Salifort Motors: Employee Turnover Prediction Project

PACE Strategy Document

Google Advanced Data Analytics Capstone

Introduction

I will use this PACE strategy document to record my decisions and reflections as I work through this capstone project. This document will serve both as a guide to help me think through my responses and reflections at different stages of the data analysis process, and as a valuable resource I can reference in the future to support my growth as a data professional.

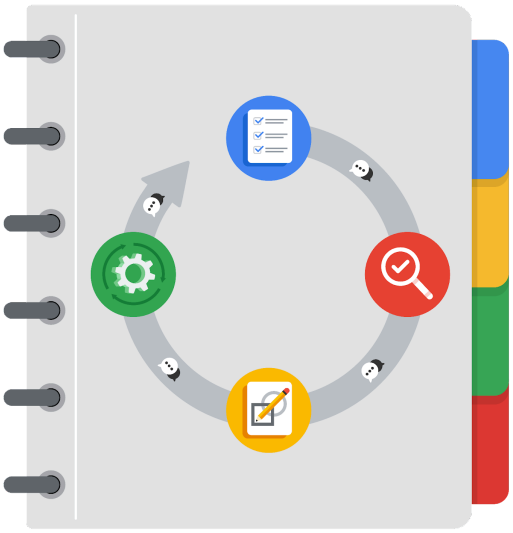
Portfolio Project Recap

Many of the goals I accomplished in my individual course portfolio projects are integrated into this Advanced Data Analytics capstone project. These include:

- Creating a clear and structured project proposal
- Demonstrating my understanding of Python's form and function
- Using Python to load, explore, extract, and organize information through custom functions
- Organizing and analyzing a dataset to uncover meaningful insights and tell the underlying "story"
- Developing a Jupyter notebook for exploratory data analysis (EDA)
- Computing descriptive statistics.
- Evaluating the model to assess its performance
- Applying machine learning techniques in a notebook environment to solve a defined problem
- Communicating results effectively by summarizing findings in an executive summary for external stakeholders

This capstone brings together all the skills I've developed across the program, allowing me to apply them in a cohesive, real-world project.

THE PACE WORKFLOW



[Alt-text: The PACE Workflow with the four stages in a circle: plan, analyze, construct, and execute.]

I will demonstrate to the company's HR team how I would apply the PACE workflow to the upcoming Salifort Motors project. For each question presented in this PACE strategy document, I will provide a structured and thoughtful response aligned with the corresponding stage of the PACE framework — Plan, Analyze, Construct, and Execute. This approach ensures clarity in my methodology, highlights my strategic planning skills, and illustrates a well-organized path from problem understanding to actionable insights.

Project Tasks

The following questions have been identified as essential to the Employee Turnover Prediction Project. I will address each question by aligning it with the most appropriate stage of the PACE workflow — Plan, Analyze, Construct, or Execute — based on the specific stage of the project workflow to which the question belongs. To ensure each response is informed and well-reasoned, I will draw on relevant materials from the project notebook, the PACE framework, and best practices for communicating analytical objectives.

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?

The key stakeholders for this project are: - Senior Leadership Team: They will be the primary recipients of the insights and progress updates throughout the project. Their role is crucial in making high-level strategic decisions based on the analysis. - HR Department: I will work closely with the HR department for domain expertise and data access. Additionally, I will collaborate with HR to share insights related to the model development, analysis, and potential action steps for improving employee retention. They will also provide valuable input on the interpretation and application of the model's findings. - Managers: As direct supervisors of employees, managers play a key role in influencing retention through day-to-day interactions and team dynamics. The insights generated from this project can help them better understand the warning signs of potential attrition within their teams and guide them in implementing targeted interventions to boost employee engagement and satisfaction.

- What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger business need?

I am trying to address the problem of employee turnover, whether due to voluntary resignations or involuntary layoffs. In both cases, turnover poses a significant issue for the company. When employees leave, the company incurs losses related to the initial investments made in their training, onboarding, and integration. Additionally, the process of interviewing, hiring, and training replacements adds to the overall cost of turnover. To mitigate these financial and operational impacts, the goal is to retain employees rather than continuously hiring new ones. By building a predictive model—whether it's logistic regression, random forest, or XGBoost—I aim to identify which employees are more likely to leave based on various features in the dataset. This can help the company take proactive measures to improve retention and reduce turnover costs.

- What questions need to be asked or answered?

At the initial stage of the project, several foundational questions need to be addressed to ensure the data is clean, well-understood, and ready for modeling:

- Are there any missing values, and how should they be handled?

- Are there any duplicate records that need to be removed?

- Are there outliers present in the dataset? If so, which mitigation techniques should be applied?

- What is the distribution of values across the different fields?

- Which features are most correlated with the left variable (i.e., employee attrition)?

- Are there strong correlations between certain variables that may impact model performance or interpretation?

- What types of visualizations can help explore these questions effectively?

Answering these questions will lay the groundwork for effective analysis, feature selection, and model development.

- What resources are required to complete this project?

To complete this project, the following resources are required:

- Python programming language and relevant libraries/packages

- Visual Studio Code (VS Code) as the primary integrated development environment (IDE)

- Jupyter Notebook extension within VS Code for writing and running code interactively

- GitHub for version control and sharing the project

- The Salifort HR dataset as the primary data source

Python Libraries and Packages Used:

- For data manipulation:

  - numpy

  - pandas

- For data visualization:
  - matplotlib.pyplot
  - seaborn

- For displaying all columns in dataframes:
  - pandas display option: `pd.set_option('display.max_columns', None)`

- For data modeling:
  - xgboost (XGBClassifier, XGBRegressor, plot_importance)
  - sklearn.linear_model (LogisticRegression)
  - sklearn.tree (DecisionTreeClassifier)
  - sklearn.ensemble (RandomForestClassifier)

- For model evaluation and utility functions:
  - sklearn.model_selection (GridSearchCV, train_test_split)
  - sklearn.metrics (accuracy_score, precision_score, recall_score,
    f1_score, confusion_matrix, classification_report,
    roc_auc_score, roc_curve, precision_recall_curve,
    average_precision_score, auc, ConfusionMatrixDisplay)
  - sklearn.tree (plot_tree)

- For saving trained models:
  - pickle

- What are the deliverables that will need to be created over the course of this project?

- The project proposal
- Advanced exploratory data analysis (EDA) visualizations and a comparative analytical study

- Machine learning model development and evaluation report

- The executive summary for technical and non-technical stakeholders

Get Started with Python

- How can you best prepare to understand and organize the provided information?

- I will begin by exploring the dataset using functions like `info()`, `describe(include='all')`, and `head()` to understand its structure and basic statistics.

- I will check for missing values, outliers, and duplicates, and apply appropriate mitigation techniques where necessary.

- I will examine data distributions through visualizations such as boxplots and histograms.

- This initial exploratory data analysis will lay the groundwork for more advanced visualizations and the comparative analytical study later in the project.

- What self-review codebooks will help you perform this work?

- The code notebook I create throughout the project will serve as a key reference, containing the questions that arise during the analysis as well as solutions and approaches.

- The PACE strategy document will also guide my reflection and decision-making, helping me stay aligned with the project objectives and analytical process.

- What are a couple additional activities a resourceful learner would perform before starting to code?

A resourceful learner would perform a couple of additional activities before starting to code:

- Plan the project workflow to ensure a clear roadmap for the coding process.

- Scrutinize the data to get an initial understanding of the structure and content of the dataset.

- Review the probable packages and libraries that will be used based on the initial understanding of the project.

- Set up the project folder and repository or notebook, ensuring version control is in place, typically on a hosting platform like GitHub, to track changes and maintain versioning.

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?

The dataset is comprehensive and includes various employee-related factors that are relevant for predicting employee turnover. Below are the key data columns and their types, along with the most relevant features for the analysis:

- Continuous Variables:

- satisfaction_level: A float representing the employee's job satisfaction level. This is a key feature to assess employee engagement and its impact on turnover.

- last_evaluation: A float indicating the employee's last evaluation score. This feature provides insights into recent performance reviews and might correlate with turnover.

- Count-Based Features:

- number_project: An integer representing the number of projects the employee is involved in. This feature can offer insight into the employee's workload and job engagement.

- average_monthly_hours: An integer that shows the average number of hours the employee works monthly. High or low hours may correlate with job burnout or satisfaction.

- time_spend_company: An integer indicating the employee's tenure at the company. A longer tenure may correlate with a lower likelihood of turnover, but it depends on various other factors.

- Binary Variables:

- work_accident: A binary variable (1 or 0) indicating whether the employee had a work-related accident. This could affect employee satisfaction and turnover.

- left: The target variable, where 1 indicates the employee left the company, and 0 indicates they stayed.

- promotion_last_5years: A binary variable indicating whether the employee received a promotion in the last 5 years. This might indicate job satisfaction or loyalty, both of which are linked to turnover.

- Categorical Variables:

 - department: A string representing the employee's department. This will likely be one-hot encoded to avoid assuming any ordinal relationship.

 - salary: An ordinal categorical feature (low, medium, high) representing the employee's salary level. The treatment of this feature depends on the modeling approach:

  - For Tree-Based Models (Random Forest, XGBoost): The salary feature can be treated as a nominal feature by encoding it as low = 0, medium = 1, high = 2. This allows the model to make splits based on numerical thresholds.

  - For Logistic Regression: The salary feature can be treated as ordinal (low = 0, medium = 1, high = 2), capturing a potential linear relationship with turnover.

Most Relevant Variables

- satisfaction_level and last_evaluation are critical continuous features likely to have a strong influence on employee turnover, as satisfaction and performance are often linked to whether employees stay or leave.

- number_project, average_monthly_hours, and time_spend_company are also important, as they provide insights into the employee's engagement and loyalty.

- work_accident, promotion_last_5years, and salary can provide additional context about factors influencing retention or departure.

- What units are your variables in?

Based on the dataset schema, here are the units or nature of each variable:

- satisfaction_level: A float between 0 and 1. This is a normalized score representing the employee's level of job satisfaction. No specific unit.

- last_evaluation: A float between 0 and 1. This is a normalized performance evaluation score. No specific unit.

- number_project: An integer representing the number of projects the employee was involved in. Unit: count of projects.

- average_montly_hours: An integer representing the average number of hours worked per month. Unit: hours.

