PREDICT FUTURE SALES

Predict Future Sales

Collaborators in no order:

Temidayo Adejobi, Beniamkem Koffi, Haiming Luo, Karen Parra, Elanchezhian Vaithianathan

Internal

Contents

Report Version History

Version No.	Description
0.1	Draft version – introduction
0.2	Update the data loading and profiling section

1 Objectives

Objective of this report is to build a prediction model and predict the total sales for every store in the next month for the 1C Company. To tackle this problem, this requires data wrangling and cleaning, data transformation and model building. Predicting the future sales of one's business can be used as a benchmark, budget planning and planning for demand and supply for specific product items and store.

The dataset chosen on which the analysis is "Predict Future Sales" dataset

The dataset was downloaded from the public dataset on Kaggle at the url, https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data

The dataset is being used under the terms of the license below.

License: This work is licensed under the Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/.

2 About Dataset

Kaggle's Predict Future Sales dataset is a time-series dataset consisting of daily sales data provided by one of the largest Russian software firms – 1C Company.

1C Company is a leading Russian software development firm specializing in development, distribution, publishing and support of mass-market software. They are known for video game development and have several internal studios. Most popular titles produced by the company are *Il-2 Sturmovik*, *King's Bounty*, *Men of War* and *Space Rangers* series. 1C Company is the official distributor of top vendors such as Microsoft, Novell, Symantex, Borland and over 100 other software vendors

Any public user can download the Google Play store data from Kaggle at no cost. Users nneeds to register with Kaggle and sign-in to access this dataset.

Brief descriptions of the column names for the datasets from Kaggle are outlined in Tables 1 and 2.

2.1 Dataset "sales_train_v2.csv"

Sales_Train_V2 dataset is the core component of the dataset. It provides month-wise sales information of the shop with price as well. Refer to below table for the dataset info. This dataset contains close 3 million records.

Feature name	Description
date	Date of the sales
date_block_num	Consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
shop_id	Overall user rating of the app (as when scraped)

Item_id	Number of user reviews for the app (as when scraped)
Item_price	Size of the app (as when scraped)
Item_cnt_day	Number of user downloads/installs for the app (as when scraped)

Table 1 – Description of the "sales_train_v2" datasets

2.2 Other supplementary dataset.

In addition to sales_train_v2.csv, Predict future sales dataset includes following dataset

1. shops.csv - Shops id to shop name
2. item_categories.csv – item name, item id and item category mapping.
3. items.csv – mapping of item name to item id
4. test.csv - dataset to the prediction model.

Following table captures the data elements available in the supplementary dataset.

Feature name	Description
item_name	Name of the item
Shop_name	Name of the shop
Item_category_name	Name of the item category

Table 2: Description of "Supplementary dataset review" dataset

2.3 Data Analysis

Leveraged Python to perform the data analysis. Python and its rich modules provide rich capabilities to analyze, transform and visualize observations. Some of the key packages of python includes Numpy, Pandas, Matplotlib, Seaborn, Sklearn. Most of these packages were used in the analysis presented in this report.

Basic analysis of the sales_train_v2 dataset is given below

Info Summary

```
RangeIndex: 2935849 entries, 0 to 2935848
Data columns (total 6 columns):
date            object
date_block_num  int64
shop_id         int64
item_id         int64
item_price      float64
item_cnt_day    float64
dtypes: float64(2), int64(3), object(1)
memory usage: 134.4+ MB
```

Head info

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

Initial analysis revealed the little need for the data cleansing. There is no null filed in the main sales dataset. Shop name, item name and all the text are in the Russian language. However, it doesn't limit implementing and executing the prediction model. Date fields requires refinement to convert into a proper date format.

New data elements

- Splitting the date field into **Year, Month and Day** fields to help time series analysis.
- Add a new element - "Sales" using item_price and item_cnt_day. It provides the net value of the sales for a given month and shop id.
- Merge "items_category" field into the main data frame on item-id. This would provide category-based analysis of the sales.

