**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 organise information through custom functions
* Demonstrate understanding of how to organise and analyse a dataset to find the “story”
* Create a Jupyter notebook for exploratory data analysis (EDA)
* Create visualisation(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**

**[Title] project proposal**

## **Overview**

*This project addresses Salifort's employee turnover issue, which involves both voluntary departures and terminations, with a focus on the 'left' y-value. The project aims to understand the key drivers behind this attrition problem, which has raised concerns among senior leadership, affecting corporate culture, employee development, and financial costs. Given Salifort's substantial investments in employee recruitment, training, and skill enhancement, it is crucial to identify the underlying factors. By analysing data related to job satisfaction, tenure, performance evaluations, project involvement, and monthly work hours, this analysis employed Random Forest modelling to identify the top 10 determinants of employee turnover. Job satisfaction emerged as the most significant influence, followed by tenure, performance evaluations, project engagement, and monthly work hours. To gain deeper insights into job satisfaction, a comprehensive survey involving current and former employees was initiated. The analysis highlighted the superior predictive performance of Random Forest cross-validation compared to the Logistic Regression model, with moderate enhancements over the Decision Tree cross-validation. These findings empower Salifort to proactively address the issue by elevating job satisfaction, fostering data-informed decision-making, and ultimately improving corporate culture while curtailing significant financial outlays.*

| **Milestones** | **Tasks** | **PACE stages** |
| --- | --- | --- |
| **1** | **Data exploration** | **Plan** |
| **2** | **Workable data manipulation** | **Plan** |
| **3** | **Data visualisations - Visualise scope** | **Analyse** |
| **4** | **Confusion matrix construction** | **Construct** |
| **5** | **Logistic regression construction** | **Construct** |
| **6** | **Decision tree and Random Forest construction** | **Construct** |
| **7** | **Analyse statistics and compare against each model for top performer** | **Construct** |
| **8** | **Visualise high association columns for individuals who left the company** | **Construct** |
| **9** | **Summary of findings and potential changes the company can take to reduce the rate of employees leaving.** | **Execute** |

**Data Project Questions & Considerations**

**PACE: Plan Stage**

**Foundations of data science**

* Who is your audience for this project?
  + HR project managers and organisation 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?
  + Provide features that indicate potential scenarios of employees leaving. Potential impact is developing incentives and guidelines to develop employees growth and retention rate.
* What questions need to be asked or answered?
  + What are the key features which signal employees to leave.
* What resources are required to complete this project?
  + The resources used were the course content and Stackoverflow for suggestions on manipulation of data.
* What are the deliverables that will need to be created over the course of this project?
  + Graphs which enable understanding of key trends associated with employee retention. Solutions and recommendations for the current solutions as well as other opportunities to gather more data.

**Get Started with Python**

* How can you best prepare to understand and organise the provided information?
  + This can be done by installing main packages which will be utilised for manipulation and visualisation. Depending on the Analysis phase, the models for prediction modelling and assessment could change and so installing of relevant packages can be done after assessing the situation.
* What follow-along and self-review codebooks will help you perform this work?
  + The follow along notebooks provided by the Google Advanced Data Analytics Professional Certificate is beneficial to mty learning and a good source of material to refer back to for the basics.
* What are a couple additional activities a resourceful learner would perform before starting to code?
  + This would be filling a document and documenting processes that will be taken and potential materials/ support items that may be necessary

**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 , last\_evaluation, number\_project, average\_monthly\_hours, tenure, work\_accident, promotion\_last\_5years, salary are part of the X.columns. The Y axis would be ‘left’.
* What units are your variables in?
  + Float64, int64 and string64.
* 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?
  + Just by looking at the scenario, Logistic regression, decision tree and random forest could be used due to the binary classification of Left or Stayed.
* Is there any missing or incomplete data?
  + There is some missing data and will have to be removed.
* Which EDA practices will be required to begin this project?
  + Exploring the data types and the head of the data.

