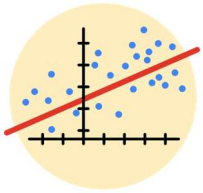
Course Five

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

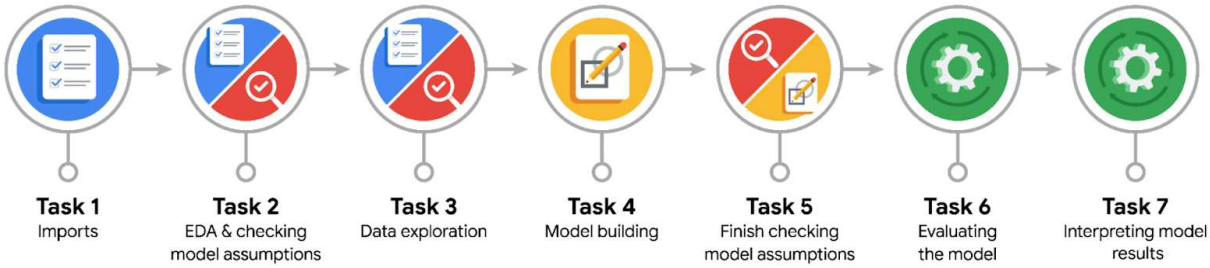
Relevant Interview Questions

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?

Data team roles:

- Willow Jaffey- Data Science Lead
- Rosie Mae Bradshaw- Data Science Manager
- Orion Rainier- Data Scientist

Cross-functional team members:

- Mary Joanna Rodgers- Project Management Officer
- Margery Adebowale- Finance Lead, Americas
- Maika Abadi- Operations Lead

- What are you trying to solve or accomplish?

This project aims to develop a predictive model capable of classifying videos as containing either opinions or claims, based on a range of features.

This milestone focuses on building a logistic regression model to identify correlations between various features and the dependent variable, verified_status. Prior analysis suggests a potential relationship between verified_status and claim_status.

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

1. The count variables exhibit significant right skew.
2. Seven columns contain 298 missing values.
3. author_ban_status and verified_status are significantly imbalanced.
4. The dependent variable, claim_status, is balanced.

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

This analysis references previous material and imports the necessary libraries, while also adhering to the Project Proposal as a general guideline.

PACE: Analyze Stage

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

EDA allows for thorough data exploration. This includes handling outliers, missing data, standardization/normalization (if necessary), imbalanced variables, and understanding variable distributions.

- Do you have any ethical considerations at this stage?

At this point, my only ethical concern is the apparent prioritization of videos containing potentially problematic content on TikTok. The most readily available videos for users often seem to be those with highly engaging, but not necessarily positive, content. If many of the videos driving increased engagement are from unverified users and banned authors, it suggests that this content may be spreading misinformation or other harmful material that violates the platform's terms of service.

PACE: Construct Stage

- Do you notice anything odd?

Many variables pertaining to video metrics, such as shares, downloads, and comments, indicate a significant right-skew. This, however, may be due to the nature of videos that achieve these high levels of engagement. These videos are likely less frequent and appear as outliers in the data.

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

The model's performance was acceptable, but it requires improvement before it can be used in production. I recommend engineering new variables from existing ones to explore whether they can provide greater predictive power.

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

At this stage, referencing previous projects has been invaluable for recalling necessary steps, packages, and methodologies for performing exploratory data analysis (EDA) and model building.

PACE: Execute Stage

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

Many of the variables have decent predictive power, but the overall model, with an accuracy score of only 65%, provides merely acceptable predictive capability. One feature, video_duration_sec, indicated that for each additional second of a video's duration, the log-odds of the user having verified status increase by 0.001.

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

The model's performance is sufficient for use in future model development and for gaining insights into the relationship between the features and the dependent variable, verified_status. However, it requires further improvement before it can be used in

production for business-level predictions and decisions. A revised model iteration with adjustments to the predictor variables is recommended.

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

The beta coefficients reveal the relationship between the variables and the dependent variable, as well as their predictive power. A positive coefficient indicates that an increase in the predictor variable corresponds to an increase in the log-odds of the outcome. Conversely, a negative coefficient indicates that an increase in the predictor variable corresponds to a decrease in the log-odds of the outcome.

- What potential recommendations would you make?

A recommended next step is to create another model iteration, incorporating newly engineered variables (if feasible), to assess whether they enhance predictive power. Additionally, exploring the impact of removing correlated predictor variables on model performance is worthwhile.

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

Model performance could be improved by investigating whether removing correlated predictor variables enhances predictive power and overall scores.

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

I would like to conduct a more thorough analysis of the video_transcription_text variable to identify potential patterns that could indicate whether a video expresses a "claim" or an "opinion." This analysis may also provide insights into whether the video contains harmful content, allowing for appropriate action.