KioMed

Sales prediction for better inventory management

Problem and goal

Problem:

Warehouses cannot meet the demand of the medicines in stores in respective cities

Goal:

Forecast the sales for the period of one month after the end of data(July(7)).This is to help ensure the company can restock its supplies of medicine in warehouses accordingly in each city for the period of one month.

—Use machine learning to solve the problem

Mindset to approach machine learning

Most of the gains comes from great features, and not great machine learning algorithms. So follow basic approaches:

1. Make sure your pipeline is solid end to end.
2. Start with a reasonable objective.
3. Add common sense features in simple way.
4. Make sure your pipeline stays solid.

– GOOGLE

Mindset to approach machine learning

Most of the gains comes from great features, and not great machine learning algorithms. So follow basic approaches:

1. Make sure your pipeline is solid end to end. = Start simple make a submission
2. Start with a reasonable objective. =set a target accordingly
3. Add common sense features in simple way. =Add more and more features
4. Make sure your pipeline stays solid. =Keep iterating(keep RMSE score in mind)

– GOOGLE

Basic info(Personal)

Work Station = Apple macbook m1 air (Ram=8, core=8)

Platform and coding language= Jupyter + python

Due to the limitations with my work station i have to improvise a lot of things. Like looking for a proper way to shorten the data and make sure it doesn’t affect my predictions.

Data Visualisation

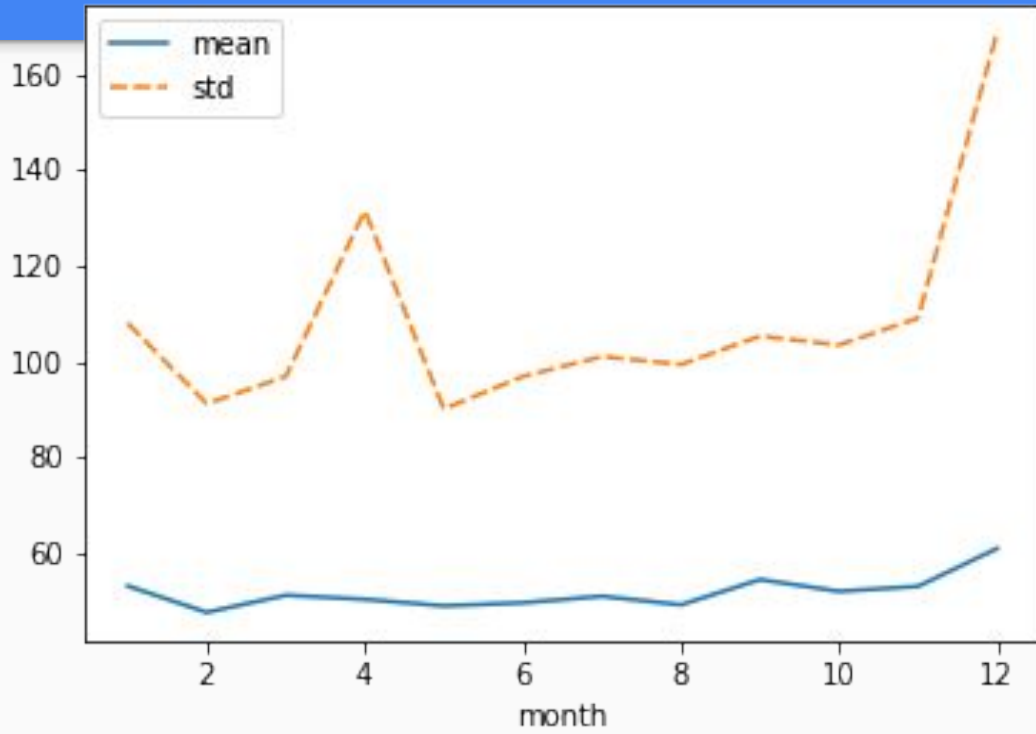
(mean and std of sales with respect to months)

Due to the fact that mean and standard deviation is very linear for months 6-8, and my target month also falls under this, i decided to use these months only.

