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

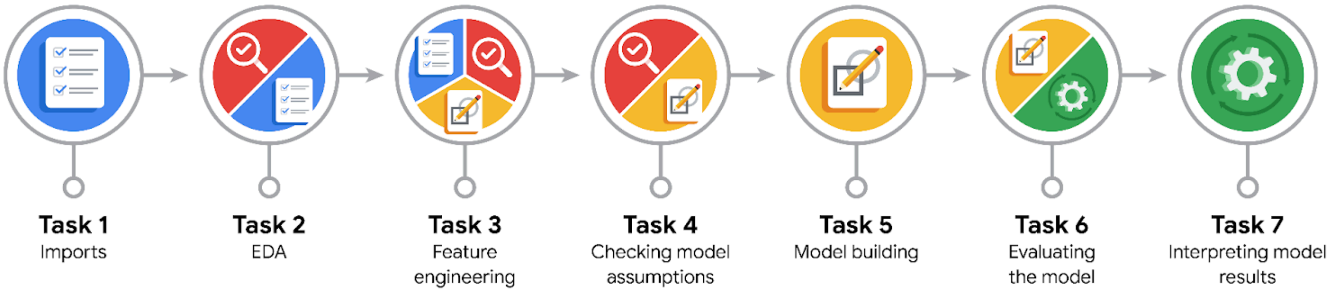
Relevant Interview Questions

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?

Developing the machine learning model(s) to predict whether the riders are generous tippers (giving tips >= 20%).

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

The New York City TLC team members.

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

Jupyter Notebook.

- Do you have any ethical considerations at this stage?

There are no ethical considerations at this stage.

- Is my data reliable?

Overall it is reliable, but there are some concerns to fix such as the negative values that have been recorded.

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

Could be the data with things that have been solved from the EDA process, preprocessing, and feature engineering. This will give the data more reliable, and the confidence to develop the model(s) with high performance.

- What data do I have/can I get?

The given dataset is from the NYC TLC.

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

Metric tries to balance the FP and FN values, which is the F1-score.

PACE: Analyze Stage

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

It’s still worked, and there’s no need to revise the plan.

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

The data doesn’t break any assumptions of the model.

- Why did you select the X variables you did?

Selecting X variables that account for the most influential to the target variable. Accounting too much or too little can affect the performance of the model which either oversimplifies or is more complex.

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

EDA is important because we want a reliable dataset first (through cleaning, handling missing values, outliers, etc.). Checking variables that account for the best influence, removing the irrelevant ones, and creating more features from the existing ones. Also, checking assumptions is required to determine whether the model might be a good fit.

- What has the EDA told you?

There are no missing values in the dataset, and since the Random Forest model can handle outliers well. There are some new features created to boost the influence on the target variable through feature engineering, as well as feature selections and transformations.

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

Jupyter Notebook for Python coding.

PACE: Construct Stage

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

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

These columns include the target variable will include training the model:

VendorID

1 passenger_count

2 RatecodeID

3 PULocationID

4 DOLocationID

5 mean_duration

6 mean_distance

7 predicted_fare

8 generous (target variable)

9 day

10 am_rush

11 daytime

12 pm_rush

13 nighttime

14 month

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

The XGBoost model produces moderate results, with a precision score is 0.69, a recall score is 0.82, an f1 score is 0.75, and an accuracy score is is 0.72.

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

We might request the NYC TLC to collect more data and create more predicted features to boost the model's performance.

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

Jupyter Notebook for Python coding.

PACE: Execute Stage

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

Both the Random Forest and XGBoost were accounted for creating, and XGBoost was chosen with a bit increase in margin. The model incorrectly predicts the high proportion of false positives (which is the rider will leave a tip >=20%, but it's less). Conversely, the false negative is low (which is the rider will leave a tip <20%, but it's more). It might be desirable for the drivers will get a surprised tip, rather than giving a tip they don't expect beforehand. Some features such as predited_fare, VendorID_2, mean_distance, mean_duration, etc. have the most influence on whether the riders will give a tip >=20%. This XGBoost model is referred to as a black-box model since it won't be explainable at all. For instance, we wouldn't know how the top features influence the target variable.

- What are the criteria for model selection?

Some questions can be asked when choosing the best model: How explainable is it? How much time will it take to train? How complex is it? Does it perform well on the unseen data? etc .

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

All the score metrics are within the range of >=70%, which is acceptable to use in this model.

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

The best approach is to collect more and more data to hopefully boost the model's performance. We might account for the relevant features, and remove the redundant ones. Feature engineering is essential, especially when creating new features that can tremendously improve the model.

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

There could be some features that won't account for too much influence after the model has been built, by looking at the feature's importance. We might account for keeping the top % of the features, remove those that bring little or no improvement to the model's performance, and keep the model to be much simpler.

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

Based on the model's results, even though there could be concerns about making incorrect predictions, it won't account for too much of the severity. We might recommend deploying the model, closely monitoring how it works, and getting feedback from the drivers for further improvements.

- 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?

Jupyter Notebook for Python coding.

- Is my model ethical?

Since the model doesn't break any assumptions, and accounts for the incorrect predictions with the proportion isn't too severe, we might choose to deploy this model.

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

In this case, we might deal with the false positive and negative values. FP is when the model incorrectly predicts the given tip is <20% as >=20%. This might lead to the disappointment of drivers since they would expect a higher tip, but it won't. FN is when the model incorrectly predicts the given tip is >=20% as <20%. This might lead drivers to not accept to take these lower-paying tips and choose the higher ones.