CarsInLoc

PREDICTION OF THE AVAILABLE NUMBER OF CARS IN A LOCATION

MIGUEL JIMÉNEZ APARICIO

DECEMBER 2019

My data set

In numbers:

- Over 6 million rows
- 18998 timestamps
- 262 different cars
- 59456 pairs of coordinates

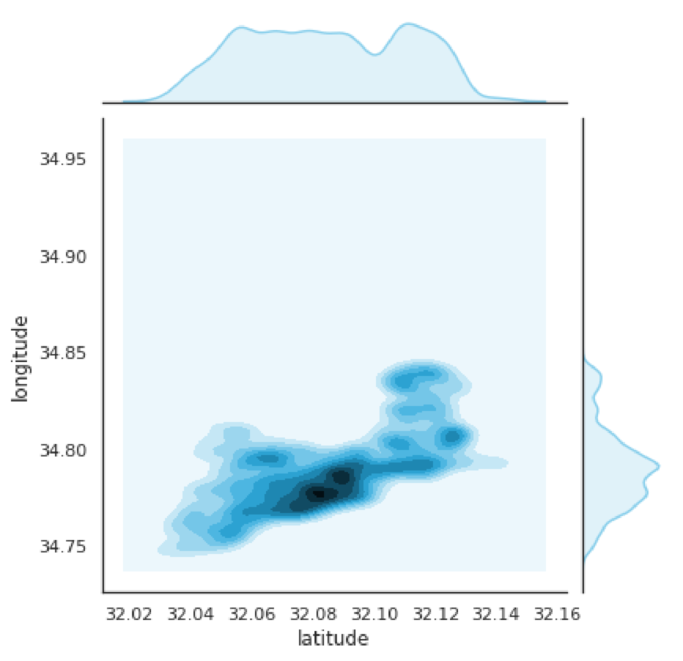
timestamp	latitude	longitude	total_cars	carsList
2019-01-10 11:45:...	32.09995	34.78794	1	[182]
2019-01-10 11:45:...	32.06567	34.79612	1	[268]
2019-01-10 11:45:...	32.06465	34.80322	1	[106]
2019-01-10 11:45:...	32.05978	34.81034	1	[180]
2019-01-10 11:45:...	32.05133	34.75089	1	[16]
2019-01-10 11:45:...	32.04223	34.7742	1	[72]
2019-01-10 11:45:...	32.04156	34.77128	1	[160]
2019-01-10 11:45:...	32.12373	34.81346	1	[210]
2019-01-10 11:45:...	32.11874	34.83406	1	[136]
2019-01-10 11:45:...	32.03351	34.75509	1	[27]
2019-01-10 11:45:...	32.14288	34.79361	1	[75]
2019-01-10 11:45:...	32.14306	34.79729	1	[132]
2019-01-10 11:45:...	32.083175	34.776552	0	[]
2019-01-10 11:45:...	32.088379	34.775111	0	[]
2019-01-10 11:45:...	32.074877	34.773515	0	[]
2019-01-10 11:45:...	32.098603	34.778565	0	[]
2019-01-10 11:45:...	32.09478	34.79728	0	[]
2019-01-10 11:45:...	32.098032	34.798089	0	[]
2019-01-10 11:45:...	32.12047	34.800318	0	[]
2019-01-10 11:45:...	32.04409	34.80421	0	[]

