PACE Strategy Document

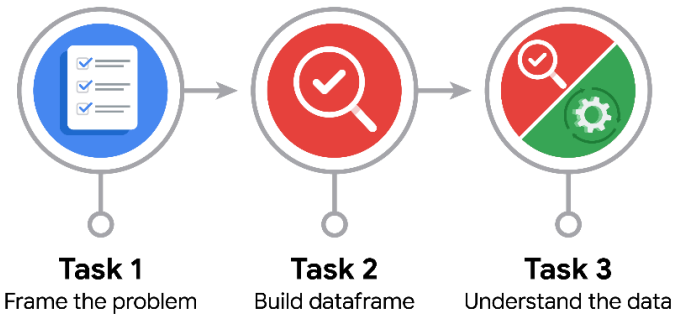
Phase 2: Exploratory Data Analysis & Engagement Insights for TikTok Claims Classification

Instructions

I will use this **PACE strategy document** to record my decisions and reflections as I work through this Phase-2 project. This document will serve as a guide, helping me consider my responses and reflections at different stages of the data analytical process.

Reference Guide

This project has three tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Plan Stage

- How can you best prepare to understand and organize the provided information?

To prepare for understanding the data, I would begin by reviewing the data fields to grasp the structure and context. I would also examine descriptive statistics for the columns I deem most relevant. This would help me identify key patterns, distributions, and potential issues with the data.

For organizing the information, I would import the data into the coding environment (Jupyter Notebook) and utilize the pandas library to create a DataFrame. From there, I would assess the organization of the data and determine any necessary preprocessing steps to ensure it is structured properly for further exploratory data analysis (EDA). This might involve handling missing values, identifying outliers, and confirming column data types to ensure the dataset is ready for analysis.

- What follow-along and self-review codebooks will help you perform this work?

For this project, the Kaggle lab work will be a valuable resource, providing step-by-step guidance throughout the process. Additionally, I will primarily use Jupyter Notebook for coding and analysis. It will serve as my main platform for exploring the data, writing the code, and documenting the steps, ensuring an organized and structured approach to the project.

- What are some additional activities a resourceful learner would perform before starting to code?

A resourceful learner would first ensure they are clear on the attributes and methods of the libraries and packages commonly used in the project, such as Pandas, NumPy, and Matplotlib. They would also take the time to explore and understand the raw data, identifying key attributes, potential issues, and the structure of the dataset. Conceptually, they would think about the problem at hand, formulate a strategy for tackling individual tasks, and plan the coding workflow to ensure an efficient and organized approach.

PACE: Analyze Stage

- Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes, based on my analysis, the dataset provides a well-rounded representation of the video content, with 12 columns capturing various features and associated metadata. These fields offer sufficient detail to achieve the goal, as they cover key aspects necessary for analysis and classification. Therefore, I believe the available information is adequate for the task at hand.

- How would you build summary dataframe statistics and assess the min and max range of the data?

To build summary statistics, I would use the `describe()` function from the pandas library on the dataframe (e.g., `data`), which provides key metrics such as mean, standard deviation, min, and max values for each numerical column. For more specific analysis of the min and max range, I could also use the `min()` and `max()` functions on individual columns to extract their respective minimum and maximum values. This would allow for a more granular assessment of the data's range.

- Do the averages of any of the data variables look unusual? Can you describe the interval data?

The mean of `video_duration_sec` appears normal, but the means of `video_view_count`, `video_like_count`, `video_share_count`, `video_download_count`, and `video_comment_count` are significantly higher than their respective medians, indicating right skewness in the data. Additionally, the difference between the 75th percentile and maximum values in these columns further suggests the presence of outliers or exceptionally popular videos. The interval data in this dataset is `video_duration_sec`, which represents the duration of the video in seconds.

PACE: Construct Stage

Note: The Construct stage does not apply to this workflow. The PACE framework can be adapted to fit the specific requirements of any project.

PACE: Execute Stage

- Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing exploratory data analysis?

I would recommend addressing the null values, as they are consistently present across most fields for the same data points. Additionally, it would be important to understand how popular videos might influence the means of numerical fields such as `video_view_count`, `video_like_count`, `video_share_count`, etc., since they may skew the data. I would also suggest focusing on outliers, as they could have a significant impact on the analysis. Lastly, it would be valuable to further investigate the correlation between `claim_status` and other engagement-related fields (e.g., views, likes, shares) to uncover any potential patterns or relationships.

- What data initially presents as containing anomalies?

The engagement fields, such as views, likes, shares, and comments, contain data points that fall well above the 75th percentile, indicating that certain videos have significantly higher counts in these areas. While not necessarily anomalies, these data points represent popular videos that can skew the overall results. These high values should be considered as outliers and handled appropriately during the analysis to avoid distorting the findings.

- What additional types of data could strengthen this dataset?

Adding data such as the author's number of followers, the number of videos from their account currently under review, and the number of reported videos would strengthen the dataset. These factors could provide more context around the author's influence, content behavior, and potential for policy violations, which would be valuable for further analysis and prediction models.