LSTM Time series for predict future sales

MOHAMED KHAIRY

Data description :
daily historical sales data for group of shops.

sales_train	Shops	items	item_categories
ID - an Id that represents a (Shop, Item) tuple within the test set	**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 - number of products sold. You are predicting a monthly amount of this measure	**shop_name** - name of shop **date** - date in format dd/mm/yyyy	**item_name** - name of item	**item_category_name** - name of item category
item_price - current price of an item	**date_block_num** - a consecutive month number, used for convenience. Jan 2013 is 0, Feb 2013 is 1, Oct 2015 is 33		

Train data

	date	date_block_num	shop_id	item_id	item_price	item_cnt_day	item_category_id
0	02.01.2013	0	59	22154	999.00	1.0	37
1	03.01.2013	0	25	2552	899.00	1.0	58
2	05.01.2013	0	25	2552	899.00	-1.0	58
3	06.01.2013	0	25	2554	1709.05	1.0	58
4	15.01.2013	0	25	2555	1099.00	1.0	56

Main objective(s) of this analysis:

Predict Future Sales :

The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge.

Cleaning Data:

Issues and Actions.

NULLS

```
1  data.isnull().sum()
```

```
date                0
date_block_num      0
shop_id             0
item_id             0
item_price          0
item_cnt_day        0
item_category_id    0
dtype: int64
```

Duplicates

```
1  data.duplicated().sum()
```

```
6
```

```
1  data = data.drop_duplicates()
```

