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~~

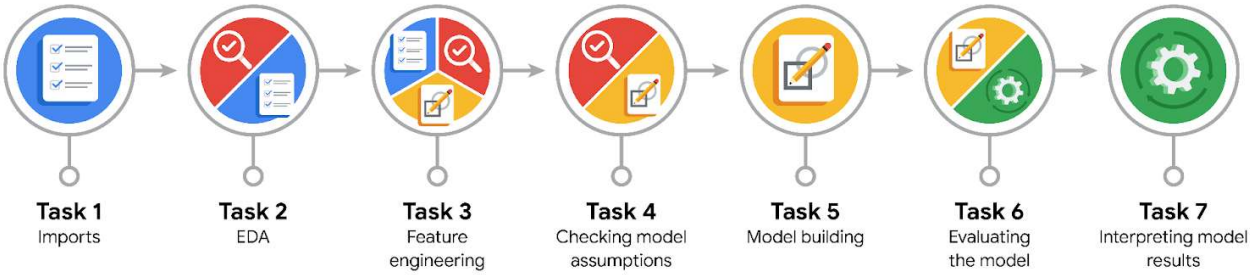
Relevant Interview Questions

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?

  To find the factors that drive user churn. and to predict whether a Waze user is retained or churned. To achieve the goal, I will build a machine learning model to predict user churn, and to get the best results, I will build and test two tree-based models: random forest and XGBoost.

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

  The external stakeholders are the Operations Manager and the Head of the Finance and Administration Department.

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

  The project dataset and Python, specifically the libraries for building Machine Learning models, and analyzing and evaluating them.

- Do you have any ethical considerations at this stage?

  At this stage the main ethical considerations regard the explainability of the model and in particular the XGBoost model, which can be considered a black box model. This means that the process and working of the model are hard to explain to stakeholders.

- Is my data reliable?

The data is first party data type, which means that it has been collected directly by Waze. This should make the data reliable because it is all internally generated and collected data.

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

In a perfect world I would like to see high quality data, all the variables that are relevant to solve the problem, sufficient quantity (large sample size, balanced classes), diverse and representative data, ethical and legal and compliance

- What data do I have/can I get?

I have a dataset with various features related to the use of the app.

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

In a classification problem the most appropriate metrics, that allow to evaluate the model on the base of false negatives and false positives, are precision and recall, and f1 as a combined measure of the two.

PACE: Analyze Stage

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

The business issue is to predict user churn and identify the factors that influence churn. The plan is to develop a random forest and XGBoost models to answer the business question. The plan does not need to be revised.

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

The target variable is imbalanced (20/80 %) but not at an extreme level. Many of the features present outliers but they are well managed by the planned machine learning models.

- Why did you select the X variables you did?

Because they are the features that potentially can have the most impact on user churn.

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

Some of the purposes of EDA are:

- cleaning the dataset (remove missing values and errors)

- verifying presence of outliers

- analyze the data to better understand the features

- What has the EDA told you?

- There were 700 missing values, but nothing indicates that they are due to a non-random process, thus being less than 5% of the observations, the corresponding rows were removed.

- Users that drove more km per driving day churned more (positive correlation with churn).

- Users that used more the app churned less, while 40% of the users who didn't use the app at all last month churned, nobody who used the app 30 days churned (activity_days negative correlation with churn).

- Users that drove more days churned less (negative correlation of driving_days with churn).

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

Mainly pandas and numpy to manipulate the dataset.

PACE: Construct Stage

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

The size of the dataset is quite small and after splitting it into training, validation and test sets, their size is reduced considerably and impacts the quality of the training. This could be improved by eliminating the validation set and using a bigger training set for training and validation. Having a separate validation and test set though improves the estimation of the future performance of the model on new data (model generalization).

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

Since tree-based models can handle multicollinearity all the features in the dataset have been included, except for "ID" which does not contain any information relevant for churn. Some new variables were added because of feature engineering.

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

The random forest model has a validation score (recall) of 13.41 while the XGBoost model has a score of 14.6. These are mediocre values for the scores which indicate low predictive values.

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

The hyperparameters have already been tuned, one possibility to improve the model is to add engineered features.

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

The scikit learn library for Random Forest modeling, evaluation metrics and model processing.,Tthe ensemble library for XGBoost

PACE: Execute Stage

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

The XGBoost model performed marginally better than the random forest model, and both performed better than the Logistic regression model. Engineered features accounted for six of the top 10 features (and three of the top five).

- What are the criteria for model selection?

The criteria for model selection are the “recall’ metric which helps to better evaluate the false negatives (users predicted not to churn when they churned).

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

The metrics of the models and in particular Recall are quite low, this indicates a poor predictive power of the model. The results are not acceptable if the models must be used to make business decisions.

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

The model could be marginally improved by adding new engineered features, but it appears that the features of the dataset are simply not good predictors. The best way to improve model performance would be to add new features.

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

“professional driver” and “device 2” had very low importance, thus they could be taken out without impacting the model.

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

The results of the analysis evidence that additional data is necessary to predict churn.

While the models developed establish a baseline, A phase 2 should focus on integrating telemetry data to improve predictive signals

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

I could address questions regarding the relative importance of the features regarding the target variable.

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

Mainly the scikit learn library (metrics) for model evaluation and the XGBoost library (plot importance) for analysis of features importance.

- Is my model ethical?

The model cannot be considered ethical towards the company given the low predictive power and the wrong indications it might provide for leadership when making decisions.

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

There are two types of mistakes:

- false positives which means that users are predicted to churn when in fact they don't, the cost of this mistake can be quite limited depending on the type of retention action performed on the user,

- false negative when the users are predicted not to churn when they do, in this case the cost is higher and corresponds to the cost of losing a user.