**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**

## **How many employees are leaving the Salifort Motor**

## **Overview**

there is a high rate of turnover among Salifort employees. (Note: In this context, turnover data includes both employees who choose to quit their job and employees who are let go). Salifort’s senior leadership team is concerned about how many employees are leaving the company. Salifort strives to create a corporate culture that supports employee success and professional development. Further, the high turnover rate is costly in the financial sense. Salifort makes a big investment in recruiting, training, and upskilling its employees.

If Salifort could predict whether an employee will leave the company, and discover the reasons behind their departure, they could better understand the problem and develop a solution.

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Tasks** | **PACE stages** |
| **1** | * **PACE proposal** * **Data Observation** | **Planning** |
| **2** | * **Data preparation** * **Exploratory data analysis (EDA)** | **Planning & Analyze** |
| **3** | * **Visualization dashboard** * **Visualization Sharing** * **Storytelling presentation** | **Analyze & Execute** |
| **4** | * **Measurement** * **Hypothesis testing** | **Analyze & Construct** |
| **5** | * **Model Assumption in linear regression** * **Linear regression** * **Model Evaluation** | **Analyze & Construct** |
| **6** | * **Model Assumption in classification** * **Machine Learning with Logistic regression, Random Forest tree with XGBoost** * **Model Evaluation** | **Analyze & Construct** |
| **7** | * **Executive summary document** * **Final presentation** | **Execute** |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

**Foundations of data science**

* Who is your audience for this project?
  + Human resource in Salifort company.
* 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?
  + Help improve environment of employee in Salifort. That why employee leave in company too much and finding the answer with data observation in HR employee dataset.
* What questions need to be asked or answered?
  + Why employee leave in company?
  + What might be frequency driving for turnover employees?
* What resources are required to complete this project?
  + Data source in csv file
  + Executive summary
  + Tableau Dashboard
  + Jupyter lab tools
* What are the deliverables that will need to be created over the course of this project?
  + Sent summarize result in Email or Stakeholder

**Get Started with Python**

* How can you best prepare to understand and organize the provided information?
  + When organization talking about some problem that mean they send requirement to data team and data team can analyze insight from problems with root cause and using data to solving information of stakeholder with always communication in each stage of PACE model
* What follow-along and self-review codebooks will help you perform this work?
  + Share my notebook to team and always communicate to team and stakeholder.
  + Writing markdown and comment on flow or step in note for understanding to self, team and stakeholder.
* What are a couple additional activities a resourceful learner would perform before starting to code?
  + Google certificate
  + Other resource on Google (Statistic and Machine Learning)

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

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

| **Column name** | **Type** | **Description** |
| --- | --- | --- |
| satisfaction\_level | int64 | The employee’s self-reported satisfaction level [0-1] |
| last\_evaluation | int64 | Score of employee's last performance review [0–1] |
| number\_project | int64 | Number of projects employee contributes to |
| average\_monthly\_hours | int64 | Average number of hours employee worked per month |
| time\_spend\_company | int64 | How long the employee has been with the company (years) |
| work\_accident | int64 | Whether or not the employee experienced an accident while at work |
| left | int64 | Whether or not the employee left the company |
| promotion\_last\_5years | int64 | Whether or not the employee was promoted in the last 5 years |
| department | str | The employee's department |
| salary | str | The employee's salary (low, medium, or high) |

* What units are your variables in?
  + Int and String
* 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?
  + Data preparation before analysis
  + Storytelling
* Which EDA practices will be required to begin this project?
  + Data preparation before analysis
    - Data dictionary
    - Data cleaning
    - Data manipulation and Data joining
    - Data analysis for insight
    - Data visualization
  + Tableau dashboard
  + Storytelling

**The Power of Statistics**

* What is the main purpose of this project?
  + Defined insight from why employee leave from Salifort company.
* 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?
  + Human resource and All employee in Salifort including leave employee.
* 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 dashboard links.)
  + <https://public.tableau.com/views/Whysalifortemployeeleavefromcompany/WhyemployeeleftinSalifortCompany?:language=th-TH&:display_count=n&:origin=viz_share_link>
* Do you have any ethical considerations in this stage?
  + Personal data protection
  + Bias data in sample data