Use case

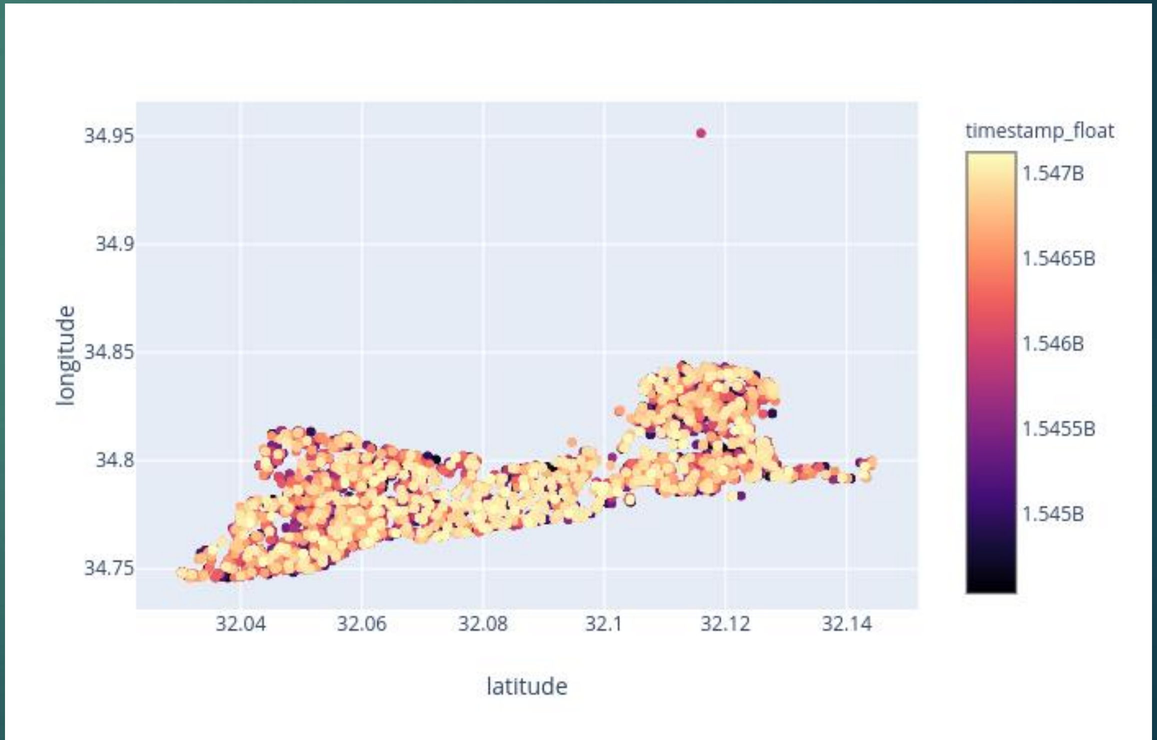
Use case I (Discontinued)

Use case II

- Prediction of the location of a single car



Density map for car 182



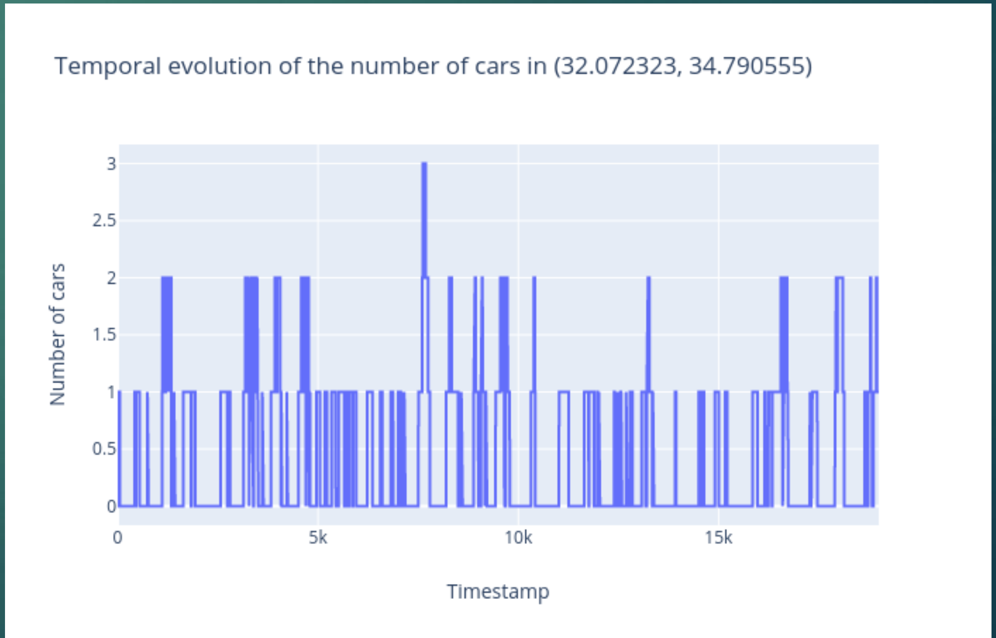
Temporal evolution of locations for car 182

Use case

Use case I (Discontinued)

