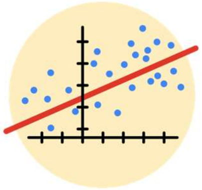
Course Five

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

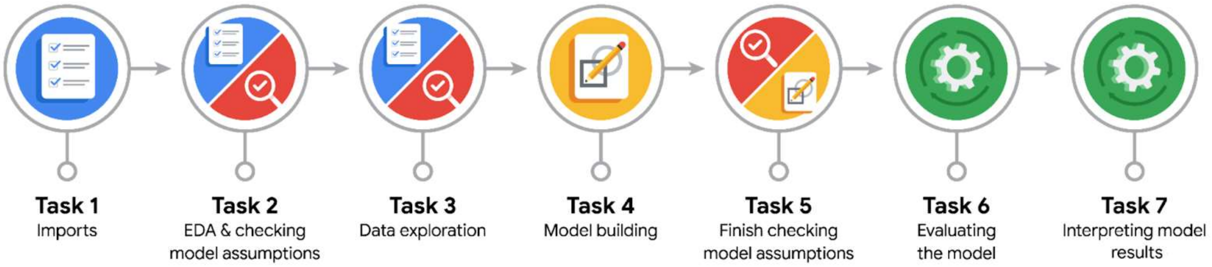
Relevant Interview Questions

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 Waze team, particularly Ursula Sayo, is the key stakeholder for this project, as they have commissioned this churn project to be completed.

- What are you trying to solve or accomplish?

The overall project goal is to identify the cause of user churn for the Waze app and develop methods to eliminate it. Additionally, identifying users at risk of churning will enable the team to target them with retention strategies. This specific stage of the project aims to develop a binomial logistic regression model utilizing various variables to predict which factors are most correlated with whether a user will churn or not.

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

Initially, an imbalance was observed in the target variable 'label.' Depending on the model type, this could negatively impact prediction accuracy. Descriptive statistics revealed potential outliers in the variables 'total_sessions,' 'drives,' 'driven_km_drives,' and 'duration_minutes_drives.'

A new variable, 'professional_driver,' was created to categorize individuals with 60 or more drives and 15 or more driving days in the past month. Comparing churn rates between professional and non-professional drivers, we found that 7.6% of professional drivers churned, while 19.9% of non-professional drivers churned. This suggests that 'professional_driver' could be a valuable predictor in the model.

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

Referring to other projects has been helpful in reminding myself of certain steps and processes. Additionally, descriptive statistics have been useful for gleaning information about the underlying data, such as outliers, missing data, distributions, and more. Checking for imbalance in the target variable has also provided insight into whether corrective methods are needed, depending on the significance of the imbalance. Lastly, engineering the new variable 'professional_driver' and comparing the two groups (professional or not) to churn frequency (retained or churned) suggests it may be a valuable predictor.

PACE: Analyze Stage

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

First and foremost, exploratory data analysis allows you to understand the data, including the variables and their types. Generating descriptive statistics, such as the mean, median, mode, standard deviation, and range, is essential. For categorical variables, frequency counts can be generated. Additionally, EDA provides insight into the distribution of variables in the dataset, such as skewness and potential outliers. Correlation analysis can help identify relationships between numerical variables and detect multicollinearity. For a categorical target variable, check for class imbalance, and for a numerical target, examine its distribution. Visualizations like scatter plots, pair plots, and bar charts can help uncover patterns and relationships. Handle missing values and outliers by deciding whether to impute, remove, or use other methods. Finally, consider feature engineering to create new predictor variables.

- Do you have any ethical considerations at this stage?

I suppose the only ethical considerations I have, are in how the data was collected, as during prior stages it was determined there may be over-represented groups in the data; the drivers in the dataset appear to drive far more than the average driver.

PACE: Construct Stage

- Do you notice anything odd?

Significant outliers were present in many of the numeric variables, particularly related to driving. This suggests that the data may include groups of drivers who drive considerably more than average.

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

The model itself does not perform well, based on the precision and recall scores. More features could be engineered with domain knowledge, or feature elimination could be employed to try different combinations of variables. Additionally, the feature importance values could be scaled to more easily see the most important predictors next to "activity_days."

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

Looking at other notebooks has helped me remember how to complete specific steps of EDA, feature engineering, training and fitting the logistic regression model, and building the necessary visualizations. Many packages have been used during these steps, such as NumPy, Pandas, Matplotlib, Seaborn, and various others.

PACE: Execute Stage

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

The logistic regression model built underperformed when predicting the dependent variable churn. With a precision score of 0.55 and recall score of 0.10, the model did not effectively predict churn rates Waze app users. That being said, ‘activity_days’ was indicated as being the strongest predictor feature in the set of independent variables, and this model, although it cannot be used to make business decisions in its current state, can be used assist in further EDA if needed.

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

Consider acquiring more granular information about user driving, such as geographic data, specific drop-off and pick-up locations, and app user behavior. Additionally, leverage domain knowledge to engineer additional predictor variables that may enhance the model's performance.

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

Beta coefficients enable you to compare the impact of independent variables on the dependent variable. For every one-unit change in an independent variable, there is an x-unit change in the dependent variable.

- What potential recommendations would you make?

At this stage, re-evaluate the model to determine if it can be improved through further feature engineering, acquiring additional data to enhance its predictive power, or employing feature elimination methods to identify optimal feature combinations.

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

While the model could be enhanced through the aforementioned recommendations, the current model's low scores suggest that additional data may be necessary.

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

Acquire more relevant, granular information on user behavior, geographic locations, and unique drop-off/pick-up points. Experiment with different combinations of existing variables to assess potential improvements in the model's predictive power. In its current state, the model can be utilized to facilitate further exploratory data analysis (EDA), as it has already provided insights into the variables currently exhibiting the highest predictive capability.

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

Questions that could be addressed include: Which features exhibited the highest predictive power? Which features displayed the highest collinearity with each other? Which features were most positively or negatively correlated with user churn? Additionally, what is the imbalance in the dependent variable? What are the proportions of true positives, false positives, false negatives, and true negatives?

- Do you have any ethical considerations at this stage?

At this stage, my primary ethical concern lies in the data collection process. The EDA conducted thus far suggests that the data may be biased toward individuals who drive significantly more than the average person. Further investigation into the data acquisition process would be beneficial.