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?
  + Predict model in machine learning that some attribute to affected to employee leaves.
* What resources do you find yourself using as you complete this stage? (Jupiter document)
* Is my data reliable?
  + If hypothesis test in research case more than 95% that mean this test case is reliable. For example.
* Do you have any additional ethical considerations in this stage?
  + If random data from train sets are bias to test model that may overfitting or underfitting
* What data do I need/would I like to see in a perfect world to answer this question?
  + Solving the question and solving to stakeholder with reliable and easy to understand answer.
* What data do I have/can I get?
* What metric should I use to evaluate success of my business objective? Why?
  + Sum square error and mean square error: In regression use find data are less error with summation differential square of y actual and y predict value in sum square error. Additionally mean square error that has sum square error divided by total y values in tables.
  + R square and adjust R-square: In regression use find data are less error and has maximum value is 1. R-square use in small dataset and Adjust R-square use in large dataset.
  + Confusion matrix: In classification use keep predict result data with test data in table that show count of predict data and actual data that show how keep data correctly when predict.
  + Accuracy: basic evaluation in confusion matrix define how correctly in overall class to predict data
  + Precision and Recall: basic evaluation in confusion matrix. In precision defined how correctly in each class to predict data and recall define rate of change data in each class to call in how to correctly.
  + F1 score: define precision and recall how to affect in overall in each class to accuracy.
  + Area Under Curve in Receiver Operating Characteristic Curve: In ROC show the data when receive from recall value with positive and negative value are valid in rate of positive and negative graph curve when random in sample and evaluation how to best random sample for predict data by using find area of under curve in ROC graph with min value is 0 and max value is 1

**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?
  + statistic process
  + Inferential statistic
  + visualization dashboard
  + my root cause question to finding answer

**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?
  + Understand root cause question before defined the answer.
  + Data cleaning
  + Data visualization from root cause question
* 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.?
  + Yes, because I will learn them like manipulating and joining.
  + If dataset is empty by null value, wrong data type and duplicate value in each row they need drop them before filtering sorting item to visualization.
* What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?
  + Pie chart, Bar chart and Line chart etc.
  + Top trend data
  + Using data visualization in root cause question when we focus on the data when analyze.

**The Power of Statistics**

* Why are descriptive statistics useful?
  + Because they are asking why data analytic using them for analysis and can describe about why this dataset is useful to analysis for finding insides before summarized.
* What is the difference between the null hypothesis and the alternative hypothesis?
  + Null hypothesis describes hypothesis is normally in sample phase. In opposite ways of null hypothesis is alternative hypothesis means hypothesis is abnormally in sample phase by using critical parameters assumption.

**Regression Analysis: Simplify Complex Data Relationships**

* What are some purposes of EDA before constructing a multiple linear regression model?
  + Define dependent value and independent value to model assumption that mean they correlated and have efficient for creating optimize solution model to use in really problems.
* Do you have any ethical considerations in this stage?
  + If use EDA but forget cleaning data before visualize data. Their data is not integrity, is dirty, make more bias on implementation.
  + If data are not manipulating or joining in some case to use plot graph with relational like aggregate data
  + If analyzed data with model assumption and wrong. They will not use in line regression and machine learning model.
  + If set wrong hypothesis. They can make data unreliable.

**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?
  + Use to define the answer from the most affected of EDA to solve optimize solution.
* What are some purposes of EDA before constructing a model?
  + Best data selection model
  + Label and Rescale feature to make model assumption
  + Make feature can be available to model solution
  + Make validation in best way like randomize with train, validate and test
* What has the EDA told you?
  + How data preparation in best way to solving really test data?
  + How train and test data with randomize have efficiency for optimize solution
* What resources do you find yourself using as you complete this stage?
  + python tutorial
  + sklearn tool
  + numpy
  + seaborn + pyplot
* Do you have any ethical considerations in this stage?
  + If analyzed data with model assumption and wrong. They will not use in line regression and machine learning model.
  + If models have not efficiency to predict data that mean something wrong in your dependent value before make model like bias (overfit case) and model can predict test data but train not well (underfit data).

**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?