Data describe before cleaning

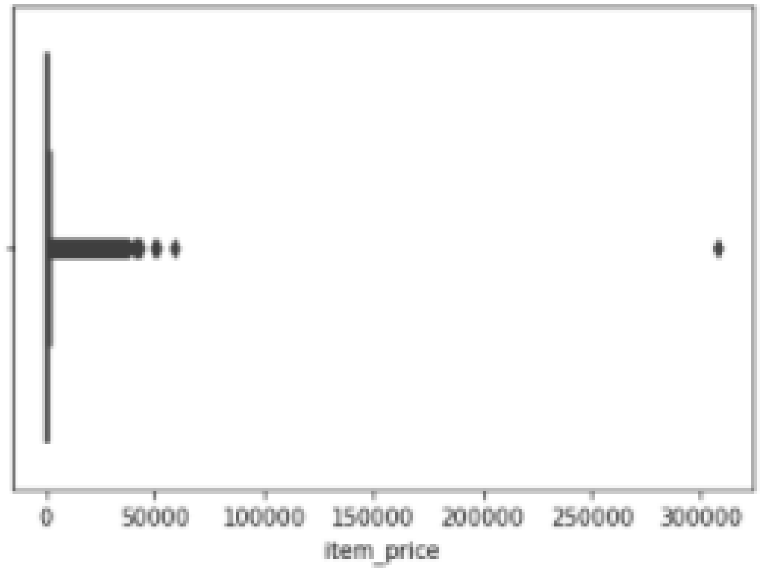
```
1  data.describe().T
```

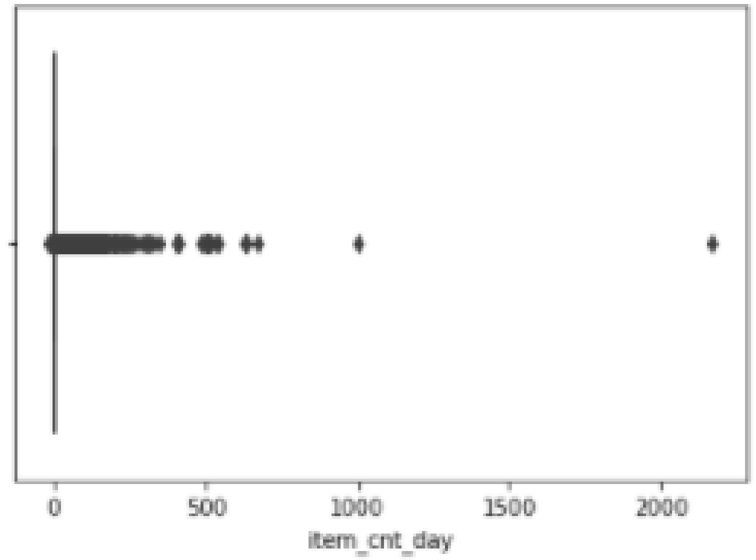
	count	mean	std	min	25%	50%	75%	max
date_block_num	2928492.0	14.569761	9.422953	0.0	7.0	14.0	23.0	33.0
shop_id	2928492.0	33.002959	16.225424	0.0	22.0	31.0	47.0	59.0
item_id	2928492.0	10200.280910	6324.396874	0.0	4477.0	9355.0	15691.0	22169.0
item_price	2928492.0	889.361584	1718.152833	-1.0	249.0	399.0	999.0	59200.0
item_cnt_day	2928492.0	1.248337	2.619588	1.0	1.0	1.0	1.0	2169.0
item_category_id	2928492.0	40.016343	17.098103	0.0	28.0	40.0	55.0	83.0

Reviewing the outlier:

The maximum for the item price 100,000, as 300,000 cannot even be due to plausible anomalies.

Some negative values which has no meaning, the other outliers can be due to plausible anomalies, for LSTM it's not necessary to remove it.



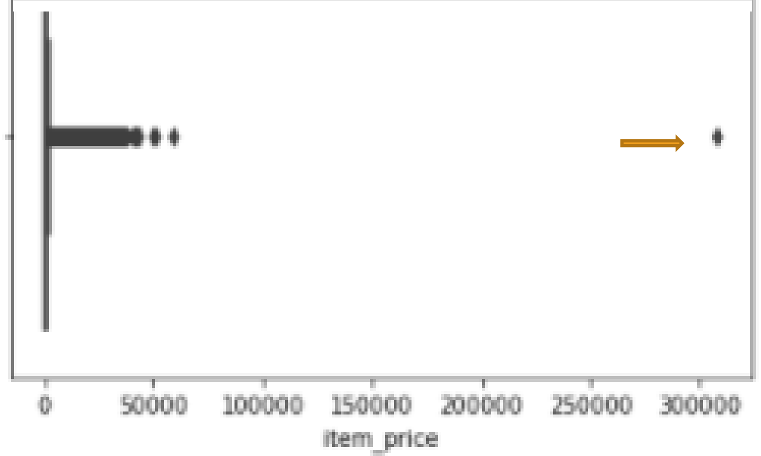


Action for outlier:

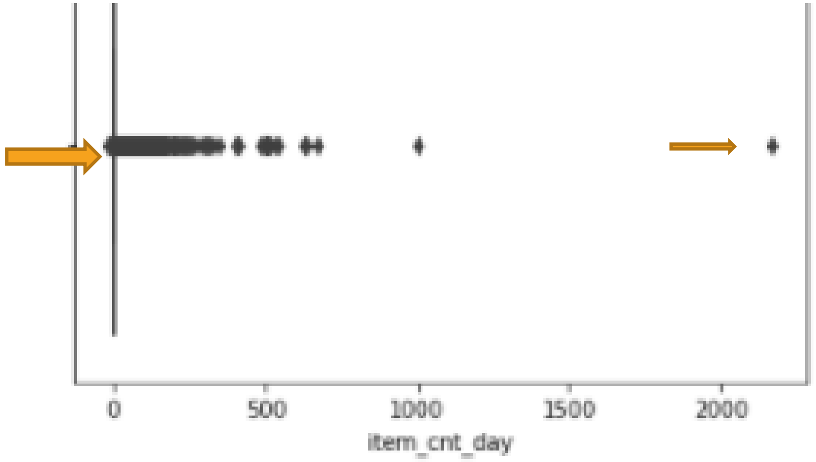
```
In [18]:  data[data['item_cnt_day']>700]
          # checked if that day with high sales is kind of cyclical event every year.
Out[18]:
```

	date	date_block_num	shop_id	item_id	item_price	item_cnt_day	item_category_id
2326929	15.01.2015	24	12	20949	4.000000	1000.0	71
2909817	28.10.2015	33	12	11373	0.908714	2169.0	9

```
data = data[data['item_price']<100000]
```



