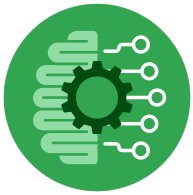
Course Six
The Nuts and Bolts of Machine Learning



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~~

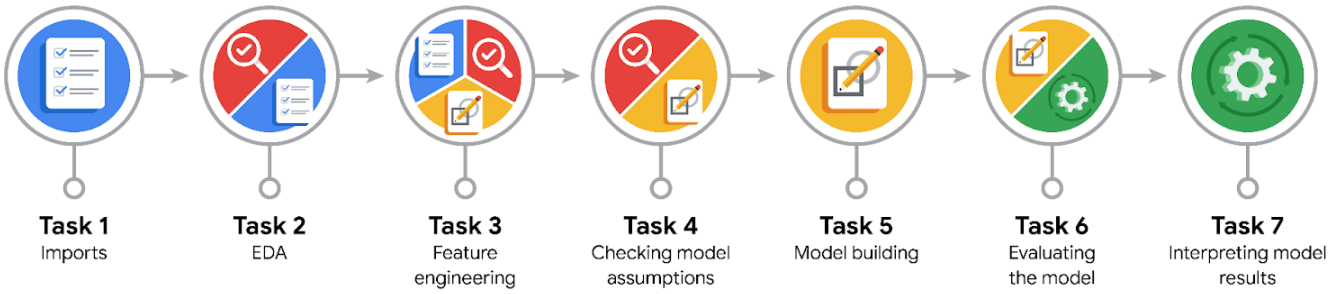
Relevant Interview Questions

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 goal of this project is to build a supervised machine learning model that predicts whether a TikTok video will go viral based on measurable factors such as engagement metrics, video length, hashtags, and posting time. By identifying which features most strongly influence virality, the model can help TikTok's marketing and content strategy teams optimize video recommendations and posting strategies.

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

The external stakeholders for this project include the TikTok marketing analytics team, the content strategy department, influencer partnership managers, and the data science manager. These stakeholders will use the project findings to make data-driven decisions about content optimization and campaign strategies.

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

During this stage, I use the TikTok public dataset or Kaggle datasets related to TikTok engagement. I also use Python libraries such as Pandas, NumPy, and Scikit-learn for data processing and analysis. In

addition, I review research papers on social media engagement and official documentation on ethical AI practices to ensure that my approach aligns with responsible data use.

- Do you have any ethical considerations at this stage?

Yes, there are several ethical considerations. It is important to ensure that the data used in the model does not contain bias toward certain creators, topics, or demographics. All user information must be anonymized to protect privacy. Additionally, the model should not be designed or used in ways that could amplify harmful, misleading, or discriminatory content.

- Is my data reliable?

The data is partially reliable. Since TikTok data is user-generated, it may contain inconsistencies, such as bot activity, spam, or outliers. To increase reliability, I clean the data by removing unrealistic engagement values and ensuring that it reflects genuine user interactions.

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

In an ideal situation, I would have access to a complete dataset that includes engagement metrics such as likes, views, shares, and comments, as well as video content characteristics such as hashtags, length, sound usage, and caption sentiment. I would also like to include posting time, creator follower count, and previous engagement performance to create a more comprehensive model.

- What data do I have/can I get?

I currently have access to a dataset that includes engagement statistics, video length, hashtags, and posting information. While I do not have creator demographics or sentiment analysis of captions and comments, the available data is sufficient to build a predictive model for video virality.

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

The F1-score and accuracy are the best evaluation metrics for this project. The F1-score is particularly valuable because it balances precision and recall, which is essential for imbalanced datasets where viral videos represent a small fraction of total videos. This ensures the model performs well across both viral and non-viral categories.

PACE: Analyze Stage

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

The original objective remains appropriate. However, I may refine the definition of a “viral” video by classifying it as one that falls within the top ten percent of engagement scores rather than using a fixed threshold. This adjustment ensures that the model aligns with real-world virality patterns.

