**FINAL 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

[Business Understanding 3](#_Toc165628558)

[Problem Statement 3](#_Toc165628559)

[Overarching Objectives 3](#_Toc165628560)

[Data Overview 4](#_Toc165628561)

[Data Cleaning and Preparation Process 5](#_Toc165628562)

[Data Cleaning Phases 6](#_Toc165628563)

[Removing Unnecessary Columns 6](#_Toc165628564)

[Removing Unnecessary Rows 6](#_Toc165628565)

[Feature Engineering 7](#_Toc165628566)

[End Results 8](#_Toc165628567)

[Descriptive Statistics 9](#_Toc165628568)

[Data Visualizations 11](#_Toc165628569)

[Outliers vs. Non-Outliers of Numerical Variables 11](#_Toc165628570)

[Histograms 12](#_Toc165628571)

[**Employee Tenure** 12](#_Toc165628572)

[**Compa Ratio** 13](#_Toc165628573)

[**Base Pay** 14](#_Toc165628574)

[Scatter Plots 15](#_Toc165628575)

[**Tenure vs. Base Pay** 15](#_Toc165628576)

[**Compa Ratio vs. Base Pay** 16](#_Toc165628577)

[Correlation Heat Map 17](#_Toc165628578)

[Implications of observed changes 17](#_Toc165628579)

[Data Cleaning Conclusion 18](#_Toc165628580)

[Data Analysis 18](#_Toc165628581)

[Data Splitting and Setup 18](#_Toc165628582)

[Feature Engineering 20](#_Toc165628583)

[Models 21](#_Toc165628584)

[Random Forest (Model 1) 21](#_Toc165628585)

[XGBoost (Model 2) 22](#_Toc165628586)

[AdaBoost (Model 3) 23](#_Toc165628587)

[CatBoost (Model 4) 24](#_Toc165628588)

[Model Performance Comparison 25](#_Toc165628589)

[Final Model Recommendation 25](#_Toc165628590)

[2 Year Hiring Pipeline 26](#_Toc165628591)

# **Business Understanding**

**Objective:** Enhance strategic HR contributions to business leadership through the integration of Predictive Analytics and Machine Learning into data analysis processes.

**Baseline Status:** The HR team at Seagate captures a wide array of employee data across multiple systems, including basic employment details and extensive feedback on employee experience accumulated over years. However, the process of integrating, cleaning, and analyzing this data from disparate systems poses significant challenges. Current reliance on Excel for data analytics is inefficient, time-consuming, and lacks scalability for accommodating updates or new data inputs.

**Need:** A more efficient and consistent data analysis methodology is crucial in supporting strategic HR choices, such as identifying the primary drivers of voluntary churn, understanding its financial impact, and precisely estimating headcount for a two-year hiring pipeline. The emphasis is on non-manufacturing specialists, excluding executive positions and manufacturing specialists.

# **Problem Statement**

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.

# **Overarching Objectives**

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.

# **Data Overview**

Navigating through the integrity and accuracy of data is a fundamental aspect of any project that aims to extract meaningful insights from available information. While this process is time-consuming, but it is essential for sifting through the diverse and often complex datasets, typically sourced from various systems. These datasets frequently exhibit a range of issues, including errors, inconsistencies, and missing values, which can significantly distort the analytical outcomes. The iterative nature of data cleaning requires continuous refinement to meet the analysis-specific demands and achieve the necessary data quality. Hence, a comprehensive data cleaning regimen is crucial for the project team to derive reliable insights and make well-informed business decisions.

As part of the Seagate HR data analysis project, we have undertaken meticulous steps to structure the data in a manner conducive to insightful analysis and decision-making. This section delineates the sophisticated data preparation and cleaning methodologies we employed, while grounded in our collective domain expertise and extensive research.

Initiating the process, we evaluated the dataset to identify and eliminate irrelevant rows and columns. This pruning was based on detailed assessment of the data's relevance on analytical goals and the feasibility of imputation for missing information. Through leveraging our analytical acumen, we were able to tailor the dataset to align precisely with the project’s objectives, thereby notably enhancing the dataset’s quality and relevance.

