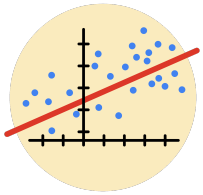
Course Five

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~~

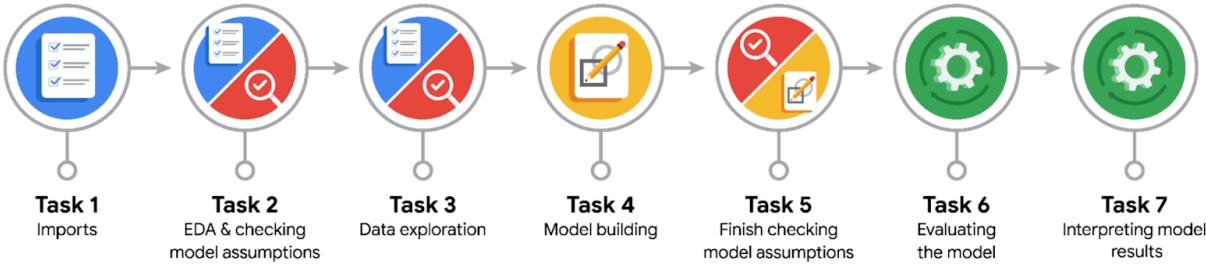
Relevant Interview Questions

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?

Waze investors, Waze users and fan base.

- What are you trying to solve or accomplish?

We are trying to accomplish building a binary logarithmic regression to help support predicting user churn on the Waze app.

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

Initially, the data looks standard out side of the 700 records missing from the labels column.

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

I find myself using a combination of packages such as numpy, pandas, seaborn, matplotlib, scikit-learn etc. as well as frequent online searches for clarification.

PACE: Analyze Stage

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

Need to get acquainted with the data and figure out summary points, the ranges of data points in each column and categorical vs, numeric data types.

- Do you have any ethical considerations at this stage?

No ethical considerations as long as the samples we are using are well distributed and fair.

PACE: Construct Stage

- Do you notice anything odd?

The 700 rows with missing information only make up 5% of the overall data

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

The 700 lines of data have been dropped

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

Lots of packages connected to python and Google searches

PACE: Execute Stage

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

It was identified that activity days was the most important category in the dataset for helping predict churn. As well as the professional driver label to distinguish occasional users from professional use.

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

Definitely will need to tighten up the recall capabilities of the model as even though the model is precise on each of the cases it identifies. Overall, it is letting a lot of examples pass by instead of correctly recalling them.

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

By analyzing the coefficients, I am able to rank them and figure out which one is most influential in the model.

- What potential recommendations would you make?

I would recommend nothing specific aside from targeting the right factors to display in the model.

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

Not really - I would need more data to improve it.

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

No additional recommendations

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

I could map user's favorite locations to drive to with Waze.

- Do you have any ethical considerations at this stage?

Possibly user privacy and being biased in samples we are using.