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

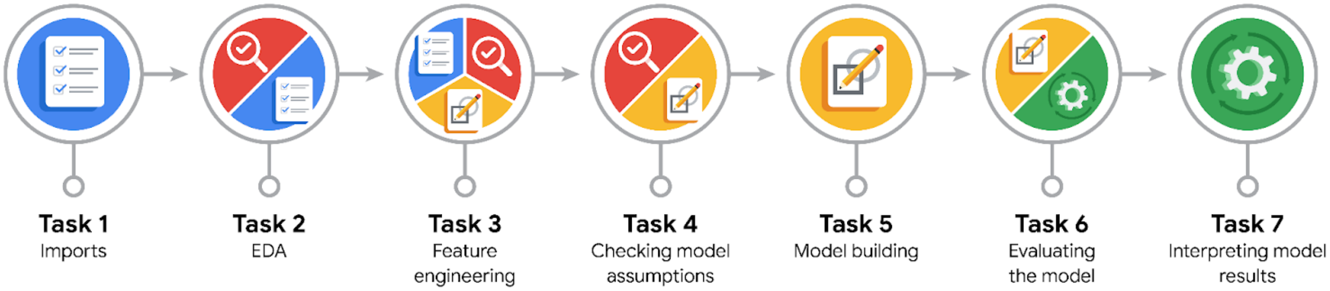
Relevant Interview Questions

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 TikTok team is trying to develop a machine learning model to assist in the classification of videos as either claims or opinions. In this specific portion of the project, the model is to be built and evaluated.

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

External of the company are TikTok users and external of the data team is the operations team.

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

At this point in the process, we are simply importing our data and thinking through the process.

- Do you have any ethical considerations at this stage?

If our model often incorrectly assigns claim videos as opinions, then it's possible that some videos that violate TikTok's terms of service will not be chosen for further review. With this thought, we know that we don't want the model to predict false negatives, meaning our main metric for evaluation will be recall.

- Is my data reliable?

Yes, the data is from TikTok and has been reviewed

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

You need data on the video claims/opinions and other features of those videos so that we can build a prediction model.

- What data do I have/can I get?

The data from TikTok which gives claim status and other identifying features of videos

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

If our model often incorrectly assigns claim videos as opinions, then it's possible that some videos that violate TikTok's terms of service will not be chosen for further review. With this thought, we know that we don't want the model to predict false negatives, meaning our main metric for evaluation will be recall

PACE: Analyze Stage

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

The plan still works after doing a quick overlook of the data.

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

No, since we’re using tree-based modeling there are no assumptions to be met

- Why did you select the X variables you did?

All of the variables seemed somewhat correlated with the outcome variable.

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

Allows you to clean the data: remove duplicates, deal with outliers, reformat data, check that the data meets any assumptions, etc.

- What has the EDA told you?

The target variable has about equal counts in each category (claims vs opinions) which means the class is balanced which is good. There were some variables we had to change from categorical to numeric. We had some missing values to remove.

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

Descriptive statistic functions in Python

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?

Basically all the variables in the dataset asides from #, video_id – because these columns are pointless, and video_transcription_text since it’s categorical.

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

Nearly perfectly, with a recall score of 0.99.

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

Only adding in more explanatory variables if they became available.

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

Going back to the labs from the course to understand how to correctly set up and train a model.

PACE: Execute Stage

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

The variables that had to do with engagement are most predictive of whether the video is a claim or an opinion. The model would have been choosing the claim status of videos based on the engagement levels each video received.

- What are the criteria for model selection?

How well the model performs in the accuracy, precision, recall, and f1 metrics.

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

Yes, the final results are acceptable

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

Not necessarily, it already performs near perfect.

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

Verified_status, video_duration, author_ban_status, and text_length weren't very important. In a future iteration, I would remove these variables.

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

I would propose that TikTok should utilize the model to identify videos that are claims or opinions. Claims should be sent for further review.

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

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

Again, past labs to understand confusion matrices and fitting the model.

- Is my model ethical?

Yes, since the model rarely assigns false negatives or false positives, I would say it is fairly ethical.

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

When a model makes a mistake, it is misclassifying a video. This could cause TikTok to review videos that are merely opinions, and miss out on reviewing claims.