NTDS Project: Bike Sharing Demand

Vincent Hardy

18/01/2017

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Acquisition of Row data

On Kaggle, Dataset from 13/10/2014 to 31/08/2016 (689 days):

kaggle

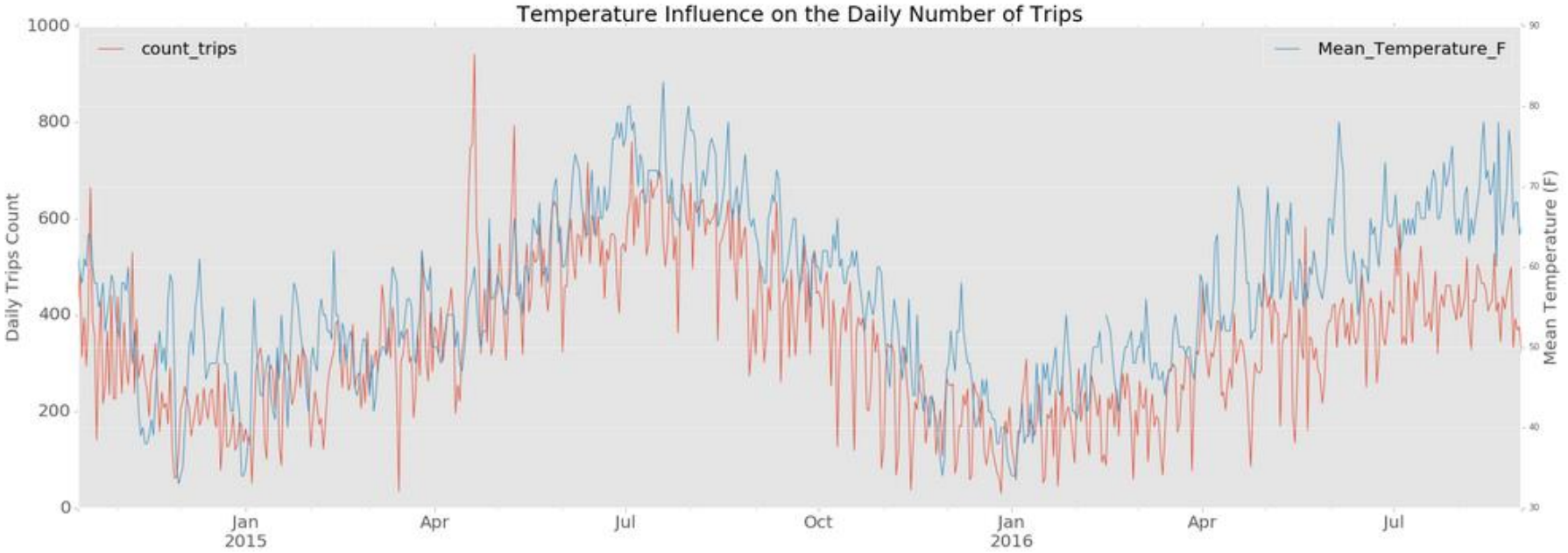
- Station.csv (58 ,8)
- Trip.csv (236065, 14)
- Weather.csv (689, 20)

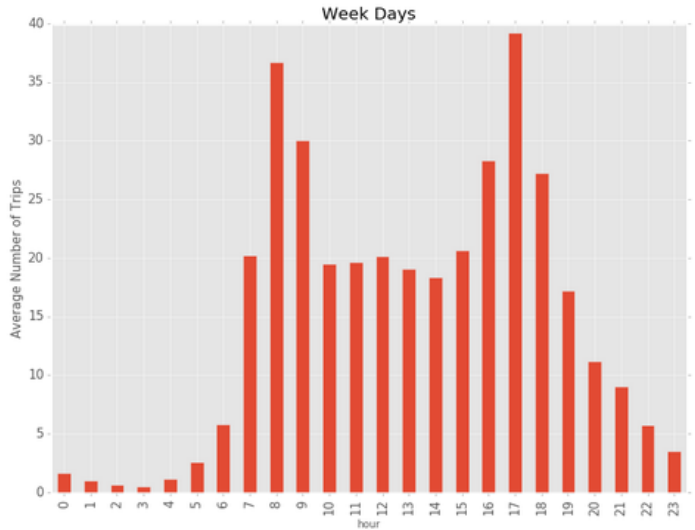
How to predict the bikes demand at each station of Seattle?

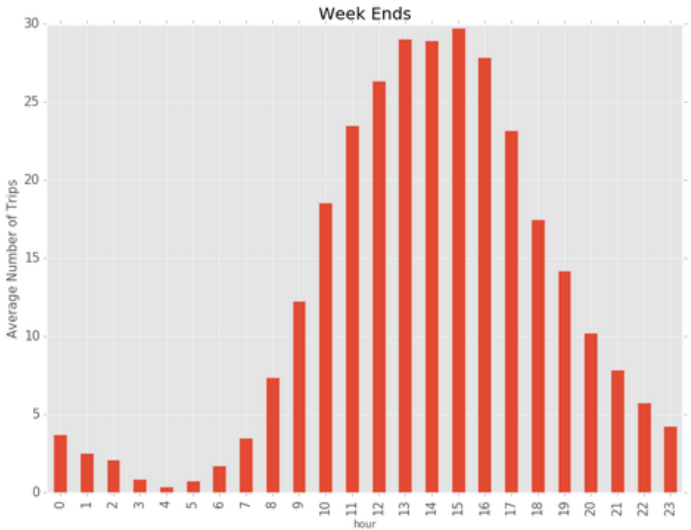
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Exploration: Users Behaviors

