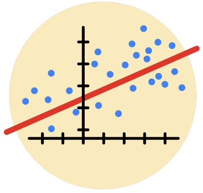
Course Five

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

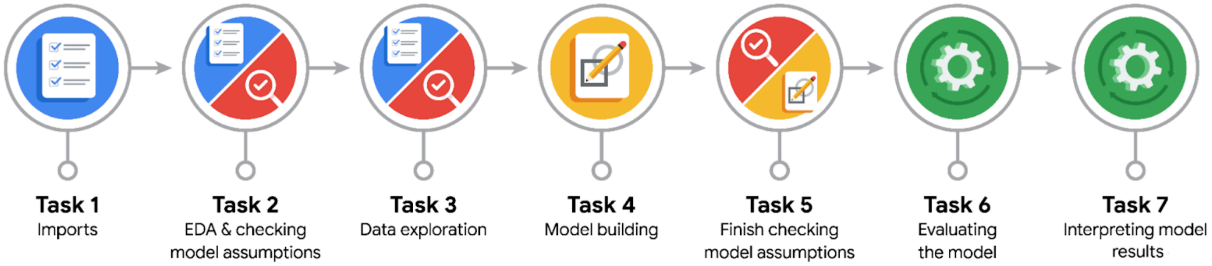
Relevant Interview Questions

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?

TikTok's Operations Lead (Maika Abadi), the wider data team, content moderators, and ultimately, TikTok users.

- What are you trying to solve or accomplish?

To build a logistic regression model that can predict whether a TikTok user is verified based on video characteristics. This model will help prioritize the review of user reports and inform the broader model for claim vs. opinion classification.

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

- The dataset contains a mix of numerical and categorical variables.
- There are missing values that need to be addressed.
- The target variable (verified_status) is imbalanced.
- There is a high correlation between video engagement metrics.
- There are outliers in the video engagement metrics.

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

Pandas for data manipulation, NumPy for numerical operations, and initial data exploration functions (e.g., `.head()`, `.info()`, `.describe()`, `.isna()`).

PACE: Analyze Stage

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

- To identify and handle missing values and outliers.
- To understand the distribution of variables and check for normality.
- To assess multicollinearity among features.
- To examine the relationship between features and the target variable.
- To check the balance of the target variable.

- Do you have any ethical considerations at this stage?

Yes, ensuring data privacy and avoiding biases in feature selection that could disproportionately affect certain user groups. Also, being aware of the potential for the model to reinforce existing biases related to who gets verified.

PACE: Construct Stage

- Do you notice anything odd?

- The strong correlations among video engagement metrics (multicollinearity).
- The imbalance of the target variable.

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

- Yes, we addressed the multicollinearity by excluding `video_like_count`.
- We also addressed the class imbalance by using resampling.
- We could test other models, and perform further feature engineering.

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

  - Scikit-learn for one-hot encoding, train-test split, and logistic regression modeling.
  - Resample from sklearn for the upsampling of the minority class.

PACE: Execute Stage

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

  - Video duration and engagement metrics are relevant predictors of verified status.
  - The model has moderate accuracy, with better recall for unverified users.
  - Multicollinearity needs to be addressed for reliable model results.

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

  - Prioritize the review of videos from potentially verified users, especially those with high engagement.
  - Investigate further feature engineering to improve model accuracy.
  - TikTok could consider using the models findings about video engagement when determining who to verify.

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

Beta coefficients indicate the strength and direction of the relationship between each feature and the log-odds of the outcome. Understanding these coefficients helps to assess the impact of each feature on the likelihood of a user being verified.

- What potential recommendations would you make?

Further refine the model by exploring other features and algorithms.

Conduct ongoing monitoring of model performance and update as needed.

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

Yes, by exploring advanced feature engineering, trying other machine learning algorithms (e.g., tree-based models), and conducting more in-depth hyperparameter tuning.

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

Integrate the model into the content moderation workflow to prioritize reviews.

Use the model to identify patterns and trends related to verified users.

Provide feedback to content creators regarding what types of content is more likely to be produced by verified users.

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

How does the model perform on different demographics or content categories?

What are the key factors that contribute to high engagement for verified users?

Can we predict future verified users based on content creation patterns?

- Do you have any ethical considerations at this stage?

Yes, ensuring fairness and transparency in how the model is used.

Monitoring for unintended biases and taking steps to mitigate them.

Communicating clearly with users about how verification decisions are made.