**The Power of Statistics**

* What is the main purpose of this project?
  + To determine the key factors associated with leaving employees.
* What is your research question for this project?
  + Predict employees likely to quit, it might be possible to identify factors that contribute to their leaving?
* 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?
  + This could lead to sampling bias, A scenario of biassed bias could be cherry picking employees from certain departments.

**Regression Analysis: Simplify Complex Data Relationships**

* Who are your stakeholders for this project?
  + HR project managers and organisation stakeholders and data science teams.
* What are you trying to solve or accomplish?
  + Predict employees likely to quit, it might be possible to identify factors that contribute to their leaving.
* What are your initial observations when you explore the data?
  + A lot of manipulation of the data to fit into certain data types and removal of missing values.
* Do you have any ethical considerations at this stage?
  + No ethical considerations at this stage.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?
  + Understanding the features which are associated with employees leaving and testing and prediction modelling to determine the accuracy of the models.
* What resources do you find yourself using as you complete this stage?
  + The follow along notebooks provided by the Google Advanced Data Analytics Professional Certificate is beneficial to mty learning and a good source of material to refer back to for the basics.
* Is my data reliable?
  + With the same database, there has been Random\_state configuration for reproducibility.
* Do you have any additional ethical considerations in this stage?
  + None at this present stage due to randomness ensured.

**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?
  + The available information is sufficient to develop an understanding of how independent categories can influence the Y-variable ‘left’.

**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?
  + First is data cleaning, to ensure that the data is in the manageable format, efficient for data manipulation.
  + Next is being able to do data exploration. This will be able to visualise the contents of the data and deduce what the data is showing.
* What initial assumptions do you have about the types of visualisations that might best be suited for the intended audience?
  + The use of box plots allow for the understanding of distributions of each categorical vairable such as wage.
  + There is a key observation of the amount of average monthly hours worked with the satisfaction level. This visualisation will aid in visualising the tight distribution of employees who had left vs who stayed with the difference in colour.

**The Power of Statistics**

* Why are descriptive statistics useful?
  + The use of descriptive statistics aids in developing key understanding of the distribution, measures of central tendency and variability of the data to ensure that the data is well represented and not skewed when taking in the full range of contextual data.
* What is the difference between the null hypothesis and the alternative hypothesis?
  + The difference between the null and alternative hypothesis is that the Null hypothesis usually supports a non change environment between determined variables whilst the Alternative hypothesis supports a change in the variables indicated that there is a correlation.

**Regression Analysis: Simplify Complex Data Relationships**

* What are some purposes of EDA before constructing a multiple linear regression model?
  + EDA helps in understanding the dataset's structure, variables, and relationships between them.
  + Identifying outliers is important as they can significantly impact regression coefficients.
* Do you have any ethical considerations at this stage?
  + No ethical considerations at this stage.

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve? Does it still work? Does the plan need revising?
  + To determine the key factors associated with leaving employees.
* Does the data break the assumptions of the model? Is that ok, or unacceptable?
  + Data does not break any assumptions as it coincides with the model. The groups are not equal ‘left’ vs ‘stayed’ but the skewness is not drastic to implement upsampling/downsampling.
* Why did you select the X variables you did?
  + The X variables were chosen on the basis of the effect these variables had on the y-axis determined by the Analysis phase of PACE.

**Data Project Questions & Considerations**

**PACE: Construct Stage**

**Get Started with Python**

* Do any data variables averages look unusual?
  + The averages do not look different as the anomalies have been ensured to be removed through the box plot - upper and lower limits method.
* How many vendors, organisations or groupings are included in this total data?
  + HR project managers and organisation stakeholders and data science teams.

