**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 to predict if a customer will not leave a tip for the driver.

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

My external stakeholders for this project are from New York City TLC. Juliana Soto, Finance and Administration Department Head; and Titus Nelson, Operations Manager.

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

I find the emails helpful which I were sent for the project. Furthermore, I have read all the instructions written on Reading section.

* Do you have any ethical considerations at this stage?

Yes, we should look for both false positive and false negative

* Is my data reliable?

Nearly, the data is reliable. But not perfect.

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

The data should be taken from customers who are loyal to their app. Constant riding history would be a great option. Then we can analyze the data of those customers to correctly predict for any new customer, who will give tip or not

* What data do I have/can I get?

The data based on loyal customers ridding history is ideal.

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

The objective is to predict which customer will not give tip to driver, so that the driver can be alerted, he will not give tip. But in this case, we must look for both false negative and false positive, that’s f1 score. Because predicting that a customer will not give tip, but actually they will, this will make customer to wait for another ride. If predicting that a customer will give tip, but actually they won’t will make driver unhappy and he will less likely to trust on app alerts.

**PACE: Analyze Stage**

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

I visited again the objective. It doesn’t work at all. Instead of considering a customer that he will tip or not, we must predict people who are particularly generous – those who tip 20% or more.

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

Data does break the assumption of our model. There is class imbalance with 1/3 are approx. defined as generous, but that’s acceptable. Since that’s not extreme and may not affect model.

* Why did you select the X variables you did?

Because we want to train the model to predict generosity before the ride, therefore we will select the varaibles for which the data will be available before the ride.

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

The EDA is done ensure to have a cleaned and high quality dataset. We do feature engineering to select relevant of predictive variables, transform and if needed and use for our model. Furthermore, we need to deep dive into data and understand it, therefore we must do EDA.

* What has the EDA told you?

EDA told me that the tip amount is only paid by the customers who paid the ride fare with credit card. So I selected only credit card customers to ensure our model is more accurate.

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

I referred back to my EDA notebooks to help me convert datetime string conversions.

**PACE: Construct Stage**

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

Yes, the performance of both the models, XGBClassifier and RandomForest Classifier, was not satisfactory. Though we could do more feature engineering and more hyperparameter tuning to improve model accuracy.

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

I chose the independent variables like VendorID, Locations IDs pickup and drop off, because these variables could give us the best possible behaviour of customers for tipping . Another variable included mean duration, mean distance etc. For the modified objective, we engineered the tip amount column to calculate the tip percent independent of tip amount. Then the customers, who tipped at least 20% were classified as generous.

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

The model is underfitted, it’s scores are not satisfactory. The model f1 score is 0.35 which is not satisfactory.

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

Yes, We could engineered a column that could identified if the trip distance was short , medium or far. Furthermore, we could do more hyperparameter tuning.

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

I looked for previous notebooks to find coding resources for this project.

**PACE: Execute Stage**

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

The Model was 0.63 accurate in predicting the customers who were generous in giving tip. The model did f1 score for 0.35. We can’t explain the model since due to algorithm is very explain. We should do more exploration for the explanation.

* What are the criteria for model selection?

The criteria for model selection is the best f1 score. Because we want to balance both the false negative and false positives.

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

The model makes sense. This model can be used to help taxi drivers to have a sense if someone will leave a tip or not.

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

Yes, the model could be improved, because our f1 score is very low and there might be some feature engineering available to us to improve model performance. We could we could try creating three new columns that indicate if the trip distance is short, medium, or far.

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

Yes, there were features that were not important at all. We could remove the pickup and drop off locations IDs features and improve our model more generalize.

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

I propose that we can use the model for prediction if urgent, but with highest uncertainty. Furthermore, the stakeholders must invest in improving the model performance.

* 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?