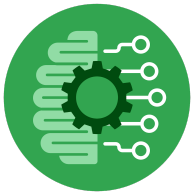
2Course Six
The Nuts and Bolts of Machine Learning



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~~

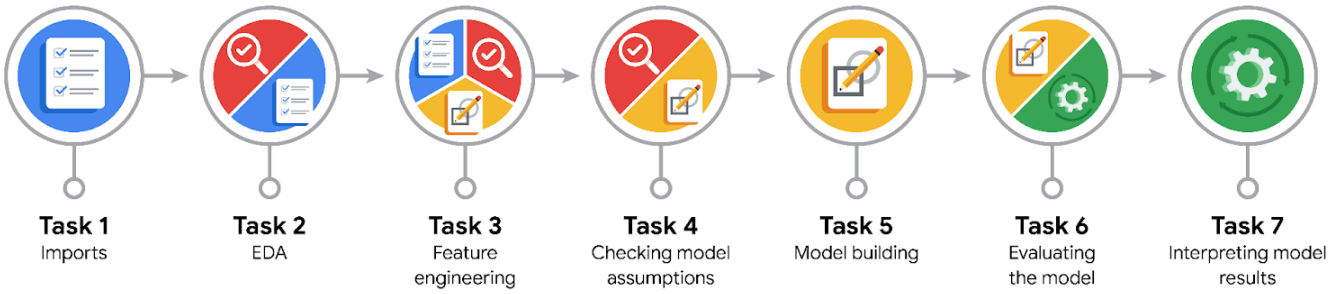
Relevant Interview Questions

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 final effective machine learning model for classifying whether a TikTok video presents a "claim" or presents an "opinion".

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

  Mary Johanna Rodgers, Project Managment Officer

  Willow Jaffey, Data Science Lead

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

  My lecture notes front this whole specialisation, a Jupyter notebook to make research, the data set metadata information and the results and executive summaries from previous steps, especially the evaluation scores from previous hypothesis testing and logistic regression models.

- Do you have any ethical considerations at this stage?

  If a video is wrongly classified, what will the consequences be for the video's user? Could their account being banned?

  Will the model be biased against non verified users over verified ones?

Will the model be biased against a type of video's content?

What have been the effect of that classification? Are the any statistics about already? What happens if TikTok commits a mistake about it?

How TikTok takes the decision of a claim is actually true or false? What kind of sources are used to to do that classification?

Although the data implies that a video labelled as a claim would have a deeper fact-checking by TikTok and would lead to consequences as bannig the author is the claim is false, the project never explicitly states what actually are the actions TikTok take about it.

- Is my data reliable?

I can't answer that with the very few information I have from how the data set I got. For instance, I don't have information about the real proportions of claim and opinion videos in TikTok platform, as well as the percentages of types of author and account status.

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

I would like to know the real proportions of the population in categories like author, account status and claim or opinion videos.

The outcome variable is almost balanced, which is a great advantage for the analysis. However, is this proportion true in TikTok?

Also, I don't know how those videos were classified as opinion or claims, That would be important information to know.

There are several null entries, so I would like to know why that information is missed.

What was the process followed to enter each cell of this data set.

I would like to know the answers to the ethical questions I did berofe.

- What data do I have/can I get?

I have several engagement metrics of a video such as views, likes, downloads, shares and comments; the video's transcription text and duration; whether a video is a claim or an opinion; the author, and the account status.

I can get the text length from the transcription text feature and I can try to get information from that transcription using tokenisation with *CountVectorizer*.

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

An opinion is less urgent and don't need as much moderation from TikTok. In contrast, claims are treated differently from opinions, likely because they can spread misinformation or require fact-checking, then they are more important to be correctly classified.

In our data set claim status must be transformed to a boolean variable. As it can be seen in the .ipynb, claim is labelled as True (1) and opinion as False (0).

For those reasons, Recall is the appropriate metric, since a video containing a claim that is wrongly predicted as an opinion could avoid fact-checking, would prevent consequences for accounts uploading videos with false claims and could potentially damage innocent audience.

However, this also have a potential censoring ethical problem.

PACE: Analyze Stage

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

To identify if a video has a claim or opinion. There is no need to change the plan for. I will train two models, a Random Forest and a Gradient Boosting Machine.

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

There are some changes to do before training the models with this data set in order to make the conditions acceptable. For instance, Random Forest needs all data to be numerical, so ‘object’ columns must be converted; null entries aren’t allowed, they must be handled beforehand. So, in this case, the decision is to remove observations containing at least one null value.

- Why did you select the X variables you did?

I used all the variables as predictor ones except by #, video_id and video_transcription_text since the first two don’t include any useful information and the third was used to create a more meaningful and numerical value as the number of characters of the transcription and tokenisation of 2-grams and 3-grams.

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

Understand relationships between variables; identify which ones behave differently or have different values depending on video’s claim status, making them useful to classify videos.

Find interesting relationships and create new variables from the given ones to extract more meaning.

To find irrelevant or missing data and outliers.

Create, modify or change questions that arise from variable’s analysis that could be interesting for stakeholders.

Let the data ready for training models according to their requirements and get meaningful results.

- What has the EDA told you?

- The mean of a video transcription length is different between claims and opinions, the former ones are longer.

- The outcome variable is almost balanced, 50.34 % of the observations are classified as claim. I don't need to resample this data set.
- All engagement metrics vary between claim and opinion videos, so they will be important to predict claim status. Claim tends to have much larger engagement rates.
- Percentage of verified users is significantly lower among videos containing a claim. So, verified status is also meaningful for the model.
- Author ban status percentages are also very different between claim and opinion videos. Proportion of active users is higher when a video is classified as an opinion, whilst banned and under review statuses are higher for videos identified with claims.

