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 objective is to build and evaluate a Random Forest classification model using the NYC TLC taxi dataset to predict whether a customer is likely to be a generous tipper (giving a tip of 20% or more), based on available ride characteristics, ultimately aiming to provide the TLC with insights into factors influencing tipping behavior.

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

The external stakeholders for presenting this project are key personnel from the New York City Taxi and Limousine Commission (TLC), specifically Juliana Soto (Finance and Administration Department Head) and Titus Nelson (Operations Manager), who are noted as having non-highly technical backgrounds.

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

Key resources for the Course 6 Plan stage include: the Course 6 project instructions and stakeholder emails, the NYC taxi data dictionary, engineered features or models from Course 5, Course 6 machine learning concepts (classification, Random Forest, XGBoost, evaluation metrics, ethics), and the Course 6 PACE template.

* Do you have any ethical considerations at this stage?

Yes, there are significant ethical considerations at this planning stage for the 'generous tipper' prediction model. Key concerns include the inherent bias in the data (tip information is typically only available for credit card transactions, not cash), the risk that features used in the model (like location, time, or even predicted fare) might correlate with and act as proxies for sensitive attributes leading to potentially biased outcomes, and the potential misuse of the model if deployed; for example, using it to profile customers pre-ride could lead to discriminatory service or unfair treatment based on potentially inaccurate predictions, impacting both riders and drivers. Careful consideration of these potential harms versus the intended benefit is crucial.

* Is my data reliable?

It has both strengths and weaknesses. The data originates from a generally credible source (NYC TLC Open Data), but the specific sample used in the lab exhibited data quality issues, such as negative and unrealistically high values in fare\_amount and duration, which required cleaning and outlier handling .Critically, for predicting tipping behavior, its reliability is limited because tip information (Tip\_amount) is only available for credit card transactions, making the data unrepresentative of cash tippers. Additionally, remember the course notes state the data was sampled and altered for pedagogical purposes, so it may not perfectly reflect real-world taxi behavior.

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

In a perfect world, to best predict generous tipping (>=20%), the ideal dataset would crucially include reliable tip amount information for all payment types, especially cash transactions, to overcome the current limitation of only having credit card tips. Additionally, having cleaner base data without the need for significant outlier capping (particularly for fare\_amount and duration), along with richer contextual features such as real-time traffic conditions, weather during the trip, and perhaps even objective service quality indicators, would likely allow for a more accurate and comprehensive model of tipping behavior.

* What data do I have/can I get?

The 2017\_Yellow\_Taxi\_Trip\_Data.csv dataset, specifically filtered to include only credit card transactions (as Tip\_amount is needed). We also have access to and merged features generated in Course 5, namely mean\_distance, mean\_duration, and predicted\_fare. Furthermore, the engineered additional features directly relevant to this task, including the binary target variable generous, and time-based features like day, month, and time-of-day bins, after dropping columns irrelevant to predicting tip behavior.

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

The F1-score for the 'generous tipper' class is likely the most appropriate metric to evaluate success toward the business objective. This metric balances Precision (the accuracy of positive predictions, minimizing falsely identified generous tippers which could lead to driver disappointment) and Recall (the ability to identify most of the actual generous tippers, maximizing insight into this group), which reflects the need to both reliably identify factors associated with high tips and understand the model's ability to capture those instances, aligning with the goal of potentially improving driver satisfaction through understanding gratuities; the F1-score was also appropriately prioritized during model tuning (refit='f1') in the lab.

PACE: Analyze Stage

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

Revisiting the objective is to build a Random Forest model predicting generous tippers (>=20%) using the available data that the plan still fundamentally works as a machine learning exercise for this course project, as the necessary data (filtered for credit card tips, with engineered features) and modeling techniques are available. However, the plan doesn't need revision as much as it requires significant caveats acknowledged during planning; the ethical concerns and known data limitations (specifically, the model only learning from credit card tips and potential biases) mean that while we can proceed with building the model as planned, its real-world applicability and interpretation must be heavily qualified, and any deployment considerations would require substantial revision or rejection due to these inherent issues.

