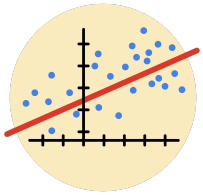
Course Five

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

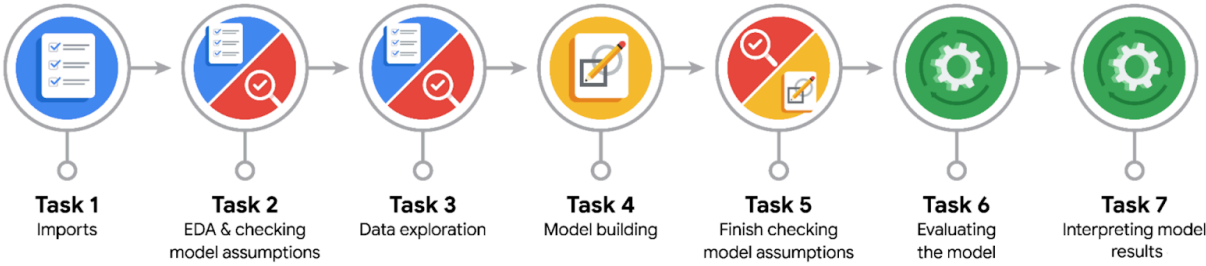
Relevant Interview Questions

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 TikTok Operations team and other non-data TikTok teams as they will be affected by the model we create.

- What are you trying to solve or accomplish?

We are trying to create a model that will predict the authors verification status based on other factors from the video.

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

There is colinearity between some engagement metrics so they cannot al be used for the model, there is also much less verified users than non verified ones so data must be resampled to create a better model.

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

Various python libraries, including statsmodels, scipy and pandas for EDA.

PACE: Analyze Stage

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

To ensure the selected features are linear and have no multicollinearity.

- Do you have any ethical considerations at this stage?

The data should not include any personally identifiable information and the model should not be used for making automatic decisions that may be biased against a certain group.

PACE: Construct Stage

- Do you notice anything odd?

There isn't many usable features and many combinations did not create great models, e.g. {view count, claim status, author ban status} as features made it only predict 0

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

It can be improved by increasing its overall accuracy but would require additional data.

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

The sklearn python library as it included everything that was necessary.

PACE: Execute Stage

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

It is difficult to accurately predict the users verification status and it will produce more false positives than false negatives.

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

The model is not ready to be used yet as its accuracy is not good enough to be used for decision making. More data is needed to construct a better model.

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

To see if there are any features which are being used much more than any others.

- What potential recommendations would you make?

Try building more models and comparing their performances to see which is best and then make informed decision on how to proceed.

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

Yes by using more advanced methods such as XGBoost or other ensemble methods or trying different algorithms entirely.

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

We can analyse which engagement metric applies to verification status the most or if there are any other hidden correlations, specifically with sentiment of text and verification status.

- Do you have any ethical considerations at this stage?

Model performs too poorly to be used as automated decision making so verification status should still be manually verified for now.