- time_spend_company: An integer indicating the total number of years the employee has worked at the company. Unit: years.

- Work_accident: A binary variable (0 or 1). Indicates whether the employee had a work-related accident. Unit: binary indicator.

- left: A binary variable (0 or 1). The target variable representing whether the employee left the company. Unit: binary indicator.

- promotion_last_5years: A binary variable (0 or 1). Indicates whether the employee received a promotion in the last 5 years. Unit: binary indicator.

- Department: A categorical variable representing the department the employee works in. Unit: text label (e.g., sales, technical, support).

- salary: A categorical variable representing the employee's salary level. Unit: text label (low, medium, high), which may be ordinal depending on the modeling context.

- What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

- I initially presumed that features like number_project, average_monthly_hours, promotion_last_5years, and salary would have the strongest influence on the target variable `left`.

- I expected that these variables would directly reflect employee workload, engagement, recognition, and compensation — all of which are key drivers of attrition.

- I also anticipated that the remaining variables (such as satisfaction_level, last_evaluation, and time_spend_company) might exhibit interaction effects, act as covariates, or require transformations to be more effectively utilized in modeling.

- These assumptions guided my approach to exploratory data analysis, where I aimed to validate which features truly correlate with employee turnover and whether any relationships or patterns are nonlinear or conditional on other variables.

- Is there any missing or incomplete data?

There are no missing values in the data.

- Are all pieces of this dataset in the same format?

No, the dataset includes a mix of data types and formats. Specifically, it contains:

- Continuous variables (`satisfaction_level`, `last_evaluation`) stored as float64.

- Count-based integer variables (`number_project`, `average_montly_hours`, `time_spend_company`) stored as int64.

- Binary variables (`Work_accident`, `left`, `promotion_last_5years`) also stored as int64, but semantically representing True/False conditions.

- Categorical string variables (`Department`, `salary`) stored as object, which will require encoding for machine learning models.

While all columns are non-null and consistently typed within their respective formats, they are not all in the same format. Different preprocessing steps will be needed depending on the variable type — such as normalization for continuous features and encoding for categorical ones.

- Which EDA practices will be required to begin this project?

I will:

- Begin with descriptive statistics to understand central tendencies, spread, and identify potential anomalies in each feature.

- Check the structure of the dataset by reviewing data types, null values, and overall completeness.

- Use visualizations such as histograms, boxplots, and count plots to examine distributions and detect outliers or skewness in continuous and count variables.

- Explore relationships between features and the target variable (`left`) using group-wise statistics and visualizations like bar charts.

- Evaluate correlations between numerical variables to identify multicollinearity or strong linear relationships.

- Examine the balance of classes in the target variable to assess potential class imbalance issues.

- For categorical features, analyze the distribution of categories and their association with the target variable using stacked bar plots.

The Power of Statistics

- What is the main purpose of this project?

The main purpose of this project is to understand why employees leave the company based on available employee data. I aim to address the issue of employee turnover, whether through voluntary resignations or involuntary terminations, both of which represent significant challenges. Turnover results in financial losses due to investments in training, onboarding, and integrating employees, as well

as the additional costs associated with interviewing, hiring, and training replacements. To reduce these operational and financial burdens, the project seeks to build a predictive model—such as logistic regression, random forest, or XGBoost—to identify employees who are more likely to leave. This predictive insight can help the company take proactive steps to improve retention strategies and minimize turnover-related costs.

- What is your research question for this project?

The research question for this project is: What factors are driving employee turnover, and how can predictive modeling help identify employees at risk of leaving the company? The leadership team has requested an analysis of employee survey data to uncover key drivers of attrition and generate actionable insights for improving retention. In addition to exploratory data analysis, the project involves selecting appropriate modeling approaches and evaluating them to identify a champion model that best predicts employee turnover. The ultimate goal is to support the leadership team in making data-informed decisions that enhance employee retention.

- What is the importance of random sampling? In this case, what is an example of sampling bias that might occur if you didn't use random sampling?

Random sampling is important because it ensures that the sample reflects the diversity of the entire employee population. It reduces the chances of overrepresenting any specific group and helps in drawing generalizable conclusions from the data. In this case, if random sampling isn't used, a potential sampling bias could be selection bias—for instance, if the survey only includes employees from a specific department, tenure group, or performance level. This would skew the results and may lead to incorrect assumptions about the factors influencing employee turnover across the company.

Regression Analysis: Simplify Complex Data Relationships

- Who are the key stakeholders involved in this project?

- HR Department: I will work closely with the HR department for domain expertise and data access. Additionally, I will collaborate with HR to share insights related to the model development, analysis, and potential action steps for improving employee retention. They will also provide valuable input on the interpretation and application of the model's findings.

- What are you trying to solve or accomplish?

The goal is to address the issue of employee attrition by building a predictive model that can identify employees who are likely to leave the company. By using models such as logistic regression, decision

trees, or ensemble-based models like random forest, the aim is to uncover key patterns and drivers behind employee turnover. This enables the company to take targeted, proactive actions to improve employee retention and reduce the financial and operational costs associated with high turnover.

- What are your initial observations when you explore the data?

My initial observations during data exploration included several key issues and insights:

- There were inconsistencies and mistakes in some column names, which I corrected for better clarity and alignment with common naming practices.

- I discovered 3,008 duplicate records in the dataset. After thorough inspection, I confirmed they were legitimate duplicates and removed them to avoid skewing the analysis.

- Outliers were detected in the `time_spend_company` (tenure) variable when I visualized the distribution using box plots.

- I further analyzed the `time_spend_company` distribution with a histogram to assess skewness, which helped inform my outlier handling strategy.

These early observations laid the foundation for cleaner and more reliable data analysis in the next phases of the project.

- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)

- I primarily use Pandas-https://pandas.pydata.org/docs/ for data importing, cleaning, manipulation, and conducting exploratory data analysis (EDA). It provides flexible tools for understanding the structure and quality of the dataset.

- Additionally, I use GitHub-https://github.com/ to track my progress and version control. Committing changes to my main project repository ensures that all updates are well-documented and can be revisited or shared easily as the project evolves.

- Do you have any ethical considerations in this stage?

Yes, my ethical considerations during the initial EDA stage include ensuring that the dataset is representative, complete, and comprehensive enough to support a valid and fair analysis. This helps prevent biases that could arise from incomplete or skewed data and ensures that insights drawn are reflective of the broader employee population.

By thoroughly checking for data quality issues—such as missing values, duplicates, and inconsistencies—I aim to maintain integrity throughout the analysis process. These early checks play a critical role in minimizing potential bias and ensuring that any models built later are based on accurate and ethical foundations.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?

I am trying to solve the problem of employee attrition. If I can predict which employees are likely to quit, it becomes possible to identify the key factors contributing to their decision to leave. Since finding, interviewing, and hiring new employees is both time-consuming and costly, improving employee retention can bring significant benefits to the company. In response to this challenge, the HR department at Salifort Motors is looking to implement initiatives aimed at enhancing employee satisfaction and reducing turnover.

- What resources do you find yourself using as you complete this stage?

I primarily use Pandas for data importing, cleaning, manipulation, and conducting exploratory data analysis (EDA) as it provides flexible tools for understanding the structure and quality of the dataset. Additionally, I use GitHub to track my progress and maintain version control. Committing changes to my main project repository ensures that all updates are well-documented and can be revisited or shared easily as the project evolves.

- Is my data reliable?

I observe that the dataset is quite comprehensive, capturing a range of employee-related factors relevant to turnover prediction. It appears to be reliable as it contains no missing values and covers a good mix of numeric, binary, and categorical variables. The structure is consistent, and after removing duplicates and correcting field names, the data seems clean and ready for analysis. However, reliability also depends on how the data was collected—if it accurately represents the entire employee population without bias, then I can be more confident in the reliability of the insights generated.

- Do you have any additional ethical considerations in this stage?

During the initial EDA stage, I focus on ensuring the dataset is representative, complete, and free of issues like missing values or duplicates. This helps prevent bias, supports fair analysis, and lays a reliable foundation for building ethical predictive models.

- What data do I need/would I like to see in a perfect world to answer this question?

The dataset already includes many fields that capture important information to help detect attrition and analyze employee satisfaction. However, in an ideal scenario, I would like to see additional demographic information about employees, such as age, gender, education level, and marital status. These features can offer deeper insights during analysis, but they must be handled with care. It is

essential to scrutinize such variables before model training to ensure that no biases or unintended patterns are learned by the model, and remove them if necessary.

Since the main question centers on understanding employee satisfaction and what factors drive employees to leave, additional data such as exit interview scores, participation in social or recreational activities, and workplace behavior or conduct would also be valuable. This kind of information could provide a more holistic view of the employee experience and enhance the predictive power of the model.

- What data do I have/can I get?

I have the HR dataset from the company, which is a random sample drawn from the company's employee database. This is the primary data I can work with. If available, additional data such as demographic information (which should be used cautiously in modeling to avoid introducing bias), employee conduct records, workplace experience feedback, and participation in fun or engagement activities could be consolidated with the existing HR dataset. Incorporating such data would provide a more comprehensive view of employee behavior and factors influencing turnover.

- What metric should I use to evaluate success of my business objective? Why?

I planned to use recall as the primary metric to evaluate the success of the business objective. Since the main goal is to detect employees who are likely to leave, it is critical to minimize false negatives—cases where a leaver is incorrectly predicted as a stayer. Misclassifying a leaver can lead to costly consequences for the company, so prioritizing recall ensures that as many actual leavers as possible are correctly identified, even if it results in a few more false positives.

Additionally, I used AUC-ROC to compare and select the best model during the tuning phase. AUC-ROC provides a single score that reflects the model's ability to distinguish between classes across all thresholds, offering a more holistic view during hyperparameter tuning. It is not tied to a specific decision boundary (like recall at 0.5), which makes it ideal for model selection.

The champion model achieved an AUC-ROC of 0.9648 during cross-validation and 0.9384 on the test set. This indicates strong and consistent ranking performance in distinguishing leavers from stayers, meaning the model has approximately a 94% chance of ranking a randomly selected leaver above a stayer, even on unseen data.