Not only did this shorten the data, but it also gave me better prediction than to use full data.(Further analysis required)

1 Linear model with full data= 74/78

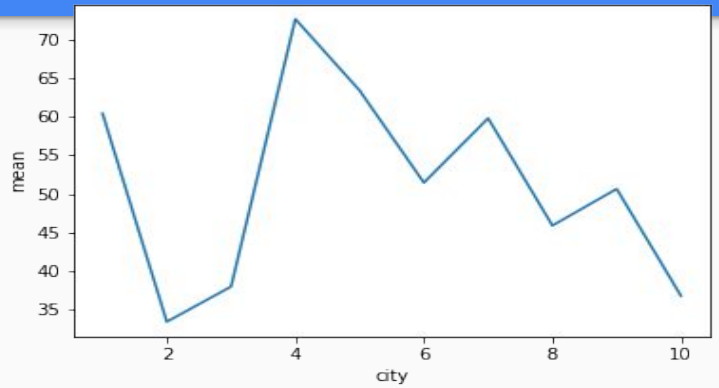
2 linear model with shorten data=72/77



Visualisation-2
(City and Day of month)

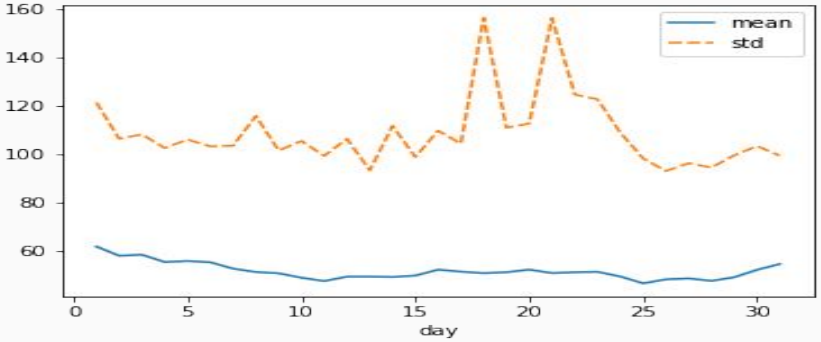
City

City 2 has the lowest sales while city 4 has the highest . The variation between different city is a lot, so it can help us a lot.



Day of month

Unlike city this follows a smooth pattern with highest sales being at start and at the end of the month. There is a sudden spike in the second half of the month so it can mean something(Further analysis required)



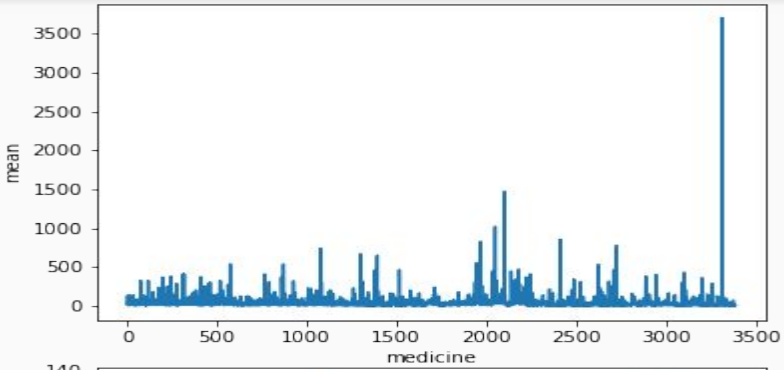
Visualisation-3
(Medicine and Day of week)

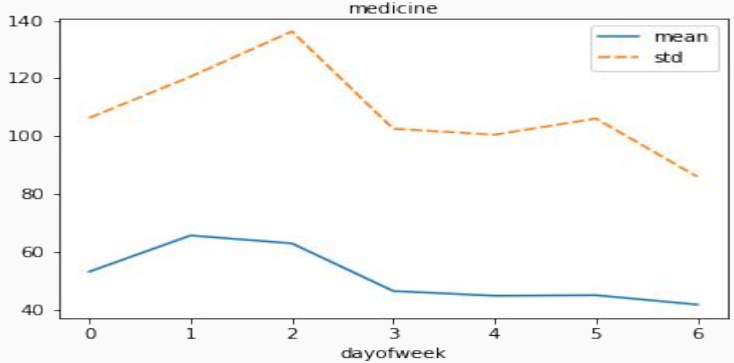
Medicine

The mean of some medicine is high, but that can be due to it being some niche and expensive medicine.

