**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 purpose** of this model is to find factors that drive user churn.

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

Finance and Administration department

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

Python and its libraries, Jupyter notebook.

* Do you have any ethical considerations at this stage?

Yes, Data source, Presence of any personal identifiable information, data fairness.

* Is my data reliable?

Yes, as it is gotten from a primary source.

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

I would need data that has attributes that accurately describe user behavioral characteristics.

* What data do I have/can I get?

Waze dataset describing characteristics of Waze users behavior.

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

**Decrease in the user churn rate.** This will provide an objective means of measuring the impact of the model. We expect to see a decrease in churn following the commencement of active user engagement strategies on model-identified users most likely to churn.

**PACE: Analyze Stage**

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

No, it does not.

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

As a result of using a tree-based model there are no assumptions to be broken.

* Why did you select the X variables you did?

We selected the X variables that we did after familiarizing ourselves with the dataset.

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

Some of the purposes of EDA include managing missing data, removing outliers and duplicate values; however for this dataset being prepared for a tree-based model only missing data was managed as it is the model takes no assumptions. Feature engineering was extensively used.

* What has the EDA told you?

There were missing values from the label column about 700 observations.

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

Jupyter notebook, python and its libraries.

**PACE: Construct Stage**

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

No, there was nothing odd.

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

The X variables chosen include: 'sessions', 'drives', 'total\_sessions', 'n\_days\_after\_onboarding', 'total\_navigations\_fav1', 'total\_navigations\_fav2', 'driven\_km\_drives', 'duration\_minutes\_drives', 'activity\_days', 'driving\_days', 'km\_per\_driving\_day', 'percent\_sessions\_in\_last\_month', 'professional\_driver', 'total\_sessions\_per\_day', 'km\_per\_hour', 'km\_per\_drive', 'percent\_of\_sessions\_to\_favorite', 'device2', 'label2'

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

It fits the data poorly, with a validation score of 0.1834 in the XGBoost model compared to a validation score of 0.1282 in the Random Forest model. The XGBoost model was the champion model.

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

Yes, the model can be improved by adjusting the predicted probabilities.

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

Python and its libraries, jupyter notebook.

**PACE: Execute Stage**

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

1. No, I would not recommend using this model at least not to drive business decisions because it has a predictive capacity even less than random guess.

2. The tradeoff made is the reduced size of data used to train the model on the other hand there is better evaluation of model performance by using test holdout data.

3. Logistic regression models are easier to interpret and show the weight of individual features in prediction of target variable.

4. It requires little to no model assumptions, so they require much less rigorous EDA.

5. I would like to do further feature engineering.

6. I would like to have other specific data on users' interaction with the app.

* What are the criteria for model selection?

For a model to be selected it needs to have a prediction rate greater than or equal to 0.5.

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

Yes it does make sense but the results are not acceptable.

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

Yes, the model can be improved by adjusting the predicted probabilities. In doing so we were able to improve on the model’s predictive capabilities for user churn by about 204%. However, a tradeoff is that we can only be certain that about 30% of its predictions will be correct.

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

No feature was completely irrelevant however, type of device, and whether or not a user was a professional driver turned out to be the least important features.

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

**Feature Engineering**: Further feature engineering could potentially enhance model performance. This could involve creating new features from existing ones, looking for interactions between features, or applying transformations that might better represent the underlying patterns in the data.

**Model Selection**: The current models appear to be underperforming. It might be beneficial to try different types of models or tune the hyperparameters of the existing models. Additionally, considering ensemble methods, which combine the predictions of multiple models, could improve performance.

**User Engagement Strategies**: Given the goal of reducing user churn, it would be advantageous to develop strategies that foster user engagement. This could encompass personalized content, regular communication, or incentives for continued use.

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

1. Which features are most important in predicting user churn?

2. How easy is it to understand the models?

3. Are the models fitting too closely to the training data?

4. Are the models too simple to capture the data's structure?

5. How well do the models perform with new data?

6. How well do the models scale with more data?

7. How much computational power do the models require?

8. What is the potential impact of improving the models?

9. Are there any ethical concerns related to the models?

10. What are some potential areas for future work?

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

Jupyter notebook, python and its libraries.

* Is my model ethical?

Yes, it does not raise any ethical concerns.

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

A mistake by the model could mean that it incorrectly predicts whether a user will churn or not. This could lead to actions being taken that are not appropriate, such as providing unnecessary incentives to a user who is unlikely to churn, or failing to take necessary action to retain a user who is likely to churn.