**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

Predicts Salifort Motors employee churn

**Project proposal**

**Overview**

The Salifort leadership team asks you to analyze the survey data and create a model which predict if employee will resign or not.

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Tasks** | **PACE stages** |
| **1** | **Establish structure for project workflow (PACE)** | **Plan** |
| **2** | **Write a project proposal** | **Plan** |
| **3** | **Conduct explanatory data analysis (EDA)** | **Analysis** |
| **4** | **Build a regression model (Logistic Regression)** | **Construct** |
| **5** | **Build machine learning model (Decision Tree/Random Forest/XGBoost)** | **Construct** |
| **6** | **Evaluate model and choose “champion” model** | **Construct** |
| **7** | **Prepare required visualizations for final insights** | **Execute** |
| **8** | **Communicate final insights with stakeholders** | **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?
  + We are trying to build model (regression or machine learning) which will predict if Salifort Motors’ employees will leave the company or not
  + The result of this work will allow the Salifort Motors stakeholders try to improve situation and come up with a solution to reduce the level of employees leaving the company
* What questions need to be asked or answered?
  + What’s likely to make the employee leave the company?
  + What reasons does employee have to leave the company?
  + What data can help us to reveal employees intention to leave?
  + Does we have sufficient amount of data to reach our goal?
  + Does the dataset is reliable and sufficient?
* What resources are required to complete this project?
  + We need results of the survey, which contain information related to the task at hand
  + We need to check how this data was collected, who is responsible for it and if it is reliable
  + Do we have possibility to obtain additional data if required or clarify any data related issues if required?
* What are the deliverables that will need to be created over the course of this project?
  + Project proposal
  + Reggression model and/or one of the Machine learning model
  + Final executive report for Salifort Motors stakeholders with our recommendations how to improve the current situation

**Get Started with Python**

* How can you best prepare to understand and organize the provided information?
  + We definitely would like to create DataFrame from dataset
  + Investigate the dataset dictionary in order to understand which information is available for us for this task
* What follow-along and self-review codebooks will help you perform this work?
  + EDA steps
  + Regression model (assumptions, building, evaluation)
  + Machine learning (best fit model for data at hand, assumptions, choosing right parameters, model building, evaluation)
* What are a couple additional activities a resourceful learner would perform before starting to code?
  + Check requirement for data and which best tools will be more suitable for such kind of data
  + Check what kind of data we are dealing with (first-hand, second-hand…)

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

* What are the data columns and variables and which ones are most relevant to your deliverable?
  + All data seems really relevant, although “promotion\_last\_5years” can be really insignificant for model, as it is very imbalanced and not presented greatly in the dataset to make understandable pattern if any
* What units are your variables in?
  + satisfaction\_level: percentage [0-1]
  + last\_evaluation: percentage [0-1]
  + number\_project: integer (>0)
  + average\_montly\_hours: integer (>0)
  + time\_spend\_company (years): integer (>0)
  + Work\_accident: binary [0/1]
  + left: binary [0/1]
  + promotion\_last\_5years: binary [0/1]
  + Department: string, 10 different departments ['sales', 'accounting', 'hr', 'technical', 'support', 'management', 'IT', 'product\_mng', 'marketing', 'RandD']
  + salary: string, 3 different levels (category) ['low', 'medium', 'high']
* 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 data contained 83% row for employee how stayed and 17% who left
  + There is no missing data, although in one column there is potential outliers (late it was confirmed)
  + The outliers is just in “years in company” column
* Is there any missing or incomplete data?
  + There is no missing data in the dataset
* Are all pieces of this dataset in the same format?
  + Two columns with percentage data
  + Three columns with different values (int)
  + Tree columns with binary data (0/1)
  + Two columns with categorical data (one with two classes, one with tree classes)
* Which EDA practices will be required to begin this project?
  + Descriptive statistic

**The Power of Statistics**

* What is the main purpose of this project?
  + Create a model which predict if employee will resign or not and make it possible to come with solution in order to improve employee satisfaction levels at the company
* What is your research question for this project?
  + What is the relationship between variables?
  + Are there any significant correlation between independent variables and dependent variable?
  + Is any significance increasing level of leaving the company from department perspective?
  + Is any significance increasing level of leaving the company from salary level perspective?
  + Is any correlation between having (not having) promotion and leaving the company decision?
* 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 make it sure that we non implement bias and increase chance to cover all different group from population
  + It can lead to undecover bias
  + If we get data for example just from one department only, and create model predicting if the employee will quit or not this model will not be sufficient for the whole company as it won’t contain all specific features of the population