Data Project Questions & Considerations

PACE: Analyze Stage

Get Started with Python

- Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes, I believe the available information is sufficient to achieve the project's goal. Although some variables did not satisfy the linearity assumption required for logistic regression, the dataset remains well-suited for alternative modeling techniques. In particular, Random Forest Round 2 was selected as the final model due to its superior performance in identifying employees likely to leave. While this choice limits the interpretability compared to a logistic regression model, the primary focus of the project is on accurate prediction.

The dataset is clean, complete, and includes a strong mix of employee-related features—such as satisfaction level, tenure, number of projects, promotions, and salary tiers—that contribute meaningfully to modeling attrition. Based on both domain intuition and exploratory data analysis, these variables are sufficient for capturing the key patterns needed to predict employee turnover effectively.

Go Beyond the Numbers: Translate Data into Insights

- What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

The EDA should stay focused on the research question at all times, as this will help uncover key insights about how the various fields influence the outcome variable (employee attrition). This will also provide an initial understanding of the modeling approach and offer a slight idea of how the model may perform.

The steps to take are primarily in the preprocessing and processing stages of the initial plan. These steps are crucial for structuring and preparing the data for the full analysis. The main focus during the "analyze" phase should be on:

- Using visualizations to explore relationships and distributions.

- Performing a comparative study of different visualizations to detect hidden patterns in the data.

However, it's important to remember that some visualizations might be misleading. Therefore, it's essential to carefully assess them in the context of the data to avoid drawing incorrect conclusions.

By following this approach, we can ensure that we gain meaningful insights and make informed decisions about the modeling process.

- Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

No, I don't need to add more data at this point. However, in an ideal scenario, I would like to see additional demographic information about employees, such as age, gender, education level, and marital status. These features could offer deeper insights during analysis. However, they must be handled carefully. It is essential to scrutinize such variables before model training to ensure that no biases or unintended patterns are learned by the model, and to remove them if necessary.

If I were to join additional data, it would need to be validated to ensure the dataset is structured correctly. The dataset should also undergo iterative validation throughout the analysis workflow to ensure its structure remains sound.

As for the current dataset, it was checked for:

- Missing values: No missing values were found.

- Duplicates: 3,008 duplicates were identified and removed after scrutiny.

- Outliers: 824 outliers were found with a lenient check, and 1,886 with a strict one. The lenient outlier check was chosen to retain as much data as possible.

- Field names: Some field names were corrected, and others were renamed for clarity.

With these steps, the dataset is now properly structured for analysis.

- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?

The best visualizations suited for the intended audience would include:

- Barplots: Useful for comparing categorical data such as employee departments or salary levels against attrition rates.

- Histograms: Helpful in understanding the distribution of numerical variables like satisfaction level, last evaluation, and average monthly hours.

- Scatterplots: Effective for exploring relationships between continuous variables, such as satisfaction level vs. average monthly hours.

- Grouped histograms: Ideal for comparing categorical variables across different groups, such as department and attrition.

- Stacked column charts: Useful for comparing the proportions of different categories within the data, like salary level and employee attrition status.

These visualizations can highlight key patterns and relationships relevant to employee attrition and are effective for communicating insights to both technical and non-technical audiences.

The Power of Statistics

- Why are descriptive statistics useful?

Descriptive statistics provide a foundational understanding of the dataset by summarizing its key characteristics. Using functions like `describe()` or `describe(include='all')` allows us to quickly observe measures such as central tendency (mean, median), spread (standard deviation, range), and shape (skewness). These insights help identify patterns, spot potential anomalies or outliers, and guide decisions about how to clean and explore the data further. Overall, descriptive statistics offer an essential first look at the structure and quality of the data before deeper analysis begins.

- What is the difference between the null hypothesis and the alternative hypothesis?

The null hypothesis (H_0) assumes that there is no effect, no difference, or no relationship between variables. It represents the default or status quo condition.

The alternative hypothesis (H_1 or Ha) proposes that there is an effect, a difference, or a relationship. It challenges the null hypothesis and suggests that a change has occurred.

In essence:

- H_0: Nothing has changed; any observed effect is due to random chance.

- H_1: Something has changed; the observed effect is statistically significant and not due to chance.

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model?

Exploratory Data Analysis (EDA) is a crucial step before building a multiple linear regression model to ensure the dataset is clean, suitable, and insightful for modeling. Key purposes include:

- Checking Assumptions:

EDA helps evaluate whether important assumptions for linear regression are met, such as linearity between independent and dependent variables, normality of residuals, homoscedasticity (equal variance), and absence of multicollinearity.

- Understanding Data Structure:

It provides a clear picture of the distributions, central tendencies, and spread of the data, as well as relationships between variables through summary statistics and visualizations.

- Identifying Data Issues:

EDA helps detect and handle missing values, outliers, and data entry errors that could otherwise distort model results.

- Feature Selection & Engineering:

EDA reveals which variables may be strong predictors and assists in transforming, encoding, or combining variables to create more effective features for the regression model.

- Guiding Model Strategy:

It informs whether certain variables should be transformed (e.g., log-transformed), dropped, or included based on observed patterns or correlations.

By performing EDA, I ensure the data is ready for modeling and can make informed decisions that contribute to a more accurate and reliable linear regression model.

- Do you have any ethical considerations in this stage?

Yes, ethical considerations were carefully taken into account during the data preparation process. Duplicate records were thoroughly assessed before removal to ensure that legitimate and unique employee experiences were not mistakenly discarded. Preserving diverse perspectives and real-life variations in employee journeys is important to maintain the integrity of the analysis.

Additionally, a comprehensive analysis of outliers was conducted before deciding on any removal. Outliers were not immediately excluded; instead, their presence was evaluated based on the distribution characteristics and the potential impact on different types of models. This ensures that any

decision to remove data is made thoughtfully and grounded in real-world variability, rather than arbitrary filtering.

Moreover, any future inclusion of sensitive demographic variables (such as gender, race, or marital status) would require careful scrutiny to avoid reinforcing bias or making unfair predictions. The goal is to ensure fairness and transparency throughout the modeling process while respecting employee privacy and maintaining the ethical use of data.

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?

I am trying to solve the employee attrition problem by predicting it, which will help mitigate the need to hire, train, and onboard new employees—activities that are costly and time-consuming.

In the Analyze stage of the PACE workflow, the primary goal is to perform advanced EDA using visualizations and comparative analytical techniques to uncover patterns, understand how different features influence the "left" field, and gain insights into employee satisfaction.

The plan still works effectively. By following the PACE workflow, I am able to iterate through stages as needed, ensuring flexibility in approach. The plan does not need revising at this point because it is grounded in a well-structured and logically sound framework developed at the outset. The adaptability of the PACE workflow supports ongoing refinement and course correction if necessary, while still keeping the overall goal on track.

- Does the data break the assumptions of the model? Is that ok, or unacceptable?

Yes, the data does break one of the key assumptions of the logistic regression model — specifically, the linearity of the logit. This assumption requires that the continuous predictor variables have a linear relationship with the log-odds of the outcome variable. In this case, variables like `satisfaction_level`, `tenure`, `average_monthly_hours`, `number_project`, and `last_evaluation` do not satisfy this requirement based on diagnostic checks.

While this violation can affect the interpretability of coefficients and potentially lead to misestimated probabilities in certain parts of the data range, the overall classification performance of the model may still be acceptable. Since the primary objective of this project is to predict employee attrition — not to make statistical inferences about individual features — the impact of this violation is considered manageable.

To address this, I could explore transformations such as polynomial terms, splines, or binning of continuous variables. These approaches can help capture non-linear relationships more effectively. However, because the non-linearity observed was moderate and the logistic regression model was more of an exploratory baseline, I chose not to implement these adjustments at this stage.

Ultimately, the plan has already leaned toward more flexible, non-parametric models like decision trees and random forests, which are not affected by the linearity of logit assumption. These models were better suited for the complexity of the attrition prediction task, and the logistic regression model served its purpose in providing initial insights into the data.

- Why did you select the X variables you did?

For the logistic regression model, I initially selected all available fields after excluding the outliers. This comprehensive inclusion allowed me to assess the relationships between predictors and the target variable while maintaining the model's interpretability. The variables were checked for multicollinearity to ensure they met the assumptions required for logistic regression.

For the Round 1 decision tree and random forest models, I followed a similar approach by including all fields. These models are less sensitive to multicollinearity, so the focus was more on identifying useful splits and assessing variable importance as a baseline.

However, after identifying a potential data leakage issue, I refined the input features in Round 2. Specifically, I dropped `satisfaction_level` (which was highly correlated with the outcome) and extracted a new feature, `overworked`, from `average_monthly_hours`. This change helped prevent leakage while preserving informative signals in the data.

Overall, the features were chosen to balance model performance with interpretability and to adhere to modeling assumptions where applicable. Feature refinement in later stages improved both robustness and ethical soundness.

- What are some purposes of EDA before constructing a model?

As part of the Analyze stage in the PACE workflow, Exploratory Data Analysis (EDA) plays a critical role in preparing the dataset for modeling. It ensures that the data is clean, well-understood, and suitable for the chosen analytical approach. Key purposes of EDA at this stage include:

- Assessing Modeling Assumptions:

I use EDA to evaluate whether the data meets key assumptions relevant to the planned models — such as absence of strong multicollinearity, appropriate distributions for numerical variables, and independence of observations.

- Understanding Data Characteristics:

EDA helps me explore the structure of the dataset by generating summary statistics and visualizations. This allows me to interpret variable ranges, spot dominant patterns, and detect potential biases or skews in the data.

- Diagnosing Data Quality Issues:

At this stage, I identify and address issues such as missing values, outliers, duplicate records, and data entry errors. These issues, if left unaddressed, could significantly impact the model's validity.

- Guiding Feature Selection and Engineering:

Through correlation analysis and data visualization, EDA helps identify which features are most relevant and predictive. This insight informs decisions about which variables to include, transform, or engineer to better capture underlying relationships.

- Identifying Class Imbalance:

For classification problems like employee turnover, EDA helps evaluate the distribution of the target variable. Detecting class imbalance early allows me to plan appropriate strategies such as resampling or adjusting evaluation metrics.

By conducting a thorough EDA during the Analyze stage, I can ensure that the data is ready for the next steps — leading to more robust model construction and clearer interpretation of results.

- What has the EDA told you?

The EDA provided valuable insights that guided both data preprocessing and the modeling approach.

