

# TikTok PACE Strategy Document V

## Regression Analysis

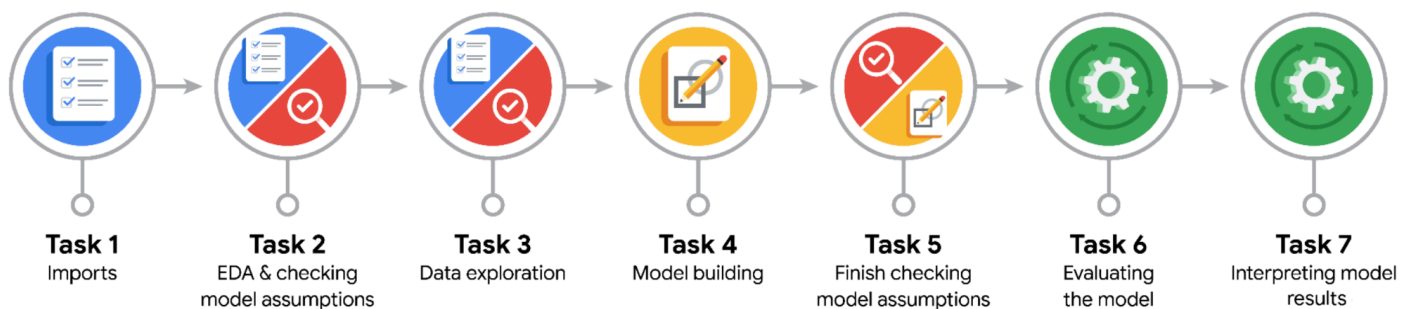
### Introduction

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

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 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 classify user submissions and understand the behavior of verified users to improve the accuracy of claim identification.

- Who are the external stakeholders for this project?

The external stakeholders include TikTok's content moderation team and possibly departments involved in managing user interactions and verified account status.

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

Initial impressions suggest some variables like video length may be key factors, and there are strong correlations among certain metrics.

- What resources would be useful for this stage?

Valuable resources include data dictionaries 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?

Potential concerns include ensuring unbiased data collection and avoiding model decisions that could unfairly favor or penalize certain user types.



### **PACE: Construct Stage**

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

The strong correlation between the number of views and the number of likes for a video was an unexpected pattern that had to be addressed to avoid multicollinearity.

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

Further improvements could include trying different model types like decision trees or random forests to see if they better capture relationships in the data.

- 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 coefficients indicate that an opinion video increases the log-odds of the user being verified by 1.70 and every additional second of video length decreases the log-odds of the user being verified by 0.002. The model showed moderate predictive performance, with a precision of 63% and a recall of 82%, though its overall accuracy of 67% was slightly lower than ideal. Overall, the model, which predicted the verified status based on video features, demonstrated adequate performance. The analysis highlighted that opinion videos are more likely to be posted by verified users and that banned users are less likely to be verified, whereas other features had less significant associations with verified status.

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

The beta coefficients indicate how each feature affects the log-odds of an outcome, providing insights into which variables most strongly influence the prediction.

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

The model could be enhanced by experimenting with other classification models or using feature engineering to refine variable selection.

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

The focus should be on promoting shorter content to verified users as this may encourage them to share opinions, aligning with TikTok's goals.

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

Interesting research questions are:

- How do fine-grained factors of user engagement (e.g. video completion) correlate with claim submission?
- How does the time of posting correlate with claim submission?

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

Transparency in how the model classifies user behavior is important, especially to avoid unintended biases that could impact user experience or moderation decisions.