**Course Seven**

# Google Advanced Data Analytics Capstone



# Instructions

Use this PACE strategy document to record your decisions and reflections as a data professional as you work through the capstone project. As a reminder, this document is a resource guide that you can reference in the future and a space to help guide your responses and reflections posed at various points throughout the project.

# Portfolio Project Recap

Many of the goals you accomplished in your individual course portfolio projects are incorporated into the Advanced Data Analytics capstone project including:

* Understand your data in the problem context
* Consider how your data will best address the business need
* Contextualize and understand the data and the problem
* Perform EDA (understand the variables and analyze relationships between them)
* Create visualizations
* Determine which models are most appropriate
* Construct the model
* Confirm model assumptions
* Evaluate model results to determine how well your model fits the data
* Interpret model performance and results
* Share actionable steps with stakeholders

**Project proposal**

**Salifort Motors project proposal**

## **Overview**

*Salifort Motors is seeking a method to use employee data to gauge what makes them leave the company.*

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Tasks** | **PACE stages** |
| **1** | **Understand the business scenario and define the problem** | **Plan** |
| **2** | **Data exploration and data cleaning** | **Plan, Analyze** |
| **3** | **Determine which models are most appropriate** | **Analyze,Construct** |
| **4** | **Construct the model** | **Construct** |
| **5** | **Confirm model assumptions** | **Analyze, Construct** |
| **6** | **Evaluate model results** | **Analyze** |
| **7** | **Interpret results and share actionable steps with stakeholders** | **Execute** |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

**Foundations of Data Science**

* Who is your audience for this project?

The audience is primarily the Human Resources (HR) department at Salifort Motors, including HR managers, business analysts, and senior leadership.

* 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?

The goal is to identify the key drivers of employee attrition and build a model that predicts whether an employee is at risk of leaving. This supports proactive HR intervention and improved retention.

* What questions need to be asked or answered?

What factors contribute most to employee attrition?

Can we predict who is likely to leave the company?

How can HR use these insights to improve retention?

* What resources are required to complete this project?

Clean HR dataset

Python libraries: pandas, scikit-learn, matplotlib, seaborn

Documentation and tutorials on ML model building and evaluation

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

A cleaned dataset

A trained and evaluated machine learning model

Key feature importance analysis

Visualizations and a 1-page stakeholder summary

**Get Started with Python**

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

Explore the structure and meaning of variables

Clean and encode data appropriately

Review model evaluation metrics

* What follow-along and self-review codebooks will help you perform this work?

Scikit-learn documentation

Google’s Machine Learning Crash Course

Course 6 and 7 sample notebooks

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

Clarify data types and business logic

Draft a checklist for EDA and feature prep

Research class imbalance strategies

**Go Beyond the Numbers: Translate Data into Insights**

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

Columns include: satisfaction\_level, last\_evaluation, num\_projects, average\_monthly\_hours, years\_at\_company, Work\_accident, promotion\_last\_5years, department, and salary. Most relevant: years\_at\_company, last\_evaluation, num\_projects, overworked, and salary.

* What units are your variables in?

Percentages (e.g., satisfaction\_level)

Counts (e.g., num\_projects)

Categorical (e.g., department, salary)

* 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?

Employees with high tenure, low recognition, or heavy workload may be more likely to leave.

* Is there any missing or incomplete data?

No missing values were detected during df.info() inspection.

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

Yes, though categorical features required label encoding.

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

Summary statistics (describe())

Histograms, boxplots

Grouping by left to compare feature distributions

Correlation matrix

**The Power of Statistics**

* What is the main purpose of this project?

To predict employee attrition and identify the key features that contribute to it.

* What is your research question for this project?

What employee characteristics are predictive of attrition at Salifort Motors?

* 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?

It ensures the training and test datasets are representative. Without it, the model might overfit to one department or seniority level, skewing results.

If all test data came from only new hires or one department, the model’s performance wouldn't generalize to the full employee base.

**Regression Analysis: Simplify Complex Data Relationships**

* Who are your stakeholders for this project?

HR executives, data analysts, and senior leadership.

* What are you trying to solve or accomplish?

Reduce attrition by understanding why employees leave and predicting at-risk individuals.

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

Employees with high project loads and no recent promotions are more likely to leave. Certain departments have higher turnover.

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

Scikit-learn documentation

Pandas guide

Google Data Analytics course materials

* Do you have any ethical considerations at this stage?

Yes. Ensure no protected attributes (e.g., gender, race) are used for predictions. Maintain employee data privacy and fairness in model interpretation.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?

Predict employee attrition based on HR data to enable targeted retention actions.

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

Course 7 capstone exemplar

Scikit-learn tutorials

GitHub notebooks on HR attrition

* Is my data reliable?

Yes. No nulls were found; features make sense and align with HR context.

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

Avoid bias in features; ensure transparency in how predictions are used.

