**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**

**Employee Turnover Prediction for Salifort Motors project proposal**

## **Overview**

*This project aims to develop a machine learning model to predict employee turnover at Salifort Motors. The goal is to identify employees who are likely to leave the company so that preventive measures can be taken to improve employee retention.*

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Tasks** | **PACE stages** |
| **Data Collection** | **Gather data from HR systems** | **Plan** |
| **Data Cleaning** | **Clean and preprocess data** | **Analyze** |
| **Exploratory Data Analysis (EDA)** | **Perform EDA to understand data distribution and relationships** | **Analyze** |
| **Feature Engineering** | **Engineer features based on domain knowledge** | **Construct** |
| **Model Selection** | **Select appropriate machine learning models** | **Construct** |
| **Model Training** | **Train models using historical data** | **Construct** |
| **Model Evaluation** | **Evaluate model performance using validation metrics** | **Execute** |
| **Model Deployment** | **Deploy the model in a Flask web application** | **Execute** |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

**Foundations of data science**

* Who is your audience for this project?

The audience includes HR managers, executives at Salifort Motors, and data analysts.

* 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?

The project aims to predict employee turnover to help the company take proactive measures to retain valuable employees.

* What questions need to be asked or answered?

Which employees are at risk of leaving?

What factors contribute most to employee turnover?

* What resources are required to complete this project?

Historical HR data, Python libraries for data analysis and machine learning (Pandas, NumPy, scikit-learn), SHAP for model interpretability, Flask for building the web application, Alibi Detect for drift detection, Online documentation, ChatGPT and forums for troubleshooting.

* What are the deliverables that will need to be created over the course of this project?  
  Data analysis report  
  Machine learning model  
  Web application for predictions

**Get Started with Python**

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

Review the data schema and understand the meaning of each feature.

Clean and preprocess the data to ensure consistency and completeness.

* What follow-along and self-review codebooks will help you perform this work?

Python documentation

Machine learning tutorials

Data science courses

ChatGPT for a lot of questions!!!

* What are a couple additional activities a resourceful learner would perform before starting to code?  
  Explore the data using visualization tools, research domain-specific factors that influence employee turnover

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

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

Relevant variables include satisfaction level, number of projects, years at the company, average monthly hours, last evaluation score.

* What units are your variables in?  
  Variables are in different units such as hours for average monthly hours, percentage for satisfaction level, and years for years at the company
* 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?  
  Higher satisfaction levels correlate with lower turnover.

Employees with extreme workloads (high or low) are more likely to leave.

* Is there any missing or incomplete data?

It does not seems to be missing or incomplete data.

* Are all pieces of this dataset in the same format?  
  No, there are categorical and continuous data.
* Which EDA practices will be required to begin this project?  
  Descriptive statistics, correlation analysis, visualization of data distributions.

**The Power of Statistics**

* What is the main purpose of this project?   
  To predict employee turnover and identify factors contributing to it.
* What is your research question for this project?  
  What factors are most predictive of employee turnover at Salifort Motors?
* 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 ensures that the model generalizes well. Without it, the model may be biased towards certain departments or job roles.

**Regression Analysis: Simplify Complex Data Relationships**

* Who are your stakeholders for this project?  
  HR managers, executives, and data analysts at Salifort Motors.
* What are you trying to solve or accomplish?  
  Predicting employee turnover to help the company implement retention strategies.
* What are your initial observations when you explore the data?  
  Satisfaction level and number of projects are key factors influencing turnover.
* What resources do you find yourself using as you complete this stage? (Make sure to include the links.)  
  Python libraries, online tutorials, data science forums, chatGPT.
* Do you have any ethical considerations in this stage?  
  Ensure the model does not discriminate against any group of employees.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?  
  Predicting employee turnover.
* What resources do you find yourself using as you complete this stage?   
  Python libraries, online documentation, machine learning courses, chatGPT.
* Is my data reliable?  
  The data seems reliable. There are no indications elsewise.
* Do you have any additional ethical considerations in this stage?  
  Ensure fairness and transparency in model predictions.
* What data do I need/would I like to see in a perfect world to answer this question?  
  Detailed employee profiles, feedback surveys, exit interview data, who left or was fired.
* What data do I have/can I get?  
  Historical HR data including employee satisfaction, workload, and tenure.
* What metric should I use to evaluate success of my business objective? Why?  
  Accuracy, precision, recall, F1 score, and AUC for evaluating model performance.

**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 available information is sufficient for initial analysis.

**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?  
  Clean data, visualize distributions, perform correlation analysis, and identify key variables.
* 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.?  
  Joining additional data sources, such as employee feedback, could enhance the analysis. Data structuring includes filtering out irrelevant records and sorting by relevant features.
* What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?  
  Histograms, box plots, correlation matrices, and bar charts for categorical data.

**The Power of Statistics**

* Why are descriptive statistics useful?  
  They provide a summary of the data and help identify patterns, trends, and anomalies.
* What is the difference between the null hypothesis and the alternative hypothesis?  
  Null hypothesis: There is no effect or relationship. Alternative hypothesis: There is an effect or relationship.

