**Assignment 3 - Modeling and Application of Project Results**

Likhita Alla

Saint Peter’s University

DS-670-HYB2-25SPTR: Capstone: Big Data & Data Science

Professor- ***Reda Mastouri***

**March 16, 2025**

**US Layoff Analysis - Modelling and Application of Project Results**

**1. Introduction**

The United States' job market has witnessed drastic fluctuations in employment patterns, and layoffs have been a recurring problem for companies and policy makers. Layoffs have multiple causatives such as recession in the economy, technological shifts, corporate restructuring, and shifts in demand in the market. Identification of the causatives of layoffs and predicting their occurrences can allow organizations to take proactive measures in curbing job loss. This research applies machine learning techniques to explore patterns in layoffs from the GV Data and DOGE Telemetry dataset. The data consists of 14,499 separation records of workers with details such as department, job category, pay grade level, employee status, and separation causes. Through predictive modelling, it aim to reveal patterns resulting in layoffs and test the effectiveness of various machine learning algorithms in predicting future separations. This work can be useful to companies in workforce planning, improving job security, and devising policies to reduce the impacts of layoffs on workers and the economy (Hossen et al., 2023).

**2. Data Preparation**

Data pre-processing plays a crucial role in enabling machine learning algorithms to learn effectively from structured data sets. The raw data set contained missing values, inconsistencies, and imbalanced class distributions and thus required thorough pre-processing and transformation. Duplicate records were first removed to prevent bias in model training. Numerical features such as "Pay Grade" had missing values imputed with the median value and categorical features such as "Reason" and "Employee Status" had the most occurring category imputed to maintain data integrity. New features were also engineered to enhance the predictive capability of the data set (Singh & Tiwari, 2025).

A "Separation Year" column was derived from the "Separation Date" to analyze the temporal distribution of layoffs and a "Recent Layoff" binary indicator to highlight layoffs in 2018 and beyond. The data set was also encoded to represent categorical data in numeric values. Target variable "Reason" was label-encoded to facilitate classification and other categorical features underwent one-hot encoding to prevent creating ordinal relationship where there wasn't any. Ultimately, the data set had an extremely skewed class distribution as some separation reasons had significantly fewer occurrences. To counter this, data augmentation techniques such as resampling and Synthetic Minority Over-sampling Technique (SMOTE) were employed to obtain a balanced class distribution. This pre-processing of the data ensured the data set became clean, balanced, and suitable for creating strong predictive models (Singh & Tiwari, 2025).

**3. Model Implementation**

The experiment employed four machine learning algorithms in predicting layoff incidents from employee attributes and separation histories. Logistic Regression, a widely applicable statistical model for binary and multiclass problems, was included as a baseline as it is easy to interpret and simple to use. Unfortunately, it suffered from convergence issues due to the nature of the data and had the max iterations increased to 1000. Random Forest, an ensemble learner where multiple trees are built and the predictions are merged, was also employed. This model can deal with non-linearity and interactions between the features and is suitable for predicting layoffs (Zhang & Han, 2024).

The Extra Trees Classifier, an extension of Random Forest with additional randomized feature selection, was also tried in order to test its influence on generalization performance. Last but not least, the Gradient Boosting Classifier was employed as a boosting learner where an ensemble of learners are built sequentially in order to minimize the error in the task of classification. Hyperparameters of the model were tuned for predictive accuracy. Below follows an overview of the models and settings:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Algorithm Type | Key Parameters | Advantages |
| Logistic Regression | Linear Model | Max Iterations = 1000 | High interpretability |
| Random Forest | Ensemble (Bagging) | Trees = 100 | Handles non-linearity, reduces overfitting |
| Extra Trees | Ensemble (Bagging) | Trees = 100 | Faster than Random Forest, better generalization |
| Gradient Boosting | Ensemble (Boosting) | Trees = 100, Learning Rate = 0.1 | Improves weak learners iteratively |

These were trained on a standard data set where numerical features were scaled by Standard Scaler in order to have equal weightage of features. 80% of the train data were set aside for model training and 20% for testing in order to test model generalization.

