Overview

In this exercise, you will use your machine learning experience to solve a straightforward but challenging prediction problem to exhibit your modeling skills.

You will be evaluated on the following — performance on the test set, feature engineering choices including features used and encoding of features, data processing, choice of models used, description of model performance and insights and observations from the model.

Problem Description

When a consumer places an order on DoorDash, we show the expected time of delivery. It is very important for DoorDash to get this right, as it has a big impact on consumer experience. In this exercise, you will build a model to predict the estimated time taken for a delivery.

Concretely, for a given delivery you must predict the **total delivery duration seconds** , i.e., the time taken from

Start: the time consumer submits the order (`created_at`) to
End: when the order will be delivered to the consumer (`actual_delivery_time`).

To help with this, we have provided

- **historical_data.csv:** table of historical deliveries
- **data_to_predict.json**: Json list of deliveries that you must predict on (for the second part)
- **data_description.txt**: description of all columns in **historical_data.csv** and details of **data_to_predict.csv**

Requirements

- Build a model to predict the total delivery duration seconds (as defined above). Feel free to generate additional features from the given data to improve model performance.
- Explain a) model(s) used, b) how you evaluated your model performance on the historical data, c) any data processing you performed on the data, d) feature engineering choices you made and e) other information you would like us to know about your modeling approach.
- Based on the findings from the model, list recommendations to reduce delivery time
- Output predictions on the instances in data_to_predict.csv — we will use the results here to evaluate your model

Deliverables

- Submit one document that includes a write-up explaining your model, choices made and discussion on the questions above.
- Your code / jupyter notebook
- Predictions for data_to_predict.csv.
    - *Should contain rows of the form <delivery_id>,<predicted duration>*

Notes

We expect the exercise to take 3-4 hours in total, but feel free to spend as much time as you like on it. Feel free to use any open source packages for the task.

Thank you for your hard work! Please let us know if you have any questions. Good luck!