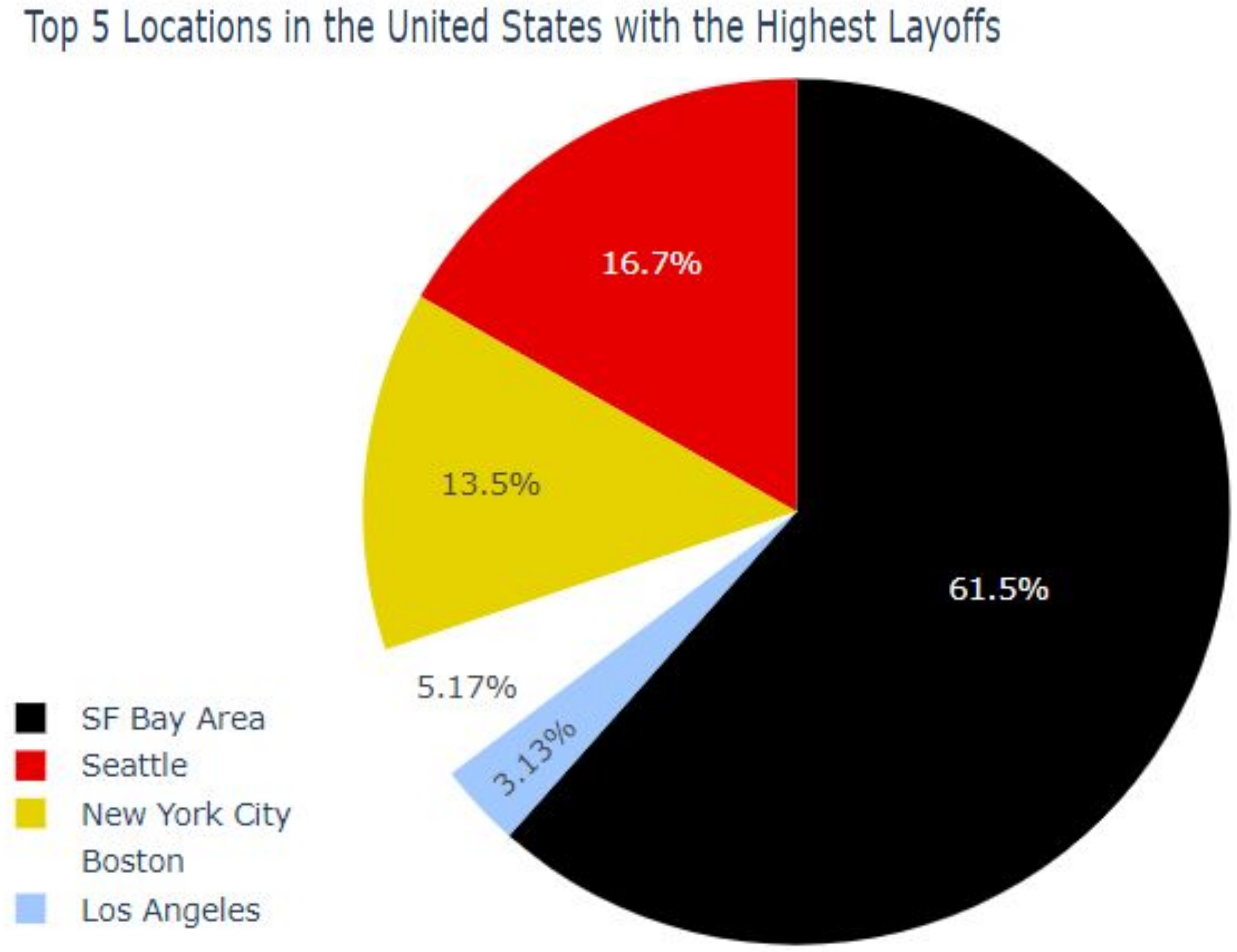


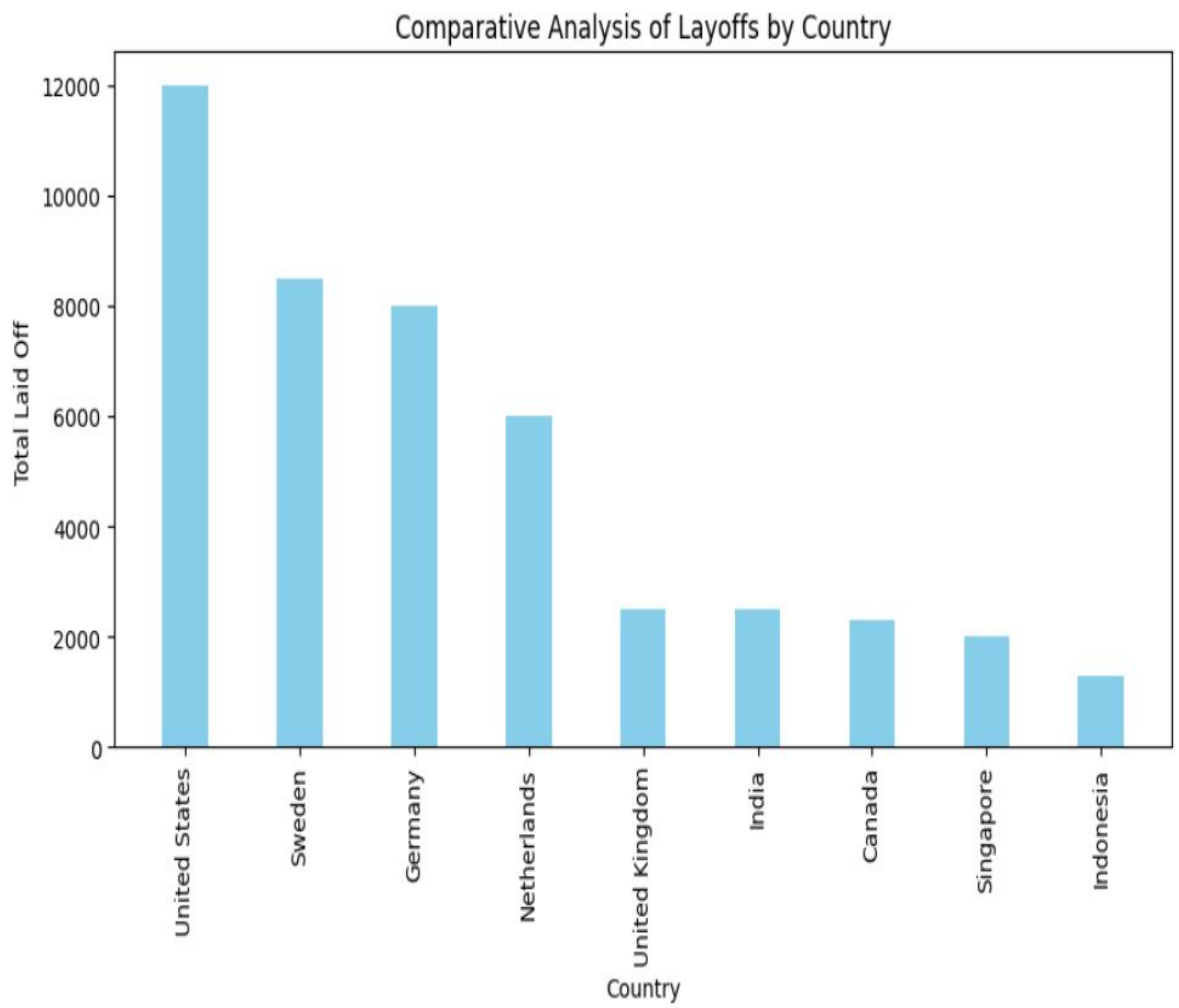
What have been the trends in layoffs across companies over the past four years?

