Predict Future Sales

Yijia Chen

Dataset

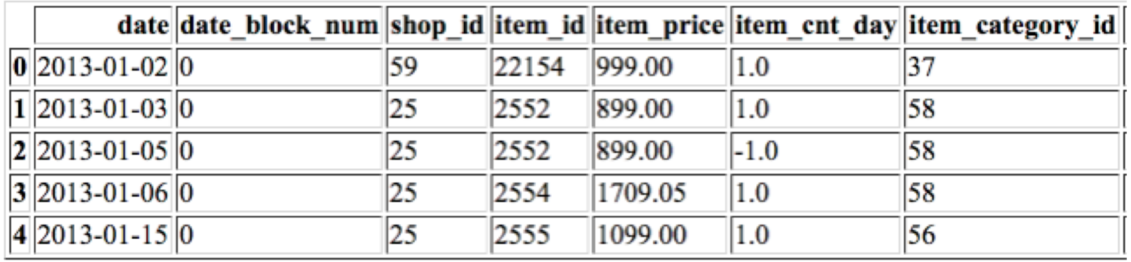
- This is a Kaggle ongoing Competition. In this competition, I will predict total sales for every product and store in next month

  About the Dataset:

- The dataset is provided by 1C Company, one of the largest Russian software firms. The dataset includes train and test dataset, as well as information about each store, category and product.

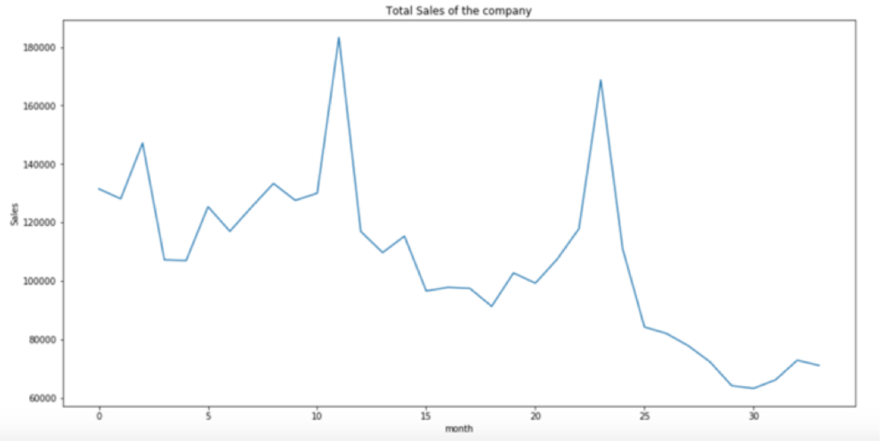
Dataset

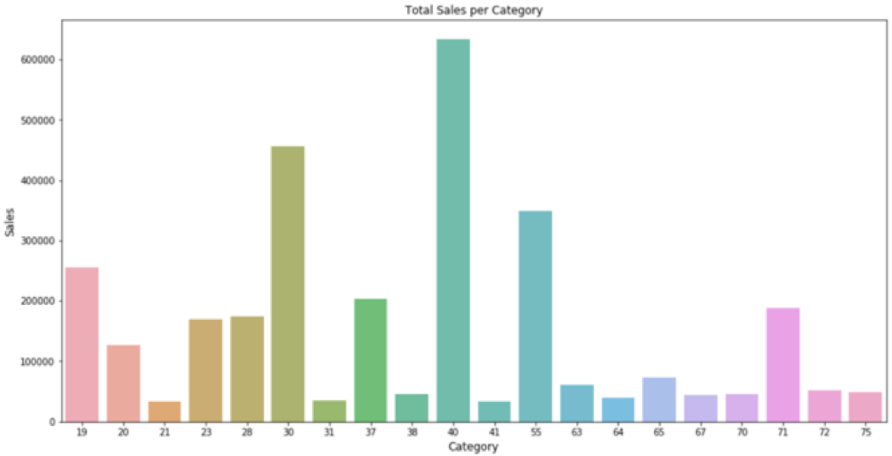
- shop_id - unique identifier of a shop
- item_id - unique identifier of a product
- item_category_id - unique identifier of item category
- Item_cnt_day - sales
- date_block_num - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33



