DATA TEST

Data Engineer Telefónica Chile

1_Data set (Required)
Built a yellow medallion taxicabs year dataset (2015) downloading raw data from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page . The final data set should have the following properties:

- Only yellow medallion taxicabs data (Yellow).
- 100K randomly selected samples from each month before filtering (1.2MM rows)
- At least 90K randomly selected samples from each month at the end of the process (>1.08MM rows).
- Data inside a 2 degree per side square geofencing (Center = latitude 40.7127, longitude -74.0059).
- Only pick ups inside geofencing.
- Only drop-off inside geofencing.
- Only trips origin and destiny between 07:00 am and 10:00 am.
- Only trips with 1-6 passengers.
- Only trips with standard fare = 1 (RatecodeID).
- Impute "Manhattan" distance column (Distance between pickup and drop-off)

Delivery: send to TID.cl@telefonica.com a link of a public repository containing a CVS final file without index, code and all the documentation explaining the process. The final dataset should have the following features: VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, pickup_longitude, pickup_latitude, RateCodeID, store_and_fwd_flag, dropoff_longitude, dropoff_latitude, manhattan_distance, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount,improvement_surcharge, total_amount .
Delivery date: March, Friday 15th. 2019

2_Bonus track (Optional)
Fare prediction (total_amount). With a sample subset, of the final dataset, train a Machine Learning model of your choice to predict the target feature "total_amount" (From the standard fare trip inside the geofencing). Training set = 70% of the subset. Compare RMSE (Root mean square error) of the predicted target feature against a baseline RMSE (For example: an average total_amount from subset).

Delivery: sent to TID.cl@telefonica.com a link of a public repository containing results, code and all the documentation explaining the process.
Delivery date: March, Friday 15th. 2019

Evaluation criteria:

1pt. Final dataset .csv format with all features.
1pt. Randomly selected samples from each month (2015). Data set rows number (>1.08MM rows).
1pt. Inside geofencing trips filter.
1pt. Time range filter.
1pt. Passenger filter.
1pt. Fare type filter.
1pt. "Manhattan" distance imputation.

0,5pt. Extra Trained model.
0,5pt. Extra RMSE comparison between target feature prediction and baseline RMSE

Data Dictionary – Yellow Taxi Trip Records

VendorID
A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.

tpep_pickup_datetime
The date and time when the meter was engaged.

tpep_dropoff_datetime
The date and time when the meter was disengaged.

Passenger_count
The number of passengers in the vehicle. This is a driver-entered value.

Trip_distance
The elapsed trip distance in miles reported by the taximeter.

pickup_longitud and pickup_latitude
Location in which the taximeter was engaged

dropoff_longitud and dropoff_latitude
Location in which the taximeter was disengaged

RateCodeID
The final rate code in effect at the end of the trip.
1 = Standard rate
2 = JFK
3 = Newark
4 = Nassau or Westchester
5 = Negotiated fare
6 = Group ride

store_and_fwd_flag
This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server.
Y= store and forward trip
N= not a store and forward trip

payment_type
A numeric code signifying how the passenger paid for the trip.
1 = Credit card
2 = Cash
3 = No charge
4 = Dispute
5 = Unknown
6 = Voided trip

fare_amount
The time-and-distance fare calculated by the meter.

extra
Miscellaneous extras and surcharges. Currently, this only includes the USD 0.50 and USD 1 rush hour and overnight charges.

mta_tax
USD 0.50 MTA tax that is automatically triggered based on the metered rate in use.

Improvement_surcharge
USD 0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.

tip_amount
Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.

tolls_amount
Total amount of all tolls paid in trip.

total_amount
The total amount charged to passengers. Does not include cash tips.