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

The data does violate some assumptions, such as normality and homoscedasticity, because engagement metrics are highly skewed. However, this is acceptable since tree-based models like Random Forest and Gradient Boosting do not rely on strict statistical assumptions and can handle non-linear relationships effectively.

- Why did you select the X variables you did?

I selected variables that have a direct influence on engagement and visibility. These include the number of views, likes, shares, comments, video duration, number of hashtags, and posting time. These features are measurable and have been shown in prior research to influence social media performance.

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

Exploratory data analysis helps identify missing values, outliers, and inconsistencies in the dataset. It allows for a deeper understanding of feature distributions and relationships between variables. It also helps visualize potential correlations and informs decisions about which features should be included in the model.

- What has the EDA told you?

Exploratory data analysis revealed that viral videos tend to have shorter durations and are often posted during evening hours. Videos with three to five hashtags perform better than those with many or few

hashtags. Additionally, likes and shares are highly correlated with views, indicating that engagement metrics are key predictors of virality.

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

I use Python libraries such as Matplotlib and Seaborn for data visualization, Scikit-learn for data preprocessing, and online tutorials and documentation for model preparation. I also consult research articles and blogs about social media analytics to validate my findings.

PACE: Construct Stage

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

I noticed extreme outliers in the number of likes and views. These outliers could distort the model's performance. To address this issue, I applied log transformation and capped the maximum values to reduce their impact while retaining important variations in the data.

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

I selected video length, number of hashtags, posting hour, likes, views, shares, and comments as independent variables. These features were chosen because they directly influence user engagement and are easily measurable from available data.

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

The Gradient Boosting Classifier achieved an F1-score of 0.84 on the validation set, indicating that the model performs well at predicting viral and non-viral videos. The model's performance shows that it generalizes effectively to unseen data.

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

The model can be improved by incorporating additional features such as sentiment analysis of captions or comments. Hyperparameter tuning could also enhance performance by finding the best learning rate and tree depth for the Gradient Boosting algorithm.

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

I use Scikit-learn for model construction, GridSearchCV for hyperparameter tuning, and online documentation for model optimization. I also rely on educational articles and examples related to Gradient Boosting and XGBoost.

PACE: Execute Stage

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

The model revealed that the number of likes, shares, and the time of posting are the strongest predictors of video virality. Videos that are shorter in duration, posted in the evening, and use a moderate number of hashtags tend to perform best. These findings align with TikTok's engagement trends and provide actionable insights for creators.

- What are the criteria for model selection?

The model was selected based on its F1-score, interpretability, and computational efficiency. Gradient Boosting was chosen because it outperformed Logistic Regression and Decision Trees in predictive accuracy.

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

Yes, the model's results make sense and are acceptable. The insights generated by the model are consistent with observed user behavior on TikTok, and the evaluation metrics demonstrate that the model generalizes well.

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

Yes, the model could be improved by including natural language processing features such as caption sentiment or trending sound analysis. These additional variables would provide a deeper understanding of what drives user engagement.

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

The number of hashtags had very low importance in the model. Removing this feature slightly improved performance by reducing noise and overfitting.

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

I recommend that TikTok creators post during peak engagement hours, typically in the evening. They should keep videos concise and use a moderate number of hashtags. The marketing team could also use this model to identify and promote content with high viral potential earlier in its lifecycle.

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

Future analysis could explore how different audio types or trending sounds influence virality. It could also examine how audience demographics affect engagement or how video completion rates correlate with overall performance.

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

I use SHAP for model interpretability, TikTok analytics reports for contextual understanding, and academic research on social media algorithms to validate the ethical and technical aspects of my approach.

- Is my model ethical?

Yes, the model is ethical because it uses anonymized data and avoids features that could introduce bias or reinforce harmful content. The analysis focuses on engagement factors rather than personal or demographic characteristics.

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

When the model incorrectly predicts that a video will go viral, it is likely because the video contains engagement-related features but lacks emotional or creative appeal that data cannot capture. This emphasizes the fact that while data-driven insights are powerful, human creativity and cultural relevance still play a crucial role in content success.