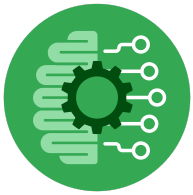
Course Six
The Nuts and Bolts of Machine Learning



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

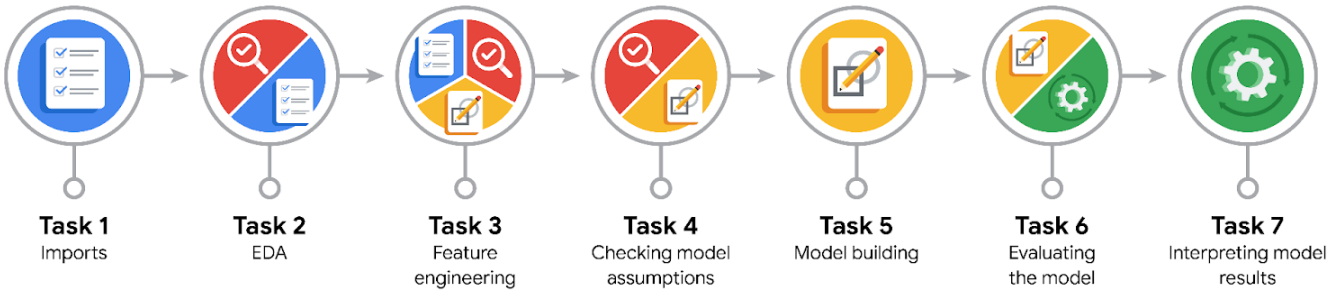
Relevant Interview Questions

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?

  The aim of the project is to build a machine learning model that can classify TikTok video content into claims and opinions.

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

  Apart from the data science team, the resource management team, quality assurance team etc. will have an interest in the outcome of the project.

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

  Python packages

- Do you have any ethical considerations at this stage?

  False positives are acceptable for this project whereas it is critical to minimize false negatives as much as possible.

- Is my data reliable?

The dataset is reliable as it has been through several iterations of EDA by the data science team and no significant issues have been raised.

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

Recall would be a good metric to evaluate the model as it gives an idea of false negatives predicted by the model.

PACE: Analyze Stage

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

Yes, the data seems to have sufficient predictive power to build an effective model.

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

No, the data does not break any of the model assumptions e.g. multicollinearity.

- Why did you select the X variables you did?

The x-variables selected represent the user engagement with videos. This has been shown in the EDA stage to have the most predictive power.

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

To check that the variables selected are appropriate, and that the data does not violate the model assumptions.

- What has the EDA told you?

The EDA has confirmed that the variables that represent user engagement with videos (e.g. view, like, share counts) have the most predictive power to build the model.

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

Python packages

PACE: Construct Stage

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

No issues have been noticed in the data or model so far.

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

Variables representing user engagement with the videos e.g. video view count, like count, share count, comment count and download count. Also, the length of the video transcription text was included as a feature.

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

The model fits the data very well with a validation recall score of 0.99.

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

As the model performs near perfect, there is no need to improve the performance.

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

Python packages.

PACE: Execute Stage

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

  The model has confirmed that variables representing user engagement with videos, can predict very well whether a video is a claim or an opinion.

- What are the criteria for model selection?

  A random forest model was chosen as a tree-based model was suggested by the data science team after the EDA.

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

  Yes, the model fits the data very well and gives an excellent performance on the test hold-out data.

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

  As the model performs near perfectly, there is no need to improve it further.

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

  Video view count and video like count are the most important features. Then with less importance, video share count and video download count also contribute to the model. Other features do not seem to contribute to the model.

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

  User engagement with the videos need to be assessed with the model, then 'suspicious' videos can be sent to reviewers/moderators to make the final classification into claims or opinions.

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

  Python packages

- Is my model ethical?

Yes the model has addressed the ethical implications, as described in the next question.

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

False positives of the model are acceptable, as this can then be classified correctly by the moderators/reviewers. However, false negatives mean that claim videos are not detected by the model - this is not acceptable. The model developed in this project have a low value for false negatives, so the model performs well.