Salifort Motors: Employee Turnover Prediction Project

PACE Strategy Document - Plan stage

Google Advanced Data Analytics Capstone

Introduction

I will use this PACE strategy document to record my decisions and reflections as I begin the **Plan** stage of this capstone project. This document will help guide my thinking around the project's goals, stakeholders, success criteria, and overall approach. It will also serve as a foundation for all subsequent stages of the PACE workflow — Analyze, Construct, and Execute. Beyond this project, it will be a valuable reference to support my continued development as a data professional.

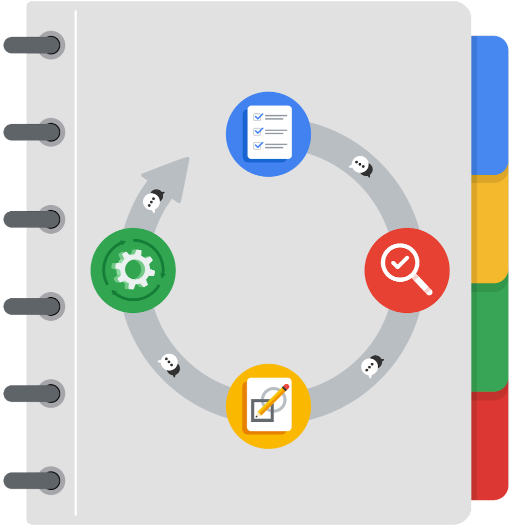
Portfolio Project Recap

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

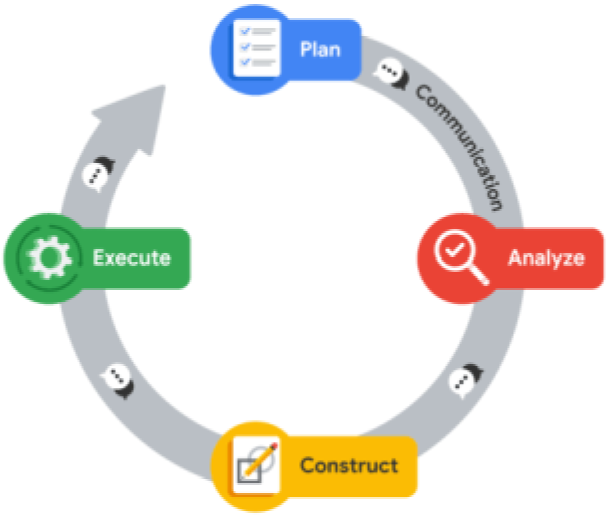


[Alt-text: The PACE Workflow with the four stages in a circle: plan, analyze, construct, and execute.]

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. These questions are primarily situated within the **Plan** stage of the PACE workflow, focusing on clarifying project goals, identifying key stakeholders, and establishing success criteria. I will address each question by aligning it with the most appropriate phase of the PACE framework — Plan, Analyze, Construct, or Execute — based on the specific project stage it reflects. 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.