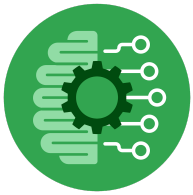
Course Six
The Nuts and Bolts of Machine Learning



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

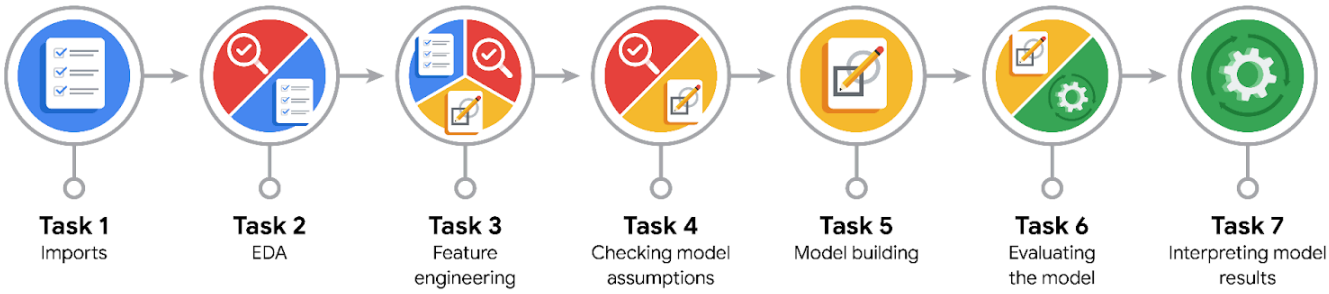
Relevant Interview Questions

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?

  We are trying to eliminate 100% reliance on human moderators when classifying video submissions on Tiktok as 'claim' or 'opinion'.

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

  There are no external stakeholders. This project is completely internal trying to improve the efficiency and effectiveness of the TikTok moderators and application.

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

  At the planning stage we find ourselves relying on the input of stakeholders and the limitations of the data that has been collected.

- Do you have any ethical considerations at this stage?

  The ethical concerns at this time is whether there are data based trends that will lead toward solutions containing baises, which will unfairly categorize certain demographics/groups/races.

- Is my data reliable?

  The data is a primary data source. It is empirical data that has been collected from observing TikTok users' submissions and the actions taken on their videos and accounts. It is very reliable data.

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

In a perfect world there would be one or two features that will allow for the perfect prediction of whether a post is a 'claim' or 'opinion'

- What data do I have/can I get?

We can get all of the primary data collected from the application. This is all historic data; however, new data is produced every day and could be incorporated into the model building and testing process as sets of reasonable size become available.

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

We would likely value the metric recall the most, since the 'cost' of a false negative, marking a submission as an 'opinion' would be high (while it is a claim). It would increase the work of the moderators to find a bad post and during that time this post that was marked incorrectly as an 'opinion' could be breaking terms of service.

PACE: Analyze Stage

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

The building of a machine learning model to categorically predict whether a post is a 'claim' or an opinion. With a successful model built we will be able to increase the efficiency of moderation on tiktok, while reducing required human interaction with each post.

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

There are quite a few features that are showing colinearity; however, if we are using a decision tree or random forest model, we should be able to move past the assumption of non-colinearity. Additionally, the data is heavily imbalanced, with few 'opinion' posts versus 'claim'. This will have to be managed through upsampling.

- Why did you select the X variables you did?

I dropped x variables that were unrelated, such as video_id, and through data cleaning processing I kept all other available data, including the categorical, which I did drop_first to keep the data set slightly cleaner.

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

The purpose of eda prior to constructing the model is to ensure that the assumptions are met and to better understand the shape of the data, as well as to glean statistical and surface level correlations and insights.

- What has the EDA told you?

EDA has shown that after cleaning the data of its outliers, we are left with an mostly balanced data set, with the length of transcripts being slightly longer for opinion videos.

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

During the analyze and EDA I found myself using pandas, matplotlib, and seaborn. I found myself really thinking whether it was better eliminate data or keep it when dealing with outliers.

PACE: Construct Stage

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

Nothing stood out as being particularly odd. One of the main issues I had was actually deciding which model was the champion, both the Random Forest and the XGB were performing extremely well and nearly identical.

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

I chose to keep video_duration_sec, video_view_count, video_like_count, video_share_count, video_download_count, video_comment_count, transcription_text_length, verified_status_verified, author_ban_status_banned, author_ban_status_under_review. The importance of each feature at this point could not be determined and all of them logically could be related to the status of a 'claim' or 'opinion'

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

The Random Forest model fit the data very well. I was worried that it would over fit. However, in the case of the validation and the testing data the model did a extremely good job predicting the target outcomes. The validation scores for both models for precision are 1.00

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

There is very little that can be improved upon with these scores. However, we could bring in other features, such as age of accounts, age of users and possible get better perspective of the trends.

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

This part of the project used scikit learn, implementing, training and reviewing the machine learning models of Random Forest and XGBoost.

PACE: Execute Stage

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

From the models we found that it is nearly 100% possible to correctly categorize submissions on features of the post and user, while using different ML models used different features to get to nearly the identical conclusion and prediction consistency.

- What are the criteria for model selection?

  The criteria for model selection was Recall and then results were sorted by max mean precision score.

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

  The results make sense and are highly accurate. They follow features that indicate high engagement by the community, which likely means that the posts are more likely to be causing upset.

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

  Honestly it would be difficult to improve this model. There might be some ways of cleaning up the data by feature engineering, such as words per second.

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

  There were a couple features with very little importance. They could be potentially eliminated, but I would need to test and compare the results to make sure it is not affecting the outputs.

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

  I would recommend to start integrating this model as soon as possible. With the high accuracy of the model and all metrics, we will likely see a smooth integration and an increase in moderator efficiency.

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

  We could also answer questions related to banned status and verification status. Especially when looking from the lens of claims/opinion we could maybe base whether a user should be verified, or banned, etc

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

  At this stage, I was using scikit learn, xgboost is from its own library. It was used in the construct phase as well. This stage also used pandas and matplotlib to manipulate and visualize the importance of the features after the models were finalized.

- Is my model ethical?

  The model appears to be ethical. We are not bringing any age, sex, race, etc demographic into the determination of the prediction. This sort of information might be interesting to reveal to better understand the users. However, the model itself is free of specifically targeting any known demographic purposefully. Without looking at these demographics with the model we will not know whether any hidden bias has snuck into the model.

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

  The most common mistake is a false negative. The number of occurrences is very low; however, these are the most important to avoid. Tuning the model to try to eliminate their occurrence could be an important place to invest, since false negative can cause content that is against terms of service to remain accessible to the community.