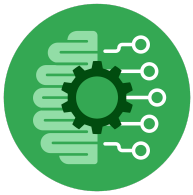
Course Six
The Nuts and Bolts of Machine Learning



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

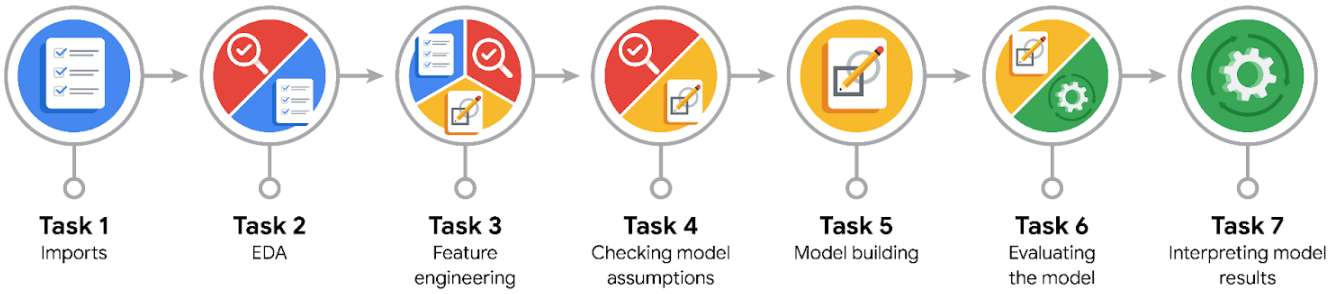
Relevant Interview Questions

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 aiming to build a predictive model using random forest to determine whether a rider will give a generous tip (>= 20%) or not based on various factors. By identifying the key variables that influence gratuity, we aim to provide actionable insights to the New York City TLC that can help improve driver satisfaction and potentially increase gratuity amounts.

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

The external stakeholders include the New York City TLC, represented by Titus Nelson, the Operations Manager, and other leadership members. Additionally, we will present our findings and results to Automatidata's leadership, including Juliana Soto, Department Head, and Udo Bankole, Director of Data Analysis.

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

We will leverage the New York City TLC dataset, which contains information about taxi rides, including variables related to gratuity, rider behavior, trip details, and other relevant factors. Our analysis and

modeling will be conducted using Python and its data science libraries, including pandas, scikit-learn, and XGBoost for the random forest model.

- Do you have any ethical considerations at this stage?

Absolutely, ethical considerations are crucial. We must ensure the responsible use of data and model predictions. Protecting rider privacy and preventing any form of bias in the model's predictions are paramount. We'll also need to address potential issues related to transparency, accountability, and fairness in the model's outcomes, especially given the potential impact on driver income and satisfaction.

- Is my data reliable?

The reliability of the data depends on various factors such as the data collection process, sources, and any potential biases or errors. It's essential to examine data quality, consistency, and potential outliers. The merged dataset appears to combine information from the original taxi dataset and the predictions dataset. While this combination can provide additional insights, we need to ensure that both datasets are accurate and well-preprocessed to ensure the reliability of the final merged data.

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

In a perfect world, it would be ideal to have comprehensive and accurately recorded data. This would include detailed information about the riders, drivers, ride circumstances, locations, and factors that might influence gratuity. Specifically, having information about rider demographics, trip satisfaction ratings, driver behavior, and external factors (such as weather, events, etc.) could contribute to a more comprehensive analysis of gratuity factors.

- What data do I have/can I get?

The current dataset includes a variety of variables such as pickup and dropoff details, trip characteristics, predicted fare amounts, and mean duration and distance. While this data provides a foundation, it may not encompass all the factors that contribute to gratuity. Additional data could potentially be gathered through surveys or customer feedback to gain insights into rider preferences and satisfaction, which could aid in understanding gratuity patterns.

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

To evaluate the success of the business objective, a suitable metric could be the "Percentage of Rides with Generous Tips" (PRGT). This metric would involve calculating the proportion of rides where the gratuity amount exceeds or equals 20% of the total fare. This metric aligns well with the objective of identifying factors that influence generous tipping behavior. It also provides a clear and actionable

insight for the TLC to improve driver satisfaction and potentially increase gratuity by focusing on the identified influencing factors.

PACE: Analyze Stage

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

Revisiting the objective of predicting generous tippers, the plan still holds. However, there might be a need to revisit the model's evaluation metric to ensure a balanced consideration of false positives and false negatives, as described earlier.

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

The data might break some assumptions of the model, especially due to its categorical and imbalanced nature. This is acceptable to some extent, as long as the model's limitations and potential biases are carefully considered and communicated.

- Why did you select the X variables you did?

The X variables were selected based on their potential relevance to predicting generous tippers. Features like pickup and dropoff locations, time of day, day of the week, and certain rates are likely to impact tipping behavior.

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

The purposes of EDA before constructing a model include understanding the distribution of variables, identifying missing values, assessing outliers, exploring relationships between variables, and gaining insights into potential patterns or trends in the data.

- What has the EDA told you?

The EDA has provided insights into the distribution of tipping behavior, showing that about one-third of customers are generous tippers. It has also revealed relationships between variables like time of day and tipping, which can be important for model interpretation.

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

As I complete this stage, I find myself using resources like domain knowledge about the taxi industry, statistical concepts for data exploration, and programming skills for data manipulation and analysis. Additionally, I might refer to documentation and online resources for specific functions or techniques,

and I'll leverage my experience to make informed decisions about data preprocessing and modeling strategies

PACE: Construct Stage

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

Looking at the results, I notice that while both the Random Forest and XGBoost models have shown improvements over the baseline, the performance metrics, particularly precision, recall, and F1 score, are still not at an optimal level. This suggests that the models may still be struggling to capture the complexities of the data. It's worth considering further feature engineering, exploring more advanced hyperparameter tuning, or even trying different algorithms to potentially address this. Additionally, the class imbalance in the target variable might be impacting the model's performance, so techniques like resampling or using different evaluation metrics could be explored.