Day of week

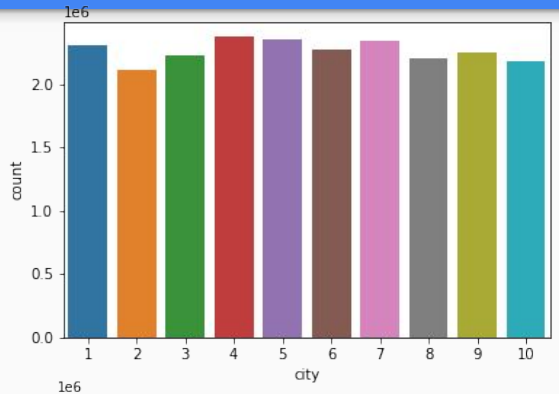
The sales start of very high in the beginning of week, and then start to become lower. This is a interesting observation and one that we need to look at in the future.

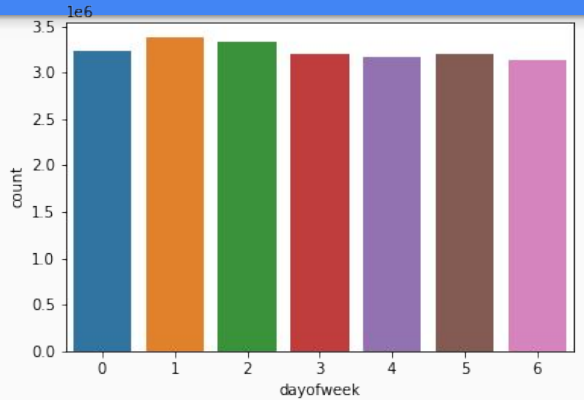


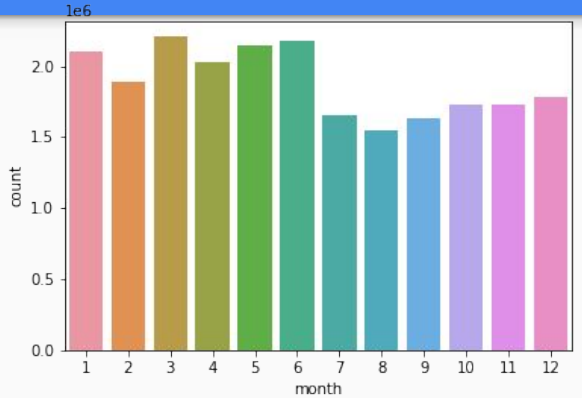


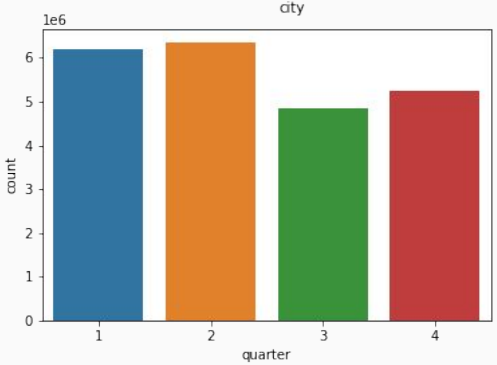
Visualisation-4

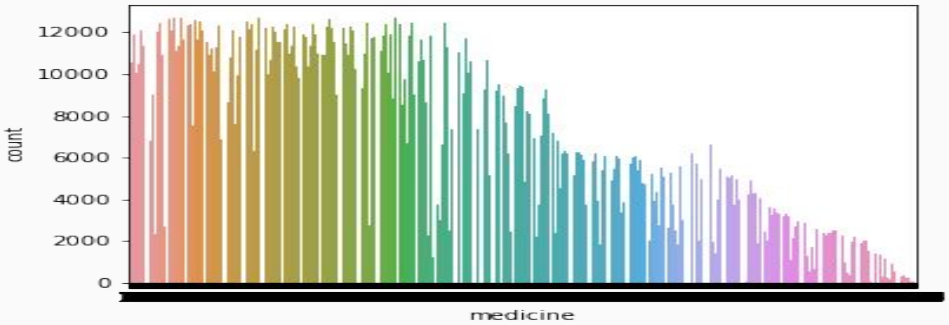
The distribution of data is good, so there is no issue there, as for medicine, more expensive medicines tends to sell less so that should be ok too.