Addressing missing values was a pivotal aspect of our data preparation, wherein we employed feature engineering to judiciously estimate values. This methodical approach ensured that each data element not only enriched our analysis but also solidified the conclusions of our study. Whether dealing with missing values in numerical fields, particularly those related to financial metrics, or confronting columns with extensive missing data, we consistently applied feature engineering to derive accurate estimates. By imputing these missing values using both categorical and numerical features, we preserved the dataset’s analytical integrity and financial depth while ensuring that no decision to remove data compromised the overall analysis. Consequently, this strategy fortified our dataset, enhancing its ability to contribute significantly to the robustness and comprehensiveness of our analysis. This approach corroborated that each piece of data principally contributed to our analysis, strengthening the validity and breadth of our conclusions.

Moreover, we labeled columns with strikingly missing yet analytically valuable data as 'Not Specified / Not Terminated,' enhancing the clarity and interpretability of our dataset. This labeling not only preserved the structural integrity of the dataset but also facilitated a more nuanced and contextual analysis.

Consequently, our data cleaning and preparation regimen, marked by strategic eliminations, exhaustive management of missing values, and targeted feature engineering, has established a robust foundation for the next stages of the Seagate HR data analysis project. Subsequent sections will outline the precise procedures we implemented, highlighting our dedication to a rigorous and data-driven analytical approach.

## **Data Cleaning and Preparation Process**

**Data Profiling and Preliminary Exploration:** To identify any issues with the quality of the data, we examined the structure of the Seagate HR dataset and the findings from this preliminary investigation were meticulously recorded, providing a foundational background for the subsequent cleaning processes.

**Addressing Missing Data:** Our examination revealed missing data across several columns. To address these gaps, we employed a combination of statistical techniques and domain knowledge, varying from simple deletion of inconsequential data points to complex imputation of values with potential analytical significance.

**Duplicate Identification and Resolution:** Duplicate records, which could skew the analysis were identified, and removed to ensure the uniqueness of each data point. The process and its implications were transparently documented to maintain integrity.

**Normalization and Data Transformation:** We transformed the data, encoding categorical variables and normalizing skewed distributions, to ensure comparability and readiness for advanced analytical techniques.

**Feature Engineering and Enrichment:** Through feature engineering, we expanded our dataset and introduced new variables, specifically in relation to financial metrics and employee involvement, to enhance the analysis and provide deeper insights.

**Outlier Detection and Treatment:** Outliers and anomalies were identified and addressed to ensure the accuracy of our data. Methods such as thresholding and data curation were used to manage these abnormalities.

**Ensuring Consistency and Reliability:** A rigorous quality assurance process was implemented. This phase ensured that the data retained its original integrity and remained a dependable and consistent foundation for our post-cleaning analyses.

**Documentation of the Final Dataset:** To conclude the data cleaning phase, we diligently documented the final state of the dataset, including the number of variables and observations retained or removed, as well as any new variables added during the cleaning process.

**Data Cleaning Validation:** Final validation checks were conducted to verify the detailed completion of all data cleaning processes and to confirm the alignment of the cleaned data with the project’s objectives and standards.

# **Data Cleaning Phases**

## **Removing Unnecessary Columns**

We removed any irrelevant or unnecessary columns from the Seagate HR dataset to make it more streamlined and insightful for further research. Columns such as Job Code and Anonymized ID were eliminated, as they did not provide any analytical value for predictive modeling or aggregate assessment. The Job Group was also omitted due to its data largely overlapping with other job-related characteristics. Furthermore, Work Region was deemed unnecessary because Work Location and Work Country adequately represent the geographic information required for our research. Given the more detailed Tenure data, Tenure Bucket was considered redundant for our modeling endeavors, which favored comprehensive continuous data over its categorized summary.

A special case was the Work Structure column, which contained a significant number of null entries. Although we could have imputed these as Not Specified, it was determined that removing this column would not substantially impact the analytical results or the robustness of our predictive models. This decision was made with the objective of preserving data integrity and modeling efficiency.

A thorough review of each column's contribution to our analysis, guided by our domain expertise, enabled this strategic pruning of the dataset, ensuring the remaining information was ideally structured for our analytical goals and free of unnecessary complexity.

## **Removing Unnecessary Rows**

A close-up of a data

Description automatically generated

A screenshot of a computer

Description automatically generated

During the data cleaning stage of our analysis, we carefully pruned the Seagate HR dataset to remove rows that were not necessary for our analysis. This process involved the elimination of rows based on logical consistency and the presence of null values. Specifically, we removed a single record that lacked a value in the 'Generation' column, as this feature is crucial to our analysis. Additionally, we discarded five entries, where the Pay Level data was missing, assessing that this would not remarkably impact our overall model.

Upon a detailed review of the Compa Ratio statistic, which compares the compensation level to the market median, we identified 120 entries with a ratio exceeding 2, materially above the expected range. Considering that such high figures are infrequent and could indicate extreme outliers or data entry errors, we opted to exclude these records to maintain the coherence of our analysis. Additionally, we identified 12 rows in the dataset where the base pay for the Russian Federation was recorded as zero. Given that these entries constituted the entire data set for this work country, and it isn’t practicable to accurately predict the values, we decided to remove them.

Furthermore, we implemented a logical check to identify any records where an employee's termination was noted without an accompanying reason. These entries were extracted from our dataset and removed. Since, it was a single entry, no effect on the analysis was noticed.

These actions resulted in a dataset that is both logically consistent and analytically robust. Our dataset now provides a well-organized and reliable foundation for subsequent phases of our project, as verified by the updated descriptive statistics reflecting these changes. This tailored dataset is poised to deliver a more accurate and focused analysis.

## **Feature Engineering**

To address the missing values in our Seagate HR dataset, we employed a consistent set of features to impute both Base Pay and Compa Ratio. These features included Tenure, Cost to Replace Employee Multiplier, Job Title, Job Function, Job Category, Pay Level, Work Location, Work Country, Gender, Employee Status, and Generation. By consistently using these features, we ensured that the predictive models for the two financial metrics were built on the same underlying data dimensions, allowing for a reasoned and uniform approach to data imputation across the dataset.

This common set of features enhanced the feature engineering phase and maintained consistency in the analytical process, enabling us to leverage the correlations and patterns within the data to accurately predict and rectify missing values. Through a systematic approach, we enhanced the robustness of our imputation process and derived accurate and reliable values for Base Pay and Compa Ratio, thereby improving the dataset's analytical utility and completeness.

We developed a Python machine learning pipeline that used a Random Forest Regressor for predictive imputation and preprocessing transformations to operationalize this approach. To tackle the heterogeneous nature of our dataset, the preprocessor distinguished between numerical and categorical data and treated each appropriately, applying a One-Hot Encoder to categorical data and sending numerical data through. To anticipate missing Base Pay data, the model was trained on observations with non-missing target values. Ultimately, the dataset's zero values were substituted with these projections, maintaining the dataset's overall coherence, and setting it up for insightful analysis. This methodical, technologically advanced outlook supports our belief that the dataset is prepared for the next phases of our study.

A screenshot of a presentation

Description automatically generated

**Note:** The same set of 1,448 records contain null values in both 'Compa Ratio' and 'Base Pay'.

Additionally, to enhance the comparability of financial data across geographical locations, we created a new column in the Seagate HR dataset that leverages the exchange rates as on date to convert Base Pay values across different country denominations to USD. This step ensures that all base pay figures are standardized, allowing for a more accurate and fair analysis of compensation levels irrespective of the employee's work location. By doing so, we can perform cross-regional comparisons and analyses more effectively, contributing to a more rational understanding of global compensation trends within the organization.

## **End Results**

The table below provides a snapshot of the data cleaning steps undertaken for each column in our dataset. Identifiers and non-essential fields were removed, missing values were addressed through deletion or imputation, and incorrect entries were corrected streamlining our data for precise and focused analysis.



# **Descriptive Statistics**

A table comparing the original and cleaned datasets for Base Pay by Work Country is displayed in the image below. It provides insights into the distribution and central tendency of compensation numbers both before and after data cleaning. It does this by breaking down counts, means, standard deviations, and the range of base pay across percentiles for each country. By allowing an evaluation of the impact of cleaning on the statistical features of the data, this comparative perspective ensures that the cleaned dataset is suitable for further analysis.

