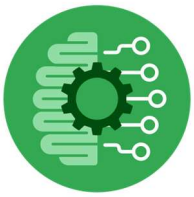
Course Six
The Nuts and Bolts of Machine Learning



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

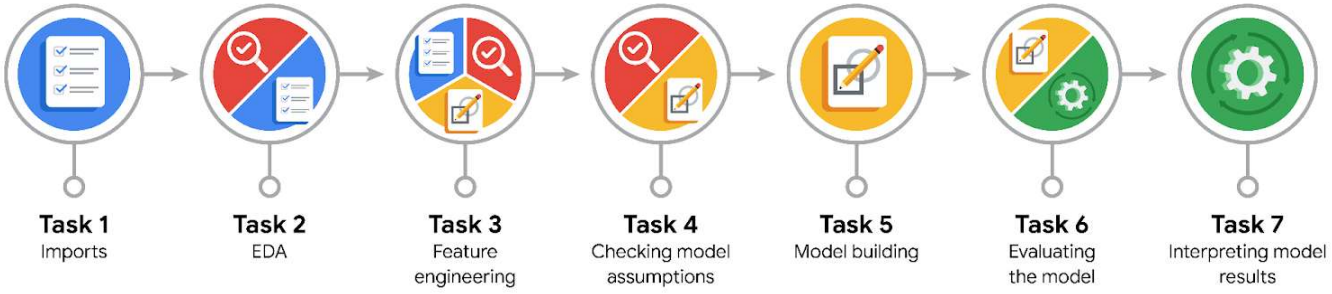
Relevant Interview Questions

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 overall objective of the project is to identify and reduce user churn in the Waze app. For this section, a machine learning model will be developed to predict user churn. Both XGBoost and Random Forest models will be built and tested to determine the superior performer. The recall score will be used as an evaluation metric, as the target variable is imbalanced and a false positive prediction is far less problematic than a false negative. Accuracy is unlikely to be a useful metric in this case, as the imbalance can lead to high accuracy for the majority class but poor prediction of the minority class.

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

1. Emerick Larson, Head of the Finance and Administration Department.
2. Ursula Sayo, Operations Manager.

These individuals are less technically savvy, so communication will be adapted accordingly.

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

Utilizing previous codebooks has been helpful for recalling code snippets and processes. Domain knowledge is also useful during this stage, as it provides insight into the significance of false positives or negatives in the context of the business problem. In this case, lacking thorough domain knowledge, further research is required, either independently or through collaboration with others.

- Do you have any ethical considerations at this stage?

When considering false negatives and false positives, false negatives are significantly more problematic. This would indicate that the model is failing to predict users who will churn, which could be very costly for the business. Beyond that, I have no ethical concerns about pursuing the development of this model, and it should provide valuable insight into the drivers of user churn.

- Is my data reliable?

Without a thorough understanding of how the data was collected, it's difficult to ascertain its reliability. Throughout the analysis thus far, concerns have arisen about the data potentially overrepresenting certain driving groups. This is evident in the high values associated with some variables, particularly those related to driving distances. Although there is some missing data, comprising less than 5% of the dataset, and previous exploratory data analysis suggests that the missing values are likely not random, the overall data structure appears sound.

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

I would like to know more about how the data itself was collected so that I can investigate further whether it violated any assumptions.

- What data do I have/can I get?

Much of the data, acquired from the Waze app, relates to driving metrics, such as distance, number of drives, number of sessions, and more. Beyond engineering new variables, it may be possible to acquire additional data related to geographic region or the number of minutes spent in the app during a session.

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

Since the risk of a false negative is far less significant than the risk of a false positive, recall will be used as an evaluation metric for this model. Missing someone who will churn is far more costly to the company than predicting someone will churn when they won't.

PACE: Analyze Stage

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

The company is attempting to address user churn within the Waze app by utilizing a machine learning model capable of accurately predicting user churn. By identifying users at risk of churning, the company can allocate resources to retention strategies, such as incentivizing users, actively seeking feedback on their app experience, and implementing other targeted interventions.

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

As far as we can understand without knowing how the data was initially collected, the data does not appear to violate any of the model assumptions. During a prior analysis, it was determined that the missing data had a high probability of being due to random causes. Additionally, although outliers exist in some of the columns, the models being used are resilient to them, so imputation is not required. While there is some multicollinearity between some of the features, these models handle it quite well, so they are being kept at this stage. The target variable exhibits class imbalance, but it's not significant enough to cause an issue at this stage. This may be investigated at a later stage.

- Why did you select the X variables you did?

