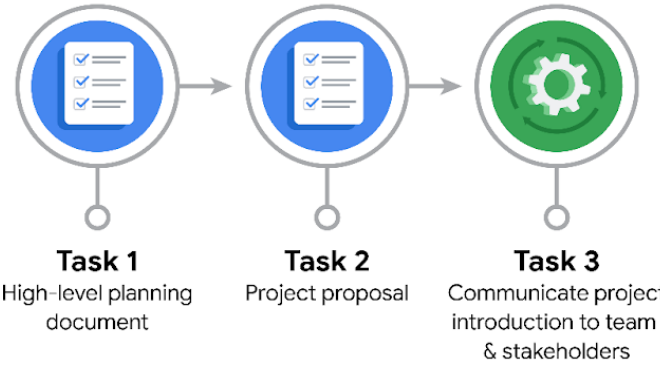
Course One
Foundations of Data Science

Reference Guide

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



Data Project Questions & Considerations

PACE: Plan Stage

- Who is your audience for this project?

TikTok Data Team (Technical) TikTok Cross-Functional Team (Non-Technical)

- What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger needs of the client?

The primary goal of the project is to develop a predictive machine learning model that can accurately classify user comments on the TikTok platform as either claims or opinions. This classification serves several key objectives:

1. **Enhanced Content Moderation**: By automating the classification of comments, TikTok aims to improve content moderation processes. This enables the platform to identify and address problematic or rule-violating content more efficiently.
2. **Report Prioritization**: The model's predictions will help prioritize user-reported content. Comments that are identified as potential claims, which often require urgent attention, can be escalated for faster review.
3. **User Experience**: By reducing the presence of misleading or harmful claims in the comment sections, TikTok seeks to enhance the overall user experience. This promotes a more positive and safe environment for its user community.
4. **Efficiency and Scalability:** Automating comment classification reduces the manual workload of content moderators. It allows TikTok to scale its content moderation efforts more effectively as the platform continues to grow.
5. **Data-Driven Insights:** The project involves exploratory data analysis (EDA) and hypothesis testing, which can yield valuable insights into user behavior and comment patterns on TikTok. These insights can inform future platform improvements and content policies.

Anticipated Impact on the Client:

1. **Improved User Safety:** TikTok's commitment to user safety and well-being is reinforced by more effective content moderation. The platform becomes a safer space for users of all ages.
2. **Enhanced User Engagement:** A positive and safe commenting environment fosters increased user engagement and interaction on the platform. Users are more likely to participate in discussions and share their thoughts.
3. **Reduced Operational Costs:** Automation of comment classification reduces the manual effort required for content moderation, leading to potential cost savings for TikTok.
4. **Data-Driven Decision Making:** The project's data analysis components provide TikTok with data-driven insights that can guide future content policies and platform improvements. This aligns with the company's mission to move at the speed of culture and adapt to user needs.
5. **Scalability:** As TikTok continues to grow, the project's automation capabilities ensure that content moderation can scale to meet the demands of a larger and more diverse user base.

- What questions need to be asked or answered?

Project Definition and Objectives:

1. What is the primary objective of this project?
2. How will the success of this project be measured?
3. What specific challenges in content moderation is TikTok trying to address through comment classification?

Data Acquisition and Preprocessing:

1. What is the source of the comment data for training the classification model?
2. What preprocessing steps are necessary to prepare the data for analysis and modeling?
3. How will missing data be handled?

Model Development:

1. Which machine learning algorithms and techniques will be used for comment classification?
2. How will the model be trained and validated?
3. What features or attributes of comments will be used for classification?

Testing and Validation:

1. What is the testing strategy to ensure the reliability of the classification model?
2. How will model predictions be validated against user-reported content?
3. How will model performance be evaluated, and what metrics will be used?

Project Timeline and Milestones:

1. What is the estimated timeline for completing the project?
2. What are the key project milestones and their deadlines?
3. How will project progress be tracked and reported to stakeholders?

Stakeholder Involvement:

1. Who are the key stakeholders involved in this project?
2. What are the specific roles and responsibilities of each stakeholder?
3. How will communication and collaboration among stakeholders be managed?

Impact and Benefits:

1. What impact is the project expected to have on content moderation efficiency?
2. How will user experience on TikTok be improved as a result of this project?
3. Are there potential cost savings or operational benefits?

Technical Infrastructure:

1. What technical infrastructure is required for data storage, model development, and deployment?
2. Are there any specific tools or platforms that will be used in the project?

Ethical and Privacy Considerations:

1. What ethical considerations should be taken into account when classifying user comments?
2. How will user privacy and data protection be ensured during the project?