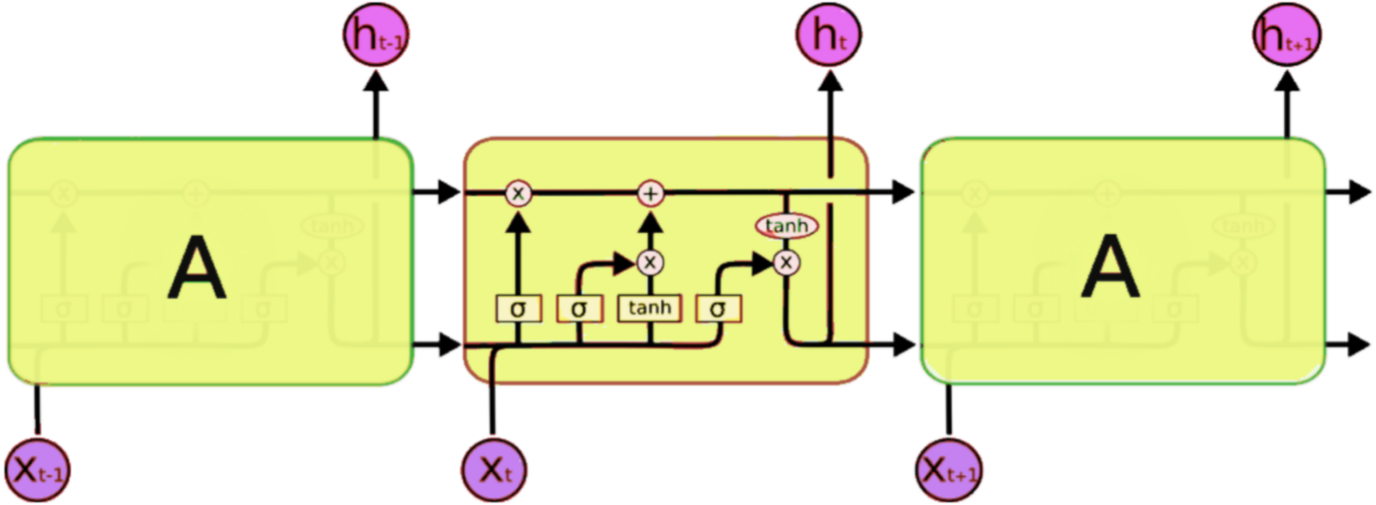
	date	date_block_num	shop_id	item_id	item_price	item_cnt_day	item_category_id
0	2013-01-02	0	59	22154	999.00	1.0	37
1	2013-01-03	0	25	2552	899.00	1.0	58
2	2013-01-05	0	25	2552	899.00	-1.0	58
3	2013-01-06	0	25	2554	1709.05	1.0	58
4	2013-01-15	0	25	2555	1099.00	1.0	56

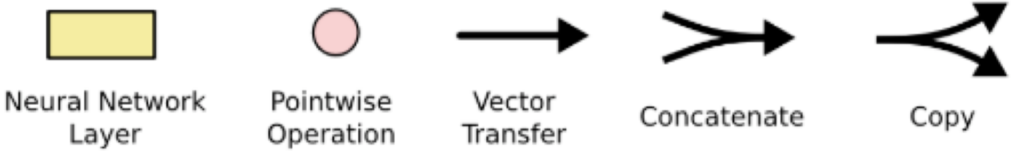
EDA


Total Sales of the company
Sales
180000
160000
140000
120000
100000
80000
60000
0
5
10
15
20
25
30
month


Total Sales per Category
Sales
600000
500000
400000
300000
200000
100000
0
19
20
21
23
28
30
31
37
38
40
41
55
63
64
65
67
70
71
72
75
Category

