Predicting Future Sales

TEAM 7
December 7, 2018

1 PROJECT OVERVIEW

Sales prediction is the process of estimating future sales . We have data for two years. In data we have item_price, item_categories, shop_ID,item_cnt_day,item_ID,date_block_num some more data sets. We train our data sets and predict sales of those items .

Problem statement : The goal is to Predict future sales for a given data
Estimating daily prices of items in the shops
′date′,′date_block_num′,′shop_id′,′item_id′,′item_price′,âĂŹitem_cnt_idâĂŹ

Metric used :- The***root-mean-square error***(RMSE) is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

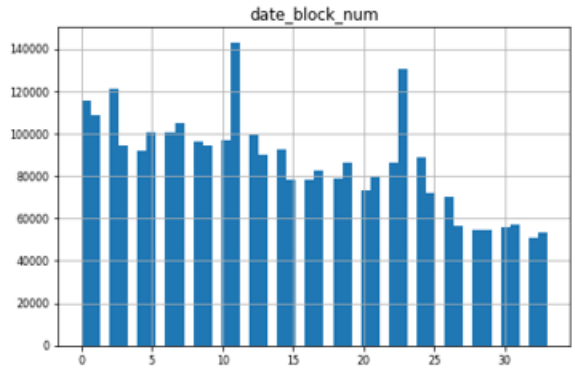
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(X_{obs,i} - X_{model,i}\right)^2}{n}} \tag{1.1}$$

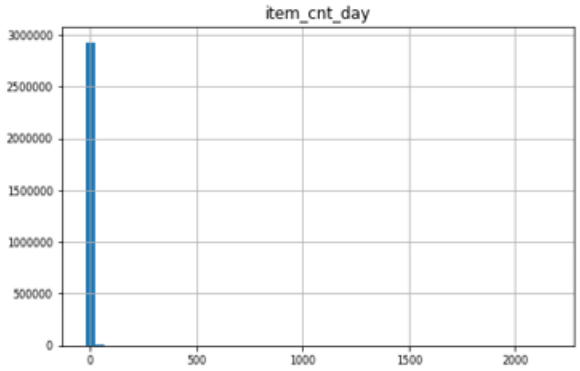
RMSE values can be used to distinguish model performance in a calibration period with that of a validation period as well as to compare the individual model performance to that of other predictive models.

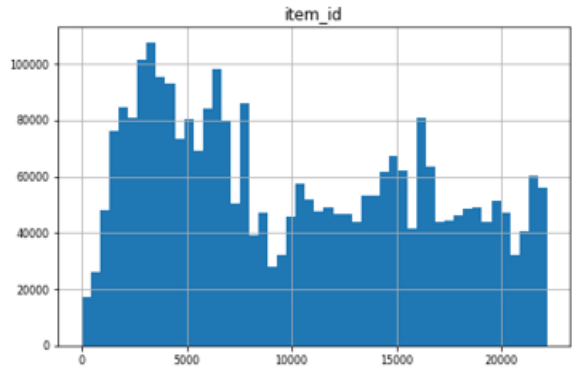
2 Data Analysis(visualization)

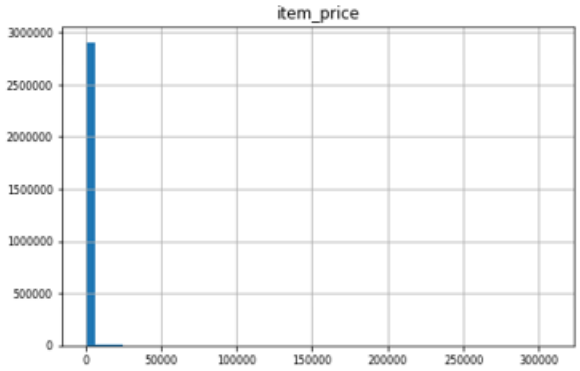
Overview of Datasets;

