**MODEL PERFORMANCE REPORT**

**A green and black logo

Description automatically generated**

**People Analytics at Seagate: A Strategic Approach**

**to Predict Voluntary Churn and Optimize the Hiring Process**

**Master of Science in Business Analytics**

**Course: Experiential Projects**

**S24 – 004 – Group 4 – Project 2**

**Prof. Noah Zikmund**

**Group Members**

**Anubhav Dubey** (anubhav.dubey@colorado.edu)

**Amandeep Bhardwaj**

(amandeep.bhardwaj@colorado.edu)

**Sandeep Reddy Modugu**

(sandeep.modugu@colorado.edu)

**Sai Arvind Atluri** (saiarvind.atluri@colorado.edu)

Table of Contents

[Introduction 3](#_Toc164808575)

[Analysis 3](#_Toc164808576)

[Data Splitting and Setup 3](#_Toc164808577)

[Feature Engineering 5](#_Toc164808578)

[Models 6](#_Toc164808579)

[Random Forest (Model 1) 6](#_Toc164808580)

[XGBoost (Model 2) 7](#_Toc164808581)

[AdaBoost (Model 3) 8](#_Toc164808582)

[CatBoost (Model 4) 9](#_Toc164808583)

[Model Performance Comparison 10](#_Toc164808584)

# **Introduction**

In the dynamic field of Strategic Human Resources management, the traditional reliance on tools like Excel has proven to be valuable, and it now provides a compelling opportunity to work with real-time data. At Seagate, the Human Resources department has amassed a substantial database from various platforms, capturing everything from detailed job records to in-depth feedback on employee experiences. This extensive collection of data initiates the stage for us to embrace cutting-edge machine learning and predictive analytics technologies. By doing so, we can significantly enhance Seagate's strategic Human Resources initiatives. Particularly for non-manufacturing workers, the use of advanced analytics promises to transform these extensive datasets into actionable insights, ultimately driving more informed and judicious decision-making within the organization. Our commitment to these innovative technologies signals a pivotal shift towards more sophisticated and impactful HR practices.

Through the adoption of advanced machine learning and predictive analytics technologies, our current project has the potential to completely transform Seagate's HR data processes. Our strategic goals are centered on innovation and empowerment:

**Churn Rate Prediction:** We developed a predictive model that will accurately forecast rates of voluntary employee attrition. Utilizing historical employee data, this innovative approach will identify key turnover factors, enabling HR to implement proactive retention strategies that reduce voluntary churn.

**Hire Rate Forecasting to Offset Churn Implications:** To mitigate the financial and operational impacts of turnover, we created a model that will precisely predict the hiring rates needed to support business operations and growth. This model aims to establish a dynamic two-year hiring pipeline that enhances operational efficiency and ensures business continuity by integrating insights from attrition rates, recruitment cycles, and business growth projections.

With this project, Seagate aims to significantly improve its Human Resources capabilities and transform rich data into actionable insights that will increase the organization's adaptability and strategic effectiveness.

# **Analysis**

## **Data Splitting and Setup**

The dataset for our predictive analytics research consists of approximately 25,856 rows, derived from various HR records at Seagate. This cleaned dataset is stored in a data frame which includes information on both active and terminated employees.

* **Data Segmenting**

We identified terminated employees with 14,168 records through a defined termination date. This subset is crucial in providing historical instances of attrition, enabling our model to learn from real-world employee departures. Subsequently, the remaining i.e. 11,688 records which do not list a termination date, represent active employees or workers who are still employed. Analyzing this subset of data enabled us to understand the characteristics of employees who remain with the company.

* **Data Splitting for Modeling**

The terminated employee’s dataset was then divided into training and validation sets. Specifically, 80% of the data from the terminated employees was used to train the predictive model, with the remaining 20% reserved for evaluating the model’s prediction accuracy. This methodology ensures a comprehensive training process and allows for an accurate assessment of the model’s predictive capacity regarding employee turnover. Furthermore, the dataset pertaining to active employees was reserved as a holdout set. This prudent measure ensures that the final assessment of the model's predictive power is conducted with the utmost integrity. By utilizing this untouched data, the model's performance can be evaluated on information that has not influenced its learning process in any way, thereby offering a clear indication of its efficacy in real-world applications. It is this rigorous approach to model testing i.e. holding back a subset of the data from all previous phases of training and validation that helps in safeguarding against overfitting and provides confidence in the model's ability to generalize from its learned patterns to novel situations and datasets it has never encountered before.