- In 2022 and 2023, major tech firms such as Google, Amazon, and Meta have experienced significant layoffs.
- Notably, Uber was the sole company in 2020 to lay off more than 5000 employees.

Which countries faced most number of layoffs?

- The United States has the highest number of layoffs, followed by Sweden. Within the USA, San Francisco is the leading state in terms of layoffs.



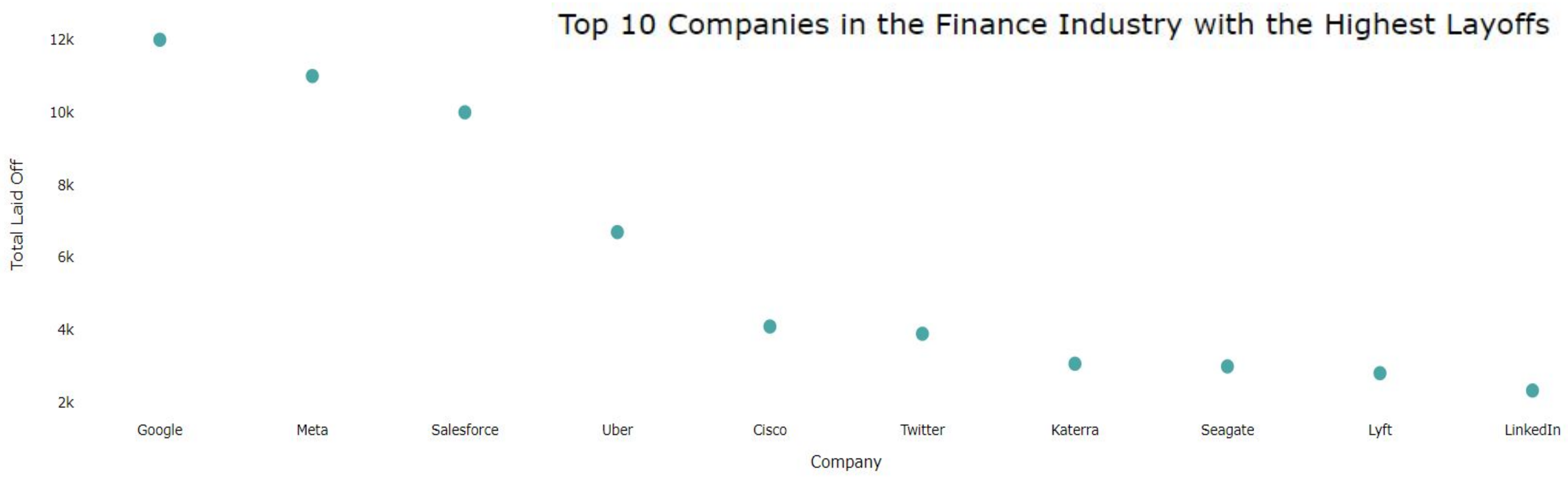


Which industries in the United States are more prone to layoffs?

- Real estate, finance, and marketing are the top three industries that have faced the highest number of layoffs in the United States.

Which companies have seen the most layoffs in the finance sector alone?

- Despite being massive technological companies, Google, Meta, and Salesforce have the highest percentage of layoffs in the finance industry.



Modeling

Job Layoffs Prediction with Random Forest & XGBoost classifier

Introduction:

- We've employed cutting-edge approaches like "Random Forest Classifier & XGBoost" to train our Job layoffs prediction system.
- Imagine this model as a super-smart virtual assistant who guide manager based on factors: **'Age', 'Job involvement', 'performance rating ', 'Years experience'' etc.**.

