# 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 Motors project proposal

## Overview

*Salifort Motors is looking for a way to analyze their employees' data to understand the reasons behind their departures from the company.*

---

| Milestones | Tasks | PACE stages |
|---|---|---|
| 1 | Define the problem at hand within the business scope. | Plan |
| 2 | Explore the data and clean it. | Plan, Analyze |
| 3 | Choose a model based on the data exploration | Analyze,Construct |
| 4 | Build the model | Construct |
| 5 | Make sure assumptions about the model are met. | Analyze, Construct |
| 6 | Evaluate the model | Analyze |
| 7 | Interpret results | Execute |

## Data Project Questions & Considerations

### PACE: Plan Stage

**Foundations of data science**

- Who is your audience for this project? Salifort Motors stakeholders
- 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?

  Im trying to create a model that predicts which employees will churn or not, this will allow the business to know what is making the employees leave.
- What questions need to be asked or answered?

  What are the important features, Is it ethical?
- What resources are required to complete this project?

  Dataset, ide, python packages.
- What are the deliverables that will need to be created over the course of this project?

  Model, visualizations, executive summary.

**Get Started with Python**

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

  Explore the data using pandas and other tools.
- What follow-along and self-review codebooks will help you perform this work?

  Past course labs.
- What are a couple additional activities a resourceful learner would perform before starting to code?

  Research the company, understanding what the features mean.

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

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

  satisfaction_level: Reflects the level of job satisfaction reported by the employee, with a range from 0 to 1.

last_evaluation: Represents the score from the employee's most recent performance review, also ranging from 0 to 1.
number_project: Indicates the total number of projects the employee is involved in.
average_monthly_hours: The average number of hours the employee works each month.
time_spend_company: How many years the employee has been working at the company.
Work_accident: Denotes whether the employee had any work-related accidents.
left: Specifies whether the employee has left the company.
promotion_last_5years: Indicates whether the employee received any promotions in the last five years.
Department: The department in which the employee works.
salary: The employee's salary level.
For the purpose of understanding why employees leave the company, the most relevant variables are likely to be:

satisfaction_level: Lower satisfaction might be directly correlated with higher turnover.
last_evaluation: Performance reviews might affect an employee's decision to stay or leave.
number_project: Both overloading and underloading employees with projects can affect their job satisfaction and decision to leave.
average_monthly_hours: Similar to the number of projects, both too many and too few hours can lead to dissatisfaction.
time_spend_company: Length of service could indicate loyalty or, conversely, a desire for change.
Work_accident: Experiencing a work-related accident might influence an employee's decision to leave.
promotion_last_5years: Lack of promotion could be a significant factor in an employee's decision to leave.
salary: Compensation is a critical factor in employee retention.

- What units are your variables in?
  Mixed variables of variables units like $, years etc.
- 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?
  There are many outliers, we need to see if those outliers are meaningful or wrong inputs.

- Is there any missing or incomplete data?

  no.
- Are all pieces of this dataset in the same format?

  Some are integers some are floats some are strings.
- Which EDA practices will be required to begin this project?

  Data exploration, analysis, correlation analysis and more.

**The Power of Statistics**

- What is the main purpose of this project?

  To figure out what makes employees leave the company and make a model that can predict it.
- What is your research question for this project?

  What features contribute the most to making an employee leave 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?

  If we only sample people that didnt leave the company then the model is gonna be really bad at predicting the ones who will leave.

**Regression Analysis: Simplify Complex Data Relationships**

- Who are your stakeholders for this project?

  Salifort Motors stakeholders
- What are you trying to solve or accomplish?

  Im trying to create a model that predicts which employees will churn or not, this will allow the business to know what is making the employees leave.
- What are your initial observations when you explore the data?

  There are some outliers in some of the features.
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)

  Past course modules.

- Do you have any ethical considerations in this stage?

  Make sure not to invade any privacy.

**The Nuts and Bolts of Machine Learning**

- What am I trying to solve?

  Create a model that predicts the employee's turnover.
- What resources do you find yourself using as you complete this stage?

  Past course modules.
- Is my data reliable?

  It is.
- Do you have any additional ethical considerations in this stage?

