PACE Strategy Document

Introduction

This PACE strategy document can be used to record decisions and reflections as you work through this end-of-course project. You can use this document as a guide to consider your responses and reflections at different stages of the data analytical process. Additionally, the PACE strategy documents can be used as a resource when working on future projects.

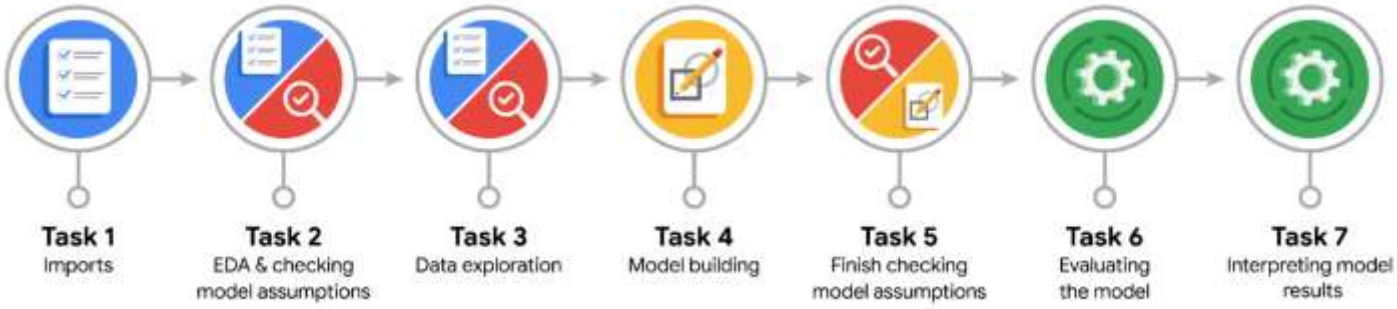
Relevant Interview Questions

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?

Titus Nelson, Operations Manager, NYC TLC Juliana Soto, Finance and Administration Department Head

- What are you trying to solve or accomplish?

The aim is to build a regression model to predict taxi fares before a ride is started.

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

A lot of suspicious values negative values negative values of distance, time, and fares. Max values were way too high. So the data had a lot of outliers.

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

Pandas, numpy, datetime, matplotlib, seaborn.

PACE: Analyze Stage

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

Handel outliers, remove errors, inspect relationships between variables.

- Do you have any ethical considerations in this stage?

Not to just delete the outlier data but to inspect them and analyze why a particular outlier occurred.

PACE: Construct Stage

- Do you notice anything odd?

Mean distance and mean duration are highly correlated.

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

We can continue with it since our prime focus is to obtain fare amounts only.

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

Sklearn libraries like preprocessing, linear_model, metrics.

PACE: Execute Stage

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

The model performs similarly on both training and test datasets. The scores are high so the model can be considered successful. The errors are on the lower side which is a good scenario. For the test data, an R2 of 0.869 means that 86.9% of the variance in the fare_amount variable is described by the model.

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

We see that the fare amount is highly dependent on mean_distance and mean_duration. For every mile traveled, the fare_amount increases by a mean of $2. Note, however, that because some highly correlated features were not removed, the confidence interval of this assessment is wider.

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

It tells which features have the most impact on our desired outcome.

- What potential recommendations would you make?

The higher the distance and duration, higher will be the fare amount.

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

Yes, the model can be improved by adding additional features. Also, this data as just of January month. If we get data of other month then we can predict with higher accuracy based on months, holidays. The rush_hour feature too can be improved by splitting in intervals. Historical traffic data too can significantly improve the model.

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

The regression model as successfully built with high accuracy and low error. The model can be deployed in booking apps to predict ride fare. Note however the model can be improved further as mentioned earlier.

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

Total amount prediction, tax prediction, Time predictions for the trip, etc.

- Do you have any ethical considerations at this stage?

Present the model results with confidence intervals and the erros.