**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?

TikTok users are submitting reports that identify videos as containing claims. Too many of these requests come in for the human moderators to review in a timely manner so a predictive model is to be developed to assist the human moderators.

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

Mary Joanna Rodgers- Project Management Officer

Margery Abebowale- Finance Lead, Americas

Maika Abadi- Operations Lead

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

Data from the organization, python, and Jupiter Lab.

* Do you have any ethical considerations at this stage?

The data used may be sensitive and so a need for privacy is important. Also, the use of a learning model to determine if videos have claims may have negative impacts on users if their videos are marked as claims incorrectly.

* Is my data reliable?

Yes the data has been cleaned so that the relatively small number of nulls is removed.

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

I would be able to use a highly accurate large language model to actively scan the transcripts of each video for instances of claims being made.

* What data do I have/can I get?

I have many engagement metrics and I know from prior steps that engagement metrics are highly predictive of if the video has claims or opinions.

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

I would rather predict too many false positives which would give humans too many videos to review rather than letting claims videos go under the radar. I will value the F1 Score as the key metric.

**PACE: Analyze Stage**

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

Yes it still works. I will create a predictive model.

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

A random forest does not have any key assumptions that the data does not conform to. Random forests have almost no assumptions which makes them resilient tests.

* Why did you select the X variables you did?

Engagement metrics such as view time, number of views, number of likes and so on are all shown to be related to if the video has claims or opinions. This was determined by previous course stages.

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

EDA ensures that the data is able to meet the assumptions, that the data is clean (no nulls and no strange data errors like duplicates), and it helps you better understand the data that you are working with.

* What has the EDA told you?

The data is already very clean. I just ensured that the duplicates are removed.

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

Just python and Jupiter labs.

**PACE: Construct Stage**

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

No issues with the data.

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

Mostly engagement metrics such as view duration, view count, and so on. These metrics were shown as relevant to claim or opinion status.

* How well does your model fit the data? What is my model’s validation score?

The model is highly accurate. The Accurately, Precision, Recall and F1 scores are all basically perfect. I did not display enough decimal points so all of the values round up to 1.00

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

There is not really anything that I can improve the model since it is basically perfect as it is.

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

Just python and Jupiter labs.

**PACE: Execute Stage**

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

The model is a random forest. A random forest is a number of decision trees trained with replacement and randomly selected features. The model shows that engagement metrics such as view duration are key in predicting if videos have claims.

* What are the criteria for model selection?

The model with the best F1 score and the least number of false negatives.

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

Yes the model shows that the key factors are also engagement metrics and the accuracy of the model is almost perfect.

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

No the model is essentially perfect. There is not really much that can be done to perfect it.

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

Some forms of data were not very predictive such as text length.

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

That we work to investigate why high engagement videos seem to be claims based videos. Is there a way we can incentivize people to view opinion videos over claim videos?

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

I could preform the inverse meaning I could use the claim vs opinion status to determine what videos are likely to perform well. I could also delve further into the transcript data to see if particular topics are more likely to be the subject of claim videos.

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

Just python and Jupiter labs.

* Is my model ethical?

Yes. The model is very accurate and the error prefers false positives. This is exactly what I want.

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

When the model makes a mistake, it is usually a false positive, videos that state opinions but the model detects as claims. This is better than false negatives because it can flag the video for review by a human.