**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?

The goal is to develop a machine learning model that can determine whether or not a customer will leave a tip.

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

The external stakeholders are the NYC TLC customers that the changes this makes to taxi operations will effect.

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

Jupyter Notebooks, the Automatidata dataset.

* Do you have any ethical considerations at this stage?

Ensuring the resulting model does not encourage drivers to pass over customers who will not tip.

* Is my data reliable?

Yes. It is provided by the NYC TLC and has all of the relevant data.

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

Behavioral data, which can give a more in-depth analysis than details like datetime and fare amount.

* What data do I have/can I get?

Data like datetime and fare amount, which is a good place to start.

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

We do not have enough information at this stage to determine the correct metric to use.

**PACE: Analyze Stage**

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

No. There are too many potential problems creating a model to tell who tips and who does not. A better approach is determining who is a generous tipper.

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

No, the data conforms to the assumptions of the model.

* Why did you select the X variables you did?

Because time of day and trip distance are the most likely factors to influence whether or not a customer will be a generous tipper.

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

To create columns relevant to the search for generous tippers and drop irrelevant columns from the frame.

* What has the EDA told you?

There is a slightly higher percentage of generous tippers than non-generous tippers.

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

The pandas and numpy libraries in Jupyter Notebooks.

**PACE: Construct Stage**

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

Nothing is out of the ordinary in the model.

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

The chosen independent variables were mean duration and distance and predicted fare, as well as several custom columns that noted rush hours. These were the variables most likely to affect whether or not a customer is likely to leave a generous tip.

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

The model fits the data fairly well. The validation score is 0.71.

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

The score can certainly be improved. The main thing to change about the model is the number of false positives.

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

The sklearn library, mainly packages such as model\_selction and metrics.

**PACE: Execute Stage**

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

The key features that lead to generous tipping are travel with Vendor 2 and the predicted fare. Sadly, we do not know how they affect tipping, due to the lack of transparency in the random forest method.

* What are the criteria for model selection?

F1 score, accuracy, recall, and precision. If the F1 score is greater than 0.65, the model is acceptable. The model with the highest score is the best choice.

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

The model makes sense and has acceptable results.

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

The model could be improved, mainly by reducing the number of false positives.

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

Several features, such as time of day and month of the year, had virtually no effect on the outcome. Taking them may skew the results.

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

I would recommend looking into the effect Vendor 2 has on the odds of leaving a good tip.

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

Whether or not removing the irrelevant features will refine the results.

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

The results of the machine learning models.

* Is my model ethical?

It is not the most ethical, as it may still lead to drivers showing bias in their customer selection based on the model’s prediction of who will and won’t be a generous tipper.

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

Mistakes in this model mean predicting someone leaving a good tip when they will not, and vice versa. This can lead to disappointment for drivers expecting a good tip.