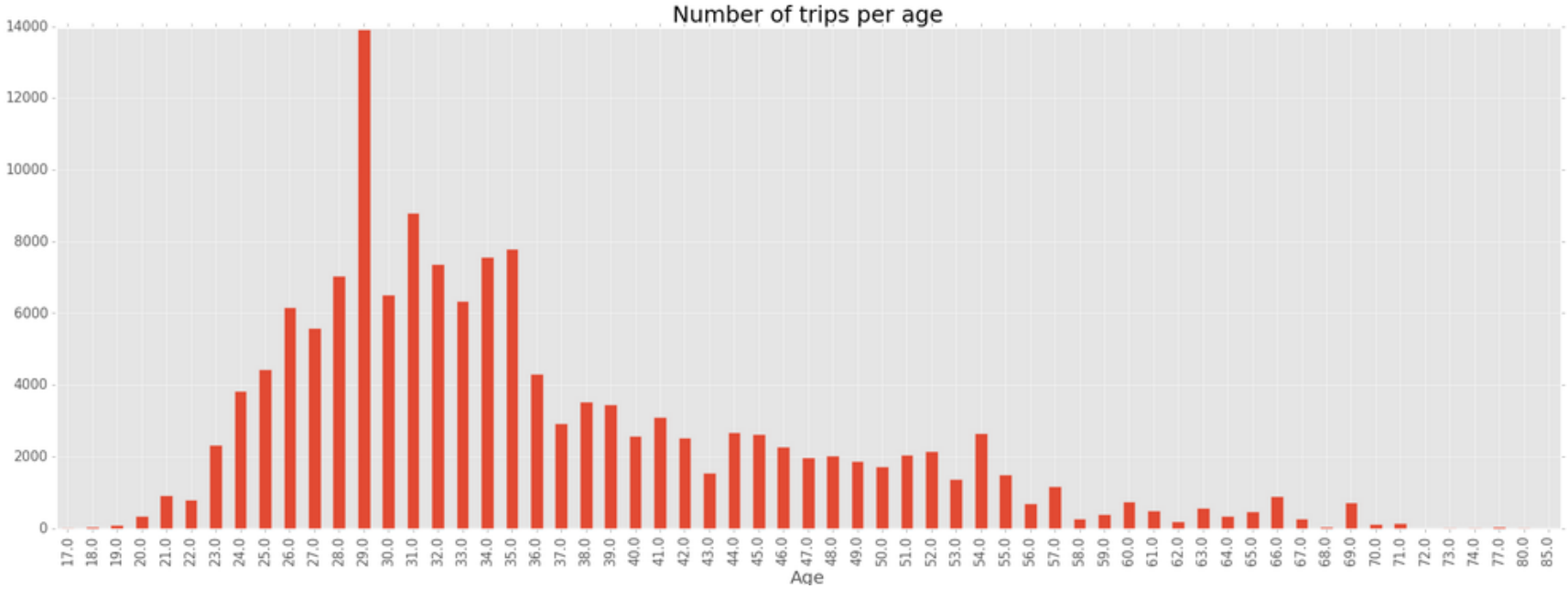
- Daily and Hourly Analysis



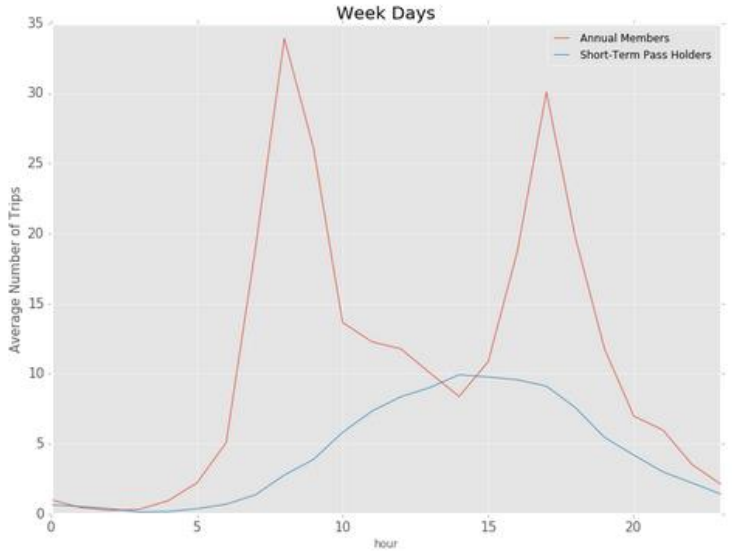


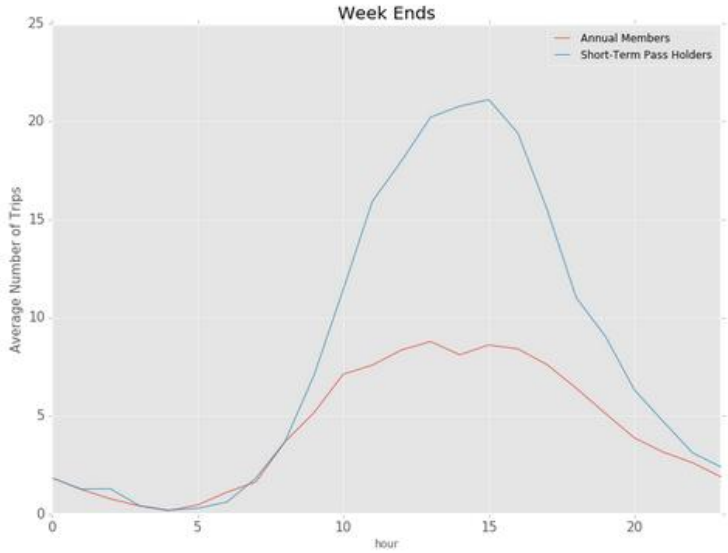


- Age Analysis



- Type of user: Annual Member or Short-Term Pass Holder (24H or 3D pass)