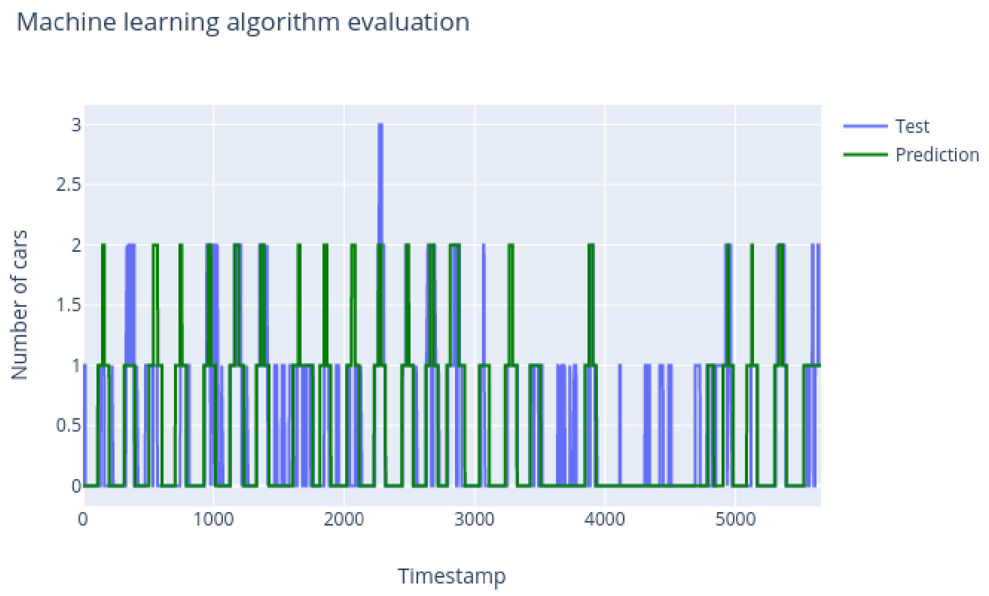
Use case II

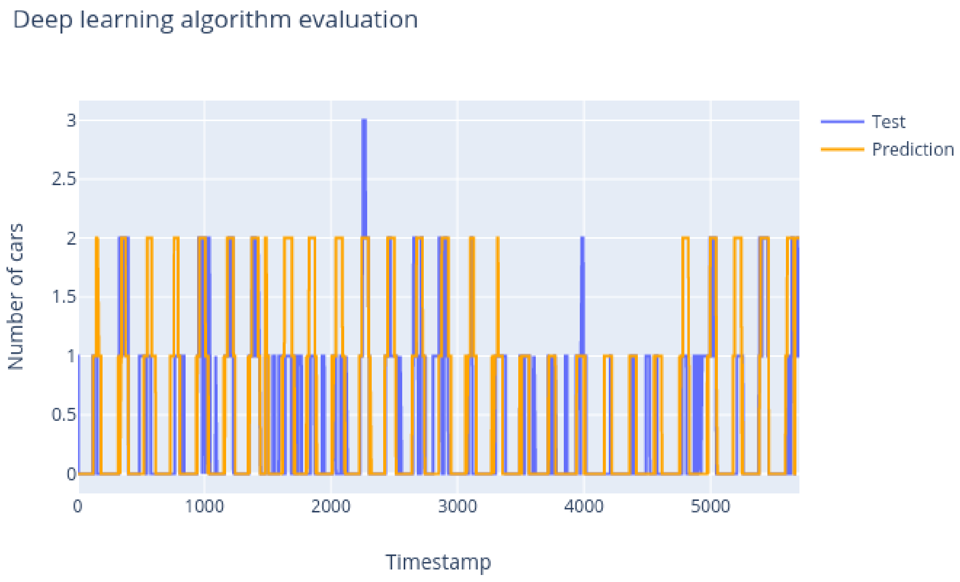
- Prediction of the numbers of cars in one particular location



Solution

	ML algorithm	DL Algorithm
Accuracy	73.07%	72.51%
Correlation	61.01%	60.06%





Architectural Decisions

- PySpark and Pandas
- Data is storaged locally in my PC as parquet files
- Keras and Apache Spark ML

Data quality assessment

Check :