A screenshot of a table

Description automatically generated

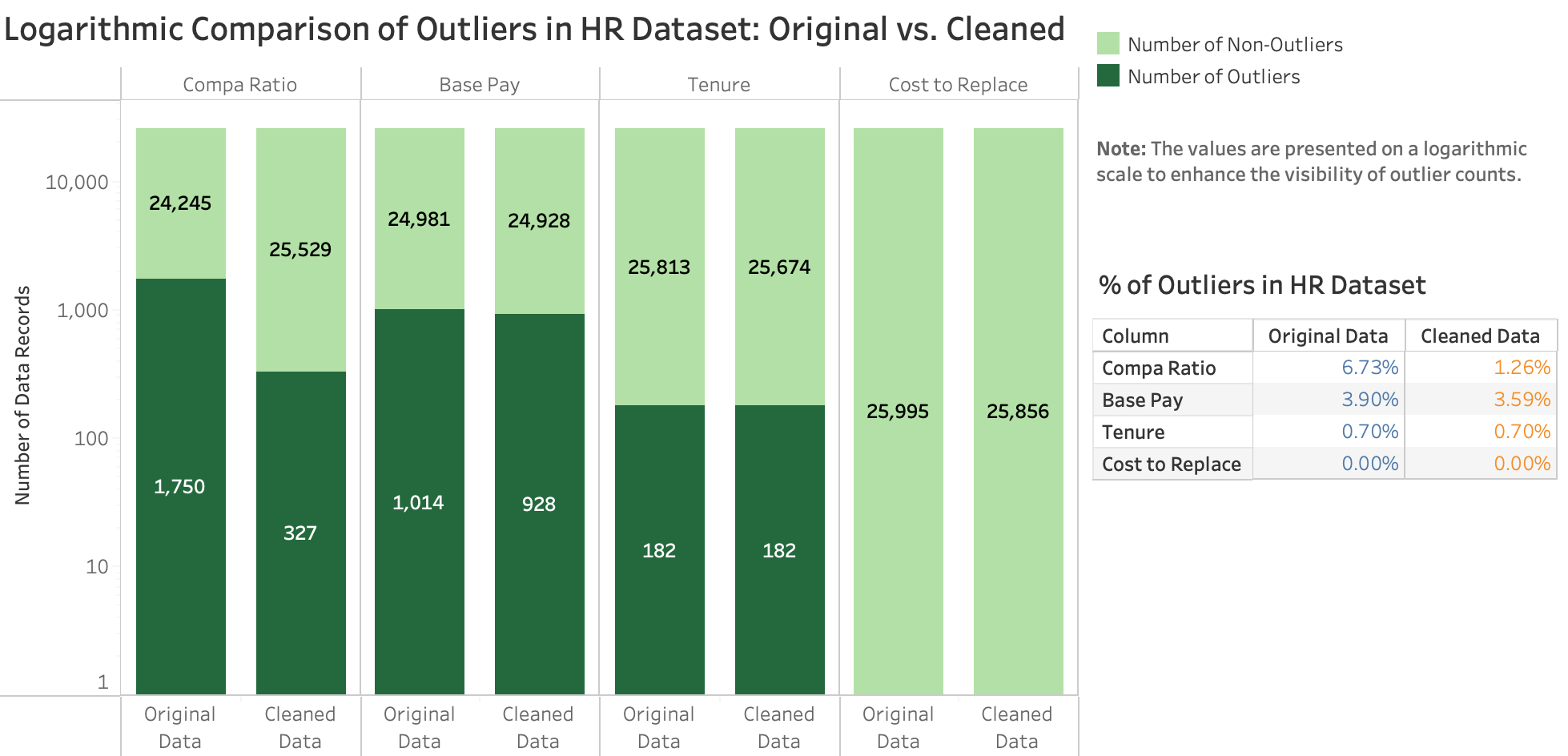
A screenshot of a data

Description automatically generated

Three metrics Compa Ratio, Cost to Replace, and Tenure are compared statistically between the original and cleaned datasets in the above table. Notably, data cleaning has raised the minimum value indicating the removal of extreme outliers and lowered the standard deviation in the Compa Ratio suggesting a decreased spread of data around the mean. As we bypassed the data cleaning owing to low percentage of outliers, the average costs to replace an employee and their tenure stayed pretty much the same. The spread of the numbers from smallest to the largest, didn't really change either. Indicative of our approach, we effectively eliminated extreme or unusual cases, maintaining the overall data's equilibrium and integrity. Moreover, these statistics demonstrate the effectiveness of the cleaning procedure, particularly in terms of normalizing the Compa Ratio variable which is essential for supplementary analysis.

# **Data Visualizations**

## **Outliers vs. Non-Outliers of Numerical Variables**



The above graph emphasizes on outlier and non-outlier values across numerical columns that considerably deviate, being much higher or lower than the rest. It demonstrates that Compa Ratio and Base Pay fields portray an apparent decrease in outliers after cleaning, indicating that data points which could have been errors or extreme values, have been addressed. Conversely, the Tenure and Cost to Replace columns exhibit minimal change in the number of outliers, suggesting that these areas either lacked predominant anomalies or that the cleaning criteria do not apply to these metrics.