View of the Train dataset post new data and merge is given below

	date	date_block_num	shop_id	item_id	item_price	item_cnt_day	day	month	year	sales	item_category_id
0	02.01.2013	0	59	22154	999.0	1.0	2	1	2013	999.0	37
1	23.01.2013	0	24	22154	999.0	1.0	23	1	2013	999.0	37
2	20.01.2013	0	27	22154	999.0	1.0	20	1	2013	999.0	37
3	02.01.2013	0	25	22154	999.0	1.0	2	1	2013	999.0	37
4	03.01.2013	0	25	22154	999.0	1.0	3	1	2013	999.0	37

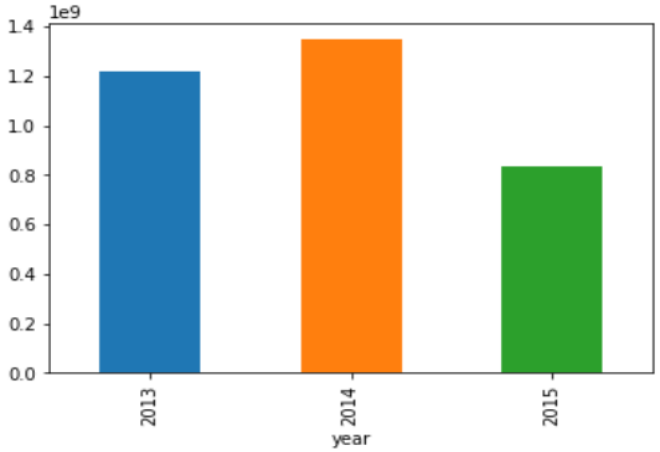
3 Feature Analysis

3.1 Seasonality Summary

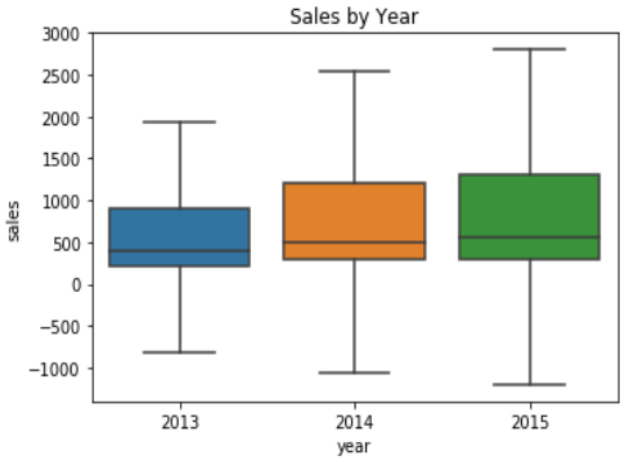
3.1.1 Sales by Year

- The year trend chart shows that 2014 total sales is higher than 2013. 2015 only contains 10-month data, so it is not comparable with the other 2 years.
- Judging from the plots, the Median, 1st quantile and 3rd quantile, as well as minimum and maximum for the daily sales for unique item for 2014 and 2015 are higher than 2013.

Total Sales by Year ($)



