**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 project proposal**

## **Overview**

*Salifort seeks a way to use a model to predict whether an employee will leave or stay in the company; the ultimate goal is to reduce turnover in the company.*

| **Milestones** | **Tasks** | **PACE stages** |
| --- | --- | --- |
| **1** | Define project scope; Establish structure for project workflow | **Plan** |
| **2** | Write a project proposal | **Plan** |
| **3** | Compile summary information about the data | **Analyze** |
| **4** | Begin exploring the data | **Analyze** |
| **5** | Data exploring and cleaning | **Plan Analyze** |
| **6** | Exploratory data analysis report | **Analyze Construct** |
| **8** | Select modeling approach | **Construct** |
| **11** | Build a machine learning model | **Construct** |
| **12** | Evaluate the model | **Execute** |
| **13** | Share results and final insights with stakeholders | **Execute** |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

**Foundations of data science**

* **Who is your audience for this project?**Salifort stakeholders.
* **What are you trying to solve or accomplish?** I will build a regression model and a machine learning model to predict whether an employee will churn the company. If I can convey to the company's stakeholders under what parameters the model was able to predict employees' churn, they will be able to improve their environment, increase employees' satisfaction, and reduce the company's turnover.
* **What questions need to be asked or answered?**

Why are the employees leaving the company?

Which variables (job titles, department, etc.) are more related to these leaves?

Which variables (job titles, department, etc.) need improvement to increase job satisfaction?

* **What resources are required to complete this project?**

The data collected by human resources.

Basic hardware and software (computers, python and its tools, visualization software, etc.)

Stakeholders’ feedback and availability.

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

Project proposal.

Exploratory data analysis report.

Visualizations.

Logistic regression model.

Machine learning model.

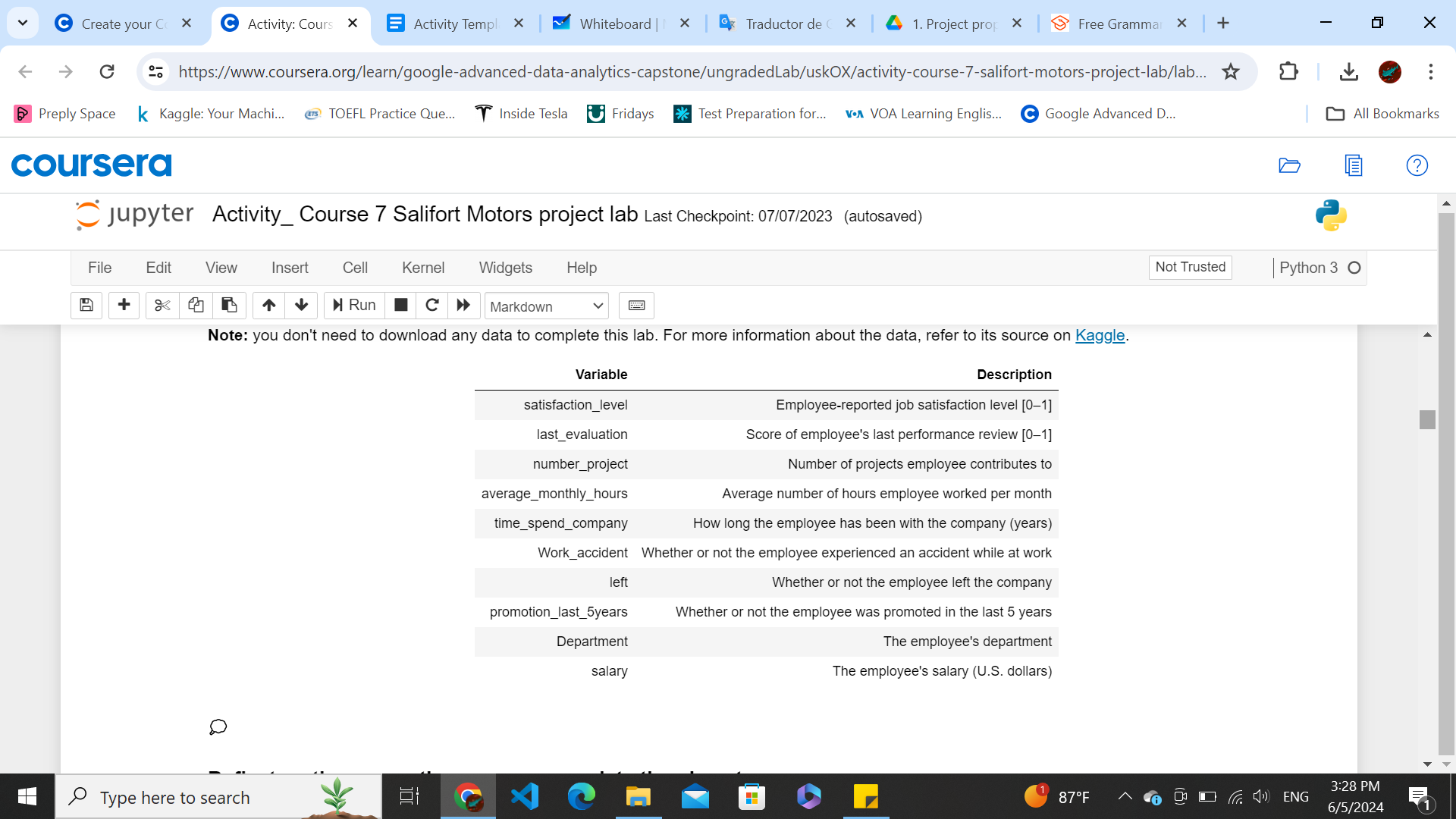
Executive summary.

