Questions and Tasks
Statistics and Machine Learning
OctoTelematics - June 2024

Summary

Dataset and environment preparation: 3
Question A: 4
Question A.1 (optional): 4
Question B: 5
Sharing results 6

All Rights Reserved

Below are reported the instructions for proceeding with the resolution and sharing of the results relating to the preliminary evaluation questions:

Dataset and environment preparation:

Find the reference dataset through the link:

https://octospa-my.sharepoint.com/:u:/g/personal/a_sauro_octotelematics_com/EZ6RSXtindlMgmvUNHANmGcB_bl8xrqEOwovbgLr1JrvYA?e=zw0PdS

Or by downloading the dataset from:

https://www.kaggle.com/datasets/microize/newyork-yellow-taxi-trip-data-2020-2019?select=yellow_tripdata_2019-04.csv (therefore referring to the 2019-Apr date)

The dataset is a csv file compressed in .zip format, it is possible to operate directly on the .zip file or perform the decompression in csv format and then load data directly from the csv file.

All questions must be solved with scripts in Python language executable on an environment with libraries installed:

- Pandas
- Numpy
- Sklearn
- Zipfile

Optional:

- Matplotlib
- Plotly

All Rights Reserved

Question A:

Calculate on the entire dataset the 5th, 50th and 95th percentiles (q05, q50, q95) on the dataset values: **'fare_amount ' , ' tip_amount ' and ' total_amount ' ** ; divided according to the **' VendorID ', ' passenger_count ' and ' payment_type ' fields**

The calculation output must be a dataFrame to be exported in CSV format organized with:

Columns: field name (on which the percentile is calculated) + "_p_" + percentile threshold

Rows (index): grouping field name + " _" + value of the group on which the percentile calculation is performed

Example output:

Index	fare_amount_p_5	fare_amount_p_50	fare_amount_p_95	tip_amount_p_5	tip_amount_p_50	tip_amount_p_95	total_amount_p_5	total_amount_p_50	total_amount_p_95
VendorID_1	4.5	9.5	35.5	0	2	6.85	8.3	14.8	49.07
VendorID_2	4.5	9.5	40	0	2	7.58	8.3	15.3	53.16
VendorID_4	4	9	33.5	0	2	7.58	8.3	14.76	49.711
passenger_count_0	4	9.5	37	0	2	7	8.3	14.8	50.7385
passenger_count_1	4.5	9.5	37.5	0	2	7.16	8.3	14.8	50.76
passenger_count_2	4.5	10	42	0	2	7.65	8.75	15.3	55.3
passenger_count_3	4.5	9.5	41.5	0	1.96	7.29	8.77	15.3	54.5
passenger_count_4	4.5	10	41.5	0	1.95	7.437	8.8	15.3	55.42
passenger_count_5	4.5	9.5	37.5	0	2	7.58	8.3	15.3	51.5
passenger_count_6	4.5	9.5	38.5	0	2	7.88	8.3	15.3	52.88
passenger_count_7	74.1	74.1	74.1	10	10	10	87.4	87.4	87.4
passenger_count_8	88.88	88.88	88.88	22.42	22.42	22.42	112.1	112.1	112.1
passenger_count_9	9.8	9.8	9.8	0	0	0	10.6	10.6	10.6
payment_type_1	4.5	10	38.5	0.62	2.45	8.48	9.36	15.95	53.42
payment_type_2	4	9	37	0	0	0	7.3	12.3	42.42
payment_type_3	-4.5	6	47	0	0	0	-8.3	9.8	50.908
payment_type_4	-14.77	5	42	0	0	0	-18.75	8.8	46.358
trip_distance>2.8	12.5	21.5	52	0	3.76	12.25	17.74	28.55	73.67
trip_distance<=2.8	4	8	14	0	1.85	3.65	8.15	13	20.75

*values shown are examples and calculated on a partial part of the dataset.