PCA Analysis on daily trends

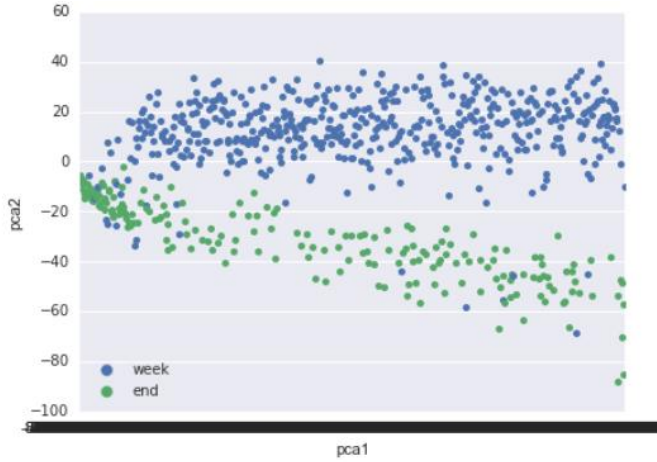
- 25 dimensions and 4 labels

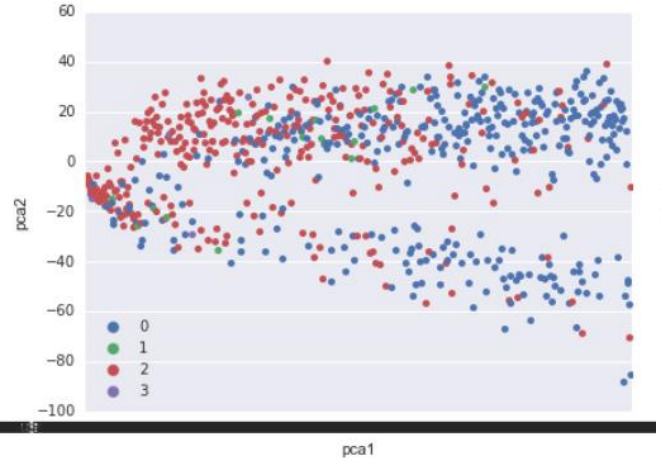
hour	0	1	2	3	4	5	6	7	8	9	...	19	20	21	22	23	Mean_Temperature_F	Precipitation_In	day_of_week	Events	total_trips
date																					
2014-10-13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	12.0	12.0	10.0	4.0	3.0	62.0	0.00	week	2	406.0
2014-10-14	0.0	0.0	0.0	0.0	0.0	1.0	5.0	22.0	35.0	25.0	...	28.0	17.0	13.0	6.0	2.0	59.0	0.11	week	2	489.0
2014-10-15	3.0	2.0	1.0	0.0	0.0	0.0	7.0	10.0	12.0	21.0	...	21.0	12.0	12.0	8.0	1.0	58.0	0.45	week	2	312.0
2014-10-16	4.0	1.0	0.0	0.0	0.0	2.0	7.0	14.0	35.0	25.0	...	22.0	21.0	8.0	6.0	6.0	61.0	0.00	week	2	389.0
2014-10-17	2.0	1.0	1.0	0.0	0.0	1.0	5.0	16.0	28.0	26.0	...	15.0	10.0	20.0	9.0	2.0	60.0	0.14	week	2	292.0

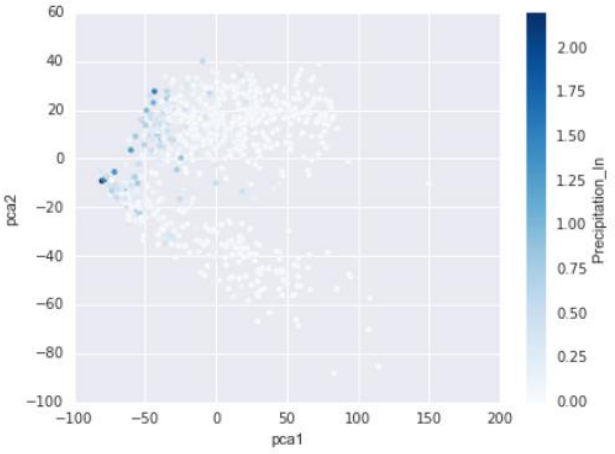
- Aim: reduce to 2 dimensions & identify similar patterns

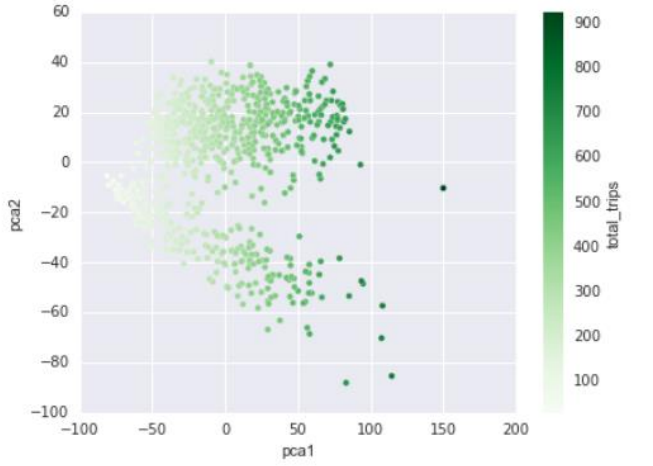
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