In the initial phase of EDA, I focused on data quality checks: identifying and handling missing values, duplicates, and outliers. Visualizations were used to examine distributions and detect skewness. Descriptive statistics (describe(include='all') and info()) helped assess the structure and completeness of the dataset. Outlier mitigation strategies were later implemented in the modeling stage.

In the advanced EDA phase, comparative visual analysis revealed several key patterns:

- There is a class imbalance in the target variable (left), with about 83% stayers and 17% leavers.

- Attrition is observed at both extremes of workload — underworked and overworked employees.

- Employees with 7 projects working 255–295 hours/month fell within the interquartile range of monthly hours.

- 3–4 projects were associated with the highest employee retention.

- A scatterplot of satisfaction vs. monthly hours revealed three clusters:

  - Low satisfaction, low hours (but still left)

  - High hours with mixed satisfaction (both high and low) — indicating burnout and disengagement

- Satisfaction vs. tenure boxplots showed two distinct groups of leavers:

  - Short tenure and low satisfaction

  - Medium tenure and high satisfaction

- Tenure histogram revealed a spike in attrition around the 3-year mark.

- In a tenure vs. salary level histogram, shorter-tenured employees were more likely to have low or medium salaries; in longer tenures, medium salary remained dominant.

- A monthly hours vs. evaluation scatterplot showed two types of leavers:

  - Slightly under baseline hours (~166.67) with lower evaluation scores

  - Overworked employees with high evaluations, pointing toward burnout

- Monthly hours vs. promotions revealed that high-hour workers were often not promoted, suggesting a lack of recognition.

- A stacked column chart by department showed HR had the highest proportion of leavers, while management had the lowest.

- Comparative analysis of two scatterplots revealed that performance alone doesn't explain attrition. Instead, a combination of low satisfaction and high workload indicated burnout.

Key takeaway:

- Satisfaction level is a leading indicator of disengagement and attrition.

- Workload without rewards contributes to hidden turnover risks.

- Monitoring both satisfaction and workload, and aligning them with performance and promotions, is essential for improving retention.

Lastly, a correlation heatmap confirmed:

- Positive correlation among number_project, monthly_hours, and last_evaluation

- Negative association between satisfaction_level and attrition (left)

These findings helped inform feature engineering, model selection, and areas of focus for intervention.

- What resources do you find yourself using as you complete this stage?

During this stage, I relied on the following resources:

- Pandas: For data import/export, cleaning, transformation, and exploratory data analysis (EDA). It's my primary tool for data manipulation in Python.

- Seaborn and Matplotlib: For data visualization, including distribution plots, box plots, and heatmaps to support EDA and pattern discovery.

- Matplotlib: Used for plotting visualizations, including decision trees, ROC curves, and confusion matrices to support model evaluation and interpretation.

- GitHub: To track version history, document changes, and commit progress to my main project repository, helping maintain a structured and reproducible workflow.

These tools have been essential in ensuring a robust and transparent data preparation process.

- Do you have any ethical considerations in this stage?

Yes, ethical considerations were carefully taken into account during the data preparation process. Duplicate records were thoroughly assessed before removal to ensure that legitimate and unique employee experiences were not mistakenly discarded. Preserving diverse perspectives and real-life variations in employee journeys is important to maintain the integrity of the analysis.

Additionally, a comprehensive analysis of outliers was conducted before deciding on any removal. Outliers were not immediately excluded; instead, their presence was evaluated based on the distribution characteristics and the potential impact on different types of models. This ensures that any decision to remove data is made thoughtfully and grounded in real-world variability, rather than arbitrary filtering.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?

The average monthly hours mean or average is 201.05 (~201 hours), which appears unusual. If I assume a standard 9–5 job with 8 working hours per day, then dividing 201 by 8 gives approximately 25 days of work per month. However, the typical number of working days in a month is around 20–23. This discrepancy suggests employees are working more than expected.

This elevated workload may be linked to attrition. A more grounded baseline—assuming a 40-hour work week and two weeks of vacation per year—puts the average monthly working hours for a full-time employee at approximately 166.67 hours (or roughly 166 hours). In comparison, the actual average of 201.05 hours is significantly higher.

As earlier analysis showed, apart from employees working on only two projects, every other group—regardless of whether they left or stayed—tended to work well above the theoretical full-time schedule. This strongly indicates that overwork could be a contributing factor to employee turnover.

- How many vendors, organizations or groupings are included in this total data?

The dataset includes several categorical groupings and variables representing different vendors, organizations, or groupings:

- Departments: There are 10 unique departments: sales, accounting, hr, technical, support, management, IT, product_mng, marketing, and RandD.

- Salary Levels: Salary is categorized into three groups — low, medium, and high.

- Number of Projects: This ranges from 2 to 7 projects.

- Time Spent in Company (Tenure): Ranges from 2 to 10 years.

- Work Accident: A binary variable indicating whether the employee experienced a work accident (0 = No, 1 = Yes).

- Attrition (Left): The outcome variable showing whether the employee left the company (0 = Stayed, 1 = Left).

These groupings provide a diverse set of features for understanding employee behavior, performance, and attrition patterns.

Go Beyond the Numbers: Translate Data into Insights

- What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

To complete the project goals, the following data visualizations, machine learning algorithms, and data outputs need to be built:

Data Visualizations:

- Boxplots: Boxplots were used to visualize the distribution of data, assess skewness, and identify outliers. In some cases, a single variable was plotted, while in other cases, both x and y axes were used to explore relationships between multiple variables, with additional segmentation based on categorical features.

- Barplots: Useful for comparing categorical data such as employee departments or salary levels against attrition rates.

- Histograms: Helpful in understanding the distribution of numerical variables like satisfaction level, last evaluation, and average monthly hours.

- Scatterplots: Effective for exploring relationships between continuous variables, such as satisfaction level vs. average monthly hours.

- Grouped histograms: Ideal for comparing categorical variables across different groups, such as department and attrition.

- Stacked column charts: Useful for comparing the proportions of different categories within the data, like salary level and employee attrition status.

- Correlation heatmaps: To check for multicollinearity and also to get an idea of which fields affect the outcome variable, how strongly, and in which direction.

- Grouped column chart (Department vs Employee Count): This chart helps visualize the distribution of employees who stayed (0) versus those who left (1) across different departments.

- Line plots: To assess the linearity of the logit assumption in logistic regression.

- Confusion matrices: For evaluating classification models.

Model Evaluation Visuals:

- AUC-PR curve: To assess precision-recall tradeoff, especially given class imbalance.

- Decision tree splits: To visualize how decisions are made in tree-based models.

- Feature importance charts: For decision tree and random forest models to understand which variables contribute most to predictions.

Machine Learning Algorithms:

- Logistic Regression: As a baseline model and for interpretability.

- Decision Trees: With and without data leakage removal to understand overfitting risks.

- Random Forest: An ensemble model to improve performance and robustness, also tested with and without leakage removal.

Model Outputs:

- Cross-validation (CV) metric results: For model selection and validation.

- Test set metric results: For evaluating model generalizability.

These components together ensure robust data exploration, model building, evaluation, and interpretation aligned with the project's objectives.

- What processes need to be performed in order to build the necessary data visualizations?

To build the necessary data visualizations, the following processes need to be carried out:

- Import Visualization Libraries

Use seaborn and matplotlib—ensure the latest versions are installed for improved functionality and aesthetics.

- Understand the Data

Perform structural data checks using .info() and .describe() to assess completeness, data types, and basic statistical properties.

- Tailor Visuals to Audience

Design visualizations based on who will view them. For technical audiences, more detailed or complex visuals like correlation maps and model evaluation metrics are appropriate. For non-technical stakeholders, simpler visuals such as bar plots or stacked columns are more digestible.

- Maintain Aesthetic and Accessibility Standards

Use consistent color themes and avoid problematic color combinations for viewers with color vision deficiencies. Choose clear labels and readable fonts.

- Understand Variables Before Visualizing

It's important to understand what each field represents to avoid misleading visuals. For instance, boxplots might mislead if the sample size is small or uneven across groups.

- Key Visualizations to Include

- Boxplots: Useful for analyzing distribution, skewness, and spotting outliers in fields like satisfaction level and evaluation scores.

- Histograms: Show distribution of numerical variables such as satisfaction, evaluation, and monthly hours.

- Grouped histograms: Used to compare attrition rates across salary levels, departments, or tenure groups.

- Scatterplots: Show relationships between continuous variables, for example, satisfaction level vs. average monthly hours.

- Stacked Column Charts: Help visualize the proportion of employees who stayed or left within departments or job roles.

- Correlation Heatmaps: Identify relationships between variables and check for multicollinearity.

- Grouped column chart (Department vs Employee Count): This chart helps visualize the distribution of employees who stayed (0) versus those who left (1) across different departments.

- Line Plots: Inspect linearity of the logit for logistic regression models.

- Confusion Matrices: Show performance of classification models with true positives, true negatives, false positives, and false negatives.

- Decision Tree Diagrams: Visualize model splits for interpretability.

- Bar Charts for Feature Importance: Rank features based on importance from decision trees and random forests.

By carefully planning and customizing visualizations in this way, the analysis remains both effective and accessible to diverse audiences while supporting model development and decision-making.

- Which variables are most applicable for the visualizations in this data project?

The most applicable variables for visualizations in this data project include both numerical and categorical features, depending on the type of visualization and the use of hue.

Numerical Variables:

These are primarily used in boxplots, line plots, scatterplots, histograms, and correlation heatmaps:

- satisfaction_level

- last_evaluation

- number_project

- average_monthly_hours

- tenure

Categorical Variables:

These are especially relevant when hue is applied or when grouped visualizations are created:

- department

- left (used as hue for comparisons across attrition status)

- promotion_last_5years

- salary

Derived/Model-based Variables:

Used in model evaluation and feature importance plots:

- Logit (log-odds of `left`) — for assessing linearity in logistic regression

- Feature importance scores — from decision tree and random forest models

- Classification outcome — used in confusion matrices (True Positive, False Negative, etc.)

Hue-specific Variables:

Hue is used to enhance interpretation in many plots, and common hue variables include:

- left

- salary

- promotion_last_5years

- Classification outcome labels (e.g., TP, FP)

In summary, the visualizations in this project make use of a wide range of variables — with numerical features forming the foundation for distribution and relationship plots, and categorical variables driving grouped, hue-based, and class-based comparisons.

- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

In the current dataset, there are no missing values, so no imputation or removal was necessary. However, if missing data had been present, I would have addressed it through the following approaches based on the extent and pattern of the missingness:

- If the missing data is minimal, I would consider dropping the affected rows using:

  - `df.dropna(inplace=True)` or assigning the result back to the original DataFrame after `df.dropna()`.

- If missing values are more widespread or occur in important columns, I would use imputation methods:

  - Forward fill (`ffill`): Fills missing values by carrying forward the last known non-null value.

  - Backfill (`bfill`): Fills missing values using the next available non-null value.

  - Mean imputation: Replaces missing entries with the mean of the column.

  - Median imputation: Uses the column's median to fill in the missing values.

Missing values like geolocation data could be identified using `isna()` or similar checks. I could also create visualizations to understand which locations or categories are associated with higher missingness. This would help pinpoint where follow-up with the data provider might be needed.

Each of these strategies would be carefully selected based on the data distribution and the role of the affected column in analysis or modeling. The goal is to preserve data integrity while ensuring analytical reliability.

The Power of Statistics

- Is hypothesis testing necessary for this analysis? Additionally, what is the difference between the null hypothesis and the alternative hypothesis?

There was no dedicated hypothesis testing required in this project, as the project's goal and objective did not necessitate it.

Difference between Null Hypothesis and Alternative Hypothesis:

- Null Hypothesis (H_0): Assumes that there is no effect, no difference, or no relationship between variables. It represents the default or status quo condition.

- Alternative Hypothesis (H_1 or Ha): Proposes that there is an effect, a difference, or a relationship. It challenges the null hypothesis and suggests that a change has occurred.

In essence:

- H_0: Nothing has changed; any observed effect is due to random chance.

- H_1: Something has changed; the observed effect is statistically significant and not due to chance.

Regression Analysis: Simplify Complex Data Relationships

- Do you notice anything odd?

Yes, there are a few unexpected patterns in the data:

- In the correlation heatmap, some intuitive relationships didn't hold up as strongly as expected. For example, tenure and salary might naturally be assumed to correlate strongly, but the data shows only a weak relationship between them.

- In the salary-by-tenure plot, there's a surprising spike in employees with exactly 3 years of tenure, which could indicate either a hiring surge or some data irregularity.

- Can you improve it? Is there anything you would change about the model?

Yes, I believe the model can be improved further despite its strong performance. Here are some key areas for potential enhancement:

- Linearity and Feature Transformation: Several continuous variables did not meet the linearity assumption, which could affect coefficient interpretability. While this doesn't impact the model's classification performance, applying transformations (e.g., quadratic terms, splines) could improve robustness, especially if the focus shifts to inference.

- Increase Recall: Improving recall remains a priority to better capture leavers. This is critical for early identification, enabling more effective retention strategies.

- Feature Engineering & Data Leakage: Further feature engineering, such as aggregating or interacting features (e.g., job role, tenure, satisfaction), could uncover stronger signals. Additionally, while data leakage was previously checked, it's worth revisiting key features (e.g., last_evaluation) to assess their true impact on model performance.

- Model Tuning: Refining hyperparameters, adjusting cross-validation settings, experimenting with resampling techniques (e.g., SMOTE), and using class weights or threshold tuning could help optimize model performance, especially in handling class imbalance.

While this model is likely near-optimal given the extensive preprocessing already conducted, these steps could offer further improvements in predictive power and insight into employee attrition.

The Nuts and Bolts of Machine Learning

- Is there a problem? Can it be fixed? If so, how?

Yes, there was a problem of data leakage, which I observed during the modeling process. I identified it when certain features like `satisfaction_level` and `last_evaluation` appeared suspiciously optimistic, suggesting that the model might be learning from information not realistically available at prediction time.

Data leakage occurs when information that should not be available during training — either because it belongs to the test set or reflects future data — is inadvertently included in the model. This can lead to overly optimistic performance metrics that do not generalize well to unseen data and would underperform in real-world deployment.

In this case, it's unlikely that a company would consistently have up-to-date satisfaction scores or be able to use average monthly hours without bias. For example, if an employee is already planning to leave, or if they've been flagged for termination, their working hours may already be reduced — effectively revealing the outcome we're trying to predict.

In the first round of modeling, I included all available features in the logistic regression, decision tree, and random forest models. In the second round, I addressed the data leakage by performing feature engineering:

- I dropped the `satisfaction_level` feature.

- I carried out feature extraction to create a new binary variable called `overworked`, derived from `average_monthly_hours`, which indicates whether an employee is working more than a typical amount.

This change helped reduce the risk of leakage while still preserving predictive value, allowing the model to generalize better to unseen scenarios.

- Which independent variables did you choose for the model, and why?

For the multiple binomial logistic regression model, I included all available features after appropriate encoding:

- I applied one-hot encoding to the `department` variable since it is nominal and has no inherent order.

- For the `salary` variable, I used ordinal encoding to reflect its natural rank (e.g., low < medium < high).

- All other variables were retained in their existing form and included as independent variables.

For the decision tree and random forest models (Round 1), I also used all features without exclusions.

However, in Round 2 of the decision tree and random forest models, I made a few adjustments:

- I dropped the `satisfaction_level` variable to explore how the model performs without it and to reduce the risk of overfitting.

- I engineered a new binary feature called `overworked`, derived from `average_monthly_hours`, to address potential data leakage. This new variable helped flag employees working significantly above average hours.

These adjusted features, alongside the rest of the dataset, were used as independent variables in the second round of modeling to test for improvements in interpretability and generalization.

- How well do your models fit the data? (What are the validation scores for each model?)

The models were evaluated based on ROC AUC scores for both cross-validation (CV) and test sets to assess how well they fit the data.

Decision Tree - Round 1 (hr_dt1)

- Cross-Validation (CV) Results:

ROC AUC: 0.9698

- Test Results:

ROC AUC: 0.9506

- Analysis:

The Decision Tree shows a strong fit, with a minor drop in test performance likely due to overfitting. Overall, it fits the data well.

Random Forest - Round 1 (hr_rf1)

- Cross-Validation (CV) Results:

ROC AUC: 0.9804

- Test Results:

ROC AUC: 0.9564

- Analysis:

The Random Forest performs excellently, with minimal drop in test results. It fits the data well and handles class imbalance effectively.

Decision Tree - Round 2 (hr_dt2)

- Cross-Validation (CV) Results:

ROC AUC: 0.9587

- Test Results:

ROC AUC: 0.9336

- Analysis:

The second Decision Tree shows some overfitting, with a more noticeable drop in test performance. It indicates a less optimal fit compared to the first round.

Random Forest - Round 2 (hr_rf2) – Champion Model

- Cross-Validation (CV) Results:

ROC AUC: 0.9648

- Test Results:

ROC AUC: 0.9384

- Analysis:

The Round 2 Random Forest provides the best performance with strong consistency between training and test sets. This makes it the best fit for the data and the chosen champion model.

In conclusion, the Round 2 Random Forest model performed consistently well on both the training and test sets, showing strong generalization, and was selected as the champion model for further use.

- Can you improve it? Is there anything you would change about the model?

Yes, I believe the model can still be improved in several ways, even though it performs well overall. Here are some areas that could lead to better performance:

- Increase Recall: Since recall is a critical metric for identifying leavers, improving it would help capture more at-risk employees. This is especially important in a business context, where detecting leavers early can lead to better retention strategies.

- Feature Engineering: There might be additional features or transformations of existing features that could better capture the patterns associated with employee turnover. For example, aggregating or interacting features like job role, tenure, or satisfaction levels might provide stronger signals.

- Investigate Data Leakage: Even though prior checks were done, it's important to continually evaluate potential data leakage. Features like `last_evaluation` and `satisfaction_level` may reflect outcomes rather than the causes of attrition. I could assess model performance with and without these features to determine their actual contribution.

- Reframe the Problem: If certain features, like `last_evaluation`, strongly correlate with attrition, it might make sense to reframe the problem. For instance, I could try predicting satisfaction or evaluation scores themselves, which could act as leading indicators of turnover, and then use those as inputs for attrition prediction.

- Model Tuning:

- Adjust the cross-validation (`cv`) parameters during hyperparameter searches to ensure the model is as robust as possible.

- Experiment with resampling techniques like SMOTE or undersampling to handle class imbalance and improve the model's ability to detect the minority class (leavers).

- Consider using class weights or threshold tuning to optimize the model based on business priorities, focusing on reducing false negatives (i.e., failing to detect leavers).

Ultimately, it's possible that this model represents near-optimal performance with the current dataset, considering the extensive preprocessing and data quality work already done. However, implementing the steps mentioned above could help unlock further improvements and provide deeper insights into the factors contributing to employee attrition.

- Do you have any ethical considerations in this stage?

Yes, at this stage, I will focus on minimizing potential bias to ensure that the model produces fair and equitable predictions. My goal is to develop a baseline model that is technically robust and also aligned with the overall problem-solving objective. I'll pay close attention to selecting and evaluating the most relevant performance metrics for the scenario to minimize the risk of misclassification and unintended consequences. Throughout this phase, fairness, transparency, and reliability will remain core priorities.

Data Project Questions & Considerations

PACE: Execute Stage

Get Started with Python

- Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing an exploratory data analysis?

Given my initial understanding of the dataset, I recommend investigating two key areas before conducting a full exploratory data analysis:

- Organizational Context:

I would ask the manager to provide recent context about the company or their specific team — for example, whether employees have appeared engaged, stressed, or disengaged during recent interactions. While this qualitative insight may not offer a comprehensive or unbiased view of the entire workforce, it can serve as a useful starting point to shape initial perspectives. Although my analysis will be grounded in the data, understanding the manager's observations may help guide more thoughtful and targeted hypotheses early in the process.

- Data Timing and Potential Leakage:

I also recommend confirming when key variables — such as satisfaction_level, last_evaluation, and average_monthly_hours — were last updated. Variables like satisfaction_level and last_evaluation could introduce data leakage if they reflect information collected after an employee had already decided to leave (e.g., after submitting their notice) or had been flagged for termination. Verifying the timing of these fields will help ensure the model is trained only on data that would have been available at the time of prediction.

Together, these investigations will help validate the data's integrity and support a more strategic, informed approach to the exploratory analysis.

- What data initially presents as containing anomalies?

Initial data exploration revealed two key areas of potential anomalies:

- Outliers in tenure:

The tenure variable (time spent at the company) exhibited a right-skewed distribution, with a long tail of employees who had unusually high tenure values. To assess these outliers, both conservative (median + 1.5 × IQR) and standard (Q3 + 1.5 × IQR) methods were applied.

- The conservative method identified 1,886 outliers, reflecting a stricter threshold suitable for skewed data.

- The standard IQR method identified 824 outliers, offering a more lenient cutoff.

Ultimately, the standard method was retained for modeling, as it preserved more data while still flagging extreme cases.

- Duplicate Records:

A total of 3,008 duplicate rows were detected — identical across all columns. After careful inspection, these were determined to be unlikely to represent valid, repeated employee records due to the number of continuous variables involved. Given the low probability of all values matching across multiple continuous fields by chance, these duplicates were removed to maintain data integrity.

These anomalies were addressed to ensure cleaner, more reliable data for the modeling phase and to prevent misleading signals during analysis.

- What additional types of data could strengthen this dataset?

While the current HR dataset provides a strong foundation for analyzing employee turnover and satisfaction, several additional data types could further enhance the analysis:

- Demographic Information:

In an ideal scenario, adding employee demographics — such as age, gender, education level, and marital status — could provide deeper insights into trends across different employee segments. However, these variables should be used cautiously to avoid introducing bias into the model. It's important to assess them critically and remove any that may lead to unintended or unethical patterns during prediction.

- Workplace Experience and Engagement:

Data reflecting the employee experience, such as exit interview feedback, participation in social or engagement activities, or employee conduct records, would offer a more holistic view of job

satisfaction and retention risks. These insights could help explain behavioral drivers behind attrition beyond what quantitative metrics reveal.

- Behavioral or Performance Indicators:

Information on peer or manager evaluations, internal surveys, or training participation might help highlight trends related to performance, growth opportunities, or burnout — all of which can impact turnover.

If such data were available, it would need to be properly validated and joined with the existing dataset. That includes ensuring structural consistency and applying ongoing validation throughout the analysis process.

As for the current dataset:

- No missing values were found.

- 3,008 duplicate records were identified and removed.

- Outlier detection on tenure was performed using both conservative and standard IQR methods. The lenient method was ultimately chosen to retain more data.

- Field names were reviewed and updated for clarity.

With these steps completed, the dataset is well-prepared for analysis — and would be further strengthened by integrating these additional data types in a thoughtful and ethical way.

Go Beyond the Numbers: Translate Data into Insights

- What key insights emerged from your EDA and visualizations(s)?

Key Insights from EDA and Visualizations

- Class Imbalance:

The target variable left exhibits a significant class imbalance, with approximately 83% of employees staying and only 17% leaving. This imbalance could affect the performance of predictive models, especially in terms of precision and recall for the minority class (leavers).

- Workload and Attrition:

Attrition was observed at both extremes of workload: employees who were either underworked or overworked were more likely to leave. This suggests that workload balance plays a critical role in employee retention.

Employees working 7 projects and around 255–295 hours/month tended to fall within the interquartile range for monthly hours, indicating a more optimal balance between workload and retention.

A spike in attrition was observed around the 3-year tenure mark, indicating that mid-tenure employees might be at a higher risk of leaving, possibly due to stagnation or unmet expectations.

- Satisfaction and Attrition:

Satisfaction level is a strong indicator of employee retention. Employees with low satisfaction and low hours worked were most likely to leave, as were those with medium tenure but high satisfaction, indicating a possible mismatch between expectations and career growth.

A scatterplot of satisfaction vs. monthly hours revealed three distinct clusters:

- Low satisfaction, low hours — leading to attrition.

- High hours, mixed satisfaction — indicating burnout and potential disengagement.

- High satisfaction, moderate hours — associated with lower attrition.

- Burnout Risk:

Employees with high monthly hours and high evaluation scores may indicate burnout rather than strong performance, suggesting that long hours without corresponding rewards or promotions lead to disengagement and turnover.

Monthly hours vs. promotions showed that employees with high monthly hours often did not receive promotions, indicating potential lack of recognition as a contributing factor to attrition.

- Department-Specific Attrition:

HR had the highest proportion of employees leaving, while management had the lowest. This department-based analysis can inform targeted retention strategies and intervention programs.

- Salary vs. Tenure:

Short-tenured employees were more likely to have low or medium salaries, while longer-tenured employees tended to maintain medium salaries throughout. This suggests that salary growth may be limited over time, which could contribute to attrition as employees seek better financial opportunities.

- Correlation Insights:

There was a positive correlation among variables like number of projects, monthly hours, and last evaluation, indicating that these factors often increase together. However, a negative association was found between satisfaction level and attrition, reinforcing the idea that satisfied employees are less likely to leave.

- Employee Satisfaction and Performance:

The combination of low satisfaction and high workload emerges as a strong indicator of burnout and potential attrition, underscoring the importance of balancing employee workload with adequate rewards and recognition.

- Monthly Hours Anomaly:

The average monthly hours of 201.05 hours per month was notably higher than the expected 166.67 hours, suggesting that many employees are working beyond the typical full-time schedule, which likely contributes to attrition risk.

These insights, drawn from the EDA and various visualizations, emphasize the critical role of workload balance, employee satisfaction, and recognition in determining turnover. They also highlight areas for potential intervention, such as reducing overwork, improving satisfaction, and ensuring that high performers are properly recognized and rewarded.

- What business recommendations do you propose based on the visualization(s) built?

Business Recommendations Based on Visualizations

- Cap the Number of Projects per Employee:

The visualizations clearly indicate that employees who are overworked (particularly those managing a higher number of projects and working excessive monthly hours) are at a significantly higher risk of attrition. Limiting the number of projects employees can handle — to a maximum of four projects — would help balance their workload, reduce burnout, and likely lower turnover rates.

- Reevaluate Employee Promotion Criteria:

The analysis of tenure and satisfaction levels reveals that employees who stay with the company for at least four years experience notable dissatisfaction. To improve retention, HR should either consider promoting employees after four years of service or conduct further investigations to understand why this group is dissatisfied. By offering more growth opportunities, employees may feel more valued and motivated to stay.

- Reassess Work Hours and Overtime Policies:

Many employees are working well beyond the expected full-time hours, contributing to burnout. To address this, the company should either reward employees for working longer hours or reduce the expectation for overtime. Additionally, it's crucial that employees are aware of the company's overtime pay policies if they aren't already. Clear communication around workload expectations, time off, and overtime pay can alleviate stress and contribute to a healthier work environment.

- Clarify Workload Expectations and Time Off:

To ensure alignment between management and employees, the company should make workload expectations and time-off policies explicit. This includes ensuring that employees understand the company's overtime policies and feel comfortable discussing workload issues without fear of reprisal. This transparency will reduce misunderstandings and promote a healthier work-life balance.

- Promote Open Discussions Around Company Culture:

The visualizations suggest that there may be an underlying issue with the company's work culture contributing to employee dissatisfaction and turnover. It's essential to hold company-wide and within-team discussions to better understand and address specific cultural or workload-related issues. Regular dialogues can help foster a more inclusive and supportive environment, improving employee morale and retention.

- Rewarding High Performance Beyond Workload:

The analysis shows that high evaluation scores are often tied to employees who work excessively long hours, yet this may not always indicate true performance. To incentivize sustainable performance, the company should consider a more proportionate scale for rewarding employees based on performance, rather than just the number of hours worked. This will prevent burnout and promote long-term engagement and loyalty.

- Incorporate Satisfaction Levels into Performance Reviews:

Satisfaction levels were identified as a strong indicator of employee attrition. HR should incorporate satisfaction surveys into regular performance reviews to identify early signs of disengagement, particularly among high-performing employees. Addressing dissatisfaction early allows the company to take proactive steps to improve morale before employees decide to leave.

- Balance Workload and Motivation:

The visualizations suggest that employees with high satisfaction and moderate workloads tend to stay longer, while those with imbalanced workloads (either too low or too high) are more likely to leave.

By regularly monitoring both workload and employee motivation, the company can uncover hidden retention risks and take corrective actions before they escalate.

- Track Burnout Indicators:

Burnout is a critical factor contributing to attrition, as revealed by the analysis of satisfaction vs. monthly hours. HR teams should incorporate burnout indicators (such as high monthly hours and low satisfaction) into performance reviews and employee feedback systems. This proactive approach will allow HR to intervene early and provide necessary support, reducing the likelihood of burnout and associated turnover.

- Model-Based Recommendations:

Based on the performance of the Random Forest model, I recommend deploying this top-performing model to enhance proactive retention strategies. This model has proven to be effective in predicting employee attrition and identifying at-risk employees.

Key Model Metrics:

- Recall: 90.36% — ensuring that most employees at risk of leaving are identified, minimizing false negatives.

- Precision: 87.04% — balancing the identification of leavers while avoiding false positives.

- Accuracy: 96.16% — overall effectiveness in classifying employees accurately.

Deploying the Random Forest model will allow HR to prioritize interventions more efficiently. The model can be integrated into the existing HR workflow to support targeted outreach and root cause analysis of potential resignations. Over time, this approach can significantly improve retention rates and support more informed workforce planning.

By leveraging this data-driven approach, the company can make more informed, proactive decisions that benefit both employees and the business as a whole.

These recommendations are designed to improve employee satisfaction, reduce burnout, and ultimately lower turnover by addressing workload imbalances, enhancing communication, and using data-driven insights to guide HR decisions.

- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?

Additional Questions for Further Research

- Is salary allocation fair relative to experience and tenure?

Exploring whether compensation is equitably distributed across different levels of experience and length of service can help identify potential disparities and inform more transparent salary policies.

- Are performance evaluations being conducted fairly?

Analyzing how evaluation scores vary across departments, managers, or employee demographics can reveal whether the evaluation process is consistent and unbiased.

- Is there a relationship between employee satisfaction and evaluation scores?

Investigating whether more satisfied employees tend to receive higher evaluations could help understand if satisfaction impacts perceived performance or if performance drives satisfaction.

These questions could guide further analysis and support strategic decision-making aimed at improving employee engagement, retention, and fairness.