**Get Started with Python**

* **How can you best prepare to understand and organize the provided information?** I will need a data dictionary to understand the data gathered, and then I will explore the data through exploratory data analysis. I may need to address questions with stakeholders.
* **What follow-along and self-review codebooks will help you perform this work?** Online documentation, past codebooks I have worked in, online forums, and any other type of information I can find.
* **What are a couple additional activities a resourceful learner would perform before starting to code?** Plan it's time, find a good place to focus and work, and gather all the necessary tools to keep a continuous workflow.

**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 left column is very important because this is going to be the target variable. I theorize that 'satisfaction\_level', 'time\_spend\_company', and 'salary' are going to be the variables with more weight in the model.



* **What units are your variables in?**

Float: 2 satisfaction level, and last evaluation.

Integer: 3 number of projects, average monthly hours, and time spent at company.

Binary: 3 work accidents, left, and promotion last 5 years.

Categorical: 2 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?**

Workers with low salaries and high work hours are more likely to go.

The workers who have reported a low satisfaction score are more likely to go.

Workers who have a bad evaluation report lower satisfaction.

* **Is there any missing or incomplete data?** There are no missing values or incomplete data.
* **Are all pieces of this dataset in the same format?** Yes, all the pieces are loaded into the same dataset.
* **Which EDA practices will be required to begin this project?** Clean missing values, look for duplicates, handle outliers, and double-check any odd values.

**The Power of Statistics**

* **What is the main purpose of this project?** Understand which variables are contributing to the employees churning. In order to be able to predict when an employee is more likely to leave the company and reduce internal turnover.
* **What is your research question for this project?** What are the variables that are contributing to employees leaving 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?** We are conducting this process on a company level; without random sampling, we take the risk of having bias that some departments or any other conditions can have. For example, we could find a low level of turnover in a department that is stewardship for a good manager if we only used the information from this department. We won't discover the real reasons why the employees are leaving the company.

**Regression Analysis: Simplify Complex Data Relationships**

* **Who are your stakeholders for this project?** Salifort HR team and managers.
* **What are you trying to solve or accomplish?** Build a logistic regression model able to predict whether an employee will churn the company.
* **What are your initial observations when you explore the data?** The data seems to be in good condition. I didn't find any missing values. All duplicates were dropped (total of duplicates dropped: 3008). I found many outliers, all of them in the time spent at the company column. This means they represent people who worked extra hours, so I will keep them in the dataset (total of outliers found: 824).
* **What resources do you find yourself using as you complete this stage? (Make sure to include the links.)** Python, Seaborn, Matplotlib, and Pandas.
* **Do you have any ethical considerations in this stage?** I decided not to drop the outliers because they represent people who worked overtime. This could be a signal of labor exploitation or an unfair balance of hours. This could be a variable of weight for the model.

