**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:

* Create a project proposal
* Demonstrate understanding of the form and function of Python
* Show how data professionals leverage Python to load, explore, extract, and organize information through custom functions
* Demonstrate understanding of how to organize and analyze a dataset to find the “story”
* Create a Jupyter notebook for exploratory data analysis (EDA)
* Create visualization(s) using Tableau
* Use Python to compute descriptive statistics and conduct a hypothesis test
* Build a multiple linear regression model with ANOVA testing
* Evaluate the model
* Demonstrate the ability to use a notebook environment to create a series of machine learning models on a dataset to solve a problem
* Articulate findings in an executive summary for external stakeholders

**Project proposal**

**Salifort Employee Retention project proposal**

## **Overview**

*This project aims to develop a Logistic Regression and Random Forest Classifier model to predict employee retention. This model will be built using the Google PACE (Plan, Analyze, Construct, Execute) framework, as a workflow, ensuring a structured and efficient approach.*

| **Business Problem** | | |
| --- | --- | --- |
| **Salifort Motors requests that we create a model to predict their employee retention, in order to reduce employee turnover, and all costs associated.** | | |
| **Milestones** | **Tasks** | **PACE stages** |
| **Week 1** | **Define the specific prediction task and target variable.** | **Plan** |
| **Week 1** | **Determine relevant features to be included in the model.** | **Plan** |
| **Week 1** | **Select appropriate evaluation metrics based on the business problem.** | **Plan** |
| **Week 1** | **Deliver completed proposal & Inform project stakeholders of progress** | **Plan** |
| **Week 2-3** | **Acquire and load the data using Python Pandas package** | **Analyze** |
| **Week 2-3** | **Perform exploratory data analysis (EDA) to understand data characteristics:**   * **Check for missing values, outliers, and data imbalances** * **Visualize feature distributions and relationships with target variable** * **Identify potential data quality issues and cleaning needs** | **Analyze** |
| **Week 2-3** | **Inform project stakeholders of findings and progress** | **Analyze** |
| **Week 3-4** | **Pre-process the data:**   * **Handle missing values (i.e. imputation, removal)** * **Address outliers (i.e. winsorization, capping)** * **Encode categorical features (i.e. OneHotEncoding)** * **Consider feature engineering to create new informative features (optional)** | **Construct** |
| **Week 3-4** | **Split data into training and testing sets (80/20 split)** | **Construct** |
| **Week 3-4** | **Define and train the Logistic Regression model using scikit-learn**  **Define and train the Decision Tree Model using scikit-learn (optional)** | **Construct** |
| **Week 5-6** | **Use the trained model to make predictions on the testing set** | **Execute** |
| **Week 5-6** | **Evaluate model performance using chosen metrics (accuracy, precision, recall, F1-score)** | **Execute** |
| **Week 5-6** | **Interpret model coefficients to understand the influence of features on predictions** | **Execute** |
| **Deliverables** | | |
| **Week 5-6** | **Well-documented Jupyter Notebook outlining the data processing, model training and evaluation steps** | |
| **Week 5-6** | **A report summarizing the project, including:**   * **Problem definition and business impact** * **Data exploration findings and data cleaning procedures** * **Model performance metrics and interpretation (if applicable)** * **Recommendations for future model improvements (i.e. hyperparameter tuning, feature selection)** | |
| **Success Criterion** | | |
| **Week 5-6** | **\*\*Model achieves a satisfactory F1 score (or other chosen metric) of at least 65% on the testing set.** | |
| **Week 5-6** | **The project deliverables are completed according to the timeline and meet the defined standards.** | |
| **Week 5-6** | **The project findings provide valuable insights for decision-making related to employee retention** | |
| **Conclusion** | | |
| **This project proposes building a Logistic Regression model using the PACE framework to address the business need for predicting Salifort Employee Retention. By following a structured approach, analyzing data thoroughly, and evaluating the model’s performance, we aim to develop a reliable solution to the Salifort Motors team.** | | |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

**Foundations of data science**

* Who is your audience for this project?
* 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?
* What questions need to be asked or answered?
* What resources are required to complete this project?
* What are the deliverables that will need to be created over the course of this project?