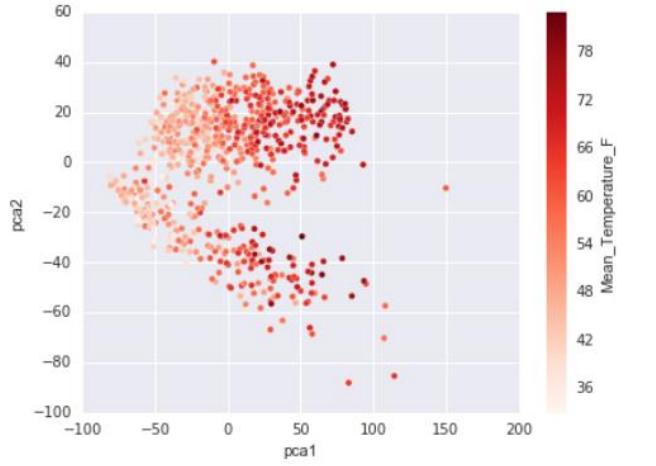
- Applying PCA & Visualization







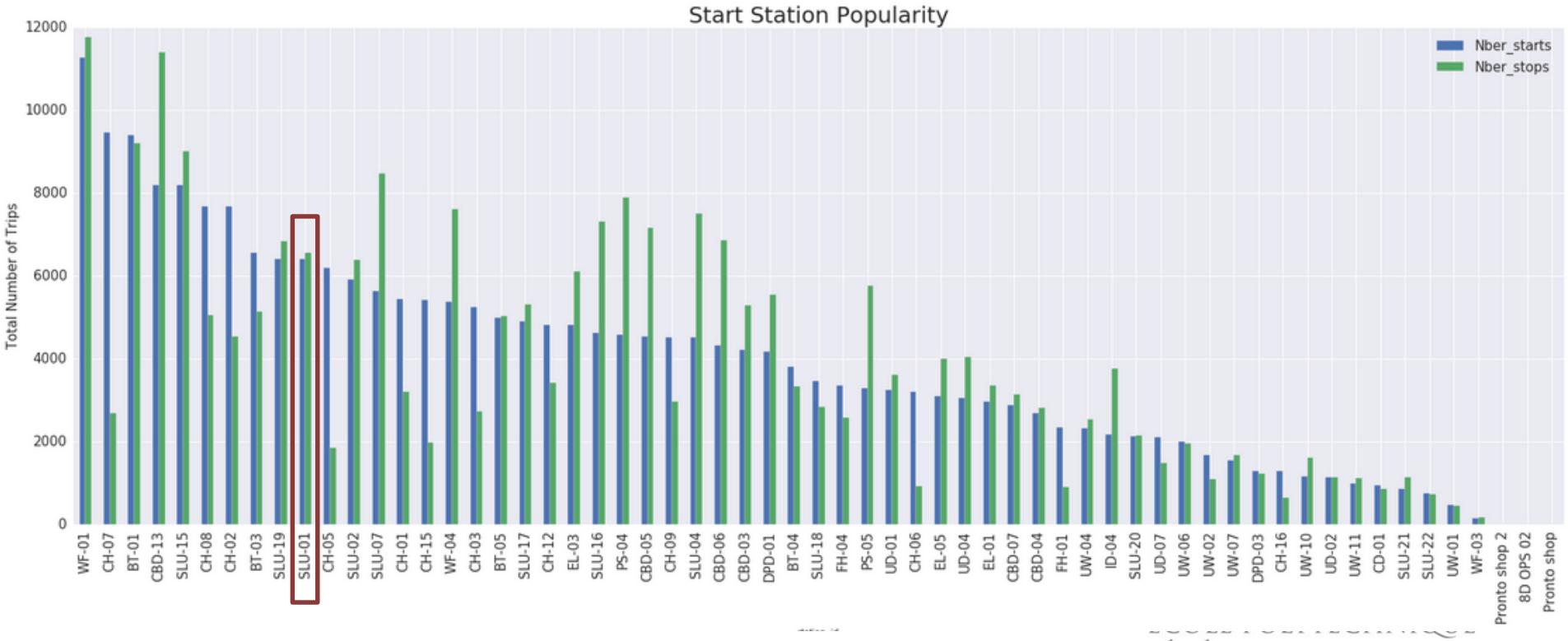




ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Bikes Variation Analysis Per Station

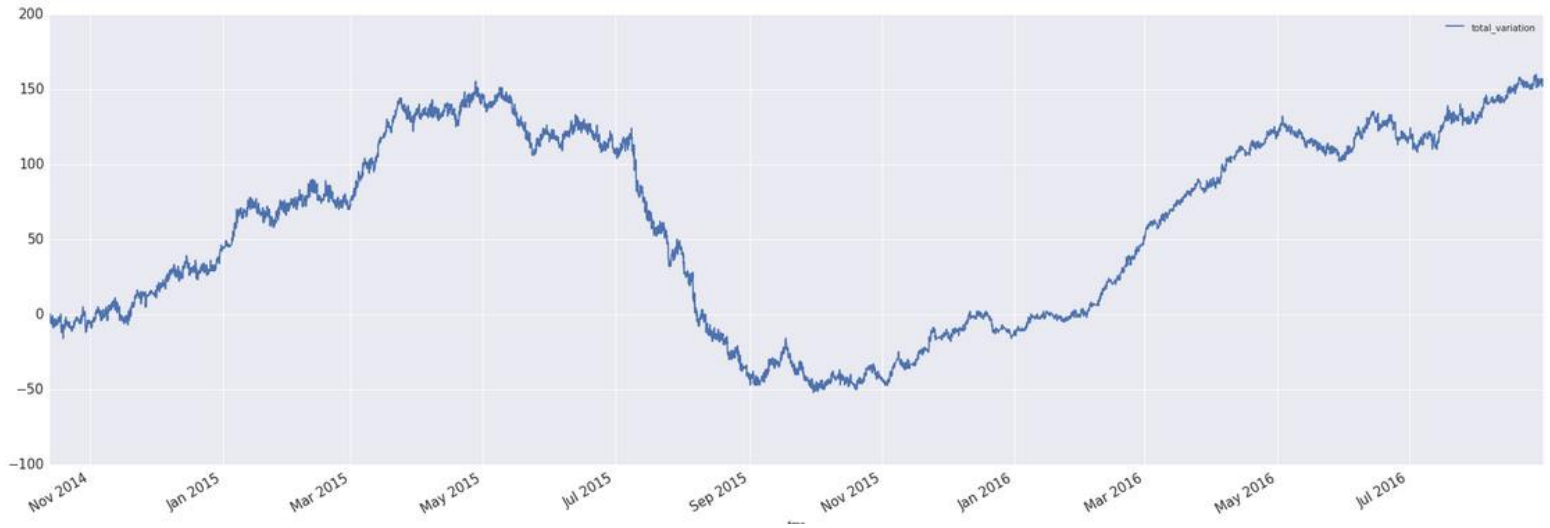
- Which are the most popular start/stop stations?