**The Nuts and Bolts of Machine Learning**

* **What am I trying to solve?** Construct a random forest model able to predict whether an employee will churn or not.
* **What resources do you find yourself using as you complete this stage?** Pandas documentation: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get\_dummies.html, and the stack exchange forum: https://datascience.stackexchange.com/questions/20199/train-test-split-error-found-input-variables-with-inconsistent-numbers-of-sam.
* **Is my data reliable?** Yes, the data I’m working with is reliable.
* **Do you have any additional ethical considerations in this stage?** No, only the same consideration I had in the past stages.
* **What data do I need/would I like to see in a perfect world to answer this question?** A categorical column that encompasses different categories because they left the company
* **What data do I have/can I get?** Only the data that the HR team gathered for me.
* **What metric should I use to evaluate success of my business objective? Why?** In this case, a false negative or positive could be equally harmful. The F1 score would be the perfect metric to evaluate this model.

**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 data has different types of variables with useful information.

**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 the original state of the data, handle outliers, missing values, and duplicates. Check the data in every column, modify them if necessary, and change the data structure as necessary. It's always wise to double-check your final dataset.
* **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 don't need to add more data.
* **What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?** To check the relationship between the 'left' column and the rest of the columns, I will need different types of visualizations. line graphs, histograms, pie charts, and box plots. I would like to check the relationship between some variables, like salary, and other variables; they shouldn't have a correlation between them because they are considered dependent variables.

**The Power of Statistics**

* **Why are descriptive statistics useful?** It is a tool that gives us the opportunity to reveal hidden correlations and patterns between the data. Descriptive statistics allow us to convey this data in more accessible ways, like visualizations and insights.
* **What is the difference between the null hypothesis and the alternative hypothesis?** The null hypothesis states that there is no statistically significant difference between two groups or more groups. When we failed to prove that the null hypothesis was true, we proved that the alternative hypothesis was true, which is exactly the opposite of the null hypothesis. This happens when the p-value is greater than the significant level.

**Regression Analysis: Simplify Complex Data Relationships**

* **What are some purposes of EDA before constructing a multiple linear regression model?** The EDA allows us to understand and prepare the data to be used for the model.
* **Do you have any ethical considerations in this stage?** Convey the results with no bias. Altering the model to show results that encourage the stakeholders to think of something that pleases them more instead of showing them the actual results would be an unethical practice.

**The Nuts and Bolts of Machine Learning**

* **What am I trying to solve? Does it still work? Does the plan need revising?** Make a random forest model that can predict when an employee is likely to churn. This model is perfect for this type of case because its assumptions fit perfectly with the data we have, so we can say with security that this plan will work.
* **Does the data break the assumptions of the model? Is that ok, or unacceptable?** The data doesn't break the model's assumptions. If it does, we should use another model.
* **Why did you select the X variables you did?** Because those are independent variables, which could have an effect on the dependent variable.
* **What are some purposes of EDA before constructing a model?** The EDA allows us to understand and prepare the data to be used for the model.
* **What has the EDA told you?** The EDA told me that there could be a close correlation between the satisfaction level and the leavings. It also told me some imperfections that the dataset had, like some duplicate values.
* **What resources do you find yourself using as you complete this stage?** Past machine learning laboratories I have worked in, include Pandas and sklearn documentation.
* **Do you have any ethical considerations in this stage?** Convey the results with no bias. Altering the model to show results that encourage the stakeholders to think of something that pleases them more instead of showing them the actual results would be an unethical practice.

**Data Project Questions & Considerations**

**PACE: Construct Stage**

**Get Started with Python**

* **Do any data variables averages look unusual?** No, all variables seemed to be in a good state.
* **How many vendors, organizations or groupings are included in this total data?** All the company's departments.

**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?** Many data visualizations using Matplotlib and Seaborn, a random forest classifier, and gridsearch.
* **What processes need to be performed in order to build the necessary data visualizations?** Grouping and sorting the variables as needed.
* **Which variables are most applicable for the visualizations in this data project?** All the variables are applicable, but I will work more with 'left' the target variable.
* **Going back to the Plan stage, how do you plan to deal with the missing data (if any)?** I will drop it.

**The Power of Statistics**

* **How did you formulate your null hypothesis and alternative hypothesis?** Does not apply.
* **What conclusion can be drawn from the hypothesis test?** does not apply.

**Regression Analysis: Simplify Complex Data Relationships**

* **Do you notice anything odd?** does not apply.
* **Can you improve it? Is there anything you would change about the model?** does not apply.

