**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 leads them to depart from the company.

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Tasks** | **PACE stages** |
| **1** | **Understand the business scenario and define the problem** | **Plan** |
| **2** | **Data exploration and data cleaning** | **Analyze** |
| **3** | **Determine which models are most appropriate** | **Analyze,Construct** |
| **4** | **Construct the model** | **Construct** |
| **5** | **Confirm model assumptions** | **Construct** |
| **6** | **Evaluate model results** | **Construct** |
| **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 for this project are the management team of Salifort Motors who have been engaged by upper-level decision makers to determine ways to understand the reasons for employee turnover.
* 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?
  + My goal is to develop a model which predicts the factors which are likely to lead employees to depart the company. Alternatively, I also hope to identify ways to enhance employee longevity and pass on some recommendations to the management team. These insights will hopefully assist them in building a more healthy and sustainable work culture.
* What questions need to be asked or answered?
  + What internal factors drive employees to leave the company?
  + What can the company improve to increase employee retention?
* What resources are required to complete this project?
  + Past and present employee data
* What are the deliverables that will need to be created over the course of this project?
  + An executive summary detailing findings with informative charts, figures and insights.
  + A detailed workbook where all PACE milestones are completed. It should display the results of our findings and evaluate each model to demonstrate their effectiveness. It should be adequately built and designed to facilitate further improvements.

**Get Started with Python**

* How can you best prepare to understand and organize the provided information?
  + Having clear objectives in mind is important to ensure all decisions are driven towards finding any and all insights explaining employee turnover.
* What follow-along and self-review codebooks will help you perform this work?
  + I mostly relied on notebooks from: ”Go Beyond the Numbers: Translate Data into Insights” and “The Nuts and Bolts of Machine Learning”.
* What are a couple additional activities a resourceful learner would perform before starting to code?
  + Have a look through the dataset to get moderately familiar with it.
  + Read through milestones and objectives to identify at an early stage which columns could provide useful insights.

**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 data columns included in the dataset include:
    - Satisfaction Level Scores
    - Last Evaluation Scores
    - Number of Projects worked
    - Average monthly hours worked
    - Time spent working for the company
    - Work Accident (whether they had one)
    - Left (whether they left the company)
    - Promotion Last 5 years (whether they were promoted during this time)
    - Department
    - Salary
  + Some engineered variables such as Tenure and Overworked (whether they worked over 175 hours a month) were key in generating the deliverables.
* What units are your variables in?
  + Continuous variables (Satisfaction Level Scores, Last Evaluation Scores, Number of Projects worked, Average monthly hours worked, Time spent working for the company)
  + Categorical variables (Work Accident, Left, Promotion Last 5 years, 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?
  + The idea that salaries and promotions could be strong reasons for employee departure.
  + The idea that some departments may have more turnover than others.
* Is there any missing or incomplete data?
  + There is no missing data.
* Are all pieces of this dataset in the same format?
  + The data are according to the different variable types (int/float for continuous and Boolean/text for categorical)
* Which EDA practices will be required to begin this project?
  + Ensure the data is bias free and it comes from a source whose integrity can be relied upon. If any irregularities are observed, more data may need to be collected, or strategies should be devised to ignore or engineer certain variables.

**The Power of Statistics**

* What is the main purpose of this project?
  + To develop a classification model which predicts employee turnover based on the most likely factors and achieve a model with a high success rate.
* What is your research question for this project?
  + What factors within the dataset could be good indicators of employee dissatisfaction and thus leading to their departure from the company?
* 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 crucial to avoid any bias from the dataset affecting our findings. This allows the samples taken to be representative of the population that is being studied.

