### Submission for Capstone Project proposal

Requirement on page:

<https://www.springboard.com/workshops/data-science-career-track/learn#/curriculum/1209>

1. What is the problem you want to solve?

To predict the duration of taxi trips in New York City

1. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

Accurate estimates of taxi trip durations can improve the taxi utilization and the satisfaction of drivers and passengers. If a taxi dispatching system knew approximately when a taxi driver would be ending their current ride, it would help identifying which drivers should be assigned to which pickup locations. Additionally, the predictive model could be used for finding optimal routes for different kinds of trips.

1. What data are you going to use for this? How will you acquire this data?

Kaggle provides a starting point dataset consisting of the records of ~1.5 million taxi trips that took place in 2016. Below is the list of the variables in dataset

.

**Test dataset:**

|  |  |
| --- | --- |
| **Variable name** | **Variable description** |
| id | A unique identifier for each trip |
| vendor\_id | A code indicating the provider associated with the trip record |
| pickup\_datetime | Date and time when the meter was engaged |
| passenger\_count | The number of passengers in the vehicle (driver entered value) |
| pickup\_longitude | The longitude where the meter was engaged |
| pickup\_latitude | The latitude where the meter was engaged |
| dropoff\_longitude | The longitude where the meter was disengaged |
| dropoff\_latitude | The latitude where the meter was disengaged |
| store\_and\_fwd\_flag | This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server. |
| Y=store and forward; N=not a store and forward trip |

**Train dataset:**

|  |  |
| --- | --- |
| **Variable name** | **Variable description** |
| id | A unique identifier for each trip |
| vendor\_id | A code indicating the provider associated with the trip record |
| pickup\_datetime | Date and time when the meter was engaged |
| dropoff\_datetime | Date and time when the meter was disengaged |
| passenger\_count | The number of passengers in the vehicle (driver entered value) |
| pickup\_longitude | The longitude where the meter was engaged |
| pickup\_latitude | The latitude where the meter was engaged |
| dropoff\_longitude | The longitude where the meter was disengaged |
| dropoff\_latitude | The latitude where the meter was disengaged |
| store\_and\_fwd\_flag | This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server. |
| Y=store and forward; N=not a store and forward trip |
| trip\_duration | Duration of the trip in seconds |

1. In brief, outline your approach to solving this problem (knowing that this might change later).

*Based on the variables available in train data ( those are not available in test data) and other common variables, trends will be analyzed and represented in the form of graphs / plots , visualizations / histograms. A predictive model will then be prepared that will be run on test data to get the trip duration(s).*

1. What are your deliverables?

Code (python notebook)