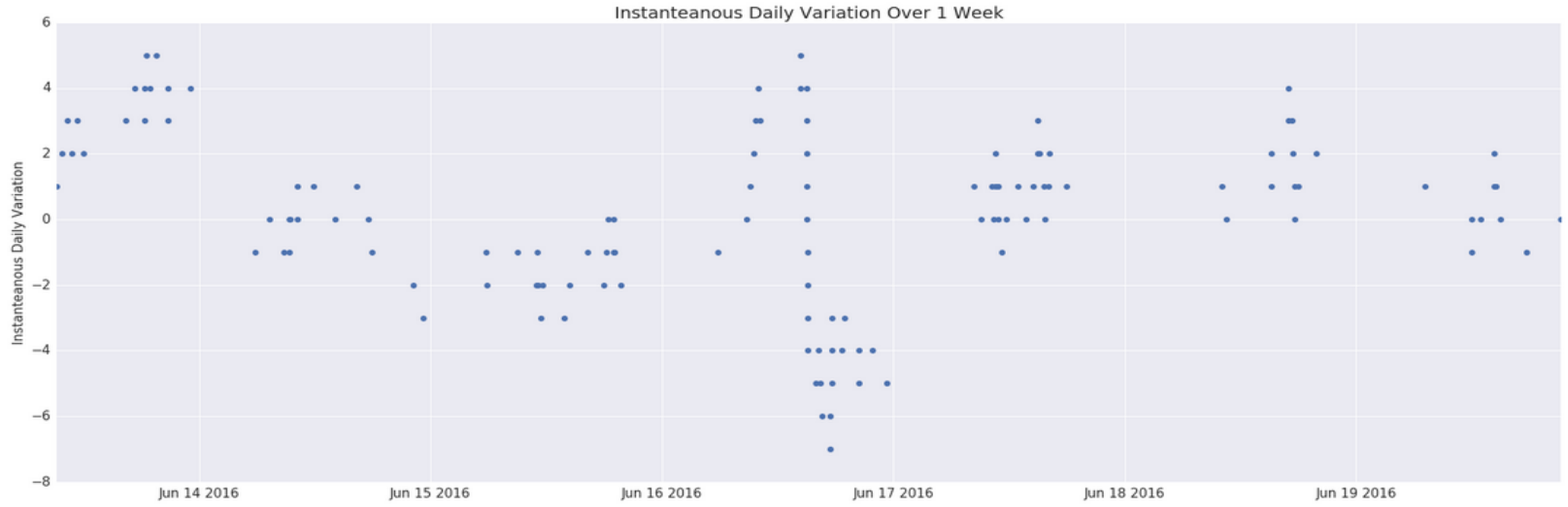
FÉDÉRALE DE LAUSANNE

- Analysis at station SLU-01

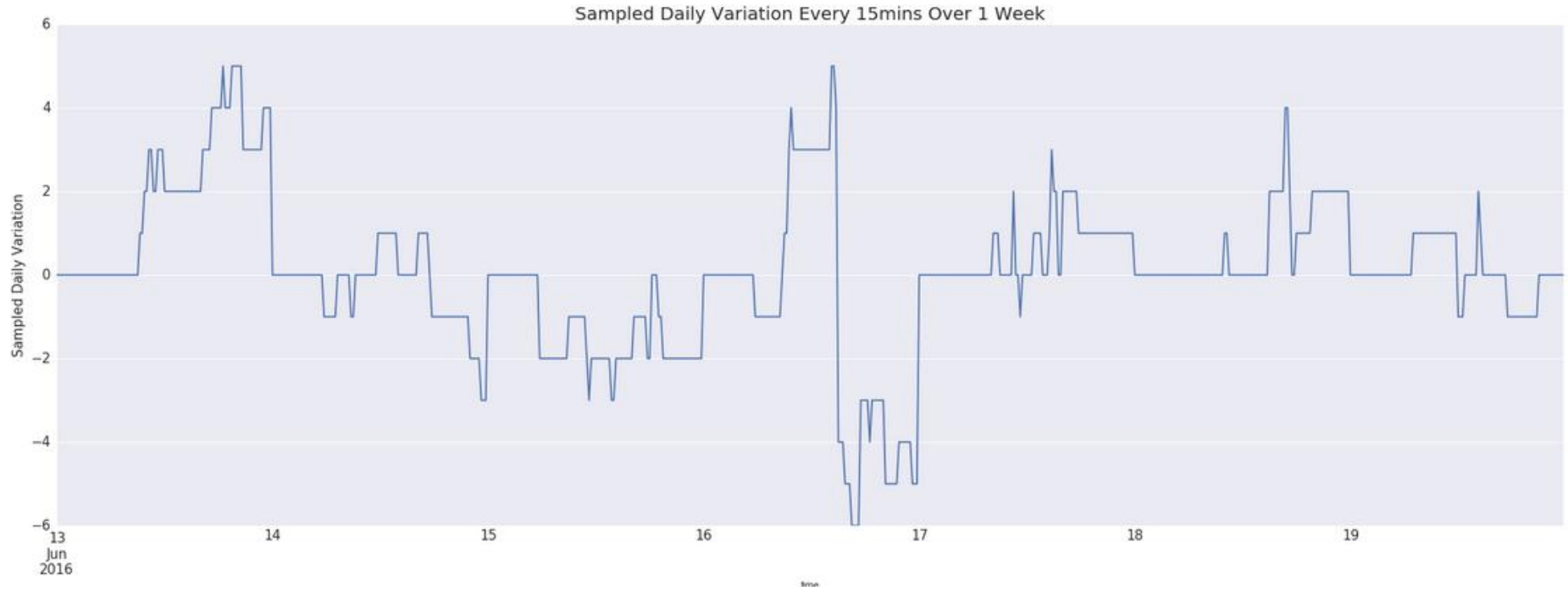
Is it possible to infer the bike availability?



… No but daily bikes variation YES!



But regression purpose requires a resampling of this daily variation:



- Computation of daily variation at large scale:
  - 45 stations
  - From 14/10/2014 to 17/03/2016
  - Number of data computed: 2 246 400

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Regression on the Daily Bikes Variation

- Knowing a dataset **X**, we would like to predict the daily variation **y** such that **f(X) = y**
- Which features choose to compose X? How to represent them?

