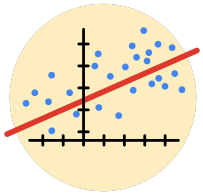
Course Five

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~~

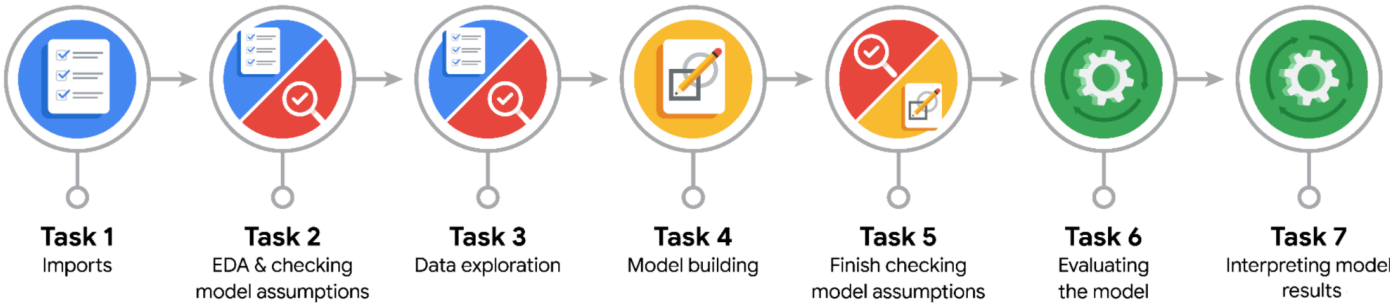
Relevant Interview Questions

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.

- What are you trying to solve or accomplish?

Conduct a logistic regression using verified status as the outcome variable.

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

93.71% of the dataset represents videos posted by unverified accounts and 6.28% represents videos posted by verified accounts. So the outcome variable is not very balanced.

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

Packages for data manipulation.
Packages for data visualization.

PACE: Analyze Stage

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

Identify outliers and extreme value that can impact model.

- Do you have any ethical considerations in this stage?

Identify outliers and extreme value that can impact model.

PACE: Construct Stage

- Do you notice anything odd?

Each additional second of the video is associated with 0.01 increase in the log-odds of the user having a verified status.

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

No. The model already decent predictive power (67% precision and 65% recall).

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

Packages for data preprocessing.
Packages for data modeling.

PACE: Execute Stage

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

Each additional second of the video is associated with 0.01 increase in the log-odds of the user having a verified status.

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

Construct a classification model that will predict the status of claims made by users.

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

Helps identify the values that are currently in use.

- What potential recommendations would you make?

Classify claims and opinions, it's important to build a model that shows how to predict the behavior of the account type (verified) that tend to post more opinions.

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

No. The model already decent predictive power (67% precision and 65% recall).

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

Need more data to improve confusion matrix result ?

- Do you have any ethical considerations at this stage?

Recommend based on finding or on model result.