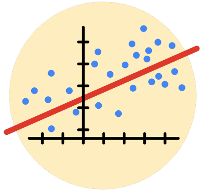
Course Five

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

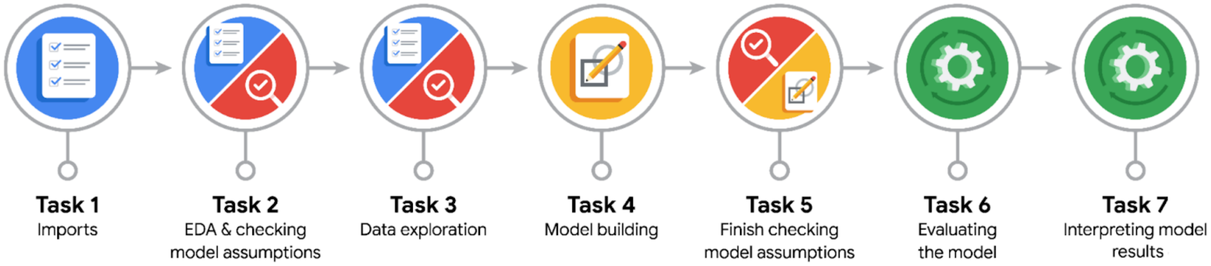
Relevant Interview Questions

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 Operations team is the main stakeholder along with the data science team.

- What are you trying to solve or accomplish?

The goal of the project is to build a logistic regression model to predict the verified status of the user given the other variables.

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

Some of the variables might be correlated with each other.

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

The Python packages for EDA.

PACE: Analyze Stage

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

Check for issues in the dataset and analyse the variables to check that the model assumptions are met.

- Do you have any ethical considerations at this stage?

Not at this stage.

PACE: Construct Stage

- Do you notice anything odd?

Video like count and video share count variables are very strongly correlated with each other (0.86).

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

In order to meet the model assumptions, only one of the correlated variables (video like count and video share count) should be used in the model construction.

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

Python package Sklearn

PACE: Execute Stage

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

The dataset has a few strongly correlated variables, which might lead to multicollinearity issues when fitting a logistic regression model. We decided to drop video_share_count from the model building. The logistic regression model had not great, but an acceptable predictive power: a precision of 61% is less than ideal, but a recall of 84% is very good. Overall accuracy is towards the lower end of what would typically be considered acceptable.

Video features included in the model have small estimated coefficients, so their association with verified status seems to be small.

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

The video features included in the current analysis are not sufficient to construct a model that can predict verified status. More feature engineering might be required.

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

Beta coefficients give a sense of how much effect a change in a predictive variable can have on the outcome variable.