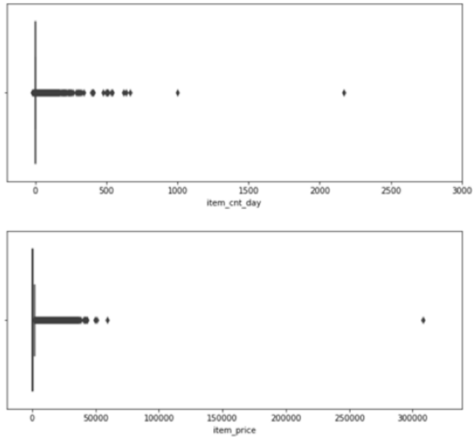
Long Short Term Memory





Data Preprocessing

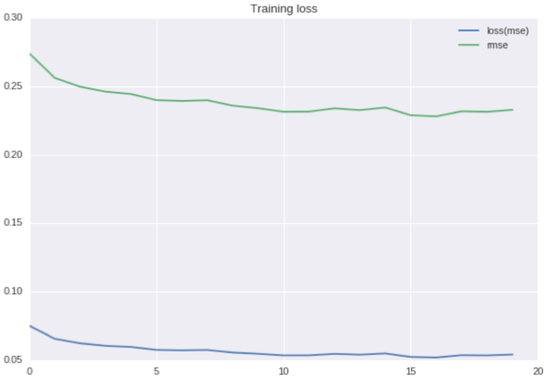
- Remove outliers, duplicates
- Min Max Scaler
- Reshape to [samples, time steps, features]

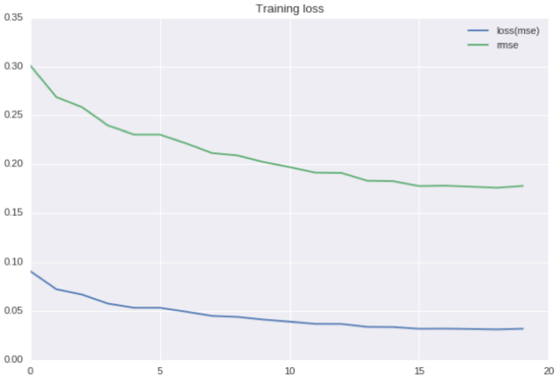


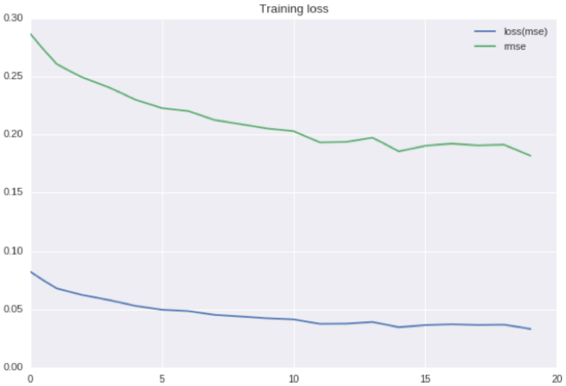
LSTM for one product

- Choose a random product
- Recreate the dataset
- Tried different lookback
- Lowest RMSE 0.2806