**The Nuts and Bolts of Machine Learning**

* **Is there a problem? Can it be fixed? If so, how?** Any big problem came across during the construction of the model.
* **Which independent variables did you choose for the model, and why?** All variables but 'left' because those independent variables are not related to each other.
* **How well does your model fit the data? (What is my model’s validation score?)** The model fitted the data well; it scored a 0.95.
* **Can you improve it? Is there anything you would change about the model?** I could improve it using a wider range of hyperparameters. However, it performed well, so it is not necessary to do it.
* **Do you have any ethical considerations in this stage?** Convey the results with no bias. Altering the model to show results that encourage the stakeholders to think of something that pleases them more instead of showing them the actual results would be an unethical practice.

**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?** Invest in increasing the satisfaction of the employees who have less than five years working in the company.
* **What data initially presents as containing anomalies?** None.
* **What additional types of data could strengthen this dataset?** Employee positions and direct supervisors.

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

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

1. As the level of satisfaction increased, the number of people who stayed in the company did as well. In the opposite way, as the level of satisfaction decreased, the number of people who left increased.
2. 59% of the people who have left the company had a low salary, while just 2.4% had a high salary. However, it's important to consider the balance: 47.87% of the dataset is in the low salary range, while 8.26% is in the high salary range.
3. Anyway, it seems that the salary is not related to the level of satisfaction because the level of satisfaction follows a similar distribution between the three ranges of salaries. It's probable that the salary is not related to the churn of the employees.
4. The two departments with more churns are sales (27.6%) and technical (19.6%). The department with the lowest number of churns is management, with a 2.6%.
5. All the departments follow a similar distribution in their salary ranges. The most of the employees are in the ranges of low and medium, and the less in high. The only department that does not follow this rule is management, where the three ranges are very similar.
6. All the departments have a similar average of monthly worked hours and have spent, on average, the same number of years in the company. The time they work does not seem to be related to why some departments have more turnover than others.
7. However, it could be a relationship between the churnings and the time. On average, employees who have worked more hours and spent more years in the company are more likely to leave**.**
8. The last point is reinforced because it seems there is a relationship between time and satisfaction level. After the 250 average monthly hours worked, the satisfaction levels decreased drastically. On the other hand, the satisfaction level decreased gradually in the first 4 years of the company; after that point, it went up gradually. The promotion after five years could be related to that.

* **What business recommendations do you propose based on the visualization(s) built?** Measure the number of projects and rises that the employees have in their first 5 years in the company. Set up the company's culture in some way so that employees feel they will grow with the company, and, in consequence, increase their satisfaction.
* **Given what you know about the data and the visualizations you were using, what other questions could you research for the team?** Which variables are related to the satisfaction level, departmental information like time or salary balances, etc.
* **How might you share these visualizations with different audiences?** Using an executive summary or presentation.

**The Power of Statistics**

* **What key business insight(s) emerged from your A/B test?** does not apply.
* **What business recommendations do you propose based on your results?** Invest in increasing the satisfaction of the employees who have less than five years working in the company.

**Regression Analysis: Simplify Complex Data Relationships**

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

**The Nuts and Bolts of Machine Learning**

* **What key insights emerged from your model(s)?** Satisfaction level and the number of projects are the two variables of importance. They directly affect whether or not an employee churns.
* **What are the criteria for model selection?** The data doesn't break the model's assumptions. Besides, random forest is a powerful model with a low level of bias, which is perfect because we are working with employee information.
* **Does my model make sense? Are my final results acceptable?** Yes, and the results are very good. Besides, the model was validated and tested.
* **Were there any features that were not important at all? What if you take them out?** Se utilizaron todas las características, pero algunas de ellas carecen de importancia.
* **Given what you know about the data and the models you were using, what other questions could you address for the team?** Which variables are related to the satisfaction level, departmental information like time or salary balances, etc.
* **What resources do you find yourself using as you complete this stage?** Only Python and Sklearn metrics and visualization tools.
* **Is my model ethical?** Yes, the model was built ethically without considering personal variables that can affect employees or manipulating the data to show other results.
* **When my model makes a mistake, what is happening? How does that translate to my use case?** When the model makes a mistake, an employee we expect to stay will leave, which means the company will have to use resources to fulfill that position. On the other hand, when the model makes a mistake, the company will invest in resources not needed to keep an employee who will anyway stay.