TikTok PACE Strategy Document II

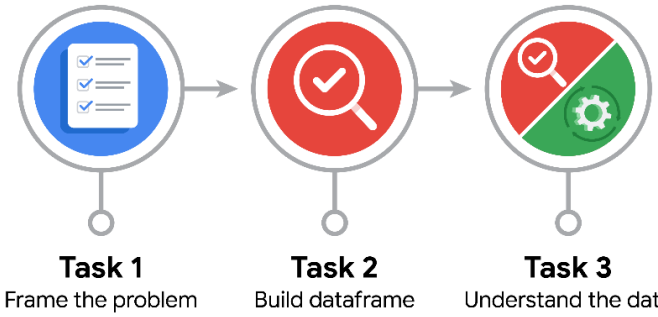
Preliminary Data Summary

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

TikTok users have the ability to report videos and comments that contain user claims. These reports identify content that needs to be reviewed by moderators. This process generates a large number of user reports that are difficult to address quickly. TikTok is working on the development of a predictive model that can determine whether a video contains a claim or offers an opinion. With a successful prediction model, TikTok can reduce the backlog of user reports and prioritize them more efficiently. 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 video 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 available data seems sufficient to build a predictive model for user-submitted claims, especially with key variables such as that claim status, the number of views the video received and the duration of the video. However, further analysis is necessary to ensure the accuracy of the data and to identify any outliers or anomalies that could skew the results.

- 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?

Based on the engagement data, the mean and median view counts for claim and opinion videos suggest that claims are heavily flavored in terms of attention. This could indicate a deeper behavioral pattern of user interaction or possible bias towards certain types of content. We also observe that some of the other count variables have outliers at the upper end of their distribution. Their standard deviations and maximum values are very large, especially compared to their quartile values.

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)?

Since claim videos have significantly higher views and engagement, it would be useful to explore why this is happening. It would be advisable to explore further the difference in engagement between banned and active authors, especially considering that banned authors seem to perform better on certain metrics. A closer look at the distribution of variables is also recommended to ensure the data's integrity before further analysis.

- Which data points seem to contain irregularities?

  The significant discrepancy between claim and opinion video view counts stands out. These could suggest that either some claim videos are extreme outliers (e.g. viral content) or that claim videos are inherently more engaging than opinion videos. The higher engagement rates for claim videos by banned authors compared to active authors may also reflect unusual user behavior.

- What additional types of data could enhance this dataset?

  Additional data such as user demographics, sentiment analysis scores for video comments or topic could provide more context for understanding the claim status.