**4. Model Evaluation**

The Evaluation of machine learning models entails a multi-dimensional approach to ascertain their ability to make accurate predictions. Different performance metrics were employed, including Area Under the Curve (AUC) to estimate the performance of classifications, Precision to estimate the accuracy of positive predictions, Accuracy to determine total correctness, and Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to estimate error margins (Naidu et al., 2023).

Evaluation results indicated the Random Forest Classifier to be better than other models with the highest accuracy (81.0%) and precision (77.2%) and consequently the best model to predict layoffs. The Gradient Boosting Classifier had excellent generalization power but had to be optimized to improve accuracy. Below follows a comparative model performance analysis:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | AUC | Precision | Accuracy | MAE | MSE | RMSE |
| Logistic Regression | 0.889 | 0.735 | 0.746 | 1.91 | 15.66 | 3.95 |
| Random Forest | 0.913 | 0.772 | 0.810 | 1.52 | 12.79 | 3.57 |
| Extra Trees | 0.908 | 0.768 | 0.805 | 1.55 | 13.04 | 3.61 |
| Gradient Boosting | 0.920 | 0.760 | 0.799 | 1.59 | 13.33 | 3.65 |

These results indicate that ensemble methods such as Random Forest and Extra Trees worked very well in identifying patterns in layoff data. Logistic regression struggled with complex feature interactions and this undermined its predictive power.

**5. Findings and Conclusions**

The results of this study reveal that separation year, department, and employee status are the most important factors in layoffs. Tree-based methods performed better and Random Forest produced the best balance between accuracy and interpretability. Convergence in logistic regression was a significant issue and could be alleviated by dimensionality reduction and feature selection. Future work should explore the impacts of workforce restructuring policies and external economic conditions in an effort to enhance predictive power. Organizations can leverage the results of this study to enhance workforce management practices, optimize talent retention programs, and make data-informed decisions to reduce unnecessary job loss (Naidu et al., 2023).

**6. Conclusion and Recommendations**

It is proposed on the basis of the findings that Random Forest Classifier would be deployed as the baseline prediction model for layoff assessment. Data collection could be improved by incorporating economic indicators to improve model stability. Deep learning techniques or newer boosting algorithms like XGBoost could be investigated in future work to improve predictions further. These findings can be used by human resources professionals and policymakers to develop proactive intervention policies such that workforce reductions are minimized and business efficiency is preserved. This report provides a systematic methodology to study layoffs and emphasizes the role of machine learning in labor market analytics.

**Reference**

Hossen, M. N., Mollah, M. A., Lipy, N. S., Hossain, G. M. S., & Rahman, M. S. (2023). Factors affecting layoff in high-tech industry: Evidence from USA. *International Journal of Information, Business and Management*, *15*(1), 1-13.

<https://www.proquest.com/openview/dff61e6a756dd2a2a006b8407833566d/1?cbl=2032142&pq-origsite=gscholar>

Naidu, G., Zuva, T., & Sibanda, E. M. (2023, April). A review of evaluation metrics in machine learning algorithms. In Computer science on-line conference (pp. 15-25). Cham: Springer International Publishing.

<https://doi.org/10.1007/978-3-031-35314-7_2>

Singh, Y., & Tiwari, M. (2025). A Comprehensive Machine Learning Approach for Early Detection of Diabetes on Imbalanced Data with Missing and Outlier Values. SN Computer Science, 6(3), 213.

<https://doi.org/10.1007/s42979-025-03751-6>

Zhang, C., & Han, W. (2024). Ensembles of decision trees and gradient-based learning for employee turnover rate prediction. PeerJ Computer Science, 10, e2387.

<https://peerj.com/articles/cs-2387/>