Predicting Grocery Sales

Marija Stanojevic
CIS Department, Temple University

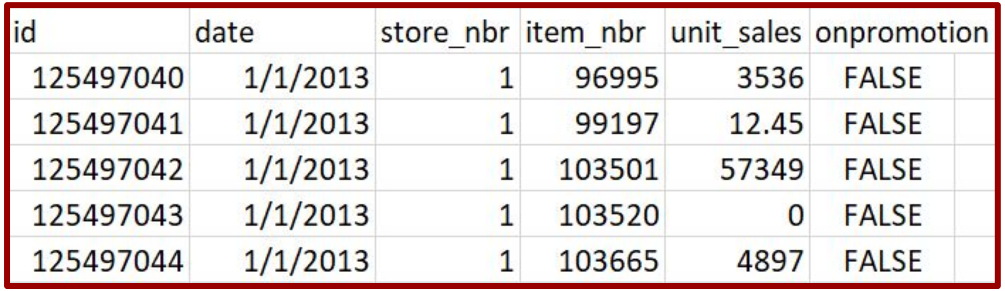
Machine Learning Course, 7th December 2017

Problem definition and dataset

- Kaggle competition: **Corporación Favorita Grocery Sales Forecasting - Nigerian retailer**
- Predict # items that will be bought in a store for date
- Growing retailer - new items and stores with time
- **Training**: 1st January 2013 - 15th August 2017
- **Test**: 16th August 2013 - 30th August 2017
- Different datasets extracted from database
- Prices of items are not given
- Items have different sizes (kg, g, package, l, gallon, l)

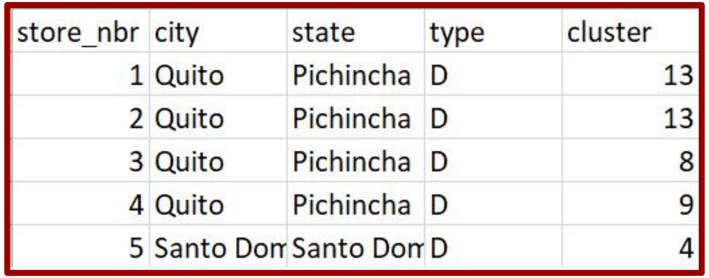
Dataset

Training



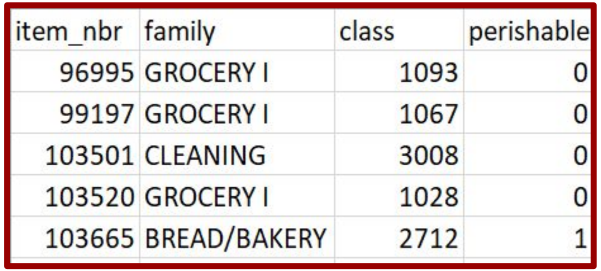
id	date	store_nbr	item_nbr	unit_sales	onpromotion
125497040	1/1/2013	1	96995	3536	FALSE
125497041	1/1/2013	1	99197	12.45	FALSE
125497042	1/1/2013	1	103501	57349	FALSE
125497043	1/1/2013	1	103520	0	FALSE
125497044	1/1/2013	1	103665	4897	FALSE

Store



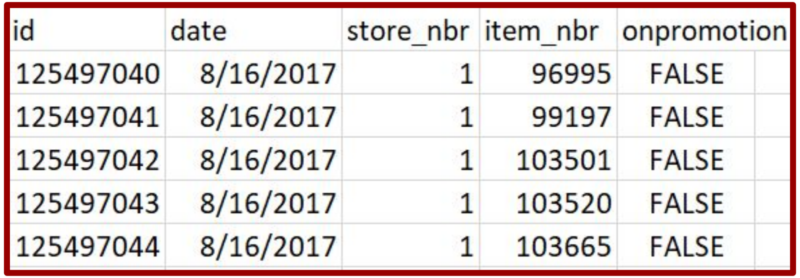
store_nbr	city	state	type	cluster
1	Quito	Pichincha	D	13
2	Quito	Pichincha	D	13
3	Quito	Pichincha	D	8
4	Quito	Pichincha	D	9
5	Santo Dom	Santo Dom	D	4

Item



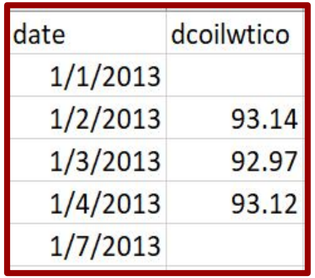
item_nbr	family	class	perishable
96995	GROCERY I	1093	0
99197	GROCERY I	1067	0
103501	CLEANING	3008	0
103520	GROCERY I	1028	0
103665	BREAD/BAKERY	2712	1