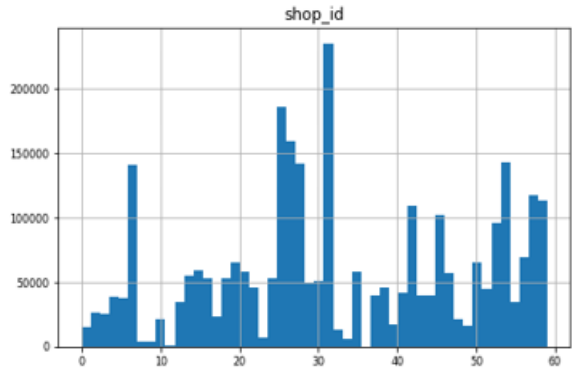
overview of datasets:-





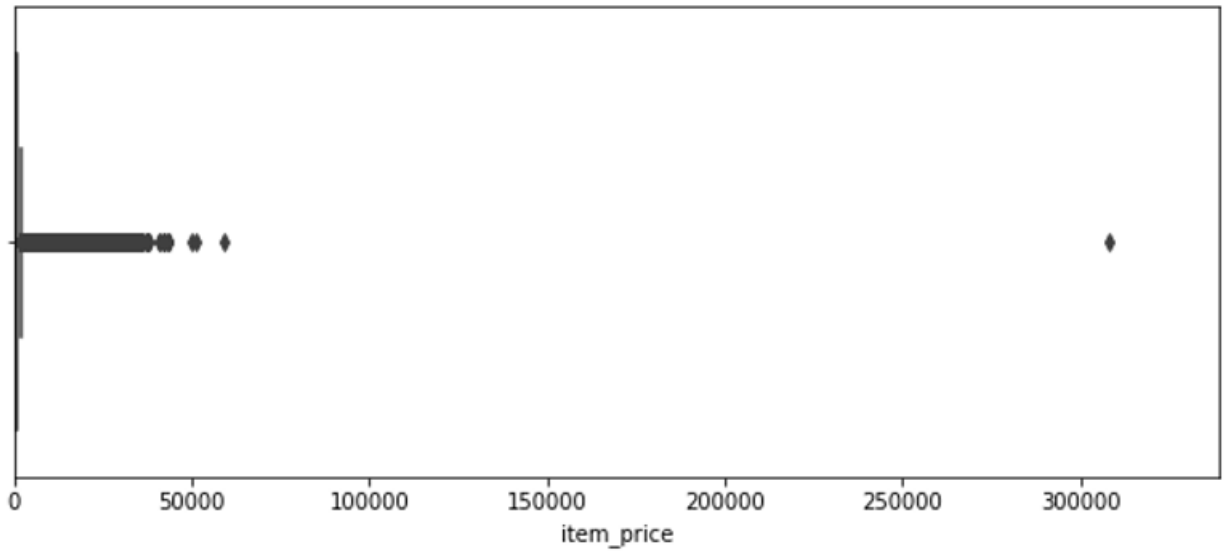






These graphs contain the overview of our datasets

Figure 2:- the following plot consits of range of item prices



In this graph we can see that most of prices are under 6000 and only one item price above 30000 so we consider it as an outlier and we are removing it for a better model

Figure 3:- the following plot consists of mean of item_cnt and monthly sum.
Mean of the month increases gradually till end of the month
Though there is up and downs in item_count in middle of the month we can see that sales of items increases in ending of the month.