When working with large datasets, employing a logarithmic scale is particularly beneficial as it aids in comparing counts across a broad range of values, thereby simplifying the visualization of outlier reduction after data cleaning. Notably, the outlier percentage for 'Compa Ratio' decreased from 6.73% to 1.26%, reflecting a more normalized dataset that is better suited for nuanced analysis.

## **Histograms**

### **Employee Tenure**

The histograms below illustrate the distribution of employee tenure prior and subsequent to data cleaning process. The striking similarities in their distributions and shapes suggest that the tenure data was already of high quality, with few errors or outliers to address. The minimal changes observed post-cleaning lends further support to this, indicating that no additional refinement measures were necessary.

A graph of blue and white lines

Description automatically generated

In the pre-cleaning histogram, there are notable number of employees with shorter tenures that gradually reduce as tenure length increases. This pattern is typical in workforce data, reflecting the dynamics of employee turnover and the higher influx of new hires compared to long-term employees.

A graph of green and black lines

Description automatically generated with medium confidence

The post-cleaning histogram displays a similar pattern, reinforcing the notion that the data cleaning was judicious, targeting only the most critical data points for removal or correction without disturbing the general trends within the Tenure column.

### **Compa Ratio**

A white background with numbers and text

Description automatically generated

The pre-cleaned histogram reveals that a significant number of both active and terminated employees had a Compa Ratio below 1, indicating that their pay was below the market median. The presence of a large initial bar at 0.50 denotes that a considerable portion of the workforce received significantly less than the market rate. As the histogram progresses to the right, the bars diminish in height, showing that fewer employees have higher Compa Ratios.

A graph of green and black bars

Description automatically generated

The post-cleaned histogram exhibits a remarkable transformation. The distribution of Compa Ratios is more uniform with a central peak. This pattern is characteristic of a normal distribution, which often emerges naturally in a variety of real-world scenarios and is generally preferred for analytical modeling. The data cleaning process appears to have moderated the extreme values that were skewing the original distribution.

### **Base Pay**

A graph of data in blue and white

Description automatically generated with medium confidence

The pre-cleaning histogram reveals the range of base salaries across the workforce. The bulk of employees, both active and terminated earned between $0 and $50,000. As base pay increases, fewer employees fall into each successive bracket indicating a substantial skew towards the lower end of the base pay range.

A graph of green and black bars

Description automatically generated with medium confidence

In the post-cleaning histogram, although there is still a peak at the lower end, the distribution across different base pay ranges is now more uniform. This indicates that the cleaning process has addressed some of the extreme values in base pay, but the overall trend of most employees earning lower base pays persist.

## **Scatter Plots**

The scatter plots below provide a granular view by illustrating how the tenure and compa ratio correlate with actual base pay across the organization and how data cleaning can influence these relationships.

## 

### **Tenure vs. Base Pay**

A screen shot of a graph

Description automatically generated

In the pre-cleaned plot, the data points are widely dispersed indicating variances in base pay across different countries and tenure levels. The spread suggests that base pay can vary predominantly for similar tenure lengths, possibly due to company specific pay scales or country-specific economic factors.

A screen shot of a graph

Description automatically generated

The data points in the post-cleaned plot seem more closely grouped around certain tenure ranges, indicating a consistent trend. This stipulates that the data cleaning process has removed or imputed some of the outliers, inaccurate entries, and null values in base pay column.

### **Compa Ratio vs. Base Pay**

A graph of a number of countries/regions

Description automatically generated with medium confidence

The pre-cleaned plot above, displays striking variation in base pay across all compa ratio levels suggesting that a substantial number of employees were compensated at rates below or above the market average.

A graph of colored circles

Description automatically generated with medium confidence

In the post-cleaned plot, the range of base pay across various compa ratios appears to be less variable and more clustered around the market rate i.e. compa ratio of 1. The consolidation of these data points demonstrate that the data cleaning process has eliminated some of the more extreme discrepancies in base pay and compa ratio columns.

## **Correlation Heat Map**

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

Concerning the pre-cleaned heat map, the compa ratio portrays minimal correlations with all other variables, indicating an absence of linear relationships. Tenure shows a slight positive correlation with the Base Pay and Cost to Replace metrics, suggesting a marginal positive relation. The strongest correlation though still nominal is between Base Pay and Cost to Replace columns, hinting that as base pay increases, the cost to replace an employee may rise proportionately.

