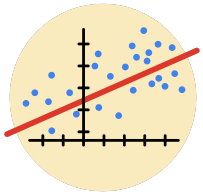
Course Five

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~~

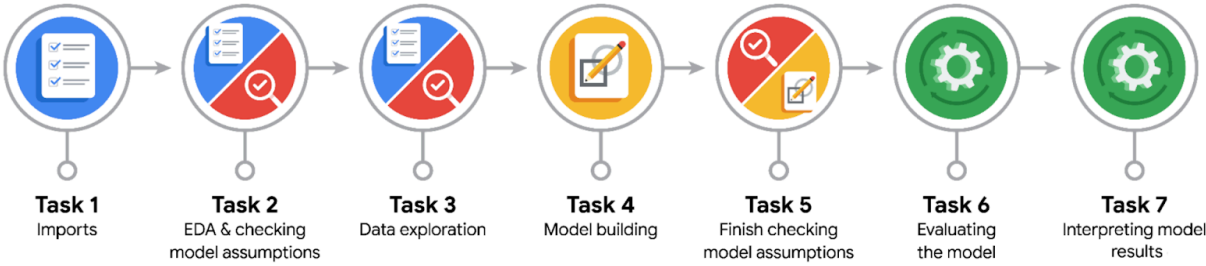
Relevant Interview Questions

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?

Project Management Officer (Mary Joanna Rodgers) – Oversees the project's progress and ensures alignment with organizational goals. Finance Lead, Americas (Margery Adebowale) – Evaluates the financial impact of the claims classification model, including cost-effectiveness and resource allocation. Operations Lead (Maika Abadi) – Assesses how the model will integrate into TikTok's operational workflows and its potential impact on content moderation and platform management.

- What are you trying to solve or accomplish?

The primary goal is to address the backlog of user reports on TikTok content, particularly videos, by developing a classification model. Initially, we aim to build a logistic regression model that classifies video content as either a claim or an opinion. In this project phase, we focused on classifying users as verified or unverified based on various features. Previous analyses suggested a connection between verification status and whether content is a claim or an opinion. By understanding the factors influencing verification, we can determine which variables directly or indirectly predict it. We first classify verification status to ensure it is a meaningful, independent feature before incorporating it into the final model. This structured approach prevents bias,

redundancy, and data leakage, ultimately leading to a more accurate and interpretable model for distinguishing opinions from claims.

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

Some fields had incompatible data types for logistic regression and other tasks, requiring conversion. The dataset was mostly complete, with a few missing values that were addressed. Outliers were present and handled using Winsorizing. Overall, the dataset provided valuable insights.

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

Key Python libraries played a crucial role in this phase. Specifically, pandas was used for data manipulation, including importing, processing, and structuring the dataset. The dataset itself served as a valuable resource for analysis.

PACE: Analyze Stage

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

Purposes of EDA Before Constructing a Logistic Regression Model: Exploratory Data Analysis (EDA) ensures data quality and improves model reliability by: Validating Assumptions – Checking for multicollinearity, ensuring independent observations, and assessing the relationship between predictors and the outcome. Understanding Data Structure – Exploring variable distributions, summary statistics, and trends to identify key patterns. Detecting Data Issues – Identifying missing values, outliers, and inconsistencies that could affect model performance. Feature Selection & Engineering – Selecting relevant predictors, transforming features, and encoding categorical variables. Addressing Class Imbalance – Evaluating target class distribution to determine if resampling is needed. Thorough EDA helps prevent potential issues, leading to a more accurate and interpretable model.

- Do you have any ethical considerations at this stage?

A key ethical consideration is ensuring that the dataset is anonymized and does not contain personally identifiable information (PII). This dataset meets that standard, as it does not include sensitive user details. Additionally, it is important to check for biases in the data that could lead to unfair predictions, ensuring transparency and fairness in model development.

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

For this phase, I relied on several essential Python libraries to perform data analysis, preprocessing, and modeling:

Data Manipulation: Used numpy and pandas for handling and processing structured data.

Data Visualization: Utilized seaborn and matplotlib.pyplot to explore relationships between variables and identify patterns.

Data Preprocessing: Applied resample to address class imbalance and ensure fair model training.

These tools were crucial in refining the dataset and preparing it for effective modeling.

PACE: Construct Stage

- Do you notice anything odd?

Yes, I noticed some missing values, which were addressed during preprocessing. Additionally, the model tends to classify a significant portion of unverified users as verified, leading to a high number of false positives (Type 1 error). This indicates the need for further tuning, such as adjusting the decision threshold or improving feature selection, to enhance model accuracy and reduce misclassification.

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

Yes, the model can be improved by:

Adjusting the decision threshold to make positive predictions more selective, reducing false positives.

Enhancing feature selection to remove irrelevant or noisy variables and improve predictive power.

Balancing the dataset if it's imbalanced, ensuring the model doesn't favor one class over the other.