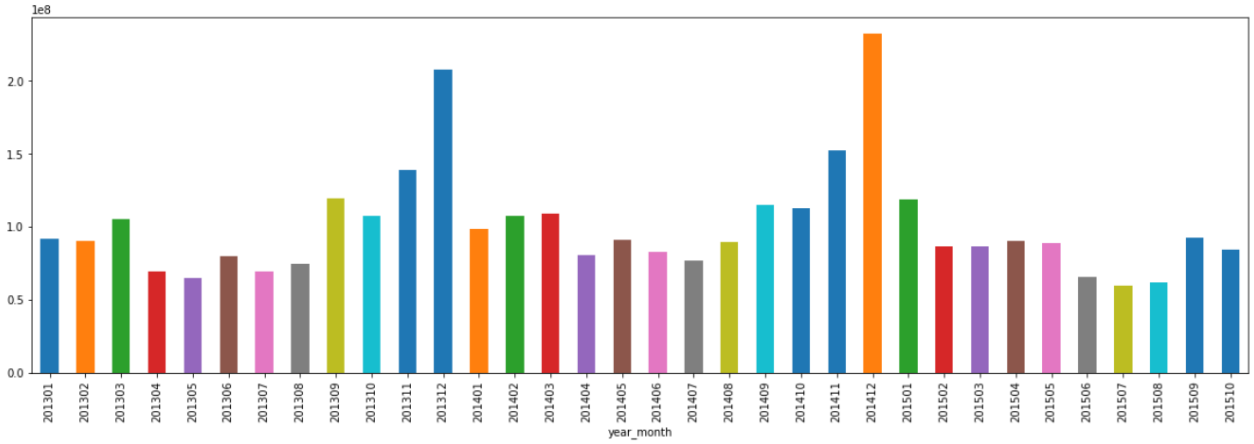
Daily Sales for unique item by Year ($)



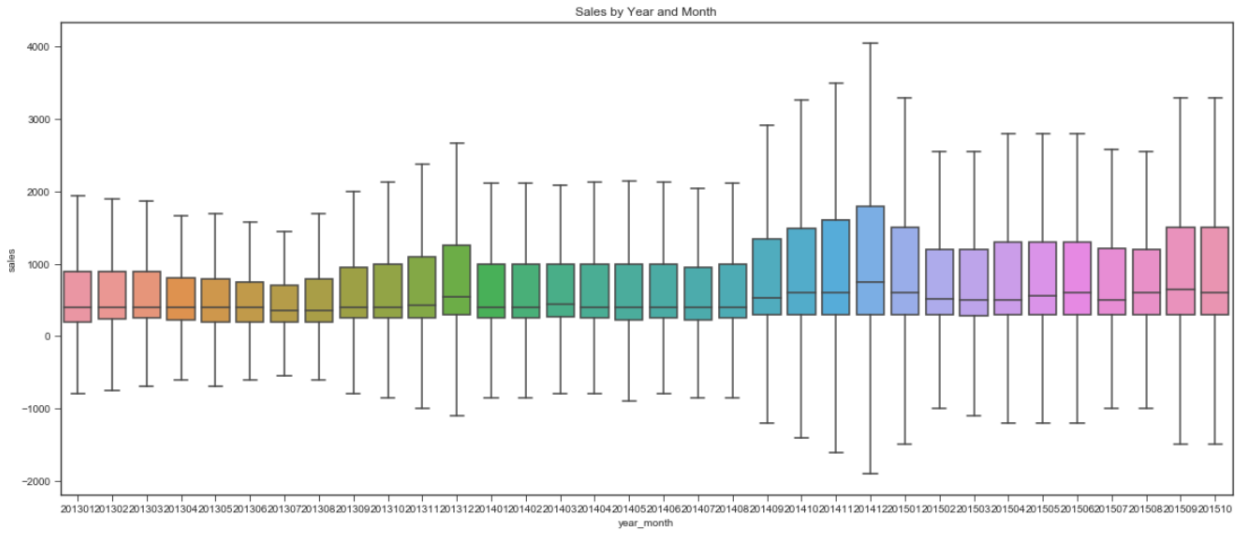
3.1.2 Sales by Year and Month

- Both the Bar Chart and the Boxplot show that the total sales or daily sales for unique item in the month of November and December are higher compared to other months.

