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

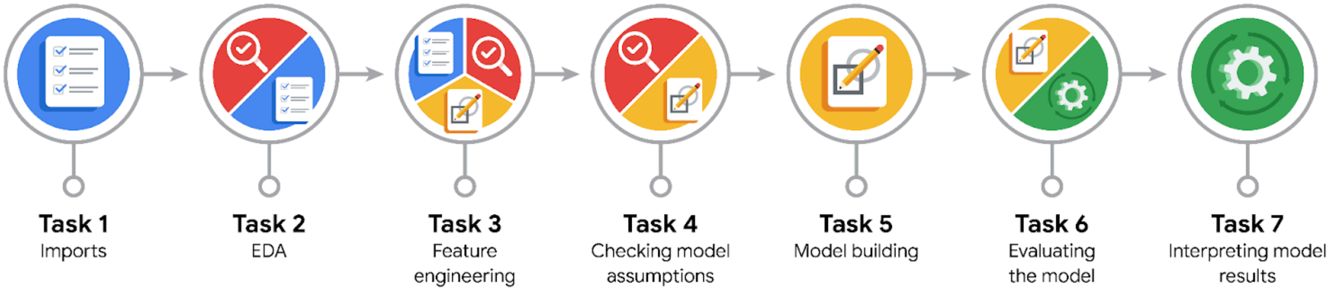
Relevant Interview Questions

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?

In the last part of the project, the data team needs to create a machine learning model that predicts whether a video is a claim or an opinion.

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

Some of the external stakeholders for this project are: TikTok's data team (Willow Jaffey- Data Science Lead, Rosie Mae Bradshaw- Data Science Manager, and Orion Rainier- Data Scientist); cross-departmental stakeholders within TikTok (Mary Joanna Rodgers- Project Management Officer, Margery Adebowale- Finance Lead, Americas, and Maika Abadi- Operations Lead); TikTok's Operations; and the leadership team at TikTok.

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

As a data professional in the Plan stage of the PACE workflow, I find myself leveraging a variety of resources to effectively conceptualize and strategize my project. Here are some of the most valuable ones: for Domain-Specific Knowledge, I reviewed Data dictionaries or metadata because these provide information about the structure and meaning of our data; for Project Management Tools, I utilized Jupyter Notebook since it is often used as a living document to outline project goals, data sources, and initial analysis plans; and for Data Exploration Tools, I considered Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn. For machine learning, I considered Python libraries such as sklearn and xgboost. To save our models once we fit them, I considered the pickle Python library.

- Do you have any ethical considerations at this stage?

Even in the initial planning stage of a data project, ethical considerations should be at the forefront. Here are some key areas to consider: for Data Privacy and Security, we ensure that we have the necessary rights and permissions to access and use the data; for Bias and Fairness, we consider potential biases in the data that could affect the results and plan for strategies to ensure that the project is fair and equitable; for Ethical Use of Data, we clearly define the purpose of the project and ensure that it is aligned with ethical principles and plan to prevent the misuse of data or results; and for Transparency and Accountability, we document our planning process to ensure transparency and accountability and involve relevant stakeholders in the planning process to gain their input and address potential concerns.

In our scenario, it appears worse for our model to predict false negatives and better for it to predict false positives when it makes a mistake. Thus, it is imperative to minimize false negatives because a claim that is misclassified as an opinion does not get reviewed when it violates the terms of service whereas an opinion misclassified as a claim goes to human review.

- Is my data reliable?

The data appears reliable for four reasons. First, the dataset has been automatically loaded in for us. Second, there are very few missing values relative to the number of samples in the dataset. Third, there are no duplicate observations in the data. Fourth, there is no need to check for and handle outliers.

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

Previous work with this data has revealed that there are ~20,000 videos in the sample. This is sufficient to conduct a rigorous model validation workflow.

- What data do I have/can I get?

Our dataset has 19382 entries with a total 12 columns. The dtypes of those columns: float64(5), int64(3), and object(4).

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

The metrics we use to evaluate success of our business/organizational objective are both precision, recall, and F1 scores. They shall help us pick the model that accurately classifies claims as claims and opinions as so.

PACE: Analyze Stage

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

The data dictionary shows that there is a column called `claim_status`. This is a binary value that indicates whether a video is a claim or an opinion. This will be the target variable. In other words, for each video, the model should predict whether the video is a claim or an opinion.

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

The dtypes of four columns are object. We encoded the target variable, two of catgorical variables using dummy encoding, and encoded the remaining object dtype using CountVectorizer. Thus, we converted the string data to dummy variables.

- Why did you select the X variables you did?

Building our model, we separated the label (y) from the features (X). We selected every X variable (or columns excluding the target variable) we did because stakeholders were interested in learning about the factors that are most important in determining whether a video is a claim or an opinion.

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

Data exploring, cleaning, and encoding are necessary for machine learning model building.

- What has the EDA told you?

