Pankaj Acharya
Capstone Project01 - EDA
May 01, 2019

Project: Predicting future sales

Introduction:

To maintain product stocks in stores is challenging for companies making those products. Having the product in stock for too long costs the companies a lot and on the other hand short supply will also lead to lost revenue, let alone customer trust. Maintaining a balance between overstock and understock is a major problem for companies in present cut-throat competition.

1C Company, one of the largest Russian software firms has provided a challenging time-series dataset consisting of daily sales for data scientists to build a model to predict future sales for every product and store in the next month. This will help the company decide how to maintain stocks for particular items in particular shops optimally.

The Dataset:

All the data for my project are downloaded from Kaggle website (https://bit.ly/2RsJ8AV). Following files are provided:

1. **sales_train.csv** - the training set. Daily historical data from January 2013 to October 2015
2. **test.csv** - the test set. I need to forecast the sales for these shops and products for November 2015.
3. **sample_submission.csv** - a sample submission file in the correct format.
4. **items.csv** - supplemental information about the items/products.
5. **item_categories.csv** - supplemental information about the items categories.
6. **shops.csv** - supplemental information about the shops.

The dataset is a real world data set and is fairly large. The training dataset has 2.9 million rows. There were no missing values. It is a time-series data, but due to multiple transactions in same day makes it challenging.

Exploratory Data Analysis:

The data has an interesting trend as shown in total daily sales (**Figure 1**) and total monthly sales (**Figure 2**). The sales tend to spike in December, obviously due to Christmas and

New Year. Another point to be noted is that the general sales trend is decreasing each year.

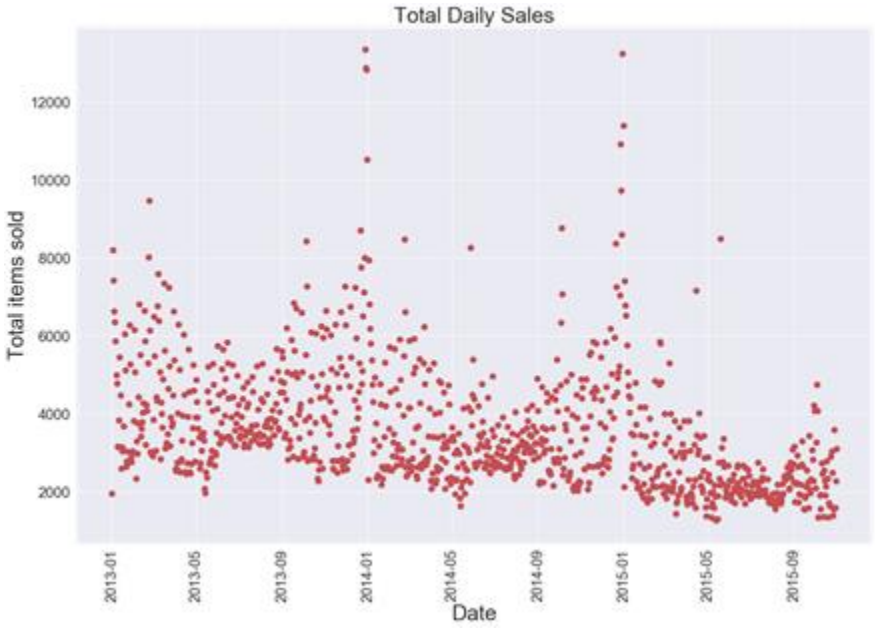

Figure 1. Total daily sales.