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

* What data visualisations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?
  + There are 2 models which could aid in this task being Logistic regression or Decision trees. I will be performing and compiling both models in order to develop my skills and for a point comparison against the appropriate metrics.
* What processes need to be performed in order to build the necessary data visualisations?
  + Processes include assigning X and y variables, placing the data in a train test split in order to train the data on the needed datasets.
* Which variables are most applicable for the visualisations in this data project?
  + Satisfaction\_leve, last\_evaluation, number\_project, average\_monthly\_hours, tenure, work\_accident, promotion\_last\_5years, salary, department\_IT, department\_RandD, department\_accounting, department\_hr, department\_management, department\_marketing department\_product\_mng, department\_sales, department\_support, department\_technical.
* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?
  + Removal of nulls through use of functions. Resting index followed.

**Regression Analysis: Simplify Complex Data Relationships**

* Do you notice anything odd?
  + I would have originally stuck with the max\_iter = 100 but this caused an attribute error. To correct this max\_iter = 500.
* Can you improve it? Is there anything you would change about the model?
  + Although the data is not equal, I would upsample the left variable as it is a low ratio.

**The Nuts and Bolts of Machine Learning**

* Is there a problem? Can it be fixed? If so, how?
  + I changed the max\_iter = 100 to max\_iter = 500 through various troubleshooting.
* Which independent variables did you choose for the model, and why?
  + Satisfaction\_leve, last\_evaluation, number\_project, average\_monthly\_hours, tenure, work\_accident, promotion\_last\_5years, salary, department\_IT, department\_RandD, department\_accounting, department\_hr, department\_management, department\_marketing department\_product\_mng, department\_sales, department\_support, department\_technical. This is because the Confusion matrix all reported a low level correlation which supports independence.
* How well does your model fit the data? (What is my model’s validation score?)
  + The logistic regression model achieved precision of 79%, recall of 83%, f1-score of 80% (all weighted averages), and accuracy of 83%, on the test set.
  + The decision tree model achieved AUC of 97.2%, precision of 97.4%, recall of 91.5%, f1-score of 94.4%, and accuracy of 98.2%, on the test set. The random forest modestly outperformed the decision tree model. The high statistics could indicate overfitting of the data.
* Can you improve it? Is there anything you would change about the model?
  + I could utilise a higher CV value for increased folds. This would run less risk of overfitting. The issue would be then increased system memory and computing power leading to increased time spent on calculation.
* Do you have any ethical considerations at this stage?
  + None 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?
  + This data that was gathered captured from which period of time and at which point was the data updated. Confirm if the data is 1st party for context and confirmation purposes.
* What data initially presents as containing anomalies?
  + Tenure contained a large amount of anomalies which was detected on the Analysis phase through the box plot method. This was corrected by excluding data that was below the lower\_limit and greater than the upper\_limit.
* What additional types of data could strengthen this dataset?
  + Future Questions could be written in a survey format in order to gain data directly from the employees and also to provide monitoring of current and prediction of future scenarios. If this survey feature was implemented, anonymity will be ensured to prevent current and past employee discrimination for ethical standards

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

* What key insights emerged from your EDA and visualisations(s)?
  + There is a spike in employees that has left the organisation with projects at 2, which decreases dramatically at 3 number of projects but steadily increases to 6 with 7 projects containing only employees which had left. This depicts that the lowest number of projects could suggest that the difficulty had been substantial or the employees felt undervalued through lack of projects. 7-projects containing only past employees suggest that the employees may be overworked or satisfied with the workload.
* What business recommendations do you propose based on the visualisation(s) built?
  + recommendations could include ensuring that there is a hard cap for employees in the number of projects that can be undertaken at any given time. Last evaluation metrics need to be looked into with more data. With the given data provided, it could be considered that the feedback given is not well received or constructive feedback needs to be given to help drive morale within the workplace and to develop a healthy workplace where directions are given to help strive for better performance.
  + Another indication from the Analysis showed that promotion is scarcely provided and that a system could be developed to help nurture individuals with a proven track record for promotion into various sectors.
* Given what you know about the data and the visualisations you were using, what other questions could you research for the team?
  + Following the recommendations, Data given in the form of a survey will be the best representatives of employees both past and present for comments of the work culture and possible implementations. This will also display value for the co workers and develop a connection with workers and the company.
* How might you share these visualisations with different audiences?
  + Depending o n the audience, the use of descriptive statistics could be restricted in order to prevent confusion and to ensure each audience is on the same page. This is not to take away from the presentation but visualisations should be selected purposely without holding information from the audience.