- Ranges
- Empty cells

avg(total_cars)
0.6432291962782604

avg(latitude)	min(latitude)	max(latitude)
32.083372700257634	31.95223	32.14566

avg(longitude)	min(longitude)	max(longitude)
34.78898859070444	34.72998	34.95142

Get useful information:

- Number of cars, locations and timestamps

Data pre-processing (Use case I)

timestamp	latitude	longitude	total_cars	carsList
2019-01-10 11:45:...	32.09995	34.78794	1	[182]
2019-01-10 11:45:...	32.06567	34.79612	1	[268]
2019-01-10 11:45:...	32.06465	34.80322	1	[106]
2019-01-10 11:45:...	32.05978	34.81034	1	[180]
2019-01-10 11:45:...	32.05133	34.75089	1	[16]
2019-01-10 11:45:...	32.04223	34.7742	1	[72]
2019-01-10 11:45:...	32.04156	34.77128	1	[160]
2019-01-10 11:45:...	32.12373	34.81346	1	[210]
2019-01-10 11:45:...	32.11874	34.83406	1	[136]
2019-01-10 11:45:...	32.03351	34.75509	1	[27]
2019-01-10 11:45:...	32.14288	34.79361	1	[75]
2019-01-10 11:45:...	32.14306	34.79729	1	[132]
2019-01-10 11:45:...	32.083175	34.776552	0	[]
2019-01-10 11:45:...	32.088379	34.775111	0	[]
2019-01-10 11:45:...	32.074877	34.773515	0	[]
2019-01-10 11:45:...	32.098603	34.778565	0	[]
2019-01-10 11:45:...	32.09478	34.79728	0	[]
2019-01-10 11:45:...	32.098032	34.798089	0	[]
2019-01-10 11:45:...	32.12047	34.800318	0	[]
2019-01-10 11:45:...	32.04409	34.80421	0	[]

Initial DF

Data pre-processing (Use case I)

```
+--------------------+--------+---------+----------+--------+
|           timestamp| latitude|longitude|total_cars|carsList|
+--------------------+--------+---------+----------+--------+
|2019-01-10 11:45:...| 32.09995| 34.78794|         1|   [182]|
|2019-01-10 11:45:...| 32.06567| 34.79612|         1|   [268]|
|2019-01-10 11:45:...| 32.06465| 34.80322|         1|   [106]|
|2019-01-10 11:45:...| 32.05978| 34.81034|         1|   [180]|
|2019-01-10 11:45:...| 32.05133| 34.75089|         1|    [16]|
|2019-01-10 11:45:...| 32.04223|  34.7742|         1|    [72]|
|2019-01-10 11:45:...| 32.04156| 34.77128|         1|   [160]|
|2019-01-10 11:45:...| 32.12373| 34.81346|         1|   [210]|
|2019-01-10 11:45:...| 32.11874| 34.83406|         1|   [136]|
|2019-01-10 11:45:...| 32.03351| 34.75509|         1|    [27]|
|2019-01-10 11:45:...| 32.14288| 34.79361|         1|    [75]|
|2019-01-10 11:45:...| 32.14306| 34.79729|         1|   [132]|
|2019-01-10 11:45:...|32.083175|34.776552|         0|      []|
|2019-01-10 11:45:...|32.088379|34.775111|         0|      []|
|2019-01-10 11:45:...|32.074877|34.773515|         0|      []|
|2019-01-10 11:45:...|32.098603|34.778565|         0|      []|
|2019-01-10 11:45:...| 32.09478| 34.79728|         0|      []|
|2019-01-10 11:45:...|32.098032|34.798089|         0|      []|
|2019-01-10 11:45:...| 32.12047|34.800318|         0|      []|
|2019-01-10 11:45:...| 32.04409| 34.80421|         0|      []|
+--------------------+--------+---------+----------+--------+
```

