

Course One

Foundations of Data Science



Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. You can use this document as a guide to consider your responses and reflections at different stages of the data analytical process. Additionally, the PACE strategy documents can be used as a resource when working on future projects.

Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

- ☐ Complete the PACE Strategy Document to plan your project while considering your audience members, teammates, key milestones, and overall project goal.
- ☐ Create a project proposal for the data team.

Relevant Interview Questions

Completing this end-of-course project will empower you to respond to the following interview topics:

- As a new member of a data analytics team, what steps could you take to get 'up to speed' with a current project? What steps would you take? Who would you like to meet with?
- How would you plan an analytics project?
- What steps would you take to translate a business question to an analytical solution?
- Why is actively managing data an important part of a data analytics team's responsibilities?
- What are some considerations you might need to be mindful of when reporting results?



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?

The audience for this project includes internal team members (data analytics team), external stakeholders (finance and operations), Waze leadership, and yourself as the data analyst.

- 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 main goal of this project is to build a machine learning model that predicts user churn for the Waze app. By accurately predicting churn, Waze aims to prevent user attrition, improve user retention, and ultimately increase growth. The impact of this work on the larger needs of the client includes enhancing user experience, optimizing retention strategies, making data-driven decisions for product development, and ultimately contributing to the overall success and growth of the Waze business.



- What questions need to be asked or answered?

Several questions need to be asked and answered to effectively address the objectives of the project:

1. What are the key factors contributing to user churn in the Waze app?
2. How can user churn be accurately measured and quantified?
3. What data sources are available for analyzing user behavior and predicting churn?
4. What machine learning algorithms are best suited for predicting user churn based on available data?
5. How can the accuracy and reliability of the churn prediction model be evaluated?
6. What actions can be taken to mitigate churn among at-risk users identified by the model?
7. How will the model's performance be monitored and maintained over time?
8. How will the insights generated from churn prediction be translated into actionable strategies for improving user retention?
9. What are the potential limitations or biases associated with the data and model, and how can they be addressed?
10. How will the project outcomes be communicated and utilized by stakeholders within the organization?



- What resources are required to complete this project?

To complete this project effectively, the following resources are required:

1. **Data**: Access to a comprehensive dataset containing user behavior and demographic information, as well as churn indicators.
2. **Computing Resources**: Sufficient computing power and storage capacity to handle large datasets and run machine learning algorithms.
3. **Software Tools**: Use of data analysis and machine learning tools such as Python (with libraries like Pandas, NumPy, Scikit-learn), R, or other relevant software packages.
4. **Domain Expertise**: Collaboration with subject matter experts to understand the nuances of user behavior and churn dynamics within the context of the Waze app.
5. **Project Management Tools**: Utilization of project management software or tools to organize tasks, track progress, and coordinate team efforts effectively.
6. **Communication Channels**: Establishment of clear communication channels for regular updates, feedback, and collaboration among team members and stakeholders.
7. **Training and Development**: Provision of training and development opportunities for team members to enhance their skills in data analysis, machine learning, and project management as needed.
8. **Documentation**: Creation of comprehensive documentation for data sources, methodologies, model development, and project outcomes to ensure transparency, reproducibility, and knowledge sharing.
9. **Stakeholder Engagement**: Engagement with relevant stakeholders, including Waze executives, product managers, and user experience specialists, to align project objectives with business goals and user needs.
10. **Quality Assurance**: Implementation of quality assurance processes to validate data integrity, model accuracy, and adherence to best practices throughout the project lifecycle.

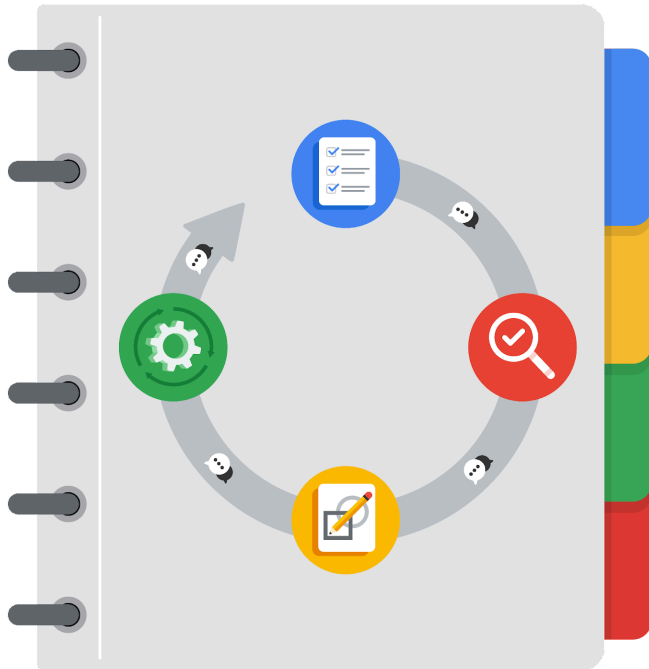
- What are the deliverables that will need to be created over the course of this project?

