# PACE Strategy Document

## Introduction

This PACE strategy document can be used to record decisions and reflections as you work through this end-of-course project. You can use this document as a guide to consider your responses and reflections at different stages of the data analytical process. Additionally, the PACE strategy documents can be used as a resource when working on future projects.

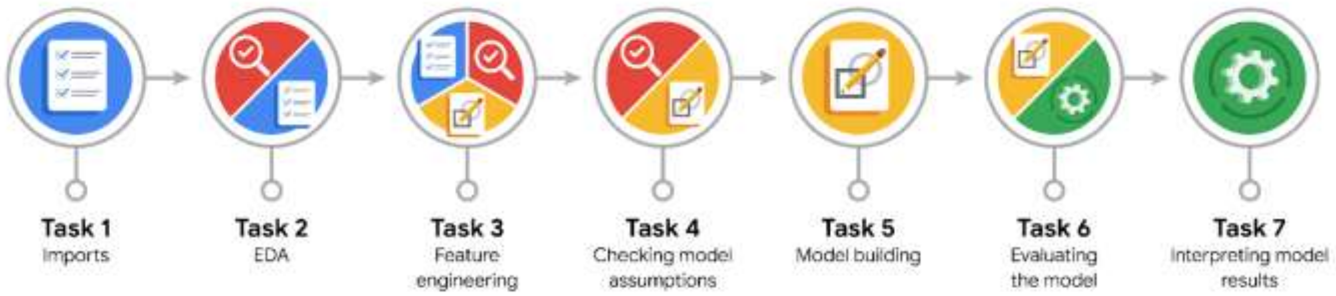
## Relevant Interview Questions

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?

To identify which variables or factors influence the amount of gratuity a rider gives a driver and thus make informed business decisions that will increase gratuities and subsequently improve driver satisfaction.

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

Juliana Soto and Udo Bankole

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

Since I'll use Random Forest Algorithm, I would need libraries like pandas, numpy, RandomForest, etc

- Is my data reliable?

Yes

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

Cash tips data should be available to conduct a full proof analysis.

- What data do I have/can I get?

The existing data is all I have currently. I would like to request cash tips data.

- What metric should I use to evaluate the success of the business/organizational objective? Why?

F1 score is good for this as this is a classification task with imbalanced dataset.

### PACE: Analyze Stage

- Revisit "What am I trying to solve?" Does it still work? Does the plan need revising?

Analyzing the ethical considerations, the objective is modified from whether a customer will give a tip or not to whether a customer will give a generous tip or not (above 20%).

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

It is not perfect but currently we have to move forward until we get the cash data too.

- Why did you select the X variables you did?

They only seemed important for our objective.

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

To analyze the most important fields and to remove unnecessary data and to get a sense of data. To handle outliers, remove errors and inspect relationships between variables.

- What has the EDA told you?

Approximately 1/3 of the customers in this dataset were "generous" (tipped ≥ 20%). The dataset is imbalanced, but not extremely so. So, we should consider F1 score for evaluation.

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

Datetime, pandas

## PACE: Construct Stage

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

Yes, there is no cash tips in the dataset. For now, it can be fixed by only including trips with payment type as credit card.

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

VendorID, passenger_count, mean_distance, mean_duration, predicted_fare, am_rush, daytime, pm_rush, nighttime, RatecodeID, PULocationID, DOLocationID.

- How well does your model fit the data? What is the model's validation score?

The model gives a low recall and in turn a low f1 score so it doesn't fit well. The score is 0.39

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

Yes I can include features like (actual time taken / mean time for those location pairs), tax percentage in total amount, etc. Also reduce less correlated features like Location ID columns.

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

Sklearn, XGBoost, pandas, matplotlib.

## PACE: Execute Stage

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

The tip amounts are highly dependent on mean duration and distance and passenger count.

- What are the criteria for model selection?

To have a high f1 score.

- Does the model make sense? Are the results acceptable?

No, the results are not acceptable as the recall score is too low.

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

To a limit yes by including features like (actual time taken / mean time for those location pairs), tax percentage in total amount, etc. But still there are untracked but really important features like behaviour of the driver, temperature inside the car, previous history of rider, etc.

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

Yes, features like Location ID columns. They could be re-engineered to be included properly.

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

To try to invent ways to reduce trip duration, maybe use better navigation. But it would be best to record more factors as told previously in trip to get better recommendations.

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

  Whether a driver would pick a customer or not.

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

  matplotlib

- Is my model ethical?

  No since it only includes trips with credit card payments.

- When the model makes a mistake, what happens? How does that translate to the use case?

  False positives are bad for cab drivers, because they would pick up a customer expecting a good tip and then not receive one, frustrating the driver.

  False negatives are bad for customers, because a cab driver would likely pick up a different customer who was predicted to tip more—even when the original customer would have tipped generously.