Total Sales by Year and Month



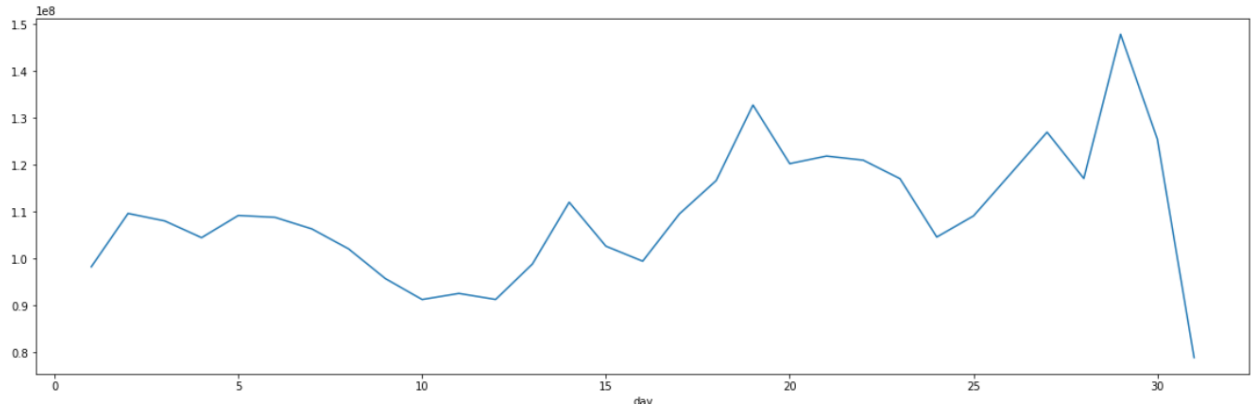
Daily Sales for unique item by Year and Month ($)



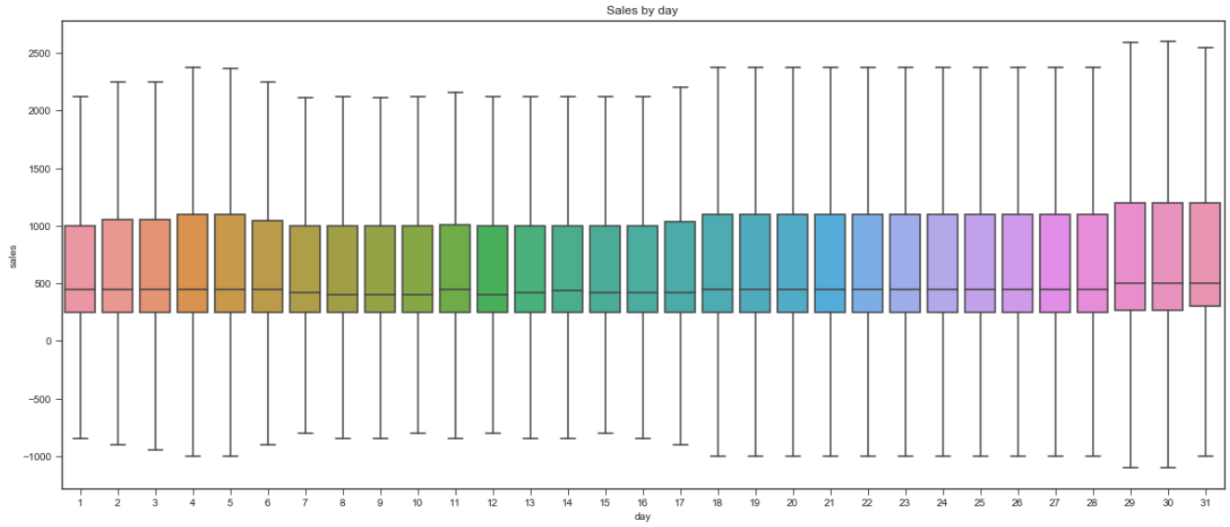
3.1.3 Sales by Day

- Bar Chart shows that the day 19 and 30 have outstanding peaks on total sales.
- Boxplot shows end of the month (day 29, 30, 31) has higher daily sales for unique item compared to other days within the month.

