Overview

In this data exercise, you will use your machine learning and modeling experience to solve a straightforward prediction problem. We require candidates to meet a high bar on this assignment to advance to the final round onsite, so **we would encourage you to demonstrate the full extent of your data science toolkit.** We will look at more than just your performance on the test set.

In particular, you will be evaluated on the following:

1. The quality of your data science process - data pre-processing, feature engineering, model selection, etc. **Please make sure to show all of your work!**
2. Your algorithm's performance on a test set where labels are withheld, evaluated on expected business impact.
3. The quality & insightfulness of your written report (**see requirements below**).

Problem Description

Before a consumer places an order on DoorDash, we show the expected delivery time. It is very important for DoorDash to get this right, as it has a big impact on consumer experience. Order lateness / underprediction of delivery time is of particular concern as past experiments suggest that underestimating delivery time is roughly twice as costly as overestimating it. Orders that are very early / late are also much worse than those that are only slightly early / late. 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 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 (your training set)
- **data_to_predict.csv**: data for deliveries that you must predict (label-free test set we will use for evaluation)
- **data_description.txt**: description of all columns in **historical_data.csv** and details of **data_to_predict.csv**

Requirements

1. Build a model to predict the total delivery duration seconds (as defined above). You'll likely find it helpful to generate additional features from the given data to improve model performance.
2. Output predictions on the instances in data_to_predict.csv — we will use the results here to evaluate your model
3. Write a short report (typically 1-2 pages) on the problem, your approach, and results containing:
    a. A high-level business summary explaining the key results to semi-technical readers. Include some feature interpretation and highlight the model's performance relative to what is likely feasible.
    b. A list of 3-5 features you believe would improve model performance if included in the training set. Please clearly articulate the value you believe adding these features would drive for the business.
    c. Assume the model you designed would replace a model already in production. How would you assess your model's performance relative to that of its predecessor before replacing the previous model?

    Please do not write about your modeling process - this is covered in the code submission.

Deliverables

1. Your code / jupyter notebook
2. Predictions for deliveries in data_to_predict.csv.
    - *Should contain rows of the form <delivery_id>,<predicted_duration>*
3. Your report

Notes

We expect the exercise to take 3-5 hours in total, but feel free to spend as much time as you would 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!