Initial DF

String processing

Separate labels into columns

```
+--------------------+---------+---------+----------+---------------+-----+-----+-----+-----+-----+-----+-----+----
-+-----+------+
|           timestamp| latitude|longitude|total_cars|       carsList|car_1|car_2|car_3|car_4|car_5|car_6|car_7|car_
8|car_9|car_10|
+--------------------+---------+---------+----------+---------------+-----+-----+-----+-----+-----+-----+-----+----
-+-----+------+
|2019-01-10 10:41:...|32.072323|34.790555|         1|           [20]|   20| null| null| null| null| null| null| nul
l| null|  null|
|2019-01-10 09:11:...|32.093871|34.785879|         0|             []|     | null| null| null| null| null| null| nul
l| null|  null|
|2019-01-10 12:04:...|32.064615|34.795787|         0|             []|     | null| null| null| null| null| null| nul
l| null|  null|
|2019-01-10 11:25:...| 32.11368| 34.79476|         1|           [39]|   39| null| null| null| null| null| null| nul
l| null|  null|
|2019-01-10 09:45:...|32.089229|34.786514|         1|          [250]|  250| null| null| null| null| null| null| nul
l| null|  null|
|2019-01-10 09:49:...|32.044725|34.767288|         1|          [168]|  168| null| null| null| null| null| null| nul
l| null|  null|
|2019-01-10 09:55:...| 32.04261| 34.76513|         1|          [155]|  155| null| null| null| null| null| null| nul
l| null|  null|
|2019-01-10 12:10:...| 32.03985| 34.77473|         2|      [88, 102]|   88|  102| null| null| null| null| null| nul
l| null|  null|
```

Data pre-processing (Use case I)

timestamp	latitude	longitude	total_cars	carsList
2019-01-10 11:45:...	32.09995	34.78794	1	[182]
2019-01-10 11:45:...	32.06567	34.79612	1	[268]
2019-01-10 11:45:...	32.06465	34.80322	1	[106]
2019-01-10 11:45:...	32.05978	34.81034	1	[180]
2019-01-10 11:45:...	32.05133	34.75089	1	[16]
2019-01-10 11:45:...	32.04223	34.7742	1	[72]
2019-01-10 11:45:...	32.04156	34.77128	1	[160]
2019-01-10 11:45:...	32.12373	34.81346	1	[210]
2019-01-10 11:45:...	32.11874	34.83406	1	[136]
2019-01-10 11:45:...	32.03351	34.75509	1	[27]
2019-01-10 11:45:...	32.14288	34.79361	1	[75]
2019-01-10 11:45:...	32.14306	34.79729	1	[132]
2019-01-10 11:45:...	32.083175	34.776552	0	[]
2019-01-10 11:45:...	32.088379	34.775111	0	[]
2019-01-10 11:45:...	32.074877	34.773515	0	[]
2019-01-10 11:45:...	32.098603	34.778565	0	[]
2019-01-10 11:45:...	32.09478	34.79728	0	[]
2019-01-10 11:45:...	32.098032	34.798089	0	[]
2019-01-10 11:45:...	32.12047	34.800318	0	[]
2019-01-10 11:45:...	32.04409	34.80421	0	[]

Initial DF

Filter by car

timestamp	latitude	longitude
2018-12-11 15:48:...	32.083	34.7806
2018-12-11 15:50:...	32.12093	34.81254
2018-12-11 15:53:...	32.04533	34.7819
2018-12-11 15:53:...	32.035393	34.75873
2018-12-11 15:57:...	32.092	34.79579
2018-12-11 15:57:...	32.106566	34.797869
2018-12-11 16:03:...	32.050287	34.752289
2018-12-11 16:03:...	32.12093	34.81254
2018-12-11 16:03:...	32.10877	34.83471
2018-12-11 16:09:...	32.056237	34.769956
2018-12-11 16:20:...	32.12093	34.81254
2018-12-11 16:22:...	32.087613	34.784496
2018-12-11 16:22:...	32.076339	34.78686

String processing

Separate labels into columns

timestamp	latitude	longitude	total_cars	carsList	car_1	car_2	car_3	car_4	car_5	car_6	car_7	car_8	car_9	car_10
2019-01-10 10:41:...	32.072323	34.790555	1	[20]	20	null	null	null	null	null	null	null	null	null
2019-01-10 09:11:...	32.093871	34.785879	0	[]		null	null	null	null	null	null	null	null	null
2019-01-10 12:04:...	32.064615	34.795787	0	[]		null	null	null	null	null	null	null	null	null
2019-01-10 11:25:...	32.11368	34.79476	1	[39]	39	null	null	null	null	null	null	null	null	null
2019-01-10 09:45:...	32.089229	34.786514	1	[250]	250	null	null	null	null	null	null	null	null	null
2019-01-10 09:49:...	32.044725	34.767288	1	[168]	168	null	null	null	null	null	null	null	null	null
2019-01-10 09:55:...	32.04261	34.76513	1	[155]	155	null	null	null	null	null	null	null	null	null
2019-01-10 12:10:...	32.03985	34.77473	2	[88, 102]	88	102	null	null	null	null	null	null	null	null

