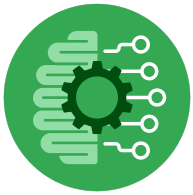
Course Six
The Nuts and Bolts of Machine Learning



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~~

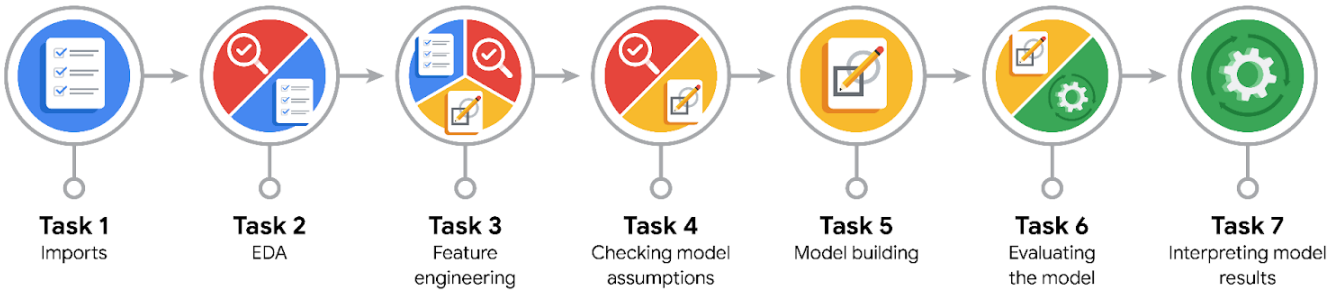
Relevant Interview Questions

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?

Predict if a customer will not leave a tip.

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

New York City Taxi & Limousine Commission (TLC).

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

Packages for numerics and dataframes.
Packages for visualization.
Packages for date conversions.

- Do you have any ethical considerations at this stage?

Drivers who didn't receive tips will probably be upset that the app told them a customer would leave a tip. If it happened often, drivers might not trust the app. Drivers are unlikely to pick up people who are predicted to not leave tips. Customers will have difficulty finding a taxi that will pick them up, and might get angry at the taxi company.

- Is my data reliable?

Yes.

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

Past tipping behavior for each customer and a lot more data.

- What data do I have/can I get?

Behavioral history for each customer.

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

accuracy, precision, recall, F-score, area under the ROC curve, or a number of other metrics.

PACE: Analyze Stage

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

It’s still working.

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

Tree-based models are resilient to outliers, so there is no need to make any imputations.

- Why did you select the X variables you did?

Because it's help predict target feature.

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

For prepare data that can impact to model.

- What has the EDA told you?

Approximately 1/3 of the customers in this dataset were "generous" (tipped ≥ 20%). The dataset is imbalanced, but not extremely so.

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

Packages for numerics and dataframes.
Packages for date conversions.

PACE: Construct Stage

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

No.

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

Generous for find customer who tipped equal or above 20%.

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

All score is too low. It is not a great model.

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

It would probably be very helpful to have past tipping behavior for each customer.
It would also be valuable to have accurate tip values for customers who pay with cash.
It would be helpful to have a lot more data.

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

Packages for confusion matrix, model_selection, ensemble, xgboost.

PACE: Execute Stage

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

After built the identified models and performed the testing, it became clear that there was not as strong a correlation as anticipated, with an F1 score of just 0.350.

- What are the criteria for model selection?

Check model's validation score.

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

  It is not a great model, but depending on how it's used it could still be useful.

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

  It would probably be very helpful to have past tipping behavior for each customer.
  It would also be valuable to have accurate tip values for customers who pay with cash.
  It would be helpful to have a lot more data.

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

  Drop redundant and irrelevant columns as well as those that would not be available when the model is deployed. This includes information like payment type, trip distance, tip amount, tip percentage, total amount, toll amount, etc.

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

  Collect/add more granular driver and user-level data, including past tipping behavior.
  Cluster with K-means and analyze the clusters to derive insights from the data

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

  Is it will be useful if the objective is only to help give taxi drivers a better idea of whether someone will leave a good tip because model not great.

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

Packages for confusion matrix.

- Is my model ethical?

Even when the model is correct, people who can't afford to tip will find it more difficult to get taxis, which limits the accessibility of taxi service to those who pay extra.

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

This is not a great model, but depending on how it's used it could still be useful. If the objective is only to help give taxi drivers a better idea of whether someone will leave a good tip, then it could be useful. It may be worthwhile to test it with a select group of taxi drivers to get feedback.