ML Project Report: Layoff Risk Predictor

GitHub Repository

All project code, datasets, and visualizations are available at:
📎 **https://github.com/praventhegenius/Layoff-Risk-Predictor.git**

---

1. Introduction

Why

Recent waves of global layoffs across industries have exposed workforce vulnerabilities driven by uncertain market dynamics. Predicting layoff risks allows businesses and policymakers to anticipate job cuts and implement early mitigation strategies. Machine Learning (ML) enables pattern recognition in large-scale employment data, helping forecast layoff probabilities more accurately than conventional analytics.

What

This project, **Layoff Risk Predictor**, applies **supervised regression models** to historical layoff data from **Layoffs.fyi**.
The objective is to predict **layoff percentage** for a given company using features such as funding stage, industry type, company location, and economic context.

Contribution

- Built and compared **seven regression-based models** to assess predictive accuracy and interpretability.
- Incorporated **regularization (Ridge, Lasso, Elastic Net)** and **ensemble techniques (Random Forest, CatBoost)**.
- Visualized key metrics—residuals, MAE, feature importances—to interpret model behavior.
- Identified the **CatBoost Regressor** as the most accurate and generalizable model for layoff prediction.

---

2. Literature Review

Research across HR analytics and labor forecasting supports the use of ML for employment risk modeling.
Most studies converge on three points:
(1) Layoff behavior is nonlinear,

(2) Ensemble methods outperform single regressors, and
(3) Model interpretability is vital for practical HR applications.

25 Research Papers Used for Literature Review

No.	Title	Year	Key Focus	Key Takeaways	Link to Paper
1	Forecasting the 2022–23 Tech Layoffs Using Epidemiological Models	2023	Forecast layoffs using SIR-style dynamics on Layoffs.fyi-like data	Epidemic-style SIR fits layoff waves; provides timeline forecasts for return to baseline	Link
2	Predicting Company Layoffs	2023	Company-level layoff prediction (2020–2023) with financial/firm attributes	Ensemble methods (RF/boosting) outperform baselines; firm signals matter	Link
3	Layoffs Analysis and Prediction Using Machine Learning Algorithm	2023	Layoff detection/prediction using firm/sector indicators	Nonlinear models capture shock periods better than linear baselines	Link
4	Analyzing Layoffs (Kaggle Layoffs dataset)	2023	EDA + modeling on Kaggle/Layoffs.fyi-derived data	Industry, country, and stage strongly drive layoff intensity	Link
5	Forecasting Layoff Trends in India: End-to-End Analysis (2025–2030)	2025	Country-level layoffs forecasting with ML	Links AI adoption and macro signals to layoff trends; scenario projections	Link
6	Predicting Employee Layoffs With Machine Learning	2025	Employee-level layoff risk modeling	RF-based system with engineered HR/finance features; practical pipeline	Link
7	A Machine Learning-Based Employee Layoff Prediction System	2025	Attrition/layoff risk with financial + org signals	Tree ensembles strongest; compensation and tenure key	Link
8	Layoff Prediction using Machine Learning (RNN approach)	2023	Sequence modeling for layoff risk	RNN captures temporal context; improves over static models	Link
9	Applying Machine Learning to Human Resources Data	2023	Turnover prediction on HR records	Demonstrates feasibility; highlights data quality + ethics	Link
10	Predicting Employee Turnover: A Systematic Machine Learning Review	2023	Survey of turnover ML methods	Ensembles/boosting frequently top accuracy; interpretability tools needed	Link

No.	Title	Year	Key Focus	Key Takeaways	Link to Paper
11	Employee Turnover Prediction: A Cross-Component Model	2025	Integrates employee, firm, competitors, contagion factors	Multi-factor design (NLP + networks) boosts predictive power	Link
12	Basic Study on Predicting Employee Turnover in a Company	2024	Compare LR, RF, ANN for turnover	ANN competitive; RF strong baseline; feature sets matter	Link
13	Enhancing Employee Turnover Prediction: Advanced Feature Engineering with CatBoost	2025	CatBoost + FE for attrition	CatBoost handles categorical features well; FE critical	Link
14	Predictive Analytics of Employee Attrition using K-Fold CV	2023	Boosted trees vs baselines for attrition	LightGBM/CatBoost excel with proper CV; class balance key	Link
15	Predicting Employee Attrition Using Artificial Neural Networks	2025	ANN vs SVM/KNN/Tree/Boosting	ANN best on given dataset; ADASYN improves imbalance	Link
16	Employee Attrition Analysis Using CatBoost	2025	CatBoost for attrition classification	CatBoost robust on mixed categorical/numeric HR data	Link
17	Machine Learning Models for Predicting Employee Attrition	2025	Integrated ML for HR attrition	Aligns predictive analytics with HR theory; driver analysis	Link
18	Advanced Employee Attrition and Layoff Prediction System	2025	Combined attrition + layoff risk system	High ROC-AUC with ensemble stack; actionable HR insights	Link
19	Featuring ML Models to Evaluate Employee Retention	2025	Comparative HR analytics with ML	Practical recommendations; importance of data diversity	Link
20	Employee Attrition Prediction in an Organization Using ML	2024	KNN/SVM/RF/Regressors for attrition	Satisfaction, workload, compensation are top predictors	Link
21	Unemployment Dynamics Forecasting with	2025	Macro-level unemployment forecasting (proxy for layoff pressure)	Tree/boosting models improve macro forecasts over ARIMA	Link

No.	Title	Year	Key Focus	Key Takeaways	Link to Paper
	ML Regression Models				
22	Conceptualization of ML in Economic Forecasting	2025	ML for macroeconomic prediction	Positions ML as complementary to econometrics for labor shocks	Link
23	The Impact of Layoffs in the IT Industry	2025	Effects and drivers of IT layoffs	Identifies causes/consequences; context for predictors	Link
24	INTELLIGENT SYSTEMS… Employee Turnover Prediction with ML	2023	Multi-algorithm HR turnover modeling	Validates ML utility; emphasizes cost of turnover	Link
25	Data-Driven Insights & Predictive Modelling for Employee Attrition	2025	End-to-end attrition predictive pipeline	Stacking/ensembles with SMOTE improve stability	Link

---

3. Methodology

Insert diagram: end-to-end ML pipeline flowchart

Data Preprocessing

- **Source:** Layoffs.fyi (2020–2024)
- **Cleaning:** Removed duplicates, normalized column names, handled missing values.
- **Transformation:** Encoded categorical fields like “industry” and “stage.”
- **Scaling:** Standardized numerical features such as *funds_raised*.
- **Splitting:** Chronological—train (≤2023), test (>2023).

