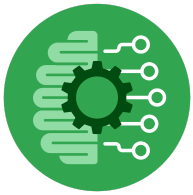
Course Six
The Nuts and Bolts of Machine Learning



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

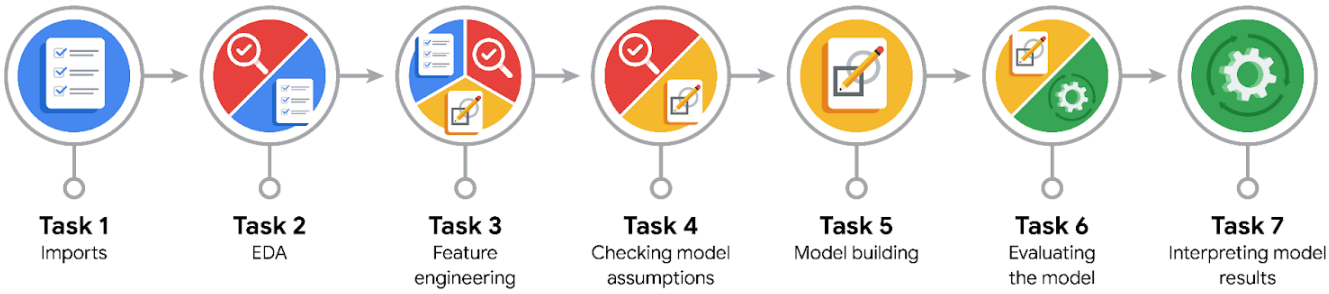
Relevant Interview Questions

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 are trying to build a machine learning model that predicts whether a taxi customer will be a generous tipper (tipping 20% or more) to help taxi drivers increase their earnings.

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

  Taxi drivers, the New York City Taxi & Limousine Commission (New York City TLC), and potentially app developers.

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

  The project notebook, the provided dataset, and Python libraries like pandas and scikit-learn.

- Do you have any ethical considerations at this stage?

  Yes, we need to consider potential biases in the model and ensure fairness and privacy. We also need to be mindful of the impact of model errors on drivers and customers.

- Is my data reliable?

  We need to assess the data for completeness, accuracy, and potential biases, especially concerning cash tips and customer demographics. The data concerning credit card transactions is reliable, but the absence of cash data is a weakness.

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

Ideally, we would have customer tipping history, cash tip data, detailed customer demographics, driver ratings, and real-time traffic data.

- What data do I have/can I get?

We have the provided taxi trip dataset with trip details, fare information, and pickup/dropoff locations. We can derive additional features like time of day and day of the week.

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

The F1-score is the most suitable metric because it balances precision and recall, which are equally important in this scenario. We must also consider the cost of false positives and false negatives equally.

PACE: Analyze Stage

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

Yes, the goal remains to predict generous tippers. The plan is still valid, but we need to ensure careful feature engineering and model selection.

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

We need to check for multicollinearity, outliers, and non-normality. We can address these issues through data preprocessing and feature engineering. It is important to remember that the data is only credit card transactions.

- Why did you select the X variables you did?

We selected features that logically influence tipping behavior, such as trip characteristics, fare-related data, and temporal features.

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

EDA helps us understand the data, identify patterns, find outliers, and discover relationships between variables.

- What has the EDA told you?

EDA helped us understand the distribution of tips, correlations between features, and the need for feature engineering, such as creating time of day and day of week features.

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

Python libraries like pandas, matplotlib, and seaborn for data exploration and visualization.

PACE: Construct Stage

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

We noticed outliers in trip duration and fare amount. We addressed this by filtering for credit card transactions and rounding tip percentage.

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

We chose features like vendor ID, passenger count, rate code ID, pickup/dropoff locations, mean duration, mean distance, predicted fare, day of the week, time of day, and month, as they are likely to influence tipping behavior.

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

The Random Forest model achieved a good F1-score on the cross-validation and test sets, indicating a reasonable fit.

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

We can further tune hyperparameters, explore other algorithms like XGBoost, and engineer additional features to potentially improve performance.