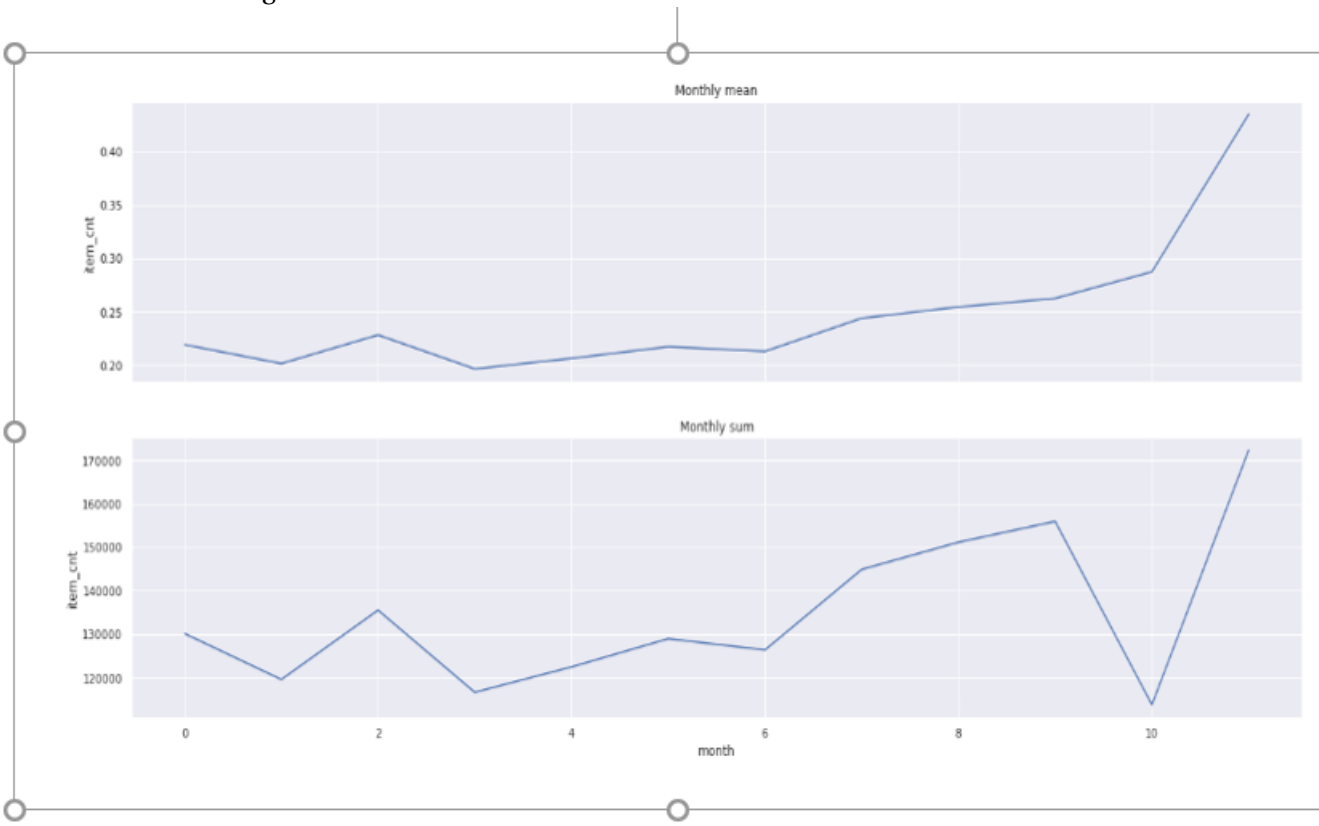
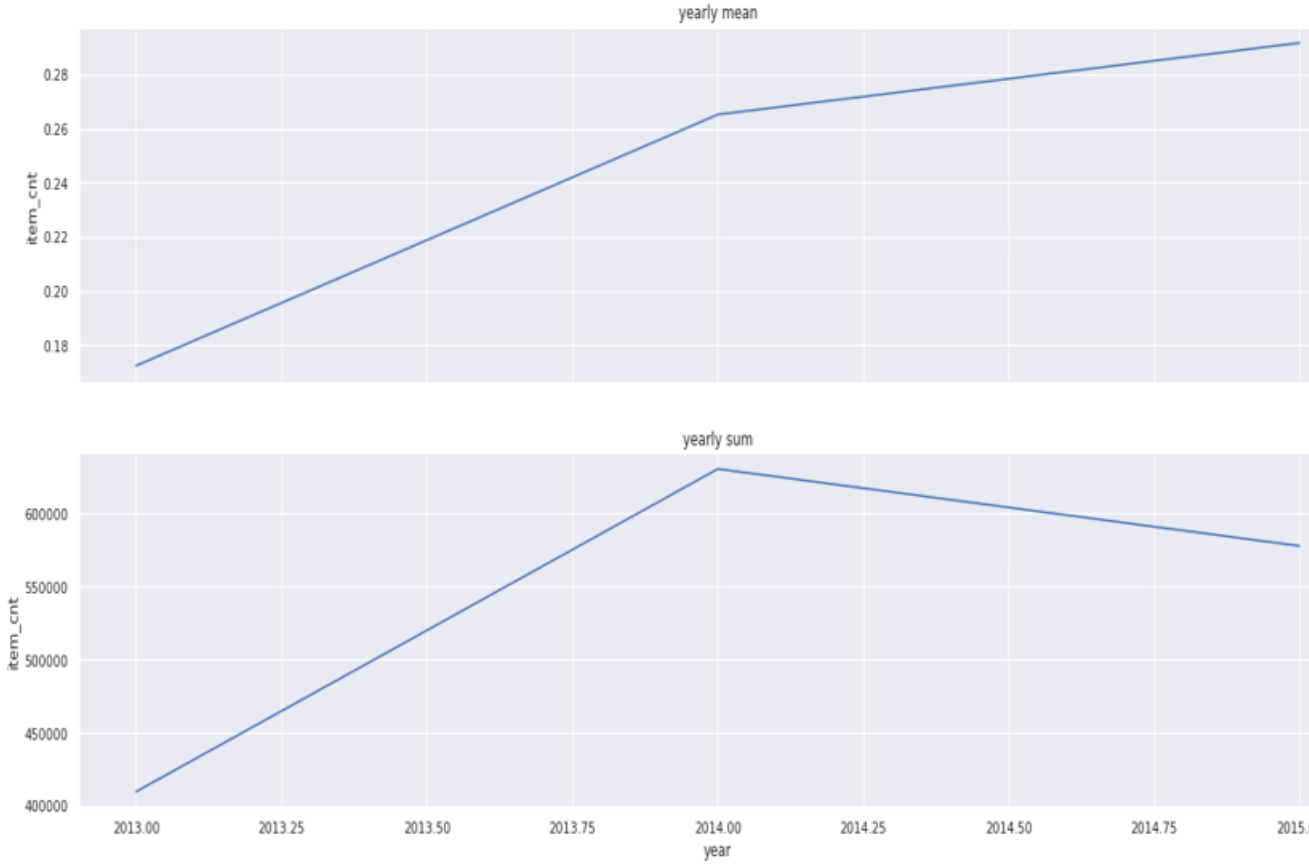
