Mid Term Report

Rahul Sangole[1]

[1] Northwestern University

Abstract

This reports the work done for the Predict 413 Section 55 Midterm assignment, Summer 2018.

Mid Term Report

Overview of Methodology Used

The input data consists of four files, detailing information about the shops, items, item categories and the daily sales information over Jan 2013 to Oct 2015. The textual data is in Russian, which is converted to English first. After data preparation activities, extensive Exploratory Data Analysis (EDA) is performed on the data. This involved univariate numerical and graphical summaries, multivariate graphical and numerical summaries, and unsupervised time series clustering. The EDA results in insights which feed in to the data preparation step, feature engineering step as well as modeling and post processing steps. A total of 8 models are built including one ensemble model. All the models follow a top-down approach as described later. Some standard model evaluation metrics are used during the model building process, though the final model selected is dependent on the Kaggle score.

Data Preparation & Exploratory Data Analysis

Translation

The text fields of the input dataset are in Russian. The first step is to convert these fields into English. This is performed passing the Russian text to a Google Translate API [**] via an R script running on an Amazon Web Services RStudio server. Since a free version of the Translate API is used, the translation activity runs at a speed of ~1 translation per second resulting in an end-to-end runtime of ~3 hours. Post translation, the shop meta data, item category metadata and item level metadata becomes readable and allows for feature engineering. Table 1 shows a sample of `items` translated into English.

Table 1

Sample Items

item_name	item_id	item_category_id
Risen [PC, Digital Version]	6146	21
* LINE OF DEATH D	11	40
OTHER WORLD (region)	11319	62
1C: Audiobooks. Mandelstam Osip. Egyptian Brand (Jewel)	311	45
ARMSTRONG LOUIS Ambassador Of Jazz Box 10CD + Book (box)	1433	31

Feature Engineering

Categorical Predictors. The `item category` can be split into two levels of information, as shown in table 2. `itemcat_lvl1` is a higher level categorization consisting of 21 different levels *(Cinema, Games, PC Games, Music, Gifts, Movies, Accessories, Books, Programs, Payment Cards, Game Consoles, Office, Elements of a food, Clean media (piece), Delivery of goods, Tickets (figure), Official, Clean carriers (spire), Android games, MAC Games, PC)*, while `itemcat_lvl2` is a lower level categorization consisting of 62 different levels. A sample is shown in table 2.

Table 2

Sample Item Categories

item_category_name	itemcat_lvl1	itemcat_lvl2	item_category_id
Books - Artbook, encyclopedia	Books	Artbook, encyclopedia	42
Games - Accessories for games	Games	Accessories for games	25
Payment Cards - Windows (Digital)	Payment Cards	Windows (Digital)	35
Music - CD of branded production	Music	CD of branded production	56
Accessories - XBOX 360	Accessories	XBOX 360	6

The `shops` table consists of some location information about the shops. This is split

into two categorical predictors as well. `loc_lvl` is a higher level categorization consisting of 32 different levels, while `loc_lvl1` is a deeper categorization consisting of 56 levels. A sample is shown in table 3.

Table 3

Sample Shops

loc_lvl1	loc_lvl2	shop_id
Moscow	TC Perlovsky	30
Krasnoyarsk	Shopping center June	18
Moscow	TC Budenovskiy (pav.K7)	24
RostovNaDonu	TRC Megacenter Horizon	39
Volzhsky	shopping center Volga Mall	4

Calendar Related Predictors. Temporal predictors are appended to the dataset, viz.,

1. `year`, `month`, `week` describing the year, month and week of the observation
2. `weekend` is a binary 0/1 variable to account for increased sales over a weekend (if any)
3. `ym`, year-month combination
4. `yw`, year-week combination
5. `is_december`, is a binary 0/1 variable to account for increased Christmas / New Year sales (if any)

Russian holiday schedules are downloaded for the years 2013, 2014 and 2015 from "Holidays in Russia, 'https://www.timeanddate.com/holidays/russia/2013#!hol=9". These are filtered to *Official and Non-Working Days* are joined to the original data.

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

Table 4

2013 Russian Holidays

date	holiday_name	holiday_type
2013-01-01	New Year's Day	National holiday
2013-01-02	New Year Holiday Week	National holiday
2013-01-03	New Year Holiday Week	National holiday
2013-01-04	New Year Holiday Week	National holiday
2013-01-07	Orthodox Christmas Day	National holiday, Orthodox
2013-01-08	New Year Holiday Week	National holiday
2013-02-23	Defender of the Fatherland Day	National holiday
2013-03-08	International Women's Day	National holiday
2013-05-01	Spring and Labor Day	National holiday
2013-05-02	Spring and Labor Day Holiday	National holiday
2013-05-03	Spring and Labor Day Holiday	National holiday
2013-05-09	Victory Day	National holiday
2013-05-10	Defender of the Fatherland Day holiday	National holiday
2013-05-10	Victory Day Holiday	National holiday
2013-06-12	Russia Day	National holiday
2013-09-01	Day of Knowledge	De facto holiday
2013-11-04	Unity Day	National holiday
2013-12-30	New Year Holiday Week	De facto holiday

Data Exploration

Time Series Exploration.

t-SNE.

Time Series Clustering.

Oddities.

Data Cleansing

removing -ve counts closed shops spikey stuff

Modeling Details

High level view of approaches

Detailed Model List

Code Details

Results Summary

Performance Evaluation

Challenges

R Packages Used

References

```
## Warning in readLines(file): incomplete final line found on 'r-
## references.bib'
```

Figure captions