X

y

	Events	Mean_Temperature_F	Precipitation_In	daily_variation	station_id	month	weekday	hour
2014-10-14 00:00:00	Rain	59.0	0.11	0.0	SLU-07	10	1	0.00
2014-10-14 00:15:00	Rain	59.0	0.11	0.0	SLU-07	10	1	0.25
2014-10-14 00:30:00	Rain	59.0	0.11	0.0	SLU-07	10	1	0.50
2014-10-14 00:45:00	Rain	59.0	0.11	0.0	SLU-07	10	1	0.75
2014-10-14 01:00:00	Rain	59.0	0.11	0.0	SLU-07	10	1	1.00

- Size of **X** : 2 242 080 x 7
- Size of **y** : 2 242 080 x 1
- Evaluation criteria: MSE & MAE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

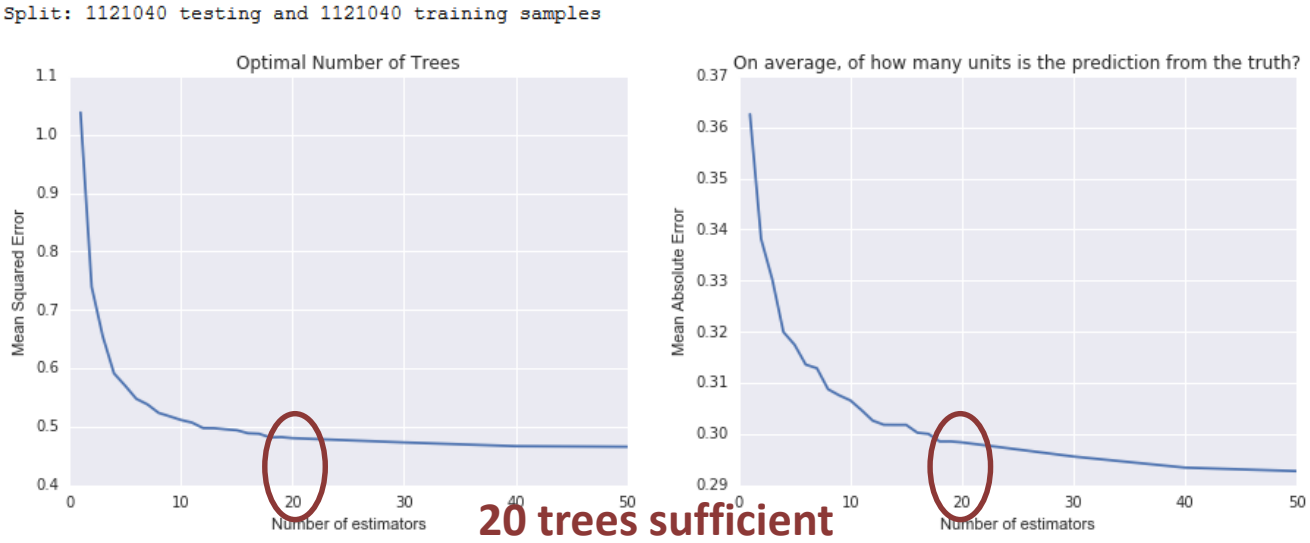
One model for all the stations

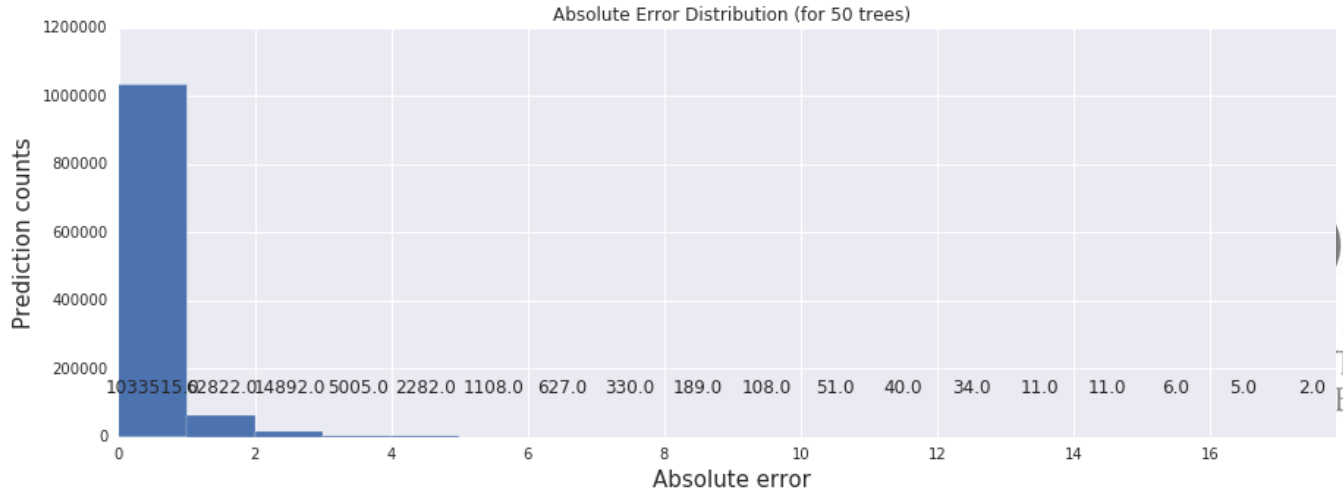
- Which regressor to use?