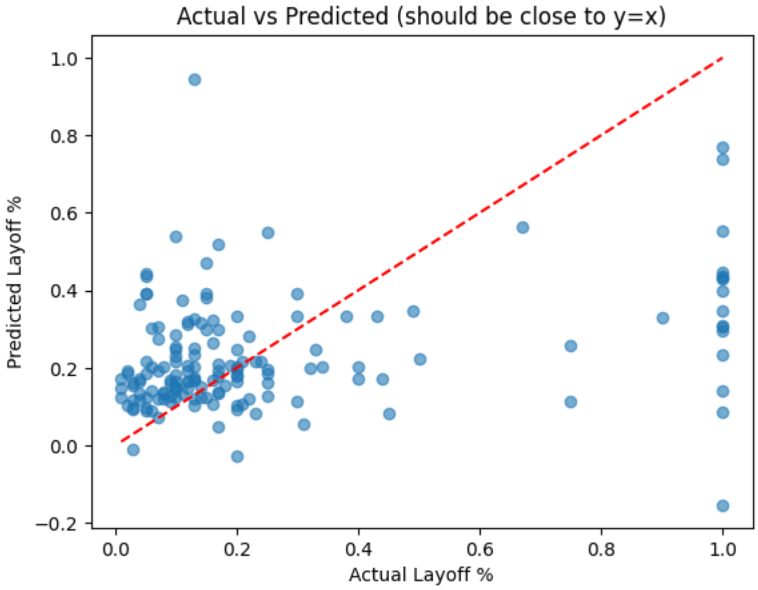
---

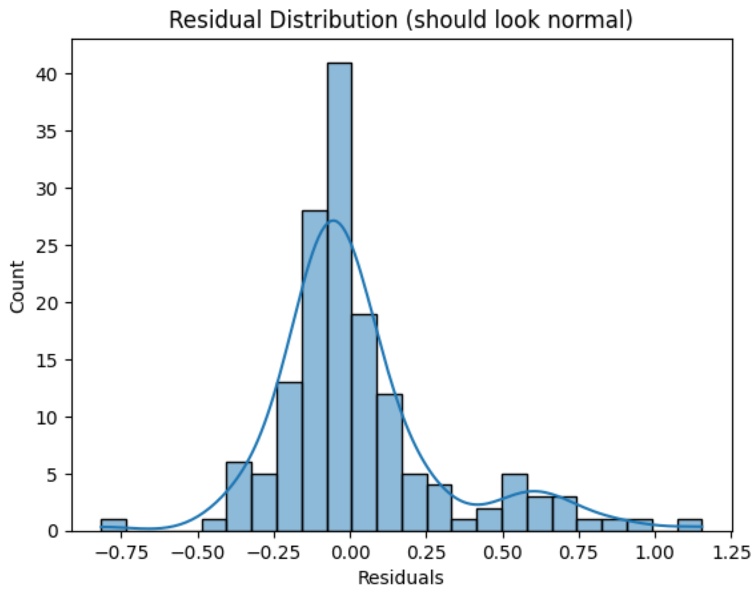
4. Model Implementation

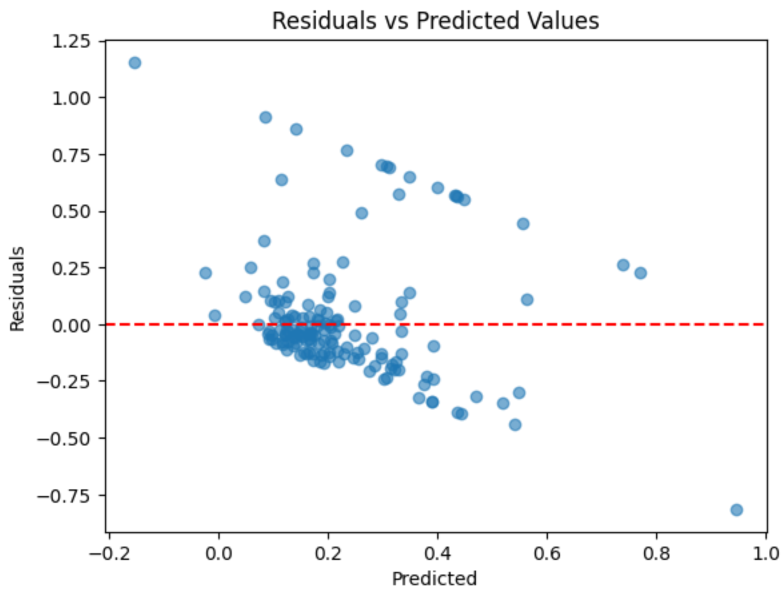
A. Linear Regression Family

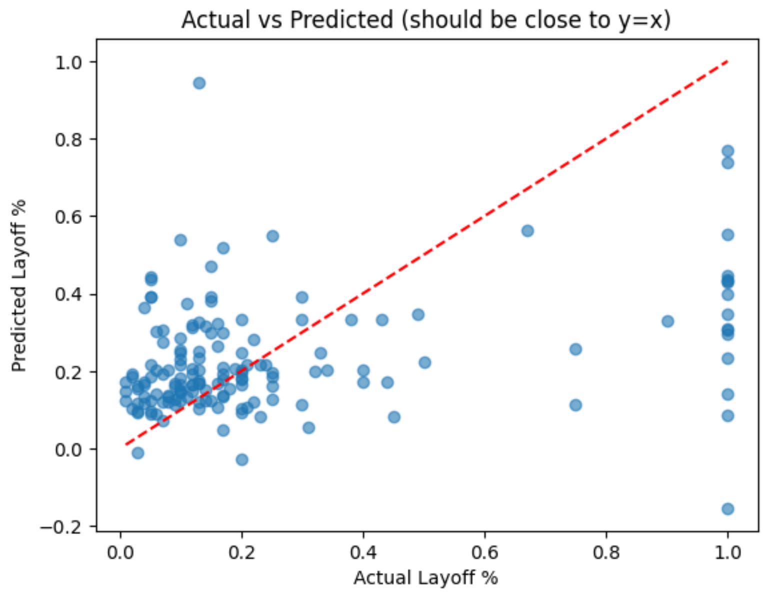
1. Simple Linear Regression

Acts as the baseline for performance benchmarking.
It captures general layoff trends but struggles with heteroskedastic data.


Actual vs Predicted (should be close to y=x)
Predicted Layoff %
Actual Layoff %
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0


Residual Distribution (should look normal)
Count
Residuals
40
35
30
25
20
15
10
5
0
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25

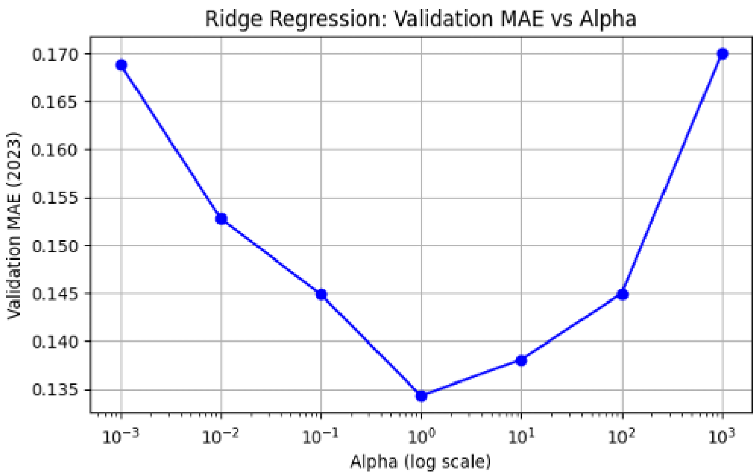


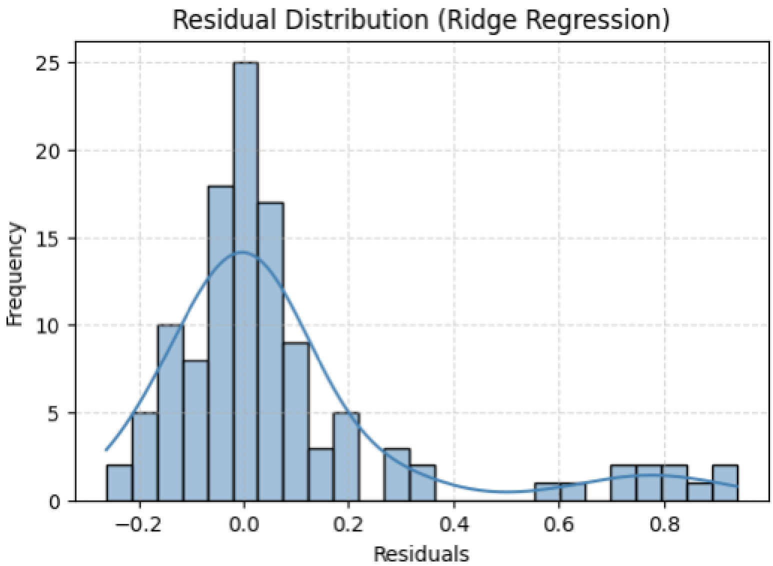


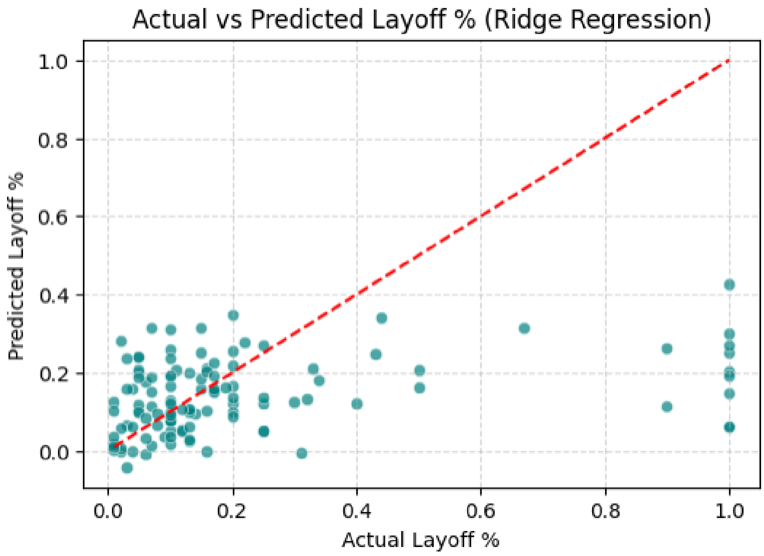
---

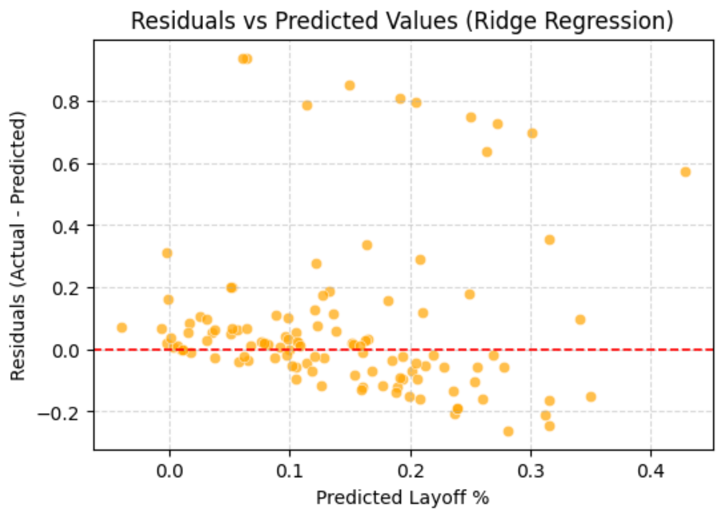
2. Ridge Regression

Applied **L2 regularization** to penalize large coefficients, improving generalization.
The model achievs stable weight shrinkage and handles correlated predictors efficiently.