Test



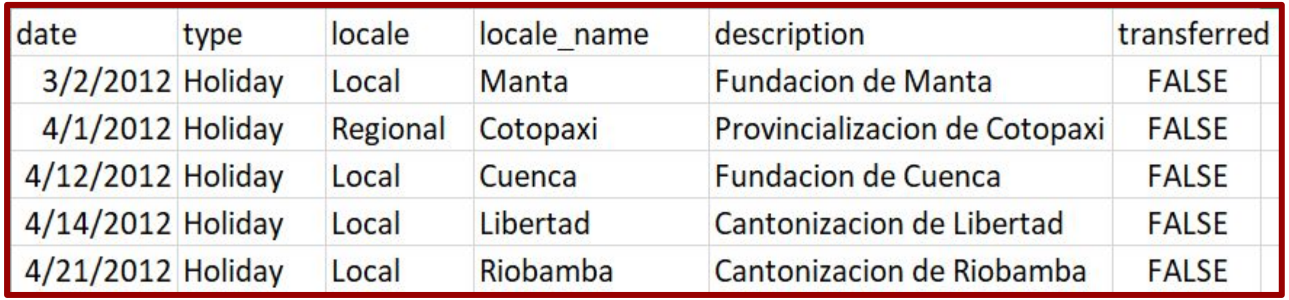
id	date	store_nbr	item_nbr	onpromotion
125497040	8/16/2017	1	96995	FALSE
125497041	8/16/2017	1	99197	FALSE
125497042	8/16/2017	1	103501	FALSE
125497043	8/16/2017	1	103520	FALSE
125497044	8/16/2017	1	103665	FALSE

Oil price



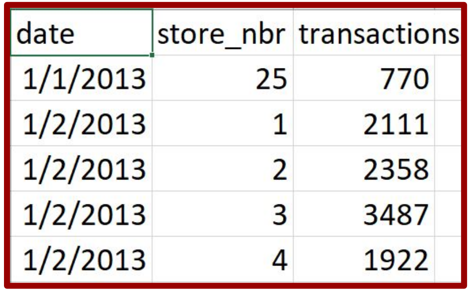
date	dcoilwtico
1/1/2013	
1/2/2013	93.14
1/3/2013	92.97
1/4/2013	93.12
1/7/2013	

Holiday

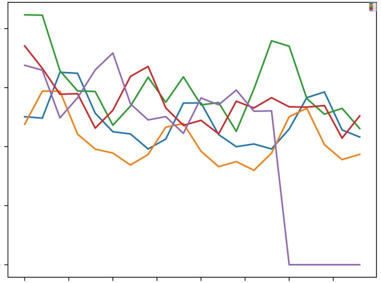


date	type	locale	locale_name	description	transferred
3/2/2012	Holiday	Local	Manta	Fundacion de Manta	FALSE
4/1/2012	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	FALSE
4/12/2012	Holiday	Local	Cuenca	Fundacion de Cuenca	FALSE
4/14/2012	Holiday	Local	Libertad	Cantonizacion de Libertad	FALSE
4/21/2012	Holiday	Local	Riobamba	Cantonizacion de Riobamba	FALSE

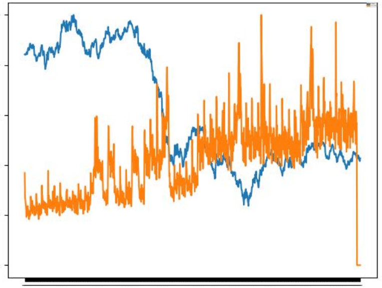
Transaction



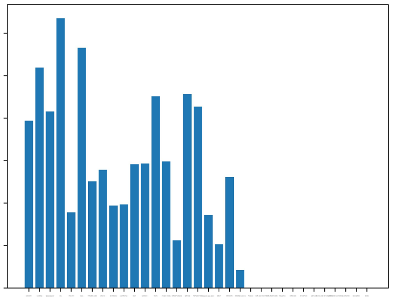
date	store_nbr	transactions
1/1/2013	25	770
1/2/2013	1	2111
1/2/2013	2	2358
1/2/2013	3	3487
1/2/2013	4	1922



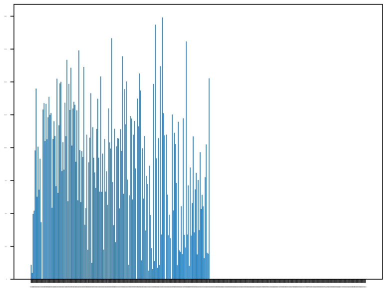
Holiday sales



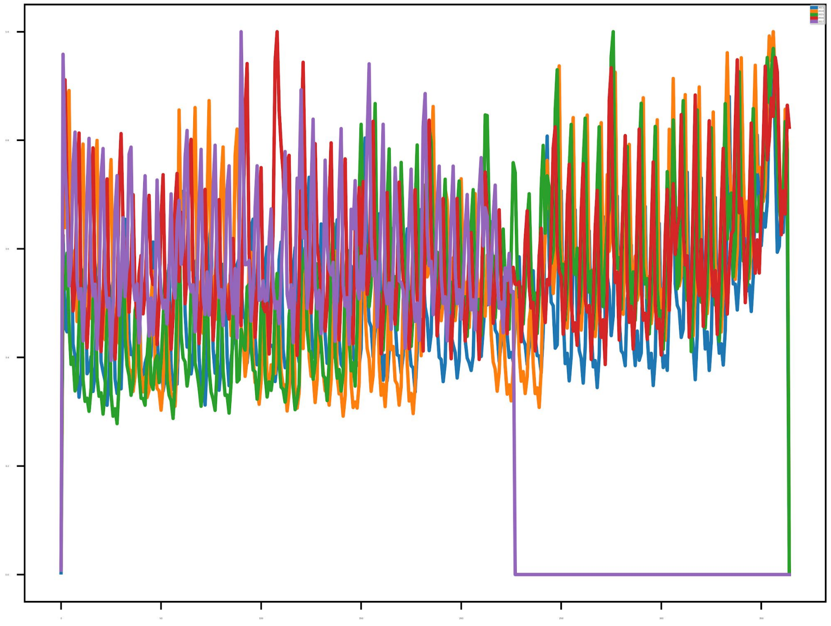
Oil price vs sales



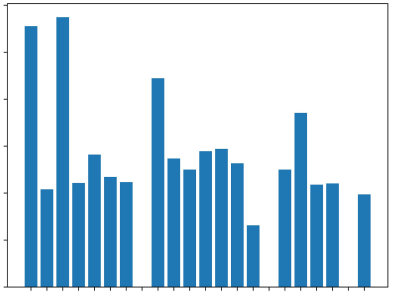
Sales vs item family



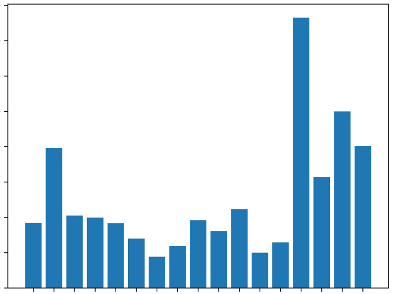
Sales vs item class



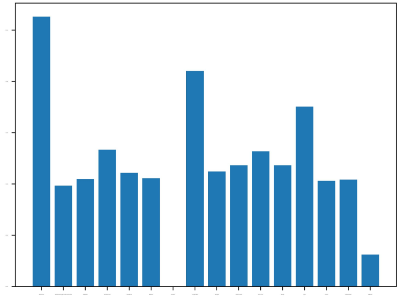
Sales during year; each year one color



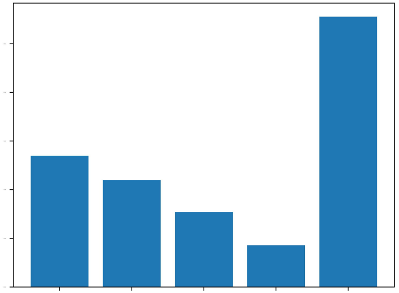
Sales vs city



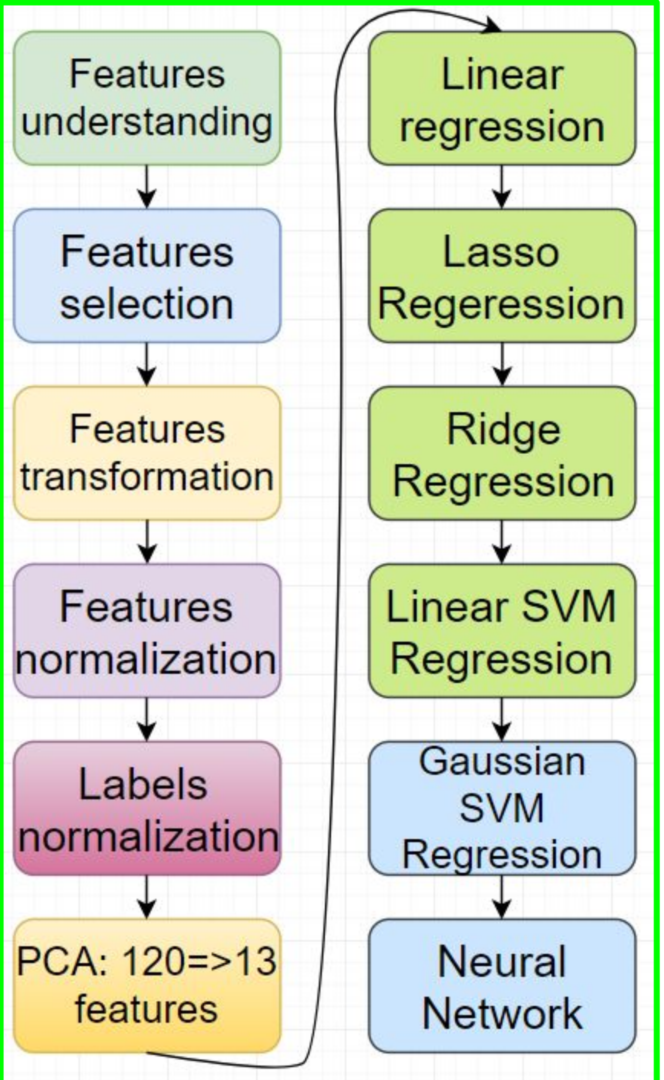
Sales vs store cluster



Sales vs state



Sales vs store type



Features understanding

Features selection

Features transformation

Features normalization

Labels normalization

PCA: 120=>13 features

Linear regression

Lasso Regeression

Ridge Regression

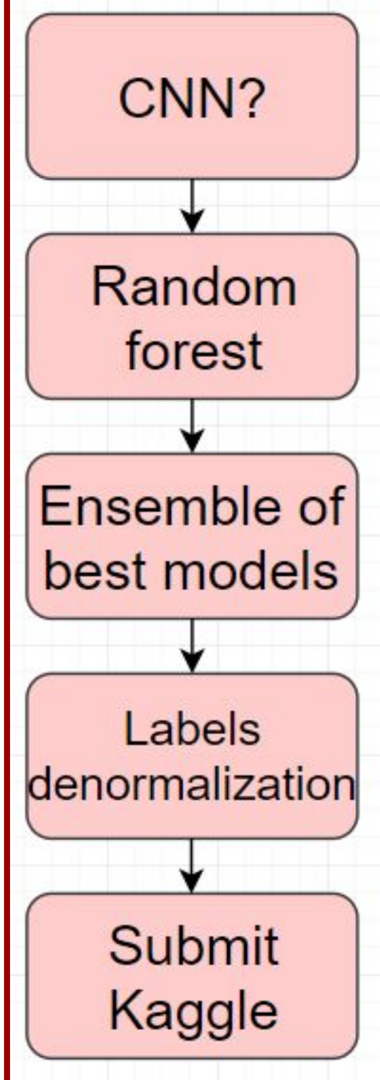
Linear SVM Regression

Gaussian SVM Regression

Neural Network

Methodology

- ❖ Mean-square loss
- ❖ Neural network with 2 and 3 fully connected layers
- ❖ Layers sizes 256, 128, 64
- ❖ Dropout 0.5 after each
- ❖ Implemented RNN, but couldn’t work with it in practice



CNN?

Random forest