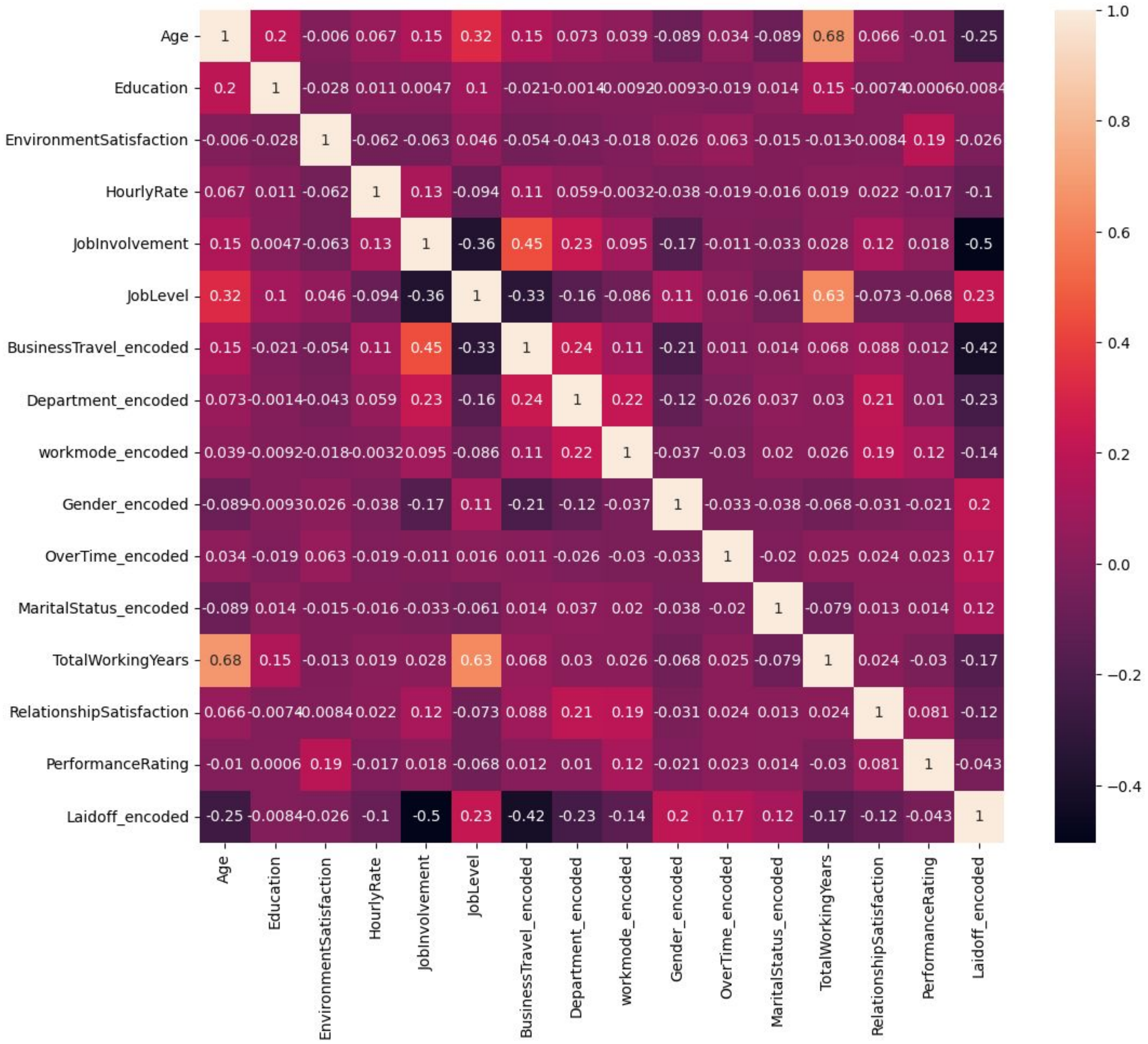
Why Random Forest and XGBoost classifiers?

Random Forest and XGBoost are both powerful ensemble learning techniques that excel in classification tasks. Here are some key reasons why these models are often chosen:

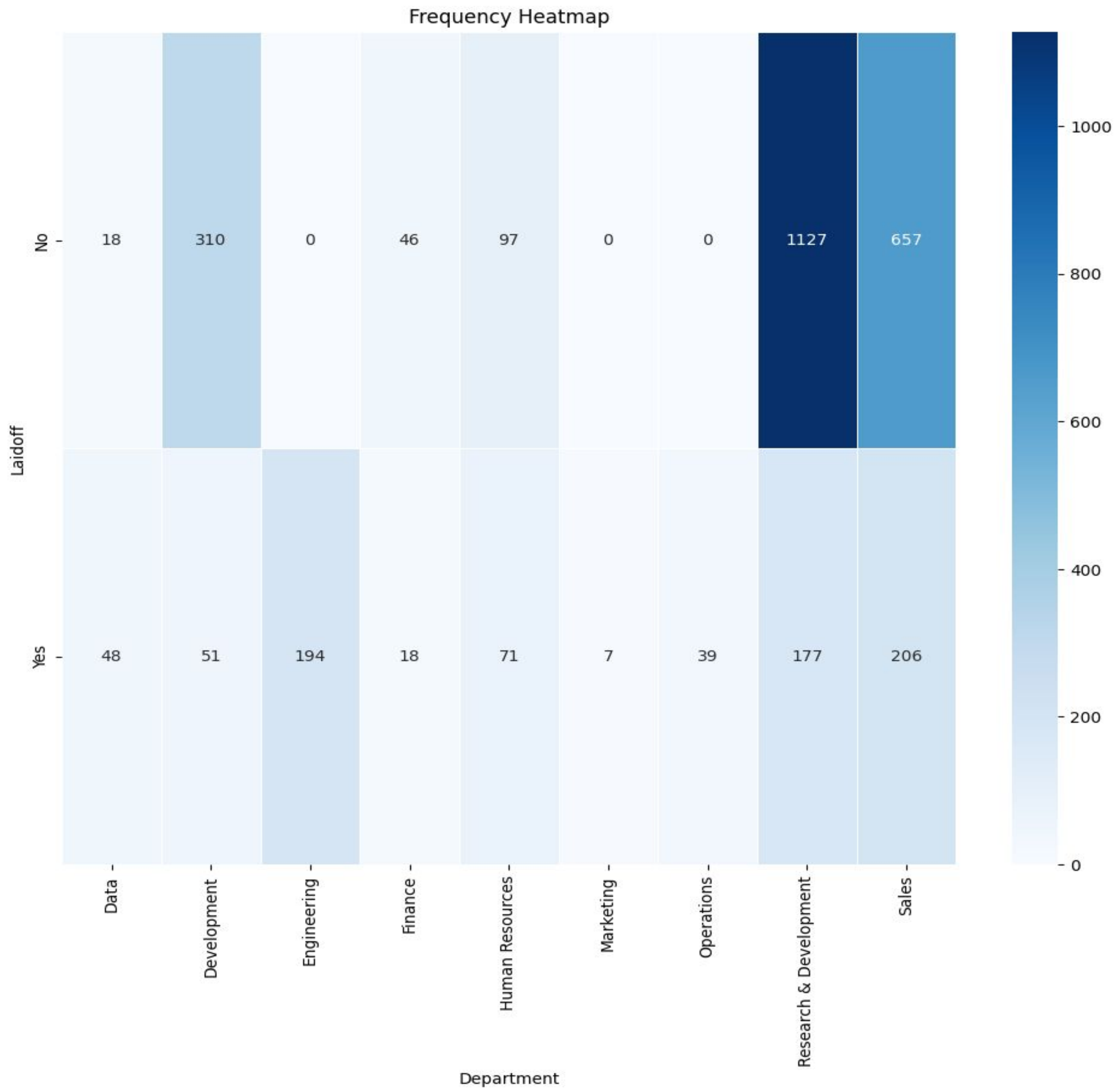
- Ensemble Learning:
  - Both Random Forest and XGBoost are ensemble methods, meaning they combine the predictions of multiple base models to improve overall performance. Ensemble methods often result in more robust and accurate models compared to individual models.
- Robustness and Accuracy:
  - Ensemble methods are known for their robustness and ability to handle noisy data and outliers. They often provide higher accuracy compared to individual models, making them suitable for complex tasks like job layoffs prediction.
- Reduction of Overfitting:
  - Random Forest and XGBoost include techniques like bagging and boosting, which help reduce overfitting. By aggregating predictions from multiple weak learners, these models are less prone to memorizing the training data and perform well on unseen data.
- Feature Importance:
  - Both models provide a feature importance score, allowing us to understand which features contribute the most to predictions. This can be valuable for interpreting the results and gaining insights into the factors influencing job layoffs.