Throughout the course of this project, the following deliverables will need to be created:

1. ****Project Proposal****: A detailed document outlining the project's objectives, scope, methodology, timeline, and deliverables. This proposal will serve as a roadmap for the project team and stakeholders.
2. ****Data Exploration and Analysis Report****: A comprehensive report documenting the findings from exploratory data analysis (EDA) conducted on the Waze user dataset. This report will include insights into user behavior, trends, patterns, and potential predictors of churn.
3. ****Machine Learning Model****: Development and implementation of a machine learning model capable of predicting user churn based on historical data. The model should be trained, validated, and tested using appropriate techniques and evaluation metrics.
4. ****Model Evaluation and Performance Report****: An assessment of the performance and effectiveness of the machine learning model in predicting user churn. This report will include metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
5. ****Visualization and Presentation Materials****: Creation of visualizations, charts, and dashboards to communicate key findings, insights, and model predictions to stakeholders. Presentation materials should be clear, informative, and visually appealing.
6. ****Final Project Report****: A comprehensive report summarizing the entire project, including methodology, findings, recommendations, limitations, and future directions. This report will provide a holistic overview of the project's objectives, processes, and outcomes.
7. ****Executive Summary****: A concise summary highlighting the key insights, conclusions, and recommendations from the project. This summary will be tailored for executive-level stakeholders who may not have technical expertise.
8. ****Documentation and Code Repository****: Documentation of data sources, preprocessing steps, feature engineering techniques, model development, and implementation code. A well-organized code repository will facilitate collaboration, reproducibility, and future iterations of the project.



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: **Analyze** ▾



Why did you select this stage for this task?

I selected the "Analyze" stage for the task "Evaluating the model" because this stage involves the critical examination of data and models to derive insights and make informed decisions. In the context of the project, evaluating the model entails analyzing its performance, assessing its accuracy, and identifying any areas of improvement. This process involves comparing the model's predictions against actual outcomes, calculating relevant metrics such as accuracy, precision, recall, and F1-score, and conducting various tests to ensure the model's reliability and effectiveness. By placing "Evaluating the model" under the "Analyze" stage, we emphasize the importance of thorough analysis and validation of the model's performance before its deployment or implementation. This ensures that the model meets the project requirements and delivers reliable results, contributing to the overall success of the project.

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

Why did you select these stages for this task?

Plan: During the planning stage, formulating hypotheses and designing the testing methodology are essential steps. This involves defining the research question, specifying the null and alternative hypotheses, selecting the appropriate statistical tests, and planning the data collection process. Planning is crucial to ensure that the hypothesis testing is conducted systematically and aligns with the project goals.

Analyze: The analysis stage involves the actual execution of the planned testing. This includes collecting and processing the data, applying statistical tests, and interpreting the results. In the context of hypothesis testing, the analysis stage focuses on examining whether the observed data provides enough evidence to reject or fail to reject the null hypothesis. This stage contributes to making data-driven decisions based on the outcomes of the hypothesis tests.

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



Why did you select this stage for this task?

I selected the "Analyze" stage for the task "Begin exploring the data" because in this stage of the PACE framework, the primary objective is to understand the data, identify patterns and trends, and explore relationships between variables. During the initial data exploration, we are not yet building models or performing more advanced analyses; we are simply seeking to grasp the nature of the available data. Therefore, the initial data exploration fits best within the analysis phase, where we can prepare the data for deeper analyses in the subsequent stages of the project.

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

Why did you select these stages for this task?

I selected the "Plan" stage because data exploration and cleaning involve initial planning and preparation, such as defining objectives, identifying data sources, and outlining data cleaning procedures. Additionally, I chose the "Analyze" stage because data exploration is a crucial step in analyzing the dataset to understand its structure, identify patterns, and detect anomalies, which are essential for informing subsequent analysis and decision-making.