Files (example):

https://octospa-my.sharepoint.com/:x:/g/personal/a_sauro_octotelematics_com/EXEr61i42VxGtyz2vvwjxeQBZUyqIGflSRTsVj8pMDpPGg?e=uNalEn

Question A.1 (optional):

Calculate the percentiles as reported for question A also for the dataset divided by trip_distance if >2.8 or <=2.8 and add the calculated values to the dataFrame with the logic reported in question A

All Rights Reserved

Question B:

Generate an ML model for estimating the " total_amount " based on the variables (as input to the model): ***' VendorID ','passenger_count ','payment_type ',' trip_distance '***

It is possible to independently define the methodology and the selection and split process of the reference dataset for training, testing and verification of the model (kf, random, train -test- valid)

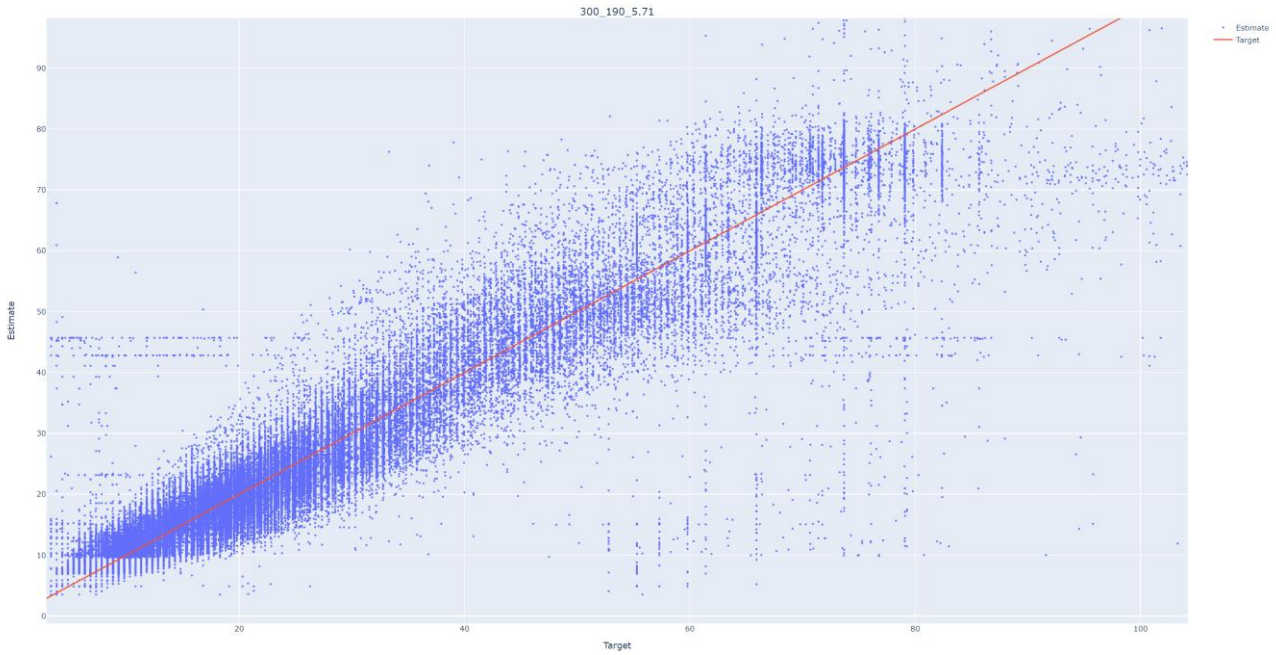
(optional) For model optimization it is recommended to calculate the RMSE on the selected partial test dataset

Export the generated model to file (ie via pickle, json ...)

*The quality assessment of the generated model will be verified through the calculation of the RMSE on a test dataset equivalent to the one used by the user in terms of format and compatible in terms of number (but not provided)

The user is given the right to use a different ML model from those present in the ***sklearn libraries,* but the generated model must be exportable, and the user must indicate the name and version of the library used (for calculation of the RMSE on the new test dataset)

Calculation example :



All Rights Reserved

Sharing results

Upon completion of the questions, generate a compressed file containing:

- python code for carrying out all the questions on files in .txt format (name it “code.txt”)
- output question A (or A + A.1) in .csv format (name it “report.csv”)
- output question B in export format, i.e. for pickle > pkl , (name it “ ml.[export format]”)

Name the zip archive file of the 3 files produced as **[name]_[surname].zip** (ie mario_rossi.zip or francesco_deangelis.zip for compound names)

Send the product file via the form : https://forms.gle/1Qyzsh6kgbkCTYQPA

If it is not possible to upload the zip file via Google Drive from the Form (you do not have a Google account), you can send the compressed file, but renamed **[name]_[surname]. txt**, via email to the following address: antonio.sauro@octotelematics.com

All Rights Reserved