Pankaj Acharya
Capstone Project01 - Data Story
April 30, 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.

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.

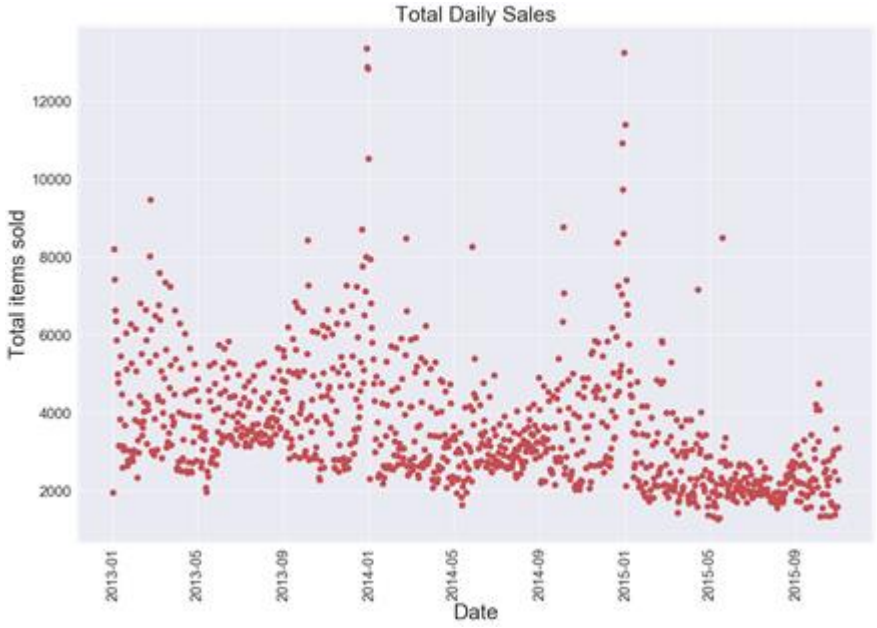

Figure 1. Total daily sales.

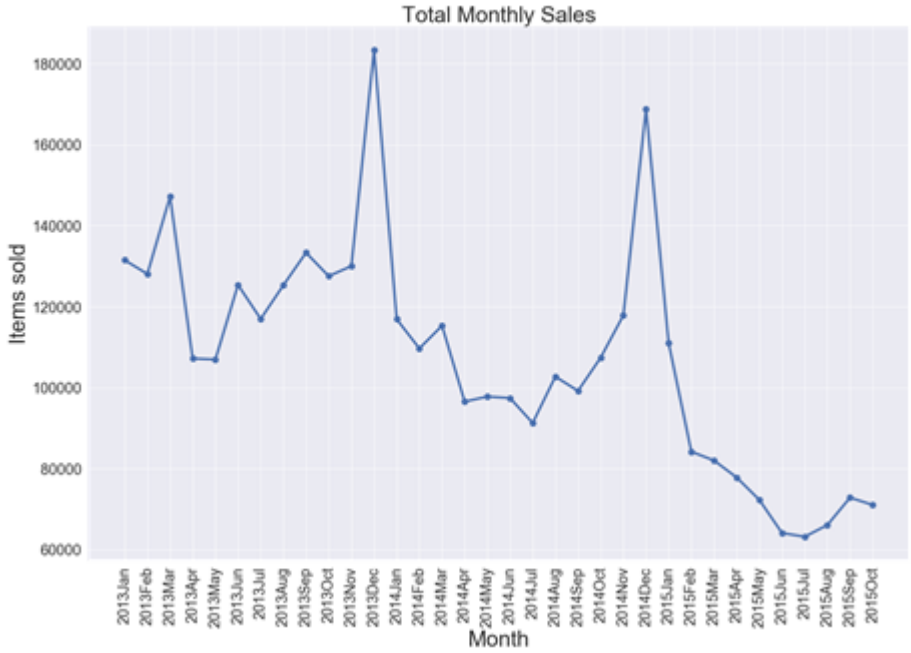

Figure 2. Total monthly sales.

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.