A diagram of a company

Description automatically generated with medium confidence

## **Feature Engineering**

We carefully considered and designed a feature set for our predictive model that aims to anticipate employee terminations, capturing a variety of significant factors. The categorical features include "Job Title," "Job Function," "Job Category," "Pay Level," "Work Location," "Work Country," "Gender," "Employee Status," and "Generation." These features are intended to grasp patterns related to attrition by providing a comprehensive understanding of an employee's role, work conditions, and demographic background.

For numerical analysis, we included metrics such as "Compa Ratio," "Tenure," "Cost to Replace Employee Multiplier," and "Base Pay Mid-Point Annualized USD." These numerical characteristics enhance the model’s ability to predict terminations based on financial factors and job satisfaction by offering quantifiable insights into an employee’s tenure, compensation relative to industry standards, and the financial impact on turnover.

In our predictive model that was aimed at forecasting voluntary employee terminations, we methodically selected and engineered a set of features to enhance predictive accuracy. Notably, while ‘Termination Date’ initially served as our target variable, we introduced a refined binary target variable, ‘Is Voluntary’, to specifically identify voluntary terminations. This new target variable is derived by marking 'Voluntary Termination' in the ‘Termination Type’ field as 1, and all other types i.e. ‘Involuntary Termination’ , ‘Other Termination’ and ‘Release Termination’ as 0. This binary approach allows for a more nuanced analysis of voluntary attrition, which is often the most actionable from an HR perspective.

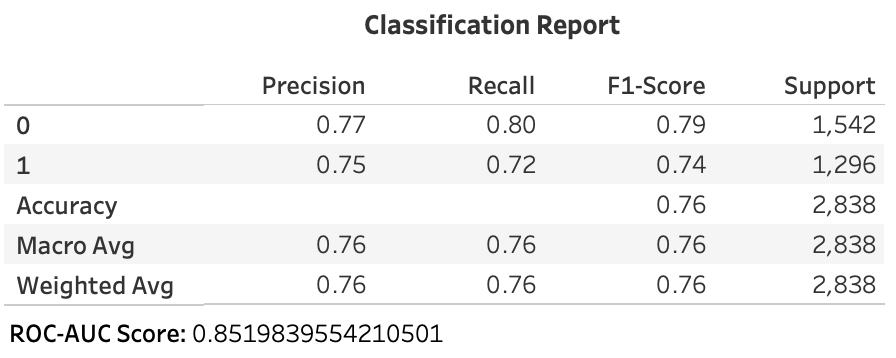
We consciously decided not to include ‘Termination Type’ and ‘Termination Reason’ in our broader feature set, despite their high correlation with the target variable. This decision was driven by the need to build a model capable of predicting terminations without prior knowledge of their context, thereby ensuring its utility in real-time prediction scenarios where the reasons or types of termination are not yet known. This approach ensures that the model can operate effectively in environments where only limited data is available at the point of prediction, thus increasing its practical applicability and robustness.

By judiciously selecting and constructing these features, we aimed to maximize our model's predictive accuracy and reliability in practical applications, making it more sensitive to underlying patterns and trends in the employee data.

# **Models**

## **Random Forest (Model 1)**

In our analytical exploration to understand the voluntary employee turnover at Seagate, we utilized a RandomForestClassifier as one of the four models to analyze this extensive dataset. During training, this ensemble model that is renowned for its high accuracy and robustness, constructs a multitude of decision trees and outputs the class that represents the average forecast, or mode of the classifications of each individual tree.