Data pre-processing (Use case II)

timestamp	latitude	longitude	total_cars	carsList
2019-01-10 11:45:...	32.09995	34.78794	1	[182]
2019-01-10 11:45:...	32.06567	34.79612	1	[268]
2019-01-10 11:45:...	32.06465	34.80322	1	[106]
2019-01-10 11:45:...	32.05978	34.81034	1	[180]
2019-01-10 11:45:...	32.05133	34.75089	1	[16]
2019-01-10 11:45:...	32.04223	34.7742	1	[72]
2019-01-10 11:45:...	32.04156	34.77128	1	[160]
2019-01-10 11:45:...	32.12373	34.81346	1	[210]
2019-01-10 11:45:...	32.11874	34.83406	1	[136]
2019-01-10 11:45:...	32.03351	34.75509	1	[27]
2019-01-10 11:45:...	32.14288	34.79361	1	[75]
2019-01-10 11:45:...	32.14306	34.79729	1	[132]
2019-01-10 11:45:...	32.083175	34.776552	0	[]
2019-01-10 11:45:...	32.088379	34.775111	0	[]
2019-01-10 11:45:...	32.074877	34.773515	0	[]
2019-01-10 11:45:...	32.098603	34.778565	0	[]
2019-01-10 11:45:...	32.09478	34.79728	0	[]
2019-01-10 11:45:...	32.098032	34.798089	0	[]
2019-01-10 11:45:...	32.12047	34.800318	0	[]
2019-01-10 11:45:...	32.04409	34.80421	0	[]

Initial DF

Filter by latitude and longitude

timestamp	total_cars
2018-12-11 15:48:...	1
2018-12-11 15:50:...	1
2018-12-11 15:53:...	1
2018-12-11 15:55:...	1
2018-12-11 15:57:...	1
2018-12-11 15:59:...	1
2018-12-11 16:01:...	1
2018-12-11 16:03:...	1
2018-12-11 16:05:...	1
2018-12-11 16:07:...	1

Feature Engineering

timestamp	total_cars
2018-12-11 15:48:...	1
2018-12-11 15:50:...	1
2018-12-11 15:53:...	1
2018-12-11 15:55:...	1
2018-12-11 15:57:...	1
2018-12-11 15:59:...	1
2018-12-11 16:01:...	1
2018-12-11 16:03:...	1
2018-12-11 16:05:...	1
2018-12-11 16:07:...	1

Separate timestamps into different columns

timestamp	total_cars	minute	hour	day	month	season	year	total_cars_int
2018-12-11 15:48:...	1	48	15	11	12	4	2018	1
2018-12-11 15:50:...	1	50	15	11	12	4	2018	1
2018-12-11 15:53:...	1	53	15	11	12	4	2018	1
2018-12-11 15:55:...	1	55	15	11	12	4	2018	1
2018-12-11 15:57:...	1	57	15	11	12	4	2018	1
2018-12-11 15:59:...	1	59	15	11	12	4	2018	1
2018-12-11 16:01:...	1	1	16	11	12	4	2018	1
2018-12-11 16:03:...	1	3	16	11	12	4	2018	1
2018-12-11 16:05:...	1	5	16	11	12	4	2018	1
2018-12-11 16:07:...	1	7	16	11	12	4	2018	1
2018-12-11 16:09:...	1	9	16	11	12	4	2018	1
2018-12-11 16:11:...	1	11	16	11	12	4	2018	1
2018-12-11 16:13:...	1	13	16	11	12	4	2018	1

Feature Engineering (for ML algo.)

