Corporación Favorita Grocery Sales Forecasting

Presented by
Mahmood Yousefi

Forecasting the unit sales for each item

- Aim is forecasting:
  - How many items will be sold in future?

- Challenges
  - Understand question and data
  - Identify influential attributes
    - Which data sources are potentially useful?
  - Can we create new attributes?
  - Which Machine Learning algorithm is a good practice?
  - How to measure our success?

Data

- Unit sales is not in test set and we are going to predict it

train

date	store_nbr	item_nbr	unit_sales	onpromotion
1/01/2017	25	115267	1.098612289	FALSE
1/01/2017	25	115891	2.564949357	FALSE
1/01/2017	25	158788	0.693147181	FALSE
1/01/2017	25	164647	0.693147181	FALSE
1/01/2017	25	222879	2.197224577	FALSE
1/01/2017	25	222975	1.386294361	FALSE
1/01/2017	25	226126	0.693147181	FALSE
1/01/2017	25	261053	1.609437912	FALSE
1/01/2017	25	305345	0.693147181	FALSE
1/01/2017	25	318935	0.693147181	TRUE

stores

store_nbr	city	state	type	cluster
1	Quito	Pichincha	D	13
2	Quito	Pichincha	D	13
3	Quito	Pichincha	D	8
4	Quito	Pichincha	D	9
5	Santo Domingo	Santo Domingo de los Tsachilas	D	4
6	Quito	Pichincha	D	13
7	Quito	Pichincha	D	8
8	Quito	Pichincha	D	8
9	Quito	Pichincha	B	6

items

item_nbr	family	class	perishable
96995	GROCERY I	1093	0
99197	GROCERY I	1067	0
103501	CLEANING	3008	0
103520	GROCERY I	1028	0
103665	BREAD/BAKERY	2712	1

oil

date	dcoilwtico
1/01/2013	
2/01/2013	93.14
3/01/2013	92.97
4/01/2013	93.12
7/01/2013	93.2

transactions

date	store_nbr	transactions
1/01/2013	25	770
2/01/2013	1	2111
2/01/2013	2	2358
2/01/2013	3	3487
2/01/2013	4	1922

Holiday_events

date	type	locale	locale_name	description	transferred
2/03/2012	Holiday	Local	Manta	Fundacion de Manta	FALSE
1/04/2012	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	FALSE
12/04/2012	Holiday	Local	Cuenca	Fundacion de Cuenca	FALSE
14/04/2012	Holiday	Local	Libertad	Cantonizacion de Libertad	FALSE

Profile, clean and transform data

- Train set includes 125,497,040 samples, from 2013 to mid of 2017
  - 23,808,260 sample from 1st Jan 2017 onward
- unit_sales zero and negative
  - Max value is 20748.0
- Less than 5% are onpromotion
- Items table and Store table are highly infromative
  - We can put other tables (e.g. oil price) for future work
- Items marked as perishable have a score weight of 1.25; otherwise, the weight is 1.0
  - This weighting is considered for performance evaluation

Weighted Moving Average

- Creating new attributes
  - simple moving average
    - $p = \frac{p_1 + p_2 + \cdots + p_n}{n} = \frac{1}{n}\sum_{i=1}^{n} p_i$
      - Where n is the window size
  - We used weighted moving average
    - weight based on recency of data point in a window
      - Window size is 15
  - Created features:
    - Per day, week, fortnightly and so on
      - Mean, median, min, max, std, mean of difference, mean of decay
  - We end up with 633 features

Supervised machine learning

- Is this problem supervised?
- Why do we not use unsupervised algorithm?

- It is a regression problem
- Data has hierarchy
  - E.g., today, in Perth in North store, item number 10 that is not on sale, Will it be sold?
- Weight the input samples based on perishable weighting
- Gradient boosting (xgboost) as a regressor
  - Decision tree
  - Ensemble algorithm
  - Drawback: many hyper-parameters
- Final prediction is based on a window of outcomes

Feature importance

- Implicit feature ranking:
  - mean_140_decay
  - mean_60_decay
  - mean_30
  - median_60
  - no_promo_mean_30
  - last_has_promo_day_in_after_15_days
  - median_30
  - promo_15
  - mean_60
  - item_promo_15
  - no_promo_mean_140
  - has_promo_mean_140
  - item_diff_30_mean

Performance metric

- Regression evaluation metric
- Normalized Weighted Root Mean Squared Logarithmic Error
  - NWRMSLE $= \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} w_i \left(\log(\widehat{y_i}+1)-(\log(y_i+1)\right)^2}{\sum_{i=1}^{n} w_i}}$
- Best result on validation: Validation MSE: 0.347
  - My scheme validation MSE: 0.358

Limitations and future work

- Limitations:
  - Large feature space and over-fitting can be there
  - Results varies based the size of predicting window
- For future
  - Other attributes from table
  - Gather more info e.g. type of promotion
  - Specialize model, that is, a model per store/city/…
  - LSTM

Many thanks

Any questions