Feature Engineering

Main Dataset=(Year,Month,Day,City,Medicine,Sales)
Using Datetime to create other variables= (Date, quarter,Day of week)
We also have Footfall data, which can be helpful for training.

Sales is our target variable, and the main dataset looks as below,

We don’t have any NA or Null values.

	Date	year	month	day	city	medicine	sales	dayofweek	quarter
0	2015-06-01	2015	6	1	1	1292	56.0	0	2
1	2015-06-01	2015	6	1	1	1	4.0	0	2

Feature Engineering

Using groupby and aggregation calculate the mean and standard deviation for (Day,month,day of week,medicines,city) on the basis of sales.
This will help us find any kind of pattern or seasonality for the respective variables.

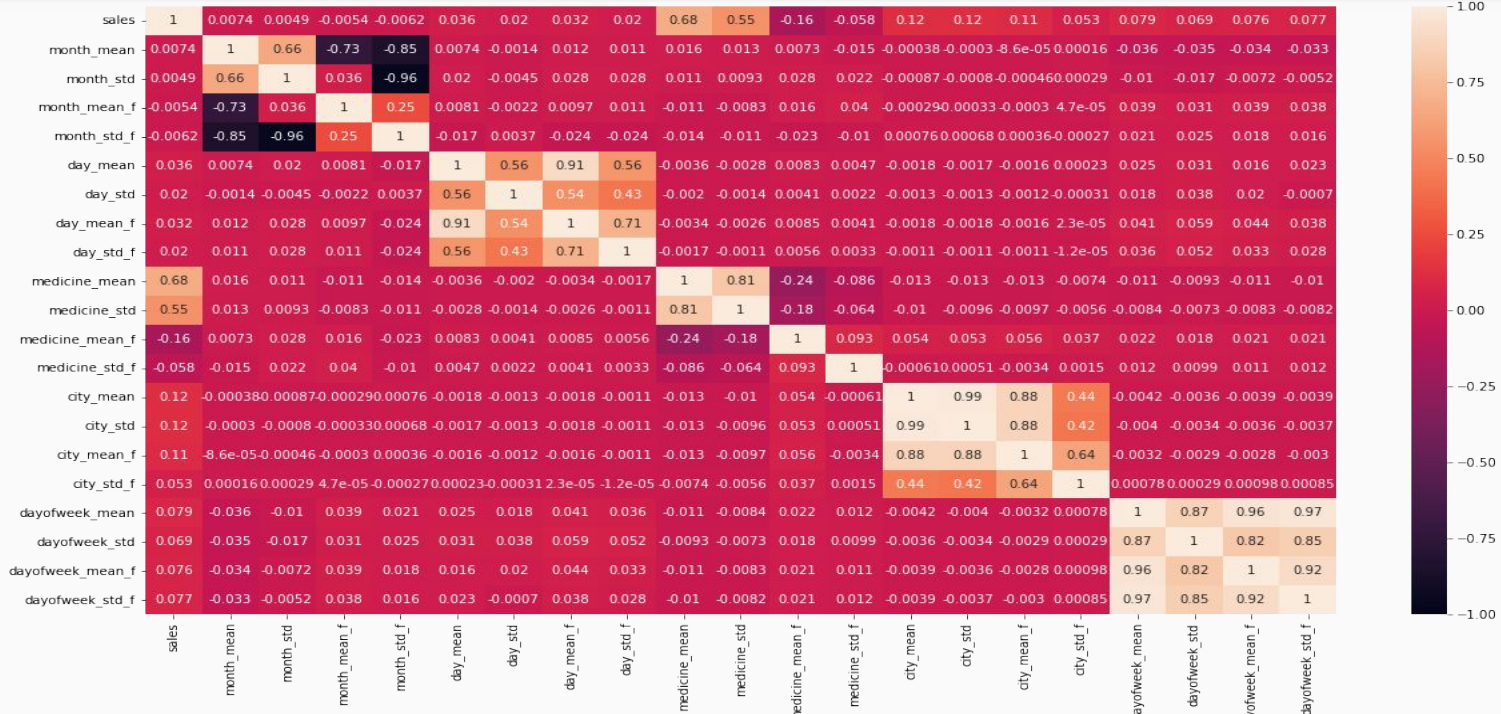
We can also use Footfall data in similar manner to find the mean with respect to all the variables, it may also help us in finding some kind of pattern.
Total features created around 22