**Regression Analysis: Simplify Complex Data Relationships**

* Who are your stakeholders for this project?
  + Salifort Motors stakeholders
* What are you trying to solve or accomplish?
  + At this stage we try to build regression model which predict if employee will resign or not and make it possible to come with solution in order to improve employee satisfaction levels at the company
* What are your initial observations when you explore the data?
  + There were 3008 duplicated rows
  + There are 83% stay and 17% who left, the class is imbalanced, we need to keep this proportion through all further tasks
  + The proportion of the employees who left is kind of similar to the proportion of the employee who have an accident at work. This can be related events, although correlation matrix did not show high correlation between these two features
  + The level of salary seems not very affected if employee left or stay, although this proportion is bigger for employee with low salary level (80% vs 20%), which is not surprise
  + The most affected departments are: sales department, technical department and support department. These department has the most employees decided to leave the company
* What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
  + On-line documentation:
    - <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rename.html>
    - <https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html>
    - <https://seaborn.pydata.org/generated/seaborn.histplot.html>
* Do you have any ethical considerations in this stage?
  + There is no personal information in the dataset. No ethical considerations in this stage

**The Nuts and Bolts of Machine Learning**

* What am I trying to solve?
  + At this stage we try to build machine learning model (one of the potential usefull algoruthms: Decision Tree, Random Forest, XGBoost) which predict if employee will resign or not and make it possible to come with solution in order to improve employee satisfaction levels at the company
* What resources do you find yourself using as you complete this stage?
  + Course 6 notes
  + On-line documentation
* Is my data reliable?
  + The data seems pretty solid
* Do you have any additional ethical considerations in this stage?
  + No, this dataset has no personal information and has no ethical issues in particular
* What data do I need/would I like to see in a perfect world to answer this question?
  + We need features with high probability suggested that employee is going to quit
* What data do I have/can I get?
  + We have dataset based on employee survey. Most of the data seems relevant and potentially may be the solid base for correct predictions
* What metric should I use to evaluate success of my business objective? Why?
  + For Classificator model we use:
    - Accuracy
    - Precision
    - Recall
    - F1 score
    - Confusion matrix
  + The main metric for this particular task will be **Recall**, as it will be more beneficial to know if employee decide to leave the company that the management can make some steps to encourage that employee to stay

**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 initial information suggest that we have relevant ans sufficient data to make quite good model

**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?
  + We begin with Discovering followed by Structuring, then Cleaning and finally Presenting
  + During all steps we will provide Validation
* 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.?
  + As we have imbalanced class for target variable (83/17) it will be great to have additional data for employees who decided to leave, although it seems that is not possible
  + We can use original dataset which has (73/29) proportion, although it is not really critical at the moment so we will use cleaned dataset with new proportion without any additional external data
* What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?
  + For duplicates – no visualization is required
  + Box plots will be helpful to determine outliers for continuous numeric variables
  + Histograms are essential to understand the distribution of variables
  + Heatmap of the correlation matrix will be helpful to visualize relationships between variables
  + Bar charts are useful for communicating levels and quantities, especially for categorical information

**The Power of Statistics**

* Why are descriptive statistics useful?
  + In general, descriptive statistics are useful because they let you quickly explore and understand large amounts of data.
* What is the difference between the null hypothesis and the alternative hypothesis?
  + The null hypothesis make assumption that observed values achieved by chance, whereas alternative hypothesis insist that it is not true, opposite to the null hypothesis