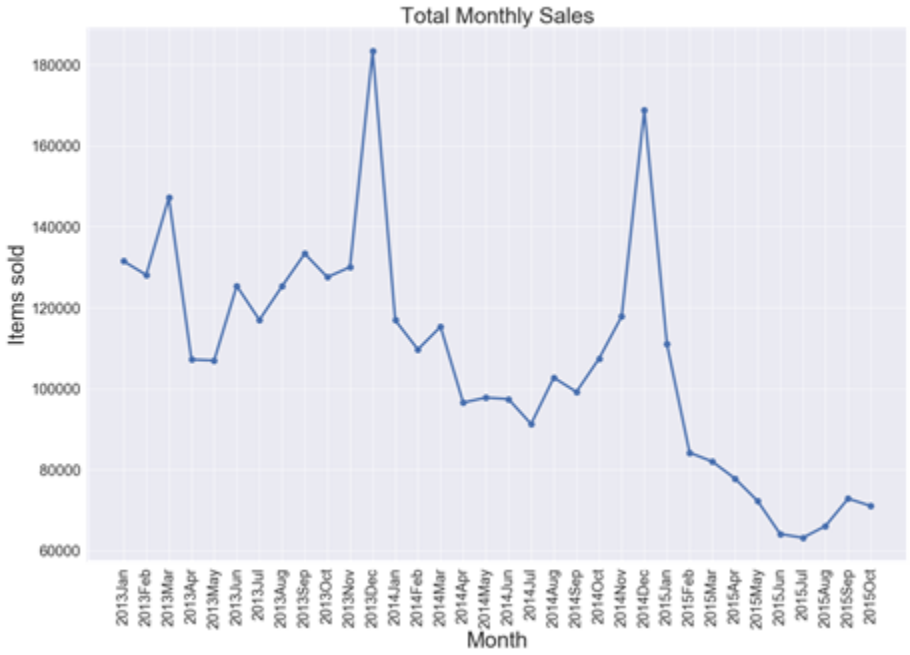

Figure 2. Total monthly sales.

Not all items are sold equally and not all shops perform similarly. When I looked at the top ten performing shops' trend, it was as expected for the most part. Interestingly for a couple of top ten performing shops, the sales was stopped completely before October 2015 (the last month in the dataset) (**Figure 3**). This could be because either the shops went out of business, which

seems less likely as they were selling well, or they were bought by somebody else and changed name so that it is not showing in the dataset.

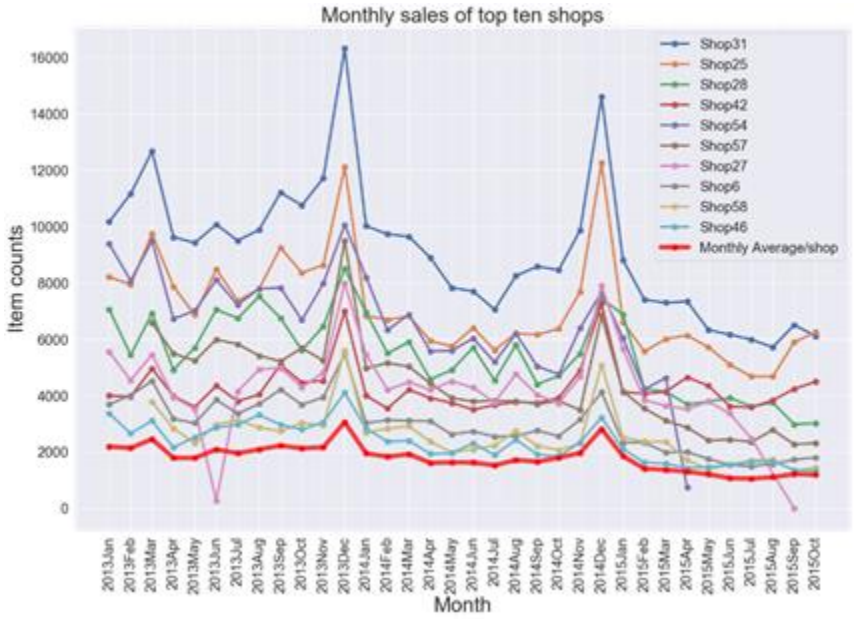

Figure 3. Performance trend of top ten shops over time.

While analyzing the top ten selling items, I found that one item (item no 20949) was by far the most popular item (**Figure 4**). Compared to it, lines of other items seemed generally flat. Some items seemed to show up in the market later than others, which could be because the software was launched later.

