Group 5 - Assignment Summary

Abstract

Objective of assignment is to build a prediction model to predict the total sales for every store in the next month for the 1C Company based on the training data. The dataset chosen for the analysis is **"Predict Future Sales"** dataset. Dataset was downloaded from the public dataset on Kaggle at the URL, https://www.kaggle.com/c/competitive-data-science-predict-future-sales/data.

Key dataset includes **"Sales_Train_V2.csv"** which provides item wise, month-wise sales information of the shop. It also includes price and the item category. Dataset contain close to 3 million records. Other support dataset includes, shops.csv, item_categories.csv, items.csv, test.csv provides further elaboration of the main data.

Data Analysis - Initial analysis revealed the little need for the data cleansing. Shop name, item name and all the text within the dataset are in the Russian language. However, it doesn't limit implementing and executing the prediction model. Created new data elements such **Year, Month and Day** fields to help time series analysis and **sales** elements from item_price and item_cnt_day. Merger of the item_category from from item_categories.csv ensured the model can manipulate category wise sales.

	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

Figure 1 : Sales_Train_V2 dataset after cleanup

Feature Analysis

Feature analysis of the time series data is provided below

Seasonality Summary

- **Sales by year** - Year trend chart shows that 2014 total sales are higher. 2015 only contains 10-month data, so it is not comparable with the other 2 years.
- **Month wise Sales - Analysis** of e Bar Chart & the Boxplot show that the Total Sales in the month of November and December are higher compared to other months.
- **Sales by Day -** Both the Bar Chart and the Boxplot show that the day 19 and day 30 have highest total sales.

Sales and volume by shops

- The Bar Chart shows that Shop ID 31, 25, 42 has the top 3 total sales.
- The box plot suggests that Shop ID 9 and Shop ID 20 has the highest daily sales for single item
- The Bar Chart for volume shows that Shop ID 31, 25, 54 has the top 3 total sales volume.
- **Sales by Day -** Both the Bar Chart and the Boxplot show that the day 19 and day 30 have highest total sales.

Prediction Model & Execution

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

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

Test data vs Prediction Model Observation

Following diagram provides the view of the model comparison. The RMSE for the regression tree was 38.68 and for the ARIMA models was 37.25

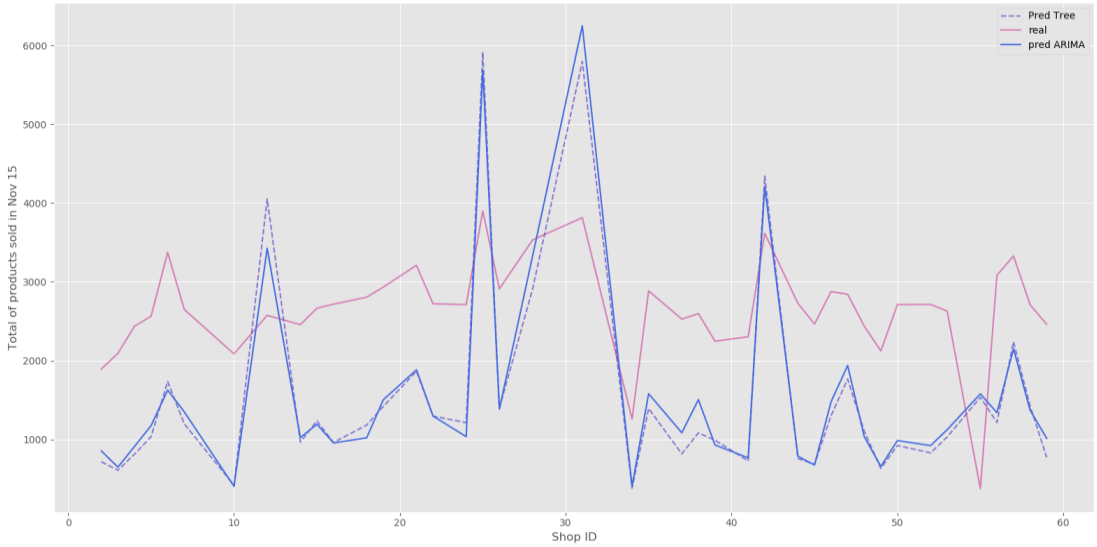

Figure 2 :Performances Regression Tree, ARIMA vs. Real products sold in each shop Nov 2015

Conclusion

We have successfully built a sales prediction model using ARIMA that predicts the future sales data for a given data set instead of the regression tree, it is important to mention that regression trees 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. Another level of complexity in the data set is that the categories and shop names initially are in russian language, can be used the library 'googletrans' and clean the cases even considering extract locations of shop names that it could be implemented tuning features to the original data set. In general, more extended techniques can be explored.