Description

Overarching goal :

The overarching goal of the project is to predict total sales for every product and store in the next month. The dataset contains time - series data consisting of daily sales data, by one of the largest Russian software firms .

datasets:

You are provided with daily historical sales data.

File descriptions

- **sales_train.csv** - the dataset. Daily historical data from January 2013 to October 2015.
- **items.csv** - supplemental information about the items/products.
- **item_categories.csv** - supplemental information about the items categories.
- **shops.csv**- supplemental information about the shops.

Data fields:

- **shop_id** - unique identifier of a shop
- **item_id** - unique identifier of a product
- **item_category_id** - unique identifier of item category
- **item_cnt_day** - number of products sold. You are predicting a monthly amount of this measure
- **item_price** - current price of an item
- **date** - date in format dd/mm/yyyy
- **date_block_num** - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
- **item_name** - name of item
- **shop_name** - name of shop
- **item_category_name** - name of item category

tasks:

1. Data Preparation, Exploratory Analysis:

CLEANING DATA

- Look for any Missing Data
  - Identify observations on Missing Data
  - Graph the representations of Missing Data
  - Decide whether to populate or to remove Missing Data
  - Identify the possible impact of it while Modelling

2. OUTLIER DETECTION:

Find out the months of sale which are considered as outliers for any shop.

3. FEATURE SELECTION/ENGINEERING:

Identify and convert Categorical columns/values to Numerical representation using Dummy Variables if suitable for modelling

4. MODELING:

Build 2 different models that uses all data from the training.csv and other files for all the months and years except October 2015.

5. Validation:

Use October 2015 data as test set and present

a. Mean squared error

b. root mean squared error (RMSE)

6. Compute confidence interval of Model 1 and Model 2 for the following different confidence levels:

80%, 90%, 95%

7. Compare these two models considering:

a. error

b. efficiency in training time (scalability)