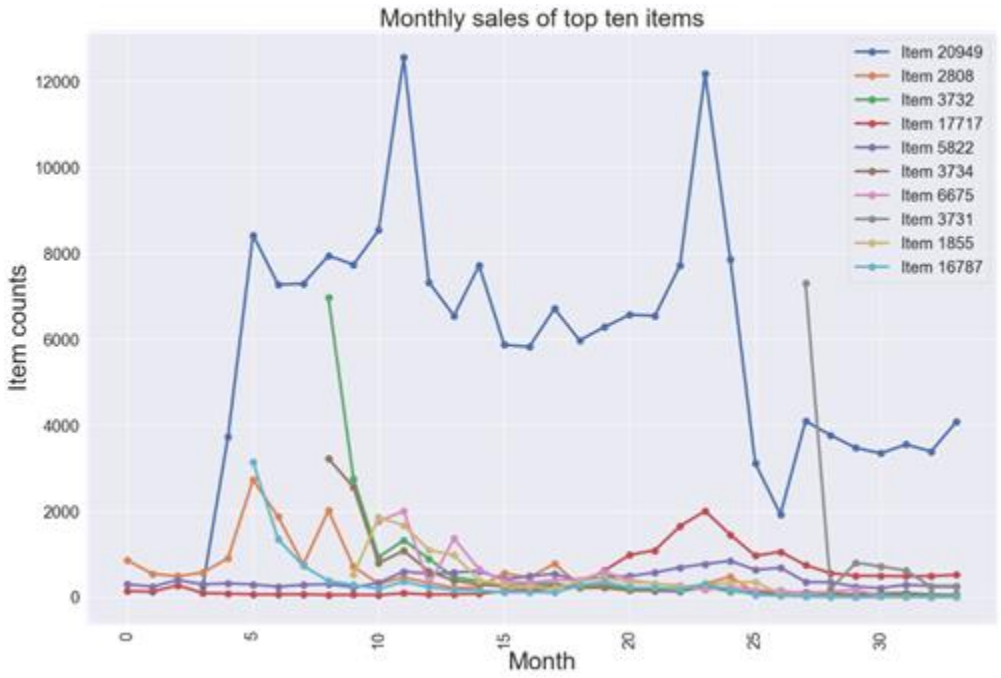

Figure 4. Top selling items.

There are some outliers as shown in **Figures 5 & 6**.

In case of items sold per day, two seemed too high, one 1000 and other over 2000. These were removed from the data. Also there were some negative numbers which were also removed. In case of item price, one was extremely high (> 300,000) which was also removed. Also there was negative item price which was also excluded from further analysis.