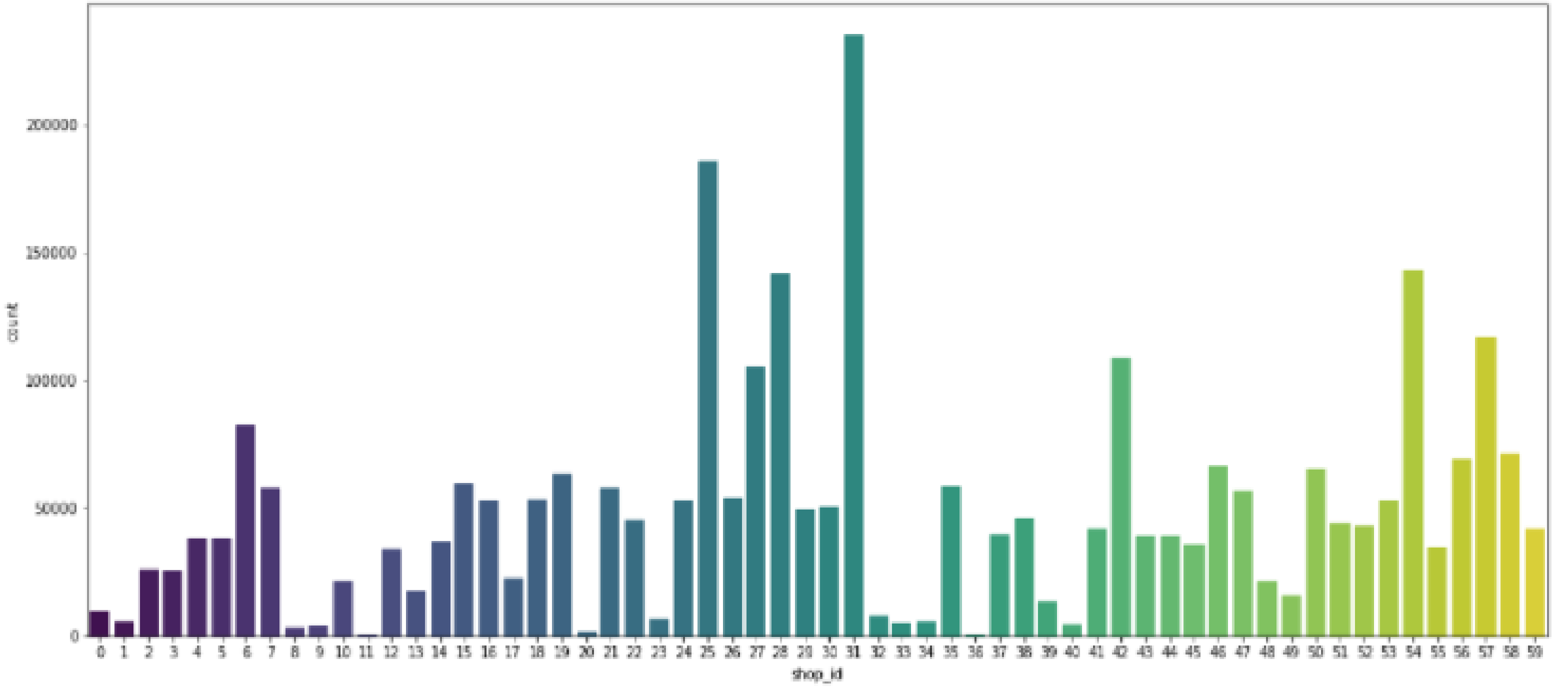

```
# removing the negative numbers and zero values in number of products sold.
train[train['item_cnt_day']<1].count()
```