Documentation and Reporting:

1. How will project documentation be organized and maintained?
2. What kind of reports and presentations will be created for different stakeholders?
3. Is there a plan for documenting code, model architecture, and findings?

Future Scalability:

1. How will the project adapt as TikTok's user base continues to grow?
2. Are there plans for ongoing model monitoring and updates?

Project Risks and Mitigations:

1. What are the potential risks and challenges associated with the project?
2. What strategies and contingencies are in place to mitigate these risks?

Feedback and Iteration:

1. How will feedback from users and moderators be collected and used to improve the classification model?
2. Is there a process for iterating on the model and making continuous improvements?

- What resources are required to complete this project?

Data:

1. High-quality and labeled comment data for training and testing the classification model.
2. Access to relevant metadata or context about comments, if available.

Technical Infrastructure:

1. Computing resources for data preprocessing, model development, and testing.
2. Data storage infrastructure to securely store and manage the comment dataset.
3. Servers or cloud-based resources for model deployment and real-time classification.

Software and Tools:

1. Data analysis and machine learning tools, such as Python libraries (e.g., pandas, scikit-learn, TensorFlow, PyTorch).
2. Development and coding tools (IDEs, version control systems, etc.).
3. Data visualization tools for exploratory data analysis (e.g., Matplotlib, Seaborn).
4. Database management systems, if needed.

Human Resources:

1. Data scientists and machine learning experts to develop and train the classification model.
2. Data engineers for data acquisition, preprocessing, and database management.
3. Project managers to oversee project planning, coordination, and reporting.
4. Content moderators to validate model predictions and provide feedback.
5. Stakeholders for project oversight and decision-making.

Documentation and Reporting:

1. Documentation tools and templates for project planning, data analysis, model development, and project reports.
2. Presentation software for creating visual materials and reports for stakeholders.

Ethical and Legal Expertise:

1. Ethical and legal experts to ensure compliance with data privacy regulations and ethical guidelines when handling user-generated content.

Budget:

1. Financial resources to cover project costs, including personnel, infrastructure, and software licensing fees.

Feedback Mechanisms:

1. Mechanisms for collecting feedback from users, moderators, and stakeholders to iteratively improve the model and content moderation processes.

Training and Education:

1. Training programs for content moderators and team members on how to use and interpret the classification model.

Quality Assurance:

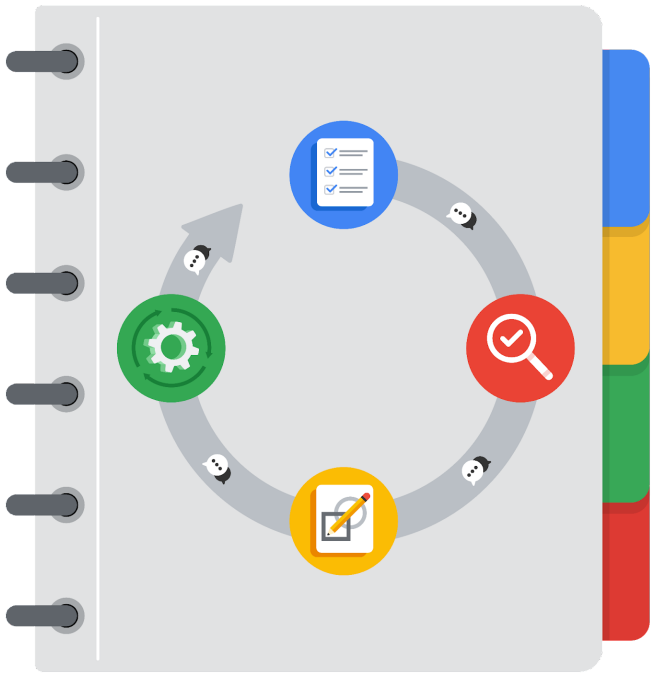
1. Quality assurance processes to ensure the accuracy and reliability of the classification model.

- What are the deliverables that will need to be created over the course of this project?

