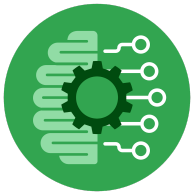
Course Six
The Nuts and Bolts of Machine Learning



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

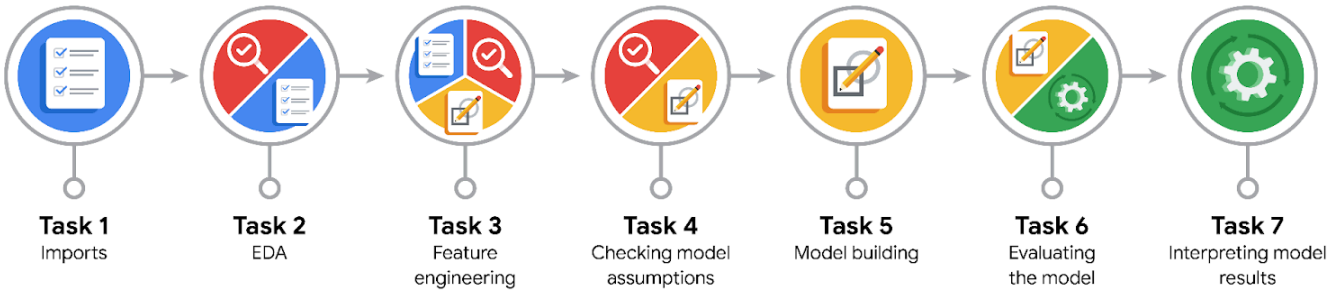
Relevant Interview Questions

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?

I am trying to build a machine learning model on how well customers tip.

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

The external stakeholders is from the Automatidata which approved of the algorithm of random forests.

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

In this activity, I would need to use extensive Python's scikit-learn machine learning modules.

- Do you have any ethical considerations at this stage?

Of course, because I can't just model on whether people paid or not, whether more generous or not.

- Is my data reliable?

Yes it is reliable, but feature engineering and exploratory data analysis must be done.

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

There are some data, but nothing is perfect. I would need to convert generous (20% tip) and get the mean distances and durations.

- What data do I have/can I get?

Predicted fares and mean distances from the previous course is the data that is vital for this.

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

Since this is a supervised model, I would need the F1 score and the other three metrics (precision, recall, accuracy). I call it “the Big Four” for supervised models.

PACE: Analyze Stage

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

Well, it won’t work that well in this stage, since I am trying to solve a complex problem.

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

No, because the assumptions on random forest is limited.

- Why did you select the X variables you did?

Because the X variables I selected are most important to predict generosity of tipping.

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

Some purposes of EDA is to tidy up data before building the model, and removing redundant rows.

- What has the EDA told you?

The EDA told me that there are quite a lot of reconstruction needed to be done, such as encoding dummies, creating new variables, and more.

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

The resources are the same as the plan stage, but I need to import Python models for specific exploratory data analysis AND I need to re-visit some date time topics.

PACE: Construct Stage

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

There is nothing odd right here.

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

The independent variables include mean distances, mean durations, predicted fares, hour of day, month of year, locations, and passenger counts.

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

Random forest: 0.748, F1: 0.722

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

Well, there is nothing to change here as the model is very reasonable.

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

I would need to use the machine learning modules throughout the course.

PACE: Execute Stage

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

Vendor ID and predicted fares are the most important features and the model is more prone to Type I errors than type II errors. The metric scores are good, but the model can't be explained easily since random forests are not easily explained.

- What are the criteria for model selection?

The F1 score, which one is higher.

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

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

Well, there is still some space for improvement, but it will be a complicated question. Maybe more explainable models such as decision trees could work.

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

Passenger counts, months, etc. are not important at all.

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

Vendors and fares are the reasons for more generous tipping.

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

Are there any other possible ways to do a machine learning model?

Can we make a more explainable model?

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

Looking over the model I created.

- Is my model ethical?

It is of course more ethical than “predict whether a customer will make a tip or not”

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

It will erroneously label some passenger as “generous” or “not generous”.