**Week-6 Report**

Likitha Alla  
Sravani yalamarthi

Saint Peter’s University

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

**Professor- *Reda Mastouri***

April 15, 2025

**Project Overview**

The objective of this project is to predict the **Reason for Employee Separation** using structured data from Iowa Executive Branch records. We aimed to evaluate multiple classification models and identify the most accurate one for potential deployment.

**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.

**Update: Integration of XGBoost Classifier**

**Model Added**

* **Name**: XGBoost Classifier
* **Library**: xgboost
* **Class**: XGBClassifier

**Model Configuration**

XGBClassifier(

n\_estimators=100,

learning\_rate=0.1,

random\_state=42,

use\_label\_encoder=False,

eval\_metric='mlogloss'

)

* Chosen for its gradient boosting capabilities and consistent performance with tabular data.
* Avoids overfitting with built-in regularization (L1 and L2).

**Model Performance Metrics**

|  |  |
| --- | --- |
| **Metric** | **XGBoost Score** |
| **Accuracy** | 0.808 |
| **AUC** | 0.926 |
| **Precision** | 0.757 |
| **MAE** | 1.545 |
| **MSE** | 12.90 |
| **RMSE** | 3.592 |
| **MAPE** | 3.82% |

These metrics were evaluated using the test set (20% stratified split from the balanced dataset).

**Comparison With Other Models**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **Model** | **AUC** | **Precision** | **Accuracy** | **MAE** | **MSE** | **RMSE** | **MAPE** | | Logistic Regression | 0.887 | 0.735 | 0.742 | 1.932 | 15.780 | 3.972 | inf | | Random Forest | 0.914 | 0.772 | **0.810** | 1.529 | 12.798 | 3.577 | inf | | Extra Trees | 0.909 | 0.769 | 0.806 | 1.556 | 13.048 | 3.612 | inf | | Gradient Boosting | 0.921 | 0.761 | 0.799 | 1.595 | 13.333 | 3.651 | inf | | XGBoost | **0.927** | 0.758 | 0.809 | 1.545 | 12.907 | 3.593 | inf | |  |  |  |

**Findings**

• The majority of separations were voluntary, with layoffs representing a smaller but notable category for analysis.

• Departments like Human Services and Corrections had the most redundancies, pointing to structural or budget issues.

• A peak in layoffs is apparent in 2017, after which it decreased, which can be attributed to policy or economic changes.

• The most successful model was Random Forest (81% accuracy, 77.2% precision), followed by Extra Trees and Gradient Boosting.

• Logistic Regression performed worse because it couldn't deal with complicated feature interactions.

• Feature engineering such as creating the "Recent Layoff" flag and extracting "Separation Year" improved model interpretability.

• Application of SMOTE eased class imbalance, thus leading to better model fairness and predictive capability.

**Results**

* **Best Performing Model**: Random Forest

Although **XGBoost** has the highest **AUC** (0.927), which means it’s excellent at distinguishing between classes overall, the **Random Forest** model shows the **best balance of performance metrics**, including:

* **Highest Accuracy** (0.810): Best at correctly predicting the layoff reasons overall.
* **Highest Precision** (0.772): Better at avoiding false positives compared to others.
* **Lowest MAE & MSE**: Indicates more accurate numerical predictions and fewer large errors.
* **Lowest RMSE**: Reflects overall model prediction error is lowest.

**5. Time-Series Analysis**

Along with the machine learning system, the project also had a time-series analysis to view layoffs in the context of how they have changed over time. This part presents the trend analysis and the significance of time-based insights.

* **Monthly and Annual Trends:** Layoffs were grouped by year and by month. Plotted were the following time-series trends:
* **Monthly Layoffs Trend:** Shows the trend in monthly layoffs, identifying peak months or seasonality of layoffs, if any.
* **Annual Layoffs Trend:** Illustrates the way layoffs have evolved year by year, providing a longer-term view of the workforce landscape.

**Major findings of time-series analysis:**

Monthly and yearly trends assist organizations in analyzing the effects of reductions in different time spans.

Determining the seasonal trends might guide future workforce planning and layoffs, which is imperative in HR and in policymaking.

**Visualizing Trends:**

Line charts are employed in order to show the trend of layoffs in time. Trends both monthly and annually give insight to how the layoffs were allocated through the years, and organizations are able to use these insights to guide data-driven decisions.

**6. Streamlit App for Interactive Visualization and Forecasting**

The last part of the project involves incorporating the results in a Streamlit app so that users can interact and view the time-series data using the model. There are two primary functionalities of the app:

* Prediction: Users are able to input employee data (including department, pay grade, and purpose of leaving) to forecast the chances of layoffs. The input is analyzed by the model and it outputs the highly likely cause of leaving based on the past data.
* Time-Series Analysis: Users are able to view trends in layoffs by time and can interact in order to view monthly and yearly trends in layoffs. This page provides users with insights into the changes in layoffs during different time spans, facilitating informed workforce planning decisions.

The Streamlit app is constructed, and it includes an easy-to-use platform through which the stakeholder can view both the predictive and time-series measures of the laying-off data.

**7. Conclusion and Recommendations**

Model Recommendation: Due to the findings of the machine learning models, Random Forest is the best choice of model to use in predicting employees being laid off. Its precision and accuracy, combined with its capabilities of working on tough data, render it a safe choice in the hands of the HR department and policymakers.

**Future Research:**

* Addition of Economic Indicators: Adding macroeconomic indicators (i.e., unemployment rate, GDP) to the data could strengthen the predictive ability of the model by increasing its robustness.
* Investigating Advanced Algorithms: Methods such as XGBoost or deep learning can further refine the model's performance by identifying non-linear trends that could elude typical machine learning models.
* Impact: This study illustrates the potential of using machine learning to examine and predict layoffs and provides a data-driven response to workforce management. These insights can enable organizations to devise proactive strategies that reduce the effects of layoffs and increase talent retention.

**8. Streamlit Code Integration**

Below is a summary of the incorporation of the Streamlit app in the project:

1. Model Loading: We load the trained machine learning model (best\_layoff\_model.pkl) and any other pre-processing components such as scalers and encoders into the app.

2. Prediction Functionality:

Users can also type in appropriate employee details, including pay grade, department, and reason for leaving.

The input is pre-processed (for example, one-hot encoded categorical variables), and the layoff cause is predicted by the model.

the result of the forecasting will appear as the likely cause of the layoff

**3. Time-Series**

* The dataset is read and preprocessed, converting the dates of separation to yearly and monthly durations.
* Monthly and yearly layoffs are charted and presented interactively so that users can examine trends.

**Final Thoughts**

This project incorporates predictive modeling and time-series analysis to present a comprehensive picture of layoffs in the US economy. By using machine learning models, time-series analysis, and interactive visualization, the project presents actionable insights to HR teams and policymakers alike. With the inclusion of a Streamlit app, the model becomes interactive and user-friendly, facilitating data-driven decision-making for workforce managers.

**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>