**Regression Analysis: Simplify Complex Data Relationships**

* What are some purposes of EDA before constructing a multiple linear regression model?
  + Get familiar with dataset (types, size, labels)
  + Understanding which variables are present in the data
  + Reviewing the distribution of features, such as minimum, mean, and maximum values
  + Check requirement for any initial data transfornation or reshaping and make such transformation
  + Check for missing data and deal with it
  + Check for outliers and deal with it
  + Plotting the relationship between the independent and dependent variables to visualize which feature is the best choice for X
* Do you have any ethical considerations in this stage?
  + The data has to be clear from any sensitive information. In case we have such information initially we should delete it from the dataset and change it for neutral data (not personal)
  + In the dataset there is no such information. No ethical consideration 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?
  + It still work, no need for changing our goal
* Does the data break the assumptions of the model? Is that ok, or unacceptable?
  + All except one assumptions are met for regression model:
    - Linearity – NOT OK. It seems that there is no linear relationship between logit and independent variable, which can be the reason our model poor performance
    - Independent observation – OK
    - No multicollinearity – OK
    - No extreme outliers - OK
* Why did you select the X variables you did?
  + We keep all available variables.
  + It is possible to drop “promotion\_last\_5years” feature as it seems to be not very influence on the target variable
* What are some purposes of EDA before constructing a model?
  + Get familiar with dataset (types, size, labels)
  + Understanding which variables are present in the data
  + Reviewing the distribution of features, such as minimum, mean, and maximum values
  + Check requirement for any initial data transfornation or reshaping and make such transformation
  + Check for missing data and deal with it
  + Check for outliers and deal with it
  + Plotting the relationship between the independent and dependent variables to visualize which feature is the best choice for X
* What has the EDA told you?
  + There were 3008 duplicated rows
  + There are 83% stay and 17% who left, the class is imbalanced, we need to keep this proportion through all further tasks
  + The proportion of the employees who left is kind of similar to the proportion of the employee who have an accident at work. This can be related events, although correlation matrix did not show high correlation between these two features
  + The level of salary seems not very affected if employee left or stay, although this proportion is bigger for employee with low salary level (80% vs 20%), which is not surprise
  + The most affected departments are: sales department, technical department and support department. These department has the most employees decided to leave the company
* What resources do you find yourself using as you complete this stage?
  + On-line documentation for particular graphs
* Do you have any ethical considerations in this stage?
  + No ethical consideration at this stage

**Data Project Questions & Considerations**

## **PACE: Construct Stage**

**Get Started with Python**

* Do any data variables averages look unusual?
  + All in the place, except “years in the company”, which initially had outliers and that was the reason for “unusual value”
* How many vendors, organizations or groupings are included in this total data?
  + We have information just for one company, although it contain information about employees from ten different 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?
  + Box plots will be helpful to determine outliers and where the bulk of the data points reside in terms of `drives`, `sessions` and all other continuous numeric variables
  + Histograms are essential to understand the distribution of variables
  + Heatmap of the correlation matrix will be helpful to visualize relationships between variables
  + Scatter plots also will be helpful to visualize relationships between variables
  + Bar charts are useful for communicating levels and quantities, especially for categorical information
* What processes need to be performed in order to build the necessary data visualizations?
  + For most of the variables we already have required format
  + For categorical data we need one-hot encoding technique
* Which variables are most applicable for the visualizations in this data project?
  + We will investigate all of the variables in the dataset
* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?
  + Request additional data, delete if not big amount, fill with appropriate values, create Nan category
  + If the missing data are missing completely at random (MCAR), meaning that the reason for missingness is independent of the data values themselves, we can proceed with a complete-case analysis by removing the rows with missing values. Otherwise, we need to investigate the root cause of the missingness and make sure it won't interfere with the statistical inference and modeling.

**The Power of Statistics**

* How did you formulate your null hypothesis and alternative hypothesis?
  + Null hypothesis we formulate about question we would like to investigate make assumptions that any interconnections between values there are just by chance.
  + The alternative hypothesis is stated opposite null hypothesis with clarifying if we want to find out if the value is not equal, less or greater than original data
  + There is no any significant correlation between dependent variable and independent variables
  + There is significant correlation between dependent variable and independent variables
* What conclusion can be drawn from the hypothesis test?
  + We need to conduct hypothesis test, although with some level of uncertainly we can propose that independent feature some how present information leading to changing dependent variable

**Regression Analysis: Simplify Complex Data Relationships**

* Do you notice anything odd?
  + There were just little amount of outliers
  + Unbalanced class in target variable
  + Linear assumption seems to be not met, which can explain why this model performs so poorly
* Can you improve it? Is there anything you would change about the model?
  + One approach is to scale the features, although this also did not get desired result
  + We also can reduce number of features and see if this will lead to any improvements (backward elimination strategy)
  + We also can try rebalanced data with one of possible strategy to make it more class balanced
  + The alternative approach is to try machine learning models to see if they can perform better on this dataset