Model (default HPs)	MSE	MAE
Linear regression	10,73	1,90
Ridge regression	10,72	1,89
SGD regression	15,36	2,74
Bayesian Ridge	10,75	1,90
Logistic regression	/// too computationally	Consuming //////
Random Forest (RF) regression	0,51	0,30

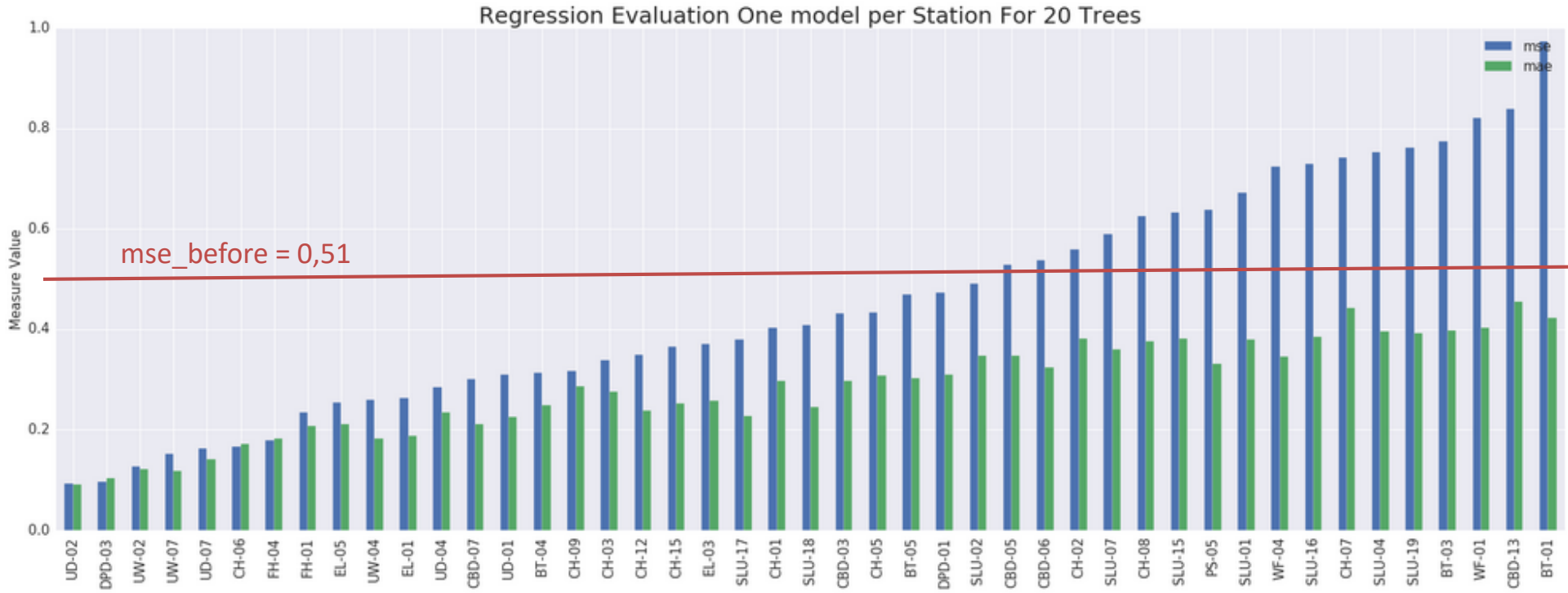
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

- Choice of hyperparameters: which number of trees to use?





One model per station



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

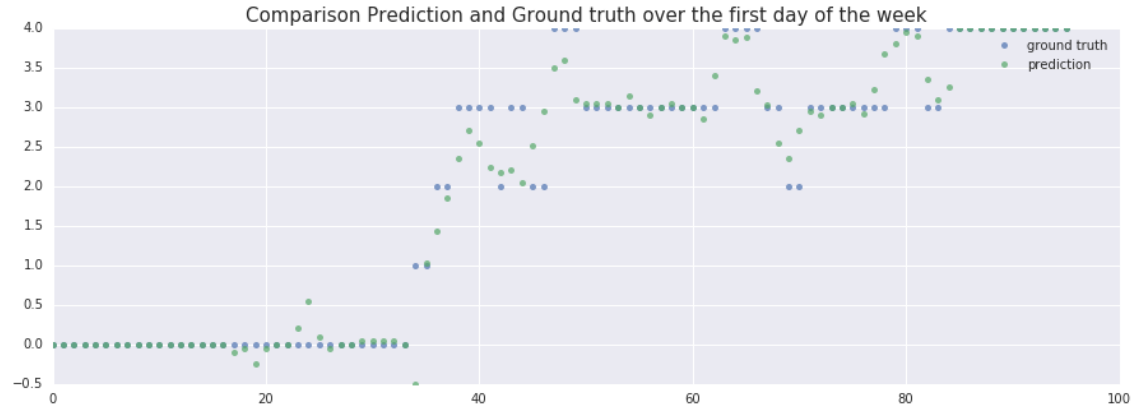
Practical application for Pronto

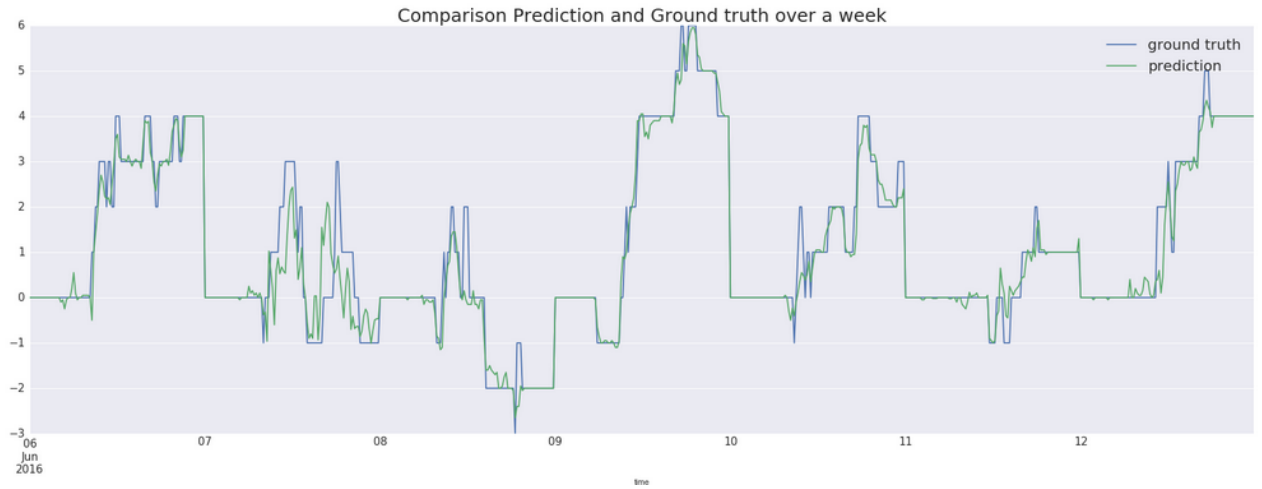
We are the weekend and given the weather forecast, what are the predictions of demand for next week at a given station?

