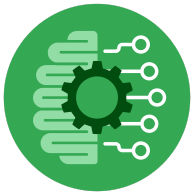
Course Six
The Nuts and Bolts of Machine Learning



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~~

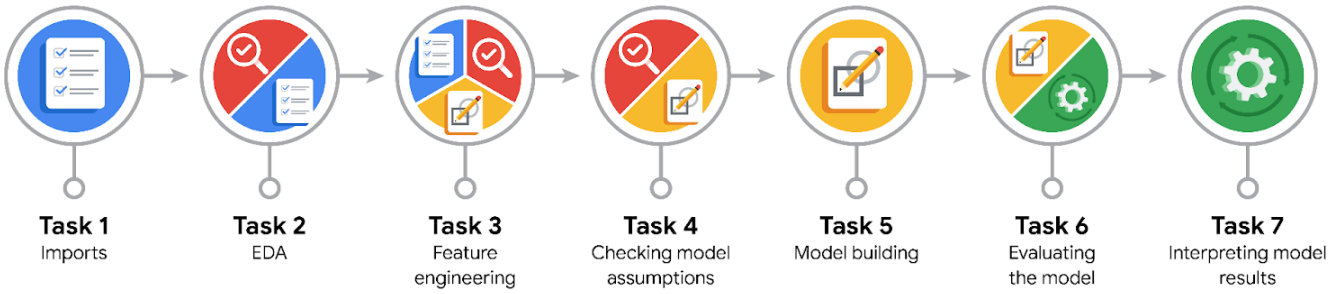
Relevant Interview Questions

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?

Build a machine learning model to help identify claims and opinions.

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

TikTok.

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

Packages for data manipulation.
Packages for data visualization.
Packages for data preprocessing.

- Do you have any ethical considerations at this stage?

It's very important to identify videos that break the terms of service, even if that means some opinion videos are misclassified as claims.
- The worst case for an opinion misclassified as a claim is that the video goes to human review.
- The worst case for a claim that's misclassified as an opinion is that the video does not get reviewed and it violates the terms of service.
- A video that violates the terms of service would be considered posted from a "banned" author.
- Because it's more important to minimize false negatives, the model evaluation metric will be recall.

- Is my data reliable?

Yes.

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

Claims and opinions data for primary and other data for help predict.

- What data do I have/can I get?

Video data.

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

accuracy_score, precision_score, recall_score, f1_score.

PACE: Analyze Stage

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

It’s still working.

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

Some columns is text-based. It is not a categorical variable, since it does not have a fixed number of possible values. Need to extract numerical features.

- Why did you select the X variables you did?

Because it's help predict target feature.

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

For prepare data that can impact to model.

- What has the EDA told you?

Tree-based models are robust to outliers, so there is no need to impute or drop any values based on where they fall in their distribution.

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

Packages for data manipulation.
Packages for data visualization.
Packages for data preprocessing.

PACE: Construct Stage

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

No.

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

Claims and opinions for predict that video status.

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

The model performs exceptionally well, with an average recall score of 0.995 across the five cross-validation folds.

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

The model currently performs nearly perfectly, there is no need to engineer any new features.

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

Packages for confusion matrix, model_selection, ensemble, xgboost.

PACE: Execute Stage

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

Both model architectures—random forest (RF) and XGBoost—performed exceptionally well. The RF model had a better recall score (0.995) and was selected as champion.

- What are the criteria for model selection?

Model score.

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

Yes, this model performed well on both the validation and test holdout data.

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

The current version of the model does not need any new features. However, it would be helpful to have the number of times the video was reported. It would also be useful to have the total number of user reports for all videos posted by each author.

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

The feature video_transcription_text is text-based. It is not a categorical variable, since it does not have a fixed number of possible values. One way to extract numerical features from it is through a bag-of-words algorithm like CountVectorizer.

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

Before deploying the model, the data team recommends further evaluation using additional subsets of user data. Furthermore, recommends monitoring the distributions of video engagement levels to ensure that the model remains robust to fluctuations in its most predictive features.

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

The model's most predictive features were all related to the user engagement levels associated with each video.
It was classifying videos based on how many views, likes, shares, and downloads they received.

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

Packages for confusion matrix.

- Is my model ethical?

The worst case for an opinion misclassified as a claim is that the video goes to human review.
The worst case for a claim that's misclassified as an opinion is that the video does not get reviewed and it violates the terms of service.
A video that violates the terms of service would be considered posted from a "banned" author.
Because it's more important to minimize false negatives, the model evaluation metric will be recall.

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

It would be helpful to have the number of times the video was reported. It would also be useful to have the total number of user reports for all videos posted by each author. For help improve the performance of model.