Key findings:

1. Top Sellers shop_id and shops name:

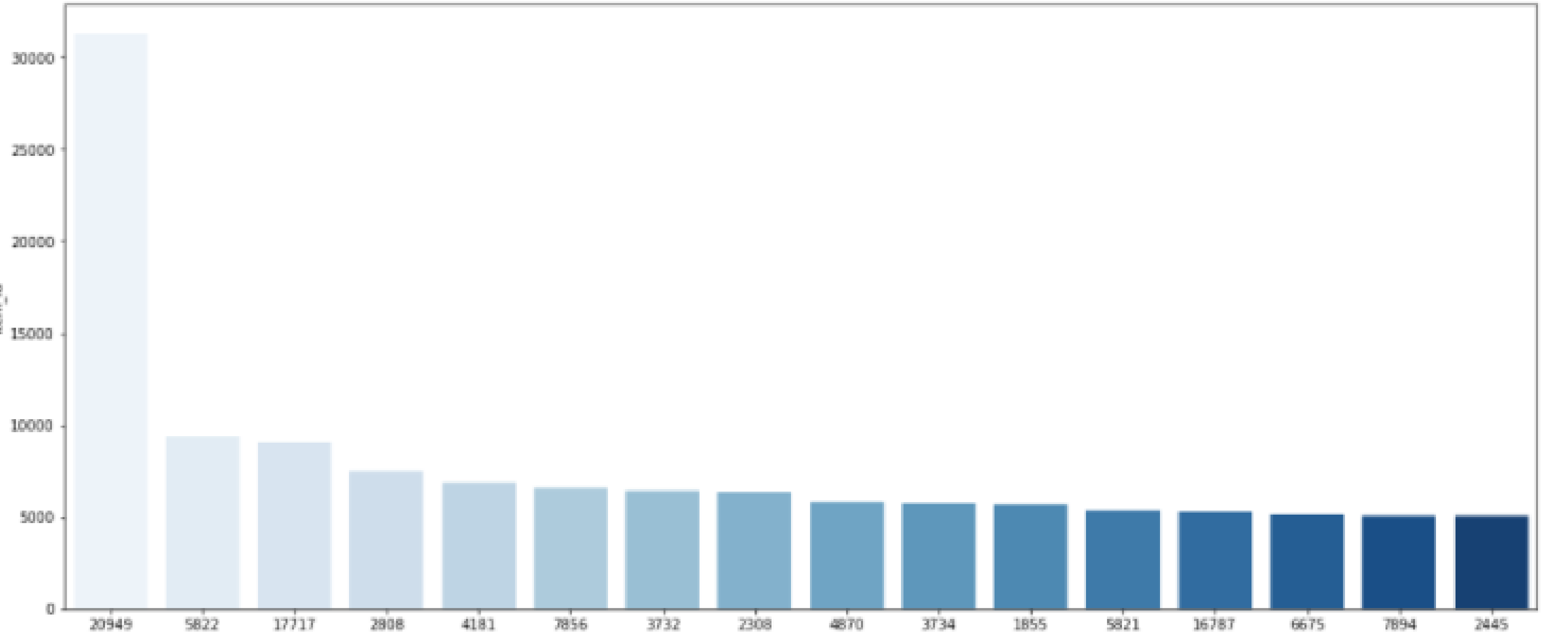
	shop_name	shop_id
6	Воронеж (Плехановская, 13)	6
25	Москва ТРК "Атриум"	25
27	Москва ТЦ "МЕГА Белая Дача II"	27
28	Москва ТЦ "МЕГА Теплый Стан" II	28
31	Москва ТЦ "Семеновский"	31
42	СПб ТК "Невский Центр"	42
54	Химки ТЦ "Мега"	54
57	Якутск Орджоникидзе, 56	57



Key findings:

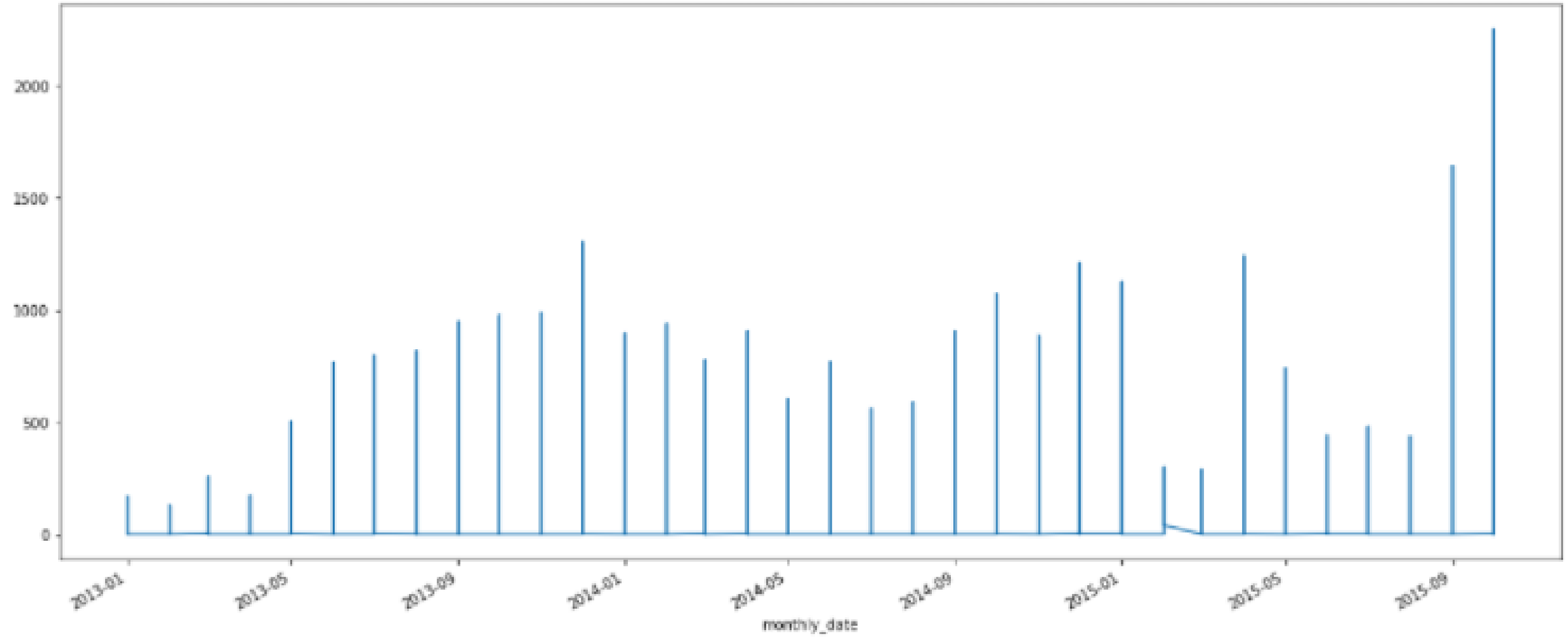
2. Top Sellers items_id and item name:

item_name	item_id
NHL 16. Legacy Edition [Xbox 360, русские субт...	5044
Nanoblock Бас-гитара	5092
Nanoblock Божья коровка	5098
Need for Speed Shift 2 Unleashed [PS3, русская...	5256
Office Home and Student 2016 Win Russian Russi...	5383
PS4: Контроллер игровой беспроводной черный (D...	5672
Pinnacle Studio 17 Plus ML (коробка)	5784
Planet mp3. Dance Collection (mp3-CD) (Jewel)	5788
SLIPKNOT Antennas To Hell CD+DVD (фирм.)	6310
Silent Hill: Book of Memories [PS Vita, англий...	6451
Skylanders. Набор из трех фигурок: Chop Chop, ...	6598
THE ROLLING STONES 1964-1969 Limited Edition...	6850
V/A Top 100 DJ 4 4CD (digibook)	7470
БУКАШКИ: ЛЕСНЫЕ СКАЗКИ	9067
Брелок металлический WOT Снайпер (100110)	9404



