**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?

We aim to build and evaluate machine learning models (Random Forest and XGBoost) to predict user churn in the Waze app, thereby helping leadership make informed decisions to improve user retention and reduce churn.

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

The primary stakeholders are Emrick Larson (Finance and Administration Department Head), Harriet Hadzic (Director of Data Analysis), and other Waze leadership team members.

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

Python libraries for data analysis and machine learning (pandas, scikit-learn, XGBoost), Jupyter Notebook for coding and documenting the process, and the provided Waze dataset.

* Do you have any ethical considerations at this stage?

Yes, ensuring that the model does not unfairly target specific user groups or make biased decisions that could negatively impact user experience.

* Is my data reliable?

While the data appears consistent and free from obvious anomalies, it is crucial to remain vigilant for any inherent biases that could influence model predictions. We must ensure that our models do not inadvertently reinforce any biases present in the data.

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

Ideally, more granular data on user interactions with the app, such as drive times, geographic locations, user feedback, and app usage patterns.

* What data do I have/can I get?

We have data on user sessions, drives, device types, total sessions since onboarding, days since onboarding, navigations to favorite places, kilometers driven, drive duration, activity days, and driving days.

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

Recall, precision, F1-score, and accuracy are crucial metrics. Recall is particularly important to ensure we identify as many users likely to churn as possible.

**PACE: Analyze Stage**

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

The goal remains to build predictive models for user churn. The plan appears sound, though iterative improvements and feature engineering may be needed.

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

Tree-based models like Random Forest and XGBoost are robust to many data issues, including non-normality and outliers.

* Why did you select the X variables you did?

The selected variables (sessions, drives, device, total\_sessions, etc.) were chosen based on their relevance to user activity and potential impact on churn prediction.

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

EDA helps understand the data distribution, identify patterns, detect anomalies, and inform feature engineering.

* What has the EDA told you?

EDA revealed some outliers, the distribution of variables, and potential feature importance. It also highlighted the need for feature engineering.

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

Python libraries (pandas, numpy, matplotlib, seaborn), Jupyter Notebook, and documentation from previous analyses and EDA.

**PACE: Construct Stage**

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

The unrealistic speeds in the `km\_per\_hour` feature were noted. Addressing this would require clarification from Waze or data cleaning.

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

Variables like sessions, drives, total\_sessions, and activity\_days were chosen for their direct relevance to user activity and engagement.

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

The XGBoost model performed best with a validation recall score of 0.173.

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

Further feature engineering and hyperparameter tuning could improve the model. Adding more relevant features could also help.

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

Scikit-learn for model building, XGBoost library, Jupyter Notebook for experimentation, and documentation for hyperparameter tuning.

**PACE: Execute Stage**

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

The XGBoost model showed the best performance, emphasizing the importance of feature engineering. Features like `activity\_days` and `professional\_driver` were significant.

* What are the criteria for model selection?

Recall was the primary criterion due to the importance of identifying potential churners.

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

Yes, the XGBoost model provided better recall and overall performance. However, there is room for improvement.

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

Yes, by further tuning hyperparameters, improving feature engineering, and potentially gathering more detailed data.

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

Features such as `n\_days\_after\_onboarding`, `total\_navigations\_fav1`, and `total\_navigations\_fav2` had minimal impact. Removing these features could simplify the model without significantly affecting performance.

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

Focus on users with high activity\_days and professional\_driver scores for retention efforts. Consider gathering more granular data for future models.

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

Investigate the impact of user engagement strategies on churn, explore geographic patterns in user activity, and analyze feedback from app interactions.

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

I used a combination of Python libraries (scikit-learn, XGBoost), Jupyter Notebook for coding and documentation, and online resources for hyperparameter tuning and model evaluation techniques. Additionally, collaboration with team members and feedback from stakeholders helped refine the approach.

* Is my model ethical?

Yes, provided it does not unfairly target specific user groups or make biased decisions. The model should be monitored for fairness and adjusted as necessary to prevent discriminatory outcomes

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

False negatives result in missed opportunities to retain users, while false positives could annoy loyal users with unnecessary retention efforts. Balancing these errors is crucial for practical application.