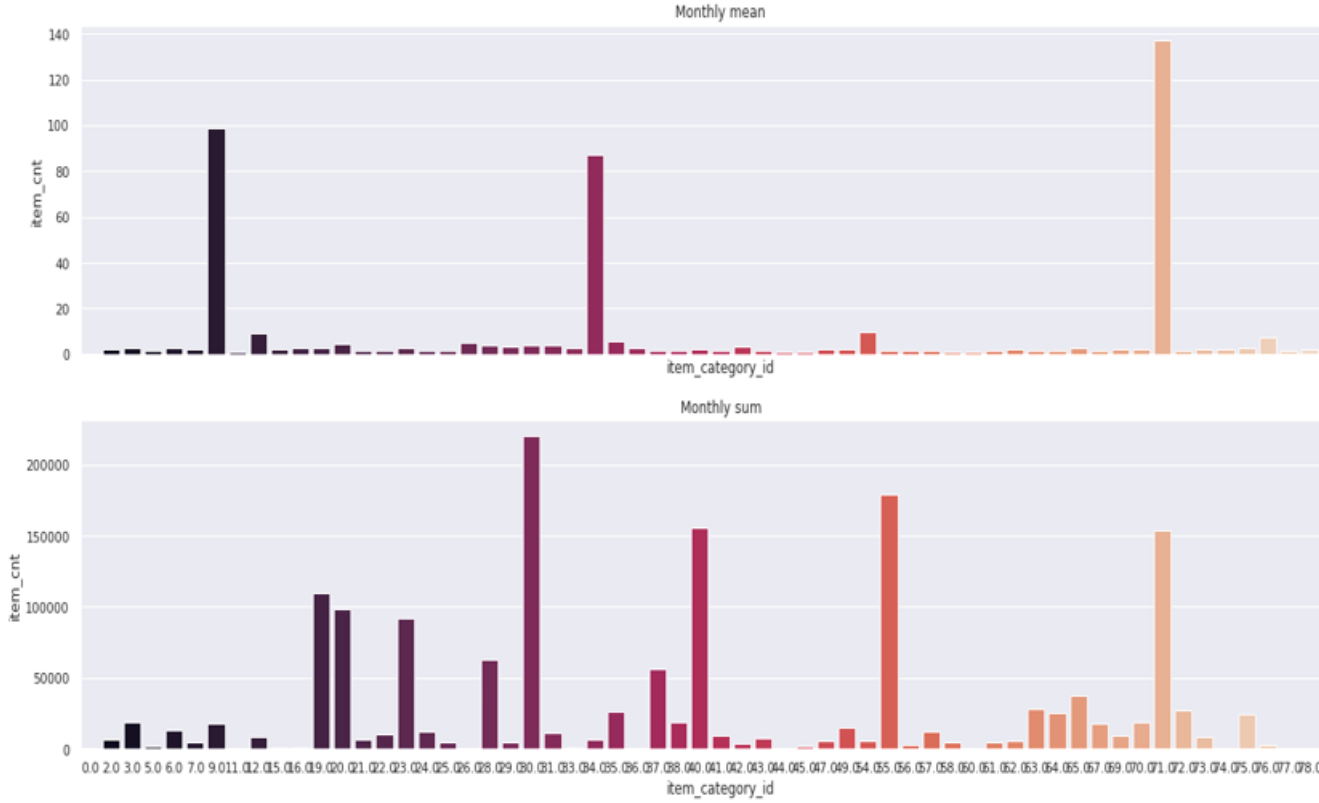