**Regression Analysis: Simplify Complex Data Relationships**

* What are some purposes of EDA before constructing a multiple linear regression model?  
  To understand the data, check assumptions, identify relationships, and detect outliers.
* Do you have any ethical considerations in this stage?  
  Ensuring that data manipulation does not introduce bias or ethical issues.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve? Does it still work? Does the plan need revising?  
  Predicting employee turnover. The plan is still valid and does not need revising at this stage.
* Does the data break the assumptions of the model? Is that ok, or unacceptable?  
  Initial checks suggest that data does not break model assumptions.
* Why did you select the X variables you did?  
  Based on their relevance and correlation with the target variable.
* What are some purposes of EDA before constructing a model?  
  To understand data distribution, identify key features, and check for outliers.
* What has the EDA told you?  
  Satisfaction level and workload are significant factors affecting turnover.
* What resources do you find yourself using as you complete this stage?  
  Python libraries, online tutorials, data science forums.
* Do you have any ethical considerations in this stage?  
  Ensuring fairness and avoiding bias in the model.

**Data Project Questions & Considerations**

**PACE: Construct Stage**

**Get Started with Python**

* Do any data variables averages look unusual?  
  No unusual averages detected.
* How many vendors, organizations or groupings are included in this total data?  
  Data includes various departments within Salifort Motors (Marketing, sales, management, etc).

**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?   
  Visualizations include histograms, bar charts, and ROC curves. Algorithms include Random Forest, XGBoost, and Logistic Regression.
* What processes need to be performed in order to build the necessary data visualizations?   
  Data cleaning, feature engineering, and model training.
* Which variables are most applicable for the visualizations in this data project?   
  Satisfaction level, number of projects, years at the company, average monthly hours, last evaluation score.
* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?  
  Imputation or removal based on the amount and importance of missing data.

**The Power of Statistics**

* How did you formulate your null hypothesis and alternative hypothesis?  
  Null hypothesis: Satisfaction level does not affect employee turnover. Alternative hypothesis: Satisfaction level affects employee turnover.
* What conclusion can be drawn from the hypothesis test?  
  Satisfaction level significantly affects employee turnover.

**Regression Analysis: Simplify Complex Data Relationships**

* Do you notice anything odd?  
  No significant oddities observed.
* Can you improve it? Is there anything you would change about the model?  
  Hyperparameter tuning and adding more interaction terms could improve the model.

**The Nuts and Bolts of Machine Learning**

* Is there a problem? Can it be fixed? If so, how?  
  No major problems detected.
* Which independent variables did you choose for the model, and why?  
  Chose variables based on their correlation with turnover and business relevance.
* How well does your model fit the data? (What is my model’s validation score?)  
  The model fits well with high validation scores (accuracy, precision, recall, F1, AUC).
* Can you improve it? Is there anything you would change about the model?  
  Further tuning and exploring additional features could improve the model.
* Do you have any ethical considerations in this stage?  
  Ensure the model is fair and does not introduce bias.

**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 satisfaction levels across different departments and project loads.
* What data initially presents as containing anomalies?  
  Outliers in average monthly hours and last evaluation scores.
* What additional types of data could strengthen this dataset?  
  Employee feedback and exit interview data. If the person was fired or left.

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

* What key insights emerged from your EDA and visualizations(s)?  
  Satisfaction level and workload are key factors influencing turnover.
* What business recommendations do you propose based on the visualization(s) built?  
  Focus on improving employee satisfaction and managing workloads to reduce turnover.
* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?   
  Explore the impact of management practices on turnover.
* How might you share these visualizations with different audiences?  
  Through interactive dashboards and detailed reports.

**The Power of Statistics**

* What key business insight(s) emerged from your A/B test?  
  Insights were not derived from an A/B test in this case.
* What business recommendations do you propose based on your results?  
  Implement targeted interventions to improve satisfaction and retention.

**Regression Analysis: Simplify Complex Data Relationships**

* To interpret model results, why is it important to interpret the beta coefficients?  
  To understand the impact of each variable on turnover.
* What potential recommendations would you make to your manager/company?  
  Focus on employee satisfaction and workload management.
* Do you think your model could be improved? Why or why not? How?  
  Yes, through further tuning and incorporating additional features.
* What business recommendations do you propose based on the models built?  
  Implement strategies to improve satisfaction and retention.
* What key insights emerged from your model(s)?  
  Key factors influencing turnover include satisfaction level and workload.
* Do you have any ethical considerations at this stage?  
  Ensure fairness and transparency in model predictions.

**The Nuts and Bolts of Machine Learning**

* What key insights emerged from your model(s)?  
  Satisfaction level and workload are significant predictors of turnover.
* What are the criteria for model selection?  
  Accuracy, precision, recall, F1 score, AUC.
* Does my model make sense? Are my final results acceptable?  
  Yes, the model is logical and the results are acceptable.
* Were there any features that were not important at all? What if you take them out?  
  Less important features could be removed to simplify the model.
* Given what you know about the data and the models you were using, what other questions could you address for the team?  
  Investigate department-specific turnover patterns.
* What resources do you find yourself using as you complete this stage?  
  Python libraries, online tutorials, data science forums, chatGPT questions.
* Is my model ethical?  
  Yes, provided it is used responsibly and transparently.
* When my model makes a mistake, what is happening? How does that translate to my use case?

Mistakes could indicate areas where the model needs improvement. Understanding these can help refine the model and its use in predicting turnover.