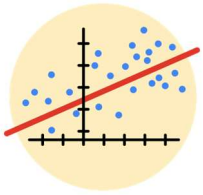
Course Five

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

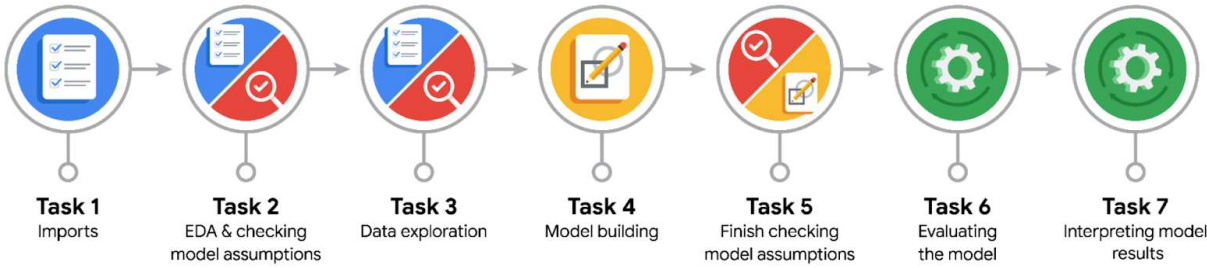
Relevant Interview Questions

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 external stakeholders for this project are the New York City Taxi and Limousine Commission (NYC TLC). They are interested in developing a robust method that can accurately predict taxi fare costs based on a variety of factors such as distance, duration, time of day, and traffic conditions.

- What are you trying to solve or accomplish?

The primary objective is to construct a multiple linear regression model that can accurately predict the fare of a taxi ride. This model will serve as a valuable tool for both the NYC TLC and taxi riders, providing a reliable estimate of fare costs before the journey begins.

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

Upon initial exploration of the data, several significant outliers were identified within the trip_distance, fare_amount, tip_amount, and total_amount columns. These outliers warrant further investigation to understand their origin and impact on the overall data. Additionally, the pickup and dropoff columns are not in datetime format, necessitating conversion for easier manipulation and analysis. As for missing values, the dataset appears to be complete. However, if any were to be discovered, they would need to be

carefully examined and addressed using appropriate methods such as imputation or removal.

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

During this stage, descriptive statistics were extensively used to summarize the dataset and identify potential issues such as outliers, missing values, and data formats. This involved the use of various tools and functions such as .describe(), .info(), .isna().sum(), and shape.

PACE: Analyze Stage

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

Exploratory Data Analysis (EDA) serves several crucial purposes before constructing a multiple linear regression model. It ensures that the data is formatted correctly, identifies potential independent variables that could be useful in the model, and addresses any outliers or missing values that could skew the results.

- Do you have any ethical considerations at this stage?

Ethical considerations at this stage include ensuring that the data was acquired in a manner that upholds the independence condition. It's also important to consider whether there are any inherent biases within the data collection process that could impact the results.

PACE: Construct Stage

- Do you notice anything odd?

Upon examination, it was observed that some of the features are highly correlated (multi-collinear) with one another. This multi-collinearity will impact the beta-coefficients of the regression model. However, it will not negatively affect the overall predictive power of the model.

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

The model could potentially be improved by incorporating other variables in the dataset that may have predictive power. Additionally, further feature engineering could enhance the accuracy of the model. Other methods of variable selection, such as backward elimination or forward selection, could be implemented to identify more variables that have a high correlation with the dependent variable.

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

During this stage, additional online documentation and code from prior coursework were invaluable resources. They assisted in setting up the training and test data, building the model, and creating relevant plots and graphs.

PACE: Execute Stage

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

The model yielded an R-Squared value of 0.89, indicating that 89% of the variation in the dependent variable can be explained by the independent variables. Although multi-collinearity exists between the two independent variables, this is not a significant issue as the model is primarily used for prediction rather than interpretation. The actual vs predicted scores are fairly close, suggesting that the model provides reasonable predictions.

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

Based on the models built, it is recommended to proceed with testing the model on additional, real-world data to ensure its reliability and accuracy in diverse scenarios.

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

Interpreting the beta coefficients is crucial as it indicates which variables have the most significant effect on the dependent variable, in this case, the fare amount. Essentially, they help us understand the relationship between the independent and dependent variables.

- What potential recommendations would you make?

One potential recommendation would be to recalculate the mean_duration and mean_distance means using only the training set, rather than the entire dataset. This would help prevent data leakage and potentially improve the model's performance.

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

The model could potentially be improved by recalculating the mean_duration and mean_distance on only the training data and not the entire dataset. This would reduce the chance of data leakage negatively impacting the model. Additionally, there may be other variables in the dataset or engineered features that could provide additional predictive power to the model.

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

One question that could be addressed is: How can we leverage this model to predict fare amounts before a ride has taken place? This could involve integrating the model with real-time traffic and distance data.

- Do you have any ethical considerations at this stage?

At this stage, there are ethical considerations related to potential data leakage caused by using the entire dataset to calculate the mean_duration and mean_distance means. It's also important to ensure that the observations are independent of one another to uphold the independence assumption of the regression model.