**Regression Analysis: Simplify Complex Data Relationships**

* To interpret model results, why is it important to interpret the beta coefficients?
  + It is important to interpret beta coefficients as this will determine the unit of change of the independent variables. In this case this would be utilised for Logistic Regression.
* What potential recommendations would you make to your manager/company?
  + The key theme is for the company to collect more data on employee feedback in the form of an anonymous survey. The survey could include areas of potential interest derived from understanding the insights that the decision tree derived. This could be the satisfaction level and how the company deals with lower performing satisfaction levels.
* Do you think your model could be improved? Why or why not? How?
  + The model could be improved by including more columns that showed overworked employees. If the number of average monthly hours was divided into weeks, then cross referenced against the government concern of ‘Overworked’ (subjective to country government perspective), then this new column could derive insights on employee behaviour and the standard of the company.
* What business recommendations do you propose based on the models built?
  + The main topic would be support for the employees to achieve a higher satisfaction level, which could be derived from the lack of promotion or training for employee career development. After each performance review, the employee could be given personalised feedback on their performance as well as support for career development to increase employee morale and loyalty to the company.
* What key insights emerged from your model(s)?
  + The key insights can be seen through the 'Random Forest: Top 10 Feature Importances for Employee Leaving', that the graph indicates Satisfaction taking the lead in importance followed by tenure, last evaluation, number\_project and monthly hours. It appears that the Satisfaction rating needs to be closely cared for within the workplace and a thorough report with answers from a survey (from past and current employees) could indicate key areas of development to improve satisfaction rating.
* Do you have any ethical considerations at this stage?
  + None at this stage.

**The Nuts and Bolts of Machine Learning**

* What key insights emerged from your model(s)?
  + Key insights suggested that the Random forest produced the best performing model when compared to the Logistic regression and only modestly outperformed the Decision tree model. The key insights can be seen through the 'Random Forest: Top 10 Feature Importances for Employee Leaving', that the graph indicates Satisfaction taking the lead in importance followed by tenure, last evaluation, number\_project and monthly hours.
* What are the criteria for model selection?
  + The criteria was the Auc score as well as the Accuracy got fitting of the model.
* Does my model make sense? Are my final results acceptable?
  + The final results are acceptable but there is a case of overfitting of the data to the model. A way to solve this for the Random forest is to use another dataset for validation requiring two sets of Train\_test\_split. This will also have to be stratified = y in order to maintain the ratio.
* Were there any features that were not important at all? What if you take them out?
  + There were no inherent features that were dropped, there were some duplicates within the data as evident by exploring the data, The duplicates were dropped with appropriate coding.
* Given what you know about the data and the models you were using, what other questions could you address for the team?
  + Questions could include further prediction methods to determine the accuracy of the data that does not include decision trees or random forest. Another question could be the next direction for the company HR team and the possibility of developing an A/B testing style for further analysis of changes.
* What resources do you find yourself using as you complete this stage?
  + Resources were utilised from the Google data analytics professional course for recounting knowledge on coding skills and possible directions based on the question and format of the results for the given question.
* Is my model ethical?
  + This method can be considered ethical due to no explicit choosing of data and promoting randomness.
* When my model makes a mistake, what is happening? How does that translate to my use case?
  + If my model makes a mistake then this would lead to the validity of the data derived from the model being skewed. This would lead to misinterpretation of the past and present workforce. If this was the case then the predictions and solutions to all questions put forward will be inaccurate and could lead to further issues with HR management.