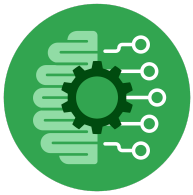
Course Six
The Nuts and Bolts of Machine Learning



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

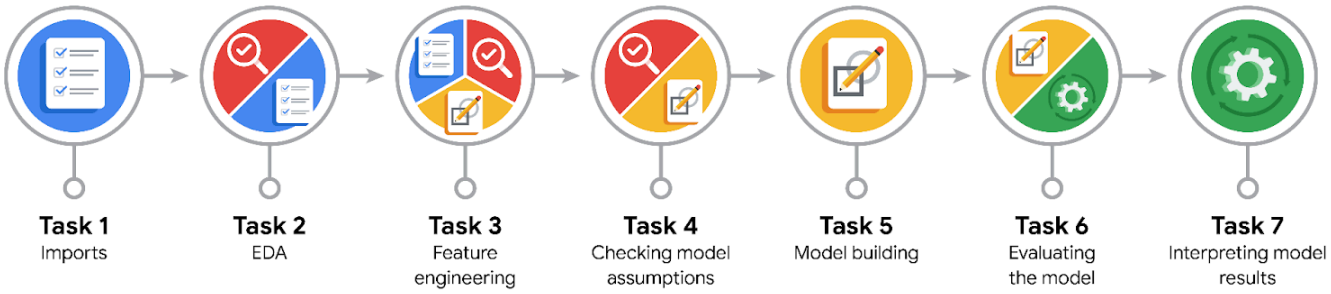
Relevant Interview Questions

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?

  We tasked with building a machine learning model to predict if a taxi customer will or will not leave a tip, helping drivers make better decisions to increase their earnings.

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

  The external stakeholders are the New York City Taxi & Limousine Commission (NYC TLC) and the taxi drivers, as they will be directly impacted by the predictions made by the model.

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

  Data (e.g., taxi trip data), documentation for machine learning libraries like scikit-learn and XGBoost, and any guidelines or ethical considerations for predicting customer behavior.

- Do you have any ethical considerations at this stage?

  Predicting tipping behavior could lead to bias and discrimination. Ethical concerns arise if drivers refuse rides to customers predicted to not tip.

- Is my data reliable?

We need to ensure that your data is complete, accurate, and representative of the tipping behavior you're trying to model. Look for missing or erroneous values.

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

Ideally, we'd want more detailed features, like customer profiles, weather conditions, or economic factors that might influence tipping behavior.

- What data do I have/can I get?

The taxi trip data contains useful features like fare amount, trip distance, and passenger count, but may lack contextual information like customer history or driver behavior.

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

F1 score, as it balances precision and recall. In this case, both false positives and false negatives carry a cost, and F1 will help balance them.

PACE: Analyze Stage

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

The goal is still to predict tipping behavior. If initial EDA shows that tipping patterns are unpredictable or unreliable, the objective may need adjustment.

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

Depending on the model used (e.g., Random Forest or XGBoost), some assumptions might not be critical (e.g., linearity), but ensure features are not highly correlated.

- Why did you select the X variables you did?

Features like fare amount, distance, and pickup/dropoff times directly correlate with the likelihood of receiving a tip, making them suitable predictors.

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

EDA helps identify trends, anomalies, and outliers in the data. It also helps you understand the distributions of your features and target variable.

- What has the EDA told you?

EDA might show whether certain times of day, locations, or trip characteristics lead to higher or lower tips. It also identifies any missing or anomalous values.

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

Python libraries like pandas for EDA, matplotlib and seaborn for visualizations, and scikit-learn for data preprocessing.

PACE: Construct Stage

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

Odd values like negative fares or tips can occur. They need to be removed or corrected for the model to perform well.

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

We might choose features like fare amount, trip distance, passenger count, pickup time, and location because they logically influence tipping behavior.

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

Model validation scores (e.g., F1 score, precision, recall) indicate how well the model is performing on unseen data, and can be improved with hyperparameter tuning.

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

Hyperparameter tuning, feature selection, and cross-validation are ways to improve model performance.

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

Python libraries for modeling like scikit-learn and XGBoost, GridSearchCV for hyperparameter tuning, and cross-validation techniques.

PACE: Execute Stage

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

Insights such as which times of day, locations, or fare amounts are most predictive of tipping behavior can emerge. The model is built to predict customer tipping behavior based on these features.

- What are the criteria for model selection?

Criteria include the model's ability to generalize, the F1 score, and how well it balances false positives and false negatives.

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

If our model's validation and test scores align and make logical predictions based on real-world data, then it is likely valid.

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

It might be improved by adding more relevant features, such as customer profiles or external factors (e.g., weather).

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

Some categorical variables or certain times of day might not contribute much to predicting tipping behavior, and could be excluded.

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

Recommend drivers focus on specific times, locations, or trip types that are most predictive of generous tipping behavior.

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

1. What factors contribute most to a generous tip?
2. How can driver behavior improve tipping outcomes?
3. What times or locations are best for maximizing tips?
4. Are there external factors that impact tipping behavior?
5. Can we predict customer satisfaction based on tipping data?
6. Is there a seasonal trend in tipping behavior?
7. How do ride characteristics (e.g., distance, duration, passenger count) influence tipping?
8. Is there a relationship between fare amount and tip percentage?
9. How can we reduce prediction errors in the model?
10. Are there demographic patterns in tipping behavior?

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

We could analyze customer satisfaction or predict when and where drivers will make the most trips.

- Is my model ethical?

It's essential to consider whether the model could unfairly disadvantage certain groups of customers. For example, predictions of no tip could lead to biased driver behavior.

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

A mistake (false positive or false negative) can frustrate both drivers and customers. A false positive could lead to a missed opportunity for a good fare, and a false negative could lead to refusing a ride to a generous tipper.