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

Random Forest models are quite flexible and don't have strict statistical assumptions like linearity or normal distributions that are common in other models, meaning our mix of numerical and encoded categorical features doesn't technically break its core requirements. While Random Forest is also relatively robust to outliers (which we know exist in the data), the most significant issue is that the data critically breaks the assumption of representativeness because it only includes credit card tips, ignoring cash tips entirely. Therefore, while it's technically ok to train the model on this data for the purpose of this specific course exercise, it's unacceptable to assume the model's findings reliably generalize to all tipping behavior, making it unsuitable for real-world deployment without addressing this fundamental data bias.

* Why did you select the X variables you did?

The features (X variables) selected for predicting generous tipping were chosen primarily based on their potential relevance to the tip amount and their availability before the tip is known. This included core trip characteristics like passenger\_count, VendorID, RatecodeID, and location IDs (PULocationID, DOLocationID), alongside engineered features capturing typical trip patterns (mean\_duration, mean\_distance, predicted\_fare from Course 5) and time-based factors ('day', 'month', and time-of-day bins like am\_rush, daytime, etc.) Columns that would cause data leakage, such as tip\_amount, tip\_percent, total\_amount, and fare\_amount, were deliberately excluded, as were original datetime columns once time-based features were extracted, and payment\_type since the data was already filtered for credit card payments.

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

The primary purposes of EDA before constructing a model are to thoroughly understand the dataset's structure, variables, and data types; identify and address critical data quality issues such as missing values, errors, or outliers; analyze individual feature distributions and relationships between variables to inform feature selection and engineering; and check basic data characteristics to help select an appropriate modeling approach, ultimately ensuring the data is clean, understood, and suitably prepared for the modeling phase.

* What has the EDA told you?

The Exploratory Data Analysis (EDA) showed that the dataset, after filtering for credit card payments and merging previous features, contained 15,265 entries with initially 27 columns and no missing values , However, crucial data quality issues were identified through .describe() and sorting, including logically impossible negative values for fare\_amount and total\_amount, zero values for trip\_distance, and extreme high outliers in these same core metrics EDA also confirmed the need to convert datetime columns established that the engineered target variable 'generous' (tip >= 20%) was relatively balanced examined distributions like VendorID counts and average tips by passenger count and provided preliminary insights into feature importance, suggesting variables like VendorID, predicted\_fare, mean\_duration, and mean\_distance were potentially strong predictors.

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

The main resources used were the Python programming language along with key data science libraries, especially pandas for loading, merging, manipulating, and encoding the dataframes (df, df1, df2), and potentially scikit-learn utilities like train\_test\_split for data preparation just before modeling. The specific datasets involved were the main 2017\_Yellow\_Taxi\_Trip\_Data.csv and the nyc\_preds\_means.csv containing features from the previous course.

PACE: Construct Stage

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

The limited hyperparameter grid searched for the Random Forest model using GridSearchCV (e.g., testing only one value for n\_estimators and max\_depth), especially compared to the wider range explored for XGBoost While this isn't necessarily a "problem" that breaks the process for this lab exercise, it means the Random Forest model might not be optimally tuned, potentially leaving performance improvements on the table. This can be easily "fixed" or improved in future iterations by expanding the cv\_params dictionary for the Random Forest GridSearchCV to include a broader range of hyperparameter values to explore.

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

The independent variables chosen for the model were selected because they represent information available before a tip is given and could plausibly influence tipping behavior. These included core trip details like VendorID, passenger\_count, RatecodeID, and location IDs, newly engineered time-based features like day, month, and time-of-day bins (am\_rush, etc.), plus relevant engineered features from the previous course such as mean\_duration, mean\_distance, and predicted\_fare.We deliberately excluded variables that would leak information about the target (like tip\_amount, total\_amount, fare\_amount, tip\_percent), those that became redundant after creating time features (original datetime columns), and payment\_type since the data was filtered only for credit card payments.

* How well does your model fit the data? What is my model’s validation score?

Based on the cross-validation performed, the models fit the data reasonably well, achieving F1-scores—the primary validation metric chosen (refit='f1')—of approximately 0.71 for the Random Forest and 0.70 for the XGBoost. These scores indicate that both models learned meaningful patterns to distinguish generous tippers better than random chance, and because these validation scores are quite close to the final test scores, it suggests the models generalize fairly well to unseen data without significant overfitting.

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

Yes, the models could likely be improved. A key change would be to perform more extensive hyperparameter tuning, especially for the Random Forest model which used a very limited search grid ,exploring a wider range of values for parameters like n\_estimators, max\_depth, and min\_samples\_split via GridSearchCV or RandomizedSearchCV might yield better results for either model. Additionally, experimenting with more sophisticated feature engineering, such as creating interaction terms between existing features (e.g., time of day and location) or trying different ways to encode categorical variables could potentially enhance performance, as could trying alternative classification algorithms beyond Random Forest and XGBoost.

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

During the Construct stage, where the main activities were building and tuning the machine learning models, the primary resources used were Python libraries within the guided Jupyter Notebook environment. Specifically, scikit-learn was essential for splitting the data (train\_test\_split), instantiating models (RandomForestClassifier), performing hyperparameter tuning (GridSearchCV), and defining evaluation metrics for the tuning process,The xgboost library was used for the XGBoost model (XGBClassifier) Pandas continued to be used for handling the feature (X\_train) and target (y\_train) data fed into the models. Implicitly, sufficient computational resources were also required to execute the GridSearchCV.fit() process for both models.

PACE: Execute Stage

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

The models, particularly the Random Forest which performed slightly better on the test set with an F1-score of about 0.72, showed a reasonable ability to distinguish between generous (>=20% tip) and non-generous tippers based on the available data, indicating they learned relevant patterns. Key insights derived from the better model (Random Forest) revealed that it tended to make more False Positive errors (predicting a generous tip which didn't happen) than False Negative errors (missing an actual generous tip) The most influential factors driving its predictions were VendorID, the predicted\_fare from the earlier regression task, mean\_duration, mean\_distance, and passenger\_count,however, a crucial overarching insight remains that all findings are limited by the dataset only containing credit card tips.

* What are the criteria for model selection?

The primary criterion for model selection, both during hyperparameter tuning within GridSearchCV (using refit='f1') and for the final choice between Random Forest and XGBoost, was the F1-score. This metric was chosen because it provides a balance between precision (correctly identifying generous tippers when predicted) and recall (finding most of the actual generous tippers), which aligns well with the project's goal and the discussion around the relative costs of different prediction errors.While other metrics like accuracy, precision, and recall were also calculated for comparison,the F1-score on the unseen test data served as the main benchmark for comparing the final performance of the tuned Random Forest and XGBoost models.

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

Yes, the model "makes sense" in the context of this project, as it learned demonstrable patterns from the provided data to predict generous tippers with better-than-chance accuracy (F1 scores around 0.70-0.72, and the features identified as important (like VendorID, predicted\_fare, mean\_duration, mean\_distance) are logically related to taxi trips. However, the model's real-world sense is limited due to the significant known bias of only using credit card tip data. Therefore, while the final results are likely "acceptable" as evidence of completing the tasks and demonstrating machine learning skills for this course assignment, they would not be acceptable for reliable real-world deployment or for drawing firm conclusions about general tipping behavior due to the modest F1 score and, more importantly, the underlying data limitations and bias.

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

Yes, the model likely could be improved, primarily because the hyperparameter tuning performed, especially for Random Forest, was quite limited.A more extensive search over a wider range of parameters (like n\_estimators, max\_depth, min\_samples\_split, etc.) using GridSearchCV or RandomizedSearchCV could potentially find a better performing configuration for either Random Forest or XGBoost. Additionally, further feature engineering, such as exploring interactions between variables or trying different encoding methods for categorical features,might uncover more predictive power, and experimenting with entirely different classification algorithms could also yield improvements, although the biggest real-world gains would likely come from addressing the underlying data bias (lack of cash tips), which is outside the scope of model changes.