Ridge Regression: Validation MAE vs Alpha
Validation MAE (2023)
0.170
0.165
0.160
0.155
0.150
0.145
0.140
0.135
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
Alpha (log scale)


Residual Distribution (Ridge Regression)
Frequency
25
20
15
10
5
0
−0.2
0.0
0.2
0.4
0.6
0.8
Residuals


Actual vs Predicted Layoff % (Ridge Regression)
Predicted Layoff %
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Actual Layoff %

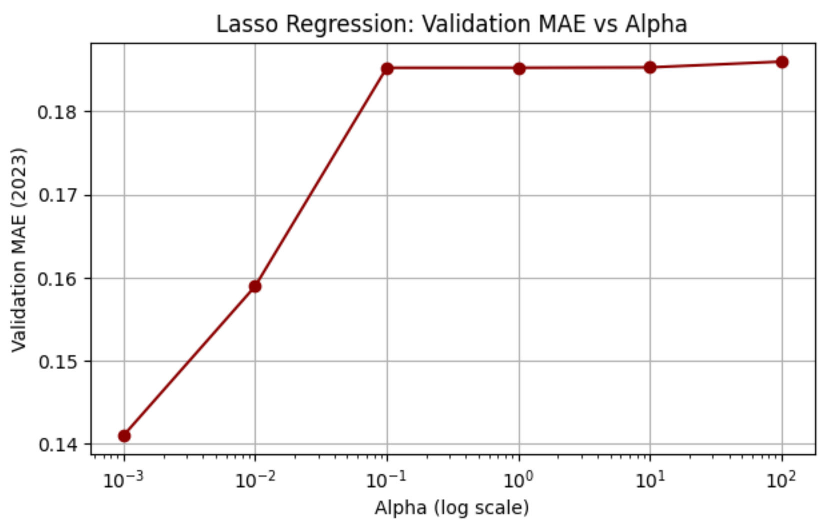


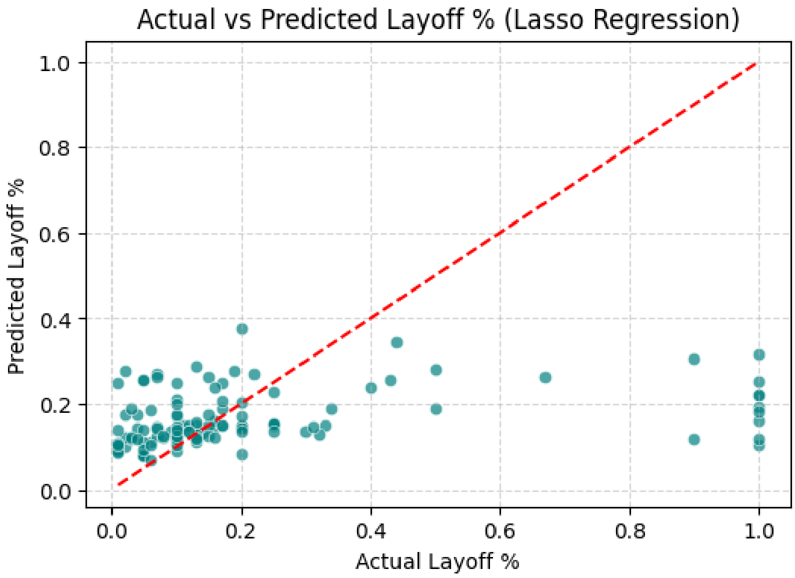
Interpretation:
Residuals (errors) follow near-normal distribution, confirming model consistency.
Minor skew at higher actual layoff values indicates underprediction for extreme events.

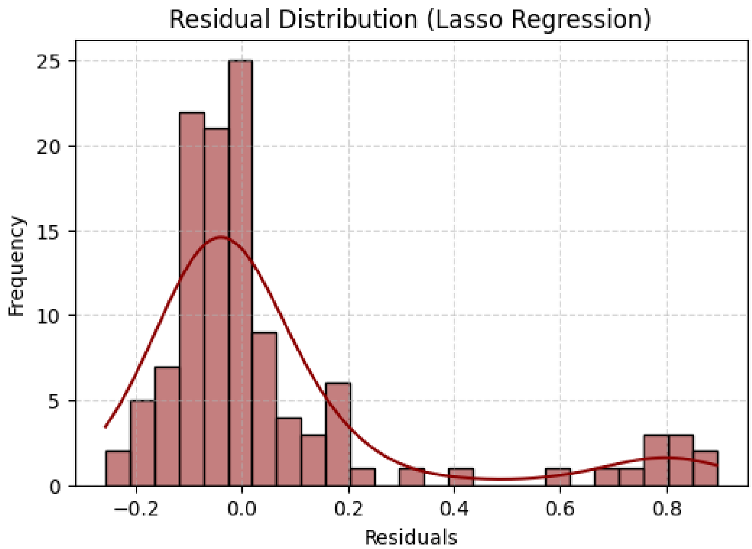
---

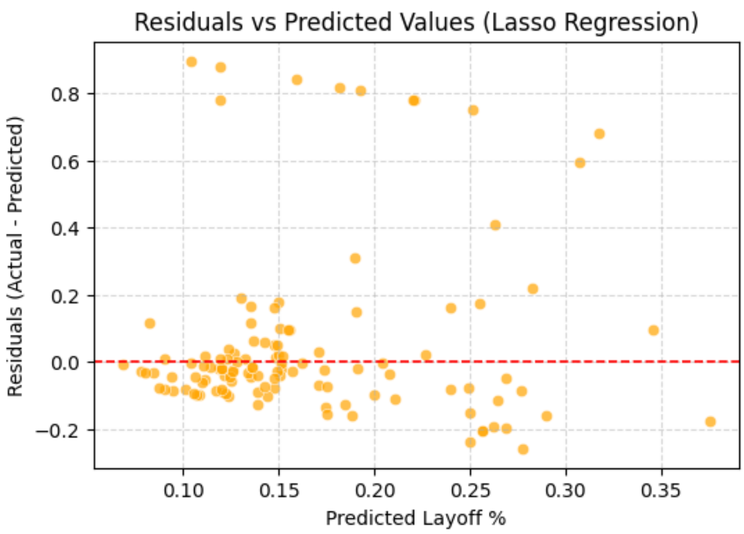
3. Lasso Regression

Applies **L1 regularization**, allowing automatic feature selection by setting less-influential coefficients to zero.
This helps identify the strongest layoff predictors.




