CASE STUDY 3

POS SALE PERFORMANCE AND SURROUNDING ATTRIBUTES

Cheng CHEN

Outline

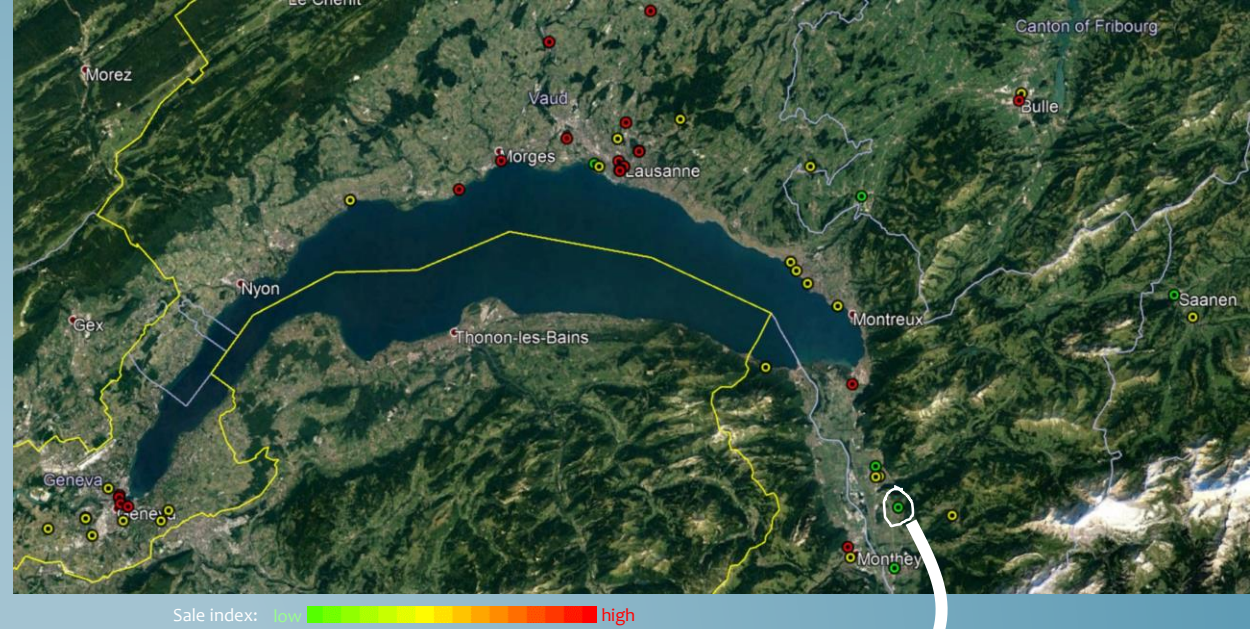
0. INTRODUCTION

Data

- **Sale**: the sales volumes of a product at particular POS
- **Surrounding**: information about different amenities (restaurants, shops, beauty salons etc.) that are in the surroundings of each POS

Questions

- **Understanding**: What are important attributes in the surrounding data that impact sales?
- **Prediction**: Can we reasonably predict a POS' sale from its surrounding info?



1.1 SURROUNDING DATA EXPLORATION

- The original surrounding data (from .json file) is nested, with information mostly on each individual amenity
- We transform this nested data into a "flat" structure, with information aggregated on level of POS (descriptor of each POS)

Original data structure

place_id
name
formatted address
address comp. (**country**, **admin. level 1/2**, **locality**, postal code)
longitude
latitude
website
phone_number
opening hours
rating
user_rating_total
reviews (rating, author name, language, text…)
icon
types

Transformed data structure (per POS)

store_id	nb. station	nb. bar	…	total amenity	mean_rating	mean_rating _total	sum_rating	sum_rating _total
3933	2	5		155	4.2	32	174	812
425	3	9		93	3.9	12	555	918
3622	0	3		64	4.5	152	522	2048
…								

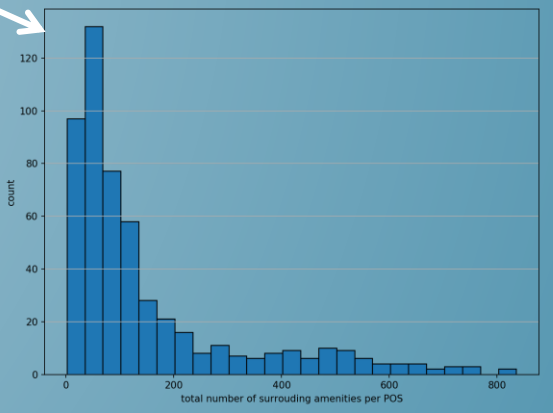
Number of amenities per category for a POS

rating statistics for all amenities of a POS

…	canton	commune	locality	longitude	latitude
	VD	Lausanne	Lausanne	6,6298	46,5165
	GE	Genève	Chêne-Bourg	6,1444	46,2101
	BS	Basel-Stadt	Basel	7.5903	47.5772

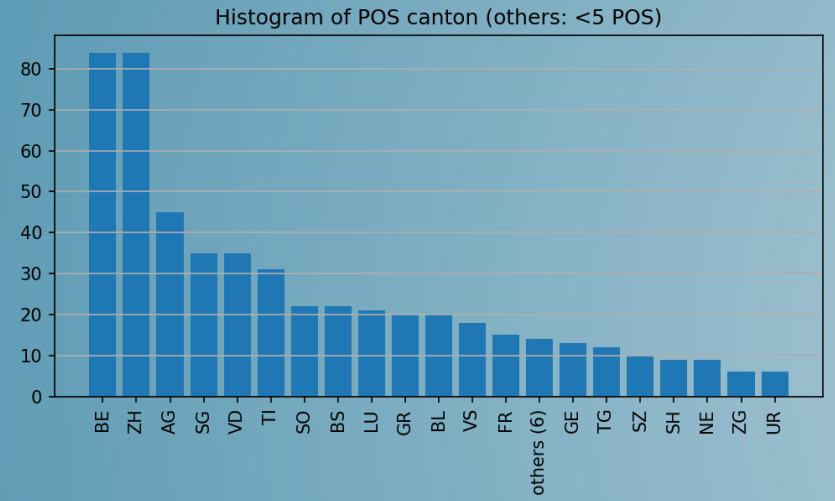
Most predominant value within all amenities of a POS

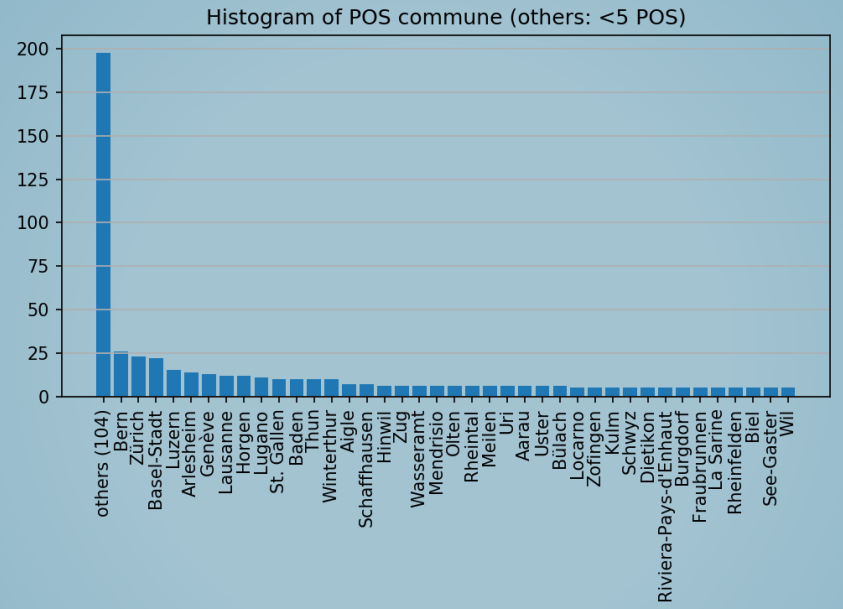
Average longitude/latitude of all amenities of a POS