Figure 4:-the graphs below tells about mean of item_cnt yearly and yearly sum



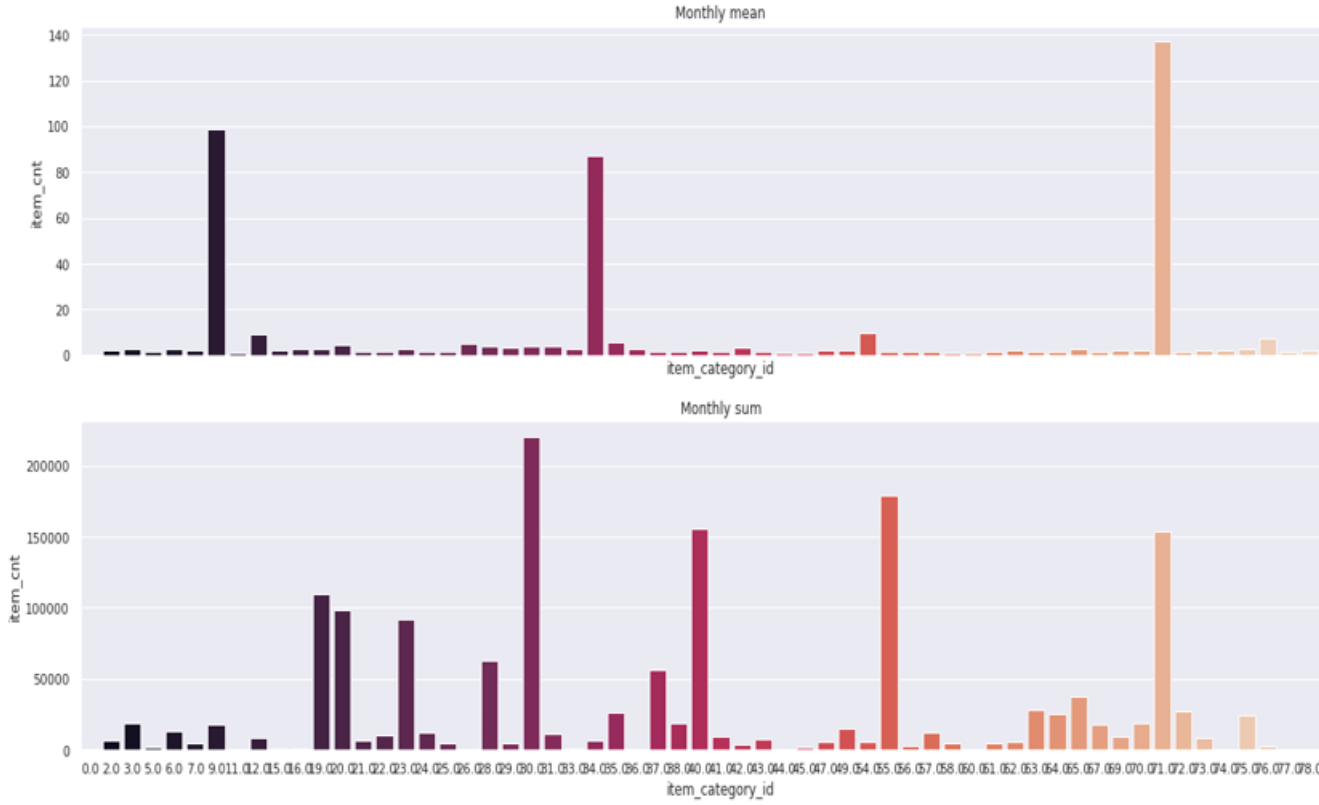
For our given item_cnt data set there is linear increase of sum till 2014 ,from 2014 the increase rate mean decreases a little
Itm_cnt increases till 2014 and decreases gradually while approaching 2015

