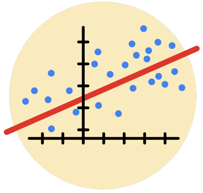
Course Five

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

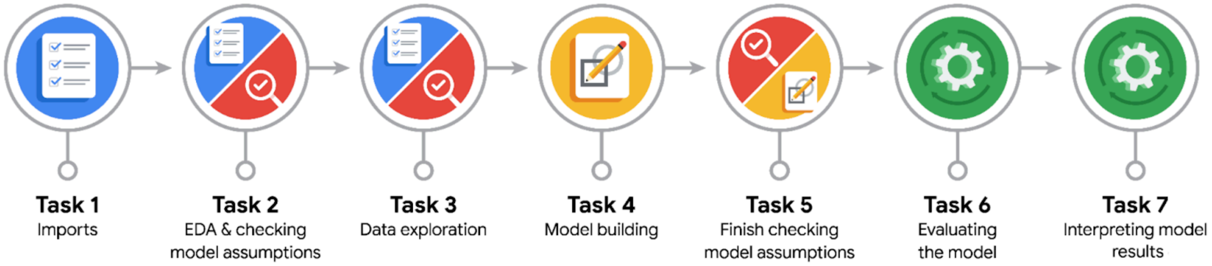
Relevant Interview Questions

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 New York Taxi & Limousine Commission (TLC) is the external stakeholder for this project.

- What are you trying to solve or accomplish?

The goal of this project is to develop a multiple linear regression model to predict taxi cab fares before a ride commences.

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

There are initially outliers with extreme large or negative data values.

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

Pandas and numpy libraries are used to explore and containerize data. Datetime library is used to convert dates with object data type to datetime format. Matplotlib and seaborn libraries are used for data visualization.

PACE: Analyze Stage

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

EDA is used to clean the dataset, make certain values fall within an interquartile range, and to reformat variables for data aggregation and calculations.

- Do you have any ethical considerations at this stage?

All the data has been fairly collected by TLC and no personal customer data like credit card numbers are present in the dataset.

PACE: Construct Stage

- Do you notice anything odd?

The continuous data is in different units. Some of the categorical data can be encoded. The linear regression model has not had its hyperparameters tuned.

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

The model should be tuned to improve model performance. Currently, it only has been loaded with its default parameters without regression techniques (Ridge, Lasso).

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

Scikit-learn libraries are used for preprocessing and modeling. Scikit-learn libraries for model scoring and seaborn for visualizing target-predictor correlations, residual distribution.

PACE: Execute Stage

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

The model explained 87% of the variance for fare amount. The mean error error was about $2, so the predicted values are off by $2 on average. The variable that had the largest impact on fare amount was mean_distance.

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

Taxi cab drivers should prioritize trips that involve long-distance travel to maximize profits.

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

Beta coefficients explain the change in the predictor variable in relation to the target variable. For a standardized predictor variable, one standard deviation in mean_distance would lead to a $7.11 increase in the fare_amount target variable.

- What potential recommendations would you make?

The number of predictor variables that would not result in data leakage was relatively small. More transformations can be performed on the dataset, such as taking the log scale of the predictor to spread out close values and reign in farther ones.

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

The model should incorporate more variables that would affect the fare amount, such as accounting for traffic for holidays or seasonal events.

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

Based on the model, taxi cab drivers should prioritize longer rides on average to increase profits.

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

Can predefined routes (Google Maps routes) be considered in relation to average trip distance? What other factors besides traffic affect average ride duration?

- Do you have any ethical considerations at this stage?

Gathering personal data on customer demographics could lead to a practice where certain groups pay higher on average based on model predictions.