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

Exit interview reasons, job satisfaction scores, peer/manager feedback, engagement survey results.

* What data do I have/can I get?

HR data including tenure, evaluation, project load, salary, department, and promotions.

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

F1 Score and ROC AUC — because we need to balance catching true positives (at-risk employees) while minimizing false positives (wasted intervention).

**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, the dataset contains meaningful features such as years\_at\_company, last\_evaluation, num\_projects, and salary, which are directly tied to employee performance and retention. These variables are sufficient to model attrition 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?

Check for null values and data types

Perform label encoding on categorical variables

Use descriptive statistics (.describe())

Visualize distributions (histograms, boxplots)

Group by attrition to identify feature differences

Generate a correlation heatmap

* 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 joins were needed; all relevant data was in a single CSV. Structuring steps included:

Encoding categorical columns (salary, department)

Creating new features (e.g., binary overworked)

Filtering out anomalies if needed

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

Bar charts to show categorical attrition rates

Boxplots for numerical feature comparison between classes

Correlation heatmap to assess linear relationships

Feature importance plots from models

**The Power of Statistics**

* Why are descriptive statistics useful?

They summarize the central tendency and dispersion of data, helping identify outliers, skewed distributions, and potential correlations with attrition.

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

Null Hypothesis (H₀): A given feature has no effect on attrition.

Alternative Hypothesis (H₁): The feature significantly affects attrition.

These help evaluate whether model-driven differences are statistically meaningful.

**Regression Analysis: Simplify Complex Data Relationships**

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

Identify multicollinearity

Detect outliers that could skew coefficients

Understand feature distribution and scale

Select and engineer meaningful predictors

* Do you have any ethical considerations at this stage?

Yes:

Avoid features that could encode bias (e.g., department culture as a proxy for gender/race).

Ensure transparency in which features are influencing model outcomes.

Maintain privacy of employees by anonymizing data.

**The Nuts and Bolts of Machine Learning**

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

The goal is still to predict employee attrition. Based on EDA and feature selection, the plan remains appropriate and valid.

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

While decision trees and random forests don’t require normality or equal variance, we handled imbalance in the target variable using stratified splits. So, assumptions were respected.

* Why did you select the X variables you did?

Variables were selected based on:

EDA findings (e.g., tenure, number of projects)

Feature importance rankings from models

Relevance to attrition prediction

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

Clean and prepare data

Select top contributing variables

Understand how the data behaves

Address imbalance and skew

* What has the EDA told you?

Employees with:

Long tenure

High evaluations without promotions

Many projects

Low salary

...are more likely to leave.

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

Seaborn documentation

Matplotlib documentation

Google’s Machine Learning Crash Course

Scikit-learn’s EDA and model selection guides

* Do you have any ethical considerations in this stage?

Remove or monitor proxy features for bias

Ensure feature selection doesn’t unintentionally disadvantage specific groups

Communicate model limits and avoid overreliance in sensitive HR decisions

**Data Project Questions & Considerations**

**PACE: Construct Stage**

**Get Started with Python**

* Do any data variables averages look unusual?

Yes — years\_at\_company has some long-tenured employees with no promotions in 5 years, which stood out during initial analysis. Also, average\_monthly\_hours was high for many employees flagged as "overworked."

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

The dataset includes several department groupings (e.g., sales, technical, support), and salary levels categorized as low, medium, and high.

**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?

Confusion matrix (model performance)

Feature importance plot (Decision Tree / Random Forest)

Accuracy, Precision, Recall, F1 Score, and ROC AUC

Decision tree structure for interpretability

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

Fitted the model on training data using GridSearchCV

Extracted and sorted feature importances

Used seaborn and matplotlib to plot charts

Used plot\_tree() for decision tree visualization

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

years\_at\_company, last\_evaluation, num\_projects, average\_monthly\_hours, salary

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

No missing values were detected (df.info() confirmed this). If there had been, appropriate imputation (mean/mode) or filtering would be applied based on variable type

**The Power of Statistics**

* How did you formulate your null hypothesis and alternative hypothesis?

Example for years\_at\_company:

H₀: Length of tenure has no effect on attrition.

H₁: Longer tenure is significantly associated with attrition.

This hypothesis was explored through correlation, feature importance, and model interpretation.

* What conclusion can be drawn from the hypothesis test?

The model and visual analysis support rejecting the null hypothesis — longer tenure, when not accompanied by promotion or reward, is strongly associated with a higher likelihood of attrition.

**Regression Analysis: Simplify Complex Data Relationships**

* Do you notice anything odd?

Yes — some employees with very high performance scores were leaving. These anomalies likely point to dissatisfaction despite good evaluations, possibly due to lack of recognition or promotion.

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

Yes. Potential improvements:

Try boosting methods like XGBoost or LightGBM

Use SMOTE to handle class imbalance

Include interaction terms between features (e.g., evaluation \* promotion\_last\_5years)

Integrate additional data (e.g., engagement scores, exit interviews)

**The Nuts and Bolts of Machine Learning**

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

The model works well, but some misclassifications occur. This could be improved with better handling of class imbalance or additional features.

