**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 are going to build a machine learning model to identify which riders are generous tippers?

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

TLC commission members

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

Visualization libraries. SkLearn, Pandas

* Do you have any ethical considerations at this stage?

No.

* Is my data reliable?

It looks like we have reliable data.

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

Rides info with tipping data.

* What data do I have/can I get?
* What metric should I use to evaluate success of my business/organizational objective? Why?

Classification matrix identifying riders that tip 20% or more of the ridding fee.

**PACE: Analyze Stage**

* Revisit “What am I trying to solve?”Does it still work? Does the plan need revising?
* We are going to build a machine learning model to identify which riders are generous tippers?
* Does the data break the assumptions of the model? Is that ok, or unacceptable?

Data looks fine

* Why did you select the X variables you did?

They might have an impact of whether a rider is a generous tipper.

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

Understand the data we are dealing with. It might have missing data or require feature transformation in order to achieve the purpose of the exercise.

* What has the EDA told you?

The dataset doesn’t have missing values. Several features are strings that require transformation to be useful in the model. Several columns should be dropped.

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

Feature transformation.

**PACE: Construct Stage**

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

It takes a long time to train the models

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

VendorID,passenger\_count,RatecodeID,PULocationID, DOLocationID, mean\_duration, mean\_distance, predicted\_fare,day, am\_rush, daytime, pm\_rush, nighttime, month . These independent variables might impact how riders feel about the ride and the tip the give the driver.

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

The validation score is 0.718. The model fits acceptably well the data.

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

Yes. We can try other hyperparameters to improve the model.

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

Grid Search for Cross Validation

**PACE: Execute Stage**

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

The XGBoost model has the best predictive ability. Predictive fare, mean distance, mean\_duration are features with greatest impact to generous tipping.

* What are the criteria for model selection?

F1 scores is the criteria for model selection. Later, I also look at other classification metrics: precision, accuracy and recall.

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

The model makes sense. Final results are acceptable.

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

Yes. The model could be improved with additional features and more data.

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

Specific Location and destination data doesn’t appear to add to the model, specially in this case were mean\_duration and mean\_distance are indirectly capturing this info. Suggest deleting these columns which would greatly simplify the dataset.

* What business/organizational recommendations do you propose based on the models built?
* Given what you know about the data and the models you were using, what other questions could you address for the team?
* What resources do you find yourself using as you complete this stage?
* Is my model ethical?

Yes. The model is ethical. Drivers might prefer longer rides as those tend to results in more generous tipping.

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