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~~

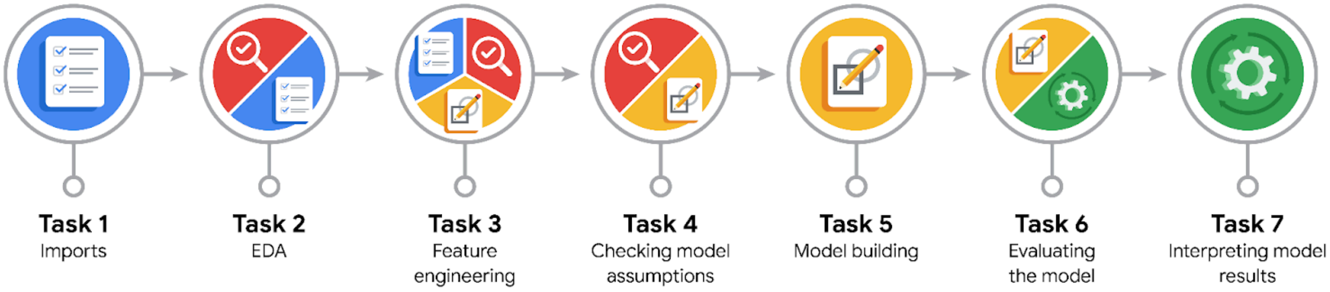
Relevant Interview Questions

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?

To develop a machine learning model that can assist in classification of videos as either claims or opinions.

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

The cross-functional team(namely Mary Joanna Rodgers- Project Management Officer,Margery Adebowale- Finance Lead, Americas and, Maika Abadi- Operations Lead.)

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

Jupyter Notebook, Python,Pandas ,Numpy,Matplotlib,Seaborn,Sci-kit learn and XGBoost packages.

- Do you have any ethical considerations at this stage?

It is better for the model to predict a false positive when it makes a mistake instead of a false negative, as it is very important to identify videos that violate terms of service, even if that means some opinions are classified as claims.

- Is my data reliable?

After conducting the EDA and dealing with missing data, the data looks reliable for further testing.

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

In a perfect world, all data regarding the number of reports for each video of each author, their verification and ban status as well as the text transcriptions and the engagement rates for the videos.

- What data do I have/can I get?

We have almost everything except the number of reports and some missing data points.

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

The F1 score would give a balanced score ,also the recall would be helpful to evaluate success for the

Business objective.

PACE: Analyze Stage

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

I think it still works , just need to focus on preparing the data for feature engineering

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

No ,not so far especially since tree models do not require handling of outliers.

- Why did you select the X variables you did?

Because they seem to be the most predictive of the claim status,especially the engagement rates

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

To handle missing data or outliers(if needed) and to structure, aggregate and validate data so we have a better understanding of it and it is ready for constructing the model.

- What has the EDA told you?

The data has very few missing rows so we can remove them and the class proportion is quite fairly balanced.

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

Python in jupyter notebook environment along with pandas ,seaborn and matplotlib packages.

PACE: Construct Stage

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

The categorical variables need to be one hot encoded or get label encoded. Yes, it can fixed right now by mapping each categorical variable or using pd.get_dummies.

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

The engagement rates,verified status ,author ban status and the video transcription text as all these seem to be predictive of the claim status.

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

The model fits well with 99% scores across major metrics especially F1.

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

I can go back and do feature engineering or hyperparameter tuning but the model is already performing near perfect,so there is no need for change

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

Pandas, numpy ,scikit-learn, and XGBoost packages.

PACE: Execute Stage

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

The model predicted near perfect with only 5 mistakes on the test data. The model predicts 1 for a claim and 0 for an opinion and does so near perfectly for the test data.

- What are the criteria for model selection?

The model selection was done based on F1 score with the random forest model having slightly higher f1 score than the XGBoost one.

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

Yes, it makes sense and the final model is acceptable.

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

It maybe can be improved but it is already at 99% so i think there is no need for it.

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

Features like video_id were already taken out. Video duration sec seems to have very few importance and could be removed as well, it can help reduce the time taken for the model to fit.

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

With these results, we can conclude that videos with higher user engagement levels were much more likely to be claims

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

Questions like the number of views for claims and opinions as no opinion video got more than 10,000 views.

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

Sci-kit learn, specifically its metrics package , matplotlib and also pickle(to save the model locally)

- Is my model ethical?

Yes , it has a really great recall score so it assists in preventing any claim go through and violate the terms of service

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

If it predicts a False negative, it mistakes a claim as an opinion and when it predicts a false positive it mistakes an opinion as a claim. Pertaining to the use case, it is better to predict a false positive instead of a false negative but with only 5 mistakes overall on the test data, it is very good for the use case.