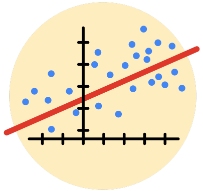
Course Five

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~~

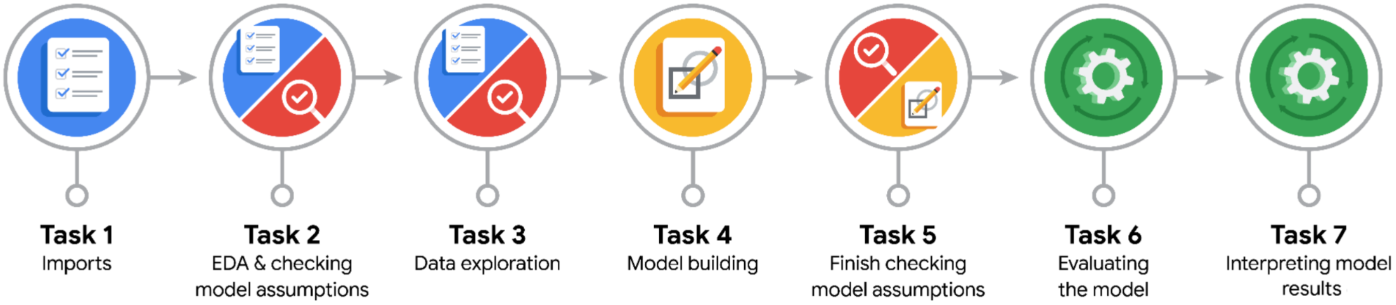
Relevant Interview Questions

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?

The external stakeholders for this project are the managers of tik-tok that are looking to fix the slow classification process of author verification.

- What are you trying to solve or accomplish?

We are trying to construct a model that can classify videos, based on their engagement metrics, if their authors are verified or not.

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

There is a substantial difference between the number of verified authors and unverified authors, this meant we had to upsample the verified author population in order to reduce the bias due to the different group size.

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

I used a lot of the resources introduced in previous courses for exploratory data analysis (EDA), as I would need to observe possible correlations between variables and remove biases from the difference between these two or more groups.

PACE: Analyze Stage

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

Observe possible multicollinearity in different variables, which can help assess what variables can be considered when testing the model.

- Do you have any ethical considerations in this stage?

The main ethical consideration at this stage concerns the the reason why so many banned authors have the most views, it could be that they are spreading harmful content in the platform and this spreads faster.

PACE: Construct Stage

- Do you notice anything odd?

Verified authors have altogether lower engagement metrics than unverified authors, it would appear the type of content differs considerably and this could lead to the banning of many authors with higher engagement metrics.

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

We don't want the model to discriminate videos with higher views or likes as a potential threat to the platform; instead we would want to look at other metrics like transcripts length or share counts which could be a better indication of this specific type of content.

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

I used several techniques for data aggregation, and visualization for detecting and displaying these trends in the data. The correlation matrix from pandas was especially helpful for the EDA practices

PACE: Execute Stage

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

Using the engagement metrics such as views, likes, shares, download and comments does not seem to classify true negatives (not verified) properly as we see the precision of the model to be barely above that of a randomizer.

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

Create a video category section, as this will illustrate the type of content presented in the videos because this could explain the leading factor for unverified and banned authors in the platform.

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

The model coefficients indicate how much the features affect the categorical variables we are trying to predict, in this case author verification status.

- What potential recommendations would you make?

Video transcription length has the highest beta coefficient; this could be indicative of type of content as the leading factor for categorizing authors in verification status.

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

Yes, the model can be improved by removing certain engagement variables which we are observing most unverified/banned authors have which could be correlated with the type of content presented in their videos.

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

Make the content creators pick a video category to which they believe their content would fit in, this will provide important insights as to the type of content each video is putting out.

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

We see a positive beta coefficient between transcription length and verification status, we need to look for keywords in these transcripts in order to find other important information about the content of these videos.

- Do you have any ethical considerations at this stage?

We need to make sure not to penalize highly successful videos as problematic given the strong indication that banned authors have the most viral videos in the platform, we need to find a way to create a classification methods that does target highly successful videos, as this could potentially hurt real content creators who might have a shot a going viral with content that is not harmful.