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

.head(), shape, info(), describe(), isna(), duplicated(), .drop(), value_counts(), CountVectorizer(), and visualisation techniques like boxplots, histograms, etc.

PACE: Construct Stage

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

The aren't problems. I just encoded the object-type variables to booleans and used drop_first=True to avoid dependency within those variables.

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

All the engagement metrics—views, likes, downloads, shares and comments—; the video's transcription text length and its 2-grams tokenisation; the author, and the account status. As I have verified, these features have different values for claims and opinions, so they pontentially have predictive power for claim status.

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

Random Forest and Gradient Boosting both excelled at fitting the data. Recall, precision, accuracy, and F1 score all exceeded 99% on the validation set. Random Forest performed slightly better than the other model—not only achieving 99.79% recall on the validation set, but also posting the highest scores across all four metrics.

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

Apart from tweaking hyper parameters to find the best combination, I am going to use the most rigorous approach to model development, including cross-validation, a separate validation set for both

models trained, and after choosing the champion's model, test it with completely new data in order to have an idea of real model's behaviour with unseen observations.

I'm also going to use a CountVectorizer to include key word pairs that could improve the model's predictions by identifying recurring expressions in the video transcriptions, depending on whether the class is claim or opinion.

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

I cross-validated data by using GridSearchCV; tuned both models with hyper parameters such as max_depth, min_samples_leaf, min_samples_split, max_features, max_samples, n_estimators, min_child_weight, subsample, colsample_bytree, learning_rate; got the best parameters from the running process by using best_params_; evaluated model results with four scores, recall, precision, accuracy, f1; and created data frames with those scores.

PACE: Execute Stage

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

The Random Forest model was the best in performance, so it was selected as the model to show to TikTok.

The model classified opinion and claim videos primarily based on engagement metrics such as the number of views, likes, shares, downloads and comments. These were the most influential variables for predicting claim status.

As I mentioned earlier, the percentage of claim videos correctly classified was over 99% in both the known data (training set) and the new data (validation and test sets). The other metrics—precision, accuracy, and F1 score—also exceeded 99%. In particular, the precision score on the test set was perfect, meaning all opinion videos were correctly classified.

- What are the criteria for model selection?

I selected recall as the most important metric for model selection. The Random Forest model outperformed the Gradient Boosting model in recall on both known and new data.

I also considered precision, since it's important to understand how many false positives occurred—that is, opinion videos incorrectly classified as claims.

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

The model's results were nearly perfect, showing no signs of overfitting—it performed exceptionally well on both the training and new data. This means TikTok can confidently use it to classify claim and opinion videos, helping reduce the backlog of user reports and enabling more efficient prioritization.

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

There is always room for improvement, but achieving over 99% in all metrics is an excellent result. The predictive power of the selected variables is very high, and adding new variables or further feature engineering might introduce complexity without delivering significant gains.

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

Feature importance results show that engagement metrics were essential for classifying videos as claims or opinions. Interestingly, several word pairs extracted from the transcriptions contributed more to the model's decisions than the original variables in the dataset, like: "media claim", "colleague learned" or "friend read". It's also clear that views, likes, shares, downloads, and comments had the strongest influence in separating claim from opinion videos—though it's worth remembering that high feature importance doesn't necessarily imply a strong causal effect.

Although I don't rule out removing variables that showed low relevance in the feature importance analysis, the results are already excellent. Eliminating them might slightly reduce the model's flexibility in borderline cases, especially with unseen data. Still, it remains a possible next step.

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

TikTok could implement it to *prioritize content moderation* efforts by flagging likely-claim videos for review. This model could help optimize the allocation of human fact-checkers, reducing their workload and accelerating response times.

Since engagement metrics are key predictors, the platform could also monitor viral videos more closely, as they're more likely to carry claims.

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

How do engagement patterns differ between banned, under-review, and active accounts?

Is there a threshold of user engagement beyond which a video is highly likely to be a claim?

How does verification status influence content type (claim or opinion)?

What role does the transcription text length and tokenized content play in identifying claims?

Is the model's performance consistent across different subsets of data (e.g., by video length, verification status, account status)?

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

I used my course notes to review theoretical explanations and code. I also leveraged artificial intelligence to deepen my understanding of certain concepts and to assist with feature engineering. Additionally, I revisited videos and readings from this Google Specialization to clarify any doubts I had while working on the project. To ensure consistency, I also reviewed my previous executive summaries to recall key conclusions, relationships, and variable influences.

- Is my model ethical?

I consider the model is designed ethically—it's transparent, avoids bias from irrelevant variables that were in the original data set and aims to improve content moderation. However, as I said earlier:

False positives (opinions flagged as claims) may result in unjustified moderation or censorship.

If the training data reflects biases in moderation decisions, the model could replicate them (e.g., treating verified/unverified accounts unfairly).

There's limited information about how videos were originally labeled as claims or opinions, which makes it hard to assess bias in ground truth.

Including human oversight and fact checking databases in the final use of the model would help mitigate them.

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

There are two kind of mistakes:

False Negatives (claims predicted as opinions): These are more critical. They may allow misleading content to go unchecked, harming users and public trust.

False Positives (opinions predicted as claims): These may trigger unnecessary moderation, leading to potential user dissatisfaction or even freedom of expression concerns.

In both cases, the stakes are high. That's why optimizing for recall (minimizing false negatives) is appropriate as the chosen model did, and keeping false positives low is also necessary to balance ethical impact.