station_id = SLU-01

- **Partially** trained during the week predicted

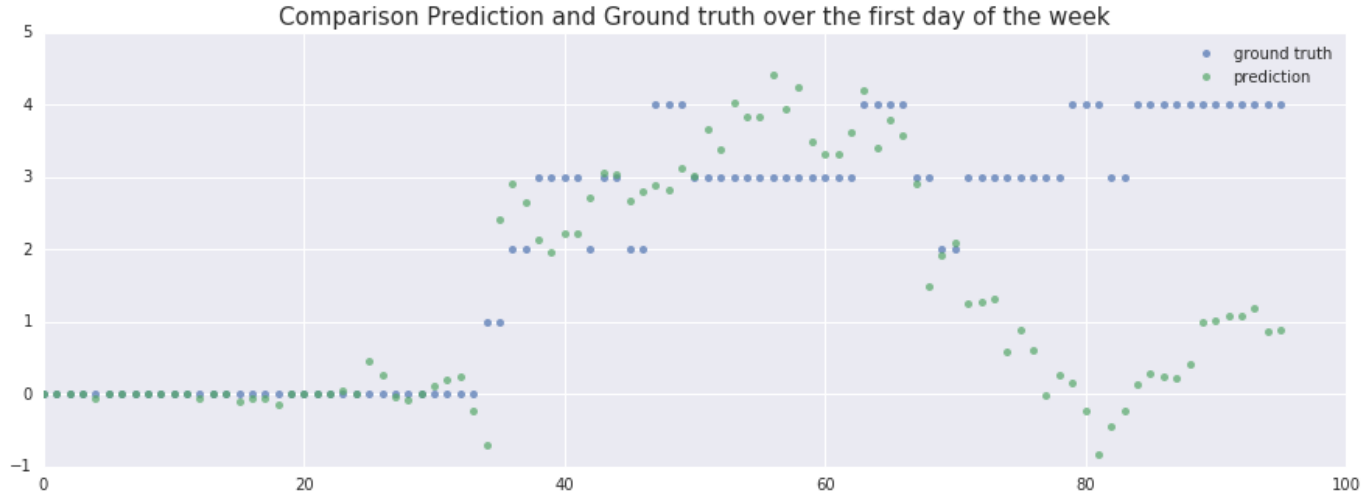
```
Split: 32928 testing and 32928 training samples
mse: 0.2346
mae: 0.2422
```

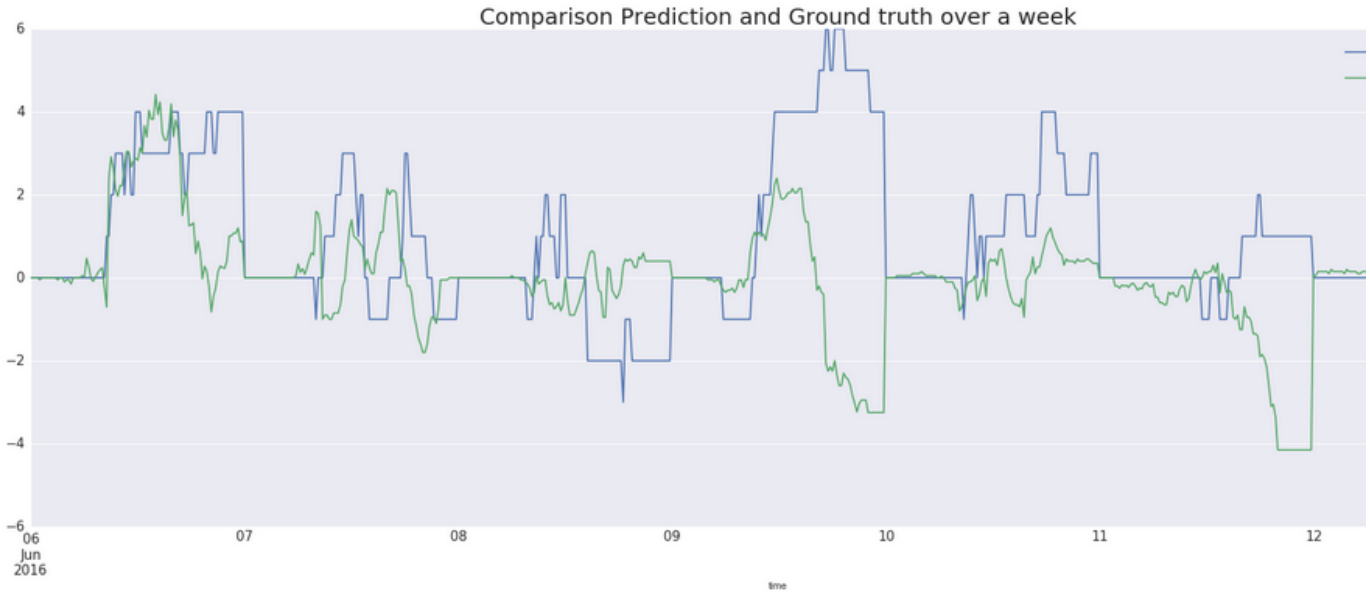




- **Not** trained during the week predicted

```
Split: 32592 testing and 32592 training samples
mse: 6.1937
mae: 1.5991
```





FEDERALE DE LAUSANNE

Conclusion

- Exploration of users trends allowed an efficient exploitation of data with PCA
- RF regressions on the daily bikes variation provided encouraging and quite precise predictions
- However this prediction tool does not seem perfectly optimized for online applications

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE