

Automatidata PACE Strategy Document VI

Machine Learning Model

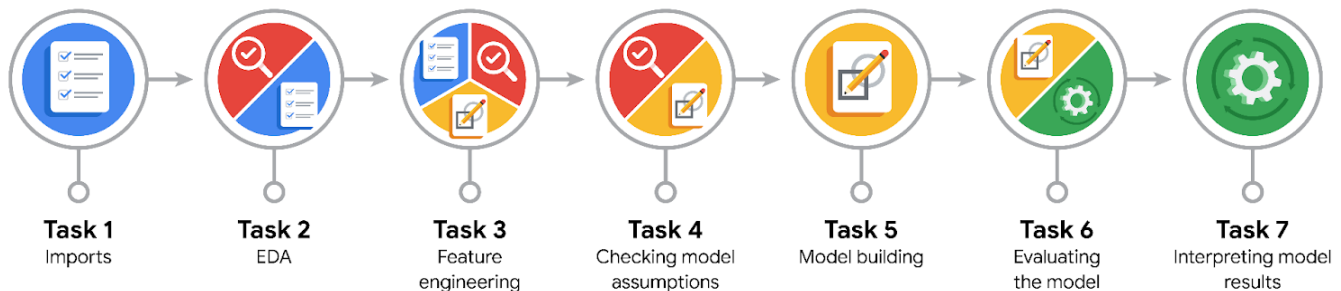
Introduction

PACE stands for Plan, Analyze, Construct and Execute. It is a framework that illustrates the foundation and structure for data analysis projects and each letter represents an actionable stage in a project. The stage “Plan” involves the definition of the project scope, the research of business data and the workflow development. The stage “Analyze” involves data scrubbing, data conversion and database formatting. The stage “Construct” involves building models and machine learning algorithms and selecting a modeling approach. The stage “Execute” involves the presentation of results to decision-makers, stakeholders and others in order to receive feedback. This framework is built upon an iterative cycle where each stage may reveal new insights, requiring the return to earlier stages. A PACE strategy document is used to record decisions and reflections at different stages of the data analytical process. It typically includes the definitions of roles and actions to ensure clarity and accountability.



Purpose

Automatidata is consulting for the New York City Taxi and Limousine Commission (TLC). New York City TLC is an agency responsible for licensing and regulating New York City's taxi cabs and for-hire vehicles. The agency has partnered with Automatidata to develop a regression model that helps estimate taxi fares before the ride, based on data that TLC has gathered. For this stage of the project, we identify seven main tasks that are presented in the following visual.



Considerations



PACE: Plan Stage

- What problem are we trying to solve or what outcome are we trying to achieve?

We aim to predict which NYC taxi riders are likely to be generous tippers ($\geq 20\%$), enhancing taxi drivers' revenue from tips.

- Who are the external stakeholders for this project?

The external stakeholders include the New York City Taxi & Limousine Commission and taxi drivers.

- What resources would be useful for this stage?

Access to more detailed driver and rider data, including historical tipping behaviors and documentation on machine learning techniques are essential.

- Are there any ethical issues to consider at this stage of the project?

Ethical concerns arise from potentially excluding customers based on predicted tipping behavior, affecting equal access to taxi services.

- Is the data accurate and reliable?

While the data seems adequate, further validation is needed to ensure accuracy, particularly regarding tipping amounts.

- What data would be perfect for answering our research question?

Comprehensive customer profiles including past tipping history, trip details and customer demographics would provide ideal insights.

- What data do we have access to or can we obtain?

We currently have trip data, estimated fares and some basic customer information but lack comprehensive tipping history.

- What metric should be used to measure the success of our business/organizational objective?

The F1 score is a suitable metric as it balances precision and recall, essential for evaluating the model's performance in predicting generous tippers.



PACE: Analyze Stage

- Based on this stage, has our research question changed? Does the plan need to be adjusted?

The research question has shifted to focus on predicting generous tippers rather than non-tippers, requiring adjustments to data collection and modeling.

- Does the data violate the assumptions of the model? Is this a significant issue?

There may be some violations regarding distribution and class balance. However, this can be addressed with appropriate preprocessing.

- Why did we choose the specific independent variables for our model?

Independent variables were selected based on their potential influence on tipping behavior, including trip characteristics and fare estimates.

- Why is Exploratory Data Analysis (EDA) important before building a model?

EDA is crucial for understanding data distributions, identifying outliers and ensuring the chosen features are relevant and informative.

- What has the data exploration revealed?

The exploration revealed that a significant portion of customers are generous tippers, indicating potential for revenue growth through targeted predictions.

- What resources would be useful for this stage?

Tools for data visualization, data cleaning and statistical analysis are needed to facilitate thorough exploration and preprocessing.



PACE: Construct Stage

- Are there any unusual or unexpected patterns in the data? Is this a problem and can it be addressed?

Patterns suggesting a vendor influence on tipping behavior were noted. This can be addressed through further investigation and model adjustments.

- Which independent variables did we choose for the model?

Key variables include Vendor ID, predicted fare, mean duration and mean distance of trips.

- How well does our model fit the data and what is its validation score?

The model's F1 score is approximately 72%, with an overall accuracy of 68%, indicating acceptable fit.

- Can we make the model better? Are there any changes we would consider?

Adding features related to trip distances and past customer behavior could enhance model performance.

- What resources would be useful for this stage?

Advanced analytics tools for feature engineering and model tuning would be beneficial.




PACE: Execute Stage

- What important discoveries have we made from our model? Can we explain our model?

The model suggests that certain variables significantly impact tipping behavior. However, random forests lack transparency which limits detailed explanations.

- What are the criteria for model selection?

Criteria include predictive performance (F1 score, accuracy), interpretability and robustness against overfitting.

- 
- Does the model make sense and are the final results acceptable?

The model's results are acceptable. However, further testing with real drivers is recommended to validate its utility.

- Do we believe the model could be enhanced? If so, how?

It could be enhanced by incorporating more features and fine-tuning model parameters.

- Were there any features that were not important at all? What would happen if they were removed?

Features with low predictive power can be removed to simplify the model, potentially improving interpretability without sacrificing performance.

- What recommendations would we provide to the organization based on the model built?

It is recommended conducting pilot testing with the model among a select group of drivers to gather feedback and iteratively improve it.

- Given our knowledge of the data and model, what other questions could we explore for the team?

Future exploration could involve examining how customer demographics influence tipping behavior and the impact of external factors like weather on tipping patterns.

- What resources would be useful for this stage?

Additional datasets for comprehensive analysis and user feedback mechanisms would be useful.

- Are there any ethical issues to consider at this stage?

Ongoing ethical considerations about data privacy and fairness in predicting tipping behaviors remain critical.

- When the model makes a mistake, what is the underlying cause? How does this impact the use case?

Mistakes may stem from inaccurate feature representations or biases in the training data, impacting trust and adoption by drivers.