**Modeling and Application of Project Results**

Sravani yalamarthi

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 US job market has experienced drastic changes with workforce reduction emerging as a significant concern for firms and policy-makers. Layoffs result from a number of factors such as economic downturns, technological changes, corporate restructuring, and changes in market demand. Understanding the trends and predicting layoffs allows organizations to apply proactive interventions to curtail job loss.

It applies machine learning techniques to study patterns of layoffs from the GV Data and DOGE Telemetry dataset consisting of 14,499 separation records of workers. These records include details such as department, job category, pay grade, employee status, and reasons for separation. This work will use predictive modeling to establish patterns of layoffs and validate the effectiveness of various machine learning algorithms in predicting separation. This data can be valuable in workforce planning, enhancing job security, and policy formulation in order to curtail the economic impact of layoffs (Saba, 2024).

**2. Data Preparation**

Data preparation is an important process in helping machine learning models learn properly. Missing values, inconsistencies, and unbalanced class distributions in the original data made it imperative to clean and transform the data. These were the following steps undertaken:

* Data Cleaning: Duplicates were removed in order to prevent model bias. Numerical column missing values like "Pay Grade" were imputed with the median value and missing values in categorical columns like "Reason" and "Employee Status" were imputed by the most frequent category.
* Feature engineering: A "Separation Year" column was derived from "Separation Date" to analyze layoffs by year. A "Recent Layoff" flag was also defined to highlight separations in and subsequent to 2018 (Hakami, 2024).
* Encoding: Categorical features were converted into numeric representations "Reason" was label-encoded for classification and other categorical features were one-hot encoded.
* Managing the Dataset Balancing: Class imbalances were addressed by utilizing resampling and Synthetic Minority Over-sampling Technique (SMOTE) in order to have an equal distribution of classes.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 preprocessing tasks resulted in a formatted data set optimized for predictive modeling.

**3. Model Implementation**

Four machine learning models were evaluated to predict layoffs:

1. Logistic Regression – A simple classifier picked for interpretability purposes but due to data sets' complexity issues, convergence issues made it necessary to set max iterations to 1000.

2. Random Forest – A machine learning approach utilizing a collection of multiple decision trees to enhance precision and avoid overfitting.

3. Extra Trees Classifier – An extension of the Random Forest with added randomness in feature selection for improved generalization.

4. Gradient Boosting Classifier – A process of iterative boosting to enhance predictive accuracy by building on weak learners (Shankar et al., 2024).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 |

Model Configurations and AdvantagesStandardScaler was employed to normalize the data and an 80-20 train-test split was performed to test the model efficiently.

**4. Model Evaluation**

The performance of the model was tested by different indicators including AUC (Area Under the Curve) to test the efficiency of classifications, Precision to verify the reliability of positive predictions, Accuracy to determine the accuracy in total, and MAE, MSE, and RMSE to estimate error margins. Random Forest had the best results with the highest accuracy (81.0%) and precision (77.2%). Ensemble methods including Random Forest and Extra Trees outperformed Logistic Regression, where the latter struggled with feature complexity (Carrington et al., 2022).

**Modeling Performance Outcomes**

Random Forest worked best with the greatest accuracy (81.0%) and precision (77.2%). Ensemble methods (Random Forest and Extra Trees) outperformed logistic regression because this latter model struggled with intricate features.

**5. Findings and Conclusions**

Key factors influencing layoffs include employee status, department, and year of separation, and all significantly contribute to layoffs. Random Forest produced the best balance between accuracy and interpretability. Logistic Regression suffered from convergence issues and could be mitigated with feature selection and dimensionality reduction. Future research could include workforce restructuring policies and economic factors to enhance model predictions. These findings can be used by organizations to enhance workforce strategy, optimize talent retention programs, and support data-based HR decisions (Sanchez et al., 2024).

**6. Conclusion and Recommendations**

Based on the results of the study, the Random Forest Classifier is proposed as the model of choice to predict layoffs. For improving predictions further, expanding data gathering to include economic indicators is suggested. Another source of improvement could arise from exploring newer algorithms such as XGBoost or deep learning techniques. Machine learning outputs should inform HR and policy formulation to curtail workforce layoffs and sustain efficiency. This study points to the potential of machine learning in labor market analytics and presents a data-based approach to employment stability management.

**Reference**

Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., ... & Holzinger, A. (2022). Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 329-341.

<https://ieeexplore.ieee.org/abstract/document/9693294>

Hakami, A. (2024). Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance. *Scientific Reports*, *14*(1), 9645.

<https://doi.org/10.1038/s41598-024-59958-9>

Saba, Z. (2024). Layoffs and corporate performance: evidence based on the US tech industry. *Journal of Economics and Finance*, *48*(3), 644-667.  
<https://doi.org/10.1007/s12197-024-09673-y>

Sanchez, D., Fleischman, G., & Sanchez, J. M. (2024). The impact of corporate social responsibility on employee layoffs, severance payments, and voluntary layoff disclosure. *Review of Quantitative Finance and Accounting*, 1-44.

<https://doi.org/10.1007/s11156-024-01363-4>

Shankar, S., Vyas, R., & Tewari, V. (2024). Applying machine learning algorithms to determine and predict the reasons and models for employee turnover. *International Journal of Information Technology and Management*, *23*(1), 48-63.

<https://www.inderscienceonline.com/doi/abs/10.1504/IJITM.2024.136187>