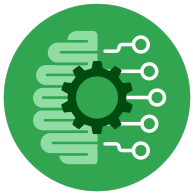
Course Six
The Nuts and Bolts of Machine Learning



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

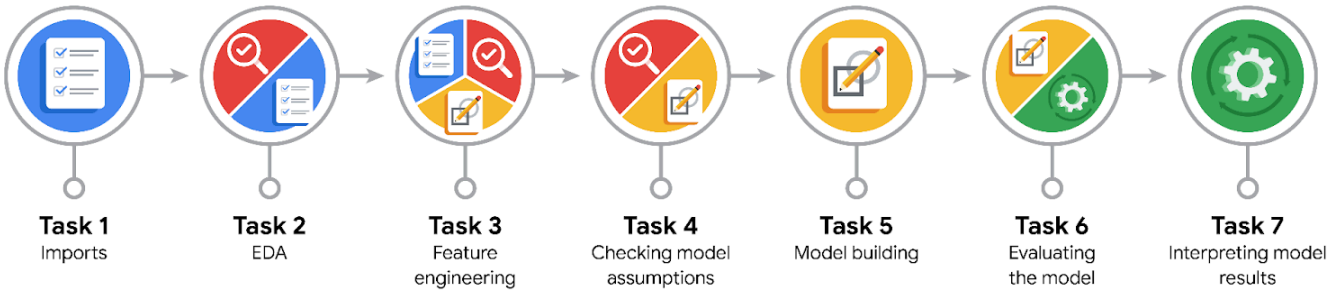
Relevant Interview Questions

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?

  I am trying to make a random forest model to predict how generous a tip amount could be for taxicab drivers.

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

  The external stakeholder is TLC.

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

  Operational libraries like pandas and numpy.

- Do you have any ethical considerations at this stage?

  If the model predicts a large number of false positives, taxicab drivers could become frustrated by selecting customers in hopes of getting larger tips and not receiving them. Certain demographics could be favored for larger tips, which is unethical in providing equal transport for all customers.

- Is my data reliable?

The data contains many features that contain target leakage. Other features need to be transformed or added to the dataset to be predictive of the target.

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

I would need data on trip reviews from the customer. Higher reviews could lead to better tips.

- What data do I have/can I get?

I have data on average trip distance and average trip duration. There are predicted fares to guide how tips can be produced that do not cause data leakage.

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

The model should focus on producing a low amount of false positives and false negatives, so both precision and recall are important. The metric that balances these metrics is f1 score.

PACE: Analyze Stage

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

Solve: random forest model Works: the model achieves higher than 65% f1 score Revising: more features that are predictive of the target need to be added

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

It is okay if the linearity assumption is broken. A random forest model should limit variance.

- Why did you select the X variables you did?

Features that described a taxi drip and did should result in data leakage were selected.

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

EDA is used to remove or impute missing values, change data types, and examine feature distributions.

- What has the EDA told you?

The EDA shows both numeric and categorical data.

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

Pandas, datetime

PACE: Construct Stage

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

The classes of the target variable are unbalanced. The target can be stratified. Irrelevant features can be dropped. Categorical features can be dummy encoded.

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

Average trip distance, average trip duration, vendor location, dropoff and pickup locations, and number of passengers were selected. These factors seemed like they would forecast customer tipping behavior based on trip conditions.

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

The model achieved an f1 score of 0.717 on the training data. The model achieved an f1 score of 0.702 on the test data.

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

Add more features that are predictive of the target. Split trip distance and trip duration into upper and lower thresholds to determine their effect on tipping.

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

Machine learning libraries like scikit-learn and xgboost. Visualization libraries like matplotlib.

PACE: Execute Stage

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

  The model shows that vendor ID had a substantial impact on removing impurity from the trees. The model weights assignments are not directly explainable.

- What are the criteria for model selection?

  The criteria involved in model selection are accuracy score, precision score, recall score, and f1 score.

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

  The model shows that factors like vendor id, mean trip duration, and passenger count contributed to how much of a tip a driver received. That is sensible. The model performs better than a random classifier, so the results are acceptable, but they can be improved.

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

  The model can be improved. Additional features can be added to the model. The log scale of continuous features can be taken for skewed distributions.

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

  Features that caused data leakage, like total_amount and payment_type were removed. The model score decreases, but this leads the model to predicting real-world data better.

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

  TLC should provided more data on why certain vendors and locations result in higher tips.

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

  What social impacts contribute to higher tips? What events do customers take part in that lead to higher tips? Are the customers tourists?

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

  Matplotlib, scikit-learn

- Is my model ethical?

  The model contains a relatively low amount of false positives and has a balanced f1 score. The model is fair for both the customer and driver.

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

  The highest error the model makes is a Type I error. This means it incorrectly predicts that a taxicab driver will receive a higher tip. This would lead to driver occasionally being upset that a customer did not leave a larger tip.