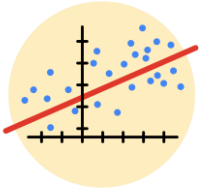
Course Five

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

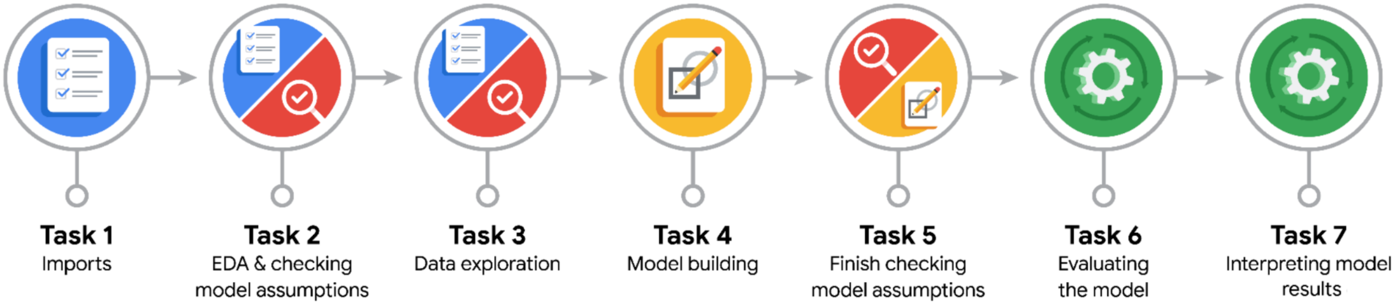
Relevant Interview Questions

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.



Task 1 Imports

Task 2 EDA & checking model assumptions

Task 3 Data exploration

Task 4 Model building

Task 5 Finish checking model assumptions

Task 6 Evaluating the model

Task 7 Interpreting model results

Data Project Questions & Considerations

PACE: Plan Stage

- Who are your external stakeholders for this project?

Maika Abadi (Operations Lead), Mary Joanna Rodgers (Project Management Officer), Rosie Mae Bradshaw (Data Science Manager), Willow Jaffey (Data Science Lead), Orion Rainer (Data Scientist).

- What are you trying to solve or accomplish?

Predicting verified status to help understand how video characteristics relate to verified users.

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

That there are some outliers in video_like_count, and that these need to be dealt with. The outcome variable is also not very balanced with 8921 rows being not verified users and 994 rows being verified users, therefore there needs to be resampling in order to create a class balance in the outcome variable.

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

The resources I find myself using are the appropriate Python Packages such as pandas, NumPy, Seaborn and scikit-learn.

PACE: Analyze Stage

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

Before constructing a multiple linear regression model, EDA helps identify patterns, trends, and relationships in the data. It allows the detection of missing values, outliers, and anomalies. EDA also helps verify that the assumptions of linear regression—such as linearity, independence, normality, and homoscedasticity—are met. In addition, it is used to assess multicollinearity between predictors and to guide feature selection and transformation.

- Do you have any ethical considerations in this stage?

During this stage, it is important to avoid introducing bias by ensuring the dataset is representative. Data privacy and confidentiality must be maintained at all times. Analysts should be transparent about any limitations, assumptions, or potential errors in the analysis. It is also unethical to manipulate data to fit a preferred outcome, and consideration should be given to how the analysis results might impact individuals or groups.

PACE: Construct Stage

- Do you notice anything odd?

That the outcome variable and author_ban_status and claim_status are a data type object after splitting the data into training and testing sets.

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

Yes, both the outcome variable and the two categorical independent variables training and test datasets can be encoded in order to make them numeric.

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

The OneHotEncoder Function from scikit-learns's preprocessing module.

PACE: Execute Stage

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

The logistic regression model revealed that opinion-based claim statuses and higher share counts are the strongest positive predictors of verification likelihood, with comment counts showing a weaker positive effect. In contrast, higher download counts, longer video durations, enforcement actions (banned or under review), and higher view counts are negatively associated with verification. These findings highlight that not all engagement boosts verification probability, shares help, but views and downloads do not. The model achieved 69% accuracy, with stronger recall for non-verified users, indicating a class imbalance.

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

To improve verification prediction and moderation decisions, TikTok should focus on enhancing features that drive positive verification signals, such as share counts and opinion-based content, while carefully reviewing cases with high views or downloads, which may indicate lower credibility. Addressing class imbalance in the data will improve model precision, and testing advanced algorithms like Random Forest or XGBoost could enhance accuracy. Finally, the model should be used as a decision-support tool alongside human review, not as the sole determinant of verification.

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

Interpreting beta coefficients is important because they indicate the direction and strength of the relationship between each feature and the prediction outcome. This allows businesses to understand which variables most influence the model's decisions, make informed changes to processes, and ensure that decisions are explainable and transparent.

- What potential recommendations would you make?

- Improve data quality by addressing class imbalance.
- Incorporate additional contextual features, such as content sentiment and audience engagement quality.
- Explore advanced machine learning models (e.g., Random Forest, XGBoost) to improve predictive accuracy.
- Integrate the verification model as a support tool in moderation workflows, ensuring human review for balanced and transparent decision.

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

Yes, the model could be improved because while it achieved moderate accuracy (69%), it shows an imbalance between precision and recall, particularly for verified users. Improvements could include addressing class imbalance in the dataset, engineering new features that capture engagement quality and context, and experimenting with advanced algorithms like Random Forest or XGBoost to enhance predictive performance. Additionally, incorporating richer metadata such as posting frequency, audience demographics, and content sentiment could improve its predictive capability.

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

The model should be implemented as a decision-support tool within the verification process, with final approval remaining in the hands of human moderators. Business strategies should focus on promoting engagement types that correlate positively with verification likelihood, such as content shares and opinion-based posts, while applying stricter review procedures to accounts with unusually high view or download counts, as these may signal lower credibility. Moderation teams should be trained to interpret the model's outputs effectively and understand its limitations, ensuring balanced and fair decision-making. Additionally, the model should be continuously retrained with updated data to adapt to evolving user behavior and platform trends.

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

- How does model performance differ across various content categories or topics?
- What effect does the time elapsed since upload have on verification likelihood?
- Could sentiment analysis of comments enhance predictive accuracy?
- Would incorporating creator reputation history improve model performance?

- Do you have any ethical considerations at this stage?

Yes, It is important to ensure that verification decisions do not unfairly disadvantage certain creators or content categories. The model's false positive rate must be carefully managed to avoid granting verification to inappropriate content. Data privacy must be upheld, and model decisions should remain explainable to maintain trust and transparency. The model should be used as a decision-support tool alongside human review, not as the sole determinant of verification status.