1. **Project Proposal**: A comprehensive project proposal outlining the project's scope, objectives, milestones, and stakeholders. This document serves as a roadmap for the project.
2. **Data Acquisition Plan:** A plan detailing how and where to acquire the comment data for training and testing the classification model. This plan may include data sources, APIs, and data collection methods.
3. **Data Preprocessing Documentation:** Documentation of the data preprocessing steps, including data cleaning, handling missing values, and feature engineering. This ensures transparency and reproducibility.
4. **Exploratory Data Analysis (EDA) Report**: A report summarizing the findings from EDA, including data distribution, patterns, and insights. Visualizations and statistics are typically included.
5. **Classification Model:** The machine learning model for comment classification, including code, model architecture, and training procedures. This model should be well-documented and reproducible.
6. **Model Evaluation Report:** A report detailing the evaluation of the classification model's performance, including metrics such as accuracy, precision, recall, F1-score, and ROC curves.
7. **Model Integration Documentation**: Documentation explaining how the classification model is integrated into the TikTok platform for real-time use, including APIs or deployment methods.
8. **Testing Framework:** Documentation of the testing framework used to validate the model's predictions and ensure its reliability. Test cases and results should be included.
9. **Project Workflow and Documentation**: Documentation of the project's workflow, including data collection processes, data analysis steps, model development procedures, and version control practices.
10. **Project Progress Reports:** Regular progress reports summarizing project milestones, achievements, challenges, and upcoming tasks. These reports are essential for communication with stakeholders.

11. **Presentation Materials:** Materials for presenting project progress and findings to the TikTok leadership team, including visuals, slides, and talking points.
12. **Hypothesis Testing Documentation:** Documentation of the hypotheses tested, statistical tests performed, and the interpretation of results. This helps in making data-driven recommendations.
13. **Documentation of Assumptions:** Documentation of any assumptions made during the project, especially in the context of regression modeling and data analysis.
14. **Documentation of Ethical Considerations**: A report detailing ethical considerations related to content moderation, data privacy, and user consent, along with steps taken to address them.
15. **Final Project Proposal:** An updated project proposal that reflects the progress and outcomes of the project, including any adjustments to the initial scope and objectives.
16. **User Feedback and Improvement Reports**: Reports summarizing user feedback on the effectiveness of the classification model and content moderation processes. This information is used for iterative improvements.

THE PACE WORKFLOW



[Alt-text: The PACE Workflow with the four stages in a circle: plan, analyze, construct, and execute.]

You have been asked to demonstrate for the company's data team how you would use the PACE workflow to organize and classify tasks for the upcoming project. Select a PACE stage from the dropdown buttons. A few tasks involve more than one stage of the PACE workflow. Additionally, not every workplace scenario will require every task. Refer back to the Course 1 end-of-course portfolio project overview reading if you need more information about the tasks within the project.

Project tasks

Following are a group of tasks your company's data team has determined need to be completed within this project. The data analysis manager has asked you to organize these tasks in preparation for the project proposal document. First, identify which stage of the PACE workflow each task would best fit under using the drop down menu. Next, give an explanation of why you selected the stage for each task. Review the following readings to help guide your selections and explanation: The PACE stages and Communicate objectives with a project proposal. You will later reorder these tasks within a project proposal.

1. **Evaluating the model:** Execute

   Why did you select this stage for this task?

After the model has been constructed, data is run through to evaluate whether it meets the project's expectations and goals.

2. **Conduct hypothesis testing:** Analyze **and** Construct

   Why did you select these stages for this task?

During the analyzing stage, it is determined that a statistical test will be used. During the construction phase, the test is carried out.

3. **Begin exploring the data:** Analyze

   Why did you select this stage for this task?

During the analysis phase, you will gain a deeper understanding of the dataset and the information within it.

4. **Data exploration and cleaning:** Plan **and** Analyze

   Why did you select these stages for this task?

Planning takes place when you first make choices about the methods needed. The cleaning process then takes place in the analyzing stage.

5. **Establish structure for project workflow (PACE):** Plan

Why did you select this stage for this task?

Planning stage. Creating an initial project PACE document outlines the workflow and helps to plan how to best approach a project.

6. **Communicate final insights with stakeholders:** Execute

Why did you select this stage for this task?

Communication is necessary at various points throughout a project. Final insights are shared with stakeholders in the execute phase of the data project workflow.

7. **Compute descriptive statistics:** Analyze

Why did you select this stage for this task?

Investigating the statistics within data takes place during analysis.

8. **Visualization building:** Analyze **and** Construct

Why did you select these stages for this task?

Visualization begins with data assessment and is created during the construction stage.

9. **Write a project proposal:** Plan

Why did you select this stage for this task?

Planning stage. A project proposal is the initial document used to define a project.

10. **Build a regression model:** Analyze **and** Construct

Why did you select this stage for this task?

During the analyzing stage, the model is examined in detail to be sure it will meet the needs of the task. The building of the regression model will take place in the construction phase.

11. **Compile summary information about the data:** Analyze

Why did you select this stage for this task?

Inspecting a dataset to compile information would take place in the analysis phase.

12. **Build machine learning model:** Construct

Why did you select this stage for this task?

The building of a data model would take place in the construct stage.