Waze PACE Strategy Document II

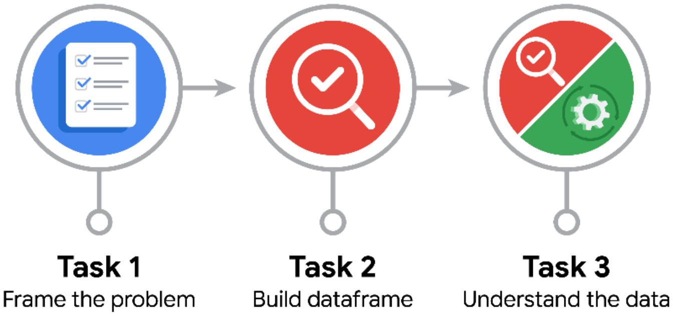
Preliminary Data Summary

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 three main tasks that are presented in the following visual.



Considerations

PACE: Plan Stage

- How can we best prepare to understand and organize the provided data?

The first step is to explore the dataset and consider reviewing the Data Dictionary. By analyzing the taxi cab data fields, we can better understand the significance of each variable. However, the main objective is to load the data into Python, examine it closely and share initial observations with the Data Analysis Manager. The next step would be to deepen our understanding and check for any irregularities in the data.

- What follow-along and self-review materials can assist in completing this task?

Helpful resources include the Data Dictionary, which explains each variable in the dataset, and the fact sheet, which provides background information on the data source. Additionally, reviewing documentation on descriptive statistics will support a better understanding of how to summarize and interpret the dataset.

- What steps might a proactive learner take before starting to code?

A proactive learner may begin by thoroughly reviewing the data and understanding the structure and meaning of each variable. They could also explore similar datasets or past analyses for guidance. Before coding, it is beneficial to identify key variables, consider potential relationships between them, and take note of any patterns or anomalies that may need special attention.

PACE: Analyze Stage

- Is the information provided enough to reach the goal based on our intuition and variable analysis?

The information presented offers valuable insights into user behavior, particularly in understanding differences between churned and retained users. However, to fully reach the goal of identifying key relationships between variables that influence churn, a deeper exploration of variables (such as time of day, geographical location or app version) might be required. This would ensure all relevant factors are captured, especially those that might explain churn among high-frequency users.

- How would we create a summary dataframe and determine the minimum and maximum values within the dataset?

To build a summary dataframe, the basic descriptive statistics functions in Python (e.g. *describe()* in *pandas*) would be used. This would provide essential metrics like mean, median and standard deviation, as well as the minimum and maximum values for each variable. Filtering the data to check for specific ranges and inspecting the distribution of values can help in identifying anomalies.

- Do any average values appear out of the ordinary? How would we describe the interval data?

Some average values stand out, particularly the high number of kilometers driven per day by churned users, which is over 240% higher than that of retained users. This figure seems to deviate from the norm and suggests that these "super-drivers" may follow a unique usage pattern. In terms of interval data (like the number of drives or distance traveled), the large variation between churned and retained users, such as kilometers driven and time spent driving, indicates that the churned group has significantly different usage patterns, potentially signaling that their driving behaviors might lead to dissatisfaction with the app.

PACE: Construct Stage

Note: This stage is not relevant to the current workflow.

PACE: Execute Stage

- Based on your current understanding of the data, what initial recommendations would we provide our manager for further investigation before diving into exploratory data analysis (EDA)?

It is essential to understand why churned users drive significantly longer distances and more frequently than retained users, yet still abandon the app. Gathering more context-specific data on their driving needs and app interactions would help pinpoint areas where Waze is not meeting their expectations. We could also consider adding variables related to app performance or in-app features used during trips. These might help clarify why certain users, especially high-frequency drivers, churn despite their frequent usage.

- Which data points seem to contain irregularities?

The high frequency of drives and longer distances covered by churned users seem unusual. Specifically, churned users driving 698 kilometers per day on average is significantly higher than retained users, suggesting that a subset of users, possibly long-haul drivers, behave differently. The imbalance in the churned user profile (longer, fewer drives) compared to retained users raises questions about whether these needs are being met by Waze.

- What additional types of data could enhance this dataset?

Incorporating feedback from churned users would provide direct insights into the reasons for their departure. This could be in the form of app store reviews, exit surveys or support ticket data. Adding geolocation data could help analyze whether driving routes or regions contribute to churn (e.g. urban vs. rural drivers). Data on which features were used in the app during drives (e.g. navigation, accident reporting) could also shed light on whether specific features (or lack thereof) correlate with churn.