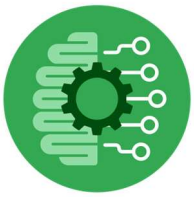
Course Six

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

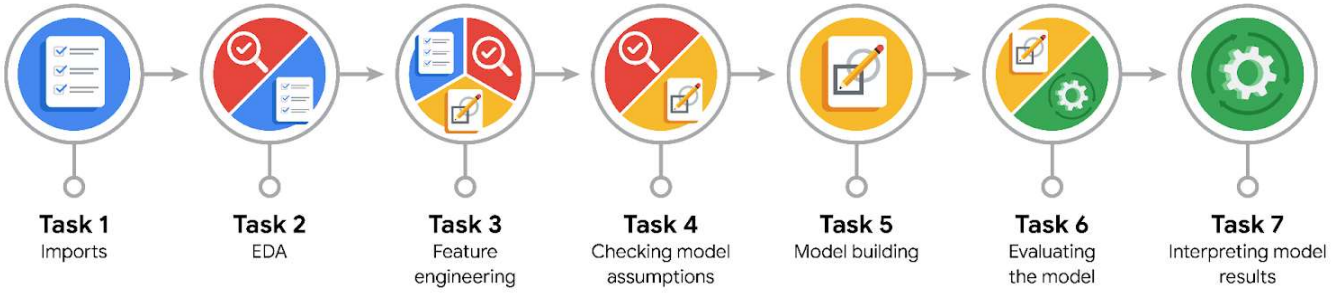
Relevant Interview Questions

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 goal is to build a Random Forest machine learning model that will aid in generating more revenue for the New York City TLC taxi drivers.

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

  The external stakeholders are Juliana Soto and Titus Nelson, both members of the New York City TLC.

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

  External research into ethical considerations around taxi tipping culture and expectations.

- Do you have any ethical considerations at this stage?

  The model inherently presents significant ethical concerns. As it currently stands, proceeding with a model that predicts whether a customer will tip or not before being picked up is likely to lead to discrimination, ultimately resulting in dissatisfaction among customers, drivers, reputational damage and significant revenue loss.

- Is my data reliable?

To the best of my knowledge, yes. While I don't have a complete understanding of the data collection process, I assume it was conducted in a way that is representative of the population. In terms of data integrity, it was cleaned, normalized, and reformatted as needed.

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

I would like to see historical tipping data for both cash and electronic payments. This would likely enhance the model's robustness by accounting for customers who don't pay electronically and potentially identifying patterns in cash tipping, such as a preference for rounding to whole numbers like $10 or $20.

- What data do I have/can I get?

The data I currently have available provides significant information on individual customer trips. Additionally, I'm able to acquire the data from the prior analysis.

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

Both false positives and false negatives can have substantial consequences. To minimize these errors and achieve optimal performance, the F1 score is a suitable metric. It effectively balances precision and recall, providing a comprehensive measure of model accuracy.

PACE: Analyze Stage

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

Considering the ethical concerns and modifying the model's approach to predicting generous tippers, the model remains viable. While the plan required adjustments, the final outcome should still deliver an effective model for the New York City TLC.

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

The model under development is robust to non-normal distributions and non-linear data. However, it requires independent observations to avoid potential multicollinearity issues. While multicollinearity can be a concern, it's less of an issue for predictive modeling. As long as the model demonstrates significant predictive power, it can still be effective. Moreover, the data exhibits minimal class imbalance, and potential outliers or problematic data have been addressed during the exploratory data analysis (EDA) phase.

- Why did you select the X variables you did?

The X variables were selected because, upon further investigation, they appeared to be the most relevant to the business problem of determining whether a customer is a generous tipper or not. Additionally, the engineered variables were created from other variables that were hypothesized to have predictive power and were relevant. In this model, introducing numerous potential relevant variables and using GridSearch and Feature Selection to identify the most predictive ones allowed me to narrow down the feature set.

At this point, based on Feature Selection rankings and other metrics and strategies, such as P-value, R, R-squared/Adjusted R-squared, further refinement of variable selection could be completed to attempt to improve the model's performance further.

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

It is imperative to clean and optimize the data before building a model. This process includes addressing outliers, normalization, missing values, scaling, and more. The majority of time is spent on this stage, ensuring that the models have a strong foundation.

- What has the EDA told you?

Most of the cleaning tasks were completed before this phase of the project. However, we engineered additional predictive variables from existing ones and the target variable 'generous,' which the model would use to classify customers as generous or non-generous tippers. Moreover, we converted the data types of some variables and adjusted their date formatting.

A limited exploratory data analysis during this stage revealed no significant outliers, missing values, or normalization needs. Additionally, the class imbalance in the newly created target variable was minimal. If it had been more pronounced, further steps would have been necessary to balance the classes.