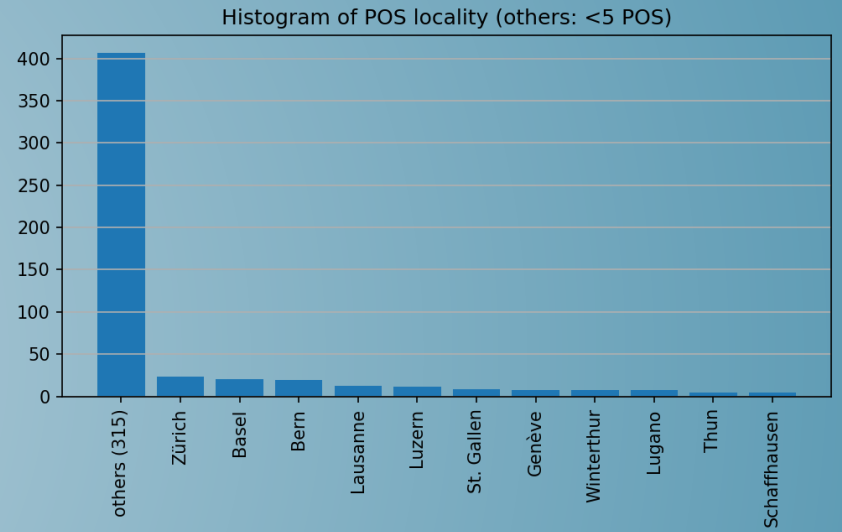


1.1 SURROUNDING DATA EXPLORATION

- There are many different canton/commune/localities, many of which only have very few POS. With so few number, it is not desirable to analyze data in such detailed level
- Therefore, we only keep the major canton/commune/localities with many POS. The others are set to one level: “others”



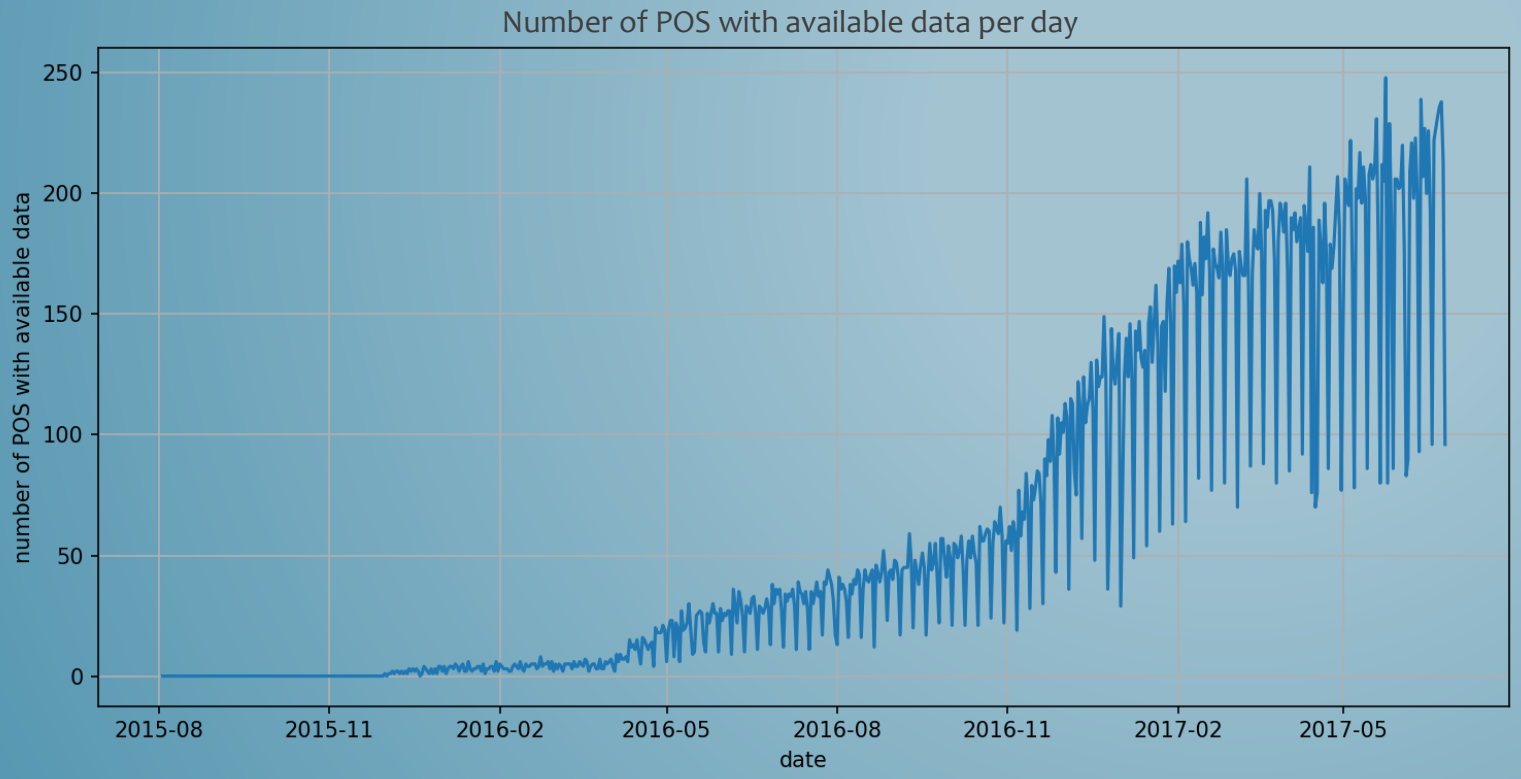




Field	Original nb. of unique values	Set to “others” if	Final nb. Of unique values
canton	26	less than 5 POS in a same canton	21
commune	142	less than 5 POS in a same commune	40
locality	325	less than 5 POS in a same locality	12