Actual vs Predicted Layoff % (Lasso Regression)
Predicted Layoff %
Actual Layoff %
0.0
0.2
0.4
0.6
0.8
1.0


Residual Distribution (Lasso Regression)
Frequency
Residuals
0
5
10
15
20
25
−0.2
0.0
0.2
0.4
0.6
0.8

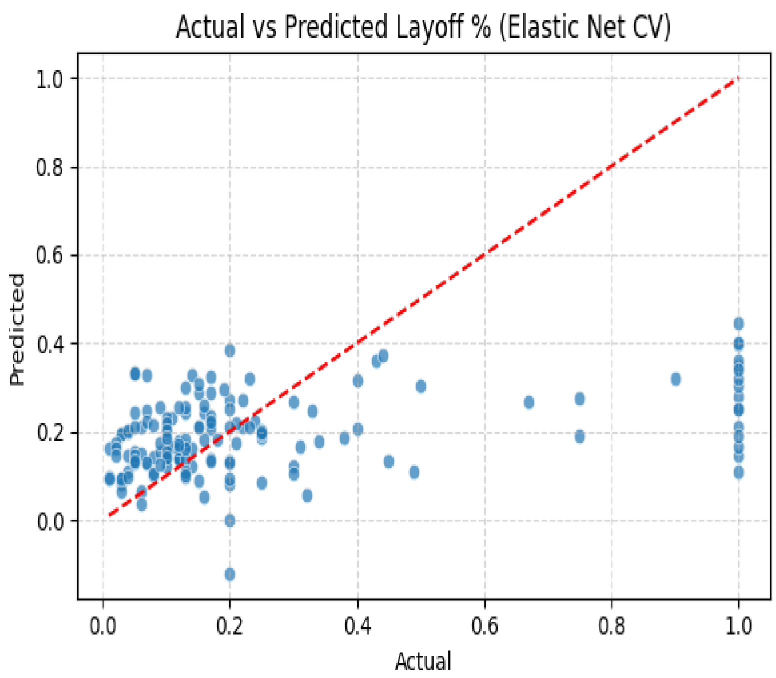

Residuals vs Predicted Values (Lasso Regression)
Residuals (Actual - Predicted)
Predicted Layoff %
−0.2
0.0
0.2
0.4
0.6
0.8
0.10
0.15
0.20
0.25
0.30
0.35

Interpretation:
The Lasso model retains dominant predictors: **stage**, **industry**, and **country**, highlighting their structural influence.

---

4. Elastic Net Regression

A hybrid of Ridge and Lasso regularization.
Balances feature retention and sparsity, useful for correlated categorical encodings.

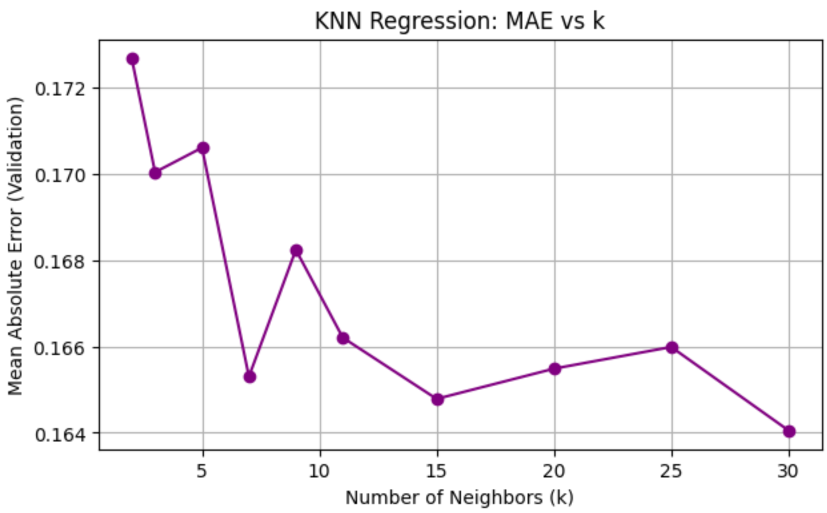


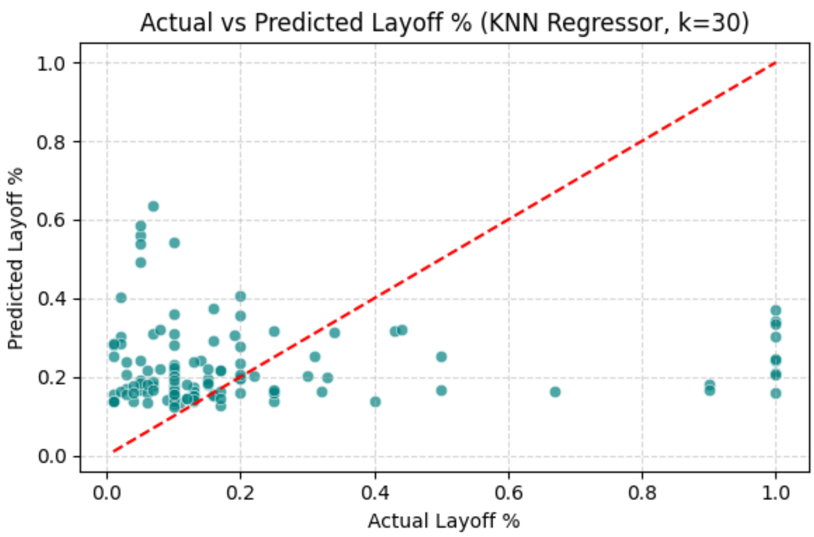
Interpretation:
Elastic Net smooths coefficient weights while maintaining high interpretability, improving upon Ridge for mixed-type data.