**The Nuts and Bolts of Machine Learning**

* Is there a problem? Can it be fixed? If so, how?
  + As either Random Forest, either XGBoost makes any assumptions of the data we need very little data preparation. All data preparation we made for Logistic regression would do with machine learning model.
  + There is no problem
* Which independent variables did you choose for the model, and why?
  + We will investigate all of the variables in the dataset as all of them seems to be relevant and have influence on target variable “left”
  + The only concern is “promotion\_last\_5years” variable which can have no influence on the target variable although that need to be checked
* How well does your model fit the data? (What is my model’s validation score?)
  + The Recall score for Fandom Forest is 92.47% on the train data and 92.46% on the test data
  + The F1 score for the Random Forest for train data is 90.48% for train data and 90.86% for the test data
  + The recall score of the XGBoost is 93.16% for train data and 92.96% on the test data
  + The F1 score for the XGBoost model is 95% for train data and 93.67% for the test data
* Can you improve it? Is there anything you would change about the model?
  + We used GridSearch approach with cross validation which allowed us to get model with the highest target score (recall in our case)
  + Adding new data with “left” class data will be helpful to make models performance better, but in general the models performance are acceptable as are
* Do you have any ethical considerations in this stage?
  + There is no ethical consideration for this dataset.

**Data Project Questions & Considerations**

## PAC**E: 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?
  + Although the target variable is imbalanced, we can proceed with this data and try to build model based on this info
* What data initially presents as containing anomalies?
  + Initially “years\_in\_company” variable had outliers, we deal with it accordingly
* What additional types of data could strengthen this dataset?
  + Additional data with information related to employee decided to leave

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

* What key insights emerged from your EDA and visualizations(s)?
  + There were 3008 duplicated rows
  + There are 83% stay and 17% who left, the class is imbalanced, we need to keep this proportion through all further tasks
  + The proportion of the employees who left is kind of similar to the proportion of the employee who have an accident at work. This can be related events, although correlation matrix did not show high correlation between these two features
  + The level of salary seems not very affected if employee left or stay, although this proportion is bigger for employee with low salary level (80% vs 20%), which is not surprise
  + The most affected departments are: sales department, technical department and support department. These department has the most employees decided to leave the company
* What business recommendations do you propose based on the visualization(s) built?
  + Investigate further if having incident really lead employee to decision to quit and leave the company or this is just coincidence
  + Investigate further if belonging to one of particular department (sales department, technical department and support department) lead to leave the company decision
  + People with low salary trend to leave the company more often than people from better salary, which is no surprise
  + The highest level of risk to quit for the employee who works from three to five years with the number of such employee decreasing with the maximum at three years
  + 40% self satisfactory level may suggest tendency to “left” decision
  + It is understandable, that employees have to choose right balance between work and personal life, so making lot of extra hours leads to "burning out" and eventually to "left" decision
  + The balance between to little and too much projects for one employee is between 3 - 5 projects. Future increasing number of projects for particular employee will lead to “left” decision
* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
  + Investigate further if having incident really lead employee to decision to quit and leave the company or this is just coincidence
* How might you share these visualizations with different audiences?
  + Depending on audience we want to highlight different features
  + It is wise especially for non-technical audience to add annotation for key points and add legends and clear titles
  + If it is required add more details, we can add text field and provide more explanation

**The Power of Statistics**

* What key business insight(s) emerged from your A/B test?
  + Did not make any A/B test within this project
* What business recommendations do you propose based on your results?
  + Did not make any A/B test within this project

**Regression Analysis: Simplify Complex Data Relationships**

* To interpret model results, why is it important to interpret the beta coefficients?
  + This coefficient shows significance of the impact to the target variable and also make it possible to understand changes in dependent variable decreasing or increasing independent variable making business decision based on that impact.
* What potential recommendations would you make to your manager/company?
  + Based on the score which Logistic Regression showed, the only recommendation would be to continue to investigate and try to improve Logistic Regression performance, or try to use alternative approach with other algorithms like Random Forest or XGBoost
* Do you think your model could be improved? Why or why not? How?
  + One approach is to scale the features, although this also did not get desired result
  + We also can reduce number of features and see if this will lead to any improvements (backward elimination strategy)
  + We also can try rebalanced data with one of possible strategy to make it more class balanced
  + The alternative approach is to try machine learning models to see if they can perform better on this dataset
* What business recommendations do you propose based on the models built?
  + The business recommendation would be not to use this model and continue research for finding best model for this particular dataset. (Later this recommendation was confirmed with XGBoost model very sufficient perfiormance)
* What key insights emerged from your model(s)?
  + As model performance is very poor any insights from this model unreliable
* Do you have any ethical considerations at this stage?
  + There is no ethical consideration for this dataset