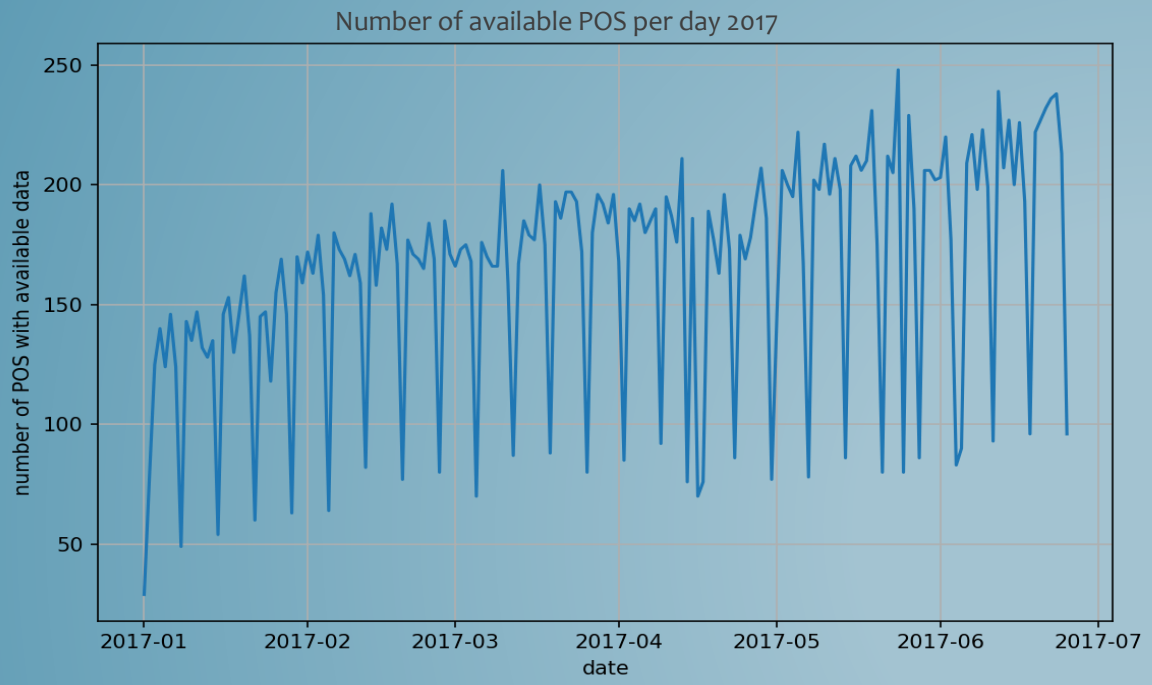
1.2 SALE DATA EXPLORATION

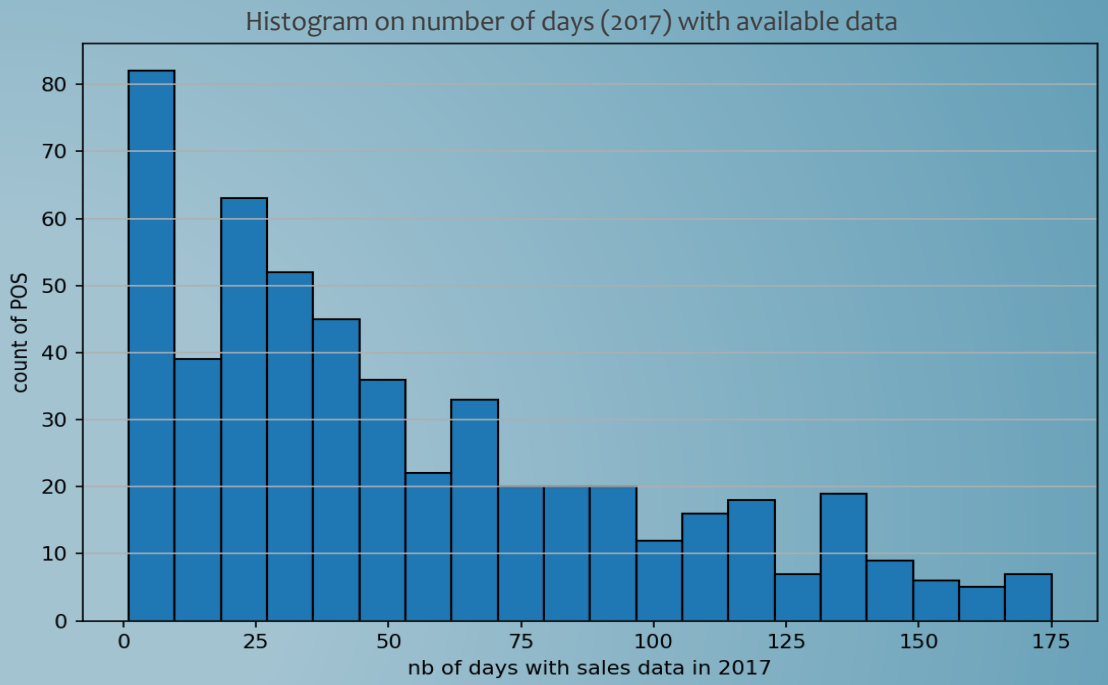
- The original sales data was hourly based sale volumes per POS.
- As a first step, we aggregate the sum of sales per day, as we probably do not want to go to details by hours
- At first glance, there are many missing sales values (na's). As we do have zeros in the data (which naturally means zero sales), we assume that the na's mean unavailable data (e.g. the POS was not open, data was not registered, etc… but not necessarily zero-sale)



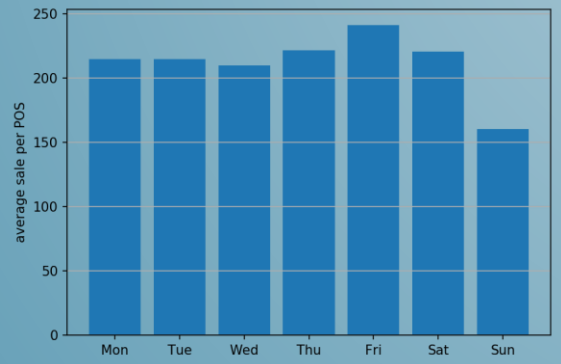
- As shown in the left, there is a clear increasing number of available POS since 2015
- We only take the sales data in 2017, as
  - There are too few POS before 2017
  - The surrounding data is probably extracted recently
  - We are probably interested in the latest sale performance

1.2 SALE DATA EXPLORATION





- We see clearly the weekly pattern and holiday effect (April and May)



- Some POS are available in many days, some are only available for very few days
- We remove POS with less than 5 days available data

1.2 SALE DATA EXPLORATION

Construction of target variable for sale performance

Option 1. Take the sum of sales for each POS
Not desirable, because each POS may be available at different number of days. For example:

POS A	POS B
available 150 days	available 80 days
Total sale: 10,000	Total sale: 10,000

We would like assign POS B a higher ranking. Reasoning: if it were available on more days, it would make more sales than POS A. The sales performance ranking should reflect the intrinsic sales potential, irrespective of opening days

Option 2. Take the mean of sales for each POS, on the days where this POS were available
Not desirable, because there is still day effect. Example:

POS A	POS B
available 100 days	available 100 days
Mostly Sun/Mon/Tue	Mostly Thu/Fri/Sat
Total sale: 10,000	Total sale: 10,000

In this case, even if the two POS were available for the same number of days and made the same sale volumes, we would like to assign POS A to a higher sale performance ranking. Because the days when POS A were open there are generally less sales. If A were open on Fri/Sat instead of Mon/Tue, it probably would made more sales.

Option 3 (our proposal). Assuming that:

- $x_p(t), p = 1 \dots P, t = 1 \dots T$ is the sale volume of POS p on day t.
- $\phi_p(t) = \begin{cases} 1, & \text{if POS } p \text{ was available on day } t \\ 0, & \text{if POS } p \text{ was unavailable on day } t \end{cases}$
- $\tau_p = \sum_{t=1}^{T} \phi_p(t)$ is the number of days where POS p was available
- $\rho_t = \sum_{p=1}^{P} \phi_p(t)$ is the number of POS available on day t

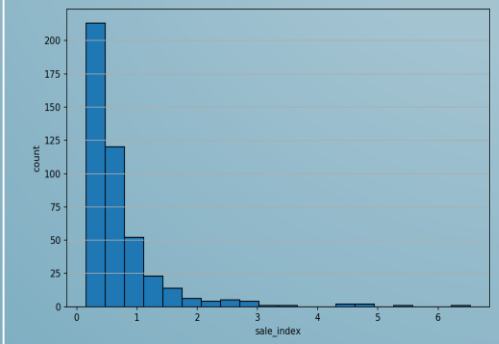
- Calculate the average sale of all POS available on each day: $\bar{x}_t = \frac{\sum_{p=1}^{P} x_p(t)}{\rho_t}$
- For each POS p
  - Calculate total sale volumes of this POS: $\sum_{t=1}^{T} x_p(t)$
  - Calculate the sum of $\bar{x}_t$ **on the days where POS p was available** $\sum_{t=1}^{T} (\bar{x}_t \times \phi_p(t))$
  - The sale_index is the ratio between the two:

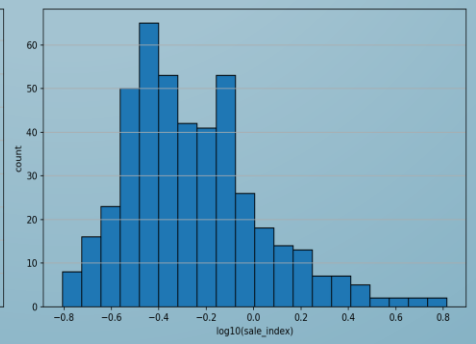
$$sale\_index = \frac{\sum_{t=1}^{T} x_p(t)}{\sum_{t=1}^{T} (\bar{x}_t \times \phi_p(t))}$$

  - As this ratio is log-normal like, we take the log

A simplified example

	day 1	day 2	day 3	day 4	option 1	option 2	option 3
POS A	100	0	n.a.	300	***400***	***133***	***-0,24***
POS B	0	n.a	500	300	***800***	***267***	***-0,05***
POS C	n.a	100	300	n.a	***400***	***200***	***-0,17***
POS D	300	300	n.a.	n.a	***600***	***300***	***0,18***
AVG	200	200	400	300			





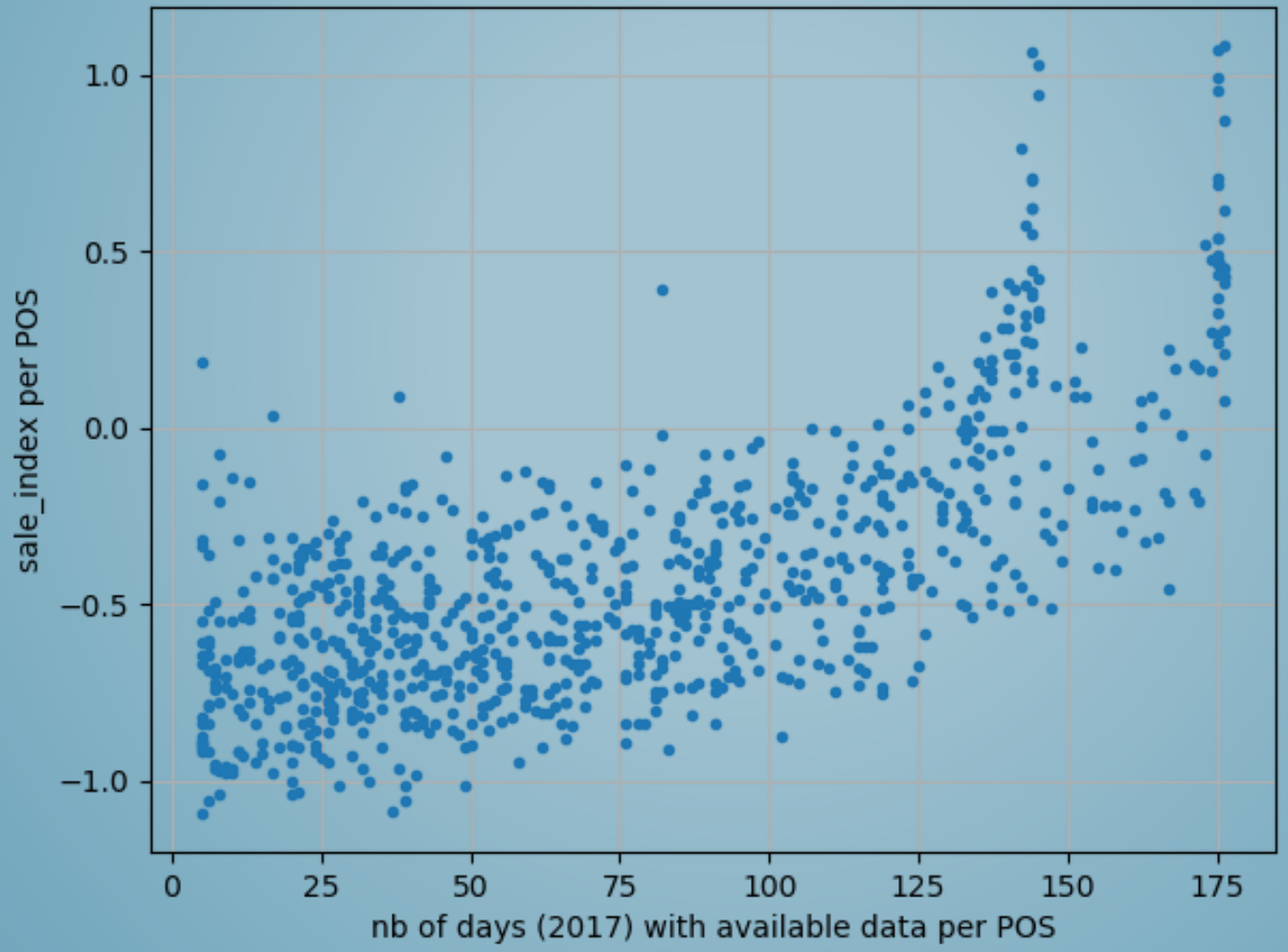
Advantages

- Insensitive to number of available days of each POS
- Insensitive to global day-to-day sale variance
- Easy to interpret
  - the higher the better
  - positive = performs better than average
  - negative = performs worse than average
- Naturally scaled (no normalization needed)