Note: Random Forest and XGBoost share similarities, they have differences in terms of their underlying algorithms, training processes, and specific strengths. That's why we are training these two models so that we can compare both the model performances.

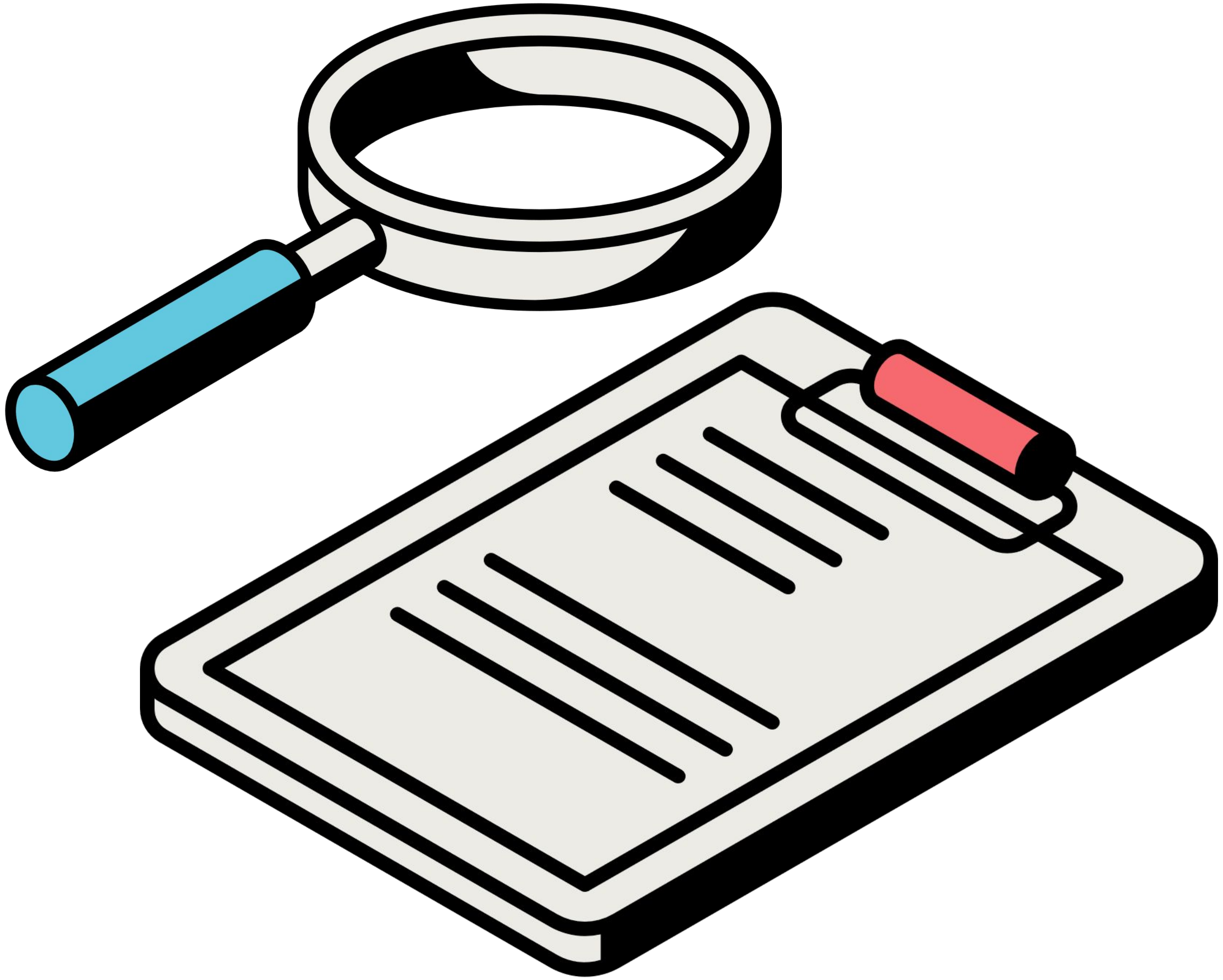
- Age, Job involvement , business travel shown the most negative correlation with our target variable.
- Gender and overtime has shown some weak positive correlation with Job layoff.



- From the frequency heatmap we can see Engineering department is most prone to layoffs.
- R&D and Sales are most safest departments in terms of layoffs.



