**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 customers will pay a tip of at least 20% for the purpose of modeling how to generate more revenue for taxi cab drivers.

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

TLC program managers

* What resources do you find yourself using as you complete this stage?
* Project specifics: check all project requirements (see project description document)
* Decide on type of model based on the objective of the project, its ethical decision properties and implications.
* Jupiter notebook: prepare data to be readily available for PACE analyze stage -> Feature engineering
* Do you have any ethical considerations at this stage?

If drivers know via app if customers who requested the taxi won’t pay a 20% or higher tip then the driver might avoid these low tip paying customers that may have issues getting a taxi when they need to.

* Is my data reliable?

2017\_Yellow\_Taxi\_Trip\_Data.csv contains data gathered by the New York City Taxi & Limousine Commission. Data quality has been verified in prior EDA (Inspect data files for missing data and Data exploration and cleaning – see PACE Analyze)

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

All data as per model built needed (which is not the case as this needs to be custom made using feature engineering

* What data do I have/can I get?
* 2017\_Yellow\_Taxi\_Trip\_Data.csv -> base data set
* nyc\_preds\_means.csv -> predicted fares and mean distance and duration from previous course
* What metric should I use to evaluate success of my business/organizational objective? Why?

Actual vs. predicted value to account for bias.

**PACE: Analyze Stage**

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

Predict whether a rider will be a generous tipper (>= 20%). Data is available, need to filter for data with tips of 20% or higher and assign Boolean int value (1/0) to it

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

We need to sort only for customers paying with credit card as tip\_amount field is populated for credit card tips only and cash tips are not included, i.e. we work with a subset of data

* Why did you select the X variables you did?

We need to set an assumption on good/relevant predictor variables. All other fields need to be excluded (see dropped in notebook).

* What are some purposes of EDA before constructing a model?
* Data cleaning, handling of missing values
* As precondition to perform feature selection, extraction, transformation to prepare the data for modeling
* What has the EDA told you?

As per findings in previous course EDA (Regression Analysis):

* There are no duplicates or missing values in the data
* Existing outliers for fields like tip\_amount and total\_amount questioning how predictive these variables are
* What resources do you find yourself using as you complete this stage?

Data Dictionary

**PACE: Construct Stage**

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

RF CV Model accuracy is very low, while one could try with different hyperparameters.

* Which independent variables did you choose for the model, and why?
* PULocationID and DOLocationID and mean distance to account for the fact that longer distances and specific locations yield higher amounts impacting als tipping
* Time of day (rush hours, night-, daytime) accounting for different customers requiring a taxi cab
* How well does your model fit the data? What is my model’s validation score?

Given low F1 score, recall and precision the RF CV model suggests high variance to the training set. While we have only 408k of rows of data we are limited in using more data to reduce model variance (at least for now). For XGB Model F1 score is just ~0.04 higher than the RF model. All metrics are listed in the notebook.

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

As we would need more data to improve model performance (fit) and as we would need more investigation on the predictor variables (trip\_distiance, trip duration and fare\_amount) it is not more improvable. For sure we would change the predictors as soon as their influential relationship to the target variable (tipping)would be better understood.

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

Mostly the notebook and the metrics are relevant to this stage.

**PACE: Execute Stage**

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

There is still more investigations on key variables required to increase model performance and avoid false positives and false negative which are financially worse than not using the app at all.

* What are the criteria for model selection?

As with prior investigations in this project and as per expectation the most predictive variables should be used.

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

It is not a useful model as it may draw false expectations on both sides – drivers and riders. As a general orientation for high tips paying customers, it may be useful to the drivers. But while the specific relationship between predictors and tipping is not more investigated it may be not much better than flipping a coin.

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

It can be done by elaborating on additional features to be engineered.

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

The daytimes (rush hours, daytime, nighttime) and weekdays were from feature importance not as expected though influential on the target variable.

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

While it is not a great model, depending on how it is 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.

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

Elaborate more on the influence of predictor variables on the tipping.

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

Notebook

* Is my model ethical?

Even with adaptation for at minimum 20% tippers there is a high risk of adverse selection on higher tips paying customers vs. those low tips paying - as the objective of the project is to increase taxi drivers’ tips by using an app predicting high tips paying customers.

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

As the question to be answered is if a passenger will not pay a tip for the ride, so the 2 types of errors can be formulated:

* False positive (i.e., when the model says a customer will give a tip, but they actually won't) – in this case drivers might distrust the app and do not use it anymore.
* False negative (i.e., when the model says a customer will not give a tip, but they actually will) – in this case customers might have issues getting a ride at all.