**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?

Increase response time and system efficiency by automating the initial stages of the claims process.

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

The project management team

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

Jupyter notebook, python and its libraries such as numpy and pandas.

* Do you have any ethical considerations at this stage?

Where is the data from?

Did the users give consent in usage of the data?

Is there any personally identifiable information in the dataset?

* Is my data reliable?

Yes, the data is reliable, it is from a primary source.

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

Data containing user video counts and user transcription text.

* What data do I have/can I get?

I do have this data.

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

RECALL.

Making a false negative prediction may predispose to high levels of customer dissatisfaction and user churn however, if we make a false positive prediction the worst scenario is that there is an increase in the workload by tiktok employees.

**PACE: Analyze Stage**

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

Yes, it does, the plan does not need a revising.

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

No, the assumptions are required for the random forest ensemble model.

* Why did you select the X variables you did?

Relevance of the X variables to the business task.

Prioritizing computational efficiency.

Decreasing the number of variables to enhance interpretability.

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

Familiarizing myself with the data;

Removal of outliers that could affect our model although this is a random forest model.

Managing duplicate values and managing missing data.

Checking for class imbalance.

* What has the EDA told you?

There are outliers in the ‘'video\_like\_count', 'video\_share\_count', 'video\_download\_count', and 'video\_comment\_count' columns.

There are no duplicate values but there were 7 columns containing 298 missing values: ‘claim\_status’, ‘video\_transcription\_text’, ‘video\_view\_count’, 'video\_like\_count', 'video\_share\_count', 'video\_download\_count', 'video\_comment\_count'.

There is relative balance between our target variables values: claim\_status - claim (50.3%) and opinion (49.7%).

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

Jupyter notebook, python and its libraries such as scikit-learn, seaborn, numpy and pandas.

**PACE: Construct Stage**

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

No, there was nothing odd.

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

The independent variables that were chosen include: video\_duration\_sec', 'video\_view\_count', 'video\_like\_count', 'video\_share\_count', 'video\_download\_count', 'video\_comment\_count', 'text\_length', 'author\_ban\_status\_banned', 'author\_ban\_status\_under review', 'verified\_status\_verified'. In addition, numerical representatives of ‘video\_transcription\_text’

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

The model fit the data accurately with a validation score of 0.996 for both random forest and xgboost models.

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

Yes, surely we can.

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

Python libraries: scikit-learn, xgboost, numpy and pandas and jupyter notebook.

**PACE: Execute Stage**

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

Users' video engagement such as view, like, share, download and comment are very high predictors for classifying videos reported as claims or opinions.

* What are the criteria for model selection?

The model that has the highest performance on validation data evaluated by key metrics: accuracy, precision, recall and f1 score.

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

Yes, it does. It had almost no false positives and very little false negatives prediction.

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

No, the model is almost perfect. Further engineering could do only little.

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

The features that were not important include: ‘video\_duration\_sec’, 'author\_ban\_status', 'author\_ban\_status', 'verified\_status'. In addition, numerical representatives of ‘video\_transcription\_text’ had almost no effect.

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

**Automate Video Classification**: This could significantly reduce the time spent manually reviewing videos and increase overall efficiency.

**Focus on User Engagement Metrics** : User engagement metrics like views, likes, shares, downloads, and comments. Therefore, it would be beneficial to focus on these metrics when analyzing potential claims or opinions.

**Continuous Model Training**: Given the dynamic nature of social media content, it would be beneficial to continuously train and update the model with new data. This will ensure that the model remains effective as trends and behaviors change over time.

**Monitor and Improve**: Regularly monitor the performance of the model and make necessary adjustments. This could involve tweaking the model parameters, adding new features, or even switching to a different model if the current one stops performing well.

Finally, while automation can save time and effort, it's crucial to have a human in the loop for critical decisions. The model should be used as a tool to assist decision-making, not replace it entirely.

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

What are the most important features driving our model's predictions?

How well does our model perform on unseen data compared to the training data?

Are there any features that seem to be overly influential or underutilized by the model?

Could we improve our model's interpretability by using a simpler model or removing underutilized variables?

How does the model's performance vary across different subsets of the data?

Are there any patterns or trends in the feature importances that could inform future data collection efforts?

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

Jupyter notebook, python and its helpful libraries including scikit-learn, numpy and pandas, visualization tools like seaborn and matplotlib.

* Is my model ethical?

Yes.

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

In this case it would be classifying a claim as an opinion when it is not.