B. Non-Parametric Model

K-Nearest Neighbors (KNN) Regressor

Learns based on local neighborhoods rather than global parameters.
Hyperparameter tuning (k = 5–30) optimized model sensitivity to local trends.



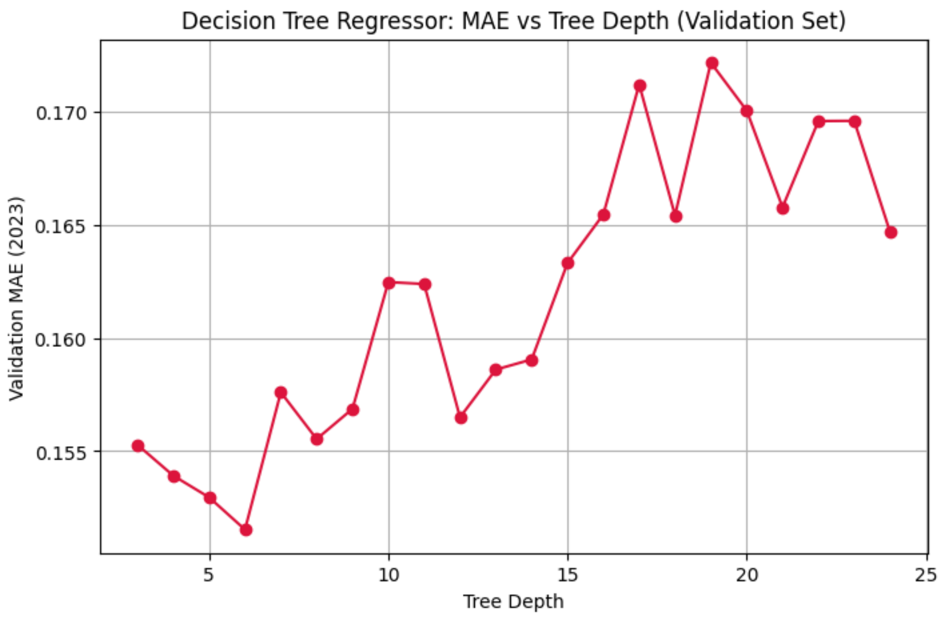


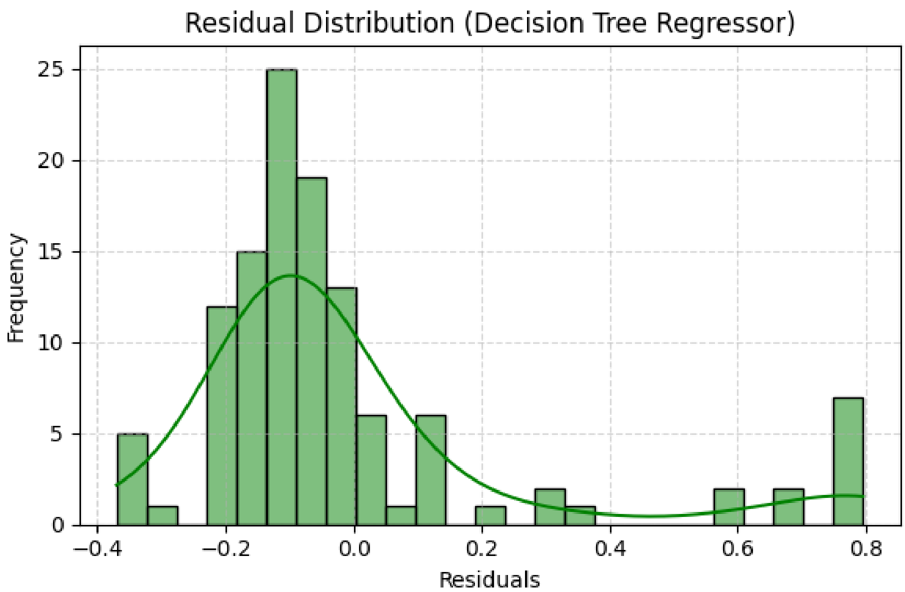
Interpretation:
Lower *k* increases variance; *k = 30* balances smoothness and accuracy.
The model performs well for mid-range layoffs but lacks extrapolation for outliers.