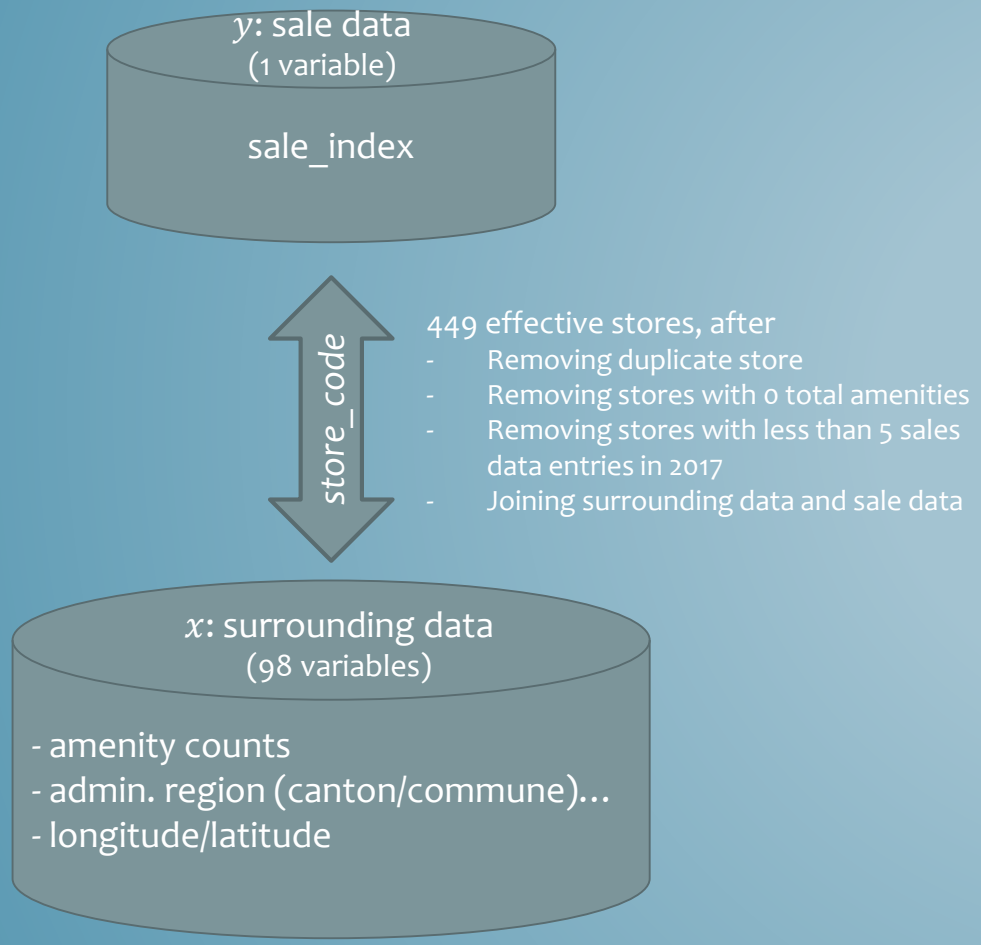
1.2 SALE DATA EXPLORATION

- Nb of days with available data per POS seems correlated to sale index
- POS with highest sale_index also has highest number of days with available data

Needs better understanding of data "unavailability" (POS closed? Sale data not registered?)



2. DATA ANALYSIS



Model / Algorithm selection

We need a regression algorithm which
- Provides some degree of "understanding" and not only black box prediction
- Can deal with both continuous and categorical variables
- Is not too data-hungry (we have a very small dataset of hundreds of POS)

ANN / deep learning is out (black-box + data hungry)

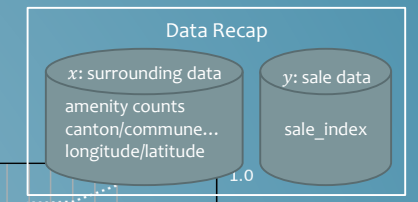
Algorithms which may be considered
- Linear regression
- Lasso / elastic net
- Tree based methods and Random Forest
- Gaussian Process regression
- K nearest neighbors
- Mahalanobis distance learning for knn (developed by myself [1])
- Feature selection by matrix trace ratio optimization (developed by myself [2])
- ……
- We adopted **Random Forest** algorithm in this case

[1] C. Chen, Y. Zhuang, F. Nie, Y. Yang, F. Wu, J. Xiao. Learning a 3D human pose distance metric from geometric pose descriptor. IEEE Transactions on Visualization and Computer Graphics (TVCG), 17(11): 1676-1689, 2011

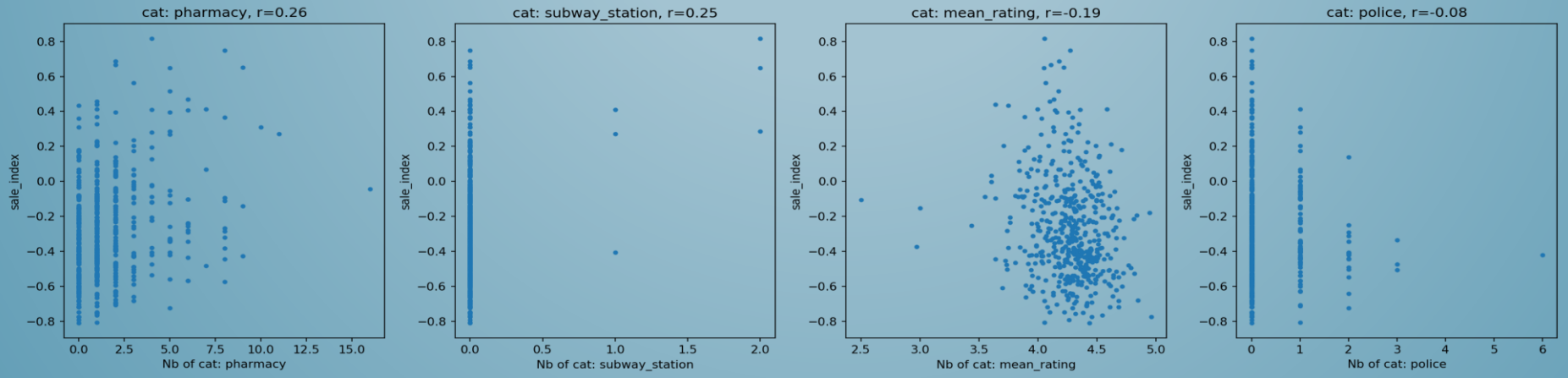
[2] C. Chen, G. Zheng. Fully Automatic Segmentation of AP Pelvis X-rays via Random Forest Regression with Efficient Feature Selection and Hierarchical Sparse Shape Composition. Computer Vision and Image Understanding (CVIU), 2014

2.1 SIMPLE CORRELATIONS: SALES VS. AMENITY COUNTS

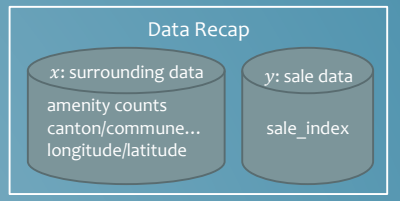
Before doing more advanced analysis, let us first look at the pretty simple pairwise correlations between X and y