  Make sure the model isnt making any biased predictions.
- What data do I need/would I like to see in a perfect world to answer this question?

  A data set thats fully representative of the population and a model of 100% accuracy.
- What data do I have/can I get?

  I have almost 15k data entries. Some are duplicates though.
- What metric should I use to evaluate success of my business objective? Why?

  The normal classification metrics like f1, accuracy,recall,precision,roc/auc.

## 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.

### 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?

  Make sure data is clean, visualize, and analyse the data.

- 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 dont need to add more data by joining, Ill probably have to filter for some attributes when I analyze univariately.

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

  Histograms and bar graphs.

### The Power of Statistics

- Why are descriptive statistics useful?

  Descriptive statistics are useful because they provide concise summaries and insights into the central tendencies, dispersion, and distribution of data, facilitating easier interpretation and understanding of large datasets.

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

  The null hypothesis is a statement that assumes no significant effect or difference exists in a given situation, implying that any observed variations are random or due to chance. In contrast, the alternative hypothesis suggests that there is a meaningful effect, difference, or relationship present that is not attributable to chance alone.

**Regression Analysis: Simplify Complex Data Relationships**

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

  EDA helps identify trends, outliers, and the relationships between variables to ensure the assumptions of multiple linear regression are met and to guide informed feature selection and transformation.

- Do you have any ethical considerations in this stage?

  None.

**The Nuts and Bolts of Machine Learning**

- What am I trying to solve? Does it still work? Does the plan need revising?
  Create a model that predicts the employee's turnover.
- Does the data break the assumptions of the model? Is that ok, or unacceptable?
  Im going to be using a random forest model and its robust to outliers and its assumptions arent broken.
- Why did you select the X variables you did?

  Average monthly hours, tenure, last evaluation all seem to have played a great part in whether an employee leaves or not.

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

  Get a good idea about which are the important features and whether or not the data follows the assumptions of the model.

- What has the EDA told you?

  Employees seem to be exiting the company due to inadequate management, with departures associated with extended work hours, excessive projects, and overall reduced satisfaction.

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

  Seaborn and matplotlib documentation.

- Do you have any ethical considerations in this stage?

  If the model classifies an employee as not leaving while he is leaving, this might cost the company a great deal.

## Data Project Questions & Considerations

### PACE: Construct Stage

**Get Started with Python**

- Do any data variables averages look unusual?

  No.
- How many vendors, organizations or groupings are included in this total data?

  Employees can be grouped by different attributes, like salary levels for example.

**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?

  Randomforests and hyperparameter tuning them, scatterplots, roc graphs.
- What processes need to be performed in order to build the necessary data visualizations?

  Some were built in the eda stage, roc/auc are built after building the model.
- Which variables are most applicable for the visualizations in this data project?

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

  No missing are present.

**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?

  There might be a chance of data leakage.
- Can you improve it? Is there anything you would change about the model?

  I could use feature engineering to figure out the issue.

**The Nuts and Bolts of Machine Learning**

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

  I like to look at my models with a skeptic eye, I might need to do more feature engineering before modeling.
- Which independent variables did you choose for the model, and why?

  Random forests choose most of the variables on its own and ranks the feature importance.
- How well does your model fit the data? (What is my model's validation score?)

  The Random Forest model, after cross-validation, demonstrates strong performance across multiple metrics. With a precision of 94.62%, it shows high accuracy in identifying true positives out of all predicted positives. The recall of 91.83% indicates the model's effectiveness in capturing a significant portion of actual positives. An F1 score of 93.20% suggests a balanced harmony between precision and recall, emphasizing the model's reliability. An overall accuracy of 97.78% demonstrates the model's capability to correctly classify both positive and negative cases. Finally, an AUC score of 98.13% highlights the model's excellent ability to discriminate between the classes, further validating its robustness in prediction.
- Can you improve it? Is there anything you would change about the model?

  I would divide the employees to two groups, overworked and under worked, and see how the model interacts with these two categories.
- Do you have any ethical considerations in this stage?

  Not exactly sure how well this model will work on new real world data.

## 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 the patterns of turnover across different departments and roles within the company to understand if specific areas are more affected than others.

  Look into the historical trends of employee turnover to see if there are specific times of the year when turnover peaks, which could indicate cyclical factors at play.