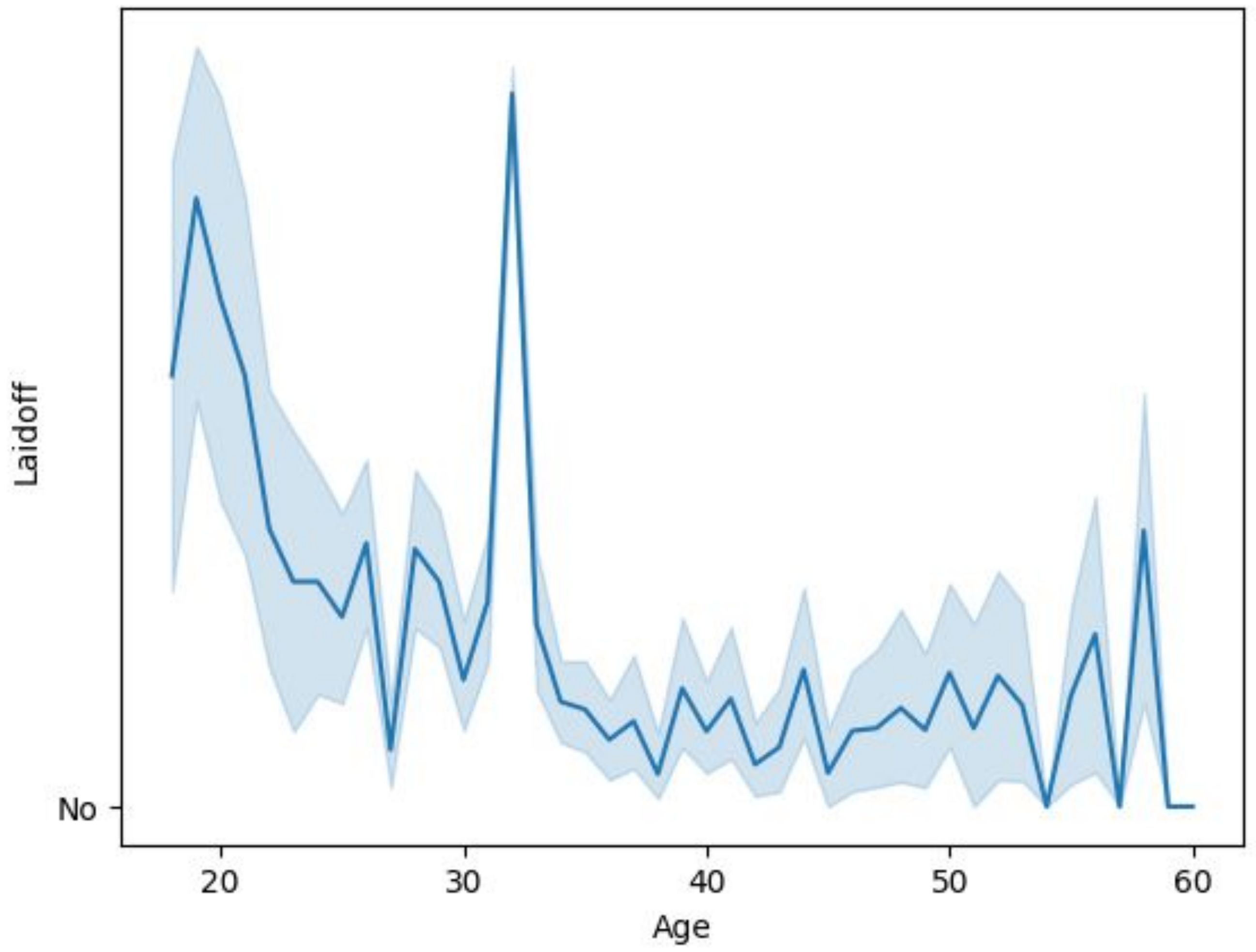
Findings & Results



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:

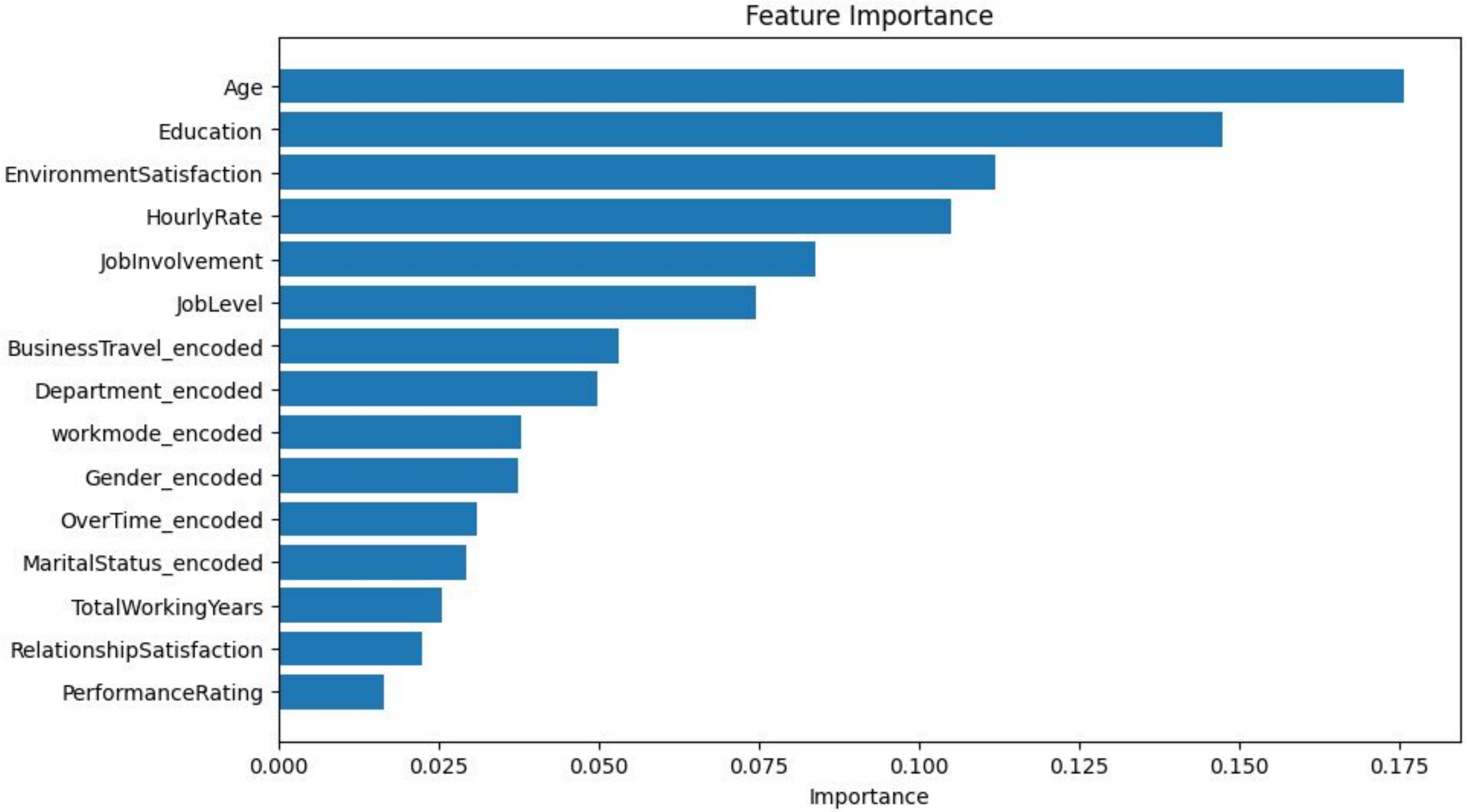
- The highest number of layoffs occurred in 2023 over the past four years.
- The United States experienced the highest number of layoffs compared to any other country.
- Sectors most susceptible to layoffs in the United States: Finance, Real Estate, and Marketing.
- Top three companies with the most 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.



