Waze PACE Strategy Document VI

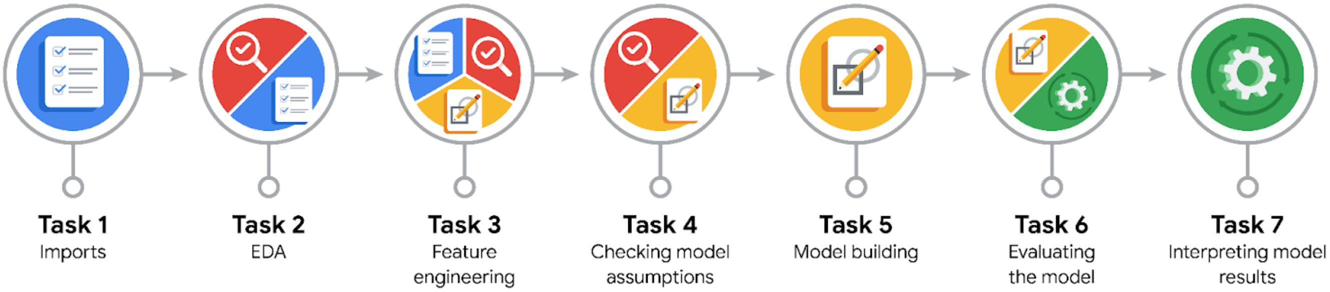
Machine Learning Model

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

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 predict user churn in the Waze app and identify factors that contribute to it, ultimately aiming to reduce churn and enhance growth.

- Who are the external stakeholders for this project?

The external stakeholders are Waze users and the marketing team.

- What resources would be useful for this stage?

Access to detailed user interaction data, data preprocessing tools and a robust computing environment for model training.

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

Ethical concerns are primarily related to handling false negatives (missed opportunities for retention) and false positives (unnecessary retention efforts).

- Is the data accurate and reliable?

It is reliable but not perfectly suitable for consistent churn prediction due to a lack of certain detailed interaction metrics.

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

A perfect dataset would include detailed drive-level metrics (e.g. driving times, locations) and interaction data like the frequency of reporting or confirming road hazards.

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

Current data includes user sessions, activity patterns and limited metrics such as average drive length and session frequency.

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

Recall should be used, to ensure that a high percentage of actual churn cases are identified and addressed.

PACE: Analyze Stage

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

The question remains the same, but the plan may need to include additional data collection for improved predictions.

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

The data is imbalanced, which may affect models like logistic regression but is manageable for tree-based models.

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

Variables were selected based on their correlation with churn and their potential to improve predictive power, like average speed and percentage of sessions in the last month.

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

It helps identify patterns, correlations and outliers, guiding feature selection and preparation for modeling.

- What has the data exploration revealed?

Engineered features such as average speed and session frequency are strong predictors, but overall data lacks depth for more precise predictions.

- What resources would be useful for this stage?

Data visualization tools, feature engineering guides and domain-specific knowledge for understanding user behavior are useful.

PACE: Construct Stage

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

The imbalance in churn data is a challenge but can be addressed through model selection and feature engineering.

- Which independent variables did we choose for the model?

Variables like average speed, total sessions per day and percentage of sessions to favorite destinations.

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

The XGBoost model had a recall score of around 18%, showing improvement over the logistic regression model.

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

Adding more detailed interaction data and refining feature selection could further enhance model accuracy.

- What resources would be useful for this stage?

More detailed data, machine learning libraries like XGBoost and domain knowledge for crafting better features.

PACE: Execute Stage

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

The XGBoost model provided better recall and utilized a wider range of features, indicating the importance of feature engineering.

- What are the criteria for model selection?

A key criterion is the emphasis on recall due to the need to accurately identify as many churn cases as possible, with a preference for models that handle imbalanced data.

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

  The results are within an acceptable range but could be improved with more detailed data.

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

  The model could be enhanced by including more granular user behavior data and exploring additional feature engineering.

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

  Features with minimal impact could be removed to simplify the model without significantly affecting performance.

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

  More detailed user interaction data should be collected and further analysis should be conducted with enhanced data to better understand churn dynamics.

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

  Interesting research questions are:

  - How do driving behaviors correlate with long-term user retention?
  - What impact do specific app features have on user engagement?

- What resources would be useful for this stage?

  Enhanced computational resources for model testing, data augmentation tools and data storage solutions would be useful.

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

  Ethical concerns around handling predictions and user engagement strategies, especially when attempting to retain potentially disengaged users.

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

  Errors often stem from limited data depth, leading to inaccurate predictions. This impacts the ability to effectively target retention efforts.