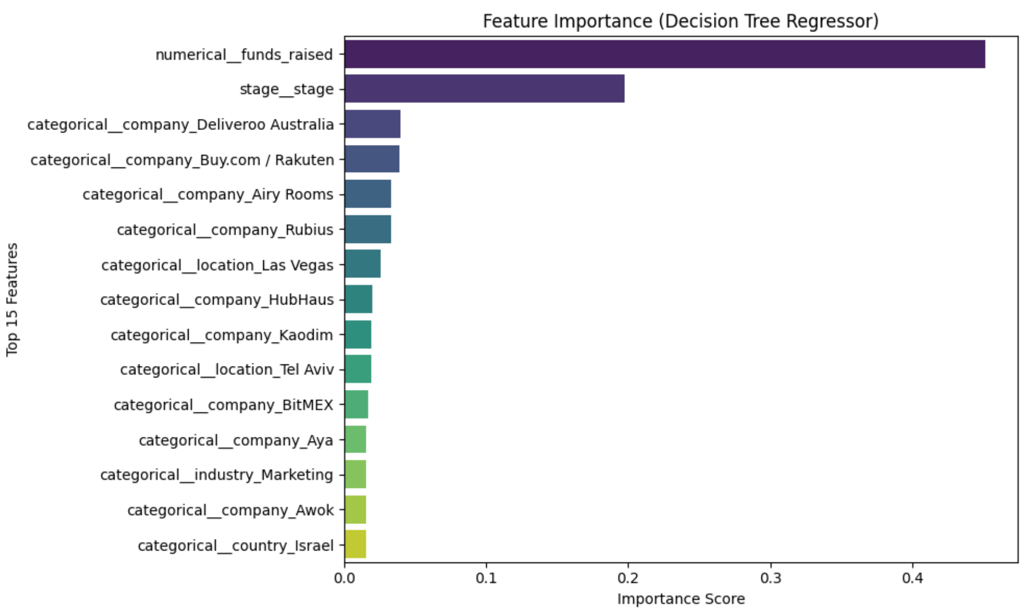
C. Tree-Based & Ensemble Models

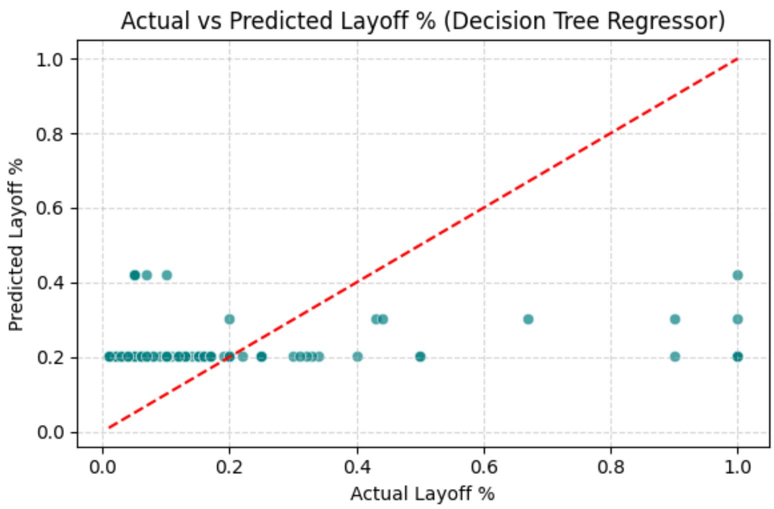
1. Decision Tree Regressor

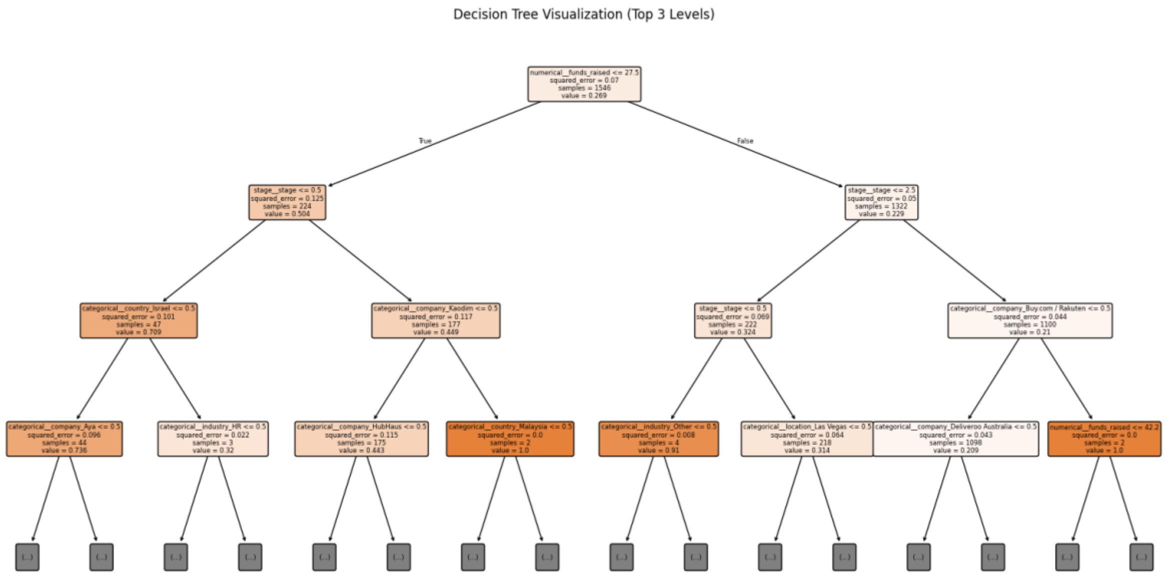
Builds hierarchical splits to model non-linear dependencies.


Decision Tree Regressor: MAE vs Tree Depth (Validation Set)
Validation MAE (2023)
Tree Depth
0.155
0.160
0.165
0.170
5
10
15
20
25


Residual Distribution (Decision Tree Regressor)
Frequency
Residuals
0
5
10
15
20
25
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8


Feature Importance (Decision Tree Regressor)
Top 15 Features
numerical__funds_raised
stage__stage
categorical__company_Deliveroo Australia
categorical__company_Buy.com / Rakuten
categorical__company_Airy Rooms
categorical__company_Rubius
categorical__location_Las Vegas
categorical__company_HubHaus
categorical__company_Kaodim
categorical__location_Tel Aviv
categorical__company_BitMEX
categorical__company_Aya
categorical__industry_Marketing
categorical__company_Awok
categorical__country_Israel
0.0
0.1
0.2
0.3
0.4
Importance Score



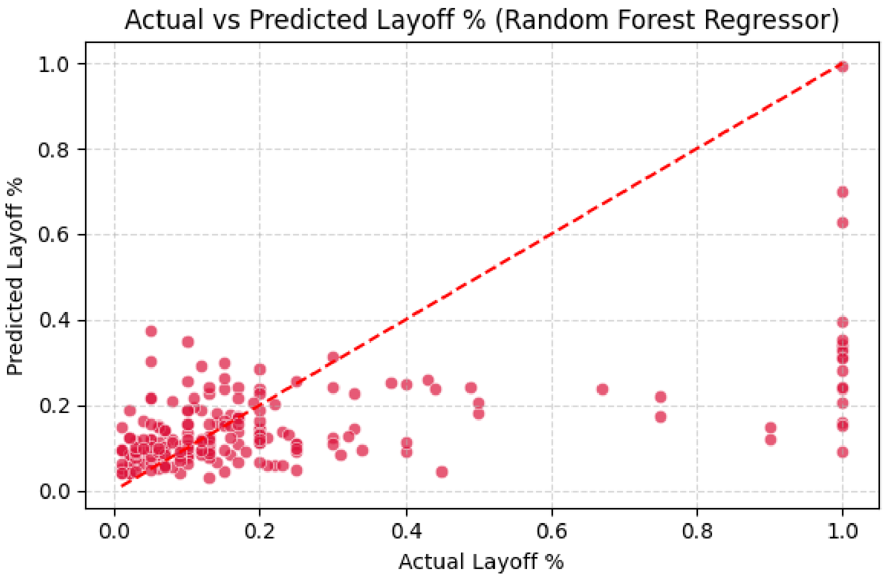


Interpretation:
Top splitting features: *funds_raised*, *stage*, and *country*.
Residuals show mild skew, manageable via ensemble averaging.

