**Course Six**

# The Nuts and Bolts of Machine Learning



# Instructions

Use this PACE strategy document to record decisions and reflections as a data professional 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:

* Understand and assess the proposed workplace scenario
* Demonstrate the ability to use a notebook environment to create a series of machine learning models on a dataset to solve a problem
* Articulate findings in an executive summary for external stakeholders

# Relevant Interview Questions

Completing the end-of-course project will empower you to be 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: Planning Stage**

* What am I trying to solve?

- find ways to generate more revenue for taxi cab drivers.

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

- Juliana Soto, Finance and Administration Department Head

- Titus Nelson, Operations Manager

* What are you trying to solve or accomplish?

- A model which can predict whether or not a customer is a generous tipper.

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

- 2017\_Yellow\_Taxi\_Trip\_Data

- XGBoost, random forest, algorithms-

* Do you have any ethical considerations at this stage?

- limiting equal access to taxis is ethically problematic, and carries a lot of risk.

* Is my data reliable?

It is somewhat reliable however it can be better.

I had already prepared much of this data and performed exploratory data analysis (EDA) in previous courses.

Ideally, we'd have behavioral history for each customer, so we could know how much they tipped on previous taxi rides. We'd also want times, dates, and locations of both pickups and dropoffs, estimated fares, and payment method.

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

Engineering additional features.

For example creating three new columns that indicate if the trip distance is short, medium, or far. We could also engineer a column that gives a ratio that represents (the amount of money from the fare amount to the nearest higher multiple of $5) / fare amount. For example, if the fare were $12, the value in this column would be 0.25, because $12 to the nearest higher multiple of $5 ($15) is $3, and $3 divided by $12 is 0.25. The intuition for this feature is that people might be likely to simply round up their tip, so journeys with fares with values just under a multiple of $5 may have lower tip percentages than those with fare values just over a multiple of $5. We could also do the same thing for fares to the nearest $10.

* What data do I have/can I get?

There are two data-frames: one containing the original data, the other containing the mean durations, mean distances, and predicted fares from the previous course's project called nyc\_preds\_means.csv.

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

This is a supervised learning, classification task.

We could use accuracy, precision, recall, F-score, area under the ROC curve, or a number of other metrics. However, we don't have enough information at this time to know which are most appropriate. We need to know the class balance of the target variable.

**PACE: Analyzing Stage**

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

- Build a model that predicts if a customer will **not** leave a tip. It will not work because it is ethically sensitive and might create some **undesired exclusion** by limiting equal access to some customers. Thus, indeed the request to be revised to make it **ethically sound** such as for example model that predicts the **most generous** customers.

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

The model is twice as likely to predict a false negative than it is to predict a false positive

The F1 score is ~0.04 higher than the random forest model. Both models are unsatisfactory, though. Not acceptable.

* Why did you select the X variables you did?
* What are some purposes of EDA before constructing a model?

EDA plays a crucial role in understanding the data, identifying data quality issues, selecting features, detecting patterns, and making informed decisions during the modeling process. It helps ensure that the data is suitable for modeling and enhances the accuracy and reliability of the subsequent analyses.

* What has the EDA told you?

Now I know that that customers who pay **cash** generally have a tip amount of $0. To meet the modeling objective, you'll need to sample the data to **select only** the customers who pay with **credit card.**

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

- Feature engineering,

- Variable encoding

- Evaluation metrics

**PACE: Constructing Stage**

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

If you get a warning that a metric is 0 due to no predicted samples

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

The variable generous was chosen.

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

A random forest classifier was used.

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

think of final model selection as another way of "tuning" your model.

All scores increased by < 0.01.

Try to improve your scores using an XGBoost model.

The F1 score is ~0.04 higher than the random forest model. Both models are unsatisfactory, though.

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

I used the features/target variable, training and testing sets, a dictionary cv\_params with several hyper-parameters, another dictionary scoring scoring metrics, use pickle to save your models and read them back in, Use the make\_results() function to output all of the scores of your model, new variable called preds. Use the plot\_importance function to inspect the top 10 most important features of your final model.

**PACE: Execute Stage**

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

-  duration, trip\_distance, and fare\_amount are the **most important features.**

* What are the criteria for model selection?

To have a decent amount of accuracy by having the minimum quantity of false negatives and false positives.

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

This is not a great model, but depending on how it's used it could still be useful. If the objective is only to help give taxi drivers a better idea of whether someone will leave a good tip, then it could be useful. It may be worthwhile to test it with a select group of taxi drivers to get feedback.

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

It would probably be very helpful to have past tipping behavior for each customer. It would also be valuable to have accurate tip values for customers who pay with cash. It would be helpful to have a lot more data. With enough data, we could create a unique feature for each pickup/dropoff combination.

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

- **tpep\_pickup\_datetime**

**- tpep\_dropoff\_datetime**

**- RatecodeID**

- store\_and\_fwd\_flag

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

- If the objective is only to help give taxi drivers a better idea of whether someone will leave a good tip, then it could be useful. It may be worthwhile to test it with a select group of taxi drivers to get feedback.

- we could try creating three new columns that indicate if the trip distance is short, medium, or far. We could also engineer a column that gives a ratio that represents (the amount of money from the fare amount to the nearest higher multiple of $5) / fare amount.

-It would probably be very helpful to have past tipping behavior for each customer. It would also be valuable to have accurate tip values for customers who pay with cash. It would be helpful to have a lot more data. With enough data, we could create a unique feature for each pickup/dropoff combination.

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

**What was your model doing? Can you explain how it was making predictions?**

**Are there new features that you can engineer that might improve model performance?**

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

Unfortunately, XGBoost is not the most transparent machine learning algorithm. We know that duration, trip\_distance, and fare\_amount are the most important features, but we don't know how they influence tipping. This would require further exploration.

* Is my model ethical?

it's better for a driver to be pleasantly surprised by a generous tip when they weren't expecting one than to be disappointed by a low tip when they were expecting a generous one. However, it's unlikely that this model would be deployed further development to significantly improve its performance.

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

Unfortunately, XGBoost is not the most transparent machine learning algorithm. We know that duration, trip\_distance, and fare\_amount are the most important features, but we don't know how they influence tipping. This would require further exploration.