The classification report revealed a precision of 0.77 for class '0' (non-voluntary terminations) and 0.75 for class '1' (voluntary terminations), indicating that the model performed well across various criteria. The F1-scores were 0.79 and 0.74, respectively, and the model's recall of 0.80 for non-voluntary terminations and 0.72 for voluntary terminations was observed. The model's ability to differentiate between the various types of terminations was confirmed by a ROC-AUC Score of approximately 0.85, and its overall accuracy was 76%.

A screenshot of a computer

Description automatically generated

The confusion matrix provided gives us insights into the performance of a predictive model that distinguishes between voluntary terminations and other types of terminations among employees. In the matrix, there are 1,234 instances where the model correctly predicted voluntary terminations (true positives) followed by 933 instances where the model correctly predicted other terminations (true negatives). However, considerable misclassification errors were also noticed. Specifically, the model predicted 308 instances of voluntary terminations as other terminations (false negatives), and 363 instances of other terminations as voluntary terminations (false positives).

## **XGBoost (Model 2)**

The second pillar in our array of predictive models, the XGBoost Classifier, was integral to our multimodal approach to understanding employee turnover. XGBoost, renowned for its effectiveness and performance, optimizes speed and accuracy by applying a gradient boosting framework to decision trees. To preserve the integrity of our feature set, the model underwent a rigorous training regimen within a pipeline that ensured one-hot encoding of categorical variables and the imputation of missing data. After training, isotonic regression was employed to calibrate the model, further enhancing its reliability by aligning the projected probabilities more closely with observed outcomes. This calibration was validated using a stratified k-fold cross-validation technique, ensuring the model’s robustness across different data subsets.

A screenshot of a graph

Description automatically generated

With an impressive ROC-AUC Score of approximately 0.867 and an accuracy of nearly 78%, the XGBoost model demonstrated its superiority in distinguishing between voluntary and non-voluntary terminations. The classification report showed a well-balanced performance, with precision, recall, and F1-scores all ranging between 0.76 and 0.80.

A screenshot of a computer

Description automatically generated

The results of the XGBoost model's predictions on voluntary terminations, in comparison to all other forms of terminations, are captured in the confusion matrix displayed. It shows that the algorithm accurately predicted 1,237 cases of voluntary terminations. These true positives (TP) demonstrate the program's ability to identify employees who are most likely to resign voluntarily. Conversely, the model misclassified 305 instances where active employees were incorrectly predicted as voluntary terminations; these are known as false negatives (FN), indicating missed opportunities where talent could have been retained with intervention. However, the model also misclassified 308 cases as false positives (FP), wherein they were voluntary terminations, suggesting that HR interventions may have been erroneously planned. Lastly, it successfully distinguished between voluntary and other forms of terminations by correctly identifying 988 cases as non-voluntary terminations, or true negatives (TN).

## **AdaBoost (Model 3)**

The AdaBoost Classifier has been a crucial tool in our analytical toolbox as we work to optimize the prediction of voluntary employee turnover. Adaptive Boosting, or AdaBoost, is an ensemble strategy that constructs a robust predictive model by combining several weak learners. Each subsequent learner is adjusted to give priority to cases that its predecessors misclassified, thereby sharpening its focus on the more challenging aspects of the prediction task. AdaBoost has been incorporated into a comprehensive pipeline for this specific application, which includes preprocessing actions such as the imputation of missing data and one-hot encoding of categorical variables. The classifier’s accuracy of approximately 74.45% after training on a subset of the data is evidence of its capability, even in the challenging field of analyzing human behavior.

A screenshot of a graph

Description automatically generated

The performance metrics from the classification report show that the recall for both classes is 0.74, with precision scores of 0.78 for non-voluntary terminations and 0.71 for voluntary terminations. This results in respectable ROC-AUC Scores of 0.8306, which highlights the model's discriminative power, and F1-scores of 0.76 and 0.73, respectively.