After training the Random forest classification model to predict employee layoffs, we identified the most influential features in the model.

Key Features:

- Most important features are Age & Education to predict the label of laid off employee.
- Layoffs also affected by hourly wages of an employee.
- Performance rating contributes the least in prediction of layoff status.



Comparison between XGBoost & Random forest

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:

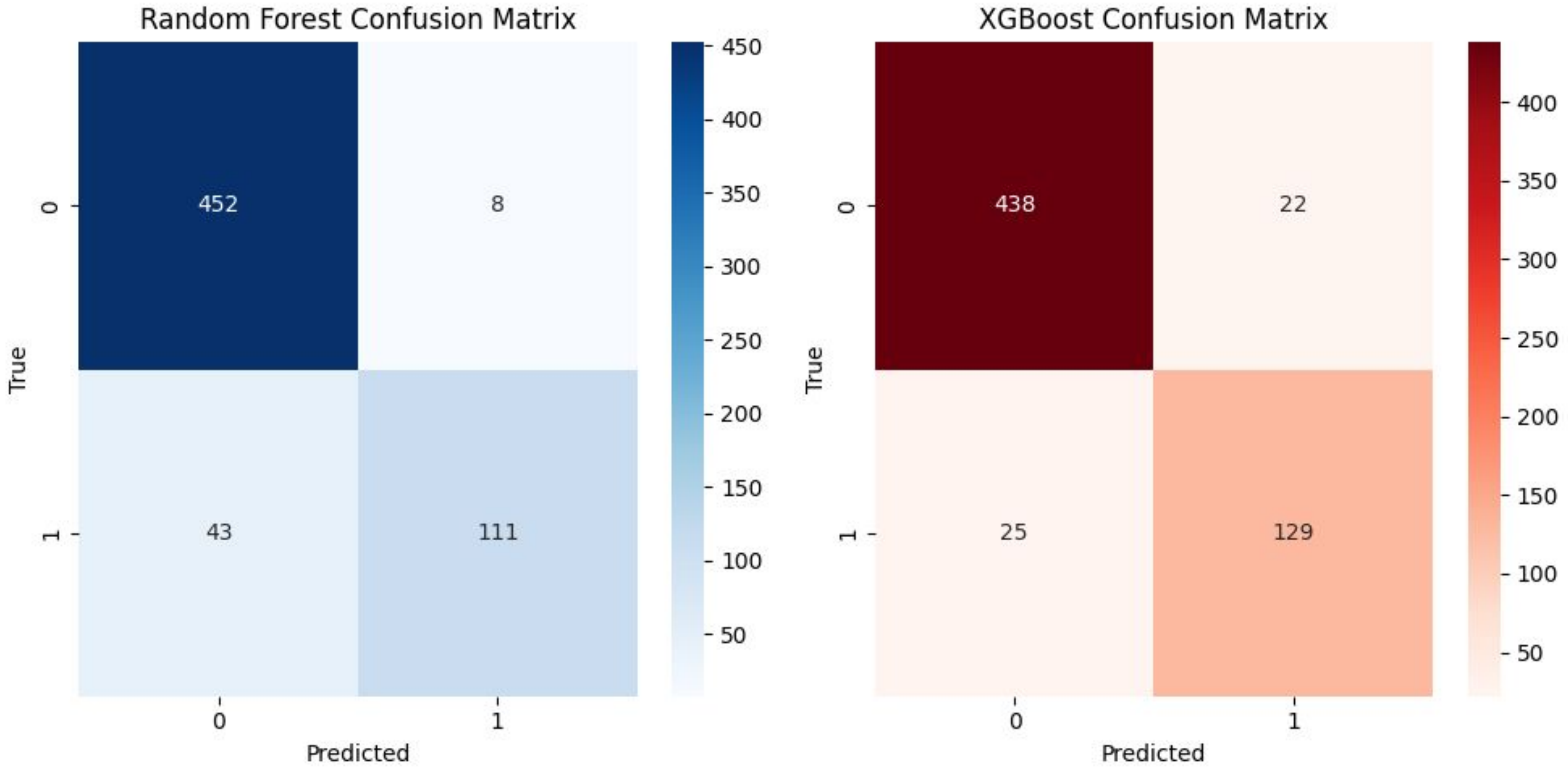
Classification Report Random Forest:

	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

Findings

We have interpret our models prediction by using Eli5, SHAP & LIME.

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.

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

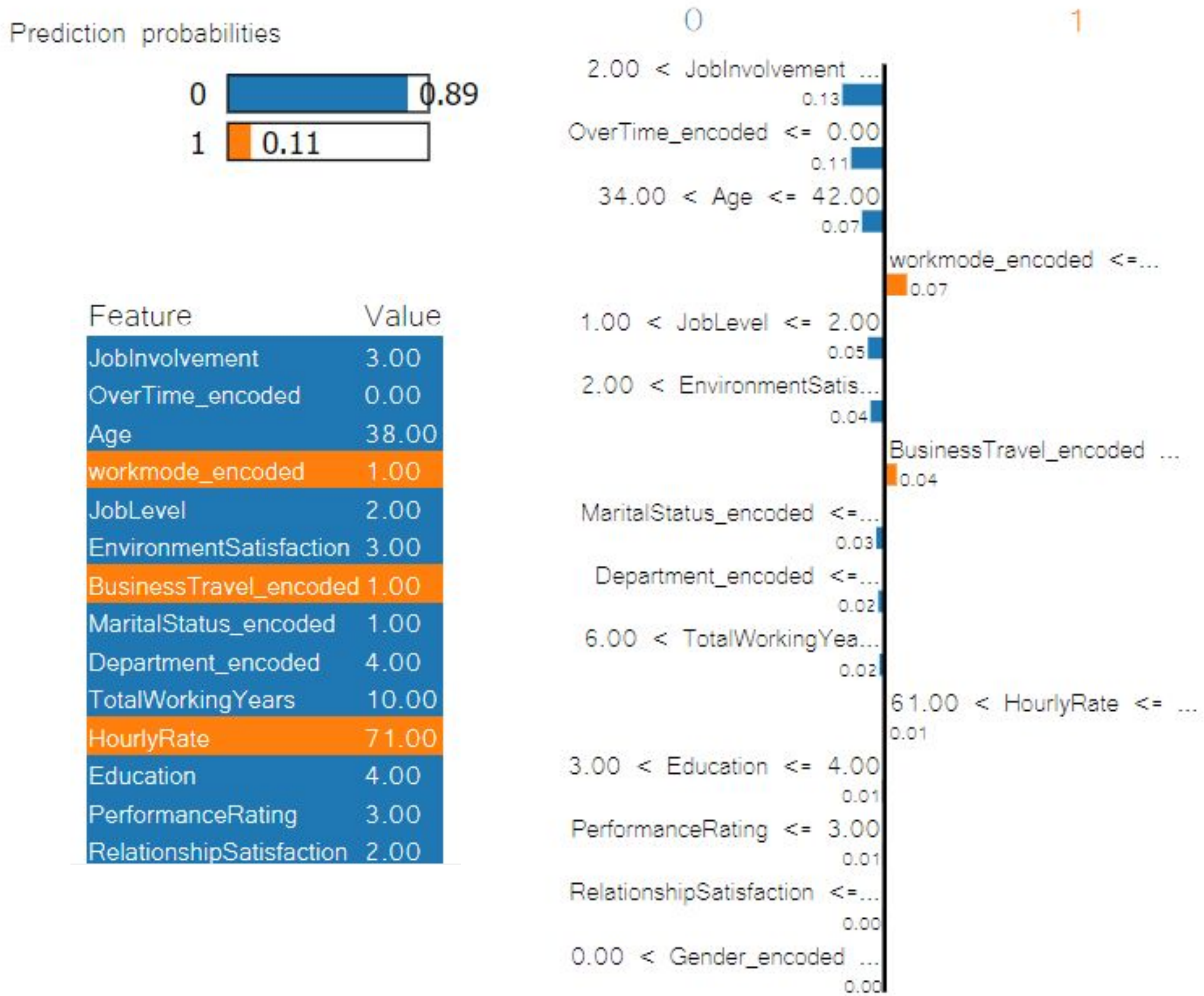
y=0 (probability **0.860**) top features