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

Selected based on:

Domain knowledge (e.g., tenure, salary, workload)

EDA findings

Feature importance from decision trees

* How well does your model fit the data? (What is my model’s validation score?)

The model has:

High accuracy

Balanced precision and recall

Strong F1 and ROC AUC scores

This indicates a good fit for predicting attrition.

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

Yes. Possible ways:

Ensemble stacking

Feature engineering

Including external feedback/sentiment data

* Do you have any ethical considerations at this stage?

Absolutely:

Model should not be used as the sole decision-making tool for HR actions

Must avoid encoding bias (e.g., department as a proxy for demographics)

Ensure fairness and transparency in model output and interpretation

**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?

Investigate patterns among long-tenured employees with high evaluation scores but no recent promotions, as these profiles showed higher attrition risk. Review internal performance policies and career advancement pathways.

* What data initially presents as containing anomalies?

Some employees with perfect or near-perfect evaluations still left the company. This appears inconsistent and suggests deeper issues like lack of recognition or job satisfaction.

* What additional types of data could strengthen this dataset?

Employee engagement or satisfaction survey scores

Manager feedback or peer reviews

Exit interview responses

Promotion history with timestamps

Sentiment data from internal communication tools

**Go Beyond the Numbers: Translate Data into Insights**

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

Employees with longer tenure and many projects were more likely to leave

High evaluation scores alone were not sufficient to retain employees

Lack of promotion within the last 5 years was a recurring trend among those who left

Departments with high workload had higher attrition rates

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

Prioritize retention efforts for employees with high tenure and no promotion

Align recognition and promotions with high evaluation scores

Monitor overworked employees and rebalance workloads

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

Which departments are consistently overworked?

What is the average time from hire to exit for high-risk employees?

Are there any predictors of top performers leaving vs. underperformers?

* How might you share these visualizations with different audiences?

HR Managers: Interactive dashboards using Power BI or Tableau

Executives: Summary slides with key insights and action items

Analysts: Full Jupyter notebook with code, models, and charts

**The Power of Statistics**

* What key business insight(s) emerged from your A/B test?

N/A

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

Reward employee loyalty with recognition or advancement

Re-evaluate promotion cycles across departments

Use model predictions to trigger early engagement efforts

**Regression Analysis: Simplify Complex Data Relationships**

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

Beta coefficients help interpret the direction and strength of influence each variable has on the probability of attrition. For instance, a positive beta for years\_at\_company would indicate higher tenure increases the odds of leaving.

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

Monitor employees at risk based on the model's predictions

Use predictive scores to prioritize retention conversations

Share insights with department heads to tailor action plans

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

Yes. Improvement areas include:

Tuning additional hyperparameters

Trying ensemble models (e.g., XGBoost)

Incorporating employee feedback or survey data

Addressing class imbalance using techniques like SMOTE

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

Implement model-backed early warning systems in HR tools

Target high-risk segments with mentoring, training, or job rotation programs

Reassess promotion criteria for highly rated yet disengaged employees

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

Top features: years\_at\_company, last\_evaluation, num\_projects, overworked, and salary

Employees who were long-tenured, overworked, and not promoted were more likely to leave

The model had balanced false positive and false negative predictions, making it useful for HR decisions

* Do you have any ethical considerations at this stage?

Yes:

Avoid misuse of model results (e.g., penalizing predicted attrition risks)

Ensure fairness by excluding bias-related features

Keep predictions confidential and use them for supportive intervention, not disciplinary action

**The Nuts and Bolts of Machine Learning**

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

Attrition is not purely performance-driven; workload, lack of promotion, and tenure play stronger roles. Predictive models effectively identified these high-risk employee profiles.

* What are the criteria for model selection?

Models were evaluated based on:

F1 Score (balance between precision and recall)

ROC AUC (classification strength)

Interpretability (especially in the decision tree)

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

Yes, the model predictions aligned with business intuition and data insights. Performance metrics were strong, with a good balance of true/false positives and negatives.

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

Yes — features like Work\_accident and some department labels had low importance. Removing them could reduce noise and simplify the model without harming performance.

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

How long do high-risk employees typically stay before leaving?

Can we predict which retention strategies are most effective?

What’s the financial cost of attrition by department?

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

Scikit-learn Documentation

Seaborn Docs

Google ML Crash Course

* Is my model ethical?

Yes, provided it's used responsibly:

No protected attributes were used

Purpose is to support, not penalize, employees

Predictions are probabilistic, not deterministic

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

False positive: Employee is predicted to leave but stays — may trigger unnecessary retention efforts.

False negative: Employee leaves unexpectedly — missed opportunity for intervention.

These highlight the importance of using the model as one tool among many in HR strategy.