Using alternative evaluation metrics like the ROC curve to better understand the trade-off between precision and recall, optimizing model performance.

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

Data Manipulation:

numpy and pandas for handling and processing structured data.

Data Visualization:

seaborn and matplotlib.pyplot for exploratory data analysis and visualizing relationships between variables.

Data Preprocessing:

OneHotEncoder for encoding categorical variables.

CountVectorizer and TfidfVectorizer for text feature extraction.

resample to handle class imbalance.

Modeling & Evaluation:

train_test_split to divide the data into training and testing sets.

LogisticRegression for classification.

classification_report, confusion_matrix, and ConfusionMatrixDisplay to assess model performance.

PACE: Execute Stage

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

Claim Status Drives Verification – Users with claim_status_opinion were significantly more likely to be verified (Odds Ratio: 5.49).

Banned & Under-Review Users Less Likely – Banned users had a lower verification probability (Odds Ratio: 0.64), with a smaller effect for under-review users (0.91).

Minimal Impact of Video Metrics – Engagement features (views, shares, downloads, comments) had little influence on verification.

Model Performance – Accuracy: 67%, strong recall for unverified users (82%), but lower precision for verified ones (74%).

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

Prioritize Claim Status for Verification – Automate verification based on claim_status_opinion to streamline the process.

Review Ban Policies – Reassess banned and under-review users for potential verification after a probation period.

Refine Verification Criteria – Since engagement metrics have little impact, consider factors like audience interaction quality.

Improve Model Precision – Incorporate additional behavioral data and explore alternative models for better accuracy.

Ongoing Model Updates – Regularly refine the model and test different verification criteria to enhance performance.

- What potential recommendations would you make?

Automate Verification – Use claim_status_opinion to streamline verification and reduce manual effort.

Reevaluate Ban Policies – Introduce reassessment mechanisms for banned or under-review users.

Expand Verification Criteria – Incorporate factors like follower growth and audience interaction beyond engagement metrics.

Enhance Model Performance – Explore advanced ML models (e.g., random forests, boosting) for better accuracy.

Tiered Verification System – Implement multiple verification levels for a more nuanced approach.

Regular Updates & Monitoring – Continuously retrain models with new data to adapt to changing trends.

Increase Transparency – Provide users with explanations for verification outcomes to build trust.

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

Yes, the model can be improved. Adjusting the decision threshold can make positive predictions more selective, reducing misclassifications. Improving feature selection can help remove noise and enhance model performance. If the dataset is imbalanced,

rebalancing techniques can improve predictions for minority classes. Additionally, using alternative evaluation metrics like the ROC curve can help find the optimal trade-off between precision and recall.

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

Refine Verification Policies – Since claim_status_opinion is the strongest predictor of verification, organizations should prioritize it in their verification process to improve efficiency and consistency.

Enhance Moderation Practices – Users with a ban or under review status are less likely to be verified. Organizations should review ban policies and consider a probationary system for re-evaluation.

Reassess Engagement-Based Criteria – Video engagement metrics have minimal impact on verification. If influence matters, consider additional behavioral indicators such as audience interaction quality and historical activity.

Improve Model Accuracy – Misclassification of verified users suggests refining the model. Exploring advanced modeling techniques or incorporating additional user behavior data could enhance precision.

Continuously Monitor & Update the Model – Regularly updating verification criteria and testing different approaches (e.g., A/B testing) will help maintain fairness and accuracy as user behavior evolves.

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

What additional factors influence user verification? – Are there other user behaviors or metadata that could improve the model's predictive power?

How does verification status impact user engagement? – Do verified users have higher engagement levels (e.g., views, shares, comments) compared to unverified users?

Are there patterns in misclassified users? – What characteristics do incorrectly predicted verified or unverified users share, and how can the model be improved to address these cases?

Does verification status correlate with content performance? – Are verified users' videos more likely to go viral, and how does verification impact content visibility?

Can alternative models improve accuracy? – Would decision trees, ensemble methods, or deep learning approaches yield better results than logistic regression?

How effective are current ban policies? – Do banned or under-review users show engagement patterns that indicate potential for future verification?

What factors contribute to long-term verification retention? – Do verified users maintain consistent activity and adherence to platform guidelines over time?

- Do you have any ethical considerations at this stage?

Ethical considerations at this stage include ensuring fairness and accuracy in the verification process. Misclassifying verified users as unverified—or vice versa—can lead to credibility issues, user distrust, and potential harm to content creators.

To address this:

Minimize Bias: Ensure the model does not disproportionately misclassify certain groups of users based on biases in the data.

Transparency: Clearly communicate the verification criteria and allow users to appeal misclassifications.

Privacy Protection: Avoid using sensitive or personally identifiable information for verification decisions.

Fair Content Moderation: Ensure banned or under-review users are treated fairly, with clear policies on re-evaluation.

Regular audits, human oversight, and continuous model improvements should be implemented to maintain ethical standards.