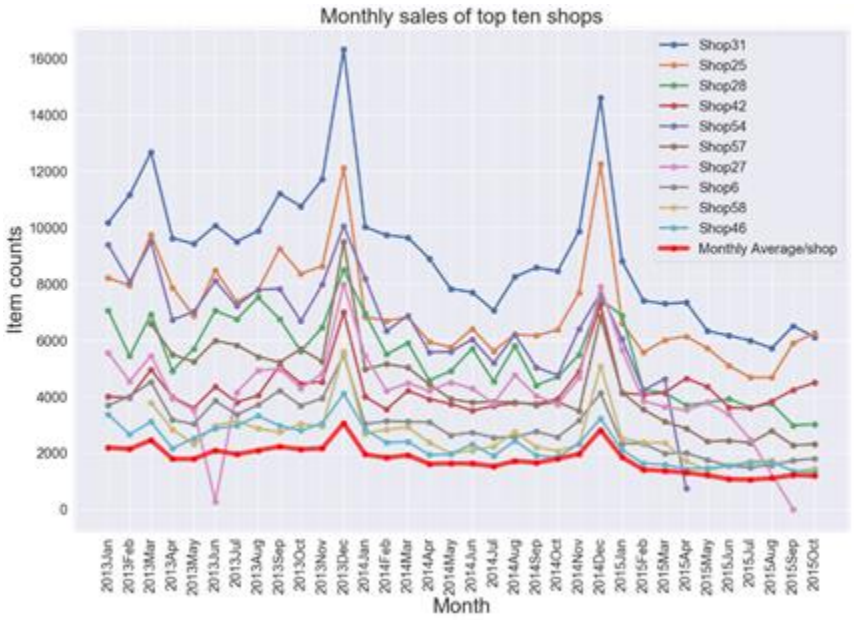

Figure 3. Performance trend of top ten shops over time.

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.

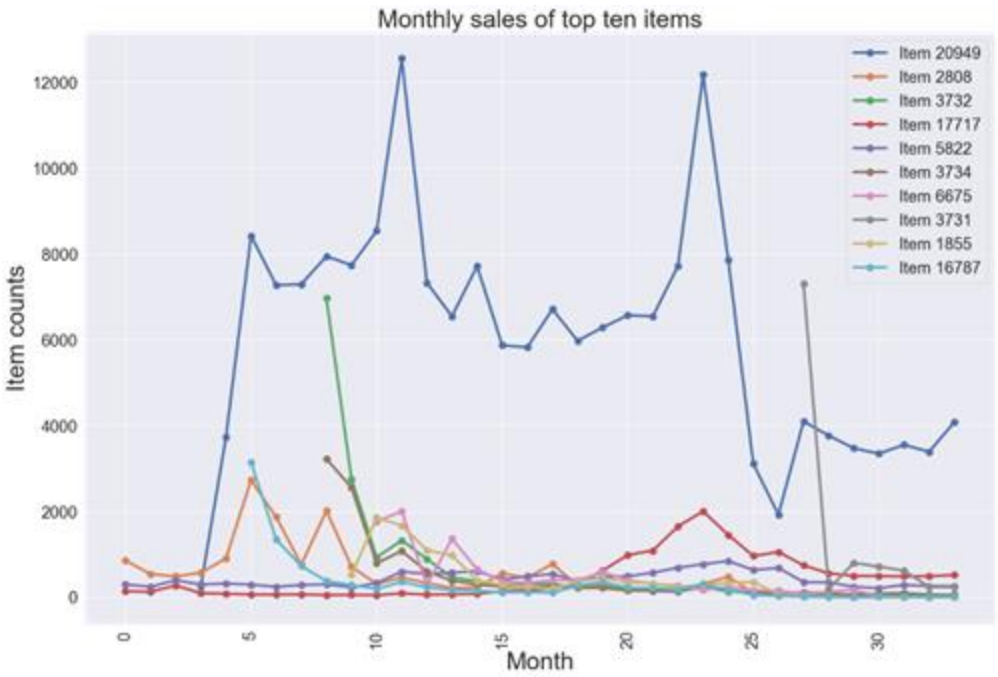

Figure 4. Top selling items.

Link to the document:

https://github.com/leukemia/Capstone_Projects/blob/master/02_Capstone_Project01_Data_Story.ipynb