Total Sales by Day ($)



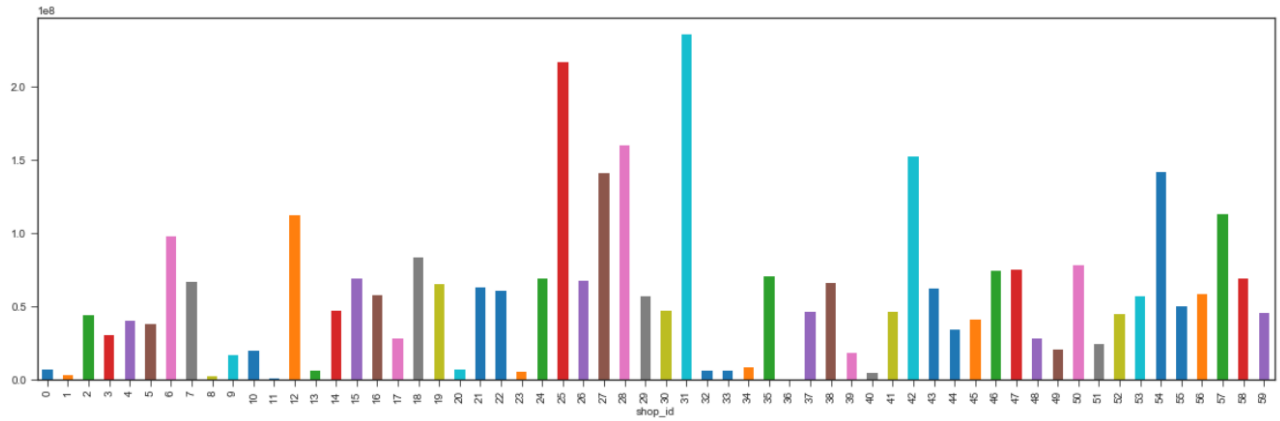
Daily Sales for unique item by Day ($)



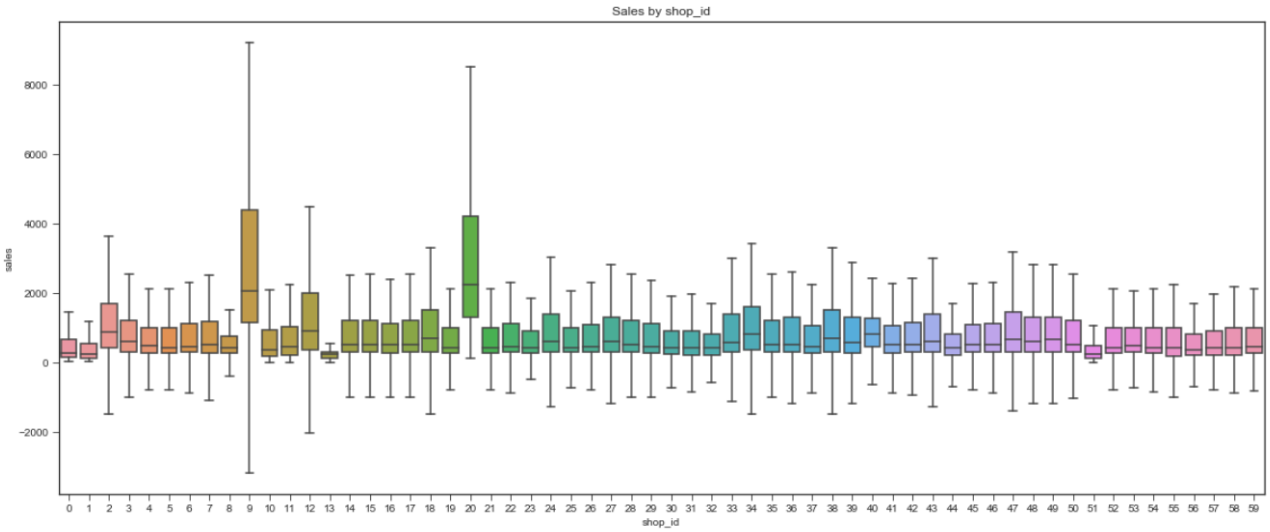
3.2 Sales Distribution by Shop

- The Bar Chart shows that Shop ID 31, 25, 42 has the top total sales.
- The Boxplot suggests that Shop ID 9 and 20 has the highest daily sales for unique single item.

