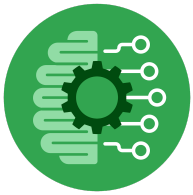
Course Six
The Nuts and Bolts of Machine Learning



Instructions

Use this PACE strategy document to record decisions and reflections as you work through the end-of-course project. As a reminder, this document is a resource that you can reference in the future and a guide to help consider responses and reflections posed at various points throughout projects.

Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

- [ ] Complete the questions in the Course 6 PACE strategy document
- [ ] Answer the questions in the Jupyter notebook project file
- [ ] Build a machine learning model
- [ ] Create an executive summary for team members and other stakeholders

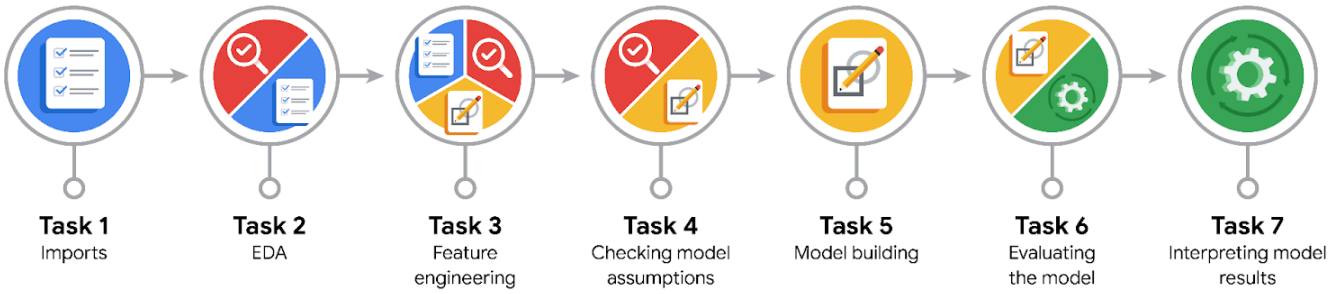
Relevant Interview Questions

Completing the end-of-course project will empower you to respond to the following interview topics:

- What kinds of business problems would be best addressed by supervised learning models?
- What requirements are needed to create effective supervised learning models?
- What does machine learning mean to you?
- How would you explain what machine learning algorithms do to a teammate who is new to the concept?
- How does gradient boosting work?

Reference Guide:

This project has seven tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Plan Stage

- What are you trying to solve or accomplish?

  The goal from our project was to create an algorithm to predict the cab fare when a client requests a trip. In this part of the stage, we're going to start the algorithm to create a machine learning model to determine if the customer will give a tip or not.

- Who are your external stakeholders that I will be presenting for this project?

  Juliana Soto, Finance and Administration Department Head

  Titus Nelson, Operations Manager

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

  We're going to use a Jupyter Notebook to create the model with the appropriate libraries for exploratory data analysis, random forest models and visualization packages.

- Do you have any ethical considerations at this stage?

  Yes, we have to stay impartial in favoring one of the parties (client and driver) to avoid unfortunate experiences.

- Is my data reliable?

  Yes, from our previous works, it seems that our data is reliable but there are some outliers that further analysis is required.

- What data do I need/would like to see in a perfect world to answer this question?

Ideally, we need data with no outliers and the ones that the regression model shows an R^2 of 1, and in the random forest, the metrics are close to 1 or equal to it.

- What data do I have/can I get?

- What metric should I use to evaluate success of my business/organizational objective? Why?

F1, precision, recall, and accuracy are the metrics to evaluate our random forest model.

PACE: Analyze Stage

- Revisit "What am I trying to solve?"Does it still work? Does the plan need revising?

Since the goal for the machine learning is to determine for the drivers which customers will give a tip of 20% from the trip. Then, it's important to find a balance between the drivers and customers needs.

- Does the data break the assumptions of the model? Is that ok, or unacceptable?

It seems that the observations are independent from each other which is an assumption for the random forest model. Although, there is kind of multicollinearity when performing regression model,

- Why did you select the X variables you did?

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

It's important to know our data and determine which variables will be useful to make some transformations in our data and create new variables that could help us in the model.

- What has the EDA told you?

There are columns in our data that need to be converted into boolean and datetime format. Also, it help us to determine missing data and basic information from the data.

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

In this part, pandas was useful to convert our data and put it into a new dataframe that will help us to create the model.

PACE: Construct Stage

- Do I notice anything odd? Is it a problem? Can it be fixed? If so, how?

Before modelling, we check the class balance from our variable `generous` and it seems almost 50/50

- Which independent variables did you choose for the model, and why?

Our y variables are generous and the x was the rest of the columns.

- How well does your model fit the data? What is my model's validation score?

From the metrics, it's reasonable

- Can you improve it? Is there anything you would change about the model?

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

PACE: Execute Stage

- What key insights emerged from your model(s)? Can you explain my model?

  The rain model forest is reasonable to work with because it presents a F1 score of 0.7194 and an accuracy of 68.22%. We also determined which features were the most important, being 'Vendor', 'predicted fare', 'mean_duration', and 'mean_distance'. Although, we don't know how they affect, so further analysis is required.

- What are the criteria for model selection?

  Confusion Matrix and the metrics (F1, precision, recall, and accuracy)

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

  The metrics are great but the confusion matrix shows that the model could predict more false positives than negatives because it means that the driver could be pleasantly surprised by a generous tip when they weren't expecting one than to be disappointed by a low tip when they were expecting a generous one

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

  Yes, a new column(s) to determine the length of the trip, the past tipping behaviour for each client, which client paid their tips with cash and were not registered, and work with more data to create a unique model.

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

  We removed some variables to create the rainforest model, and in the end, we determined the feature importances. So, further analysis need to be done to know how these variables affect

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

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

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

- Is my model ethical?

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