Ensemble of best models

Labels denormalization

Submit Kaggle

Results

Model	Parameters	Error
Linear Regression	**-**	0.6797
Lasso Regression	L = any	0.8724
Ridge Regression	L => 1	0.6789
Ridge Regression	L => 20	0.6788
Gaussian SVR	L = 1 , e < 0.02	0.4724
Gaussian SVR	L = 12 , e < 0.02	0.5631

Model	Parameters	Error
Gaussian SVR	L = 1, e > 0.04	0.5844
Gaussian SVR	L = 12, e > 0.04	0.6184
NN-2 layers	sgd, lr: 0.001, 10 epochs, dp:0.5, relu	0.8668
NN-3 layers	sgd, lr: 0.001, 10 epochs, dp:0.5, relu	0.8905
NN-2 layers	adam, lr: 0.001, 10 epochs, dp:0.5, relu	0.4541
NN-3 layers	adam, lr: 0.001, 10 epochs, dp:0.5, relu	0.4308

Conclusion

- In general NN performs best, then Gaussian SVR
- Items have different best models => ensembly
- Linear SVR is too slow
- SGD is converging slowly, so ADAM optimizer is suitable
- Best l.r.=0.001, but doesn’t change much
- Adding additional layers doesn’t change much
- Big difference in accuracy for different items with NN
- Possible last layer activations: linear and relu

Thank you

Questions?