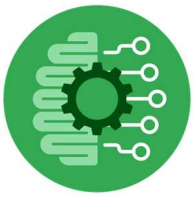
Course Six
The Nuts and Bolts of Machine Learning



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

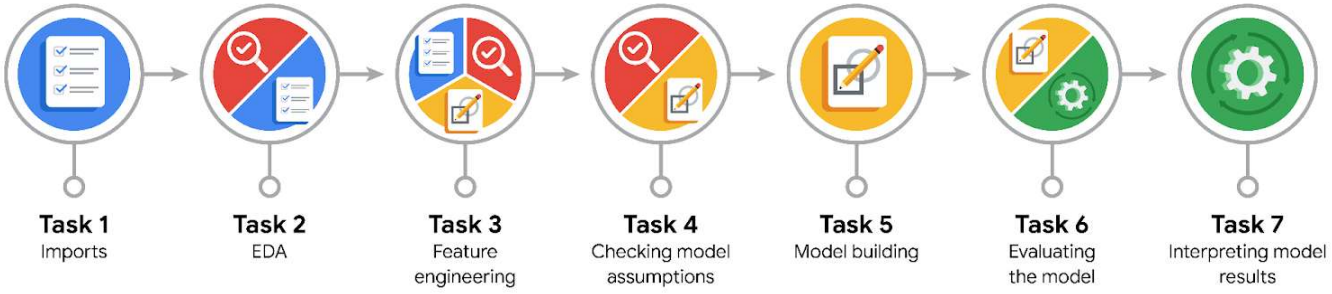
Relevant Interview Questions

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 project goal is to create a model capable of accurately classifying video content as either claim or opinion.

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

  The main stakeholders for this project are Harriet Hadzic and May Santner, Director and Data Analysis Managers.

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

  I have used files from other projects as a guide for this one. Furthermore, independent research into several topics has facilitated my understanding of new concepts that have arisen.

- Do you have any ethical considerations at this stage?

  My primary ethical concern at this stage relates to the content of videos containing claims. These videos are clearly viewed and shared much more widely than those containing opinions, and platforms like TikTok need to invest significantly more in ensuring harmful content is flagged and blocked.

- Is my data reliable?

There is no reason to believe the data is unreliable. While an imbalance existed in the 'verified_status' variable, this was addressed by upsampling the minority class.

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

How was the data collected, and what methods were used?

- What data do I have/can I get?

The data primarily consists of video engagement metrics, such as likes, shares, views, downloads, and comments. I'd also like to see data on the videos' geographic origin, such as the region they come from.

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

The objective is to create a model that predicts whether a video contains a claim or an opinion. This model's performance can be evaluated using several metrics, including precision and the F1 score. Precision measures the probability that a video predicted to contain a claim actually does so, while the F1 score represents the harmonic mean of precision and recall, where recall measures the model's ability to correctly identify all actual claims.

PACE: Analyze Stage

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

By this stage, the method for solving the problem should still be relevant and workable. Any further revisions will likely occur later in the analysis phase.

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

• Independence of Observations: Currently, there's no reason to suspect that the data points in the dataset are not independent. However, more information about the data collection process would be helpful in confirming this. • Multicollinearity: Some variables within the data exhibit multicollinearity. One highly correlated variable, in particular, was removed as a feature. • Sufficient Data: Approximately 20,000 rows of data should be sufficient for the analysis. • Relevance of Features: All features have some relevance to the target variable. Exploratory data analysis (EDA) will identify the most relevant features.

- Why did you select the X variables you did?

The X variables chosen in earlier stages of the analysis process were determined to have some degree of relationship with the target variable and were therefore included in the model.

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

Exploratory data analysis (EDA) allows for a thorough examination of the data, including identifying and addressing outliers (by removal or normalization), checking distributions, removing erroneous data, handling missing data (through imputation or removal of rows), and examining relationships between variables. It is a critical step in the process before moving on to model development and gaining insights.

- What has the EDA told you?

EDA revealed a significant imbalance in the verified_status variable but very little imbalance in the claim_status variable. It also showed that many of the video engagement metrics are highly right-skewed and contain numerous outliers.

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

Referencing similar material, as well as researching topics online as needed.

PACE: Construct Stage

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

The only anomaly I noticed was that the author_ban_status variable, when encoded, used TRUE and FALSE values instead of 1s and 0s. I corrected this to prevent potential issues during model development.

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

All video engagement metrics were included, along with verified_status, author_ban_status, and a newly engineered variable containing tokenized text from the video_transcription_text variable.

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

Both the Random Forest and XGBoost models fit the data nearly perfectly. The metrics used, precision and recall, are very high, indicating that the models do an excellent job of predicting video claims and opinions.

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

At this stage, any further improvements will be minimal, the model already performs almost perfectly.

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

Referencing similar material, as well as researching topics online as needed.

PACE: Execute Stage

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

The features that were determined as being the most predictive, the video engagement metrics such as views, likes, shares, downloads and comments were utilized by the model as predicting whether a video contained a claim or opinion. The scores obtained were very high, and indicate the model does an almost perfect job classifying a video as either containing a claim or opinion.

- What are the criteria for model selection?

The model must have high precision when predicting whether a video contains a claim and must minimize false negatives (classifying a video as containing an opinion when it actually contains a claim). This is crucial because such misclassifications could be problematic, especially if the video contains harmful content and is widely shared.

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

Based on what is known about the video engagement metrics prior to building the model, it is not surprising they are correlated with whether a video contains a claim or opinion. The final results are acceptable, and the model performs very well.

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

  At this point, the model is nearly perfect.

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

  Many features, compared to the video engagement metrics, appeared to have low importance and could likely be removed through feature elimination to potentially improve the model further. However, the relative importance of these seemingly less important features compared to the video engagement metrics may not be accurate.

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

  Because of the model's strong performance on the validation and holdout (test) sets, it is suitable for deployment in a production environment. Ideally, however, it should first be tested on a completely independent dataset of real-world data.

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

  Video engagement metrics, such as likes and views, are strongly correlated with whether a video contains a claim or an opinion. verified_status is also related to whether a user posts a video containing a claim or an opinion.

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

  Referencing similar material, as well as researching topics online as needed.

- Is my model ethical?

  The model itself does not appear to violate any ethical rules. Rather, it has illuminated ethical considerations, such as how widely claim videos (which may contain harmful, misleading, or misinforming content) are spread and viewed by users of the application, particularly those who may be at higher risk.

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

  When the model makes an error, it misclassifies a "claim" as an "opinion," or vice versa.