# 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~~

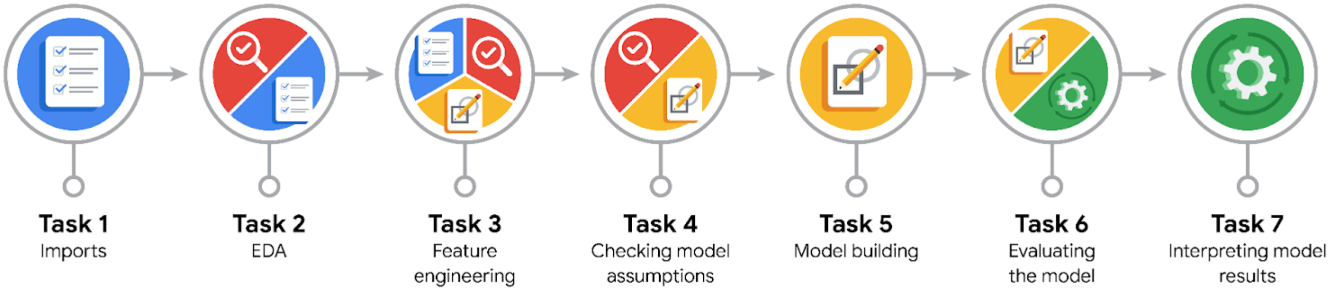
## Relevant Interview Questions

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 create a classifying algorithm that can reduce the backlog for video claim status. |

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

| |
|---|
| The external stakeholders are the Tik-Tik management team from external dpt who also make business decisions regarding the future of the company. |

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

| |
|---|
| The reference guides for using and calling the classifiers such as the Random Forest and the Gradient Boosting. |

- Do you have any ethical considerations at this stage?

There are several ethical considerations that need to be considered at this stage, for instance the results of our model will determine the exposure a video from a content creator gets from Tik-Tok. This can greatly affect the financial stability of several content creators.

- Is my data reliable?

The data used for this study is reliable for the purposes of this study because it is sourced from Tik-Tok and it provides unbiased information about important video metrics.

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

It would be useful to have video category classifications and maybe author history with previous videos claim status and ban status.

- What data do I have/can I get?

With the dataset provided I can look at the video transcription length and possible rates of engagement such as comments per views, or shares per view, such that we can quantize how active users are with the content.

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

Good metrics to use are the standard 4 metrics which are Precision, Recall, F1, and Accuracy. This metrics measure the  model's capability of classifying correct true positives, true negatives and minimizing false positives and false negatives.

## PACE: Analyze Stage

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

We are trying to develop a model to classify videos as claims or opinions using the engagement metrics provided by Tik-Tok.

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

Unfortunately, most of the engagement metrics such as view count, share count, comment count download count and like counts are all right skewed therefore we cannot assume these variables are normally distributed, and the most descriptive statistics used are the median and the interquartile range.

- Why did you select the X variables you did?

I selected all the engagement metrics previously mentioned with the inclusion of the video transcription length. This feature can be very important to determine the type of content the video presents to the platform.

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

The purpose of performing EDA is to ensure that the data used for the model is of the highest quality possible. It's important to remember that a model is as good as the data used for training.

- What has the EDA told you?

There are 298 rows with missing values on them, there were no duplicated rows in the whole dataset, a log transformation on the skewed data did not remove the asymmetry from the engagement metrics. Using the IQR rule to determine outliers showed there were 3589 outliers in the whole dataset, in all the columns. Video transcription length is the least correlated feature like count is the overall most correlated feature.

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

I used some of my previous notebooks for feature engineering, and from previous courses on handling missing data points, outliers and multicollinearity in the selected features.

## PACE: Construct Stage

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

The video transcription length was the only engagement metric that followed a normal distribution and claim videos had longer transcription lengths on average than  opinion videos.

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

I chose all the engagement metrics from the previous labs plus the length of the transcription length, and the author verified/ban status. These features provide information about the video's success, its content and its author.

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

The model fits that data very well, as we see the recall, precision, f1 and accuracy scores being very close for the validation sets.

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

We could potentially improve the model by using custom indexing with the validation set instead of using a 5 fold cross validation methodology.

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

The previous notebooks for implementing random forests, and XGBoosters were incredibly useful.

## PACE: Execute Stage

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

Some of the key insights provided by the model are that video view count, video transcription length and video like count are the best features to estimate a video's claim status. We also see that it is more likely to misclassify an opinion video as a claim but not a claim as an opinion.

- What are the criteria for model selection?

The criteria used for model selection was to use the model who would produce the least amount of incorrect classifications such as false positives and false negatives, therefore, the most important metric for model selection was the F1 score.

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

The results of the model make sense because the feature importance outlined by the model have logical explanations behind each one of them, thus it can be inferred how certain features affect the classification procedure.

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

The model could be improved if we chose a hyper-tuned random forest because this would remove the "black box" characteristic of the XGBooster and could provide further insight into the relation of specific features to the target variable.

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

I took out the video id number as this is not significant toward the classification of the claim of a video.

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

We would recommend to invest time and resources into classifying content type in Tik-Tok as this is the only type of data we had a the least amount of and it had one of the highest importance ratings in our model.

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

We could expand the hyper-parameter space of the model to see if further improvements can be achieved through these fine tuning decisions.

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

The lab for XGBoosting and the reference guide for evaluating models based on metrics.

- Is my model ethical?

I believe my model is ethical as I trained it to no be biased, manage randomness and the dataset itself had no features which could lead to misrepresentation for a specific characteristic/group in Tik-Tok

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

Sometimes mistakes like those could be indication of the quality of data, or bias in the data used to train the model itself