ALDI
VIBRIANI

Improving Employee Retention by Predicting Employee Attrition

Aldi's Data Science Portofolio

Agenda Overview

01 Background

02 Business Insight

03 Exploratory Data Analysis

04 Data Preprocessing

05 Model Machine Learning

06 Business Recommendation

Background

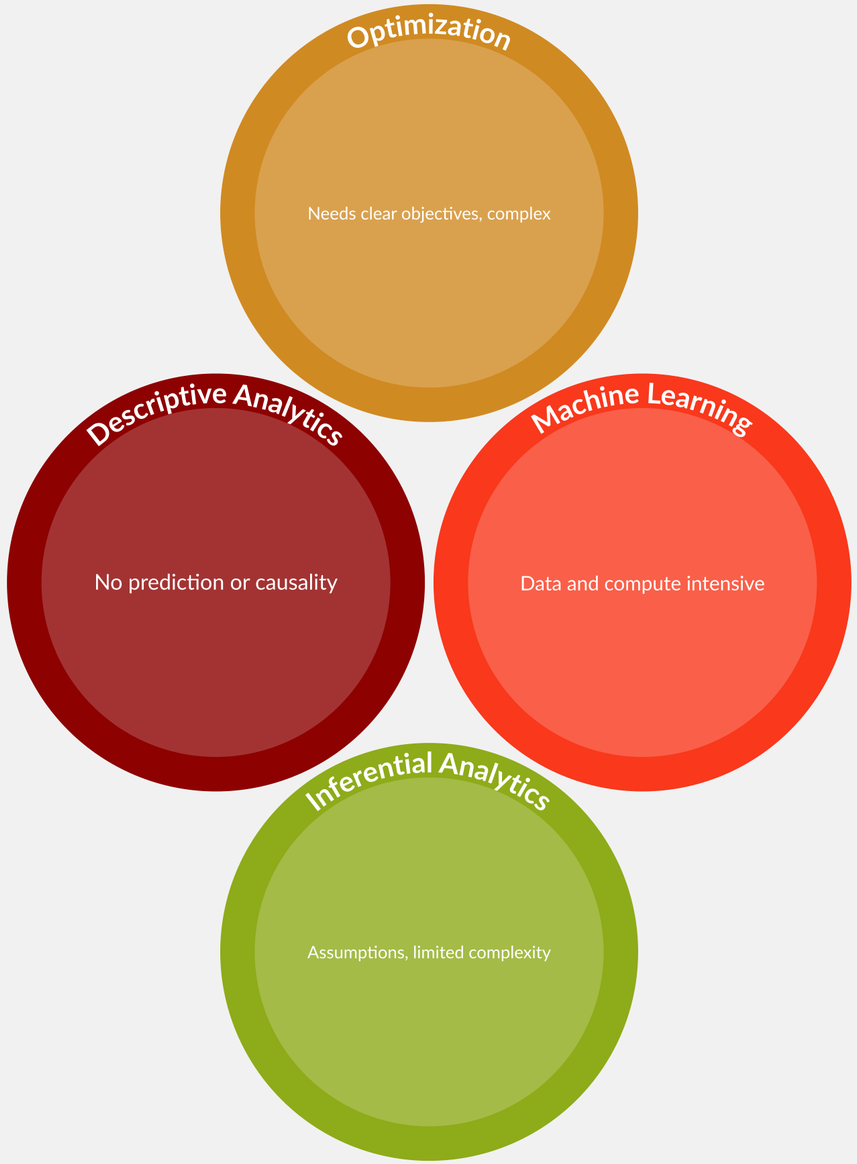
Background

A technology start-up is currently facing a significant challenge with a **high rate of employee resignations**. Despite the growing number of departures, the company has yet to implement effective measures to address this issue. This project aims to analyze the current state of the workforce, identify key factors contributing to employee attrition, and propose strategies to enhance employee retention. Through the use of descriptive data visualization and storytelling, combined with inferential analysis employing statistical methods and machine learning frameworks such as interpretable and explainable AI, the findings will be translated into actionable insights to support the company’s efforts in reducing turnover and fostering workforce stability.

There are several methods to solve this problem

Solution Type	Purpose	Key Techniques	Strengths	Limitations
Descriptive Analytics	Understand past and current data	Reporting, visualization	Simple, foundational	No prediction or causality
Machine Learning	Predict and classify data	Algorithms (e.g., Random Forest)	Handles complexity, scalable	Data and compute intensive
Inferential Analytics	Make inferences about populations	Hypothesis testing, regression	Statistical rigor, causality	Assumptions, limited complexity
Optimization	Find best decisions/solutions	Linear programming, heuristics	Actionable, handles constraints	Needs clear objectives, complex

Limitation in each solution



Machine Learning delivers greater impact than other solutions because it can model complex patterns and relationships in data, enabling accurate predictions and actionable insights despite its higher data and computational demands.

Four study shows that Machine Learning is Better than others

Traditional vs Machine Learning Approaches

Written by: Sadler, B.
This paper highlights machine learning provide , ata accuracy and handle complex data pattern more effectively than traditional model pattern, which are often limited to univariate and linear data.

Comparison of Conventional Statistical Methods with Machine Learning Techniques in Predictive Analytics

Written by: Rajula, H. S. R., Verlato, G., et al.
The paper highlights that ML techniques are more flexible and free from strict assumptions required by traditional statistical methods, leading to better performance in predictive tasks.

