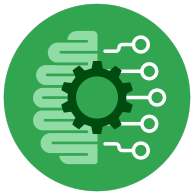
Course Six
The Nuts and Bolts of Machine Learning



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~~

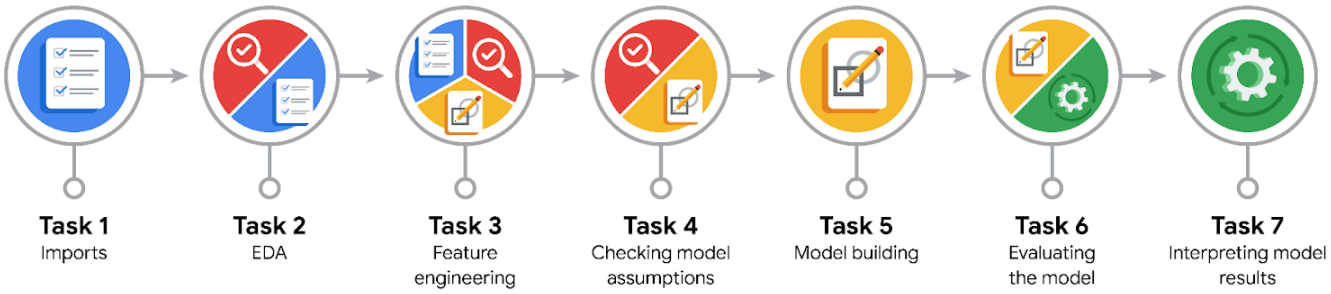
Relevant Interview Questions

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?

To help the New York City Taxi & Limousine Commission (TLC) understand what factors influence whether a customer will be a generous tipper (≥ 20%).
The goal is to build a predictive model that classifies trips as “generous” or “not generous,” enabling TLC to make data-driven decisions that improve driver satisfaction and tipping outcomes.

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

Juliana Soto, Finance and Administration Department Head, NYC TLC

Titus Nelson, Operations Manager, NYC TLC

Udo Bankole, Director of Data Analysis, Automatidata

Automatidata Data Analytics Team

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

TLC trip dataset (cleaned and prepared previously)

Python libraries: Pandas, NumPy, Scikit-learn, XGBoost

Jupyter Notebook for model development

Course documentation and Automatidata project instructions

- Do you have any ethical considerations at this stage?

Yes — ensuring the model does not introduce bias against any vendor, route, or time of day.

We must also safeguard trip data privacy and not make predictions that could unfairly influence driver allocation or pay.

- Is my data reliable?

Yes, the data is sourced from the official NYC TLC dataset and previously cleaned during earlier Automatidata projects (missing values handled, features standardized).

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

Weather, traffic conditions, passenger demographics, driver experience, and trip purpose — these could enhance prediction accuracy.

- What data do I have/can I get?

Trip-level data: fare amount, duration, distance, passenger count, vendor ID, day, month, and time features (e.g., rush hour).

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

F1 Score, because it balances precision and recall.

In this case, both false positives (predicting generous when not) and false negatives (missing generous riders) are equally important.

PACE: Analyze Stage

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

Yes — the goal of predicting tipping behavior remains relevant and well-supported by the data.

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

No major issues. Random Forest and XGBoost do not assume linearity or normality, so they are robust to feature distributions.

- Why did you select the X variables you did?

Selected based on logical connection to rider tipping: fare amount, trip duration, distance, passenger count, vendor ID, and temporal factors (day/month).

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

To understand data distribution, detect outliers or missing values, and identify correlations that might influence tipping.

- What has the EDA told you?

Higher fares and longer trips correlate with generous tips.

VendorID_2 trips show slightly higher tipping rates.

Rush hour and nighttime trips show distinct patterns.

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

Python (Pandas, Matplotlib, Seaborn), Jupyter Notebook visualizations, and prior project cleaning work.

PACE: Construct Stage

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

Slight class imbalance (fewer generous tips).

Handled by using balanced parameters in model training and cross-validation.

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

`predicted_fare`, `mean_duration`, `mean_distance`, `passenger_count`, `VendorID_2`, `day_of_week`, `month`, `rush_hour`

These were chosen for their direct influence on ride cost and tipping opportunity.

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

Both models performed similarly with F1 ≈ 0.71 and accuracy ≈ 0.68 on test data.

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

Adding contextual features (e.g., weather or traffic) could improve performance. Hyperparameter tuning or ensemble averaging could further refine accuracy.

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

Python (Scikit-learn, XGBoost), GridSearchCV for tuning, and make_results/test_scores helper functions.

PACE: Execute Stage

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

Trip characteristics (fare, duration, distance) strongly influence tipping.

Vendor differences also play a role.

The XGBoost model performed slightly better than Random Forest in generalization and precision.

- What are the criteria for model selection?

Accuracy, F1 score, interpretability, and ability to generalize.

XGBoost was selected as the preferred model.

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

Yes, consistent and interpretable. F1 ≈ 0.71 is acceptable for behavioral prediction with limited contextual data.

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

Yes — by adding external data (weather, driver/passenger factors) and retraining periodically.

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

Some time features (specific months) had low importance — removing them had minimal impact on model performance.

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

Use XGBoost predictions to identify likely generous tippers.

Consider incentive programs for drivers based on model predictions.

Collect richer trip context to improve future models.

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

How do seasonal patterns influence tipping?

Can we use the model to forecast overall tipping revenue trends?

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

Python notebook outputs, feature importance plots, and confusion matrix analysis.

- Is my model ethical?

Yes — predictions are based solely on trip characteristics, not personal or demographic data.

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

False Positive: Predicting a generous tip when it isn't — may lead to misplaced incentives. **False Negative:** Missing a generous tipper — potential missed opportunity for optimizing driver assignments. Overall, the model's balanced precision and recall minimize both risks.