Automatidata PACE Strategy Document V

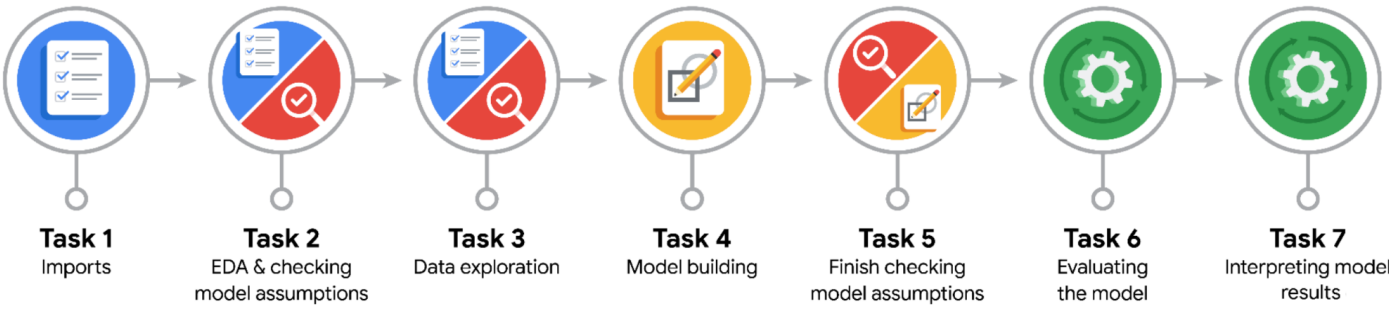
Regression Analysis

Introduction

PACE stands for Plan, Analyze, Construct and Execute. It is a framework that illustrates the foundation and structure for data analysis projects and each letter represents an actionable stage in a project. The stage "Plan" involves the definition of the project scope, the research of business data and the workflow development. The stage "Analyze" involves data scrubbing, data conversion and database formatting. The stage "Construct" involves building models and machine learning algorithms and selecting a modeling approach. The stage "Execute" involves the presentation of results to decision-makers, stakeholders and others in order to receive feedback. This framework is built upon an iterative cycle where each stage may reveal new insights, requiring the return to earlier stages. A PACE strategy document is used to record decisions and reflections at different stages of the data analytical process. It typically includes the definitions of roles and actions to ensure clarity and accountability.

Purpose

Automatidata is consulting for the New York City Taxi and Limousine Commission (TLC). New York City TLC is an agency responsible for licensing and regulating New York City's taxi cabs and for-hire vehicles. The agency has partnered with Automatidata to develop a regression model that helps estimate taxi fares before the ride, based on data that TLC has gathered. For this stage of the project, we identify seven main tasks that are presented in the following visual.



Considerations

PACE: Plan Stage

- What problem are we trying to solve or what outcome are we trying to achieve?

We want to predict accurate taxi fare amounts before rides begin using a regression model.

- Who are the external stakeholders for this project?

The external stakeholders include the New York City Taxi & Limousine Commission and taxi drivers.

- What are our first impressions of the data when we start examining it?

There are significant outliers and some nearly constant variables.

- What resources would be useful for this stage?

Valuable resources include data dictionaries, domain knowledge (e.g. flat-rate fares) and documentation on regression analysis.

PACE: Analyze Stage

- Why is Exploratory Data Analysis (EDA) important before building a multiple linear regression model?

It helps identify outliers, missing data and correlations among variables, ensuring a more accurate regression model.

- Are there any ethical issues to consider at this stage of the project?

Ethical considerations include addressing biases, ensuring fair representation of different routes and rides and preventing overfitting that may disadvantage some passengers.

PACE: Construct Stage

- Are there any unusual or unexpected patterns in the data?

There are high outliers in fare amount and duration and a fixed rate of $52 for certain airport trips.

- Can we make the model better? Are there any changes we would consider?

We can cap extreme fare values and address data leakage.

- What resources would be useful for this stage?

Data visualization libraries and documentation on statistical tests and performance metrics can be proven useful.

PACE: Execute Stage

- What important discoveries have we made from our model(s)?

Ride distance and duration are the most influential factors in determining fares. However, these two features are heavily correlated with each other. When we isolate the trip distance, the analysis suggests an average fare increase of about $9.5 for each additional mile of ride distance. When we isolate the trip duration from the distance, we observe an average fare increase of about $9 for each additional minute of ride time. Approximately 85.1% of the variation in fare amounts is explained by the model.

- Why is understanding the beta coefficients crucial for interpreting model results?

Understanding them helps to interpret the contribution of each variable to the fare prediction.

- Do we believe the model(s) could be enhanced? If so, how?

Addressing data leakage and improving feature selection could yield better results.

- What recommendations would we provide to the organization based on the model(s) built?

An app can be developed for riders to estimate fares based on model predictions. Additionally, more data for underrepresented routes should be collected.

- Given our knowledge of the data and model(s), what other questions could we explore for the team?

Interesting research questions are:

- How do fare estimates change during different times of the day?
- Could adding more real-time traffic data improve predictions?

- Are there any ethical issues to consider at this stage of the project?

We should ensure transparency in the fare estimation process and avoid bias against certain regions or time periods in the model's predictions.