First, there are very few missing values relative to the number of samples in the dataset; therefore, observations with missing values were dropped. Second, there are no duplicate observations in the data. Third, there are outliers in the dataset; however, observations with values deviating significantly from the norm remained because Tree-based models are robust to them. Fourth, there are few columns whose data types are 'object;' thus, these columns were encoded.

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

Some resources I find myself using as I complete this stage are: for data manipulation, importing and utilizing packages such as numpy and pandas; for data visualization, importing and utilizing packages such as matplotlib and seaborn; and for data preprocessing and data modeling, importing and utilizing packages such as sklearn and xgboost for OneHotEncoder, CountVectorizer, train_test_split, GridSearchCV, RandomForestClassifier and XGBClassifier to name a few.

PACE: Construct Stage

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

If a model is evaluated using samples that were also used to build or fine-tune that model, it likely will provide a biased evaluation. A potential overfitting issue could happen when fitting the model's scores on the test data.

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

All of these variables seem like meaningful predictors of whether the video is a claim or an opinion. In particular, those related to engagement levels generated by the video seem to be correlated with the target dependent variable.

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

The data team chose to develop two different machine learning models (i.e., random forest and XGBoost) to cross-compare results and obtain the model with the highest predictive power. The Random Forest model fits the data better than XGBoost model with f1-score of 1.

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

Because the model currently performs nearly perfectly, there is no need to engineer any new features.

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

Developing tree-based models since Tree-based models are robust to outliers or values that deviate significantly from the norm. Plus, when fitting and tuning our classification models, data professionals aim to minimize false positives and false negatives

PACE: Execute Stage

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

Splitting the data, fitting models and tuning hyperparameters on the training set, and performing final model selection on the validation set, we see from the recall , precision, and f1 scores of both the Tuned Random Forest and Tuned XGBoost that both performed exceptionally well. Yet, the Random Forest model fits the data better than XGBoost model with f1-score of 1. And according to the bar plot of the relative importance of the features in the model, the most predictive features were all related to engagement levels generated by the video. Both "video view count" and "video share count" accounted for the top two important features in the model.

- What are the criteria for model selection?

As this is a binary classification problem, it will be important to evaluate not just accuracy, but the balance of false positives and false negatives that the model's predictions provide. Therefore, precision, recall, and ultimately the F1 score will be excellent metrics to use.

In our scenario, since it is more important to minimize false negatives, the model evaluation metric will be “recall” and “f1”.

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

Our machine learning model provides an exceptional framework for predicting binary outcomes indicating whether a video is a claim or an opinion.

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

The model correctly predicts true negatives and true positives; yielding almost all true negatives and true positives and almost no false negatives or false positives.

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

Removing unimportant features (e.g., those not related to engagement levels generated by the video) from a random forest model can be a valuable strategy for improving its performance, interpretability, and efficiency. By carefully considering the techniques for identifying and removing unimportant features, data professionals can create more effective and practical machine learning models.

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

The data team recommends sending this machine learning classification model to our operations team to predict the status of claims made by users since the model successfully classified claims as claims and opinions as so.

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

Because of the nature of the data provided in this TikTok project, rather than excluding the `video_transcription_text` variable from our analysis in building our tree based model we may find it useful to extract numerical features from this variable. Although it is not a "categorical" variable, we may split text found in the `video_transcription_text` variable utilizing sklearn package and its CountVectorizer module.

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

A machine learning model would greatly assist in the effort to present human moderators with videos that are most likely to be in violation of TikTok's terms of service.

- Is my model ethical?

Any ethical considerations at this stage I may have are: for Model Fairness and Bias, we both ensure that our models are free from biases that could lead to unfair or discriminatory outcomes and continuously monitor model performance to identify and address biases; for Ethical Use of Results, we both ensure that the results of our analysis are used ethically and for its intended purposes and be transparent about the limitations and potential biases of our models; and for Accountability, we maintain records of the entire execution process for accountability and transparency. Thus, it seems to me that the model is ethical.

In our scenario, since it appears worse for our model to predict false negatives and better for it to predict false positives when it makes a mistake, our model appears ethical since our model almost has no false negatives or false positives

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

When a machine learning model, such as a random forest or XGBoost, makes a mistake, it means that it has misclassified or predicted an incorrect outcome for a particular data point. This can happen for several reasons: for Data Quality Issues, missing values can introduce bias or hinder the model's ability to learn accurate patterns; for Model Complexity, a model that is too complex may fit the training data too closely, leading to overfitting and poor generalization on new data; for Model Limitations, the model may make assumptions about the data that are not met in reality; and for Randomness, some algorithms, like random forest, involve randomness, which can lead to slight variations in results.

In our scenario, the risk of Overfitting exists where our model fits the training data well but makes poor generalizations with respect to new data.

Now, the impact of model mistakes on our specific use case is incorrect predictions leading to misclassifications, such as incorrectly identifying opinions as claims and claims as opinions. To mitigate the impact of such model mistakes, we experiment with different hyperparameters to optimize our model performances and use metrics like accuracy, precision, recall, F1-score, and accuracy to assess model performances. Also, it helps us to analyze the types of mistakes the model is making to understand the underlying reasons.