The post-cleaned heatmap reveals a slightly more pronounced weak positive correlation between Compa Ratio and Tenure. The increased negative correlations of Compa Ratio with Base Pay and Cost to Replace columns suggest that a higher Compa Ratio may correlate with lower values in these variables. Furthermore, the weak positive correlation between Base Pay and Cost to Replace fields intensified marginally, indicating that data cleaning process may have unveiled a more pronounced association between these metrics.

# **Implications of observed changes**

The rigorous data cleaning process has notably enhanced the quality of the Seagate HR dataset, carrying substantial implications for the reliability and validity of subsequent analyses. Below is a detailed examination of these changes:

**Improved Data Quality:** Outliers, particularly in the Compa Ratio and Base Pay variables, have been eliminated to the most extent. This reduction in noise helps prevent the distortion of trend analyses and predictive modeling. The Compa Ratio now clusters more tightly around the market rate, manifesting that the data more accurately reflects the company's compensation landscape.

**Enhanced Analytical Precision:** By judiciously removing irrelevant rows and columns, the dataset has become more streamlined, potentially leading to more precise machine learning models. This refined dataset helps to improve the model’s generalizability while reducing the risk of overfitting.

**Increased Representativeness:** With the imputation of missing values through utilizing a consistent set of features, the dataset now presents a fuller picture. This is crucial for cross-country comparisons, where incomplete data could lead to biased interpretations of compensation trends.

**Better Decision Making:** The enhanced Seagate HR dataset fosters superior decision-making by providing a reliable, accurate foundation for supplemental analyses. This refinement ensures that decisions align with market trends and data integrity, leading to better alignment with project’s goals.

# **Data Cleaning Conclusion**

To encapsulate, the data cleaning process applied to the Seagate dataset exemplifies a conscientious approach that balances the need for advanced analytics while maintaining the original data's integrity. The resulting dataset is not only cleaner but also primed to yield more insightful analyses.

The salient outcomes below reflect the effectiveness of the data cleaning process:

**Predictive Analysis Readiness:** The dataset is now better suited for advanced analytical techniques, such as hiring optimization and predictive modeling for voluntary turnover. The models constructed will be underpinned by the cleanest data possible, enhancing precision.

**Foundation for Strategic Decisions:** With more transparent data, we can make more informed strategic decisions. The refined dataset offers a dependable foundation for critical initiatives such as developing retention strategies.

The overall data cleaning process underscored the importance of data quality and its direct impact on analytical conclusions. The Seagate HR dataset now stands as a testament to the power of thorough data cleaning. The implications of the refined data resonate through every aspect of the analysis that follows, ensuring that any conclusions drawn are based on the most sound and robust information available.

# **Data 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 employees' dataset was then divided into training and validation sets. Specifically, 11,334 rows, representing 80% of the data from the terminated employees, were used to train the predictive model, with the remaining 2,838 rows, or 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, encompassing 11,668 rows, 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.

A screenshot of a graph

Description automatically generated

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.

# **Final Model Recommendation**

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.

# **2 Year Hiring Pipeline**

Through analyzing the Seagate HR dataset, we project that 1,616 people will leave voluntarily over the next two years, with an estimated financial impact of $66.8 million. This prediction includes both direct and indirect costs, such as severance pay, diminished productivity, and the cost of knowledge drain, based on typical salaries, turnover rates, and hiring multipliers for specialized professions.

In response, Seagate should strategically allocate a considerable sum for hiring, training, and onboarding processes to address this issue. Additional incentives, including relocation packages and signing bonuses, should be offered to attract top talent and keep recruiting procedures competitive.

Seagate should prioritize its budget to support the creation of new positions that are essential for the growth and sustainability of the organization, as well as to cover the costs of replacing key roles. By closely analyzing turnover trends to pinpoint high-risk areas, Seagate should adopt targeted strategies aimed at reducing attrition and boosting employee retention. Such strategies could involve enhancing workplace conditions, providing opportunities for professional growth, and offering competitive benefits packages.

Furthermore, Seagate should maintain an adaptive financial approach that starts with these projections, enabling the company to adjust dynamically to the evolving market and business environments. By ensuring a workforce that is in alignment with its strategic objectives, Seagate can maintain its agility and responsiveness, regularly fine-tuning its strategies based on real-time data and emerging trends. To support sustained performance and operational efficiency, Seagate must continue to exercise prudent financial management and detailed planning.