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

Yes, based on the feature importance plot for the Random Forest model ,many features appeared to have very low importance compared to the top predictors like VendorID or predicted\_fare; these low-importance features likely included many of the individual day-of-week or month dummy variables, and potentially some of the time-of-day bins or less influential location IDs that didn't make the top 15 list shown. Removing these features with very low or near-zero importance could simplify the model and slightly speed up training, but because Random Forest is generally quite robust to irrelevant features, taking them out might not significantly improve (and could even slightly decrease) the overall F1-score or accuracy; it's often worth experimenting with removal if model simplicity is desired, but substantial performance gains wouldn't necessarily be expected from removing only the least important features.

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

Based on the models built, the primary recommendation is not to deploy this specific model for operational purposes like predicting individual tipping behavior or identifying generous customers in real-time, mainly due to the significant, acknowledged bias stemming from the training data only including credit card tips and the associated ethical concerns. Instead, the model serves as a methodological demonstration and potentially offers limited, internal insights for exploration; for instance, the key features identified (like predicted\_fare, mean\_duration, VendorID) could spark internal discussion at TLC or Automatidata about factors potentially correlated with higher credit card tips, warranting further, cautious investigation. The most valuable organizational recommendation, if understanding tipping is a priority, would be to explore methods for capturing more representative tip data across all payment types or to redirect modeling efforts towards business questions less impacted by this data limitation.

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

Given the insights and limitations, several other questions could be addressed: we could delve deeper into why features like VendorID and predicted\_fare were so important for predicting credit card tips; analyze the characteristics of the trips the model misclassified (especially the false positives to understand its weaknesses; explore alternative modeling approaches, such as predicting the actual tip\_percent using regression (understanding its limitation to credit card data) or classifying tips into multiple brackets instead of just binary 'generous'; or conduct a more granular analysis focusing on how tipping behavior might vary significantly across different pickup or dropoff location zones, which would require mapping the LocationIDs.

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

During the Execute stage, which focused on evaluating the models, interpreting results, and drawing conclusions, the main resources used were the Python libraries within the Jupyter Notebook, particularly scikit-learn for calculating final performance metrics on the test set (like confusion\_matrix, f1\_score, etc.) and accessing model properties like feature importances, along with matplotlib (or libraries using it, like ConfusionMatrixDisplay) for visualizing the confusion matrix and feature importances]. Additionally, pandas was used to organize and compare the final results tables.The fitted models and their outputs (predictions, scores, importances) were key inputs, guided by the Course 6 lab notebook instructions and reflective questions, and the Course 6 PACE Strategy Document was a resource for reflection.

* Is my model ethical?

Considering the significant ethical concerns raised during the planning stage,primarily the inherent bias resulting from using only credit card tip data which excludes cash tippers and makes the data unrepresentative,the model built in this lab cannot be considered ethical for real-world deployment or for drawing reliable conclusions about general tipping behavior. While constructing the model served its purpose as a valuable learning exercise within the course, using its predictions in any real-world application impacting drivers or customers would be ethically problematic due to this fundamental, unresolved data bias and the potential risks of misuse that were identified.

* 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 predicts a generous tip (>=20%) when one isn't given (a False Positive) or predicts a non-generous tip when the customer actually is generous (a False Negative). Based on the Random Forest confusion matrix [cite: Nuts 21.png], False Positives (802 instances) were more common than False Negatives (355 instances). In the context of this use case, a False Positive translates to potentially setting incorrect expectations for a driver or misidentifying factors associated with high tips based on non-tippers, while a False Negative means the model fails to identify an actual generous tipper, potentially causing missed opportunities to understand or learn from those positive instances.