Predict Future Sales

Predict Future Sales

Collaborators in no order:

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

**Contents**

[1 Objectives 3](#_Toc26213368)

[2 About Dataset 3](#_Toc26213369)

[2.1 Dataset “sales\_train\_v2.csv” 3](#_Toc26213370)

[2.2 Other supplementary dataset. 4](#_Toc26213371)

[2.3 Data Analysis 4](#_Toc26213372)

[3 Feature Analysis 6](#_Toc26213373)

[4 Prediction model implementation 6](#_Toc26213374)

[5 Execution of Model on Test dataset 6](#_Toc26213375)

[6 Conclusion 6](#_Toc26213376)

[7 Associated files 6](#_Toc26213377)

[8 References 6](#_Toc26213378)

**Report Version History**

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

# Objectives

Objective of this report is to build a prediction model and predict the total sales for every product and store in the next month for the 1C Company. To tackle this problem, this requires data wrangling and cleaning, data transformation to make it stationary and supervised 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/.

# 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 n needs 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.

## 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**

## 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 |

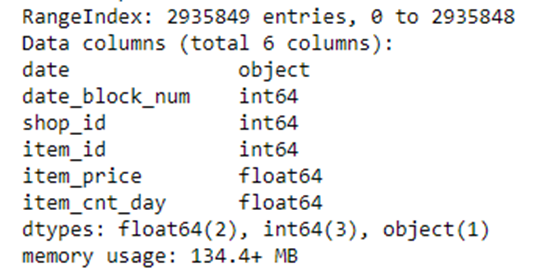
**Table 2: Description of "Supplementary dataset review" dataset**

## Data Analysis

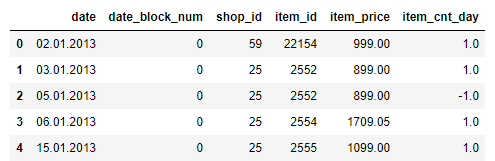
Leveraged Python to perform the data analysis. Python and its rich modules provide rich capabilities to analyze, transform and visualize observations. Some of the key packages of python includes Numpy, Pandas, Matplotlib, Seaborn, Sklearn. Most of these packages were used in the analysis presented in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

Basic analysis of the sales\_train\_v2 dataset is given below

**Info Summary**

****

**Head info**

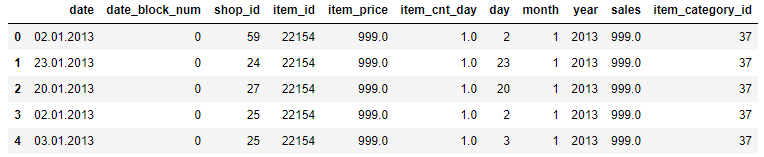
****

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



# Feature Analysis

# Prediction model implementation

# Execution of Model on Test dataset

# Conclusion

# Associated files

The files associated to this report are,

|  |  |
| --- | --- |
| **File name** | **Description** |
| Group5-Assign\_PredictFutureSales\_V1.0.ipynb | All the code is in this Jupyter notebook |
| ‘sales\_train\_v2.csv |  |
| Shops.csv |  |
| item\_categories.csv |  |
| Items.csv |  |
| test.csv.csv’ |  |

# References

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