Add features columns

timestamp	total_cars	minute	hour	day	month	season	year	total_cars_int	features
2018-12-11 15:48:...	1	48	15	11	12	4	2018	1	[48.0,15.0,11.0,1...
2018-12-11 15:50:...	1	50	15	11	12	4	2018	1	[50.0,15.0,11.0,1...
2018-12-11 15:53:...	1	53	15	11	12	4	2018	1	[53.0,15.0,11.0,1...
2018-12-11 15:55:...	1	55	15	11	12	4	2018	1	[55.0,15.0,11.0,1...
2018-12-11 15:57:...	1	57	15	11	12	4	2018	1	[57.0,15.0,11.0,1...
2018-12-11 15:59:...	1	59	15	11	12	4	2018	1	[59.0,15.0,11.0,1...
2018-12-11 16:01:...	1	1	16	11	12	4	2018	1	[1.0,16.0,11.0,12...
2018-12-11 16:03:...	1	3	16	11	12	4	2018	1	[3.0,16.0,11.0,12...
2018-12-11 16:05:...	1	5	16	11	12	4	2018	1	[5.0,16.0,11.0,12...
2018-12-11 16:07:...	1	7	16	11	12	4	2018	1	[7.0,16.0,11.0,12...
2018-12-11 16:09:...	1	9	16	11	12	4	2018	[illegible]	[9.0,16.0,11.0,12...
2018-12-11 16:11:...	1	11	16	11	12	4	2018	1	[11.0,16.0,11.0,1...
2018-12-11 16:13:...	1	13	16	11	12	4	2018	1	[13.0,16.0,11.0,1...
2018-12-11 16:16:...	1	16	16	11	12	4	2018	1	[illegible]16.0,16.0,11.0,1...
2018-12-11 16:18:...	1	18	16	11	12	4	2018	1	[1[illegible].0,16.0,11.0,1...

Feature Engineering (for DL algo.)

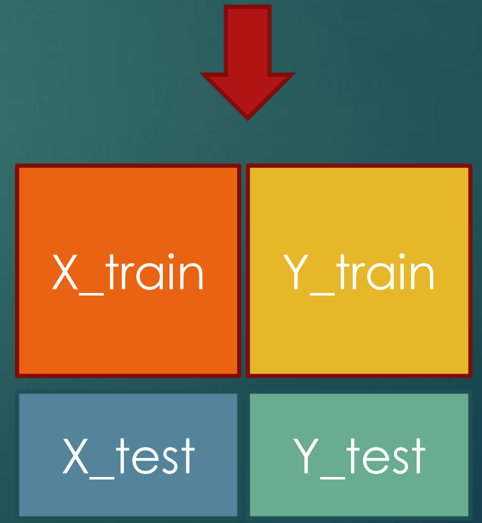
One-hot encoder for season column

Standard scaling

Split into X and Y (input and output)

Split into Train and Test set

timestamp	total_cars	minute	hour	day	month	season	year	total_cars_int
2018-12-11 15:48:...	1	48	15	11	12	4	2018	1
2018-12-11 15:50:...	1	50	15	11	12	4	2018	1
2018-12-11 15:53:...	1	53	15	11	12	4	2018	1
2018-12-11 15:55:...	1	55	15	11	12	4	2018	1
2018-12-11 15:57:...	1	57	15	11	12	4	2018	1
2018-12-11 15:59:...	1	59	15	11	12	4	2018	1
2018-12-11 16:01:...	1	1	16	11	12	4	2018	1
2018-12-11 16:03:...	1	3	16	11	12	4	2018	1
2018-12-11 16:05:...	1	5	16	11	12	4	2018	1
2018-12-11 16:07:...	1	7	16	11	12	4	2018	1
2018-12-11 16:09:...	1	9	16	11	12	4	2018	1
2018-12-11 16:11:...	1	11	16	11	12	4	2018	1
2018-12-11 16:13:...	1	13	16	11	12	4	2018	1



Model Algorithm

Machine Learning

- Decision Tree Regressor
  - Feature importance

	idx	name	score
1	1	hour	0.726897
2	2	day	0.271512
0	0	minute	0.001591
3	3	month	0.000000
4	4	season	0.000000
5	5	year	0.000000