Total Sales by Shop ID Day ($)



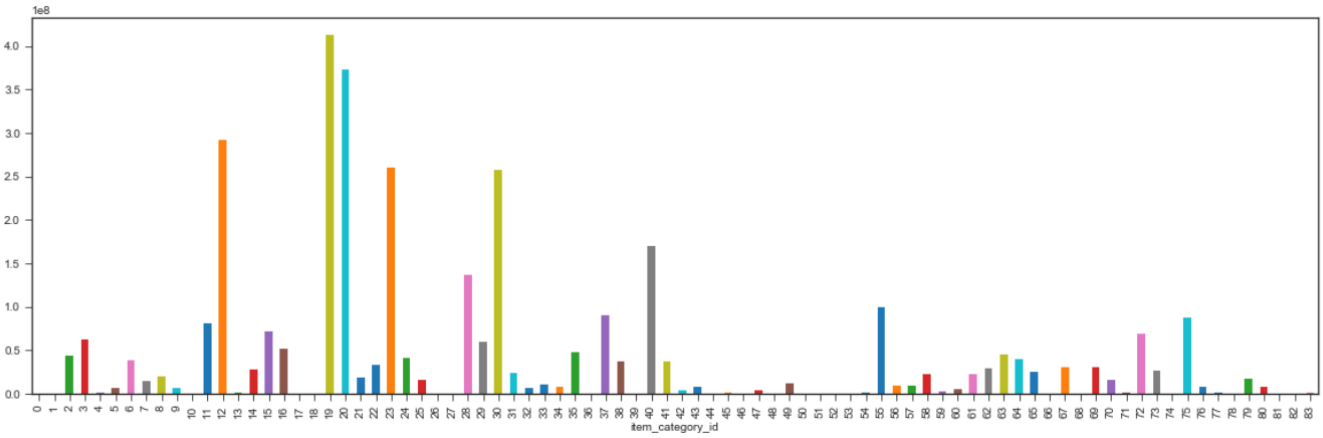
Daily Sales for unique item by Shop ID ($)