- How might you share these visualizations with different audiences?

To effectively communicate the visualizations to different audiences, I would adjust the complexity of the visuals based on the technical knowledge of each group. Here's how I would approach it:

For Technical Stakeholders:

- Boxplots: Useful for analyzing distributions, skewness, and identifying outliers in variables like satisfaction level and evaluation scores.

- Correlation Heatmaps: Visualize relationships between variables and identify multicollinearity, aiding in understanding interdependencies.

- Bar Charts for Feature Importance: Display the ranking of features based on their importance from decision trees and random forests.

- Line Plots: Analyze the linearity of the logit in logistic regression models, helping to evaluate model assumptions and AUC-PR.

- Decision Tree Diagrams: Show how decision trees split data, providing interpretability of model logic.

- Confusion Matrices: Assess classification model performance, highlighting true positives, false positives, true negatives, and false negatives.

For Non-Technical Stakeholders:

- Histograms: Show the distribution of key variables like satisfaction level, evaluation scores, and monthly hours, making data easy to interpret.

- Grouped Histograms: Compare attrition rates across salary levels, departments, or tenure groups to spot trends.

- Stacked Column Charts: Visualize proportions of employees who stayed or left within departments or job roles, simplifying turnover patterns.

- Grouped Column Charts (Department vs. Employee Count): Display the distribution of employees who stayed versus those who left across departments for clear, comparative insights.

For Both Audiences:

- Scatterplots: These show relationships between continuous variables (e.g., satisfaction level vs. monthly hours). For technical stakeholders, scatterplots can be used for deeper analysis, while for non-technical audiences, they should be explained more intuitively. When presented with context and a narrative, scatterplots can be a helpful tool for both groups to understand correlations and patterns.

By tailoring the visualizations to the specific audience, I can ensure effective communication of the findings while making the insights accessible to both technical and non-technical stakeholders.

The Power of Statistics

- Did any part of the project involve or require A/B testing, or would it align with the overall goals of this project? If so, how would I design it, and what key business insights could potentially be gained from it?

No part of my project involved A/B testing, nor was it required to meet the goals of this analysis. A/B testing wasn't part of any current or previous phase of the project, and it doesn't align with the core objective — which was to build a predictive model for employee attrition.

That said, A/B testing could be considered in the future if the company decides to test specific interventions inspired by the model's insights. For example, if HR wants to assess the impact of capping employees at 4 projects, an A/B test could be set up:

- Group A (Control): No changes — employees continue under current conditions.

- Group B (Treatment): Limit employees to a maximum of 4 projects.

Metric: Compare attrition rates between the two groups over a 6-month period.

This kind of testing would fall outside the scope of my current project but could provide valuable insights into which actions are most effective at reducing turnover.

- What business recommendations do you propose based on your results?

Based on the visualizations and analysis, I propose the following business recommendations to improve employee retention and satisfaction:

- Cap the Number of Projects per Employee: Limiting the number of projects to four will help balance workloads, reduce burnout, and lower turnover rates.

- Reevaluate Employee Promotion Criteria: Consider promoting employees after four years of service or investigate why this group is dissatisfied to improve retention.

- Reassess Work Hours and Overtime Policies: Either reward employees for overtime or reduce expectations to avoid burnout, while ensuring employees understand the company's overtime policies.

- Clarify Workload Expectations and Time Off: Ensure clear communication about workload and time-off policies to promote a healthier work-life balance.

- Promote Open Discussions Around Company Culture: Regular team discussions to address cultural or workload issues can improve morale and retention.

- Reward High Performance Beyond Workload: Incentivize sustainable performance by rewarding results, not just hours worked.

- Incorporate Satisfaction Levels into Performance Reviews: Regularly measure satisfaction to identify disengagement early and take proactive steps.

- Balance Workload and Motivation: Monitor workload and motivation to prevent imbalance and hidden turnover risks.

- Track Burnout Indicators: Implement burnout indicators into performance reviews to intervene early and reduce turnover.

- Deploy the Random Forest Model: Integrating the top-performing Random Forest model into HR processes will enable targeted interventions and improve retention rates by identifying at-risk employees.

These recommendations can help the company create a more supportive, balanced, and proactive work environment, ultimately reducing attrition and improving retention.

Regression Analysis: Simplify Complex Data Relationships

- To interpret model results, why is it important to interpret the beta coefficients?

Interpreting beta coefficients is crucial because they provide insights into the relationships between the independent variables and the dependent variable in a regression model. Here's why:

Intercept and Slope Interpretation:

- The intercept (constant term) represents the predicted value of the dependent variable when all independent variables are zero.

- The slope (beta coefficient) represents the change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant. This tells us the strength and direction of the relationship.

Model Insights:

- Beta coefficients allow us to interpret how each predictor variable influences the outcome. For instance, in a logistic regression model, a positive beta coefficient indicates that as the independent variable increases, the likelihood of the outcome increases. Conversely, a negative coefficient means the outcome is less likely as the predictor increases.

Significance:

- When paired with statistical tests, the beta coefficients help us understand whether a relationship is statistically significant. A non-zero coefficient typically indicates that the predictor has a meaningful impact on the dependent variable, assuming the p-value is below a certain threshold (e.g., 0.05).

Comparison Across Variables:

- By interpreting beta coefficients, we can compare the relative importance of different variables. A larger magnitude of a coefficient typically suggests a stronger effect on the dependent variable, providing a guide for prioritizing interventions or decisions.

In summary, beta coefficients are essential for understanding the direction, magnitude, and significance of relationships in the model, offering actionable insights beyond just model performance metrics like accuracy or R-squared.

- What potential recommendations would you make to your manager/company?

Here are the key recommendations to improve employee retention and reduce burnout:

- Cap Projects per Employee: Limit employees to four projects to prevent overwork and burnout.

- Promote After Four Years: Consider promoting employees after four years or investigate their dissatisfaction to improve retention.

- Review Overtime Policies: Reward or reduce excessive overtime, and ensure employees are aware of overtime pay policies.

- Clarify Workload Expectations: Make workload and time-off policies transparent and encourage open discussions.

- Foster Open Culture Conversations: Hold regular discussions to identify and address cultural or workload-related issues.

- Reward True Performance: Tie high evaluation scores to actual performance, not just hours worked.

- Consider Satisfaction in Attrition Predictions: Include satisfaction levels in performance reviews to spot disengagement early.

- Monitor Workload and Motivation: Regularly track employee workload and motivation to prevent burnout.

- Include Burnout Indicators in Reviews: Use burnout and satisfaction data in performance reviews to enable early intervention.

These recommendations aim to balance workloads, improve satisfaction, and reduce turnover.

- Do you think your model could be improved? Why or why not? How?

Yes, I believe the model could still be improved. While it performs well, there are several potential areas to enhance its performance further:

- Increase Recall: Since recall is critical for identifying at-risk employees, improving this metric would help capture more leavers and reduce the likelihood of false negatives.

- Feature Engineering: Additional or modified features could help better capture turnover patterns. By creating new features or refining existing ones, the model might find stronger signals and improve its predictive power.

- Investigate Data Leakage: While previous checks were made, revisiting potential data leakage is essential. Certain features like last_evaluation and satisfaction_level may reflect outcomes rather than

causal factors. Evaluating model performance with and without these features could help understand their true influence.

- Alternative Modeling Focus: If features like last_evaluation are strong predictors of attrition, it may be worth reframing the problem. For instance, predicting evaluation or satisfaction scores themselves could be useful as leading indicators of turnover.

- Model Tuning:

 - Adjusting cross-validation parameters can ensure more robust hyperparameter optimization.

 - Resampling techniques (such as SMOTE or undersampling) could address class imbalance and improve model performance.

 - Class weights or threshold tuning can help optimize the model for specific business needs, such as minimizing false negatives and prioritizing retention efforts.

Despite these possible improvements, it's also possible that this model is close to achieving the best performance with the current dataset, given the thorough preprocessing and quality checks already performed. However, exploring these areas could still unlock valuable insights and performance improvements.

- What business recommendations do you propose based on the models built?

Based on the models I built, especially the top-performing Random Forest model, I recommend the following:

- Deploy the Random Forest Model: With high recall (90.36%), precision (87.04%), and accuracy (96.16%), this model is effective at identifying employees at risk of leaving. It's reliable for supporting proactive retention strategies.

- Integrate the Model into HR Workflow: Use the model to flag at-risk employees early, allowing HR to prioritize interventions and investigate root causes of dissatisfaction before resignations occur.

- Use Predictions to Guide Actions: Model outputs can help inform more targeted and timely outreach efforts, improve retention strategies, and support better workforce planning.

Overall, the model enables data-driven decision-making and allows the company to act before attrition becomes a larger issue.

- What key insights emerged from your model(s)?

The key insights from the models are as follows:

- Logistic Regression Insights:

- Good Performance for Stayers: The logistic regression model accurately identified employees who were likely to stay, with a strong performance for true negatives (stayers). However, it struggled with the minority class (leavers).

- Challenges with Attrition Prediction: The model misclassified a significant number of leavers, with low recall (26%) and precision (44%) for predicting employees who would leave. This is problematic for the business, as the primary goal is to identify at-risk employees to prevent attrition.

- Decision Tree Insights (Round 1):

- Better Performance on Attrition: The decision tree model outperformed logistic regression, especially in recall and F1-score for predicting leavers. This is crucial for detecting employee turnover.

- Strong Model for Imbalanced Data: With a high ROC AUC (0.9698), the decision tree showed good handling of class imbalance. It also demonstrated robustness without significant overfitting.

- Clear Interpretability: Decision trees provide easy-to-understand, interpretable results, making it easier for stakeholders to grasp how the model makes predictions.

- Random Forest Insights (Round 1):

- Best Overall Performance: The random forest model was the highest performing across all evaluation metrics (ROC AUC: 0.9804), with great recall for both leavers and stayers. This makes it ideal for dealing with the class imbalance.

- Effective Handling of Imbalance: By using ensembling techniques, random forest reduced overfitting and improved generalization. It handled the class imbalance better than logistic regression and decision trees, making it particularly valuable for attrition prediction.

- Post-Data Leakage Model Insights (Round 2):

- Improved Model Performance: After addressing data leakage and refining features, the random forest model continued to outperform the decision tree model, with a higher precision-recall balance and reduced false negatives.