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

Scikit-learn for model building, evaluation, and hyperparameter tuning.

PACE: Execute Stage

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

The Random Forest model, which outperformed XGBoost in this scenario, highlighted several key insights:

- **VendorID** is the most influential feature, indicating significant differences in tipping patterns between vendors. This suggests potential variations in customer demographics, service quality, or pricing strategies.

- **Predicted fare, mean duration, and mean distance** are also crucial predictors, implying that trip characteristics strongly correlate with tipping behavior.

- The model operates as an ensemble of decision trees, aggregating their predictions to provide a robust and accurate outcome. It captures complex, non-linear relationships within the data, which is essential for predicting nuanced behavior like tipping.

- The model has more false positives than false negatives. Meaning it is more likely to predict a generous tip, when one is not given.

Essentially, the model identifies patterns in trip and vendor data to predict the likelihood of a generous tip.

- What are the criteria for model selection?

The primary criterion for model selection was the **F1-score**, as it provides a balanced measure of precision and recall. Given that the costs of false positives and false negatives are relatively equal in this use case, the F1-score offered the most appropriate evaluation metric. We also considered accuracy, but the F1-score was given priority. We also used the confusion matrix to evaluate the types of errors that the model was producing.

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

Yes, the model's results are sensible and acceptable. The Random Forest model achieved a solid F1-score and accuracy, demonstrating its ability to capture meaningful patterns in the data. The model is also able to correctly identify a high percentage of generous tippers. The fact that the model is more likely to give false positives is also acceptable, because it is better for the driver to be pleasantly surprised, then disappointed. However, continuous monitoring and potential refinements are necessary to ensure its long-term effectiveness.

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

Yes, the model can be further improved through:

- **Enhanced feature engineering:** Incorporating more granular location data, real-time traffic information, and customer feedback could provide valuable insights.

- **Additional data:** Gaining access to cash tip data and customer tipping history would significantly enhance the model's predictive power.

- **Hyperparameter tuning:** More extensive tuning of the chosen model, and exploration of other models, could lead to better results.

- **Addressing class imbalance:** If the class imbalance becomes more pronounced in future data, techniques like oversampling or undersampling could be employed.

- Collecting Driver data: Collecting data about the driver, and the vehicle they are driving, could provide more predictive power.

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

Some location IDs exhibited relatively low importance. Removing these features could simplify the model and potentially improve its generalization without significantly impacting performance. However, it's crucial to evaluate the impact of feature removal on model performance through rigorous testing. The model could run faster, and be less complex, if the less important features are removed.

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

I recommend:

- Deploying the Random Forest model in a pilot program to gather real-world feedback and assess its impact on driver earnings and customer satisfaction.

- Implementing a system to collect cash tip data and customer tipping history to enhance the model's predictive power.

- Providing drivers with clear explanations of how the model works and its limitations to manage expectations and build trust.

- Continuously monitoring the model's performance and retraining it as needed to adapt to changing data patterns.

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

We could explore:

- How tipping behavior varies across different times of day, locations, and events.
- The impact of driver ratings and vehicle types on tipping.
- The effectiveness of different incentive programs for drivers based on model predictions.
- How real time traffic, and event data, influence tip amounts.
- How customer feedback can be used to improve driver performance, and tip amounts.

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

I used Python libraries such as scikit-learn for model evaluation and visualization tools like matplotlib and seaborn to present the results. I also used documentation for interpreting model metrics and generating reports.

- Is my model ethical?

The model is designed to assist drivers in increasing their earnings, not to discriminate against customers. However, it's essential to:

- Ensure transparency in how the model works.
- Continuously monitor for potential biases.
- Avoid using the model in ways that could lead to discriminatory practices.
- Be mindful of customer privacy.

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

When the model predicts a generous tip that doesn't happen (false positive), the driver might experience disappointment. When it misses a generous tip (false negative), the driver misses an opportunity. These errors can impact driver morale and trust in the app. Therefore, minimizing both types of errors is crucial for the app's success. The model is more likely to produce false positives, which means drivers are more likely to be disappointed, then miss out on a good tip.