The X variables were selected based on their relationship to the dependent variable, label2 (retained or churned). Through domain expertise and analysis of the independent variables' interactions with the dependent variable and their distributions, these features were identified as potential candidates. However, after model creation and testing, it's likely that some of these features will be removed through further analysis.

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

Some purposes of Exploratory Data Analysis (EDA) include: Understanding the distribution of variables, analyzing the relationships between variables and the dependent variable, handling missing values and outliers and engineering new features from existing ones, leveraging domain knowledge or statistical properties.

- What has the EDA told you?

The EDA conducted thus far has revealed the presence of missing values in the dataset. However, based on previous analysis and the fact that these missing values constitute less than 5% of the data and do not exhibit a non-random pattern, the corresponding rows were removed. Additionally, some engineered variables have generated extreme values that are unlikely to represent typical driver behavior. Further information regarding user groups is necessary to address this issue.

The dependent variable exhibits moderate class imbalance. Given that false negatives are less problematic than false positives, recall score will be used as the primary evaluation metric instead of accuracy or precision.

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

Other than referencing similar notebooks I've created, utilizing online resources and package documentation has been useful.

PACE: Construct Stage

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

During this phase, particularly when building the RandomForest and XGBoost models, one interesting thing was when finding the optimal hyperparameters, one can greatly influence the other. Finding a balance between hyperparameters to maximize the model evaluation score, while taking into account potential overfitting and underfitting of the training data can be a tricky process.

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

Effectively all independent variables were used in the development of the model, with the exception of 'id', 'device' and 'label.' By incorporating all independent variables, I was able to extract which features had the most predictive signal with the target variable, 'label2', which contained information on whether the client churned or was retained.

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

The initial two models had poor recall scores on the training, validation, and test sets. This is problematic as false negatives are far more significant than false positives. The threshold was then modified to a more appropriate value of 0.14 from the default of 0.5, which was more practical for this business problem. This raised the recall score to 0.499. However, it's important to note that the model would only successfully identify half of the people who will actually churn (recall) and would be correct only 30% of the time (precision).

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

I would conduct further investigation into which features have the most predictive signal, rather than employing feature elimination, whether backward or forward, to determine the most relevant and performative features. Additionally, further hyperparameter tuning can be completed, as well as additional data gathered, such as total drives since onboarding or other relevant engineered features.

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

Other than referencing similar notebooks I've created, utilizing online resources and package documentation has been useful.

PACE: Execute Stage

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

  The final model, with the modified threshold of 0.14, on average, predicts approximately 50% of people who will actually churn and is only correct approximately 30% of the time. The XGBoost model is an ensemble model that works by sequentially correcting the errors of previous models, typically decision trees. The final prediction is a weighted sum of predictions from all models.

- What are the criteria for model selection?

  For this particular business problem, where a false negative is significantly worse than a false positive, the criterion for model selection is the recall evaluation score.

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

  The final recall score achieved after testing two models and modifying the threshold was 0.499, with a precision score of approximately 30%. This essentially means the model can identify half of the people who will actually churn, but is only correct in identifying whether someone will churn 30% of the time. This is likely not high enough to be used in production, and at the very least, can be used as a tool for further investigation.

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

  The model can most certainly be improved through further hyperparameter tuning, acquiring more relevant data, engineering more predictive features, and through feature elimination methods.

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

  The feature importances extracted indicated there were some features that were less important than others, however, this is not the only gauge of a features predictive strength. Other metrics, such as their R^2 should also be taken into account, and through feature elimination methods these optimal features can be found.

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

  The model performs satisfactorily, but the predictive strength and consistency of it could be improved. Since false positives are relatively minor in impact, the model could still, to a degree, be used to try and identify users that may churn, even if it's only correct 30% of time.

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

  Insights can be provided into some of the most important features indicated by the model, and what further information could be acquired to potentially improve its predictive strength. Information can also be provided regarding how many false positives and negatives were predicted.

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

  Other than referencing similar notebooks I've created, utilizing online resources and package documentation has been useful.

- Is my model ethical?

  The model does not appear to violate any ethics rules. If the model is leveraged as it is, some users would receive some additional follow-up, such as receiving a survey or call with the intent of tracking their current satisfaction.

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

  When the model makes a false negative mistake, this is problematic, as it means a user who will churn was not predicted as such, which if missed, will have a far more significant impact on the business. If the model makes a false positive mistake, it means the model has predicted a user as going to churn, which has very little negative impact on the company and the user, as the most they may receive is a call or survey link inquiring about their current satisfaction with the app.