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

I leveraged several common Python libraries, including NumPy, Pandas, Scikit-learn, and XGBoost, for this project. To train the model, I combined previously prepared data and engineered additional features. Moreover, I consulted previous projects, particularly those involving Random Forest and XGBoost models, as a reference for building our own.

PACE: Construct Stage

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

During this phase, I didn't identify any significant anomalies or issues within the data. However, after training the model and evaluating its performance on the test set, I observed a higher rate of false positives (Type I errors) compared to false negatives (Type II errors). In this context, a false positive is

less desirable than a false negative because it's preferable for a driver to receive an unexpected generous tip rather than not receiving an expected one.

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

The independent variables were chosen based on their assumed relevance to the engineered target variable, "generous." Some variables were carried over from a previous stage of the project, where they had been engineered earlier, while others were left unchanged in the dataset. This was previously investigated by determining correlations between independent variables, checking statistical significance through p-values, checking model fit with r-squared/adjusted r-squared, and other methods. Additionally, some new variables were engineered from existing ones.

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

The model performs satisfactorily. The F1-score, which was 0.748967, and an accuracy of 0.711186 were used to assess its performance. The model seems to predict more false positives (Type 1 errors) than false negatives (Type 2 errors), which is more problematic for this business problem. That being said, the model performs decently on unseen data, as demonstrated by its performance on the test set. Ideally, a separate validation set should have been used to select the best model, and then the test set could have been used to obtain the final result.

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

I believe the model can be further improved by using additional feature selection techniques to focus on only the most influential independent variables. This could be achieved through backward or forward elimination, recursive feature elimination, or by examining p-values and adjusted r-squared values. Additionally, the model's hyperparameters could likely be further tuned to enhance performance, and new features could potentially be engineered that are more predictive than those currently used in the model.

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

First and foremost, I reviewed prior documentation and code to refresh my memory of the steps already taken with the data. I utilized many of the standard statistical, visualization, and modeling packages, conducted additional exploratory data analysis on existing and newly engineered features, and created a confusion matrix to assess the model's performance. Moreover, I obtained various performance metrics, including recall, accuracy, precision, and F1 score.

PACE: Execute Stage

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

The model performed satisfactorily on the test data, achieving an F1 score of 0.748967 and an accuracy score of 0.711186. While the model is more prone to false positives than false negatives, which is a potential concern, it's not a major issue given its overall performance. It's recommended to test the model on a small subset of taxi drivers to assess its practical performance and reception.

- What are the criteria for model selection?

For this model, the criteria used were how well it performed when assessed based on F1 and Accuracy scores, and the outcome of the confusion matrix. The confusion matrix provided valuable insights, particularly regarding the false positives and false negatives that this model is more likely to predict.

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

Although Random Forest models are inherently complex and difficult to interpret at a fundamental level, we can leverage feature importance extraction to gain a higher-level understanding of which features were deemed most important to the final outcome. Furthermore, the minimal feature imbalance aligns with the results obtained from the Confusion Matrix and the Accuracy, Precision, Recall, and F1 Score metrics. Therefore, while a complete understanding of the model's internal workings may be elusive, the final results are acceptable given the insights gained.

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

There is potential for further model improvement. This could involve engineering new, relevant features from existing data, concatenating additional datasets, fine-tuning hyperparameters, or exploring more feature selection methods to retain only the most relevant features. Additionally, investigating other models, such as XGBoost or Support Vector Machines, might lead to a better overall final model.

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

Several features were dropped because they were either irrelevant to the business question or became redundant after new features were engineered.

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

At this stage, the final model performed reasonably well on the unseen test data. Further testing should be conducted with a small subset of taxi drivers to gather their feedback and assess its performance in a production environment.

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

How well did the model perform on the test set and the sample taxi drivers the deployed model was tested on? What aspects of the model could be further improved? Is there additional data that could enhance the model during a future iteration? How significant are the false positives and false negatives predicted by the model, and what potential negative business impacts could they have?

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

I've been referencing previous project documents and metrics from the final models' performance evaluation.

- Is my model ethical?

The first iteration of the model was ethically questionable, as it would have incentivized drivers to prioritize customers who tipped. This would have had a significant negative impact on the business, as many customers would have been overlooked, leading to a loss of revenue and declining customer satisfaction.

To address this issue, the model was modified to predict whether a customer was a generous tipper or not. While this approach may still have some drawbacks, it is a significant improvement overall.

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

Based on the confusion matrix and metrics, the model appears to be more prone to false positives than false negatives. This is a less desirable outcome, as it means the model incorrectly predicts a customer as a generous tipper. This can negatively impact driver morale and potentially erode trust in the model over time.