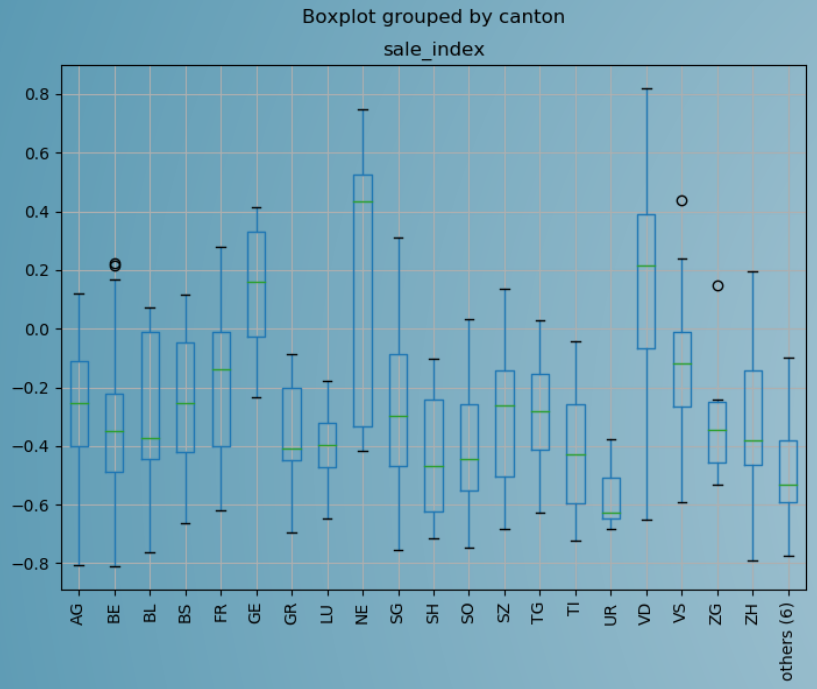


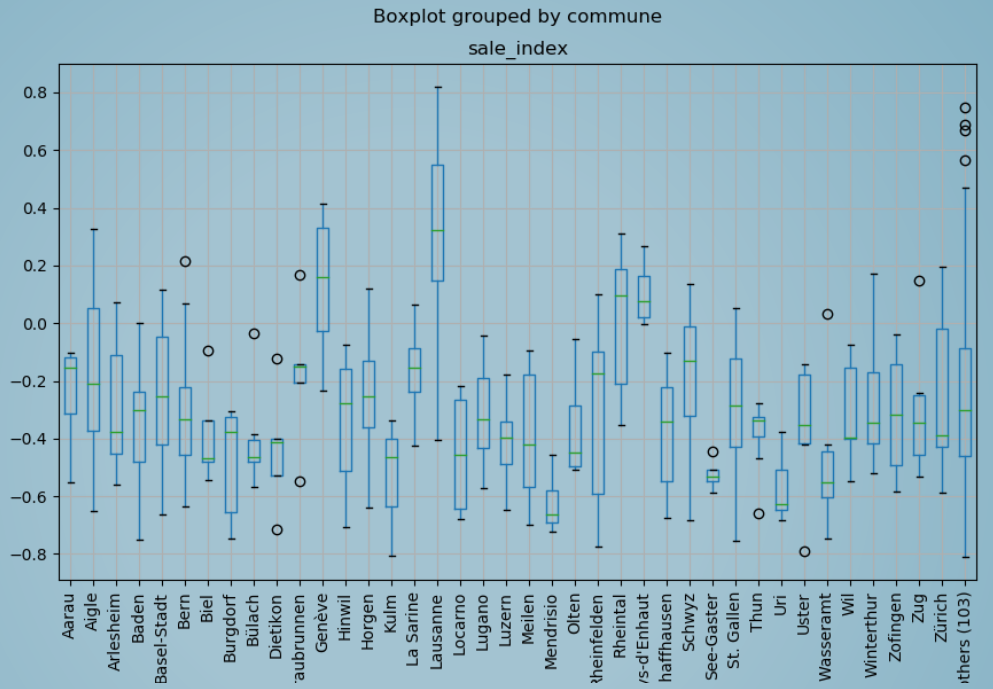
Some amenity categories seems to have some correlation with sale_index (e.g. pharmacy, subway_station, bakery, café…), but generally speaking, the effect is not strong

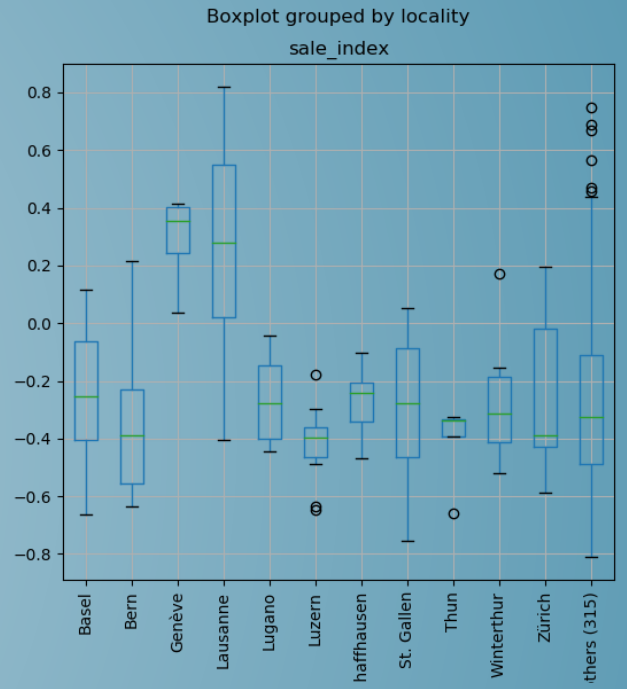


2.1 SIMPLE CORRELATIONS: SALES VS. ADMIN. REGION









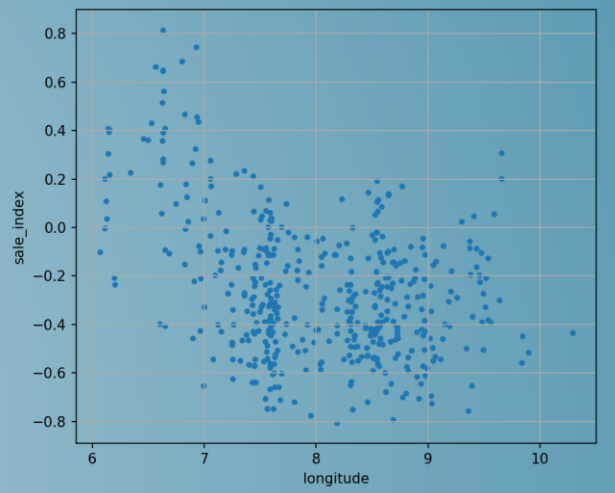
Sale is generally higher in some regions such as VD/GE/NE, Lausanne, Genève

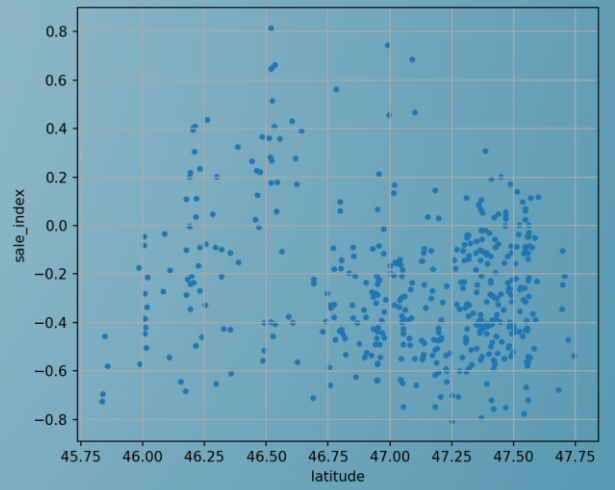
2.1 SIMPLE CORRELATIONS: SALES VS. LONGITUDE/LATITUDE

Data Recap

x: surrounding data	y: sale data
amenity counts canton/commune… longitude/latitude	sale_index

Sale index: low high





The sale is generally higher in western Switzerland (Suisse Romande)

2.2 RANDOM FOREST MODELING

Data Recap

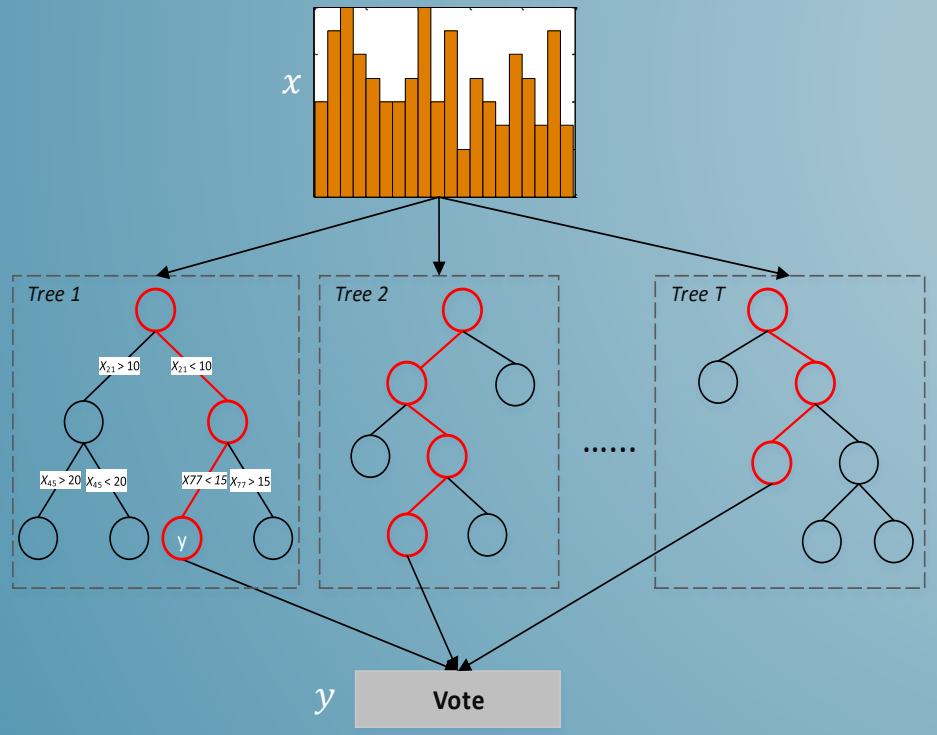
x: surrounding data
amenity counts
canton/commune…
longitude/latitude

