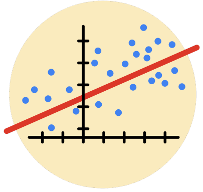
# Course Five

## 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

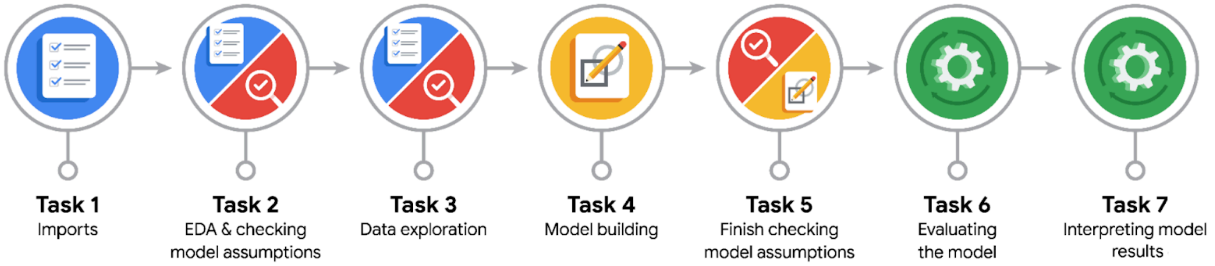
## Relevant Interview Questions

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?

Some of the external stakeholders for this project are: TikTok's data team (Willow Jaffey- Data Science Lead, Rosie Mae Bradshaw- Data Science Manager, and Orion Rainier- Data Scientist); cross-departmental stakeholders within TikTok (Mary Joanna Rodgers- Project Management Officer, Margery Adebowale- Finance Lead, Americas, and Maika Abadi- Operations Lead); TikTok's Operations; and the leadership team at TikTok.

- What are you trying to solve or accomplish?

In this part of the project, the data team needs to conduct regression analysis on the provided data set and build a logistic regression model that predicts verified_status.

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

The claim_status and author_ban_status features are each of data type 'object'. In order to work with the implementations of logistic models through sklearn, these categorical features will need to be made numeric. One way to do this is through one-hot encoding.

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

As a data professional in the Plan stage of the PACE workflow, I find myself leveraging a variety of resources to effectively conceptualize and strategize my project. Here are some of the most valuable ones: for Domain-Specific Knowledge, I reviewed Data dictionaries or metadata because these provide information about the structure and meaning of our data; for Project Management Tools, I utilized Jupyter Notebook since it is often used as a living document to outline project goals, data sources, and initial analysis plans; and for Data Exploration Tools, I considered Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn.

## PACE: Analyze Stage

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

The purposes of EDA before constructing a logistic regression model are: 1) to identify data anomalies such as outliers and class imbalance that might affect the modeling;; 2) to verify model assumptions such as no severe multicollinearity.

- Do you have any ethical considerations at this stage?

Any ethical considerations at this stage I may have are: for Data Privacy and Security, it is imperative to secure data storage and access, implementing measures to prevent unauthorized access; for Data Bias, we ensure that the data used is representative and free from biases; for Data Misinterpretation:, we consider the context, avoid overfitting, and communicate our findings clearly; and for Data Misuse, we consider the potential consequences of your analysis.

## PACE: Construct Stage

- Do you notice anything odd?

There does not seem to be any duplicates; However, the number of outliers for video_like_count and video_comment_count were 3468 and 3882 respectively. Also, approximately 94% of the dataset represents videos posted by unverified accounts and 6% represents videos posted by verified accounts. So the outcome variable is not very balanced. Finally, the dataset has a few strongly correlated variables, which might lead to multicollinearity issues when fitting a logistic regression model. We decided to drop `video_like_count` and  from the model building because the heatmap showed that the following pair of variables were strongly correlated: video_view_count and video_like_count (0.87 correlation coefficient) and video_comment_count and video_like_count (0.85 correlation coefficient)

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

Performing class balancing using upscaling optimized the model since the data was imbalanced 93.71/6.29 (not verified/verified). Plus, given the large number of outliers (3468 and 3882), deletion might not be ideal. We chose a Handling Outlier Strategy such as Capping/Trimming where we replaced outliers with a threshold value.

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

Some resources I find myself using as I complete this stage are: for data manipulation, importing and utilizing packages such as numpy and pandas; for data visualization, importing and utilizing packages such as matplotlib and seaborn; and for data preprocessing and data modeling, importing and utilizing packages such as sklearn.

## PACE: Execute Stage

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

According to the classification report, the logistic regression model achieved a precision of 63%, a recall of 83.9%, and f1-score of 72%. It achieved an accuracy of 67% as well. These scores are taken from the "not verified" row of the output because it is the target class that we are most interested in predicting. In addition, based on the estimated model coefficients from the logistic regression, the video feature of 'video_duration_sec' compared to the other video features has a relatively larger estimated coefficient in the model. Each additional second of the video is associated with 0.005 increase in the log-odds of the user having a verified status.

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

The logistic regression model provides a framework for predicting binary outcomes such as the verified_status of a user submission.

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

In the Execute stage of the PACE workflow, interpreting the beta coefficients from a binomial logistic model is crucial for understanding the relationship between the predictor variables and the binary outcome variable. These coefficients represent the log odds ratio, which is the change in the log odds of the outcome for a one-unit increase in the predictor variable. Interpreting beta coefficients, we understand the direction of relationships, quantify the strength of relationships, and identify important predictors. Building predictive models, the beta coefficients can help identify the most important predictors in the model, and during feature selection coefficients with small or non-significant values may be considered for removal to simplify the model. Thus, by interpreting the beta coefficients, data professionals can gain valuable insights into the factors influencing the outcome variable and build more accurate and interpretable models.

- What potential recommendations would you make?

The results of a logistic regression model can be used to express the relationship between variables. It provides a generally strong and reliable fare prediction that can be used in downstream modeling efforts. Thus, our data team recommends using the key insights from this project milestone to guide further exploration.

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

I think our model could be improved given the logistic regression model had not great, but acceptable predictive power. The estimated model coefficients from the logistic regression have small estimated coefficients. It appears that using machine learning may improve that and, as a result, our model.

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

The data team recommends constructing a classification model that will predict the status of claims made by users. With the helpful context we discovered around user behavior, we may better construct and analyze the results of that classification model.

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

Seeing the confusion matrix, the model correctly predicts relatively more true positives than the other quadrants. It indicates that Type I errors are more likely to occur. Thus since a perfect model yields all true negatives and true positives (and no false negatives or false positives), it seems to me that our goal should be to find a model using machine learning that predicts relatively more true negatives and true positives.

- Do you have any ethical considerations at this stage?

Any ethical considerations at this stage I may have are: for Model Fairness and Bias, we both ensure that our models are free from biases that could lead to unfair or discriminatory outcomes and continuously monitor model performance to identify and address biases; for Ethical Use of Results, we both ensure that the results of our analysis are used ethically and for its intended purposes and be transparent about the limitations and potential biases of our models; and for Accountability, we maintain records of the entire execution process for accountability and transparency.