**Regression Analysis: Simplify Complex Data Relationships**

* Who are your stakeholders for this project?
  + The management team.
* What are you trying to solve or accomplish?
  + We are trying to find which characteristics impact employees negatively and eventually lead to them leaving the company. These insights will help the management team in molding the work culture which they hope to create.
* What are your initial observations when you explore the data?
  + The statistics for employees are relatively similar across all departments.
  + There does not seem to be any missing data.
* What resources do you find yourself using as you complete this stage?
  + I mostly relied on notebooks from:” Regression Analysis: Simplify Complex Data Relationships” and “The Nuts and Bolts of Machine Learning”.
  + I referred to the TikTok and Waze projects which I have completed earlier in this program.
* Do you have any ethical considerations at this stage?
  + There are some factors not included in the dataset which may be relevant such as age, marital status, number of children, and an employee’s average job tenure across their career. To ensure no misclassification or discrimination occurs, these may need to be collected and used in future versions of this model.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?
  + I am trying to identify the characteristics that have the highest correlation with employees leaving the company.
* What resources do you find yourself using as you complete this stage?
  + I used the Random Forest notebooks in “The Nuts and Bolts of Machine Learning”.
* Is my data reliable?
  + We are using first party data which was generated strictly for this project and thus is unlikely to be unreliable and misrepresenting of the employee population at the company.
* Do you have any additional ethical considerations in this stage?
  + We need to ensure the training; test splits are representative of all employees at the company. It has to not be skewed in any direction and not result in overfitting or underfitting.
* What data do I need/would I like to see in a perfect world to answer this question?
  + Ideally a larger dataset encompassing all branches (if possible) and past years’ employee data.
  + A few extra columns in the dataset for factors such as: such as age, marital status, number of children, an employee’s average job tenure across their career, and distance travelled to reach work.
* What data do I have/can I get?
  + Data on 12000 employees for 10 columns.
* What metric should I use to evaluate success of my business objective? Why?
  + I will use the four metric of success (accuracy, recall, precision, f1) to grade the quality of the insights generated from the model.
  + Transparency is crucial to ensure these insights and metrics are emphasized to stakeholders to give them confidence when making data-driven decision for the company.

**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?
  + I believe the features provided should be more than enough to provide some insights as to why employees are leaving the company. More importantly, it includes most of the features which can be impacted by the management team (i.e no private factors).

**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?
  + There are several outliers that need to be removed.
* 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.?
  + I do not believe any additional data is necessary to derive insights for the stakeholders. The current features available seem to have already indicated numerous major factors leading to departure from the company.
* What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?
  + Histograms, bar charts and scatter plots should be useful to compare the distributions across different departments and other categorical variables.
  + Heat maps should also be great to demonstrate correlation and confusion matrices can validate results.

**The Power of Statistics**

* Why are descriptive statistics useful?
  + It allows for data to be presented in meaningful and understandable ways thus enabling better transparency.
* What is the difference between the null hypothesis and the alternative hypothesis?
  + The null hypothesis basically states that there are no statistically significant differences between two or more experimental or control groups. By contrast, the alternative hypothesis indicates that statistically significant differences occur between two or more experimental or control groups.

**Regression Analysis: Simplify Complex Data Relationships**

* What are some purposes of EDA before constructing a multiple linear regression model?
  + EDA ensures that the data is free of bias and errors, thus guaranteeing that the results produced by regression are as accurate as possible.
* Do you have any ethical considerations at this stage?
  + During the EDA process, features might be removed or engineered to facilitate future steps. However, we should be wary of not introducing new biases or removing key details which yield more accurate results.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve? Does it still work? Does the plan need revising?
  + The analysis so far shows a correlation between the hours worked and the number of employees leaving the company. There are a few more factors which may show correlation such as promotions and last evaluation scores. The plan seems to have been fairly correct and we will continue exploring these relationships.
* Does the data break the assumptions of the model? Is that ok, or unacceptable?
  + So far, the data is still faithful to the assumptions.
* Why did you select the X variables you did?
  + They were chosen based on good correlation coefficients to the target variable.
* What are some purposes of EDA before constructing a model?
  + EDA helped us with targeting the features which are more likely to have strong correlation coefficients with the correct charts to quickly and visually explore any relationship if present.
* What has the EDA told you?
  + Some variables were slightly correlated and thus we engineered new features to bring forward those relationships.
* 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 at 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?