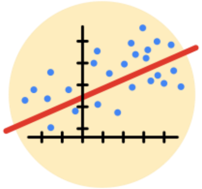
Course Five

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

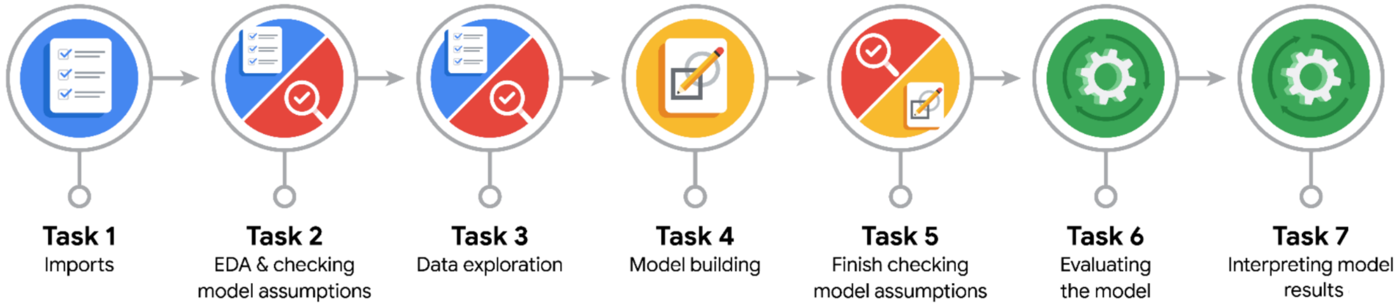
Relevant Interview Questions

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?

Ursular Sayo and May Santner

- What are you trying to solve or accomplish?

To build a logistic regression model and evaluate its performance in classifying user churn.

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

That there is some missing data in the target variable, there are outliers in the data and categorical variables need to be encoded.

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

Pandas, NumPy, matplotlib, seaborn and scikit-learn.

PACE: Analyze Stage

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

Some of the purposes of EDA before constructing a logistic regression model are to check for outliers and extreme data values as this can significantly impact logistic regression models. After visualising the data, a plan is made for addressing the outliers such as by dropping the rows, substituting extreme data values with average data or removing data values greater than 3 standard deviations. Another purpose of EDA before model construction us to help the analyst identify missing data which in turn helps to make the decisions of their exclusion or inclusion by substituting values with dataset means, medians and other methods. Additionally, it can be useful to engineer more variables via multiplying the variables together or by calculating the ratio between two variables.

- Do you have any ethical considerations in this stage?

Yes, to make sure that the model is fair as the target variable label is quite imbalanced.

PACE: Construct Stage

- Do you notice anything odd?

Yes, a small number of features dominate the model, in particular activity_days and professional_driver which have very high negative coefficients, with activity_days having -0.10 and professional_driver having -0.03. In contrast, the majority of the other variables have coefficients close to zero, thereby suggesting that some of the variables may in fact add limited incremental value once overall engagement is accounted for.

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

Yes the model can be improved by reducing the feature redundancy of the model, such as through feature selection or regularisation.

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

Pandas, NumPy, seaborn, matplotlib.pyplot, scikit-learn, scikit-learn LogisticRegression() function

PACE: Execute Stage

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

That activity_days and professional_driver had the biggest influence on the models predictions. The model achieved an accuracy of 83%, precision of 57% and a recall of 10%.

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

To focus the retention efforts on increasing the user engagement, particularly among the new and low-activity users due to higher activity levels are strongly associated with lower churn. Another business recommendation is to use the churn risk scores to target interventions where they are most likely to improve retention of the user.

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

It is important to interpret the beta coefficients as they show the direction and strength of each feature's impact on churn, thereby allowing stakeholders to understand why the model makes predictions.

- What potential recommendations would you make?

One potential recommendation I would make is to validate the logistic regression model on unseen data to reduce the risk of overfitting. Another is to engineer new features and reconstruct the model with different combinations of the features to reduce noise from the unpredictive features. Finally, to test alternative models such as tree-based models to compare their accuracy and interpretability.

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

Yes the model can be improved and one of the ways in which it can be is through the engineering of new features to attempt to generate a better predictive signal. It could also be useful to scale the independent variables and to reconstruct the model with different combinations of the predictor variables in order to reduce the noise from the unpredictive features.

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

To look more into engagement variables and to engineer new features which can be used to build several iterations of the model in order to improve classification performance.

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

Is the model good at classifying user churn?

Would you recommend this model to be used by Waze?

- Do you have any ethical considerations at this stage?

Yes, my ethical consideration is that this model is accurate, fair and explainable.