

Waze PACE Strategy Document V

Regression Analysis

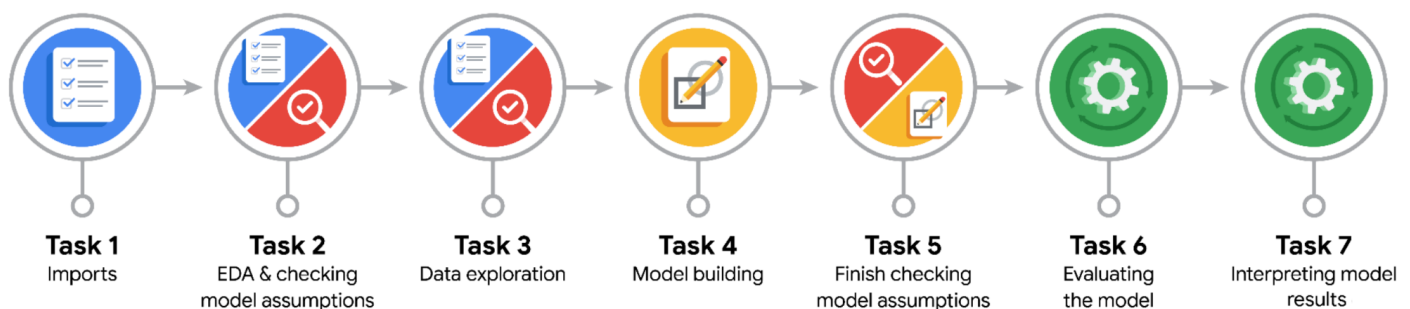
Introduction

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

Waze’s free navigation app makes it easier for drivers around the world to get to where they want to go. We want to analyze user data and develop a machine learning model that predicts user churn on the Waze app. Churn quantifies the number of users who have uninstalled the Waze app or stopped using the app. The project focuses on monthly user churn. Developing a churn prediction model will help prevent churn, improve user retention and grow Waze’s business. An accurate model can also help identify specific factors that contribute to churn and answer questions such as “Who are the users most likely to churn?”, “Why do users churn?” and “When do users churn?”. 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?

The goal is to minimize user churn to increase overall growth for the Waze app.

- Who are the external stakeholders for this project?

The external stakeholders are Waze users and the marketing team.

- What are our first impressions of the data when we start examining it?

Data contains some outliers and missing values, which could affect model building.

- What resources would be useful for this stage?

Valuable resources include data dictionaries, domain knowledge (e.g. driver types) and documentation on regression analysis.



PACE: Analyze Stage

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

It helps identify outliers, missing data and correlations among variables, ensuring a more accurate regression model.

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

Ethical considerations include ensuring user privacy when analyzing sensitive data like driving patterns and behaviors.



PACe: Construct Stage

- Are there any unusual or unexpected patterns in the data?

Some variables like distance driven per driving day were expected to be more significant but turned out to have low importance in the model.

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

Introducing new features, scaling predictors or adjusting variable selection could enhance model performance.

- What resources would be useful for this stage?

Data visualization libraries and documentation on statistical tests and performance metrics can be proven useful.



PACe: Execute Stage

- What important discoveries have we made from our model(s)?


The number of activity days emerged as the most significant predictor in the model, exhibiting a negative correlation with user churn. This was expected due to its strong relationship with the number of driving days, which was identified in the EDA as having a similar trend. Although the variable for the distance per driving day showed a strong positive correlation with churn during EDA, it had minimal impact in the final model.

- Why is understanding the beta coefficients crucial for interpreting model results?

They reveal how changes in predictors affect the likelihood of an outcome, crucial for model interpretation.

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

Adding domain-specific features and re-evaluating variable selection could improve its accuracy and recall.

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

The focus should be on further data collection to enhance model features and refine target user profiles.

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

Interesting research questions are:

- How do specific user behaviors, such as reporting road hazards, correlate with long-term retention?
- How do different types of drivers correlate with user retention?
- What type of drivers are the most and least likely to churn?

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

We should ensure transparency in how user data is collected and used for model predictions, especially to avoid unintended biases that could impact user experience or marketing decisions.