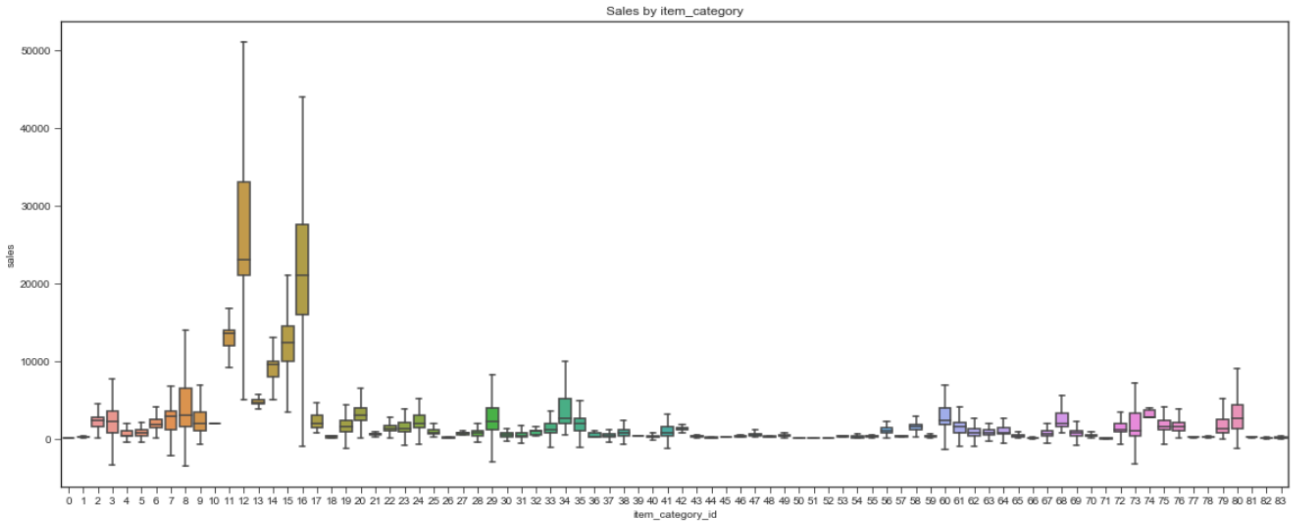
3.3 Sales Distribution by Item Category

- The Bar Chart shows that Item_catogory_id 19, 20 has outstanding high total sales.
- The Boxplot shows that Item_catogory_id 12 and 16 has the highest daily sales for single item.

Total Sales by Item Category ID ($)



Daily Sales for unique item by Item Category ID ($)



3.4 Shops and Item Variety

- Item Variety is calculated as the count if the unique item ID for Each Shop ID
- The Boxplot shows that he more variety of items the Shop has, the more the total sales.

- However, the following trend do not show obvious trend for the Shop that has more unique Item Category ID.
- In some shops, the more unique category ID, the less the sales.

4 Prediction model implementation

4.1 Considerations over train and test set

The data set present conditions that will determine a better adjustment of the predictions, it was considered exclude those items that do not appear in the shop during the time of study on the train set but appear in test set because it do not have history these items will be excluded. In total are 102,796 items that do not have history in each respective store during Jan 2013 - Oct 2015.

Another important review is the variable 'item_cnt_day' is defined like the number of products sold by each shop during a day, there are 2,941 registers after evaluating the first consideration that present 'item_cnt_day' <0 these cases were excluded from the study.

Analogously, there are items that haven't been sold in a specific shop in the last 7 months, this can be reflected in the pivot table considering the variable 'date_block_num ' that indicates a unique number for month and year vs. 'item_id' (See reference Table 4.1)

	item_id	0	1	2	3	4	5	6	7	8	...	24	25	26	27	28	29	30	31	32	33
0	30	0	600	394	105	45	46	30	16	12	...	12	13	4	4	5	4	4	6	2	1
1	31	0	466	155	45	25	21	13	20	18	...	22	11	10	13	4	10	6	52	9	18
2	32	225	156	143	80	53	65	79	65	45	...	36	25	34	19	19	25	21	30	19	22
3	33	42	30	26	13	11	40	38	29	37	...	17	21	20	12	11	11	15	14	17	16
4	38	0	0	0	0	0	0	0	0	0	...	4	4	1	0	3	2	4	7	3	0

Table 4.1 Pivot table to review last items sold

After the analysis, were discarded 168 items outdated with no sells in the last 7 months.

4.2 Modelling

Since the point of view of the objective to predict the total items sold by shop for Nov 2015, the focus of the model prediction is considering at the level of shops instead of item_id that could be an enhance of our approach. Both train and test set were grouped by month, shop. After seeing the summaries by month, it was detected some shops that does not have enough monthly history, it were considered only those shops with at least 4 months of history. The approach was considered in two ways considered a regression tree and an ARIMA adjustment by shop level defined by