	Date	year	month	day	city	medicine	sales	dayofweek	quarter	footfall	...	medicine_mean_f	medicine_std_f	city_mean	city_std	city_mean_f	city_std_f	dayofweek_mean	dayo
0	2015-06-01	2015	6	1	1	1292	56.0	0	2	14356.0	...	12429.260535	3285.969458	59.353949	113.002345	12579.50207	1276.953338	51.787851	1
1	2017-07-03	2017	7	3	1	1292	32.0	0	3	14444.0	...	12429.260535	3285.969458	59.353949	113.002345	12579.50207	1276.953338	51.787851	1
2	2015-08-03	2015	8	3	1	1292	52.0	0	3	14452.0	...	12429.260535	3285.969458	59.353949	113.002345	12579.50207	1276.953338	51.787851	1
3	2018-06-04	2018	6	4	1	1292	20.0	0	2	14564.0	...	12429.260535	3285.969458	59.353949	113.002345	12579.50207	1276.953338	51.787851	1
4	2016-07-04	2016	7	4	1	1292	16.0	0	3	12480.0	...	12429.260535	3285.969458	59.353949	113.002345	12579.50207	1276.953338	51.787851	1

5 rows × 30 columns

Feature selection

Selecting the most meaningful and helpful features with the help of correlation matrix and heatmap. Anything with the correlation of above 70-75 we can consider heavily correlated and choose appropriately.



Feature selection

We can also check correlation with our target variable(sales) to get some understanding.

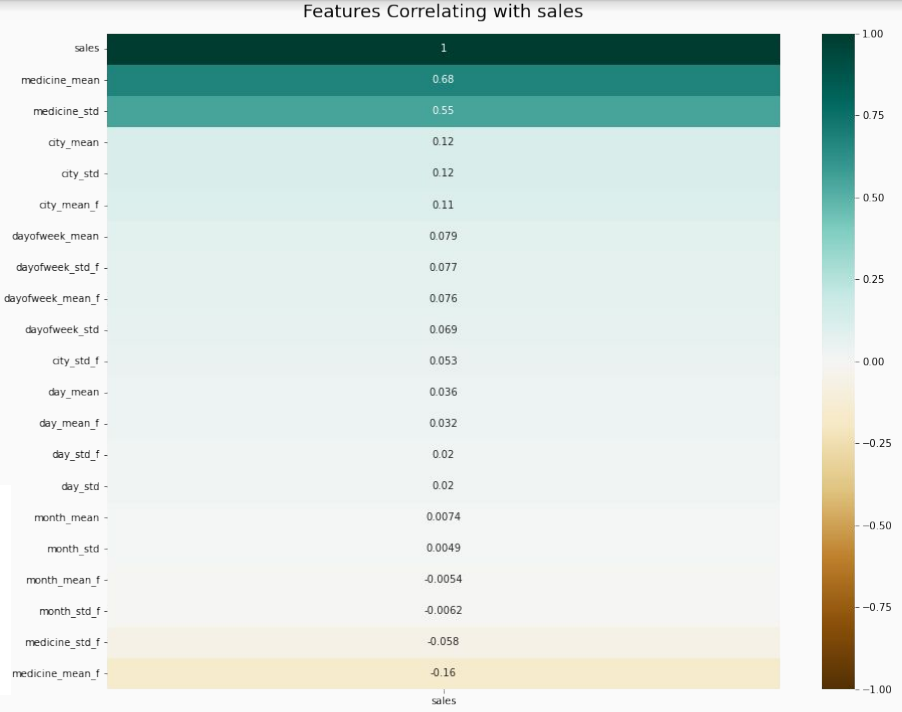
This gives us some understanding on how our numerical variables correlate with our target variable.

