# Sprint Retrospective

## SEAGATE EMPLOYEE CHURN PROJECT


Submitted by:

Charles Biddle Porter
charles.biddleporter@colorado.edu
Harshit Gole
harshit.gole@colorado.edu
Manan Dhanteja
Manan.dhanteja@colorado.edu

Duc Hoang
duc.hoang@colorado.edu
Kirti Vatsh
kirti.vatsh@colorado.edu

# TABLE OF CONTENTS

## PRODUCT BACKLOG

Following are all the user stories picked by the development team in Inception phase, Sprint 1, and Sprint 2.

*Table 1: Completed user stories from Product Backlog*

| Sprint Name | Tasks | User Stories | Size |
|---|---|---|---|
| Inception | Problem Statement | As a User, I want to comprehend the Problem Statement, so that I can better address the challenges faced by Seagate's HR team regarding predictive analytics of employee churn. | S:20 |
| | Objective | As a User, I want to know the Objective, so that I can focus on developing a scalable, integrated data analysis framework tailored to Seagate's HR requirements | S:20 |
| | Scope | As a User, I want to grasp the Scope, so that the team understands the tasks required to achieve the goal of this project. | S:20 |
| | Assumptions | As a User, I want to be aware of the Assumptions, so that I can base my analysis and predictions on a consistent understanding of factors affecting employee churn. | S:20 |
| | Metrics/KPIs | As a User, I want to understand Metrics/KPIs, so that I can measure the efficiency of data integration, process automation impact, and forecasting accuracy effectively. | S:20 |
| | Risks | As a User, I want to acknowledge the Risks, so that I can mitigate the risk of deriving contradictory insights and ensure the quality of decision-making. | S:20 |
| | Benefits | As a User, I want to identify the Benefits, so that I can contribute to streamlined processes for HR data analysis and enhance precision in decision-making and future planning. | S:20 |

| | | | |
|---|---|---|---|
| | Agile Roles | As a User, I want to know about the Agile Roles, so that I can understand my role and responsibilities within the project team and how it fits into the agile methodology. | S:20 |
| | Key takeaways | As a User, I want to learn from the Key Takeaways, so that I can enhance my adaptability, collaboration, and understanding of iterative progress in project management. | S:20 |
| Sprint 1 | Understanding Data | As a User, I want to grasp the process of Understanding Data, so that I can accurately interpret data for better decision-making. | S:20 |
| | Metadata | As a User, I want to learn about Metadata, so that I can use this information to manage and organize data effectively. | M:40 |
| | Descriptive analysis | As a User, I want to perform descriptive analysis, so that I can summarize the main features of our data, aiding in initial insights and further analyses. | L:60 |
| | Issues with data | As a User, I want to identify Issues with data, so that I can address data quality and integrity problems early in our process. | L:60 |
| | Team discussion | As a User, I want to participate in Team discussion, so that I can collaborate with the team on addressing data issues and strategizing on analyses. | S:20 |
| Sprint 2 | Literature survey | As a User, I want to engage in Literature survey, so that I can base our modeling efforts on established research and methodologies. | M:40 |
| | Model 1 - Preliminary stage | As a User, I want to develop and refine predictive Model 1 at various stages (Preliminary, Intermediate, Final), so that I can help in accurately forecasting employee churn and optimizing the hiring pipeline. | XL:80 |
| | Model 1 - Intermediate stage | | XL:80 |

| | | | |
|---|---|---|---|
| | Model 2 - Preliminary Stage | As a User, I want to develop and refine predictive Model 1 at various stages (Preliminary, Intermediate, Final), so that I can help in accurately forecasting employee churn and optimizing the hiring pipeline. | XL:80 |
| | Model 2 - Intermediate stage | | XL:80 |
| | Team Discussion - M1 and M2 Preliminary Stage | As a User, I want to participate in Team Discussion for M1 and M2 Preliminary Stage, so that I can contribute to the initial development and setup of our models, ensuring a solid foundation for further refinement. | S:20 |
| | Team Discussion - M1 and M2 Intermediate Stage | As a User, I want to engage in Team Discussion for M1 and M2 Intermediate Stage, so that I can collaborate on analyzing interim results, addressing any issues, and making necessary adjustments to our predictive models. | S:20 |

Following tasks are still pending from sprint 2:

*Table 2: Pending user stories from Sprint 2 to be completed in Sprint 3.*

| Sprint Name | Tasks | User Stories | Size | Percentage completion |
|---|---|---|---|---|
| Sprint 2 | Model 1 - Final Stage | As a User, I want to develop and refine predictive Model 1 at various stages (Preliminary, Intermediate, Final), so that I can help in accurately forecasting employee churn and optimizing the hiring pipeline. | XL:80 | 20% |
| | Model 2 - Final Stage | As a User, I want to develop and refine predictive Model 2 at various stages (Preliminary, Intermediate, Final), so that I can help in accurately forecasting employee churn and optimizing the hiring pipeline. | XL:80 | 20% |
| | Team Discussion - M1 and M2 Final Stage | As a User, I want to be involved in Team Discussion for M1 and M2 Final Stage, so that I can help finalize our models, ensuring they are optimized, accurate, and ready for deployment in forecasting and decision-making processes. | S:20 | 0% |

# LESSONS LEARNED

## A. WHAT WENT WELL

1. Sprint 1: The descriptive analysis provided insights into the general distribution of salary data. It revealed an unusual number for salary deviation, indicating that salaries were not scaled uniformly. This issue was brought to the client's attention, and they provided a conversion rate to normalize the salaries in USD. This solution was implemented at the start of sprint 1.
2. Sprint 1: The exploratory time analysis during sprint 1 was conducted effectively. We addressed missing data and applied one-hot encoding to create a preliminary data report, which included a correlation heatmap. The analysis was enhanced by using histograms to visualize data distributions, providing clear insights into how data values were spread over time.
3. Sprint 2: Conducting a literature survey was helpful for the project. Engaging with established research ensured that the project is grounded in reliable theories and practices, enhancing both credibility and effectiveness. This foundational understanding helped in designing and implementing the project based on proven strategies rather than conjecture.

## B. WHAT DID NOT GO WELL

1. Sprint 1: The selection of features presented significant challenges, particularly in balancing computational costs with the relevance to our project deliverables. A critical delay was experienced while waiting for the client to provide currency conversion rates for the base pay, which became a universal bottleneck. Additionally, although we considered employing techniques like Principal Component Analysis (PCA) for dimensionality reduction, these were not adequately utilized. This underutilization prevented us from achieving a more streamlined and pertinent dataset, which would have reduced computational demands.
2. Sprint 2: Determining an effective method to divide the training and test sets posed considerable difficulties. We struggled to establish a non-arbitrary, non-iterative standard for splitting the dataset into training, testing, and validation segments. Despite the computational cost, a more structured and iterative approach to dataset splitting could have been developed to maximize model accuracy and minimize overfitting. Furthermore, deriving clear HR insights and adequately measuring and addressing overfitting between training and validation datasets was challenging. Exploring a broader range of ensemble

learning techniques beyond just random forest classifiers might enhance the accuracy of our model.

3. Sprint 2: During the construction of the AI model using AdaBoost and Random Forest methods, we encountered problems due to inconsistencies in categorical levels between the training and test datasets. Specifically, the "Gender" column in the training dataset included four categories (M, F, D, O), whereas the test set contained only three (M, F, O). This mismatch hindered the proper functioning of the machine learning model, as the algorithm could not handle the unexpected discrepancy in data structure.

## C. WHAT COULD BE IMPROVED ON?

1. Sprint 1: Feature engineering is crucial in predicting employee churn because it enhances model accuracy and effectiveness. By transforming and creating new features from existing data, the model gains access to more relevant and insightful information, which can significantly impact its predictive power. For instance, deriving categorical ranges from continuous tenure data or calculating interaction effects between variables like salary adjustments and performance ratings can uncover subtle patterns that directly influence churn. Additionally, well-crafted features help the model handle complexities of human behavior and organizational dynamics more adeptly, leading to more robust and actionable insights. Overall, effective feature engineering not only improves the model's performance but also provides strategic value by highlighting key factors contributing to employee turnover, thus enabling targeted interventions to improve retention.
2. Sprint 1: In the analysis of employee churn data, a deeper comparative study using statistical tests such as t-tests for continuous variables and chi-squared tests for categorical variables could substantially enhance our understanding of the differences between employees who leave and those who stay. This approach would allow for a more rigorous examination of how distinct features correlate with employee turnover.
3. Sprint 2: The process for team discussions on the models we've created can be significantly improved. Up to now, our meetings have consistently been constrained by time limitations, largely because we are simultaneously developing multiple models to compare their accuracy. To address this, we could implement a more structured agenda for our discussions, prioritizing key issues and decisions.