Machine Learning Research

Applying Machine Learning to Human Resources Data

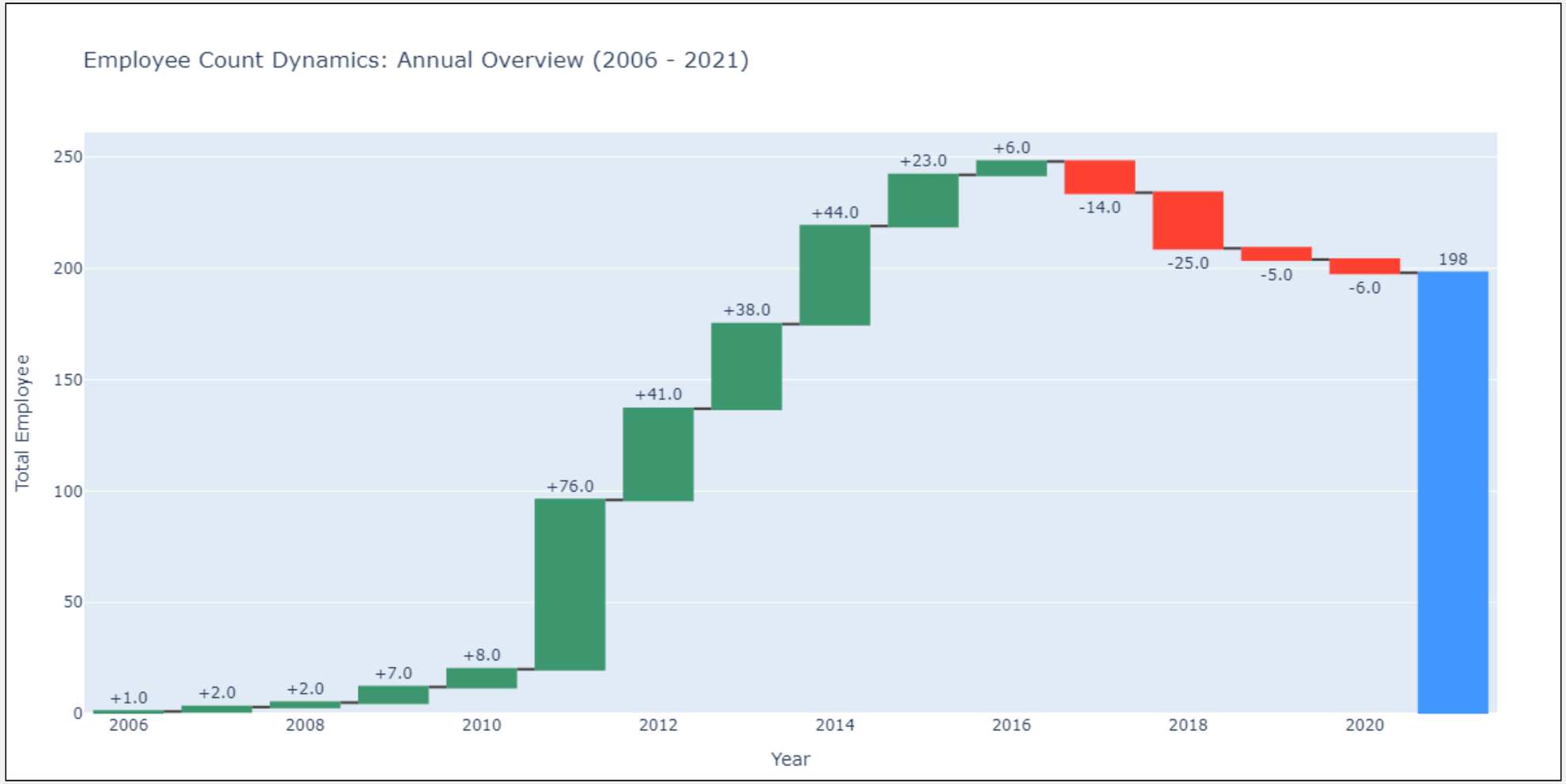
Written by: Fukui, S., Wu, W., Greenfield, J. et al.
This study demonstrates that ML models, particularly random forests, achieve high predictive accuracy (AUC > 0.8) in forecasting employee turnover, outperforming traditional regression models. It also identifies key predictors of turnover using ML.

Machine Learning Models for Predicting Employee Retention and Performance

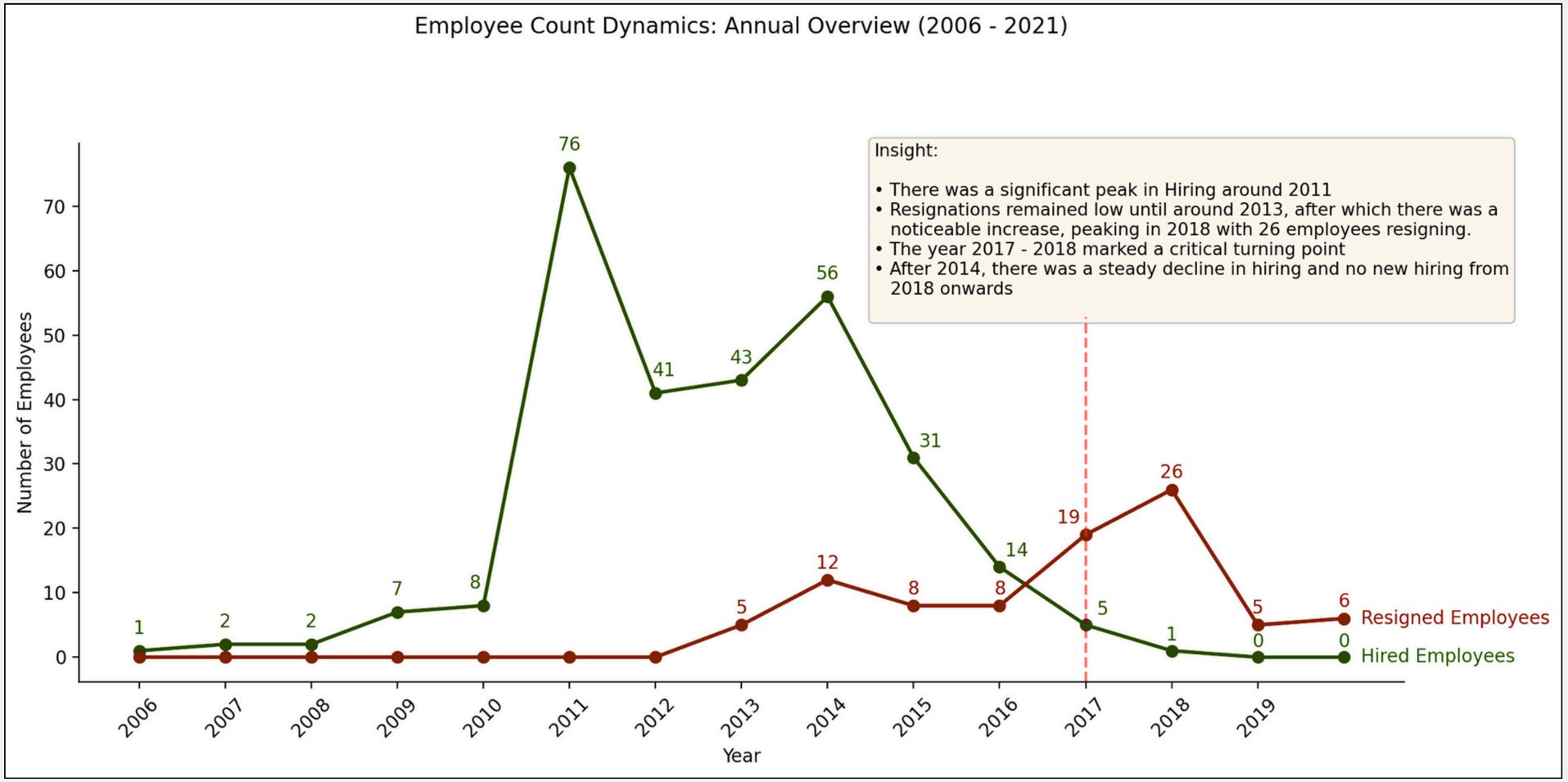
Written by: Nalla, N.R.
This study reports accuracy rates of various ML models in employee turnover prediction, with random forests achieving 84% accuracy and neural networks up to 91%, outperforming logistic regression and SVM models. It also notes that behavioral data integration improves retention outcomes.

