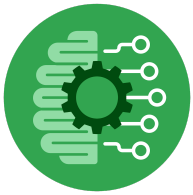
Course Six
The Nuts and Bolts of Machine Learning



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

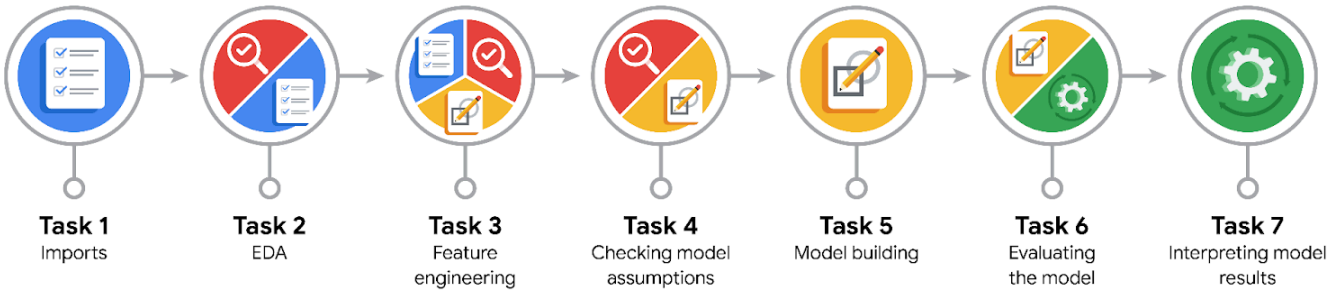
Relevant Interview Questions

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?

Building machine learning models (Random Forest and XGBoost) to predict user churn, improve the retention rate, and grow the business.

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

Data analysis team and cross-functional team members.

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

Using Jupyter Notebook to complete the tasks.

- Do you have any ethical considerations at this stage?

About the class imbalance, which might affect the model's predictions and accuracy.

- Is my data reliable?

Concerns such as missing values, outliers, and class imbalance might exist.

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

The data that all the hard work is done (well-cleaned, preprocessing, feature engineering, etc.)

- What data do I have/can I get?

The data of 15,000 users and the associated features.

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

Accuracy might be a choice, but it won't be for this objective due to the class imbalance. The recall score might be effective since we want to minimize the FN value, which the model incorrectly the churned users as retained.

PACE: Analyze Stage

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

It still works and is on the right track, so there's no need to revise. However, it's still worth monitoring on the track and revising on stuff if necessary.

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

This data won't break any assumptions of the model.

- Why did you select the X variables you did?

Select X variables that might have the most influence on the target variable. Things that might stand out when building these kinds of models like they can handle multicollinearity of variables, and they use a random subset of variables rather than accounting for all.

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

Performing EDA before constructing a model will make the model perform the best it can by dealing with things such as cleaning data, removing missing values, etc. Also, EDA helps to understand thoroughly what the data is given, understand variables, and address the problems that might arise. Besides, EDA helps to point out if the chosen model(s) will meet the objective and assumptions.

- What has the EDA told you?

By performing EDA, some variables might influence the target variable. Also, from the feature engineering, creating some new features that are worth predicting on the target one, as long as the selection of features and transformations.

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

Jupyter Notebook for Python code.

PACE: Construct Stage

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

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

Except for the target variable, ID, and object columns, the rest that I'll choose as independent variables. The advantage of building these kinds of models is that they can take up more variables to

have diversity in each estimator of the model. In other words, each estimator won't contain the full range of variables, but instead use a subset of it.

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

In both models using cross-validation, the accuracy and precision scores tend to be higher, but the recall score is pretty low. RF has the lowest recall score than the XGBoost model, and this applies the same when using the validation set to predict. In this case, since the recall score is the main metric, the XGBoost will be the champion model. But overall, both the models produce bad results which account for little improvement besides the logistic regression model.

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

Tweaking the hyperparameters might be the only option to improve the model's performance.

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

PACE: Execute Stage

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

Both RF and XGBoost perform better than the previous logistic model. However, both account for the bad results of predictions. XGBoost is then chosen as a final model. Using all the 3 sets of data, it accounts for the recall score (the main metric) below 20%. That's why it makes more negative values than positive ones. We might account for the top 10 features that influence the target variable, and half of them come from feature engineering. Further examining these features might be worth it. For sure, XGBoost isn't as easy to interpret as the logistic regression, and we can refer to it as a black-box model.

- What are the criteria for model selection?

It involves multiple criteria such as: how explainable is the model? How complex is it? Does it perform well on the unseen data? Do the score metrics produce good results? Etc.

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

Since the final model, XGBoost doesn't produce good results, it won't be acceptable for a business scenario. However, if it is used for further analysis, then it should be.

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

It might be worth improving, but it can come with a trade-off. For instance, we can adjust the threshold of probability (default is 0.5) to predict the retained and churned users. If we choose to improve the recall score, which is the main metric, we might accept to decrease in the precision score. After all, it all depends, so further examination with the data team might be helpful.

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

Some features wouldn't affect much to the target variable. We might account for the top % of features to discuss about their importance. Then, it might be worth to remove the unimportant features.

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

Since the model doesn't produce the best results, we won't account for deploying it for business scenarios.

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

It's worth further examining the feature importance to the model. Additionally, getting more granular data from users might be helpful to have diversity.

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

- Is my model ethical?

It's completely not. Based on the main metric used in this case, which is pretty low. If it chooses to deploy, it will hurt the business.

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

Since the recall score of ~15%, which is the model correctly predicts the churned rate. In the business case, this translates to an increase in churned users which is assumed from the model as retained.