Figure5:-this plot shows mean item categories_id with itm_cnt and monthly sum of item count versus item_categoriy_id



From the graph we can see that there only few categories maintain the most of the sales so this data set is in not that use full

Figure:-6 the below plot is abount monthly mean of shop_id and item_id and montlysum of shop_id and item_id



Most of the shop id have almost same count but 3-4 shops have much higher rate, this may be a indicative of the shop size.
As the shop size is bigger than other shops, monthly sum of these shop is higher than other shops

.

3 APPROACH

Overview of our approach: -

We chose linear regression as our base model. We obtained an RMSE of 2.3. Which was a better result than the previous sample CSV.

Then we shifted to using XG_BOOST and did some more preprocessing. The result was not desirable, so started using feature extraction.

We divided the date into day, month and time using: train['X'] = pd.DatetimeIndex(train['date']).X

Where X can be month, day or year.

We varied between using the features for a better result. We decided to use a combination of day and month because using year with labels resulted in the model getting overfitted. So, we stopped using year label. We obtained an result of 1.6 RMSE.

For better results, we shifted to random forest model.

Using the preprocessing and feature extraction under random forest model, the result was higher than XG_BOOST, so we are shifted back to XG_BOOST and dropped ID label, date label.

We fixed itm_cnt_id range from 0 -1100, we increased number of estimators to 1500 and varied max_depth from 25-35 and fixed 30 so that our model doesnâĂŹt overfit

4 METHODOLOGY

FEATURE EXTRACTION AND PRE PROCESSING :-

> We removed duplicates
> Removed null values
> Replace zero values with most commonest price (30)
> Given range for itm_cnt_day from 0-1100(clip(0,1100))
> Splited date into day,month,year
> For xg_boost last 250 rows ared considered
> Dropped date ,Id ,item_cnt_day

Algorithm and techniques :-

Algorithms used:-

> Linear regression
> Xg_boost
> Random forest

With these algorithms we trained our data sets .

Validations and results

Best validation for linear regression is 2.65 (appx)

Best result for linear regression is 2.18

Best validation for random forest is 1.34(appx)

Best result for random forest is 2.90

Best validation for xg_boost is 1.8805

Best result for xg_boost is 1.209

Link For Pickle File;

https://drive.google.com/open?id=1ISNvDgRA776OMnvFSehUZvN2zhFB5rtK