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

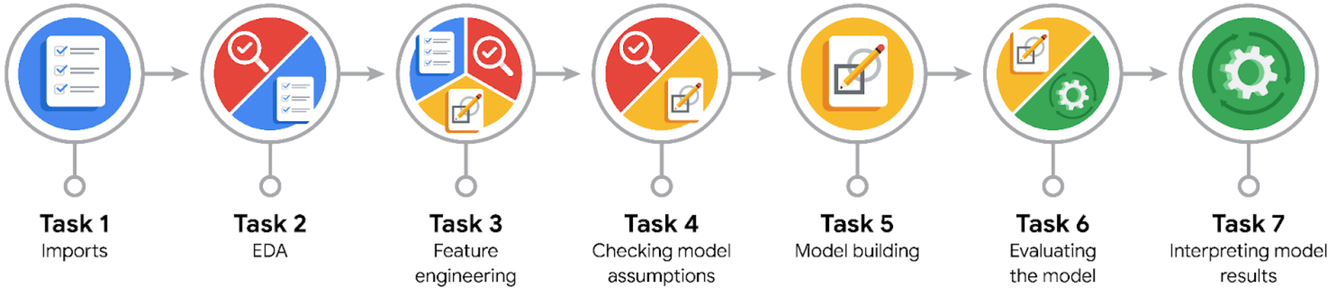
Relevant Interview Questions

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.



Task 1 Imports

Task 2 EDA

Task 3 Feature engineering

Task 4 Checking model assumptions

Task 5 Model building

Task 6 Evaluating the model

Task 7 Interpreting model results

Data Project Questions & Considerations

PACE: Plan Stage

- What are you trying to solve or accomplish?

I am being asked to prepare the data for modelling, compute feature engineering and to build two tree-based models and evaluate their performance to classify whether a user will churn or be retained.

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

Harriet Hadzic (Director of Data Analysis) and Emrick Larson (Finance and Administration Department Head).

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

Pandas, NumPy, matplotlib, seaborn, xgboost and scikit-learn.

- Do you have any ethical considerations at this stage?

Yes my ethical considerations are of that if the model predicts a false negative or a false positive and its affect on Waze. Other ethical considerations are transparency and the ethical use of the model.

- Is my data reliable?

Yes my data is reliable because it is primary data from the Waze company.

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

The additional data that I would like to see is drive level information such as drive times and geographic locations, as well as more granular data so that we know how the users interact with the Waze app. Additionally, it can be insightful to know the monthly count of the unique starting and ending locations of each of the drivers inputs for their journeys.

- What data do I have/can I get?

The data I can get is through feature engineering new features that are constructed from a combination of original features. An example of this is km_per_driving_day which is constructed from driven_km_drives and driving_days.

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

The metric that should be used to evaluate the success of the business objective of predicting and preventing user churn is False Negative count and Recall. This is because False Negatives is when the model has predicted a user as retaining when in fact they actually left, and so predicting user churn accurately in order to in act preventative measures is the core of the business objective. Therefore, monitoring the False Negative count is extremely important, and in turn making recall the focused evaluation metric as it is costly to have a False Negative compared to a False Positive.

PACE: Analyze Stage

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

The plan still works and does not need revising.

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

No, the data does not break the assumptions of the model as the model is resilient to outliers and the categorical variables have been encoded.

- Why did you select the X variables you did?

I chose the X variables I did due to their relevance, such as not including label and label2 as this is the target variable, and not including device as this is the categorical version of device2 and so can not be used by the model.

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

Some of the purposes of EDA before constructing a model is to help the analyst identify missing data which in turn helps to make the decisions of their exclusion or inclusion by substituting values with dataset means, medians and other methods. Additionally, it can be useful to engineer more variables via multiplying the variables together or by calculating the ratio between two variables.

- What has the EDA told you?

That there are 700 missing rows in the target variables, there are a majority of retained users compared to churned users and that there are outliers in many of the features due to extreme users.

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

NumPy and Pandas.

PACE: Construct Stage

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

Yes both of the models evaluation metrics are lower than I would have expected but are still considerably higher than the previous logistic regression's performance, with the XGBoost model performing the highest. Yes this can be fixed with some more data such as with new features being engineered as to improve the predictive signal of the model.

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

I chose the label variable, but encoded as label2, this was because it was the variable responsible for stating whether a user churned or retained with the Waze app.

