**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’m building the model to predict if a customer tip 20% or more.

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

Finance Department, Operation Department, CEO of the company

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

I have used two datasets 2017\_Yellow\_Taxi\_Trip\_Data.csv and nyc\_preds\_means.csv

* Do you have any ethical considerations at this stage?

Yes at first the model is destined to predict whether a customer will give a tip or not but this question is ethically controversial so we approach it in a different way and more ethical

* Is my data reliable?

Yes, is it extracted from the Automatidata and I have been working with it for 6 courses

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

I want too see more about customer demographic and trip details, historical tipping behavior of customers.

* What data do I have/can I get?

The dataset is fully enough with what I need but if any, add the historical tipping behavior of customers

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

AUC, ROC, F1, Recall, Precision, Accuracy – Focus on F1 because we want to balance between FN and FP so that both customers and our taxi drivers are not feeling satisfied

**PACE: Analyze Stage**

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

No it’s okay now.

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

Not at all

* Why did you select the X variables you did?

Because for me they are the best estimators for our prediction

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

To understand more about the dataset and do some preprocessing such as handling missing values, outliers, duplicates as well as having basic about model assumptions so that we can choose the best predictors for model and transforming data if necessary

* What has the EDA told you?

There’s a lot of unnecessary variables so we can drop them for our model

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

Only Python

**PACE: Construct Stage**

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

The only problem is waiting for GridsearchCV to fit parameters is too long and boring.

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

The instruction told me to choose variables and I feel reasonable, we have enough about necessary to predict the model.

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

Overall, the models appear to perform similarly on both the validation and test sets, which is a positive sign. The precision and F1 scores are also reasonably good, suggesting a balance between correctly identifying positive cases and minimizing false positives.

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

Yes I can improve it. I will do more EDA to understand more about the dataset and reselect the predictors as well as running more on GridsearchCV with different hyperparameters so see difference.

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

I use sk-learn libraries mostly

**PACE: Execute Stage**

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

The models are effective at identifying customers who tip 20% or more.

Models balance between correctly identifying positive cases and minimizing false positives.

* What are the criteria for model selection?

Focused on F1 score to balance false negatives and positives.

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

Models demonstrate reasonable performance and balance but to apply it into real-world we still need to do more.

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

Potential improvement through more feature engineering and ensemble techniques. We also can use ensemble stacking to have the best prediction.

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

Some features might not contribute significantly to predictions.

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

Recommend using the models for targeted promotions or incentives.

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

Optimize fare, allocate drivers, segment customers.

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

NYC taxi datasets, preprocessing techniques, model evaluation.

* Is my model ethical?

The model addresses ethical concerns by predicting likelihood of tipping, not amount.

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

Mistakes could be misclassifying customers, affecting driver earnings and customer satisfaction.