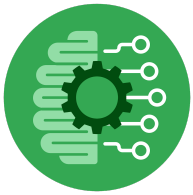
Course Six
The Nuts and Bolts of Machine Learning



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

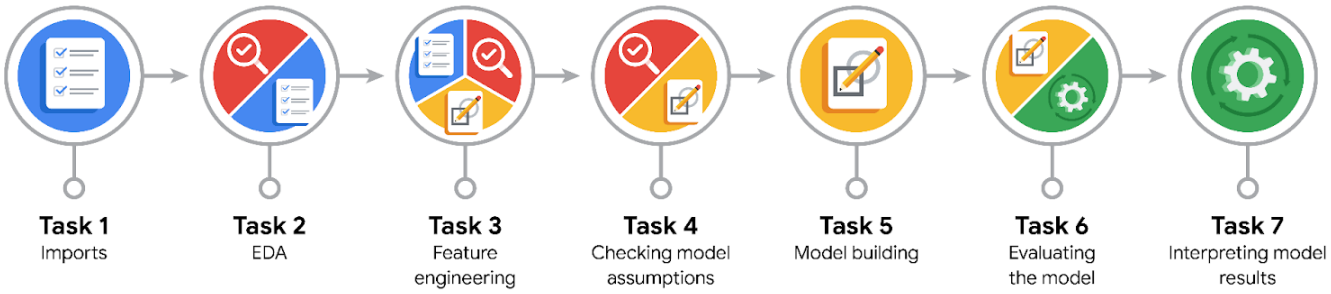
Relevant Interview Questions

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 develop and test machine learning models, specifically random forest and XGBoost, to predict user churn for Waze. The goal is to provide leadership with tools to make informed business decisions, preventing user churn, improving retention, and ultimately contributing to the growth of Waze's business.

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

The primary external stakeholders include Emrick Larson from the Finance and Administration Department and Harriet Hadzic, the Director of Data Analysis at Waze. These individuals will be crucial in the decision-making process based on the results of our machine learning models.

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

During this stage, we will leverage the Waze user data, the insights gained from previous analyses, and the expertise of the data team. Additionally, we may need to explore external resources related to machine learning methodologies and best practices.

- Do you have any ethical considerations at this stage?

Ethical considerations include ensuring the privacy and security of user data, transparent communication about the use of predictive models, and avoiding biases in the algorithms that could impact certain user groups unfairly. These considerations will guide our approach throughout the project.

- Is my data reliable?

Ensuring the reliability of the Waze user data is paramount. We will revisit the dataset, validate variables, and run necessary checks to guarantee the quality and accuracy of the data before building the machine learning models.

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

In an ideal scenario, having comprehensive user data with detailed information on user interactions, preferences, and behaviors would be valuable. This would include data on app usage patterns, feedback, and perhaps external factors influencing user engagement.

- What data do I have/can I get?

We currently possess Waze user data that has undergone cleaning and analysis, including the use of a binomial logistic regression model. We will assess the existing dataset and, if necessary, explore additional data sources that may enhance the predictive capabilities of our models.

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

The success of our business/organizational objective will be evaluated using key performance indicators (KPIs) such as model accuracy, precision, recall, and F1 score. These metrics will provide a comprehensive view of the model's predictive performance, ensuring its effectiveness in preventing user churn and enhancing user retention.

PACE: Analyze Stage

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

The initial goal is to predict user churn using machine learning models. The plan involves using both Random Forest and XGBoost models with hyperparameter tuning through GridSearchCV. The plan seems reasonable, but it may need adjustment based on the actual performance of the models on the validation data.

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

The data preprocessing and exploration stages, including handling missing values and creating new features, were designed to address potential issues. It's essential to assess if the assumptions of the models (e.g., independence of observations) hold given the data characteristics. Deviations from assumptions may impact the model's performance, and it's crucial to determine if it's acceptable in the context of the problem.

- Why did you select the X variables you did?