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

The independent variables chosen for the model were a combination of those that were deemed relevant based on domain knowledge and exploratory data analysis. These included features such as day of the week, time of day, month, passenger count, and location identifiers. The rationale behind selecting these features was that they could potentially capture patterns related to customer behavior, travel patterns, and tipping habits. However, it's important to note that feature selection is an iterative process, and further experimentation could be done by adding or removing features to see how they impact the model's performance.

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

The model's performance on the validation data has been captured using various metrics like precision, recall, F1 score, and accuracy. While the scores have improved compared to the baseline, they still indicate room for enhancement. The F1 score, which balances precision and recall, provides a good overall measure of the model's effectiveness. In the case of the best XGBoost model, the F1 score is around 0.34. This indicates that the model is making a trade-off between precision and recall and may need further fine-tuning.

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

There's certainly potential for improvement. To enhance the model's performance, we could consider the following steps:

- Experiment with more sophisticated feature engineering techniques to capture additional information from the data.

- Explore alternative algorithms beyond Random Forest and XGBoost to see if they can capture the data's patterns more effectively.

- Address the class imbalance using techniques like oversampling or undersampling to ensure the model learns from both classes equally.

- Conduct more extensive hyperparameter tuning using a wider range of values to fine-tune the model's parameters.

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

During this stage, various resources were used to inform decisions and guide the modeling process. These include:

- Domain knowledge: Understanding of the taxi industry and customer behavior helped in selecting relevant features and interpreting model results.

- Exploratory data analysis (EDA): EDA provided insights into data distribution, patterns, and potential relationships among variables, guiding feature selection and preprocessing.

- Documentation and literature: References to documentation of libraries (such as scikit-learn and XGBoost) and relevant research papers on model optimization and evaluation were valuable references.

- Online communities: Platforms like Stack Overflow and forums for machine learning enthusiasts were helpful in troubleshooting and gaining insights from the experiences of others.

- Trial and experimentation: Iterative experimentation with different algorithms, hyperparameters, and feature combinations allowed for hands-on learning and model refinement.

PACE: Execute Stage

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

The key insights from the models indicate that predicting generous tips in the taxi industry is a complex task. While the models have shown improvements over the baseline, they still struggle with achieving high precision and recall simultaneously. The Random Forest and XGBoost models both identify certain patterns, such as the impact of passenger count, day of the week, and time of day on tipping behavior. However, the models also encounter challenges in capturing more subtle nuances that influence tipping decisions. The models' performance suggests a trade-off between correctly identifying generous tips and avoiding false positives.

- What are the criteria for model selection?

The criteria for model selection involve a balance between precision and recall, given the importance of both for the business case. While accuracy provides an overall sense of the model's correctness, precision and recall focus on specific aspects. The selection of the best model is based on the F1 score, which harmonizes precision and recall. This choice acknowledges the need to consider both false positives (misclassifying generous tippers) and false negatives (missing potential generous tippers) in the context of the taxi driver-customer relationship.

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

The model's performance and results make sense in the context of the complexity of the tipping behavior prediction task. While the results have improved over the baseline, achieving high precision and recall simultaneously is challenging due to the inherent uncertainty in human behavior. The final results are acceptable given the trade-offs and the balanced evaluation metrics considered. However, there's still room for improvement, especially if the business objective requires more specific precision or recall levels.

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

Yes, there is potential for improvement. Feature engineering could be refined to capture more subtle patterns, such as customer preferences, trip context, and other relevant factors. Hyperparameter tuning can be further explored to optimize the models' performance. Additionally, addressing class

imbalance using techniques like resampling or utilizing more advanced algorithms may enhance the model's ability to differentiate between the classes.

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

Yes, some features may not contribute significantly to the model's predictive power. For instance, certain location-based features or specific day-month combinations might not play a substantial role in determining generous tips. Removing such less impactful features could potentially streamline the model and lead to a more efficient and focused predictor.

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

Based on the models, recommendations for the business could include:

- Providing targeted training or incentives to drivers during peak tipping hours or on specific days.

- Promoting excellent service during weekends and evenings to cater to potentially generous tippers.

- Offering personalized experiences for repeat customers based on historical tipping behavior.

- Leveraging the model to identify potential high-tipping customers and strategically assign drivers for improved earnings.

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

Further analysis could involve:

- Exploring the impact of external factors such as weather, special events, or holidays on tipping behavior.

- Investigating whether specific neighborhoods or routes tend to have higher tipping rates.

- Predicting tipping behavior based on different payment types or customer demographics.

- Assessing the model's performance across different time periods to identify potential changes in tipping patterns.

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

During this stage, valuable resources include:

- Domain knowledge of the taxi industry and customer behavior.

- Online forums and communities for troubleshooting and best practices in machine learning.

- Documentation of libraries like scikit-learn and XGBoost for implementation guidance.

- Research papers and articles on feature engineering, model evaluation, and hyperparameter tuning.

- Is my model ethical?

The model itself appears to be ethically sound as it aims to predict tipping behavior based on available data. However, it's crucial to ensure that the model's predictions are not used in a way that discriminates against specific groups or individuals, and that any potential biases are carefully considered and addressed.

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

When the model makes a mistake, it is either misclassifying a customer's tipping behavior as generous (false positive) or missing a customer who would tip generously (false negative). In the context of the taxi business, a false positive could lead to drivers being disappointed when they don't receive expected tips, potentially affecting morale. A false negative might result in drivers missing out on opportunities for higher earnings. Balancing these errors based on the business priorities is crucial for optimizing the model's use case.