**Get Started with Python**

* How can you best prepare to understand and organize the provided information?
* What follow-along and self-review codebooks will help you perform this work?
* What are a couple additional activities a resourceful learner would perform before starting to code?

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

* What are the data columns and variables and which ones are most relevant to your deliverable?
* What units are your variables in?
* 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?
* Is there any missing or incomplete data?
* Are all pieces of this dataset in the same format?
* Which EDA practices will be required to begin this project?

**The Power of Statistics**

* What is the main purpose of this project?
* What is your research question for this project?
* 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?

**Regression Analysis: Simplify Complex Data Relationships**

* Who are your stakeholders for this project?
* What are you trying to solve or accomplish?
* What are your initial observations when you explore the data?
* What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
* Do you have any ethical considerations in this stage?

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?
* What resources do you find yourself using as you complete this stage?
* Is my data reliable?
* Do you have any additional ethical considerations in this stage?
* What data do I need/would I like to see in a perfect world to answer this question?
* What data do I have/can I get?
* What metric should I use to evaluate success of my business objective? Why?

**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?

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

**The Power of Statistics**

* Why are descriptive statistics useful?
* What is the difference between the null hypothesis and the alternative hypothesis?

**Regression Analysis: Simplify Complex Data Relationships**

* What are some purposes of EDA before constructing a multiple linear regression model?
* Do you have any ethical considerations in this stage?

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve? Does it still work? Does the plan need revising?
* Does the data break the assumptions of the model? Is that ok, or unacceptable?
* Why did you select the X variables you did?
* What are some purposes of EDA before constructing a model?
* What has the EDA told you?
* What resources do you find yourself using as you complete this stage?
* Do you have any ethical considerations in this stage?

**Data Project Questions & Considerations**

**PACE: Construct Stage**

**Get Started with Python**

* Do any data variables averages look unusual?
* How many vendors, organizations or groupings are included in this total data?

**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?
* What processes need to be performed in order to build the necessary data visualizations?
* Which variables are most applicable for the visualizations in this data project?
* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

**The Power of Statistics**

* How did you formulate your null hypothesis and alternative hypothesis?
* What conclusion can be drawn from the hypothesis test?

**Regression Analysis: Simplify Complex Data Relationships**

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

**The Nuts and Bolts of Machine Learning**

* Is there a problem? Can it be fixed? If so, how?
* Which independent variables did you choose for the model, and why?
* How well does your model fit the data? (What is my model’s validation score?)
* Can you improve it? Is there anything you would change about the model?
* Do you have any ethical considerations in this stage?

**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?
* What data initially presents as containing anomalies?
* What additional types of data could strengthen this dataset?

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

* What key insights emerged from your EDA and visualizations(s)?
* What business recommendations do you propose based on the visualization(s) built?
* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
* How might you share these visualizations with different audiences?

**The Power of Statistics**

* What key business insight(s) emerged from your A/B test?
* What business recommendations do you propose based on your results?

**Regression Analysis: Simplify Complex Data Relationships**

* To interpret model results, why is it important to interpret the beta coefficients?
* What potential recommendations would you make to your manager/company?
* Do you think your model could be improved? Why or why not? How?
* What business recommendations do you propose based on the models built?
* What key insights emerged from your model(s)?
* Do you have any ethical considerations at this stage?

**The Nuts and Bolts of Machine Learning**

* What key insights emerged from your model(s)?
* What are the criteria for model selection?
* Does my model make sense? Are my final results acceptable?
* Were there any features that were not important at all? What if you take them out?
* Given what you know about the data and the models you were using, what other questions could you address for the team?
* What resources do you find yourself using as you complete this stage?
* Is my model ethical?
* When my model makes a mistake, what is happening? How does that translate to my use case?