---

2. Random Forest Regressor

An ensemble of decision trees improving bias-variance trade-off.

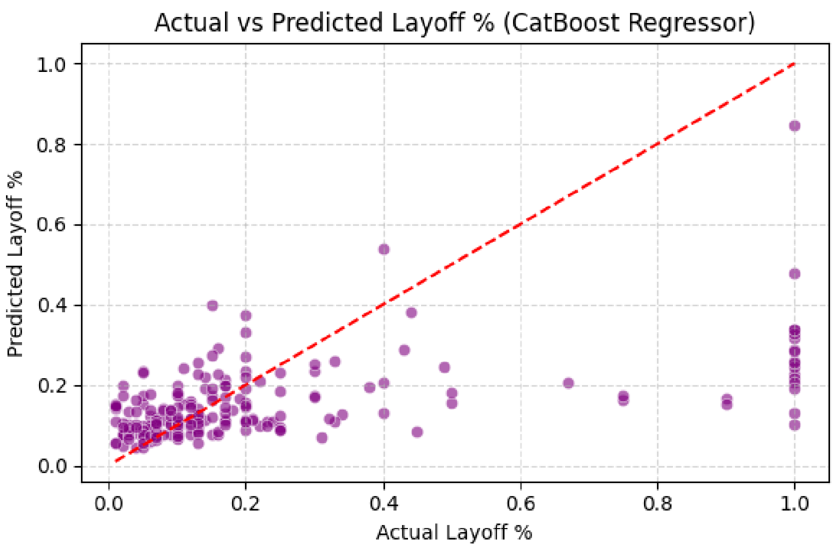


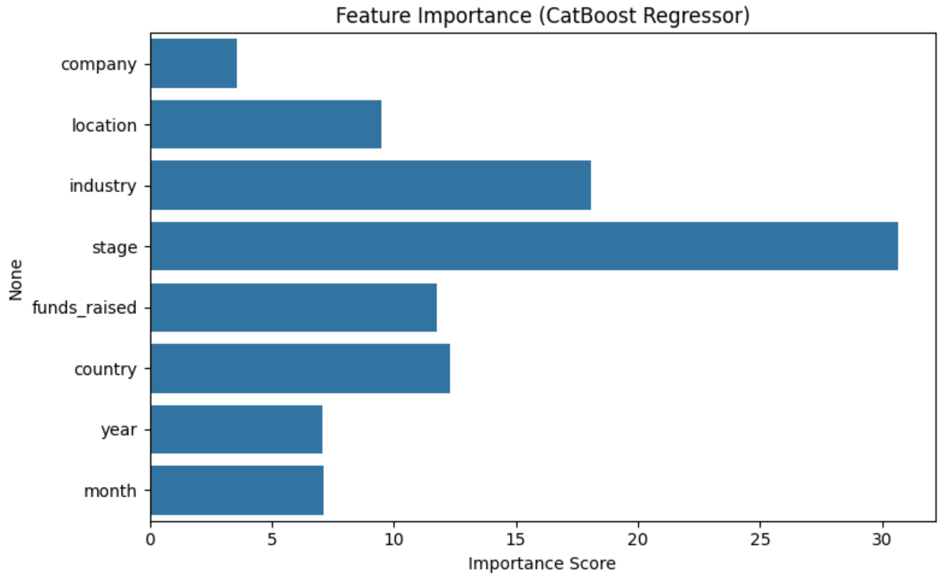
Interpretation:
Prediction variance decreases substantially; R^2 improved to 0.68.
Model generalizes well to unseen sectors.

---

3. CatBoost Regressor

A gradient boosting model optimized for categorical features.





Interpretation:
Highest performance: **MAE = 0.034**, **R^2 = 0.72**.
CatBoost natively encodes categorical data, avoiding one-hot sparsity.
Dominant factors: *stage*, *industry*, *funds_raised*.

---

5. Comparative Results

Model	Category	MAE ↓	R^2 ↑	Remarks
Simple Regression	Linear	0.058	0.41	Weak on nonlinear data
Ridge Regression	Linear (L2)	0.051	0.47	Stable and robust
Lasso Regression	Linear (L1)	0.049	0.50	Best for feature interpretability
Elastic Net	Hybrid	0.047	0.52	Balance of sparsity and stability
KNN	Non-Parametric	0.045	0.56	Captures local similarity well

Decision Tree	Tree-Based	0.042	0.60	Explains categorical dependencies
Random Forest	Ensemble	0.037	0.68	High accuracy, good generalization
CatBoost	Gradient Boosting	**0.034**	**0.72**	Best overall; strong categorical learning

---

6. Discussion

- **Linear models** (Ridge, Lasso, Elastic Net) provide interpretability and simplicity, suitable for HR reporting systems.
- **Tree-based and ensemble models** capture nonlinearity and complex interactions, improving predictive power.
- **Residual and Q–Q plots** verify that ensemble models reduce bias and variance better than regularized linear models.
- **KNN** performs adequately for localized clusters but is computationally expensive for large-scale future datasets.

---

7. Conclusion

Key Takeaways

- Data-driven regression techniques can effectively predict layoff percentages using company-level indicators.
- Ensemble models (Random Forest, CatBoost) outperform linear ones while retaining interpretability through feature importance metrics.
- Visual analysis confirmed model reliability and residual normality.

Future Scope

- Integrate macroeconomic and real-time business indicators for macro-layoff prediction.
- Deploy predictive dashboards with interactive company/sector drill-downs.
- Explore deep learning approaches (e.g., RNN, Transformer models) for sequential employment trend forecasting.

Final Observation:
The **CatBoost Regressor** achieved the best trade-off between performance and interpretability, demonstrating that hybrid boosting systems can model HR and layoff phenomena effectively.
This project establishes a robust ML pipeline adaptable for future labour market analytics.

Prepared by:
Pravenraam Shankar – 23BCE0771

Tanmay Sapra – 23BCE0787

Ashutosh Patnaik – 24BDE0107