**The Nuts and Bolts of Machine Learning**

* What key insights emerged from your model(s)?
  + For XGBoos model the most important feature is average monthly hours, following by satisfactory level, employee performance, number of projects and years spent in the company
  + Average monthly hours confirm our preliminary suggestion that this feature has great impact on the "leave" decision and little hours leads to "boredom" of the employee whereas great amount of hours lead to employee's "burning out"
  + Satisfactory level was also one we noted during EDA as well as employee performance and number of projects. So, we have no surprise here and all followed our expectations
  + Belonging to one or another department have little effect on "leave" decision, as well as having promotion
  + The model's most predictive features were **average monthly hours**, following by **satisfactory level**, **employee performance**, **number of projects** and **years spent in the company**
  + “**average monthly hours**”: it is important to keep “gold medium” as to little hours leads to "boredom" of the employee whereas great number of hours lead to employee's "burning out"
  + **“satisfactory level”:** Extremely low (~0%), a little bit less than average (38% - 48%) and to optimistic (76% - 90%) are the notably values which can suggest that employee is required attention and encourage measure in order not to leave the company
  + **“employee performance”:** mediocre performance (48% - 58%) and extraordinary good performance (>80%) is the value for additional attention. Mediocre performance may suggest “burning out”, whereas great performance may required additional motivation for employee to stay and not to look after new job
  + **“number of projects”:** the balance between too little and too much projects for one employee is between 3 - 5 projects
  + “**years spent in the company**”: The highest level of risk to quit for the employee who works from three to five years with the number of such employee decreasing with the maximum at three years
* What are the criteria for model selection?
  + For this particular task we focus on “Recall” metric trying to decrease unawareness about employee decision to leave the company that the company management make take some measures in order to improve situation and prevent employee from leaving encourage him to stay
* Does my model make sense? Are my final results acceptable?
  + The model performance is good with 92.96% recall score for test data and 93.67% f1 score (XGBoost model)
  + For the test data there were just 50 incorrect results in total (XGBoost)
  + 22 out of 2001 were falsely identify as leaving, where as they are employees who not consider to leave the company (twice less than with Random Forest)
  + 28 out of 398 were falsely identify as not leaving and actually left the company (XGBoost)
  + The result is very sufficient and this model will be recommended as a "champion" and for the real case usage (this result is acceptable)
* Were there any features that were not important at all? What if you take them out?
  + For XGBoost belonging to the particular department and having (no having) promotion for the last 5 years having very little effect on the final model decision.
  + Despite that fact I would not recommend to remove this information as these features have little but effect on the decision, and may worse model performance in case of removing
* Given what you know about the data and the models you were using, what other questions could you address for the team?
  + It will be important to ask domain experts why they believe this feature scores so highly in this model.
  + It will be helpful to have new data with employee who decided to leave company in order to improve balance in the target variable and improve model performance
* What resources do you find yourself using as you complete this stage?
  + On-line documentation for Random Forest and XGBoost
  + Notes from the course (both digital and manual notes)
* Is my model ethical?
  + There is no ethical consideration for this dataset
* When my model makes a mistake, what is happening? How does that translate to my use case?
  + When model predict negative class (meaning employee is going to stay in the company) whereas in reality it is positive class (employee decided to leave the company), the company management has no opportunity to provide any measures and encourage employee to stay. So, the company will lose employee which is bad for business operation and bad for business in general, especially when the company lose high-quality employee
  + When model predict positive class (meaning employee is going to leave the company) whereas in reality it is negative class (employee have no desire to leave the company), the management can provide some encouraging measures and try to improve employees working conditions which can be kind of surprise for employee and can boos employee’s performance level which is very good for business, so in this case the company would benefit from this.