Based on everything ,these are the final variables i selected for training.

Categorical =5

Numerical =7 Total = 13

	month	day	city	medicine	dayofweek	day_mean	medicine_mean	medicine_mean_f	city_mean	month_mean	month_mean_f	dayofweek_mean
0	7	1	1	1292	6	60.471096	23.960425	12429.260535	59.353949	50.944719	12322.888339	40.892422
1	7	1	1	1	6	60.471096	23.596844	12467.489240	59.353949	50.944719	12322.888339	40.892422
2	7	1	1	2	6	60.471096	69.643944	12334.403706	59.353949	50.944719	12322.888339	40.892422
3	7	1	1	3	6	60.471096	123.892388	12309.486877	59.353949	50.944719	12322.888339	40.892422
4	7	1	1	4	6	60.471096	22.927114	12477.046647	59.353949	50.944719	12322.888339	40.892422



Model training, validation, inference

Train/Validation split: Time splits ,last 20% used as validation(no shuffle to keep time aspect).

Training: LightGBM(Gradient boosting Trees)

Reason to use LightGBM:

1 Faster training speed and higher efficiency.

2 Lower memory use.

3 Better accuracy

4 Support of parallel, distributed and gpu learning

5 Capable of handling large scale data.

To know further = (https://lightgbm.readthedocs.io/en/latest/Features.html#optimal-split-for-categorical-features)

Model training, validation, inference.

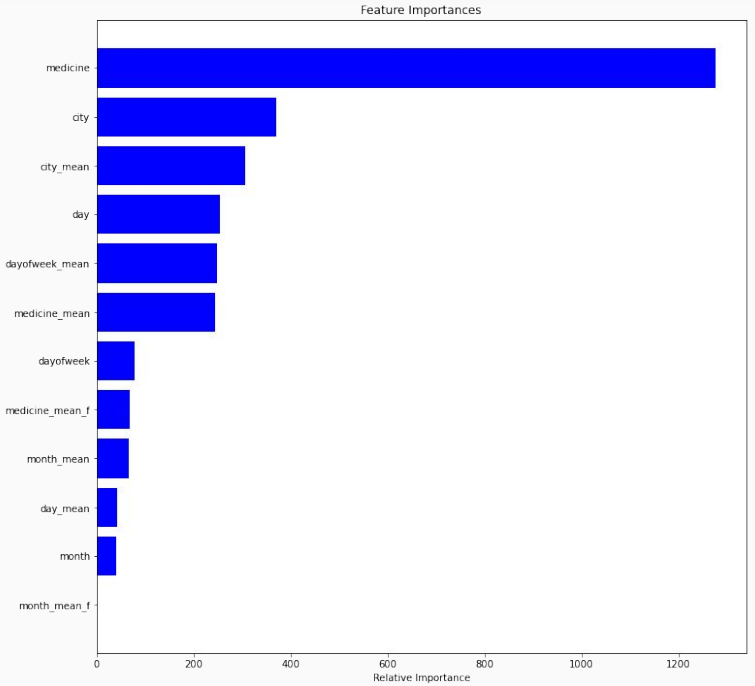
Test score(Best)= 50.9(RMSE)

Feature importance

Medicine is the strongest feature, followed by city,city mean,day, day of week mean and medicine mean.

Footfall mean by medicine is a little helpful, but in general footfall data does not helps a lot . month doesn't matter much coz we are using only 3 month of data.

With better machine and more time further advancement can be make.



Problems faced during the process

Merge function= whenever we merge 2 dataframes, the function will change the order based on the on function,(eg, on= month, then the order will become monthly.)

But that is a big issue when we want to predict in a particular order.

Other issue is the machine, which is not powerful hence limits me in what i can do. Still i tried my best .

Thank you for the opportunity.