y: sale data
sale_index

- Random Forest (RF) is a type of ensemble method in machine learning for classification/regression problems.
- Instead of training a strong learner, the idea is to train many weak learners (trees) in a **random** and **de-correlated** manner. The classification or regression result is aggregated from all trees

Advantages of RF

- Can easily learn non-linear relationships
- Naturally deals with (mix of) continuous and categorical variables
- Robust (adding trees will not cause overfitting)
- Can deal with $n < p$ problems
- Has a natural way to calculate the feature importance



Each tree is trained separately on a **random** training set generated by bagging (**b**ootstrap **agg**regat**ing**) from the full training set.

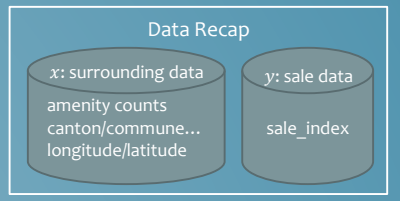
To train each node in each tree, a **random** subset of features are considered, and a best split is determined by a certain criteria (e.g. information gain)

For each tree, the full training set is divided into

- in-bag training samples
- out-of-bag samples (oob samples)

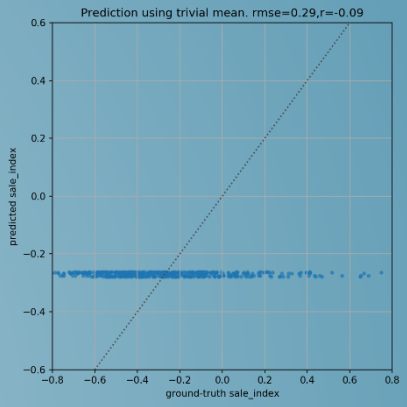
OOB samples can be seen as a validation set for each tree, and provide a convenient way to estimate the validation error of individual trees

2.2 RANDOM FOREST MODELING – PREDICTION RESULTS



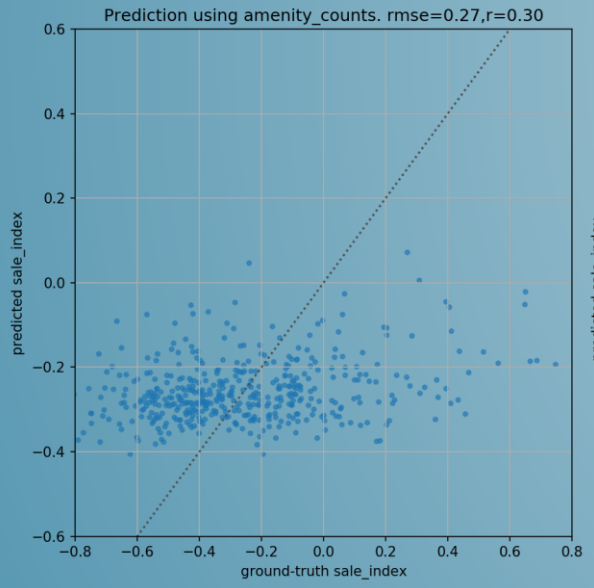
Why are we interested in the **prediction** accuracy, while our primary goal is to **understand** the factors which effects sale?

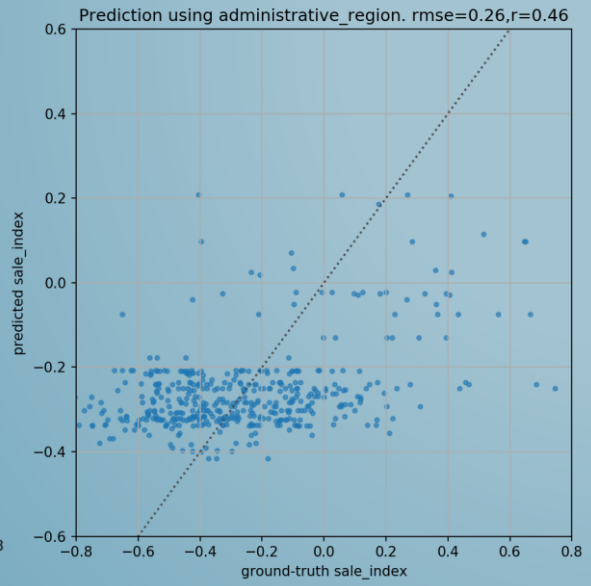
- A model which "understands" well the data should also "predicts" well on unseen data

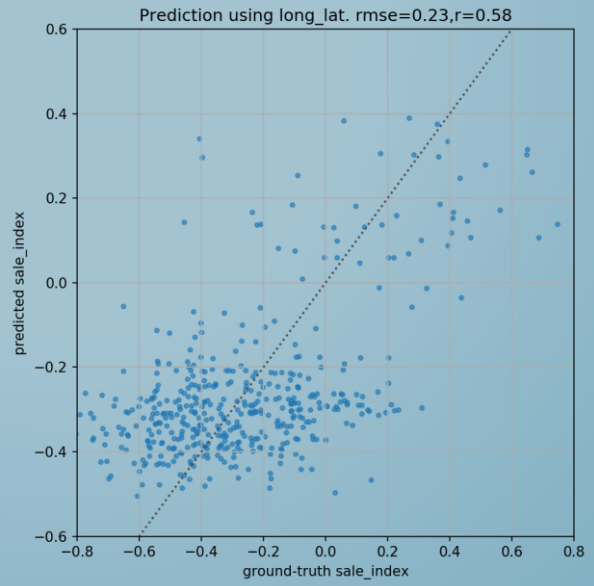


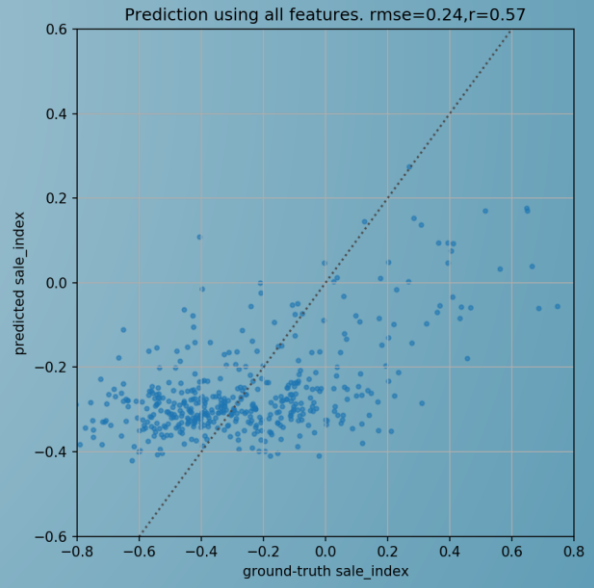
The prediction evaluation was done on **5-fold cross validation**
Metric: RMSE (Root Mean Squared Error)