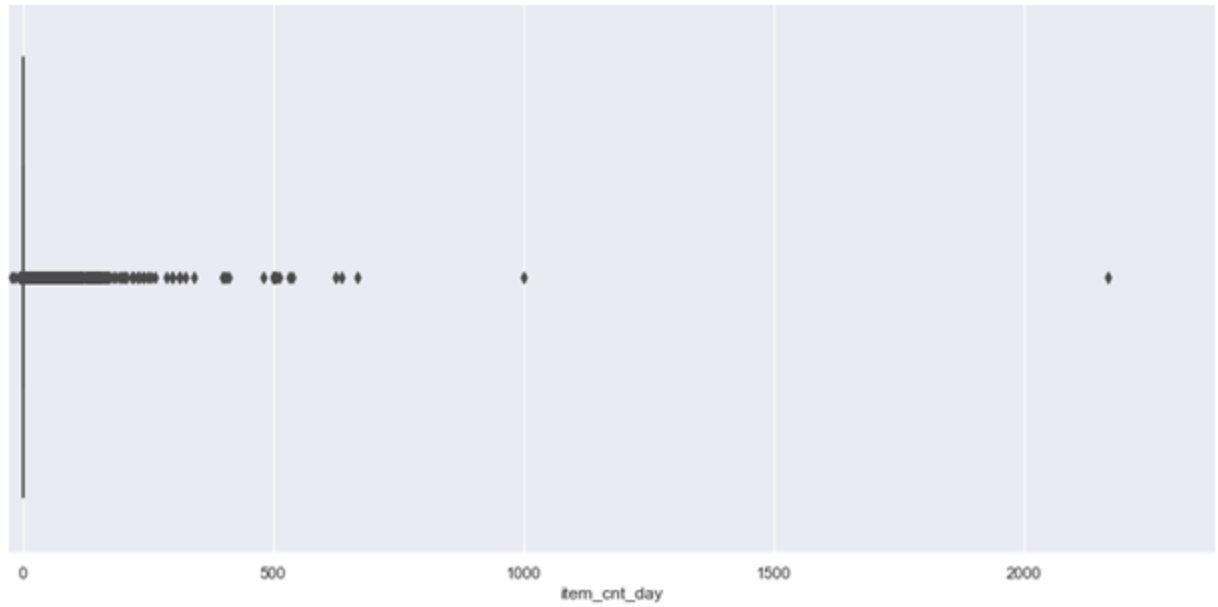

Figure 5: Number of items sold per day

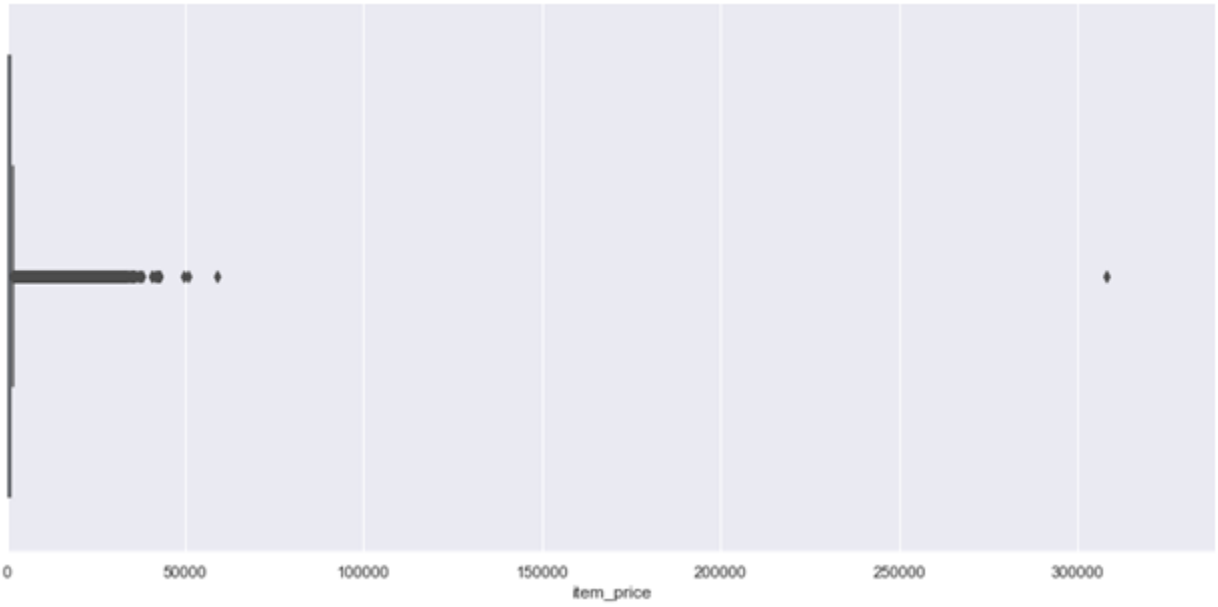

Figure 6: Item price

The goal of the project was to predict future sales. In order to make better predictions I looked at the trend of sales. I broke the dataframe into three (one for each year 2013, 2014 and 2015). Table 1 shows the total items sold per month in each of those years.

Table 1:

Month	2013	2014	2015
Jan	131,850	117,243	110,299
Feb	128,382	109,975	84,419
Mar	147,438	115,501	82,233
Apr	107,439	96,744	77,948
May	107,147	97,938	72,435
Jun	125,583	97,623	64,237
Jul	117,165	91,505	63,316
Aug	125,586	102,910	66,196
Sep	133,551	99,427	72,989
Oct	127,815	107,623	69,015
Nov	130,270	118,050	NaN
Dec	183,669	169,055	NaN

The above table is visualized in **Figure 7** below.

Figure 7: Monthly sales trend each year.

It seems that year 2013 and 2015 are much less correlated (corr coeff 0.31) with each other than to year 2014 (corr coeff 0.92 and 0.75 respectively). It is also obvious from Figure 7 that :- (i) sales are declining each year, (ii) year 2014 seems to be closer to 2015 than is 2013 with 2015. This suggests that it would give better prediction if we take more recent data. Thus, I excluded year 2013 entirely from the dataset for modeling.

Link to the document:

https://github.com/leukemia/Capstone_Projects/blob/master/03_Capstone_Project01_EDA.ipynb