The selected features were chosen based on their potential relevance to predicting user churn. Features like session statistics, drive information, and activity days were considered as they might capture patterns indicative of user behavior. The final feature set aimed to provide meaningful information for the models without introducing unnecessary complexity.

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

Exploratory Data Analysis (EDA) serves several purposes:

- Identify patterns and trends in the data.

- Assess the distribution of variables and check for outliers.

- Examine relationships between variables.

- Inform feature selection by understanding variable importance.

- Detect potential issues like missing data or skewed distributions.

- Gain insights into the characteristics of the dataset, guiding model selection and hyperparameter tuning.

- What has the EDA told you?

EDA revealed insights into the distribution of features, relationships between variables, and potential patterns in user behavior. It guided decisions on feature engineering, such as creating new variables like "km_per_drive" and "percent_sessions_in_last_month." Understanding these patterns helps in constructing models that capture the underlying dynamics of user engagement.

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

During this stage, I've relied on various resources:

- Documentation and guides for the specific machine learning libraries used (e.g., scikit-learn, XGBoost).

- Reference materials on best practices for model evaluation and hyperparameter tuning.

- Online forums and communities for problem-solving and gaining insights from experiences of others.

- Domain-specific knowledge to interpret the relevance of features and potential implications of model predictions in the context of predicting user churn.

PACE: Construct Stage

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

The model's validation scores went down from the training scores across all metrics, but only by very little. This means that the model did not overfit the training data.

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

Same identified in previous EDA.

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

XGBoost validation scores : ~81% accuracy, ~44% precision, ~17% recall, ~24% score f1

Random Forest: ~81% accuracy, ~43% precision, ~13% recall, ~19% score f1

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

Hyperparameters optimization, Feature engineering.

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

Docs.

PACE: Execute Stage

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

Models demonstrate a critical need for additional data in order to more accurately predict user churn.

Engineered features accounted for six of the top 10 features: km_per_hour, percent_sessions_in_last_month, total_sessions_per_day, percent_of_drives_to_favorite, km_per_drive, km_per_driving_day.

The XGBoost model fit the data better than the random forest model. Additionally, it's important to call out that the recall score (17%) is nearly double the score from the previous logistic regression model built in Milestone 5, while still maintaining a similar accuracy and precision score.

The ensembles of tree-based models in this project milestone are more valuable than a singular logistic regression model because they achieve higher scores across all evaluation metrics and require less preprocessing of the data. However, it is more difficult to understand how they make their predictions.

- What are the criteria for model selection?

model	accuracy	precision	recall	f1

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

The XGBoost model fit the data better than the random forest model. Additionally, it's important to call out that the recall score (17%) is nearly double the score from the previous logistic regression model built in Milestone 5, while still maintaining a similar accuracy and precision score.

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

The ML models developed for Milestone 6 demonstrate a critical need for additional data in order to more accurately predict user churn.

This modeling effort confirms that the current data is insufficient to consistently predict churn. 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.

Since engineered features are a proven valuable tool for improving the performance of ML models, the Waze team recommends a second iteration of the User Churn Project.

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

Not really, the last feature is 'if the user is considered a professional driver' with an importance of 1 and before is 'if the device of the user is an iphone' with an importance on the F-score of

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

The ensembles of tree-based models in this project milestone are more valuable than a singular logistic regression model because they achieve higher scores across all evaluation metrics and require less preprocessing of the data. However, it is more difficult to understand how they make their predictions.

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

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.

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

Docs, Data

- Is my model ethical?

As long as its used to understand what drives user churn and the limits of such models are understood. It shall not be u

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

If the model predicts a false positive, a user who isn't likely to churn will join the churn funnel and will receive messages targeted at users who are likely to churn. This can be a mild annoyance for the user but it can also drive even more churn due to unsolicited and untargeted messages, so Waze needs to be very careful before using such a model to systematically predict if any user is likely to churn.