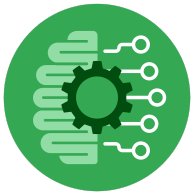
Course Six
The Nuts and Bolts of Machine Learning



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~~

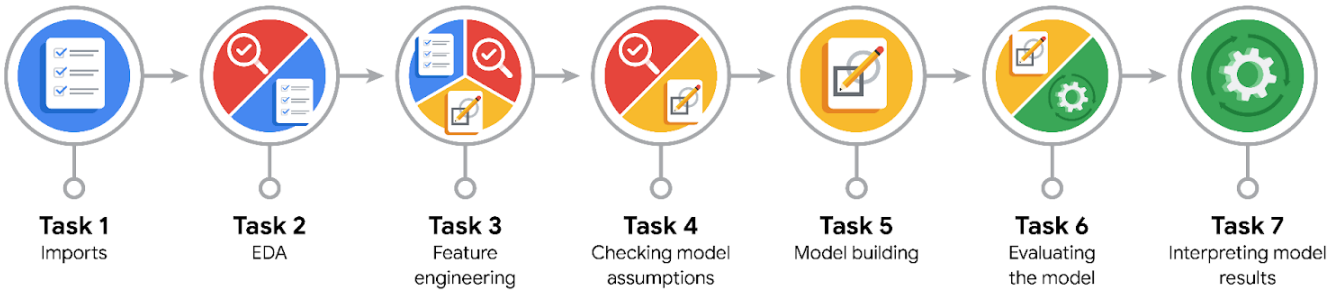
Relevant Interview Questions

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?

Predict if a customer will churn or be retained.

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

Waze.

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

Packages for numerics and dataframes.
Packages for visualization.
Packages for date conversions.

- Do you have any ethical considerations at this stage?

Waze will fail to take proactive measures to retain users who are likely to stop using the app.
Waze may take proactive measures to retain users who are NOT likely to churn. This may lead to an annoying or negative experience for loyal users of the app.

- Is my data reliable?

Yes.

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

Customer with churn or be retained label data.

- What data do I have/can I get?

Customer data.

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

accuracy, precision, recall, F-score, area under the ROC curve.

PACE: Analyze Stage

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

It’s still working.

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

Tree-based models are resilient to outliers, so there is no need to make any imputations.

- Why did you select the X variables you did?

Because it's help predict target feature.

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

For prepare data that can impact to model.

- What has the EDA told you?

Approximately 18% of the users in this dataset churned. This is an unbalanced dataset, but not extremely so. It can be modeled without any class rebalancing.

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

Packages for numerics and dataframes.
Packages for date conversions.

PACE: Construct Stage

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

No.

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

Label2 after convert churn and retained data with dummy variable.

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

All score is too low. It is not a great model.

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

New features could be engineered to try to generate better predictive signal, as they often do if you have domain knowledge.
In the case of this model, the engineered features made up over half of the top 10 most-predictive features used by the model.
It could also be helpful to reconstruct the model with different combinations of predictor variables to reduce noise from unpredictive features.

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

Packages for confusion matrix, model_selection, ensemble, xgboost.

PACE: Execute Stage

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

The XGBoost model fit the data better than the random forest model.

- What are the criteria for model selection?

Check model's validation score.

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

It is not a great model, but depending on how it's used it could still be useful.

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

New features could be engineered to try to generate better predictive signal, as they often do if you have domain knowledge.
In the case of this model, the engineered features made up over half of the top 10 most-predictive features used by the model.
It could also be helpful to reconstruct the model with different combinations of predictor variables to reduce noise from unpredictive features.

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

The only feature that can be cut is ID, since it doesn't contain any information relevant to churn.

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

It depends. What would the model be used for? If it's used to drive consequential business decisions, then no. The model is not a strong enough predictor, as made clear by its poor recall score. However, if the model is only being used to guide further exploratory efforts, then it can have value.

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

Will logistic regression be better ? because logistic regression models are easier to interpret and understand even with the same result.

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

Packages for confusion matrix.

- Is my model ethical?

Waze will fail to take proactive measures to retain users who are likely to stop using the app.
Waze may take proactive measures to retain users who are NOT likely to churn. This may lead to an annoying or negative experience for loyal users of the app.

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

It would be helpful to have drive-level information for each user (such as drive times, geographic locations, etc.).
It would probably also be helpful to have more granular data to know how users interact with the app. For example, how often do they report or confirm road hazard alerts?
Finally, it could be helpful to know the monthly count of unique starting and ending locations each driver inputs.