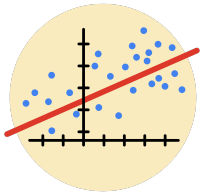
Course Five

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

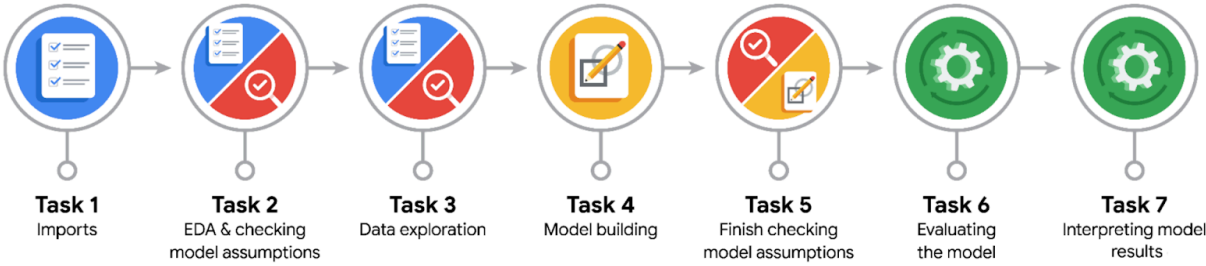
Relevant Interview Questions

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?

My external stakeholders are the taxi company's management, particularly those involved in fare setting, operational planning, and financial analysis. Additionally, customers who rely on accurate fare estimations are indirect stakeholders.

- What are you trying to solve or accomplish?

I am trying to build a reliable multiple linear regression model that accurately predicts taxi fare amounts. This will enable the company to optimize fare setting, improve customer satisfaction through accurate estimations, and gain insights into the factors that influence fare prices.

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

• Strong positive correlations between trip distance, duration, and fare amount. • The presence of outliers in the fare amount and trip data. • Variations in fare amounts based on pickup and dropoff locations. • The existence of rate codes that have a fixed price. • The existence of rush hour effects. • Vendor ID having an effect on price.

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

• Pandas and NumPy for data manipulation and exploration. • Seaborn and Matplotlib for data visualization. • The data dictionary for understanding the meaning of each column. • Online resources and documentation for statistical concepts and libraries.

PACE: Analyze Stage

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

• Identifying relationships between variables. • Detecting outliers and anomalies. • Understanding the distribution of data. • Selecting relevant features for the model. • Checking for multicollinearity. • Validating assumptions of linear regression.

- Do you have any ethical considerations at this stage?

• Ensuring data privacy and avoiding the use of sensitive information. • Being transparent about any data cleaning or manipulation. • Avoiding biased interpretations of the data that could lead to unfair fare practices.

PACE: Construct Stage

- Do you notice anything odd?

I noticed the horizontal lines in the scatter plot of `mean_duration` vs. `fare_amount`, which were explained by the fixed fare for JFK trips and the imputed maximum fare for outliers. Also, I noticed the high correlation between mean duration and mean distance.

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

• Further feature engineering to capture more complex relationships. • Exploring non-linear regression models if linear assumptions are violated. • Gathering more data to increase the model's robustness. • Testing and validating the model with more up to date data, to check for model drift.

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

• Scikit-learn for model building and evaluation. • Statsmodels for statistical analysis. • Online tutorials and documentation for regression techniques.

PACE: Execute Stage

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

• `mean_distance` and `mean_duration` are strong predictors of fare amount. • Vendor ID and rush hour have a lesser, but still measurable, impact on fare. • The model has a high R^2 value, indicating a good fit.

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

• Use the model for fare estimation and planning. • Analyze the impact of rush hour and vendor ID on fares. • Consider dynamic fare adjustments based on predicted trip characteristics. • Investigate the causes of longer than average trip durations, to see if efficiency can be increased.

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

Beta coefficients quantify the relationship between each independent variable and the dependent variable. Understanding these coefficients is crucial for determining the impact of each feature on the predicted fare.

- What potential recommendations would you make?

• Implement a fare estimation tool for customers. • Optimize driver dispatching based on predicted trip durations. • Conduct further analysis on the impact of external factors like traffic and weather.

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

Yes, the model could be improved by: - Including more relevant features, such as traffic data. - Using more complex modeling techniques, such as polynomial regression. - Gathering more data, to increase the models robustness. - Continually checking the model against new data.

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

• Use the model for strategic pricing decisions. • Develop driver training programs to improve efficiency. • Invest in technology to collect more granular trip data.

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

• How do external factors like weather and events affect fares? • Can we predict customer demand based on time and location? • What is the optimal fleet size and distribution? • Can we predict the likelihood of tip amount?

- Do you have any ethical considerations at this stage?

• Ensuring fairness in fare setting and avoiding discriminatory practices. • Being transparent with customers about how fares are calculated. • Protecting customer privacy and data security.