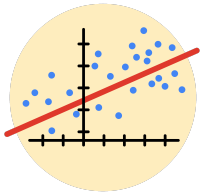
Course Five

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~~

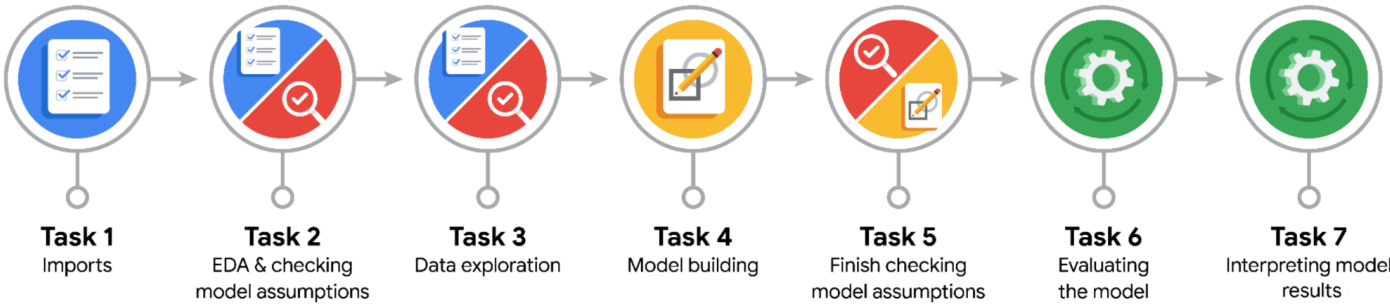
Relevant Interview Questions

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.

- What are you trying to solve or accomplish?

Build a regression model to predict user churn based on a variety of variables.

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

Some variables contain outliers.
The churn rate for professional drivers is 7.5%, while the churn rate for non-professionals is 19.8%.

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

Packages for numerics and dataframes. Packages for visualization.

PACE: Analyze Stage

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

Understanding data and distribution. Check outlier, extreme value, and missing value that can impact model.

- Do you have any ethical considerations in this stage?

Check outlier, extreme value, and missing value that can impact model.

PACE: Construct Stage

- Do you notice anything odd?

Yes, it's activity_days.

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

Yes, by add new features it could be generate better predictive signal. One of the engineered features (professional_driver) was the third-most-predictive predictor. It could also be helpful to scale the predictor variables, and/or to reconstruct the model with different combinations of predictor variables to reduce noise from unpredictive features.

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

Packages for Logistic Regression & Confusion Matrix.

PACE: Execute Stage

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

User churn rate increased as the values in km_per_driving_day increased. The model is not a strong enough predictor, as made clear by its poor recall score. However, if the model is only being used to guide further exploratory efforts, then it can have value.

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

It would be helpful to have drive-level information for each user (such as drive times, geographic locations, etc.). It would probably also be helpful to have more granular data to know how users interact with the app.

- 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?

Drive-level information for each user (such as drive times, geographic locations, etc.).

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

Yes, by add new features it could be generate better predictive signal. One of the engineered features (professional_driver) was the third-most-predictive predictor. It could also be helpful to scale the predictor variables, and/or to reconstruct the model with different combinations of predictor variables to reduce noise from unpredictive features.

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

Is there other engineered features such as professional_driver that could also be helpful ?

- Do you have any ethical considerations at this stage?

Recommend based on finding or on model result.