- Critical Role of False Negatives: The random forest model achieved excellent results in minimizing false negatives, correctly identifying 450 out of 498 leavers. This is crucial, as the business goal is to proactively identify employees at risk of leaving.

- Acceptable Trade-Offs in False Positives: While the model misclassified some stayers as leavers, this trade-off was acceptable for the business, as it allows for proactive intervention in at-risk employees.

Final Insight:

The Random Forest Round 2 model emerged as the champion model. It outperformed both logistic regression and decision trees, particularly in predicting leavers (attrition) while maintaining high accuracy and recall. This model will serve as the basis for deploying targeted retention interventions, supporting the business goal of improving employee retention by identifying at-risk employees early.

- Do you have any ethical considerations at this stage?

Yes, at this stage, I will focus on minimizing potential bias to ensure that the model produces fair and equitable predictions. My goal is to develop a baseline model that is technically robust and also aligned with the overall problem-solving objective. I'll pay close attention to selecting and evaluating the most relevant performance metrics for the scenario to minimize the risk of misclassification and unintended consequences. Throughout this phase, fairness, transparency, and reliability will remain core priorities.

The Nuts and Bolts of Machine Learning

- What key insights emerged from your model(s)?

The key insights from the models revealed that while logistic regression performed well in identifying employees who would stay, it struggled with predicting attrition. Specifically, it had low recall and precision for the "leaver" class, misclassifying a significant number of employees who were likely to leave. Decision trees provided better performance for attrition prediction, with improved recall and F1 scores, but showed slight signs of overfitting. The model also handled class imbalance well, offering strong interpretability, making it easier to understand the decision-making process.

However, the Random Forest model emerged as the most effective model, outperforming both logistic regression and decision trees across all metrics. It showed the best overall performance in identifying leavers, with high recall, precision, and ROC AUC, while effectively managing class imbalance. After addressing potential data leakage in Round 2, the Random Forest model continued to perform exceptionally well, minimizing false negatives and correctly identifying most leavers, which aligns with the business goal of improving employee retention. This model is now the recommended choice for predicting attrition and supporting targeted employee retention interventions.

- What are the criteria for model selection?

The criteria for model selection in this analysis are primarily based on performance metrics and the business goal of predicting employee attrition. Several factors were considered to determine the best model:

- ROC AUC: This metric is important for assessing the model's ability to discriminate between employees likely to leave (leavers) and those likely to stay (stayers). Models were optimized for ROC AUC through hyperparameter tuning using GridSearchCV with a focus on maximizing this metric, as it reflects overall model performance in imbalanced datasets.

- Recall: Since the business objective is to predict employee attrition effectively, recall is the most crucial metric. It indicates the model's ability to correctly identify employees who will leave, minimizing false negatives (leavers misclassified as stayers). Given the importance of not missing potential leavers, high recall was prioritized.

- Precision, F1 Score, and Accuracy: While precision and F1 score were also considered, they were secondary to recall in this case. A balance between precision and recall ensures that the model not only correctly identifies leavers but also minimizes the number of false positives. Accuracy was considered but was not the primary selection criterion due to the class imbalance in the dataset.

- Overfitting Risk: Models were evaluated for overfitting, especially with the decision tree model. Random Forest, being an ensemble method, showed a stronger ability to generalize, which led to its selection over the decision tree model.

- Class Imbalance Handling: All models were evaluated for their effectiveness in managing the class imbalance between stayers and leavers. Random Forests handled this well, maintaining strong performance across both classes, making it the top choice.

The Random Forest model (Round 2) emerged as the champion due to its superior performance across these criteria, especially in identifying potential leavers with high recall and robust performance metrics.

- Does my model make sense? Are my final results acceptable?

Yes, the model makes sense, and the final results are acceptable, especially given the context of employee attrition prediction. The Random Forest model, after addressing potential data leakage and performing rigorous hyperparameter tuning, outperformed both logistic regression and decision tree models in identifying employees likely to leave. This aligns well with the primary business goal of improving employee retention by focusing on attrition prediction.

The Random Forest model demonstrated high accuracy, strong recall, and a good balance between precision and recall, particularly for the "leaver" class. The low number of false negatives (employees

misclassified as stayers) is crucial, as missing potential leavers can have significant consequences for retention efforts. Furthermore, the ensemble nature of Random Forest helped reduce overfitting while effectively handling the class imbalance in the dataset. The results from both the training and test sets suggest the model is robust and generalizable, making it suitable for deployment in real-world scenarios.

Therefore, the final results are acceptable, and the model can be confidently used for supporting targeted retention interventions.

- Were there any features that were not important at all? What if you take them out?

Based on the analysis above, it seems that tenure is one feature that appears to have very weak correlations with most other variables. Specifically, its correlation with other features like satisfaction level (-0.15) and number of projects (-0.13) is quite low, suggesting that tenure may not be strongly predictive of employee attrition, at least not in relation to the other features.

If we were to remove tenure from the model:

- The impact would likely be minimal in terms of performance, given its weak correlations with other features.

- Model simplicity could improve as it might reduce noise, especially if tenure is not contributing significantly to predictions.

- However, tenure could still have some explanatory power that is not fully captured by the correlation coefficients alone, so we'd need to validate through model performance (e.g., cross-validation or feature importance) before definitively removing it.

It would be useful to evaluate the model performance with and without tenure and observe if it has a noticeable impact on accuracy, precision, recall, or other relevant metrics. In any case, tenure seems to be a weaker predictor compared to other features like satisfaction level or average monthly hours.

- Given what you know about the data and the models you were using, what other questions could you address for the team?

Given the data and models, I could address several additional questions for the team, including:

- Is salary allocation fair relative to experience and tenure?

By analyzing salary distribution across different levels of experience and tenure, I can uncover any disparities that might exist, helping to inform more transparent and equitable compensation policies.

- Are performance evaluations being conducted fairly?

Investigating whether evaluation scores differ across departments, managers, or demographics will reveal if the evaluation process is consistent and unbiased.

- Is there a relationship between employee satisfaction and evaluation scores?

Exploring the correlation between satisfaction and evaluation scores can help determine whether higher satisfaction influences performance evaluations or if the reverse is true.

These questions could lead to actionable insights that promote fairness and enhance employee engagement and retention.

- What resources do you find yourself using as you complete this stage?

As I complete this stage, I find myself relying on the following resources:

- Pandas: For data import/export, cleaning, transformation, and exploratory data analysis (EDA). It's my primary tool for data manipulation in Python.

- Seaborn and Matplotlib: For creating various visualizations, including distribution plots, box plots, heatmaps, and decision trees to support both EDA and model evaluation.

- Matplotlib: Used for additional visualizations, particularly for model evaluation tools like ROC curves and confusion matrices.

- Scikit-learn: This library is central for training and evaluating models, including logistic regression, decision trees, and random forests. It helps with dataset splitting, hyperparameter tuning, and performance metrics like accuracy, precision, recall, F1 score, and AUC-PR.

- XGBoost: I use this framework for gradient boosting and to visualize feature importance in tree-based models, which helps improve predictive performance.

- Pickle: For saving and loading trained models, which ensures reproducibility and avoids retraining from scratch.

- GitHub: Essential for tracking version history, documenting changes, and maintaining a structured and reproducible workflow.

These resources are crucial for building, tuning, and evaluating machine learning models effectively at this stage.

- Is my model ethical?

Yes, my approach to model development and evaluation is ethical. I've taken several important steps to ensure that the model is both fair and transparent. Here are the key aspects that demonstrate the ethical considerations I've incorporated:

- Transparency in Communication:

I am committed to communicating the model's results with full transparency. By providing both strengths and limitations of the model, I avoid overstating its effectiveness and acknowledge potential weaknesses. This helps ensure that stakeholders have a clear, realistic understanding of what the model can and cannot do.

- Minimizing Bias:

I am actively working to minimize bias in my model. This includes careful selection of features and attention to potential sources of bias that could skew the predictions, particularly when it comes to sensitive factors like employee attrition. Ensuring fairness by addressing and reducing bias is a core ethical responsibility.

- Balanced Evaluation:

My focus on constructive criticism of the model's performance, rather than presenting an overly optimistic view, is a sign of ethical integrity. This approach helps prevent the model from being used inappropriately or irresponsibly.

- Consideration of Unintended Consequences:

By emphasizing the importance of avoiding harmful or misleading decisions, I am considering the broader impact of my model. Recognizing that misclassification of employee attrition could lead to ineffective interventions or potentially harmful HR decisions shows a commitment to responsible AI use.

- Focus on Fairness and Equity:

I'm mindful of ensuring the model produces fair and equitable predictions. This includes considering the ethical implications of using features such as demographic data and making sure that the model's recommendations don't inadvertently discriminate against certain employee groups.

In summary, my approach aligns with ethical best practices in data science. By prioritizing fairness, transparency, and accountability, I'm ensuring that the model serves the broader goal of responsible decision-making and minimizes the risk of harm or bias.

- When my model makes a mistake, what is happening? How does that translate to my use case?

When the model makes a mistake, it can either misclassify leavers as stayers or stayers as leavers, both of which have different implications for your use case:

- False Negatives (Leavers Misclassified as Stayers):

- What's Happening: The model predicts an employee is likely to stay when, in fact, they are at risk of leaving. This means the model fails to identify some employees who are at high risk of attrition.

- Implication for Use Case: False negatives are particularly critical because the business goal is to proactively address attrition risk. Failing to identify these employees could lead to missed opportunities for retention strategies, such as targeted interventions or engagement activities, which might prevent their departure.

- False Positives (Stayers Misclassified as Leavers):

- What's Happening: The model predicts an employee is at risk of leaving when, in fact, they are likely to stay. This means the model incorrectly identifies some employees as at-risk leavers.

- Implication for Use Case: While false positives lead to unnecessary interventions or engagement efforts for employees who aren't at risk, this is generally less problematic than false negatives. The business can still reach out to these employees to confirm satisfaction and perhaps prevent any potential dissatisfaction before it becomes a larger issue. This approach is less costly than failing to act on potential leavers.

In summary, the primary risk of the model's mistakes lies in false negatives — failing to identify employees likely to leave, which would result in missed opportunities to intervene and retain those individuals. However, false positives, while requiring some additional effort, are more manageable because they allow the business to engage with employees preemptively.