A screenshot of a computer

Description automatically generated

When evaluating our AdaBoost Classifier's ability to forecast voluntary employee turnover, the confusion matrix provides a clear and comprehensive picture of its performance. It shows that the model was effective in identifying employees who were at risk of leaving voluntarily, as evidenced by the 1,148 cases (True Positives) of actual voluntary terminations that are correctly detected. However, it also shows that 394 real voluntary terminations were incorrectly labeled as other types (False Negatives), pointing to potential areas where the algorithm is overly conservative. Conversely, 331 employees were predicted by the algorithm to leave voluntarily when they did not (False Positives). This indicates a conservative approach that may require further fine-tuning to minimize overpredictions. Interestingly, the model successfully identified 965 cases of non-voluntary terminations (True Negatives), demonstrating its ability to discriminate between various termination scenarios.

## **CatBoost (Model 4)**

Among the models used to forecast voluntary employee turnover, the CatBoost Classifier stands out as an advanced algorithm adept at managing categorical data. With an accuracy score of approximately 77.66%, CatBoost employs gradient boosting on decision trees and is designed to deliver quick, scalable, and high-performance outcomes. Our model benefited from a structured preprocessing stage that included imputation and one-hot encoding. This was followed by a fitting procedure that leveraged CatBoost's built-in capability to handle categorical features effectively. Following training, we conducted a calibration phase using cross-validation and isotonic regression to align predicted probabilities with actual outcomes, enhancing the model’s reliability.

A screenshot of a graph

Description automatically generated

The classification report that was produced provides a balanced view, with recall rates of 0.78 and 0.77, and precision rates of 0.80 for non-voluntary and 0.75 for voluntary terminations, respectively. The F1-scores and an ROC-AUC Score of 0.863 further illustrates this balance, demonstrating a strong ability to differentiate between the two outcomes. The metrics derived from CatBoost offer significant insights that facilitate a comprehensive understanding of the model’s effectiveness within the broader framework of employee attrition study.

A screenshot of a computer

Description automatically generated

The CatBoost model's confusion matrix provides insight into how effectively the model differentiates between voluntary and involuntary employee terminations. The model has skilfully detected a considerable number of voluntary terminations, with 1,209 true positives, indicating strong sensitivity to patterns that may signal an employee's tendency to quit voluntarily. On the other hand, the 333 false negatives represent missed voluntary termination instances where the model's predictions did not align with actual worker behaviour. Conversely, the 301 false positives highlight areas where the algorithm might have raised false alarms; these are employees as to whom the model anticipated would voluntarily terminate but did not. Finally, the model's effectiveness in identifying stable employee scenarios is confirmed by the 995 true negatives, which demonstrate its ability to accurately identify those not inclined to terminate voluntarily.

# **Model Performance Comparison**

A screenshot of a graph

Description automatically generated

Despite XGBoost's marginally better performance metrics, we have chosen to employ the CatBoost classifier as our final prediction model after carefully analyzing several machine learning classifiers. CatBoost outperformed the other models with respectable precision (0.75), the highest recall (0.77), a strong F1 Score (0.76), and an outstanding ROC-AUC Score (0.86). These figures demonstrate CatBoost’s superior predictive performance, particularly in its ability to accurately detect voluntary terminations and its potential for balanced categorization.

Although XGBoost did provide better statistics, CatBoost's array of benefits makes a compelling case for its selection. Thanks to the algorithm's natural handling of categorical features, our complex dataset benefits greatly from reduced pre-processing time and preserved data integrity. Its ability to resist overfitting also suggests that the model will function reliably outside the testing environment in real-world applications.

Real-time deployment scenarios greatly benefit from CatBoost’s faster prediction time post-training, which is a critical issue in operational efficiency. CatBoost ensures that large-scale data can be processed more effectively, speeding up the training phase without sacrificing model quality, especially when combined with its improved GPU utilization.

The interpretability of the model cannot be compromised, and CatBoost’s user-friendly tools for analyzing the effects of features align with our requirements for transparency and adherence to the legal standards. Because of the platform’s ease of use, fine-tuning complex models do not require the steep learning curve that is sometimes associated with them. This shortens the development cycle and facilitates the transition from development to deployment.

No matter the variety of data encountered in future applications, our model will remain compatible due to CatBoost’s resiliency across different data distributions. With a balance between operational practicality and analytical quality, CatBoost classifier is the optimal choice for achieving our goals, as theoretically proven by all these criteria alongside the model’s proven real-time predictive performance.