**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 TLC wants to know which customer will not give a tip.

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

The TLC team are our stakeholders.

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

Sklearn, xgboost

* Do you have any ethical considerations at this stage?

Yes, TLC originally wanted to find out which customers did not give tips at all. This notion can impact the way drivers pick up customers and customers having trouble getting rides. Hence, it was not ethical to create such a model.

* Is my data reliable?

Yes, the data is complete with no missing values. AlTHOUGH, EDA needs to be done to prepare the data for training.

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

In summary, the "perfect" data would be comprehensive, high-quality, and contextually relevant, allowing us to fully capture and analyze the factors influencing the answer to your question.

* What data do I have/can I get?

We have the original data combined with the mean data gathered through the last course. Combining this data gives us almost a 50-50 distribution of our target variable.

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

This is a supervised learning, classification task. We could use accuracy, precision, recall, F-score, area under the ROC curve, or a number of other metrics. However, both models were later refitted to F1 score.

**PACE: Analyze Stage**

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

We can build a model that predicts the most generous customers. This could accomplish the goal of helping taxi drivers increase their earnings from tips while preventing the wrongful exclusion of certain people from using taxis.

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

No, the data doesn’t break any assumptions.

* Why did you select the X variables you did?

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

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

EDA gives insights into the complexity of data patterns and relationships, which can help guide model selection. DA helps you decide if the problem might be suited to linear models, tree-based models, or more complex models like neural networks.

* What has the EDA told you?

EDA told us to create a Random Forest and XGB Classifier.

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

Pandas functions to select/transform features, drop irrelevant columns, adding new columns.

**PACE: Construct Stage**

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

A little over half of the customers in this dataset were "generous" (tipped ≥ 20%). The dataset is very nearly balanced hence, nothing to worry about.

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

Passenger\_count, mean\_duration, mean\_distance, predicted\_fare, generous, am\_rush, daytime, pm\_rush, nighttime, VendorID, RatecodeID, PULocationID, DOLocationID.

All these variables will help determine which customers will give generous tips.

* How well does your model fit the data? What is my model’s validation score?

My model's F1 score = 0.8214509295246202. It is a fairly good model.

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

Yes, it can be improved with tweaking the hyperparameters.

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

from sklearn.model\_selection import GridSearchCV, train\_test\_split

from sklearn.metrics import roc\_auc\_score, roc\_curve

from sklearn.metrics import accuracy\_score, precision\_score, recall\_score,\

f1\_score, confusion\_matrix, ConfusionMatrixDisplay, RocCurveDisplay

from sklearn.ensemble import RandomForestClassifier

from xgboost import XGBClassifier

**PACE: Execute Stage**

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

The model is almost twice as likely to predict a false positive than it is to predict a false negative. Therefore, type I errors are more common. This is less desirable, because it's better for a driver to 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. However, the overall performance of this model is satisfactory.

* What are the criteria for model selection?

A good F1 score and Cost of Errors.

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

Yes, this model performs acceptably. Its F1 score was 0.82 and it had an overall accuracy of 0.69. It correctly identified ~78% of the actual responders in the test set, which is 48% better than a random guess. It may be worthwhile to test the model with a select group of taxi drivers to get feedback.

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

Yes, with more/different hyperparameters we could test how they impact our model’s results.

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

Rush hour columns, Month columns

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

The model is nearly twice as likely to predict a false positive compared to a false negative. Testing it with a small group of taxi drivers for feedback could be beneficial.

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

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. With enough data, we could create a unique feature for each pickup/dropoff combination.

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

Documentation and Technical References, Experiment Tracking Tools, Hyperparameter Tuning and Optimization Libraries

* Is my model ethical?

Yes, since we build a model that predicts the most generous customers.

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

When a model makes a mistake, it means it has incorrectly predicted an outcome or classified an instance.