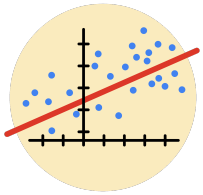
Course Five

Regression Analysis: Simplifying Complex Data Relationships



Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. As a reminder, this document is a resource that you can reference in the future, and a guide to help you consider responses and reflections posed at various points throughout projects.

Course Project Recap

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

- ☑ ~~Complete the questions in the Course 5 PACE strategy document~~
- ☑ ~~Answer the questions in the Jupyter notebook project file~~
- ☑ ~~Build a multiple linear regression model~~
- ☑ ~~Evaluate the model~~
- ☑ ~~Create an executive summary for team members~~

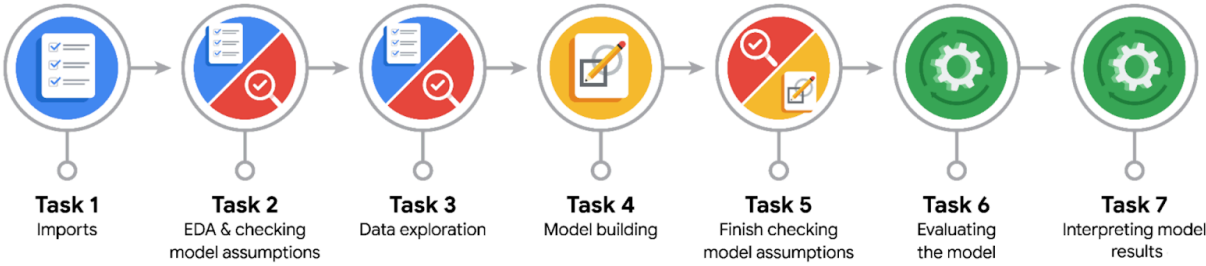
Relevant Interview Questions

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

- Describe the steps you would take to run a regression-based analysis
- List and describe the critical assumptions of linear regression
- What is the primary difference between R^2 and adjusted R^2?
- How do you interpret a Q-Q plot in a linear regression model?
- What is the bias-variance tradeoff? How does it relate to building a multiple linear regression model? Consider variable selection and adjusted R^2.

Reference Guide

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



Data Project Questions & Considerations

PACE: Plan Stage

- Who are your external stakeholders for this project?

The main stakeholders include TikTok's Operations Lead (**Maika Abadi**) and the broader Data Science and Trust & Safety teams. These teams are interested in understanding what factors predict whether a user is verified. The analysis will also inform leadership on strategies related to user engagement and content moderation.

- What are you trying to solve or accomplish?

The goal is to determine which video features best predict a user's verification status. Specifically, the project aims to explore how engagement metrics such as video likes, comments, and duration are associated with verified accounts. The logistic regression model will help TikTok understand engagement patterns that distinguish verified users from non-verified ones.

-
-

- What are your initial observations when you explore the data?

Initial data exploration revealed that verified users generally had higher engagement levels, particularly in terms of comment activity. Some numerical variables had large value ranges and needed scaling, while categorical variables required encoding. The dataset appeared balanced enough for classification, though there was moderate variability in engagement metrics.

- What resources do you find yourself using as you complete this stage?

Python libraries such as **pandas**, **numpy**, **scikit-learn**, and **matplotlib**. Course readings on logistic regression, model evaluation, and feature encoding. The provided **TikTok claims classification dataset** and **Jupyter notebook template**. Prior coursework on hypothesis testing and exploratory data analysis (EDA).

PACE: Analyze Stage

- What are some purposes of EDA before constructing a multiple linear regression model?

Identify variable relationships, outliers, and potential multicollinearity. Verify data quality and detect missing or skewed values. Guide feature selection and inform decisions about encoding and scaling. Ensure assumptions (such as linearity, independence, and homoscedasticity) are considered before model construction.

-
-
-

- Do you have any ethical considerations at this stage?

Yes. It's essential to ensure that variables used in the model do not introduce bias or unfairly influence verification outcomes. For instance, demographic or sensitive user information should not be included in predictive modeling. Transparency and data privacy are key ethical considerations.

PACE: Construct Stage

- Do you notice anything odd?

During model construction, some variables (e.g., like count) showed unexpected weak or negative relationships with verification. This suggests that simple popularity metrics are not necessarily predictive of verified status.

- Can you improve it? Is there anything you would change about the model?

Yes. Improvements could include:

Adding more relevant variables, such as follower count or posting frequency.

Applying regularization (e.g., L2 penalty) to handle potential multicollinearity.

Exploring nonlinear models like Random Forests or Gradient Boosted Trees for comparison.

- What resources do you find yourself using as you complete this stage?

Scikit-learn documentation for logistic regression and model evaluation functions.

Visualization tools (matplotlib, seaborn) for confusion matrix and coefficient interpretation.

Peer examples and course notebooks to validate approach consistency.

PACE: Execute Stage

- What key insights emerged from your model(s)?

Comment count was the most significant positive predictor of verification.

Like count and **video duration** had small negative coefficients, indicating they are less reliable predictors.

The model achieved **63% accuracy** and **88% recall**, suggesting strong identification of verified users but some false positives among non-verified users.

- What business recommendations do you propose based on the models built?

- Encourage engagement-driven strategies that promote meaningful audience interaction, not just likes or views.
- Use comment activity as a stronger indicator of user influence and verification potential.
- Apply this model framework to help automate early verification suggestions or content classification.

- To interpret model results, why is it important to interpret the beta coefficients?

The beta (log-odds) coefficients reveal the direction and strength of each predictor's relationship with the outcome. Interpreting these helps stakeholders understand **which variables truly influence verification** and how changes in those features affect the probability of a user being verified.

- What potential recommendations would you make?

Refine the model using more diverse engagement variables. Conduct regular model audits to check for performance drift or bias. Integrate model predictions into broader trust and safety review systems.

- Do you think your model could be improved? Why or why not? How?

Yes. Model accuracy can be improved by: Adding more explanatory variables (e.g., follower growth rate). Testing alternative algorithms with better generalization (e.g., ensemble methods). Applying feature engineering techniques to transform engagement ratios into more meaningful predictors.

- What business/organizational recommendations would you propose based on the models built?

Develop policies encouraging authentic engagement through meaningful interactions (comments, shares).

Use insights from this analysis to improve TikTok's verification processes.

Align model findings with influencer and creator support strategies to maintain platform trust and authenticity.

- Given what you know about the data and the models you were using, what other questions could you address for the team?

What other engagement metrics might better predict verification (e.g., duet count, follower-to-like ratio)?

Can content topics or hashtags influence verification probability?

How does verification affect future engagement growth?

- Do you have any ethical considerations at this stage?

Yes. It's crucial to avoid using models that could unintentionally reinforce bias or misrepresent users. Predictions should **not** determine verification eligibility without human oversight. Additionally, user privacy must be maintained, and sensitive attributes should be excluded from future analyses.