t-3	t-2	t-1	t
2	3	1	2
3	1	2	5
1	2	5	2
2	5	2	4

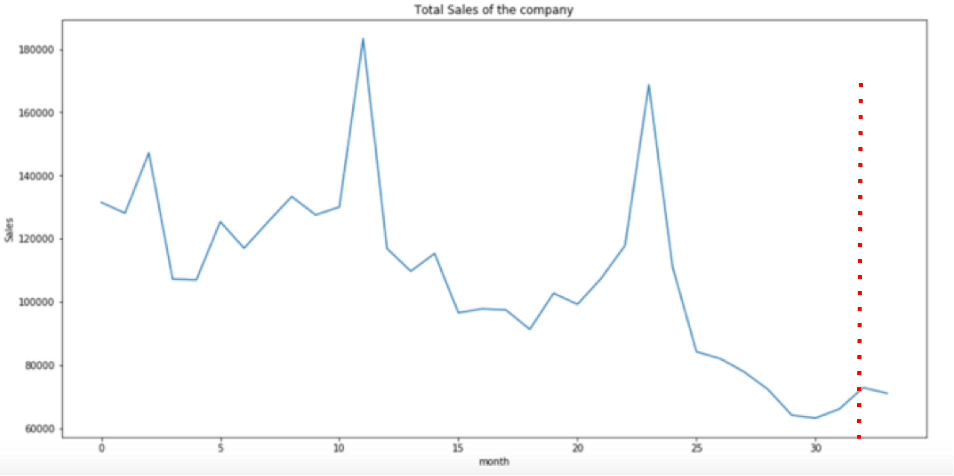






LSTM to predict sales

- Set train, test, target
- The last month ‘2015-10’ is the target to predict



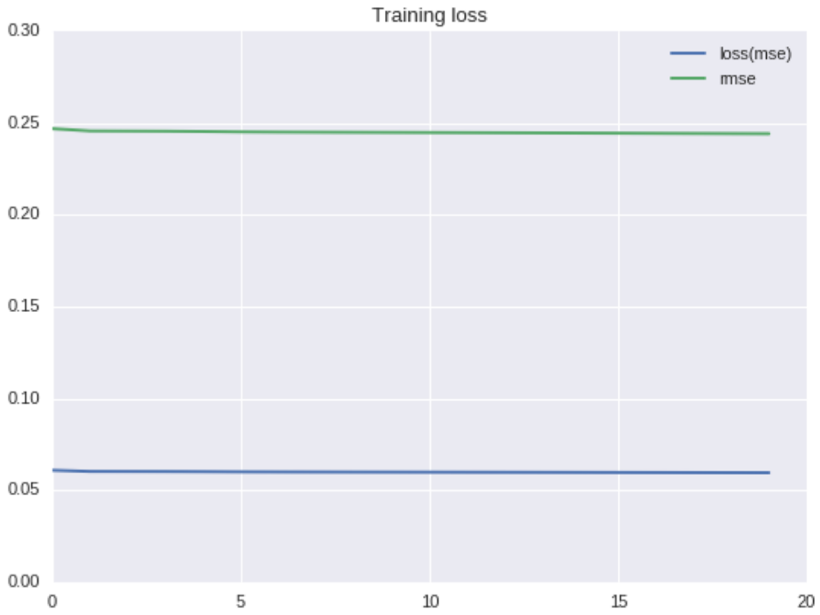
Model compares

- Epochs = 20
- Optimizer: Adam

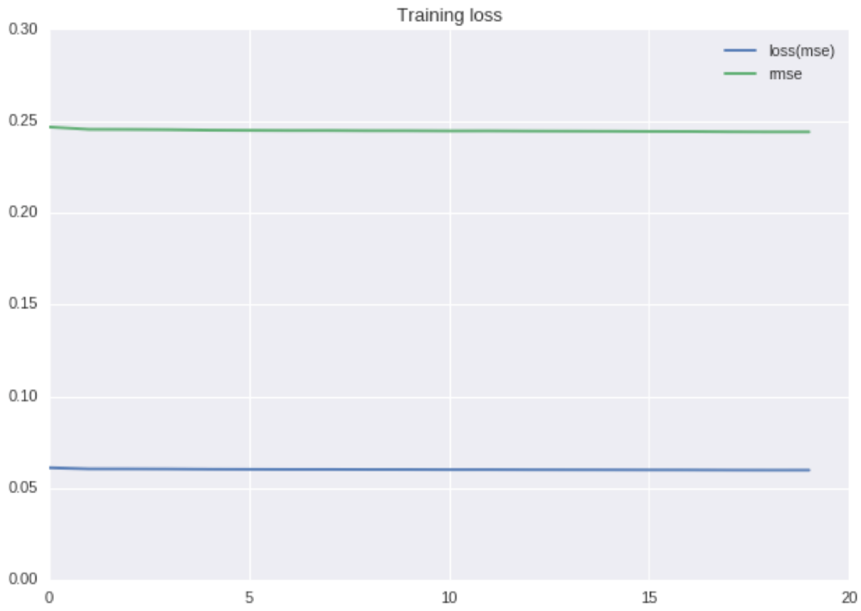
	Hidden neuron	Batch size	stateful	shuffle	RMSE on test	Computation time
Model 1	32	1260	False	False	0.05941	88.8275
Model 2	64	1260	True	False	0.05945	93.122
Model 3	64	1260	True	True	0.05953	95.9197
Model 4	64	1260	False	False	0.05948	96.5655

Stateful vs Stateless

Stateless



Stateful



Conclusion & Future Work

- LSTM is a useful model to handle time series
- Stateful and stateless is not showing a big difference in this dataset
- Adding price as a feature