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

Why did you select this stage for this task?

I selected the "Plan" stage for the task "Establish structure for project workflow (PACE)" because it involves laying out the overall framework and organization for how the project will be executed. This task focuses on setting up the structure, milestones, and timelines for the project, which is a fundamental aspect of project planning. In the "Plan" stage, we define the project goals, scope, and objectives, as well as identify the resources and strategies needed to achieve them. By establishing the structure for the project workflow at the outset, we can ensure that the project progresses smoothly and efficiently towards its goals.

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



Why did you select this stage for this task?

I selected the "Execute" stage for the task "Communicate final insights with stakeholders" because this task involves the final phase of the project, where the focus shifts from analysis to sharing the outcomes with relevant stakeholders. In this stage, the primary objective is to effectively communicate the insights, findings, and recommendations derived from the data analysis process to the stakeholders. This includes preparing presentations, reports, dashboards, or any other relevant deliverables to ensure that the insights are clearly understood and actionable for decision-making purposes. By choosing the "Execute" stage, I aim to emphasize the importance of effectively disseminating the project's results to drive meaningful outcomes and actions within the organization.

7. Compute descriptive statistics: **Analyze** ▾

Why did you select this stage for this task?

In this stage, data analysts examine the dataset to understand its characteristics, patterns, and relationships. Computing descriptive statistics, such as measures of central tendency, variability, and distribution, is crucial for gaining insights into the data's basic properties and preparing for further analysis. Descriptive statistics provide a summary of the dataset's key features, enabling analysts to identify trends, outliers, and potential areas of interest for deeper exploration and modeling. Therefore, this task is foundational to the analysis process and aligns with the "Analyze" stage's focus on understanding the data.

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



Why did you select these stages for this task?

I selected the "Analyze" and "Construct" stages for the task of "Visualization building" because:

1. **Analyze Stage**: This stage involves exploring the data to understand its structure, patterns, and relationships. Building visualizations during the analysis stage allows us to gain insights into the data and identify important trends or outliers that may inform subsequent steps in the project.
2. **Construct Stage**: In this stage, we begin to develop solutions or models based on the insights gained from the data analysis. Visualization building continues in the construct stage to communicate the findings effectively to stakeholders, aiding in decision-making and further development of the project. Visualizations created at this stage help in presenting the results of our analysis and proposed solutions in a clear and understandable manner, facilitating discussions and alignment with project goals.

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

Why did you select this stage for this task?

In this stage, we define the project scope, objectives, and deliverables, as well as establish a roadmap for how the project will be executed. This involves outlining the tasks, timelines, and resources required to complete the project successfully. Additionally, during the planning stage, we identify potential risks and develop strategies to mitigate them. Overall, the project proposal sets the foundation for the entire project, guiding the team towards achieving its goals efficiently and effectively.

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



Why did you select this stage for this task?

Analyze: In this stage, we delve into the data to understand its characteristics, identify patterns, relationships, and potential predictors that could be relevant for the regression model. This involves data exploration, descriptive statistics, and possibly hypothesis testing to gain insights into the variables and their interactions.

Construct: Once we have thoroughly analyzed the data and identified the relevant variables, we move on to the construction stage. Here, we actually build the regression model using techniques such as ordinary least squares regression, logistic regression, or other appropriate methods based on the nature of the problem and the data. This stage involves selecting the appropriate model, fitting it to the data, and evaluating its performance.

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

Why did you select this stage for this task?

This stage involves exploring the dataset to understand its structure, characteristics, and key attributes. Compiling summary information about the data includes tasks such as calculating descriptive statistics (e.g., mean, median, standard deviation), examining data distributions, identifying missing values, and assessing data quality.

By selecting the "Analyze" stage for this task, we acknowledge the importance of gaining insights into the dataset before proceeding with further analysis or model building. This initial exploration lays the foundation for making informed decisions throughout the project.

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

Why did you select this stage for this task?



In this stage, the focus is on creating the machine learning model based on the insights gained from data analysis. Tasks include selecting appropriate algorithms, preprocessing data, feature engineering, model training, validation, and fine-tuning hyperparameters.

By selecting the "Construct" stage for this task, we indicate that the primary objective is to build and refine the machine learning model based on the analyzed data. This stage involves implementing the planned approach and developing a model that effectively predicts user churn.