Deep Learning

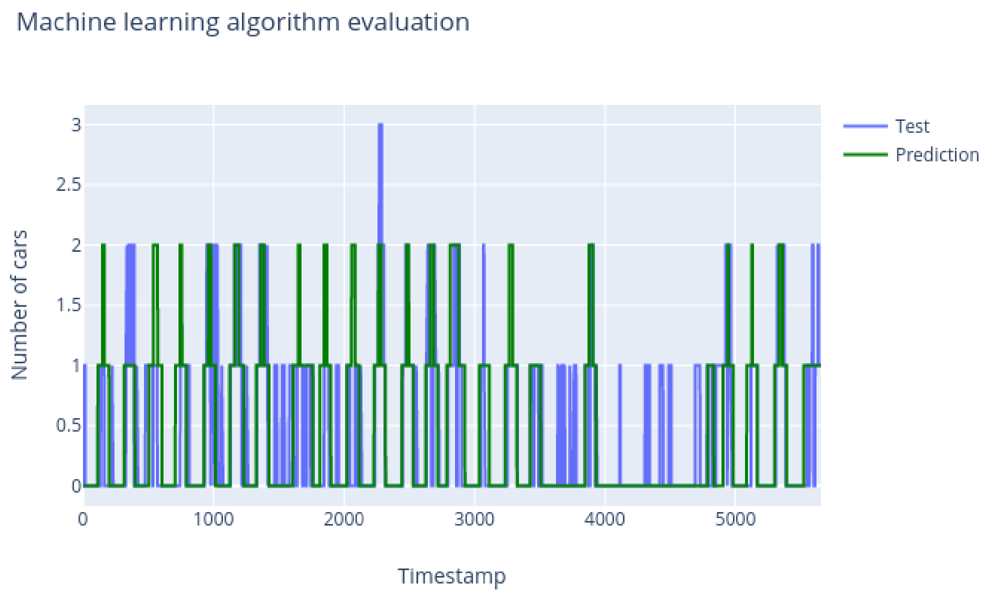
- 4 dense layers (2 relu, tanh and sigmoid)
- Compiler:
  - Optimizer: adam
  - Loss: binary_crossentropy
  - Metric: accuracy
- 10 epochs
- Final step:
  - Loss: 0.1364
  - Accuracy: 0.7408

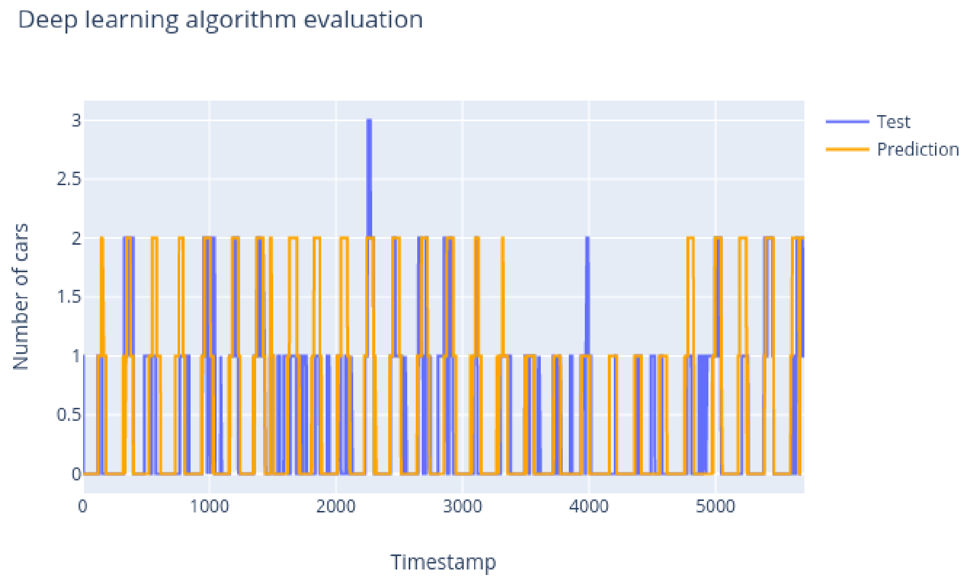
Model performance and indicators

$$Accuracy = \frac{Timestamps\ in\ which\ the\ prediction\ is\ correct}{Number\ of\ timestamps}$$

	ML algorithm	DL Algorithm
Accuracy	73.07%	72.51%
Correlation	0.6101	0.6006

Pearson correlation between the test set and the predictions





Thanks!