It seems that using longitude/latitude information we achieve the best prediction results

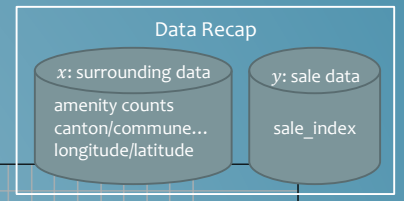


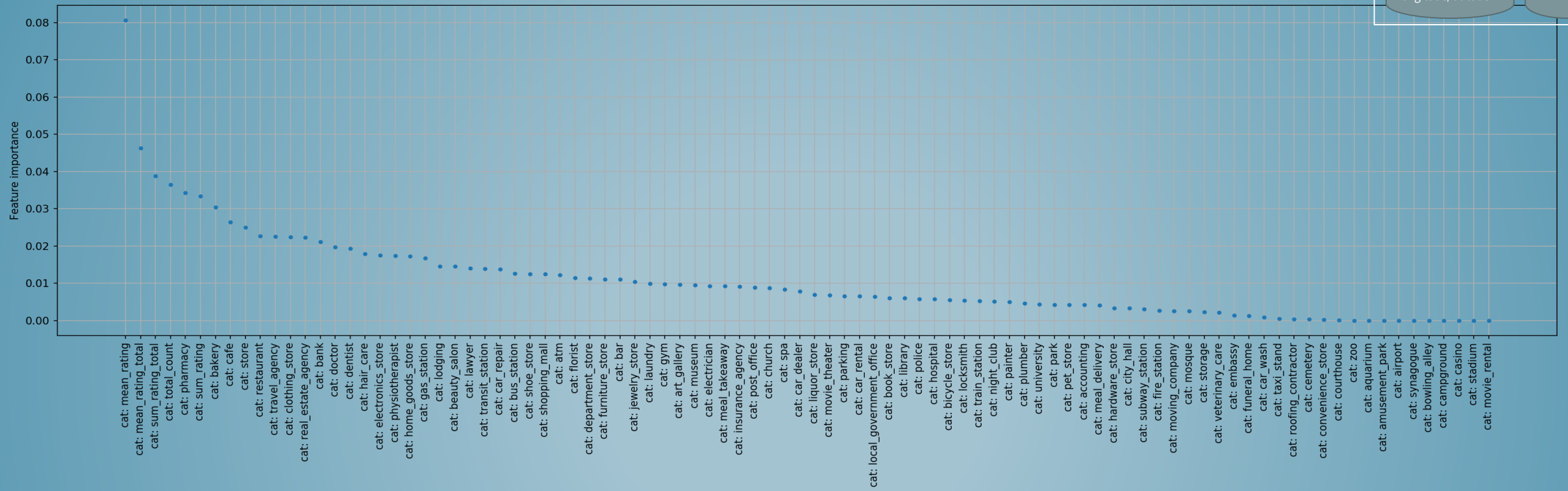


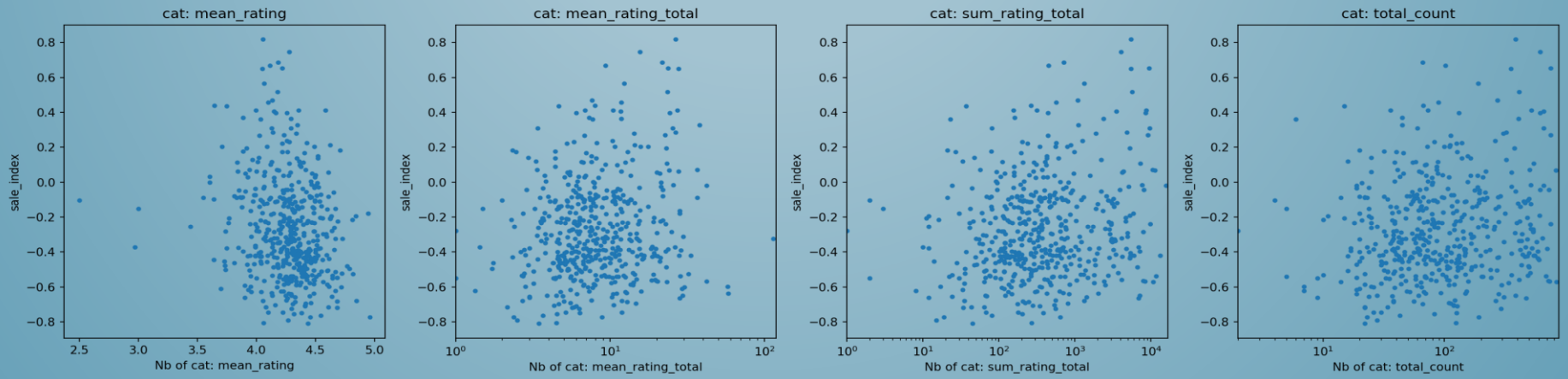




2.2 RANDOM FOREST MODELING – FEATURE IMPORTANCE







2.2 RANDOM FOREST MODELING – FEATURE IMPORTANCE

Note: how feature importance is calculated in Random Forest

1. Train a tree using the in-bag samples, note down the average prediction error (MSE) on OOB samples of this tree
2. Now randomly permute the values of the i^{th} variable on OOB samples, run these samples down the tree, and note down the new average prediction error (MSE) of these OOB samples on this tree
3. Calculate the change of MSE when the ith variable is permuted (MSE in step 2 minus MSE in step1)
4. Repeat 1-3 on all trees. The average of MSE increase over all trees is then the importance of the i^{th} variable

	Tree 1	**Tree 2**	**…**	**Tree *t***	**…**	**Tree *T***
Intact oob error	e_1^0	e_2^0		e_t^0		e_T^0
oob error when i^{th} var. permuted	e_1^i	e_2^i		e_t^i		e_T^i
Error increase	$\delta_1^i = e_1^i - e_1^0$	$\delta_2^i = e_2^i - e_2^0$		$\delta_t^i = e_t^i - e_t^0$		$\delta_T^i = e_T^i - e_T^0$

e_0^t: the prediction MSE on OOB samples on tree t, when the data is intact
e_i^t: the prediction MSE on OOB samples on tree t, when the values of i^{th} variable on the OOB samples are permuted

Importance of i^{th} feature: $\frac{1}{T}\sum_{t=1}^{T}(e_t^i - e_t^0)$

3. CONCLUSIONS

- **Amenities information has some effects on sale performance**
  - *Factors positively correlating with the sale*
    - *Ratings: mean/total number of user ratings*
    - *Total number of amenities (all categories confounded)*
    - *Number of some particular categories of amenities (pharmacy, café, bakery, store, restaurant…)*
  - *Generally speaking, the effect is not strong and does not allow accurate sale prediction of unseen POS*

- **Geographical data has stronger correlations with sale**
  - *Sale index per POS is generally higher in the west of Switzerland (Suisse Romande)*
  - *Sale index per POS is generally higher in some cantons/communes (e.g. VD, GE, NE, Lausanne, Genève)*

- **Next steps**
  - *Better understanding of unavailable data (POS closed? Data missing?)*
    - *Nb of days with available data per POS seems correlated to sale index (see right figure)*
  - *More detailed look into individual amenities (e.g. opening time, detailed reviews…)*
  - *External data from geographical information (e.g. population, commute pattern…)*
  - *Density/proximity between POS's*