Business Insights

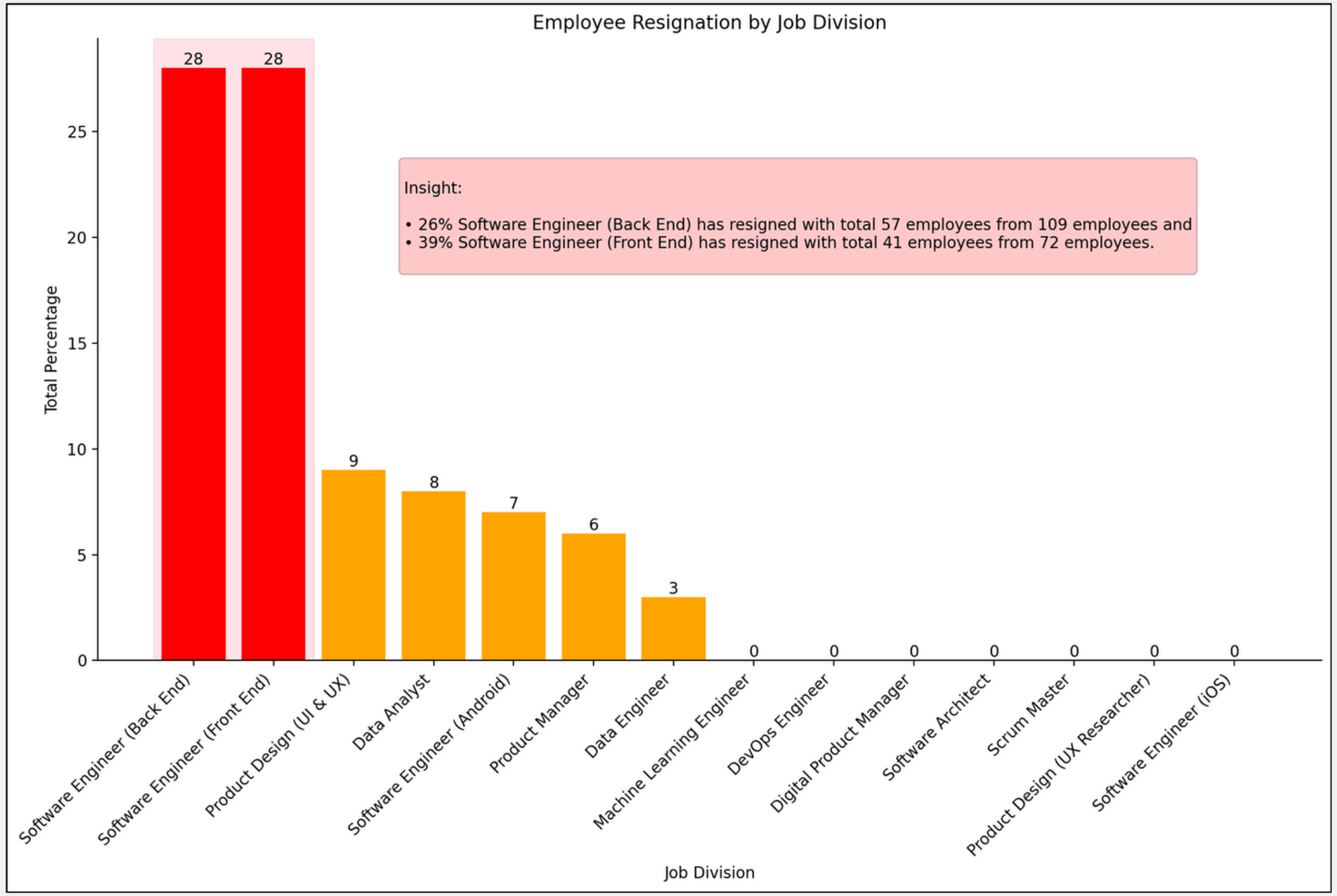
The overall employee count increased steadily from 2006 to 2016, reaching a peak, followed by a gradual decline from 2017 to 2021, ending with a total of 198 employees.



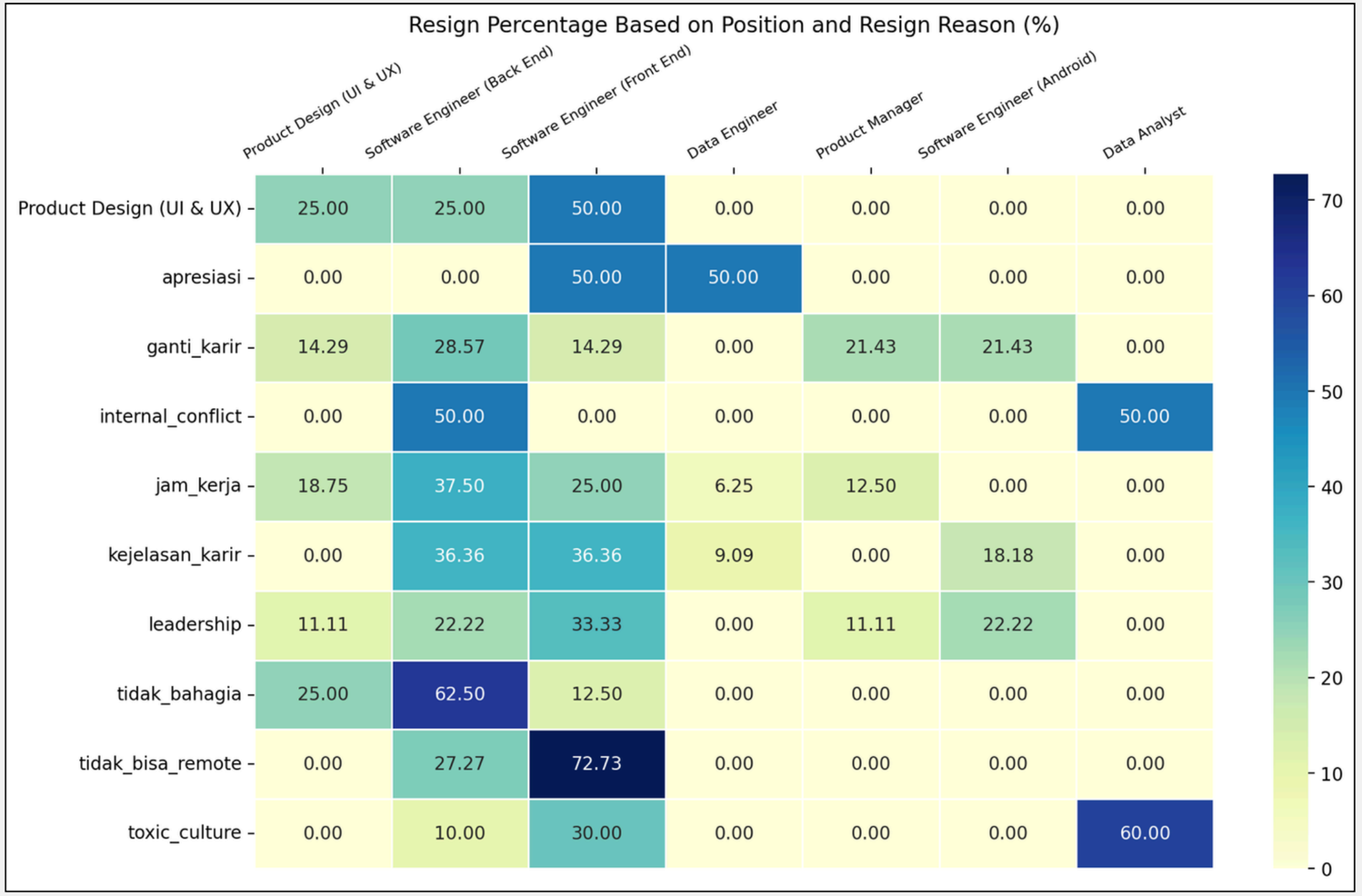
The year 2017 marks a critical turning point where employee resignations began to surpass hiring, signaling urgent need for strategic intervention to address workforce retention and prevent further talent loss.



The highest employee resignation rates occurred in the Software Engineer roles, with 26% of Back End and 39% of Front End engineers leaving, indicating a critical retention challenge in these key engineering divisions.

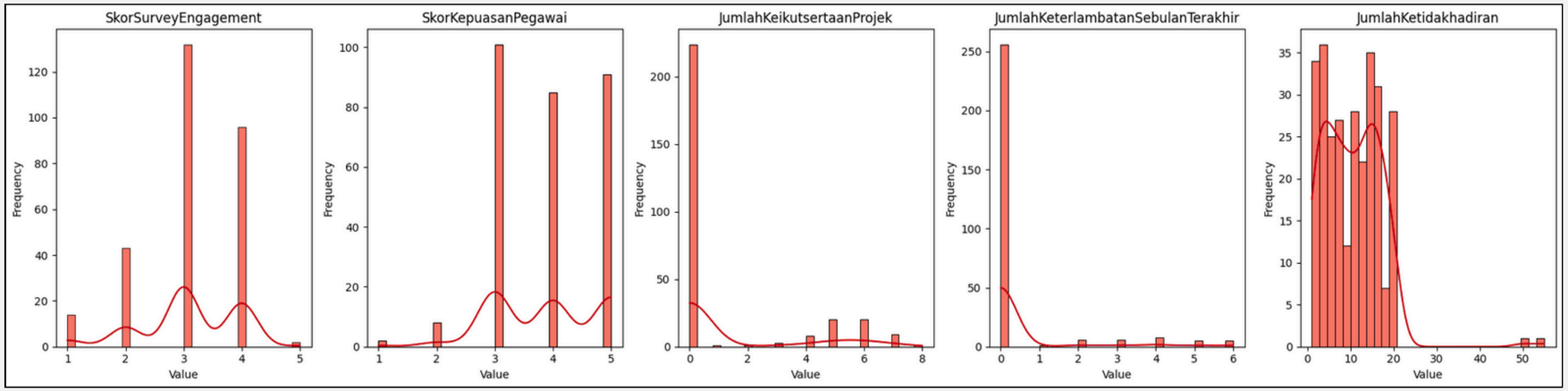


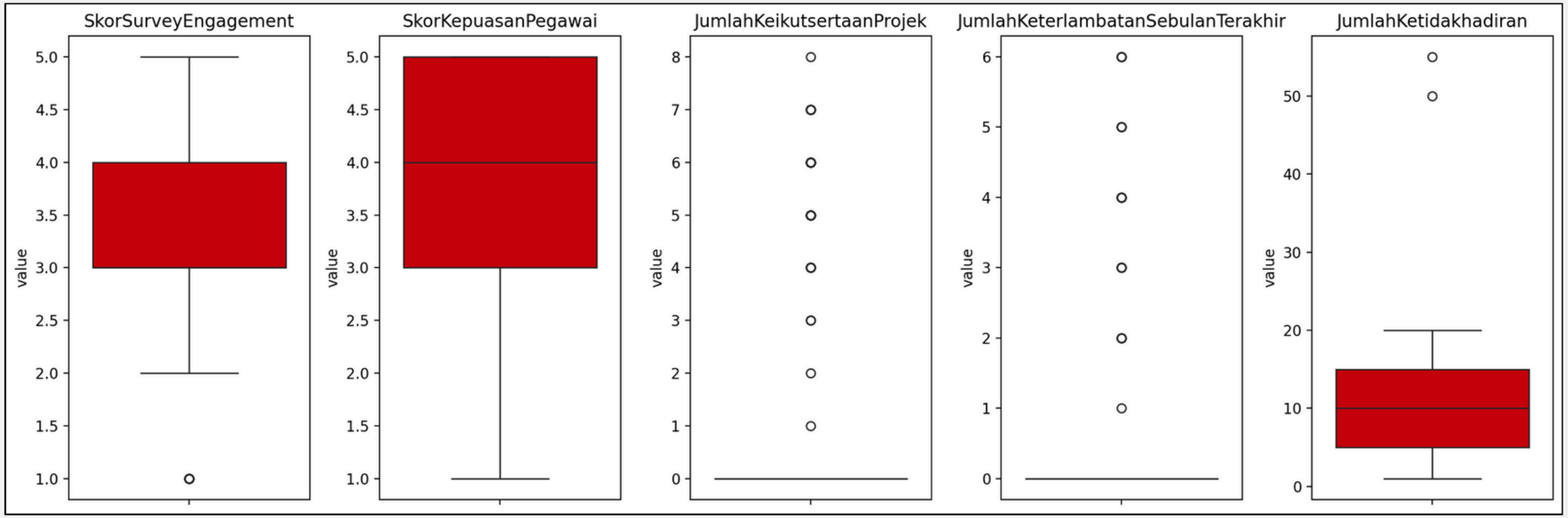
The resignation results are predominantly driven by dissatisfaction factors, highlighting critical areas for organizational improvement to enhance employee retention.

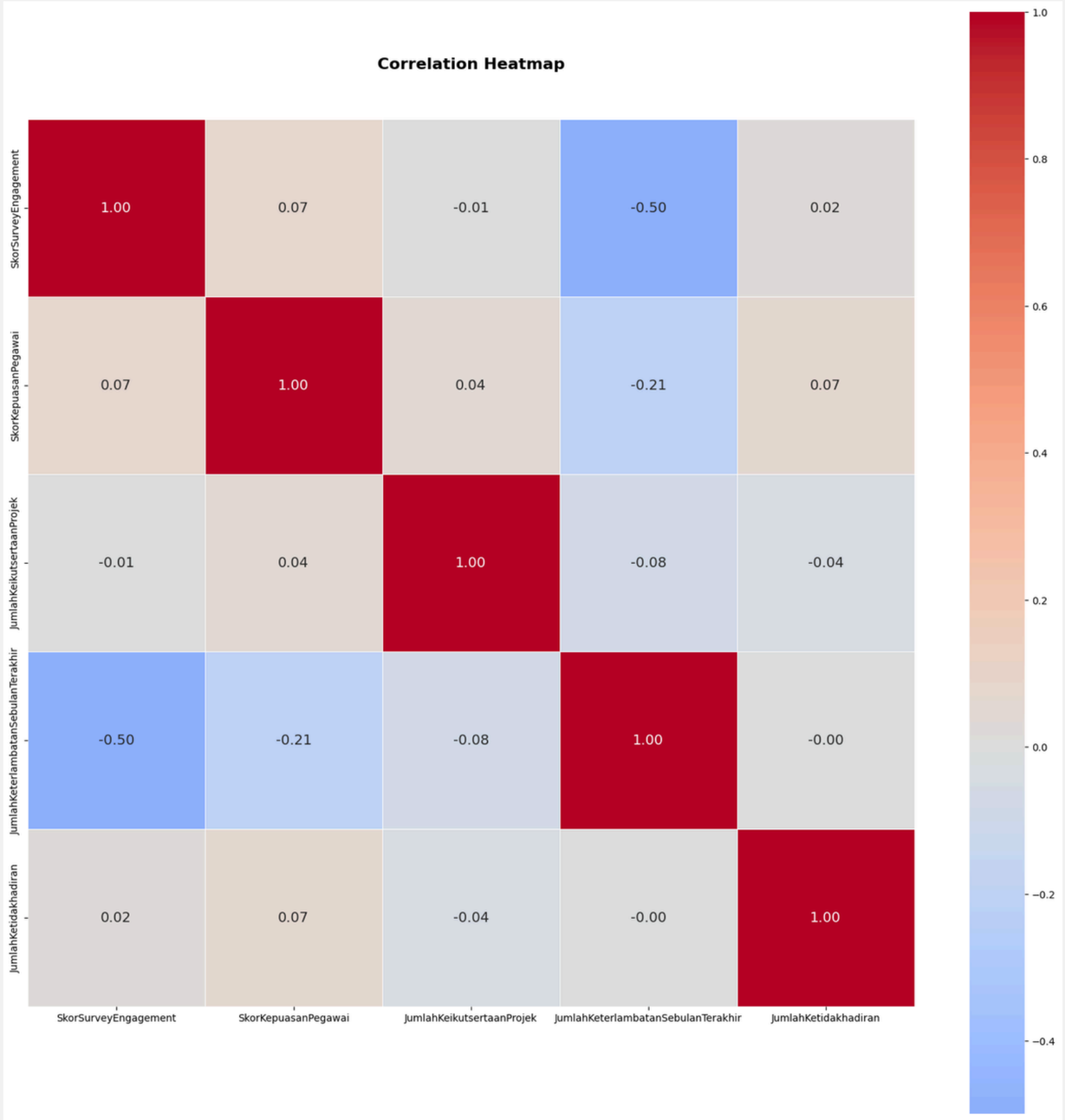


Exploratory Data Analysis

Most of the employees does not take part of project. Then, there are an extreme outliers where some employee absence more than 50 times.







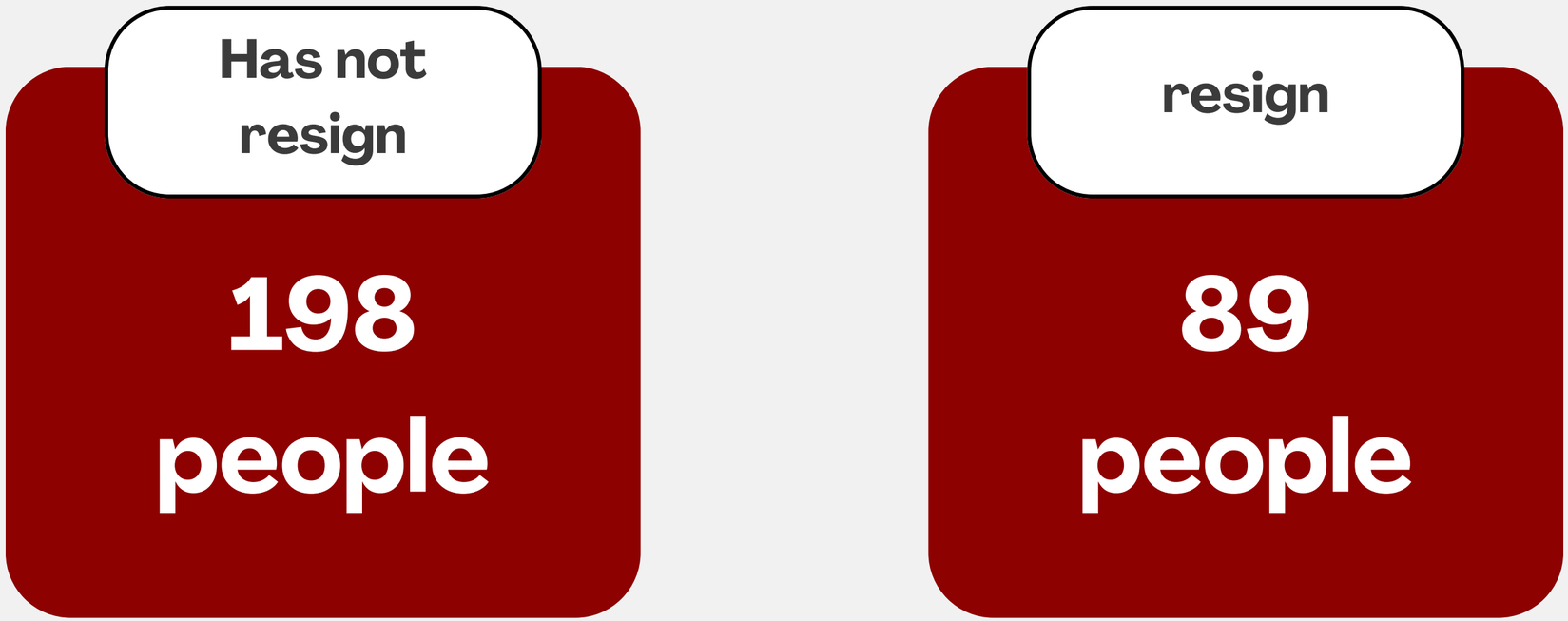
“Jumlah Keterlambatan Sebulan Terakhir” feature shows a negative correlation with “Skor Survey Engagement” (-0.50) and “Jumlah Keikutsertaan Projek” (-0.21).

Data Preprocessing

Feature Engineerning

We want to predict employee resignation so we need a label consist of employees who had resign and not.

“is_resign” feature

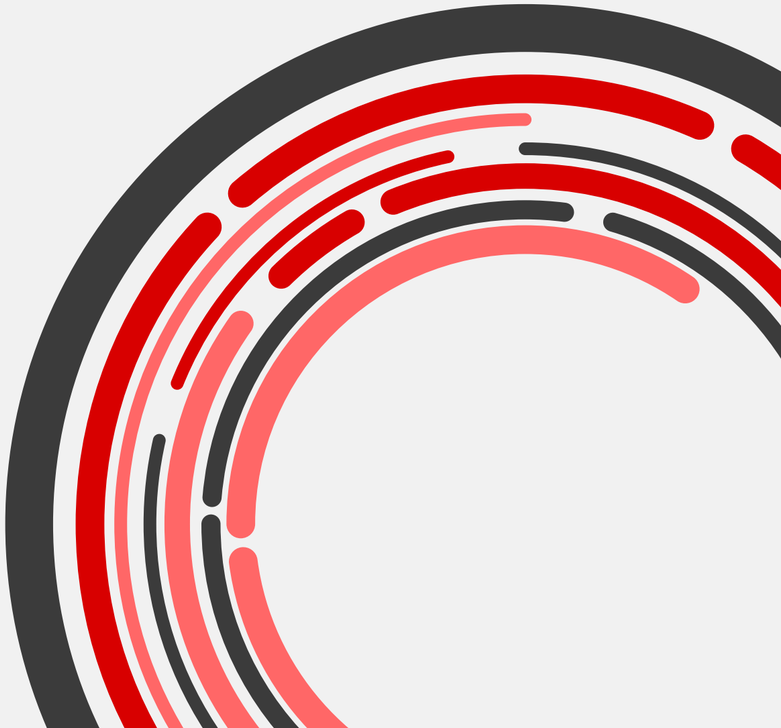


Feature Engineerning

We also extract and create a new features:

- **keikutsertaanproject_boolean**
- **Tahun_Hiring**
- **Tahun_Resign**
- **lama_bekerja**
- **usia_hired**
- **jarak_penilaian_tahun**

Then, we dropped **User name, Nomor Hp,** and **Email** feature because these features has many unique values and does not give much information for machine learning.



Data Cleaning

Missing Values

There is no missing value found in each feature

Duplicates Data

No duplicates data found

Outliers

There are many features have outliers but only “**JumlahKetidakhadiran**” and “**Jarak_penilaian_tahun**” that being handled using IQR

Feature Encoding

There are two methods in handling categorical feature:

- **Label Encoding:**
  - TingkatPendidikan
  - JenjangKarir
  - PerformancePegawai
- **One-hot-encoding:**
  - StatusPernikahan
  - AsalDaerah
  - Pekerjaan
  - StatusKepegawaian

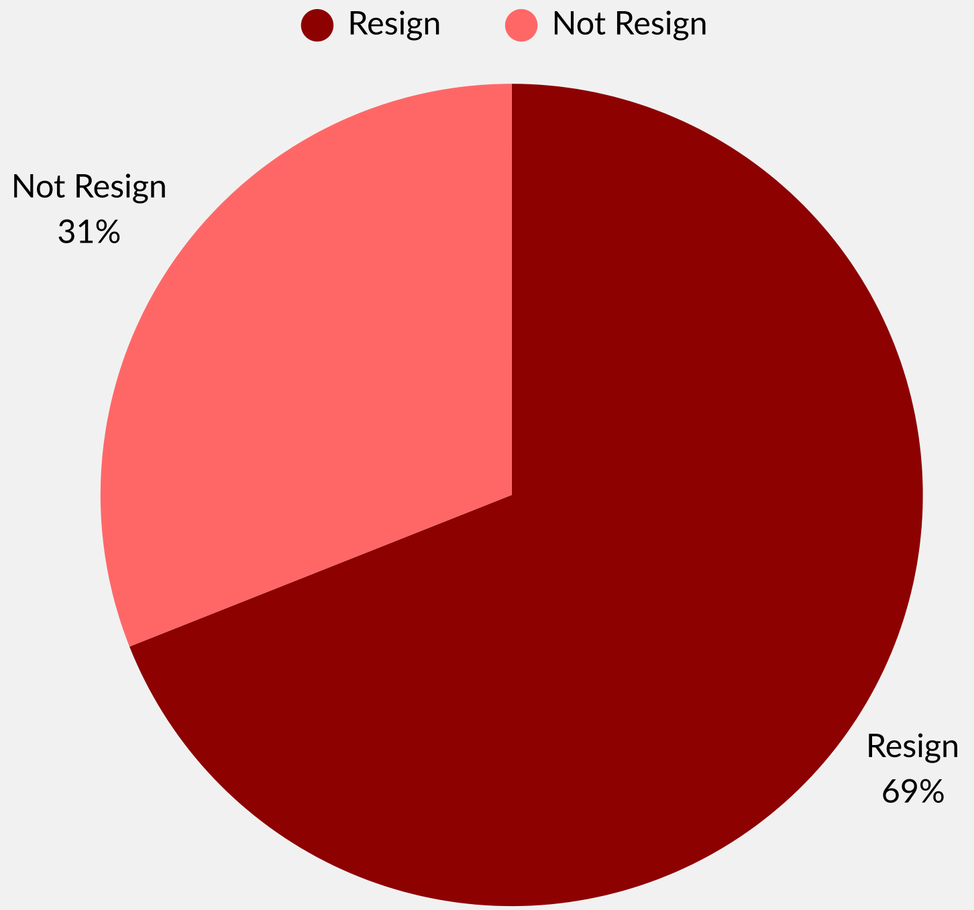
Model Machine Learning

Handling Imbalance Class

To avoid bias on the majority values, we tested to find the best handling imbalance strategy

	Imbalance_Strategy	Accuracy	Precision	Recall	AUC	Training_Time
0	TomekLinks	0.950237	0.958506	0.927024	NaN	00:00:05
1	RandomUnderSampler	0.918775	0.907649	0.905327	NaN	00:00:00
2	EditedNearestNeighbours	0.906601	0.896724	0.899226	NaN	00:00:00
3	SMOTEENN	0.931779	0.931899	0.915506	NaN	00:00:00
4	SMOTETomek	0.945810	0.951624	0.923899	NaN	00:00:00

The result shows **SMOTENN** (combination SMOTE and Edited Nearest Neighbours) give a the best results and performance

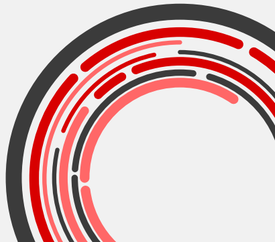


is_resign feature

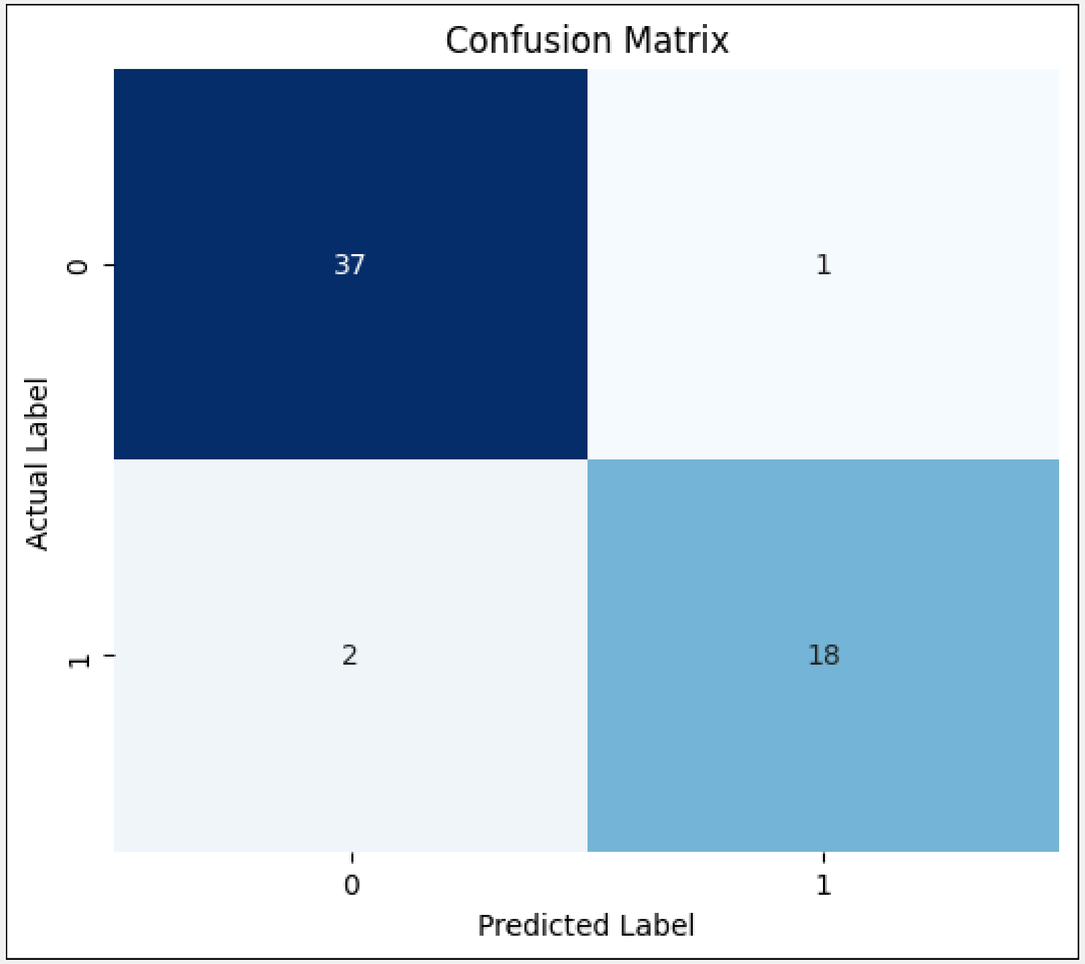
Machine Learning Model Test

	ML_Model	Accuracy	Precision	Recall	AUC	Training_Time
12	CatBoostClassifier	0.940250	0.943156	0.919940	0.939583	00:00:16
11	LGBMClassifier	0.934519	0.935699	0.915774	0.937426	00:00:00
0	GaussianNB	0.943215	0.947360	0.924702	0.934177	00:00:00
2	RandomForestClassifier	0.943281	0.952363	0.918006	0.932143	00:00:00
5	BaggingClassifier	0.929974	0.923730	0.916667	0.929241	00:00:00
6	GradientBoostingClassifier	0.927009	0.922139	0.910565	0.923363	00:00:00
3	LogisticRegression	0.922727	0.929932	0.889236	0.914236	00:00:00
1	SVC	0.854348	0.850523	0.828224	0.903571	00:00:00
4	DecisionTreeClassifier	0.912516	0.904749	0.900149	0.900149	00:00:00
7	AdaBoostClassifier	0.905270	0.893450	0.896280	0.896280	00:00:00
9	MLPClassifier	0.852372	0.870499	0.831597	0.880456	00:00:00
8	KNeighborsClassifier	0.711594	0.686952	0.711409	0.770585	00:00:00
10	XGBClassifier	0.934585	0.936963	0.915774	NaN	00:00:00

By comparing the metrics and training time, **GaussianNB** shows the best results. Thus, we will use this model.



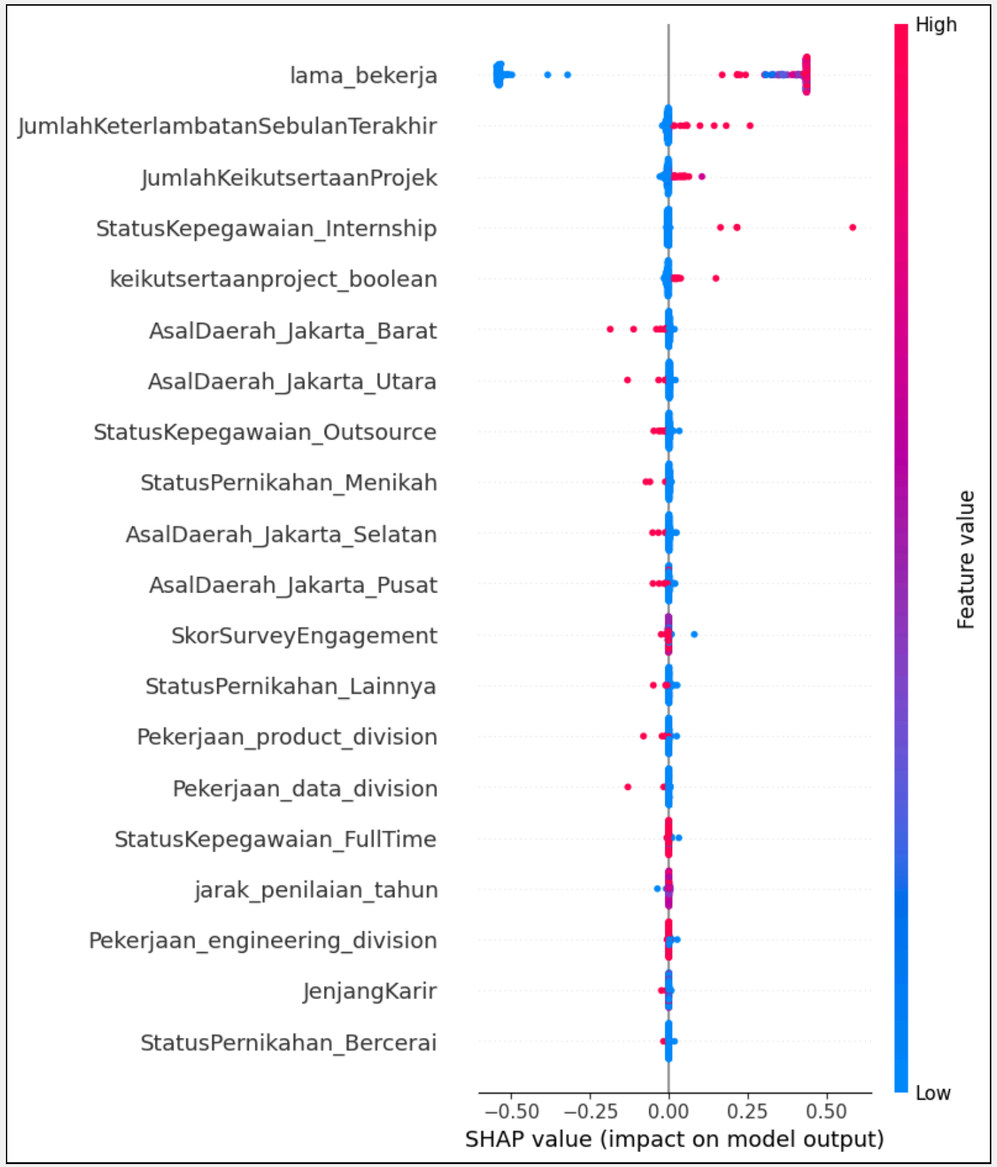
Confusion Matrix



Model prediction result:

- **37** employees were **correctly** predicted as **not resigning**.
- **18** employees were **correctly** predicted as **resigning**.
- **1** employee was **incorrectly** predicted to resign but actually did not (false positive).
- **2** employees were **incorrectly** predicted to stay but actually resigned (false negative).

Shap Values



Based on the impact of each feature, we found the following insights:

- The longer an employee has been employed, the higher the likelihood they will resign.
- Employees who are frequently late are more likely to resign.
- Interns have a high chance of resigning, or more precisely, not extending their contracts.
- Employees who participate in projects have an increased chance of resigning.
- Employees living in Jakarta Barat and Jakarta Utara are less likely to resign compared to those living in Jakarta Selatan and Jakarta Pusat.

Business Recommendation

Business Recommendation

Based on the Exploratory Data Analysis, we found that Software Engineers (Front End) exhibit high resignation rates, primarily linked to career changes and challenges with remote work. Data Analysts tend to resign due to a toxic work culture. Additionally, SHAP value analysis indicates that employees' length of service strongly influences their likelihood to resign or stay. Frequent tardiness and involvement in projects also impact resignation outcomes. Therefore, we recommend the company to:

Focus Area	Key Actions		
1. Support High-Risk Staff	- Help employees with short tenure and high risk (e.g., frequent lateness, low engagement). - Watch for burnout in those handling multiple projects.		
2. Address Resignation Causes	Software Engineers: • Improve remote work options. • Offer clear career paths. • Build a positive, supportive culture.	Data Team: • Give more recognition. • Promote a friendly work environment.	Other Roles: • Update work hours. • Train leaders. • Clarify career growth.
3. Increase Engagement	- Run regular surveys on job satisfaction and workplace quality.		
4. Predict and Act Early	- Use updated models to spot at-risk employees and intervene promptly.		

Conclusion

The analysis highlights key factors driving employee resignation, including career changes, remote work challenges, toxic work culture, and employee tenure. By focusing retention efforts on **high-risk employees**, **addressing specific resignation causes by role**, **enhancing engagement**, and **leveraging predictive monitoring**, the company can proactively **reduce turnover** and foster a healthier, more supportive work environment. Implementing these targeted actions will help **retain valuable talent** and **improve overall organizational stability**.

Thank You

Contact:

http://wa.me/6282280471417

aldivibriani@gmail.com

Social Media:

Github Account 

Linkedin Account