**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?

I am trying to build a model that will be used in an app with the aim to improve driver satisfaction by alerting drivers about whose rider is likely not to leave a tip.

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

The external stakeholders include Juliana Soto, Department Head at the New York City TLC and Titus Nelson who is a member of the New York City TLC.

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

For the PLAN stage, I find myself referring to the email correspondence between myself and the team coupled with information shared by the NYC TL team. I am constantly referring to the project purpose and objectives as defined by the stakeholders.

* Do you have any ethical considerations at this stage?

Yes. My ethical consideration centers around data collection/consent, privacy and most importantly bias. The model making errors can result in bias which can have significant effects on every stakeholder most especially the rider's prospects and lives. Biased models can reinforce discrimination against a certain group of people based on their tip preference.

* Is my data reliable?

Given the data was provided by the client, and has been explored previously, my response would be that the data is reliable.

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

In a perfect world, income, gender, education and employment data would be helpful to answer this question. Knowing if a person is employed and their average income can be a good gauge to predict their generosity. However since these sort of information could lead to lead to ethical issues,, it's better to do without

* What data do I have/can I get?

I have data about ride history from 2017 that includes, pickup and drop off time, passenger time and more. Also there is data about mean duration, mean distance and predicted fare.

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

Since this is a binary problem, i.e. will leave tip or not, measures like recall, precision, f1 score would be more important than accuracy to use to evaluate success of the business objective.

**PACE: Analyze Stage**

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

At this stage it still works. I am still trying to solve the problem of predicting if a customer will not leave a tip.

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

Because I am exploring a tree based model approach to this classification problem, there are no model assumptions that need to be adhered to. Of course, it is important to check for missing values as tree models do not deal well with missing data.

* Why did you select the X variables you did?

My choice of X variables is informed by the business problem. For example, one could infer that the riders location time of day. For example if a ride originates from an area dominated by the high class in society, the chance of tip is greater compared to location with less influential people.

Additionally, month(people tend to be generous during holiday season), trip duration, mean trip distance can be effective variables from answering the business problem

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

EDA serves several purposes in constructing a model.

EDA helps with understanding the data such as learning about the structure, characteristics of data as well as potential issues. Detecting patterns and relationships present in the data is done at EDA. This can include learning about correlations between variables, detecting trend over terms and also learning about hidden patterns.

Because real life data are messy, EDA help to identify outliers and anomalies in the data that indicate error. By visualizing data in EDA, I can identify values that are outside the range of the allowable limits

EDA plays a crucial role in features selection and engineering which are important steps in building predictive models.

* What has the EDA told you?

EDA revealed that only customers who paid by credit card tip the driver after the ride. It also reveal through the summary statistic, variables with outliers and unreasonable values. For example, the fare amount shows a negative value. This is clearing an anomaly as fare amount can not be negative.

EDA reveals the data types of the variables.

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

I find myself using libraries and packages such as pandas, numpy and matplotlib to complete this stage. I also find that I refer to the business objective and problem statement so as to keep my analysis focus

**PACE: Construct Stage**

* Do I notice anything odd? Is it a problem? Can it be fixed? If so, how?
* Which independent variables did you choose for the model, and why?

I choose the following independent variables : VendorID, passenger count, rate code id, location id, mean trip duration, mean distance, predicted fare, day of the week, month of the year, hour of the day. The reason behind my choices are based on EDA and business problems.

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

The model fits the data relatively well. The model validation score in terms of accuracy, precision and recall are 0.693477, 0.681523, 0.783972 respectively. This indicates that the model fits the data well.

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

Yes, I would like to investigate other features in the dataset and feature engineer others. Also,I could tune the hyperparameters with a different set of initial parameters.

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

I reference the sklearn documentation a great deal at this stage of modeling. I also find myself reading blogs post, response on stackoverflow about how to select initial hyperparameters to train the model

**PACE: Execute Stage**

* What key insights emerged from your model(s)? Can you explain my model?
* What are the criteria for model selection?
* Does my model make sense? Are my final results acceptable?
* Do you think your model could be improved? Why or why not? How?
* Were there any features that were not important at all? What if you take them out?
* 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?
* When my model makes a mistake, what is happening? How does that translate to my use case?