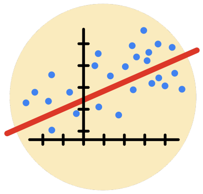
Course Five

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

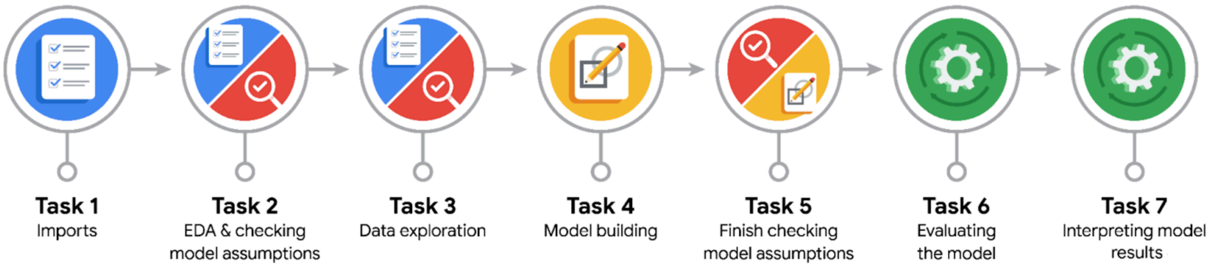
Relevant Interview Questions

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?

Ursula Sayo, Waze's Operations Manager. May Santner, your supervisor. The Waze leadership team.

- What are you trying to solve or accomplish?

Build a binomial logistic regression model to predict user churn. Evaluate the model's performance. Provide insights and recommendations to Waze leadership.

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

Missing values in the `label` column. Potential outliers in `sessions`, `drives`, `total_sessions`, etc. Imbalanced target variable (`label`). High multicollinearity between variables like `sessions` and `drives`.

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

Python libraries (pandas, NumPy, matplotlib, seaborn, scikit-learn). The provided dataset and notebook. Documentation for logistic regression and data analysis.

PACE: Analyze Stage

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

Understand data structure and identify issues (missing values, outliers). Check logistic regression assumptions (multicollinearity). Inspire feature engineering. Assess data quality.

- Do you have any ethical considerations at this stage?

Data privacy. Avoiding biased analysis. Transparency about data limitations.

PACE: Construct Stage

- Do you notice anything odd?

High multicollinearity between `sessions` and `drives`, `activity_days` and `driving_days`. The need to impute outlier values. The need to create dummy variables.

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

Address multicollinearity by dropping redundant variables. Impute outliers. Create binary `device2` variable. Address class imbalance.

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

Scikit-learn for model building and data splitting. Pandas and NumPy for data manipulation. Logistic regression statistical knowledge.

PACE: Execute Stage

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

`activity_days` is the most influential predictor. Model has decent accuracy but low recall for churn. Model performance metrics (precision, recall, f1-score).

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

Focus on user engagement (increase `activity_days`). Target retention efforts for at-risk users. Further investigate the impact of professional drivers. Refine the model to improve recall.

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

Beta coefficients show the direction and strength of the relationship between predictors and the log-odds of churn.

- What potential recommendations would you make?

Proactive churn interventions. Further analysis of churn reasons. Continuous model monitoring.

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

Yes, by: - Addressing class imbalance. - Exploring non-linear relationships. - Collecting more data. - Trying other model types.

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

Allocate resources for churn reduction. Targeted marketing campaigns. Improve user experience.

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

Churn variation across user segments. Root causes of churn. Effectiveness of retention strategies. Predicting churn earlier.

- Do you have any ethical considerations at this stage?

Avoid discrimination. Transparency about model limitations. Responsible use of predictions.