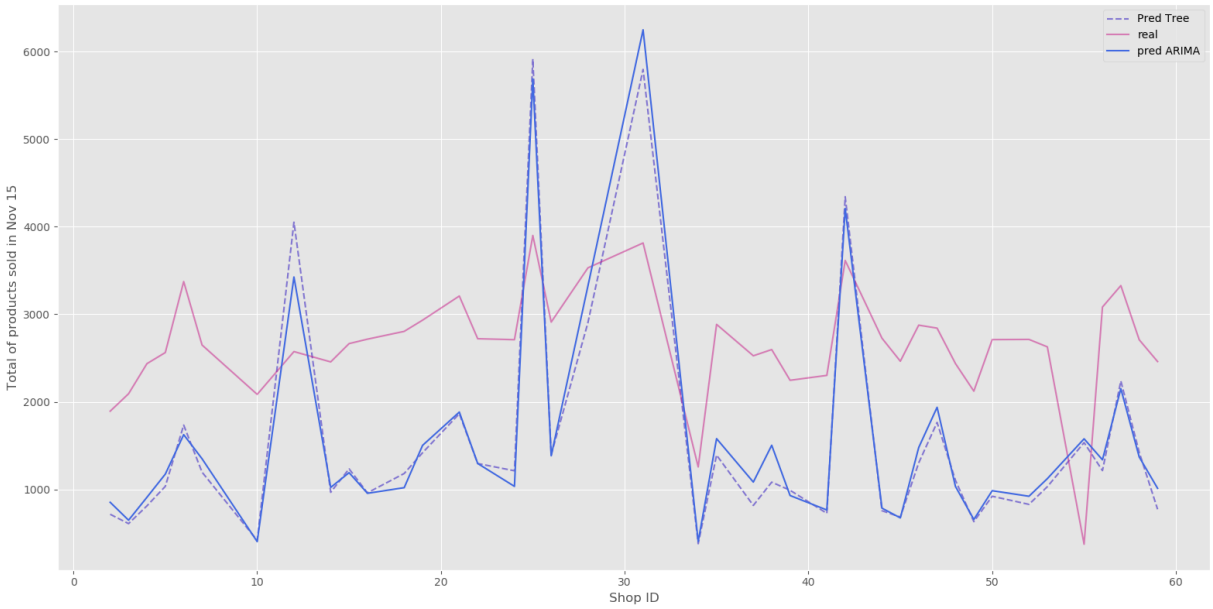
$$\widehat{y_{t,i}} = \mu + \phi_1 y_{t-1,i} + \ldots + \phi_p y_{t-p,i} - \theta_1 e_{t-1,i} - \cdots - \theta_q e_{t-q,i} \quad where\ i = shop\ level = \{\overline{1,41}\}$$

The RMSE for the regression tree was 38.68 and for the ARIMA models was 37.25

5 Execution of Model on Test dataset

According with the RMSE performances, the ARIMA got a better RMSE with 37.25

Graphically we can see the performance of the ARIMA and tree predictions vs. real test values for the 41 shops with products sold in Nov 15



6 Conclusion

The forecasting of the quantity of products sold by shop for Nov 2015 were predicted better with an ARIMA implementation with a RMSE of 37.25 instead of a regression tree with a RMSE of 38.68 since as the tree not detect patterns in trend if the serie is non stationary. Understanding in deep the approach of this model, it is important to remark that at the level of item_id can be obtained better performances since a detailed analysis can be done including hierarchal models using bayes analysis could be another technique to forecast at the level of shop and item id, for time considerations, it was considered to focus at the shop level, however, more extended techniques can be explored.

The data set even in training and test set were treated before applying the modelling phase related to some error factors in the original data set key to achieve better results like observe the shops with the items sold through the time and detect which items appear on the test set but do not have history, one approach that could be implemented is a cluster analysis at the level of shops to detect the more closer ones in groups and determine predictions according to each similar groups, including also the analysis of the categories and shop names initially are in russian language and can be used the library googletrans and clean the cases, in the project it was initially used the translations for shop names ought to the volume of products to categorize and cleaning of texts we decided not to use it for the project and work with the ID of products and shops.

7 Associated files

The files associated to this report are,

File name	Description
Group5-Assign_PredictFutureSales_Final.ipynb	All the code is in this Jupyter notebook
'sales_train_v2.csv	the training set. Daily historical data from January 2013 to October 2015
Shops.csv	Supplemental information about the shops
item_categories.csv	Supplemental information about the items categories
Items.csv	Supplemental information about the items/products
test.csv.csv'	Need to forecast the sales for these shops and products for November 2015

8 References

Available at: https://www.kaggle.com/c/competitive-data-science-predict-future-sales/