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

Continue..

- Lime Showing the weightage of job involvement is the highest in the prediction of class label 0 (no layoff).
- Gender has no contribution to predict any of the class.
- Workmode has the high weightage toward positive class prediction.



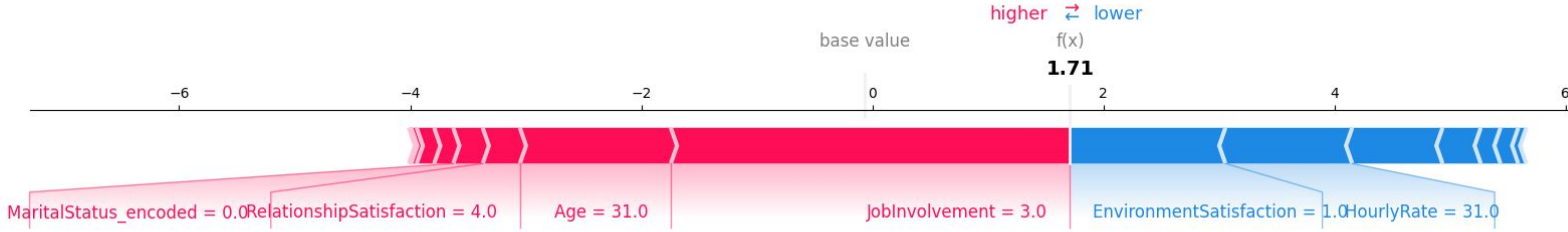
Feature	Value
JobInvolvement	3.00
OverTime_encoded	0.00
Age	38.00
workmode_encoded	1.00
JobLevel	2.00
EnvironmentSatisfaction	3.00
BusinessTravel_encoded	1.00
MaritalStatus_encoded	1.00
Department_encoded	4.00
TotalWorkingYears	10.00
HourlyRate	71.00
Education	4.00
PerformanceRating	3.00
RelationshipSatisfaction	2.00

Continue..

- SHAP force plot showing the influence of each factor in predicting a positive and negative class.
- Feature in Red forcing model prediction toward positive class i.e. 1 while feature in blue forcing prediction toward negative class i.e. 0.

