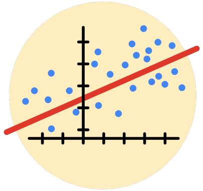
Course Five

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

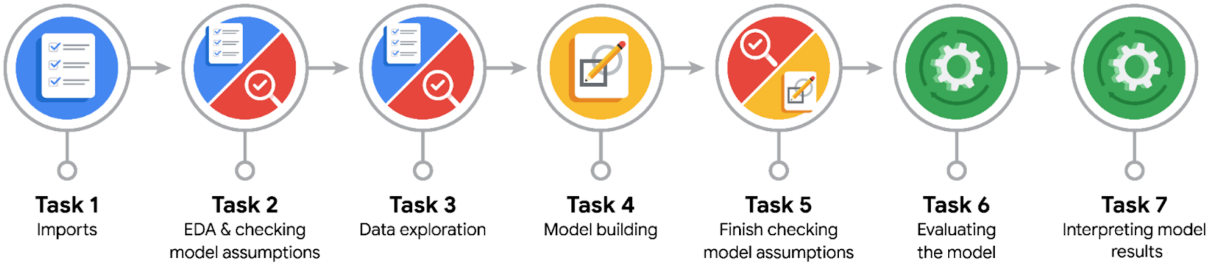
Relevant Interview Questions

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?

Ursula Sayo: Operations Manager

May Santner: Data Analysis Manager

- What are you trying to solve or accomplish?

The main goal of the project is to build a binomial logistic regression model to predict user churn. This model will help Waze understand the factors that influence user retention and take proactive measures to reduce churn.

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

In this stage, your primary focus should be on understanding the dataset. Here are some initial observations you can make:

Check for missing data and handle it if necessary.

Examine summary statistics and distributions of key variables.

Identify any potential outliers.

Explore relationships between variables through visualizations.

Preliminary analysis to identify potential predictors for the logistic regression model.

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

During this stage, you will typically use resources like:

Python programming for data analysis (e.g., Pandas, NumPy).

Data visualization libraries (e.g., Matplotlib, Seaborn).

Jupyter notebooks for data exploration and analysis.

Statistical knowledge and domain expertise.

PACE: Analyze Stage

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

Exploratory Data Analysis (EDA) before constructing a multiple logistic regression model serves several purposes:

Identifying relationships: EDA helps you understand how variables are related to each other and to the target variable (user churn in this case).

Outlier detection: EDA helps in spotting unusual data points that might affect the model.

Feature selection: EDA can guide the selection of relevant features for the logistic regression model.

Data preparation: EDA often reveals data preprocessing tasks, such as handling missing values or encoding categorical variables.

- Do you have any ethical considerations at this stage?

In this stage, you should consider ethical aspects related to data usage and model deployment. Some key ethical considerations include:

Privacy: Ensure that user data is handled with respect to privacy regulations.

Fairness: Ensure that the model doesn't discriminate against any user group based on sensitive attributes (e.g., race or gender).

Transparency: Clearly document data sources, data processing steps, and model decisions for transparency.

Accountability: Define who is responsible for model outcomes and how to address any issues.

PACE: Construct Stage

- Do you notice anything odd?

The log-odd (logit) of the model tends towards -4 the more active days a user had.

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

By investigating the data distribution and relationships. This could involve checking for outliers, exploring feature engineering options, or re-evaluating the model's hyperparameters. Additionally, assessing if there are any non-linear relationships or interactions that the model is not capturing could be beneficial.

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

Documentation for libraries and Data.

PACE: Execute Stage

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

The efficacy of a binomial logistic regression model is determined by accuracy, precision, and recall scores; in particular, recall is essential to this model as it shows the number of churned users.

The model has mediocre precision (53% of its positive predictions are correct) but very low recall, with only 9% of churned users identified. This means the model makes a lot of false negative predictions and fails to capture users who will churn.

Activity_days was by far the most important feature in the model. It had a negative correlation with user churn.

In previous EDA, user churn rate increased as the values in km_per_driving_day increased. In the model, distance driven per day was the second-least-important variable.

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

Due to the model results, we recommends using the key insights from this project milestone to guide further exploration.

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

Beta coefficients represent the change in the log-odds of the dependent variable for a one-unit change in the predictor variable. Understanding these coefficients is crucial for interpreting the impact of each predictor on the outcome.

- What potential recommendations would you make?

This model should not be used to make significant business decisions; however, it has valuable insights insofar as it demonstrated a great need for additional data (features) that correlates with user churn, and also a possible need to better define the user profile Waze seeks to target in their aim to increase overall growth by preventing monthly user churn on the app.

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

Additional data (features) that correlates with user churn, and also a possible need to better define the user profile Waze seeks to target in their aim to increase overall growth by preventing monthly user churn

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

This model should not be used to make significant business decisions; however, it has valuable insights insofar as it demonstrated a great need for additional data (features) that correlates with user churn, and also a possible need to better define the user profile Waze seeks to target in their aim to increase overall growth by preventing monthly user churn on the app.

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

Better describe the targeted user profile.

- Do you have any ethical considerations at this stage?

This model could be used to make erroneous significant business decisions.