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

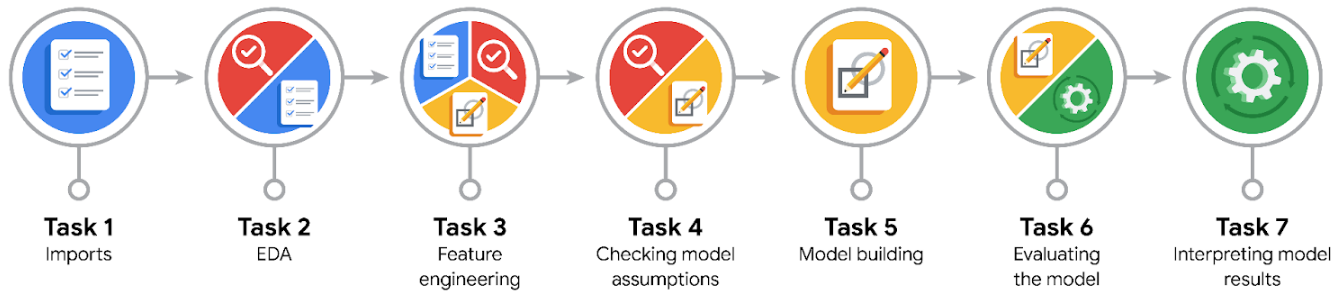
Relevant Interview Questions

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.



Task 1 Imports

Task 2 EDA

Task 3 Feature engineering

Task 4 Checking model assumptions

Task 5 Model building

Task 6 Evaluating the model

Task 7 Interpreting model results

Data Project Questions & Considerations

PACE: Planning Stage

- What are you trying to solve or accomplish?

  Predict whether the customer will leave a tip higher than 20% of the trip's value.

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

  I will use the data disposed, the Jupyter Notebook environment, several libraries from for operations, preparations, to produce the model with Sklearn and the boost which is XGB classifier.

- Do you have any ethical considerations at this stage?

  Whether the model will cause discrimination among users and affect people who can and cannot afford a generous tip.

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

  We could use accuracy, precision, recall, F-score, area under the ROC curve, or a number of other metrics, but currently there is not enough information to know which is the most appropriate.

PACE: Analyzing Stage

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

The approach of the task would be redirected to determine those customers who are considered generous clients, so they will have that feature, something that does not affect other clients.

- Why did you select the X variables you did?

Because they have significant relevance with the objective variable, so there were deleted unnecessary variables such as taxes.

- What has the EDA told you?

There are 2 times more “no generous” customers than generous customers. Additionally, it seems that precision could be the best metric to evaluate the model.

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

I would be mainly using pandas and the help of the datetime package.

PACE: Constructing Stage

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

Yes, there were string variables and variables that needed to be booleans to become part of the model. To solve this problem, the date variables were changed and distributed into schedules and the function get_dummies was utilized.

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

I focused on the location, time, fare, and other factors because of their proximity with a potential tip.

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

At the beginning of the process the model was poor in performance, having a validation score of: 0.68.

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

With the application of Xgboosters the model would improve its metrics. However, after applying boosters the model won’t have an explicit logic criteria to define understandable patterns.

PACE: Execute Stage

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

The cost, trip distance, and trip duration were the most relevant factors for tips higher than the 20% of the trips value. Additionally, the model was modestly precise, which means that has a good performance, but it is propense to errors.

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

Yes, It would be useful to analyze the history of the client and the exact amount that was used for a tip, there can be some patterns such as a maximum tip per trip regardless of the distance and duration of the trip, which means that there could be a roof in the curve.

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

The hour and location of the trip had almost no influence on the model. It would be useful to clean the location because it increases the size of the dataset. On the other hand, the hour was more useful and had a good effect on determining trends.

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

I would not recommend this model because it is not as precise as it can be expected. However, if the results are taken into account I would strongly encourage to display the precision of the model, so the drivers would understand its limits. Additionally, further analysis in the response of the drivers is recommended to evaluate the performance of the model among users.

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

How classification models such as clusters could be applied to solve the problem, and could they perform better than the current model?