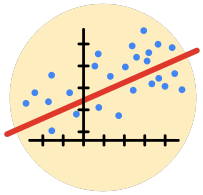
Course Five

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

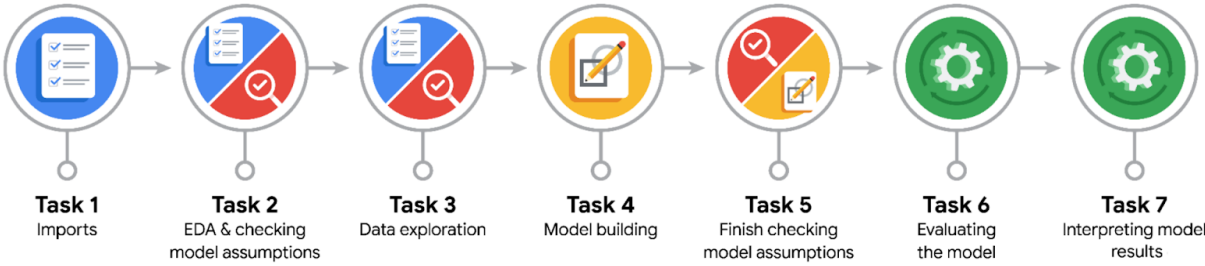
Relevant Interview Questions

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 Waze data team and cross-functional team members.

- What are you trying to solve or accomplish?

The goal of this project is to create the binomial logistic regression model to predict the probability that the users will be churned or retained based on the selected variables.

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

One thing is about missing values of one of the columns, which we might address later.

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

Using Jupyter Notebook to complete the coding work of building the model.

PACE: Analyze Stage

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

Constructing the EDA process is important before building the model. We have to get a general understanding of the whole dataset, what each variable represents, how many rows, columns, etc. We then perform some descriptive statistics, cleaning, transforming, and calculating the variables to have a reliable dataset to work with. Afterward, we need to check the model assumptions of the model we want to build to determine if it all meets the requirements.

- Do you have any ethical considerations at this stage?

Some variables have a high variation. It might indicate the process of collecting data showing values are greater than normal.

PACE: Construct Stage

- Do you notice anything odd?

The model is pretty bad since the majority of negative values appear the most than the positive. Moreover, the false negatives are pretty high, which is the reason we might not choose to use the model.

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

I might recommend carefully choosing the predictors that might have a significant impact on the dependent variable

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

PACE: Execute Stage

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

The results of the model contain the majority of negative values over the positive, which indicates the effectiveness of this model. The false negative values are pretty high (580) that the model predicts as retained, but actually churned users.

The activity days variable has the most impact on the target variable, which shows a negative correlation (the same as the driving days variable). In particular, if we increase one unit of activity days, we might expect the odds of the churned rate to decrease by 10%.

Besides the activity days variable, other variables don't show much effect on the target variable. From the score results (decent precision, but low recall), we might not implement this model in business decisions.

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

Since the result of the model is ineffective (although there are variables that have an impact on determining the churned or retained users rate), we might not want to implement the model in business decisions.

Our Waze data team might have to discuss more with the model results we've made. We might address the issues in the model's performance to further examine the reasons. Perhaps we might need to perform another model to determine if it's more effective than the previous one.

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

It's important to interpret the beta coefficients to determine which variable(s) have significant impact to the target variable. Also, how much it increases and decreases, based on the coefficients.

- What potential recommendations would you make?

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

It could be improved. For instance, we might carefully choose the variables that show a significant impact on the target variable. Additionally, it might lead to overfitting if we include so many variables. Also, scaling the variables is essential to have the same scale for all which leads to better performance of the model.

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

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

We might implement additional features that can better impact the target variable. And like I said before scaling features are important to get a better performance of the model

- Do you have any ethical considerations at this stage?