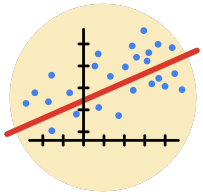
Course Five

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

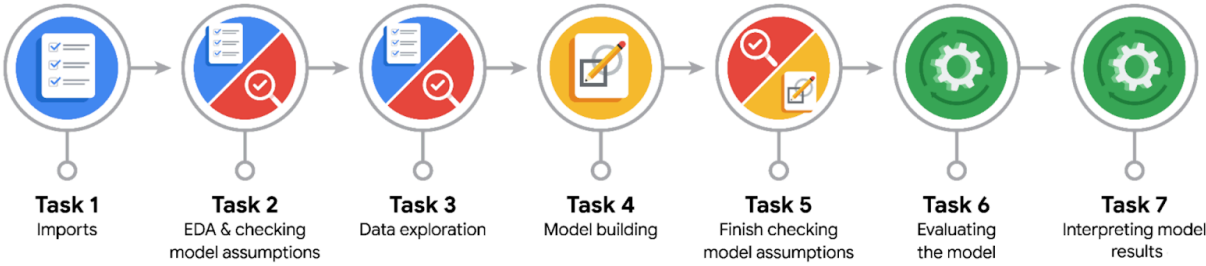
Relevant Interview Questions

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?

Mary Joanna Rodgers- Project Management Officer, Margery Adebowale-Finance Lead, Americas and Maika Abadi- Operations Lead

- What are you trying to solve or accomplish?

To find the features or variables that can help determine the verified status

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

The verified status outcome variable is binomial and the engagement rates could be something to keep an eye on for predicting the verified status

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

Sci-kit learn libraries, matplotlib,pandas,seaborn and numpy

PACE: Analyze Stage

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

To clean and validate data,develop a better understanding of it as well as convert the data into appropriate formwell as convert data for the regression model.

- Do you have any ethical considerations at this stage?

Yes, to make sure we consider all variables worth considering and we take both outcome variables , verified and non verified as equally important outcome predictions, reducing bias.

PACE: Construct Stage

- Do you notice anything odd?

The ‘video like count’ variable is severely correlated with rest of independent variables

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

We exclude it from the model features.

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

Sci-kit learn library packages.(metrics,linear_model,preprocessing,model_selection)

PACE: Execute Stage

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

shorter videos tend to be associated with higher odds of the user being verified.

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

The next step is to 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?

Because it helps understand the difference one unit change in the independent variable makes in the odds of probability of the dependent variable

- What potential recommendations would you make?

To further explore the relationship between verified status and claim status.

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

Yes , because Overall accuracy is towards the lower end of what would typically be considered acceptable.We can improve it by providing more data.

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

As shorter videos tend to have better odds for verified status, and verified users are much more likely to post opinions, exploring this relationship further can help streamline the classification process for claims.

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

Answer questions like video like count is related to other engagement rates,how shorter videos tend to have better odds for verified status,how video text transcript length relates to the verified status or if at all significantly

- Do you have any ethical considerations at this stage?

To appropriately manage the bias-variance trade off.