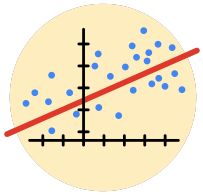
Course Five

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
- [x] ~~Build a multiple linear regression model~~
- [x] ~~Evaluate the model~~
- [ ] Create an executive summary for team members

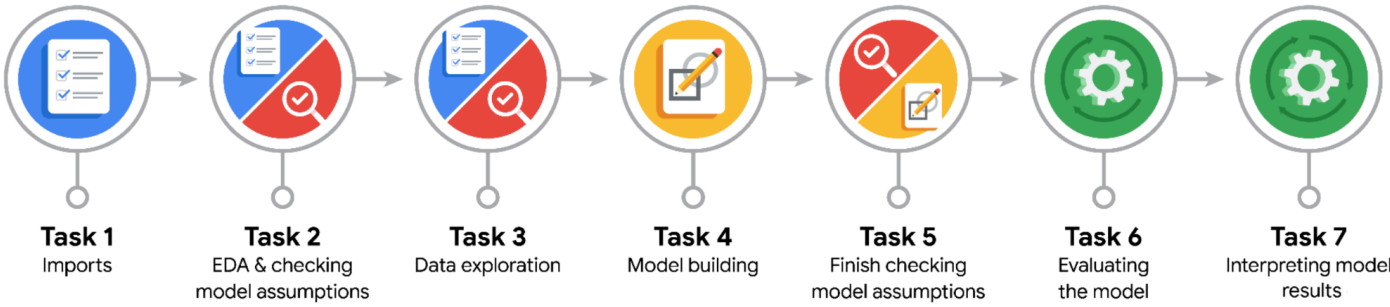
Relevant Interview Questions

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?

- Maika Abadi, Operations Lead - Rosie Mae Bradshaw, Data Science Manager

- What are you trying to solve or accomplish?

We are trying to build a linear regression model on verified status. We need to know what variables impact verified status and if we can predict verified status through linear regression model.

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

We observe no severe multicollinearity with the data. We observe multivariate normality, and assume independent variables. Homoscedasticity is observed.

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

I find myself going back to videos, notes, and documentation of libraries such as numpy, sklearn, train_test_split, etc.

PACE: Analyze Stage

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

The purpose of EDA is to find and analyze the most relevant variables before constructing a linear regression model. This step serves to prevent as many unnecessary variables as possible from making it into the final model.

- Do you have any ethical considerations in this stage?

We consider that the content of the captions may need consideration. Some opinions may be more controversial than others which may contribute to some biases in the model.

PACE: Construct Stage

- Do you notice anything odd?

Yes, we notice outliers associated with comment_count and like_count

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

We will remove outliers.

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

Seaborn's histplot and boxplot applied to all the variables in order to locate the outliers.

PACE: Execute Stage

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

Video duration in seconds, claim status opinion, and video share count were the three most impactful, positively correlated variables in the model.

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

I recommend that we conduct research on how we can increase video duration in seconds without affecting churn!

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

It's important to interpret beta coefficients because they serve as predictor variables for our linear regression model. We can base our model improvements on the most impactful variables.

- What potential recommendations would you make?

I recommend we study the three most impactful beta coefficients in order to better understand the correlation between verified status and

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

Our model could be improved if we can find more beta coefficients that can increase accuracy of our predictive model. We can also use more complex models or even a neural network!

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

Based on the models, it would be beneficial for the organization to research video length in seconds if we want to more accurately predict verified status.

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

I can flip the model around and predict claim status as the dependent variable to see if we can gather further information.

- Do you have any ethical considerations at this stage?

I'm considering the text may have certain weights associated with them. For example, some videos may have more inflammatory statements than others which can affect model results or even render the model useless!