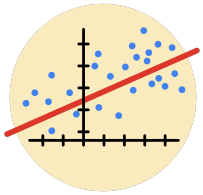
Course Five

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

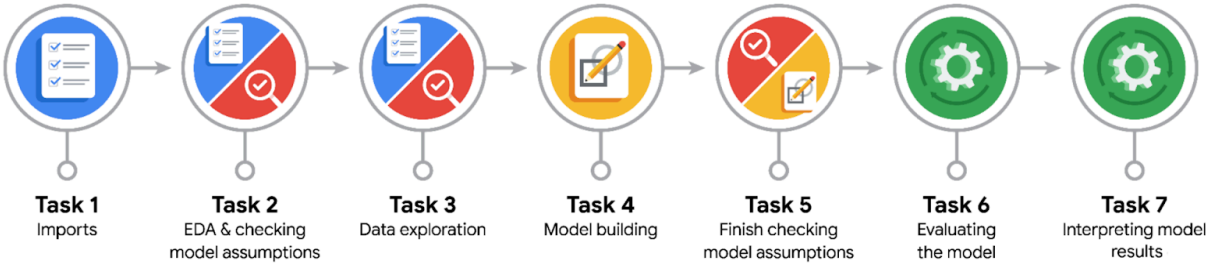
Relevant Interview Questions

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 stakeholders are the management and moderator staff of TikTok. There are no stakeholders outside the company. This new ML tool we are aiming to build will improve our capabilities to manage our product as it is used and consumed.

- What are you trying to solve or accomplish?

We are solving the issue of moderators having to review every single post made by the Tiktok users. The classifying ML model will be able to better categorize the user submissions by ‘claim’ and ‘opinion’ based on the other factors of each post and user account connected to it.

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

There were many outliers due to the viralness of many submissions. The number of verified accounts is much smaller than unverified. Categorical variables need to be shifted to numerical so as data analysis can be completed.

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

I was using my knowledge of pandas, matplotlib, seaborn, and scikit-learn to observe, clean, modify, and perform manipulations for smoother, cleaner, and more interpretable code and work further in the analysis.

PACE: Analyze Stage

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

The purpose of EDA at this time of the project is to better understand the shape and distribution of the data as well as to confirm data linearity, normality, and no colinearity

- Do you have any ethical considerations at this stage?

We are taking submissions from users and reviewing and moderating it. A major concern would be the with the development of this model, if a bias is trained into it, there could be inadvertent and unwarranted censoring of creative content.

PACE: Construct Stage

- Do you notice anything odd?

My first model build was extremely inaccurate, when I was conservative with the number of features I included. With the inclusion of video duration the model started acting much better. It seems like a very important, pivotal feature to include.

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

There is the threat of collinearity with many of the features, especially with video likes. However, this is one of the more correlated features relative to user verification. I would like to try to include it and exclude the other collinear features. This could potentially tighten up the efficacy of the model.

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

I used the train_test_split and LogisticRegression from scikit-learn. Pandas is an important tool for the selection moving of the data set.

PACE: Execute Stage

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

Many of the features have definite to strong correlation, which may cause collinearity conflicts with the model. The most correlative feature is video_like_count with correlation values of as high as 0.85 with other features. It has been left out of the model's training.

Video duration has a beta of 0.0079, meaning for each second there is 0.0079 increase in log-odds probability of the user being verified. Claim status has a beta of 0.00036, meaning for a value of 1 (a claim) there is a 0.00036 increase in log-odds probability of the user being verified. The video download count's beta is -0.00023, meaning for each additional download there is a -0.00023 decrease in log-odds probability of the user being verified.

Precision: 59%,Max recall: 88%, Max f1-score: 71%, Max accuracy: 64%, the weighted and macro averages are all down in the 60s. The model did perform; however, it does feel like there is room for improvement, since there was a lot of misclassification.

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

With this regression model, we have some insight into the data relationship of the data set to user verification. With an understanding of this landscape we can state that for each second there is 0.0079 increase in log-odds probability of the user being verified, and if the person is verified, they are more likely to make an 'opinion' than a 'claim'. Therefore, we can more confidently move forward with the building of the machine learning model with the current data set pointing towards that claim status can be discerned.

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

It is important to interpret betas, since this gives insight into what each step of a beta will impact. This gives the interpreter the understanding generally of the model function without having to run the whole system.

- What potential recommendations would you make?

I would recommend some further investigation of feature use choices, maybe a rebuilding of the regression model including the video likes. I would also recommend to move forward with the machine learning model exploration and development.

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

I am sure it can be improved. This is due to certain features being eliminated from the model due to concerns of collinearity. A balance could likely be found with a better model outcome, while still avoiding collinearity.

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

I would propose moving forward with the machine learning model development. Additionally, I would encourage more investment in the verification team so as to bring about more accountability within the user community.

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

We can use verification for a feature and look at the logistic model for claims and opinion, while only changing two lines of code. We could also use a different logistic regression model classifier and see if this brings out a better accuracy.

- Do you have any ethical considerations at this stage?

With a low accuracy, this model is has the potential skewing the output data due to introduced bias from data that it was trained on. I would be worried to make any definite decisions while having only a 64% accuracy.

We did upsample the number of verified data points. This may bring some insight to the fact that there were so many False positives, while so few false negatives.