- What data initially presents as containing anomalies?

  Anomalies could be present in the form of employees with extremely high or low values in 'average_monthly_hours', 'number_project', or 'last_evaluation'. These could indicate data entry errors or genuinely unique cases worth exploring further.

- What additional types of data could strengthen this dataset?

  External benchmarking data on industry turnover rates and practices could also provide context for the company's turnover rates.

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

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

  Key insights could include correlations between high turnover and specific factors such as low job satisfaction, lack of promotions, high work pressure (evident from high average monthly hours or number of projects), and low compensation.

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

  Recommendations include implementing targeted retention strategies for high-risk groups, such as career development opportunities for those with high performance but low satisfaction, or workload adjustments for those with high average monthly hours.

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

  What are the common characteristics of employees who choose to stay with the company long-term?

- How might you share these visualizations with different audiences?

For executive stakeholders, summarize key insights and recommendations in a high-level dashboard with the option to drill down into specific data points.

For managers and team leaders, provide detailed reports and visualizations focused on their specific teams, highlighting areas of concern and actionable insights.

**The Power of Statistics**

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

  None was made.

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

  N/A

**Regression Analysis: Simplify Complex Data Relationships**

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

  Interpreting the beta coefficients helps understand the direction and magnitude of the impact that each predictor variable has on the response variable. This insight is crucial for identifying the most influential factors contributing to employee turnover and for making informed decisions to address those factors.

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

  I would recommend implementing targeted interventions aimed at improving job satisfaction and managing workload, as these are the most significant predictors of churn according to the model. Additionally, setting up a system to regularly collect and analyze employee feedback could preemptively identify dissatisfaction and prevent potential turnover.

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

  The model could potentially be improved by exploring additional features, employing different encoding techniques for categorical variables, or trying alternative machine learning algorithms for comparison. Regular updates and retraining with new data will also help maintain its relevance and accuracy over time.

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

  Stakeholders should focus on the factors identified by the model as the most predictive of churn, like employee satisfaction levels and workload, and develop targeted retention strategies to address these issues.

  Recommendations include implementing career development opportunities for those with high performance but low satisfaction, or workload adjustments for those with high average monthly hours.

  Ensure evaluations are fair and constructive, as their influence is significant on employee decisions to stay or leave.

Set a cap on the number of projects an employee can take at a time.

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

  The model's insights indicate that employee satisfaction, the number of projects, and the last performance evaluation are pivotal in predicting turnover. This suggests that how employees perceive their role and their experiences at the company directly influence their decision to stay or leave.

- Do you have any ethical considerations at this stage?

  Its important to use the model's predictions responsibly, focusing on supporting employees rather than penalizing them based on the model's output.

**The Nuts and Bolts of Machine Learning**

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

  The model's insights indicate that employee satisfaction, the number of projects, and the last performance evaluation are pivotal in predicting turnover. This suggests that how employees perceive their role and their experiences at the company directly influence their decision to stay or leave.

- What are the criteria for model selection?

  The model was selected based on its performance metrics (precision, recall, F1 score, accuracy, AUC) and its ability to handle the nuances of the dataset, such as non-linear relationships and interactions between features.

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

  The final results are acceptable given the high performance metrics, but continual monitoring is essential to ensure the model remains fair and accurate as the company and workforce evolve.

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

  According to the Random Forest model, the most important factor was satisfaction level, with the number of projects, last evaluation, tenure, and average monthly hours also being important, while the rest of the features were less significant.

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

  Further questions could involve exploring the impact of external factors on turnover, such as market trends and the competitive landscape, or how internal policies and management practices influence employee satisfaction and retention.

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

  Resources likely included Python libraries such as Scikit-learn for model building, Pandas and NumPy for data manipulation, and Matplotlib and Seaborn for visualization.

- Is my model ethical?

  The model appears ethical in its current state, but continuous ethics reviews should be conducted, especially as new data is included and the model evolves.

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

  When the model makes mistakes, it could be due to anomalies, unexpected interactions between features, or shifts in the underlying data distribution. Understanding why these mistakes occur is crucial for refining the model and ensuring it remains useful and fair in practical applications.