Salifort Motors: Employee Turnover Prediction Project

PACE Strategy Document - Analyze stage

Google Advanced Data Analytics Capstone

Introduction

I will use this PACE strategy document to record my decisions and reflections as I move into the Analyze stage of this capstone project. This document will help guide my exploration and understanding of the dataset, shape my questions, assumptions, and hypotheses, and inform the analytical methods I apply. It will also support the transition into the Construct and Execute stages by grounding those efforts in well-reasoned insights. Beyond this project, it will serve as a valuable reference to support my continued growth as a data professional.

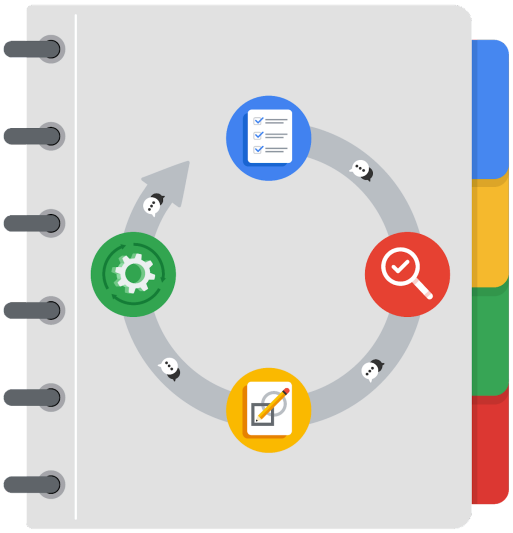
Portfolio Project Recap

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. These questions are primarily situated within the Analyze stage of the PACE workflow, focusing on exploring the dataset, identifying trends and relationships, and refining questions and hypotheses. 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 in data exploration and analysis.

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 column charts: 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.