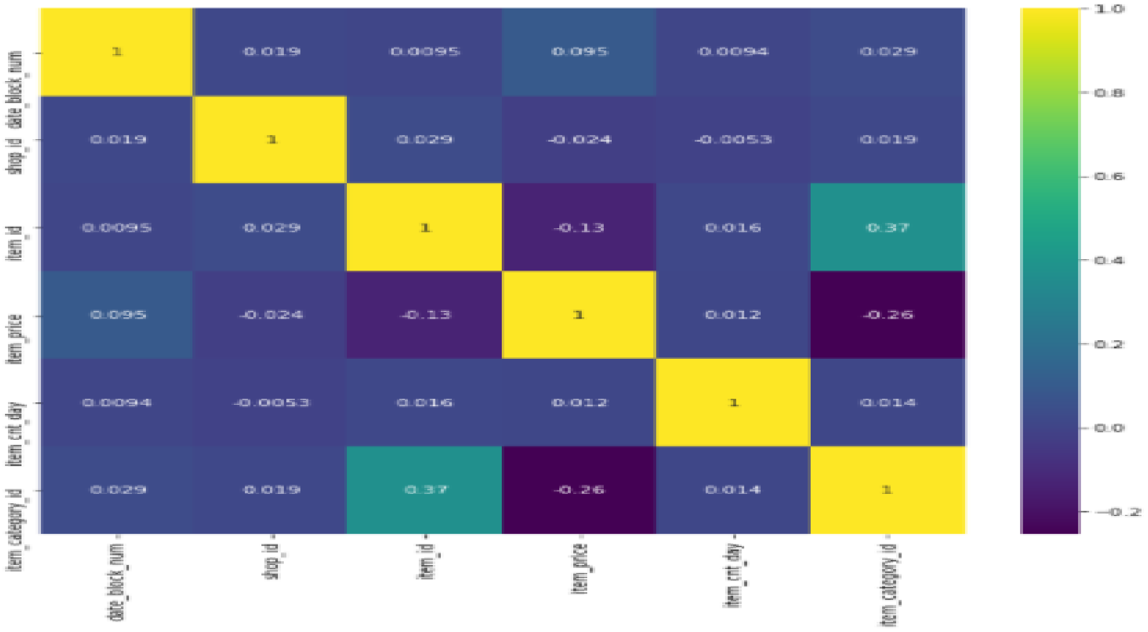
Key findings:

3. Last 2 month score more high in the whole 3 years:



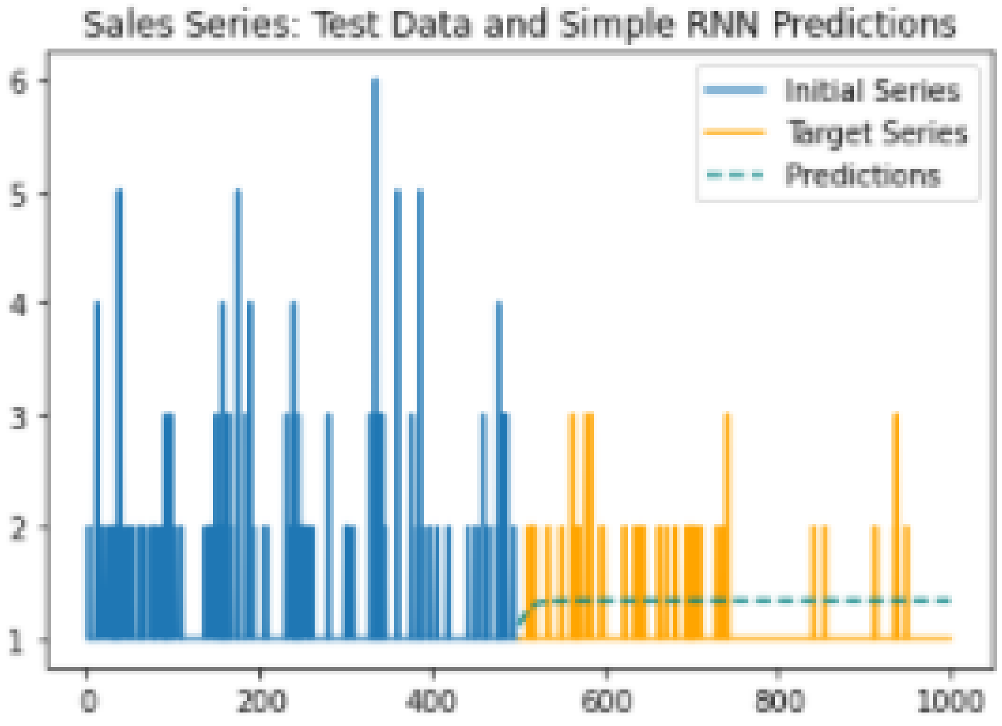
Correlation:

4. Heatmap for showing the correlation between the data features:

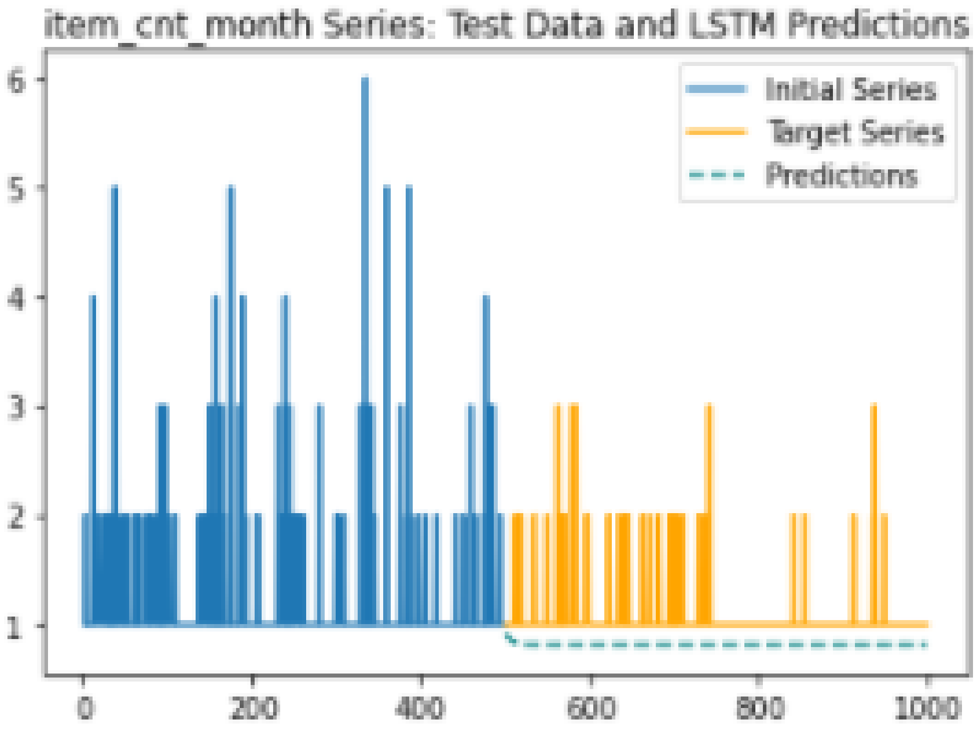


Reviewing the models :

RNN model performance:



LSTM model performance:



Next steps

The 2 models looks like underfit:

1. We need to add more layers and increase the epochs.
2. Get dummies for the categorical features .

That will require more time and some computer with high RAM.

Project link:

https://github.com/khairy84/Sales-LSTM-Time-series