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

Both of the models fit the data very well and this is evident in the difference in their evaluation metric scores from the GridSearch to the validation set. The differences was a slight decrease in the evaluation scores on the validation set, thereby meaning that the model did not overfit the data. The validation score of the champion XGBoost model was an Accuracy of 80%, a Recall of 21%, a Precision of 36% and a F1-Score of 27%.

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

Yes, the data itself can be improved in order to improve the predictive signal for the model, by creating new features from current ones via feature engineering. Another way is to get more data which could provide better predictive insight for the model such as user drive level information and granular data. Additionally, the models performance can be improved with the optimization of the decision threshold for the model, so that it is optimized for Recall and catches more users who actually churns even with this causing an increase in misclassifying users as churning when they actually retained.

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

Pandas, NumPy, scikit-learn and xgboost.

PACE: Execute Stage

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

Some key insights that emerged from my model was its evaluation metrics on the test set, the champion model achieved a Accuracy of 80%, a Precision of 37%, a Recall of 21% and a F1-Score of 27%. From the confusion matrix of the champion model on the test set, the model classified 2166 True Negatives, 110 True Positives, 187 False Positives and 397 False Negatives, which can be due to the lack of predictive signal in the dataset and so the model was not able to deliver consistently accurate predictions. Finally, the feature importance plot from the model placed n_days_after_onboarding as the most importance feature relevant to the models predictions, with km_per_hour, duration_minutes_drives, percent_sessions_in_last_month and total_sessions_per_day coming afterwards.

- What are the criteria for model selection?

That the model produce a great Recall score on the validation data while maintaining appropriate accuracy, which the XGBoost model did.

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

Yes my model does make sense and the final results are acceptable due to their being a lack of predictive signal in the dataset and even the most complex of algorithms will not be able to deliver consistently accurate predictions with data that lacks predictive signal.

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

Yes I do think the model could be improved starting with the data that the model is trained on, such as the data for the model could be improved as to generate more predictive signal. For example, majority of the top features in the feature importance plot were engineered via feature engineering thereby showing how it can increase model performance. Finally, decision threshold tuning could be carried out on the champion model as to optimize for Recall so that the model classifies more of the users who actually churn even with this causing an increase in False Positives.

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

Device2 did not seem important at all, with it being the 2nd to last in the feature importance plot and so I would remove this variable as it contributes nothing to the overall predictions and the variable itself being evenly distributed across churned and retained users. I would also remove professional_driver as this was the last variable in the feature importance plot as contributed nothing to the models predictions.

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

The recommendation I propose based on the models built is to not use this model to drive consequential business decisions as it is not suitable as a predictor as made clear by the many misclassifications. The model however can be kept for exploratory reasons as it can help develop further insight and improved modelling that may achieve the business objective. Additionally, I recommend that new features be engineered to increase predictive signal in the data for the model and the decision threshold for the model be optimized for Recall as to improve classification performance. Finally, to design and test the retention strategies by working in collaboration with the Marketing and Production teams to develop targeted retention strategies for high-risk users, and perform A/B testing in order to measure the effectiveness of these interventions.

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

What are the most important variables influencing user churn? How did the model perform on unseen data? What features were engineered to improve prediction performance?

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

Pandas, xgboost, scikit-learn and matplotlib.

- Is my model ethical?

Yes my model is ethical as it predicts user churn solely based on behavioural and engagement metrics and does not use sensitive data that display personal attributes such as race, gender, sexuality or income, thereby reducing the risk of discriminatory bias. This actual ethical risks comes from the models False Negatives and False Positives but due to the model being optimised for Recall thereby prioritising the classification of at-risk users, it aligns with the business objective of predicting and reducing user churn while minimising user harm. Finally, to ensure that the models ethical standards are upheld it is important that the model is transparent about not just the model but how its predictions will be used.

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

When the model makes a mistake it either classifies a False Positive or a False Negative. The way in which this translates to our use case is that when the model classifies a False Positive, this is when a user is predicted as being churned when in reality they retained and so has the preventative measures in acted on them, such as a push notification or email. In the case of False Negative, this is when the model predicts a user as being retained when in actuality they churned, which is very bad in our case as we want to capture as many people who actually churned as possible. If a user has been classified as a False Negative, then they won't have any of the preventative measures placed on them and will be dissatisfied with the Waze without Waze leadership understanding or knowing why.