**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 build a model which accurately identifies generous tippers at the moment of ordering the car.

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

The workers of NYC TLC are my external stakeholders.

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

I use a lot of code documentation, and resort to the help of chatGPT

* Do you have any ethical considerations at this stage?

The data available is not private and no discrimination is done based on other possible variables

* Is my data reliable?

Yes, it is. Its source is the TLC’s historical database.

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

All data available now, and personal facts, such as person’s demographics and financial information.

* What data do I have/can I get?

I can ask for some more personal data about customers.

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

In the given context, f1 score is the most appropriate, it balances out both precision and recall, so the interests of drivers and passengers are taken into the account

**PACE: Analyze Stage**

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

Everything is fine. I keep training the model which labels the customer as the generous tipper based on the data available at the moment of ordering the car.

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

No, it does not.

* Why did you select the X variables you did?

I filtered out those X variables which are accessible at the moment of ordering the car

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

It allows to identify the anomalies in the data beforehand, so these can be removed and normalized

* What has the EDA told you?

The data values are normal, the data types of some of the columns must be changed.

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

Code documentation, data dictionary, chatGPT

**PACE: Construct Stage**

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

For some reason the model performed extremely well on the test data. Probably, coincidence. Therefore I retrained the model. Also, I modified the list of predictor variables. Now the results are lower but more reflect reality.

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

There is a long list of them, the main condition was availability at the moment of ordering the car

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

It is decent. The best f1 score is 0.45

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

I try out the other algorithm - XGBoost Classifier.

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

ChatGPT

**PACE: Execute Stage**

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

Fare, duration, distance and the hour of the order are the most important variables regarding the predictive power. The model is a bit better at identifying true positives - it is better for drivers as they are able to be more sure in that the customer will give them the tip over 20% percent of the price.

* What are the criteria for model selection?

It must be able to perform logistic binary classification tasks. Preferably, not requirable regarding data assumptions

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

The model makes sense. However, it must be improved in the future and the key metric is to be identified after the discussion with the external stakeholders from NYC TLC.

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

Surely, it could be. Some demographic and other personal data about the customers could help massively. Some feature engineering as well. Also, the historical data of the passenger’s trips would improve the model’s performance drastically either.

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

Yes, there were. The model’s predictive power is not going to drop significantly if I remove them

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

I can’t propose anything yet as the priority is to be identified(driver, passenger or the balance) and additional improvement to be made.

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

I would ask for general advice as the other perspective always brings new solutions and insights into the project, making the process of its completion more productive.

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

Still the same, documentation, chatGPT, data dictionary.

* Is my model ethical?

No personally sensitive data, so yes, it is.

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

If the model fails to identify generous tipper, then the likelihood of picking him up is lower(unsatisfied customer), otherwise, in the case of labeling the customer as the generous tipper and him not paying it the driver is upset.