Positive label

Negative Label

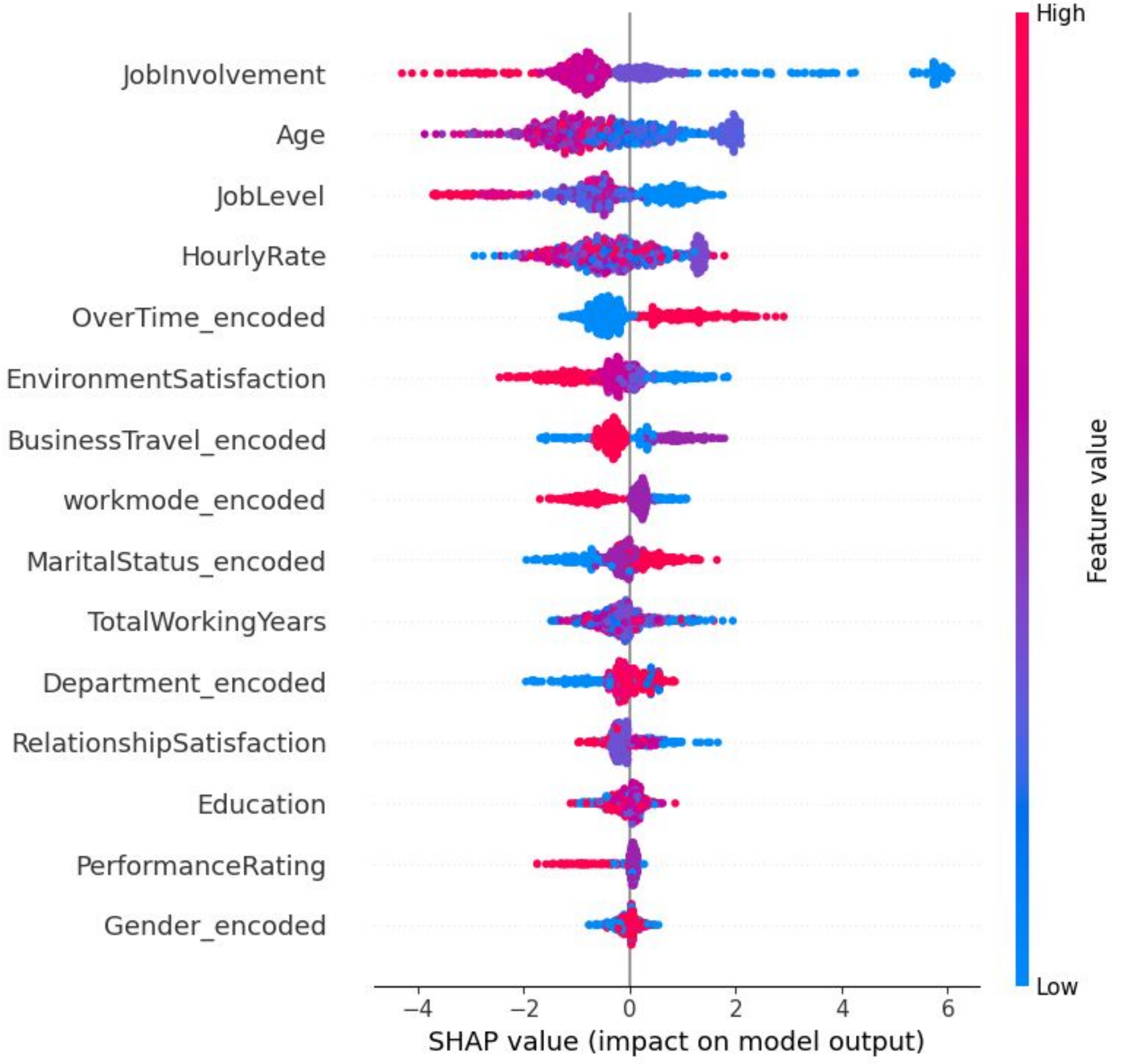


Continue..

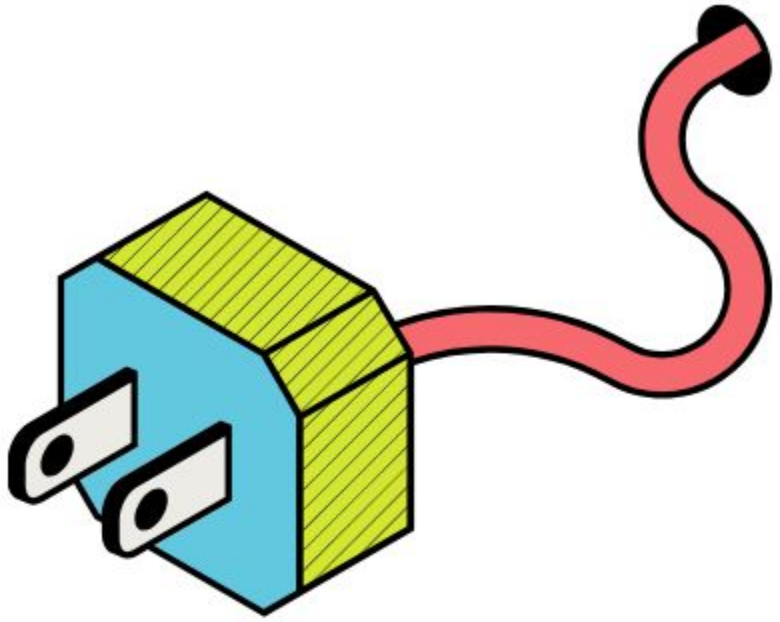
Interpreting SHAP Summary for XGBoost Model

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.



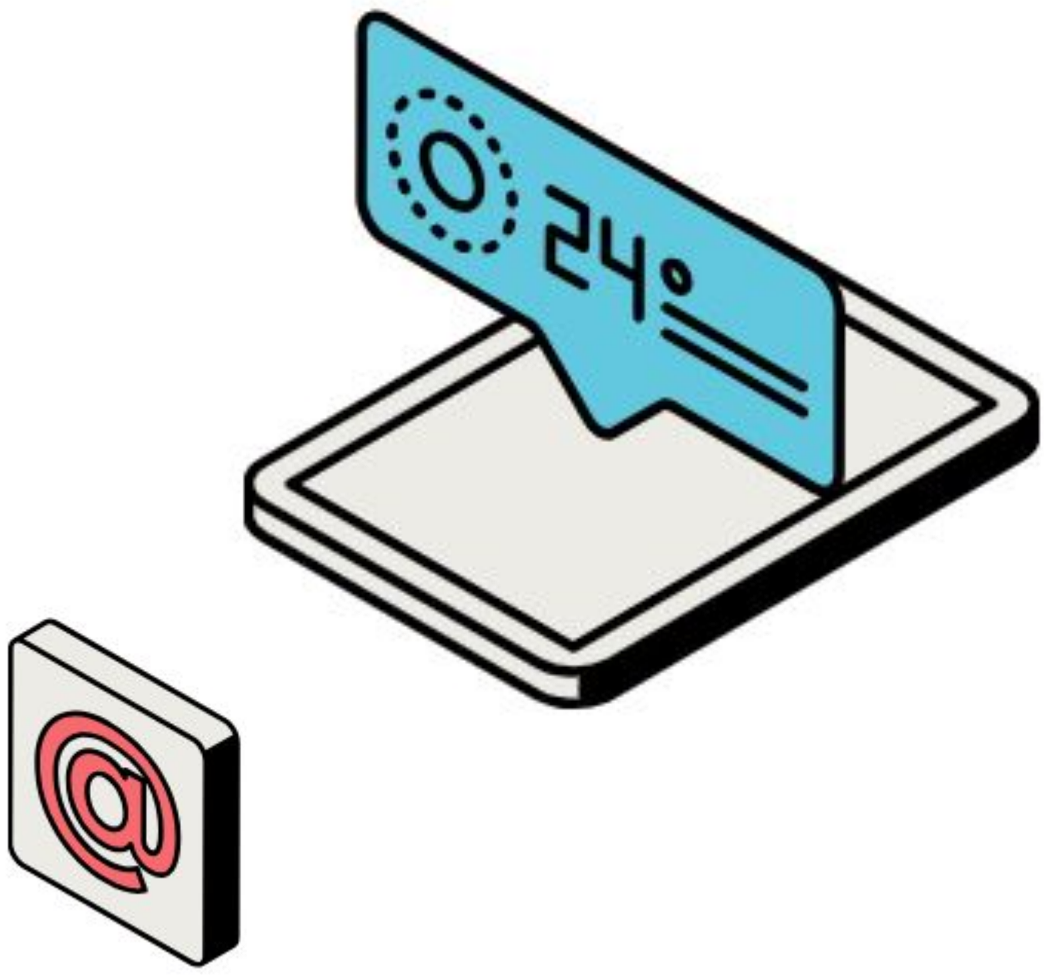
Limitations



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

Conclusion & 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.



Thank you.