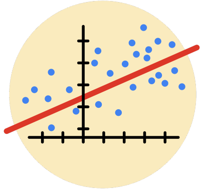
Course Five

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 logistic regression model~~
- ☑ ~~Evaluate the model~~
- □ Create an executive summary for team members

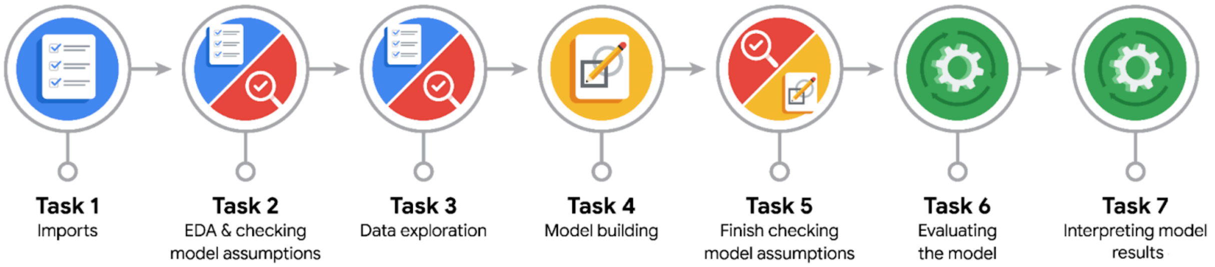
Relevant Interview Questions

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?

Maika Abadi- Operations Lead

Mary Joanna Rodgers- Project Management Officer

- What are you trying to solve or accomplish?

How different variables are associated with whether a user is verified?

How to predict ‘verified_status’?

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

- There are 19382 videos in the database and 12 features. All of the videos are unique and 298 have null almost all their information (including verified_status, which is the dependent variable). Then those videos must be removed.
- ‘video_duration_sec’ is an ‘int’ type, so for the sake of performing logistic regressions, it will need to be converted into a ‘float’ type.
- We have four ‘object’ type variables, so if they wil be include, then they will need to be converted into numerical variables, e.g. using dummy variables.

- ‘verified_status’ is not balanced. Just 6.29 % of the videos are from a verified account in this data set.

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

I used resample to change the imbalance ‘verified_status’; histogram for understanding transcription length; boxplots to check for outliers and differences by verified status; heatmaps to graph correlations among numerical variables; length and mean for transcription texts; loc to change outliers values.

PACE: Analyze Stage

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

To have a general understanding of the data set; to check requirements for a model; identify null values; use correct types in columns for making them work for models; handle outliers before models, etc.

- Do you have any ethical considerations at this stage?

All the numeric the metrics, but length of a video’s transcription and video’s duration, are strongly correlated to each other. Specially number of comments and downloads (0.91); number of likes with number of views (0.86), downloads (0.87) and shares (0.89); and downloads and shares (0.80) . Those correlations violate one of the assumptions for performing a logistic regression.

Also, if this study could lead to ban or increase prejudice from users and their content depending on whether the account is verified. Furthermore, anonymity, which is a main difference between not verified and verified users, could also be affected depending on the results of this analysis and how they could be used.

Could resampling the number of videos from verified users to bias the study?

PACE: Construct Stage

- Do you notice anything odd?

- y_train and y_test aren't in (#,) array form in order to be used in a logistic regression, so they need to be converted with .ravel() or .iloc[].

- As said before, the dependent variable 'verified status' must be one-hot encoded to use it in a logistic regression.

- Due to multicollinearity, the independent variables chosen for the logistic regression were: 'claim_status', 'author_ban_status', 'video_comment_count', 'video_share_count', 'video_view_count', 'video_duration_sec'. Number of video's likes and downloads were excluded to remove high collinearity.

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

When analising p-values at 5 % significance level, the only not significant estimated coefficient was for number of comments in a video. Then improving the model must include removing this variable from the study.

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

I split the data in train and test sets, converted variables to 1 and 0 and used logistic regression.

PACE: Execute Stage

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

- There are more false positives than true negatives. Specifically, 54 % of the actual negatives were wrongly identified.

- Although model's Recall is 84 %, which means about 8 of 10 videos from verified users were correctly identified as coming from verified accounts, its Precision is just 61 %, that is, only around 6 of 10 videos identified as done by verified users were correctly predicted.

- Even worse, 56 % of all videos from not verified users were incorrectly predicted as coming from verified users, which affects model's reliability.

- For each additional second of a video's duration, the estimated odds of coming from a verified user increase in 0,85 %.

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

Because they explain how the dependent variable's odds vary as percentages whilst a unit of each independent variable increases one unit.

- What potential recommendations would you make?

- Remove number of comments in a video to see if that makes the logistic regression.

- If possible, to use a model not as sensitive to collinearity as a logistic regression.

- Create models for other variables that can also be used to analyse users, like author ban status.

- Length of video transcription's text distribution is not different enough to be used as predictor of verified status. As an option, further research on identifying key words in those texts to identify, not only verified users, but also if they post claim or opinion videos.

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

When analising p-values at 5 % significance level, the only not significant estimated coefficient was for number of comments in a video with a p-value at 58.1 %. Then improving the model must include removing this variable from the study.

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

I consider this model should not be used yet until getting better results, specifically on detecting not verified users at a high percentage, since it could lead to fraudulent activities as impersonation.

Further research and using other kind of models could be a possibility to improve results. Increasing the number of samples for verified users could help to have more data to train the chosen model.

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

Most common reasons for an user to become verified in TikTok?

Could further research on identifying key words in those texts help to find, not only verified users, but also if they post claim or opinion videos?

How identifying whether verified user status is going to be used? What's the purpose of it?

- Do you have any ethical considerations at this stage?

The model isn't good enough to correctly identify not verified users, which could lead to fraudulent activities as impersonation.

How this could affect the main purpose of this claim classification analysis.