Rossmann Sales Prediction Milestone Report

1 Introduction

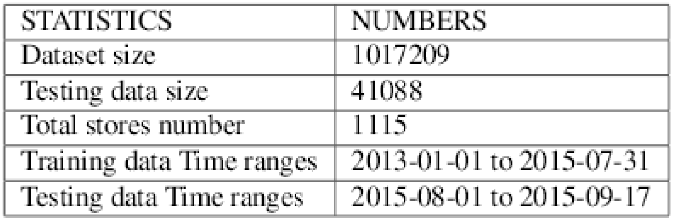
Rossmann, a German drug store has around 1115 stores across the country and has provided the past sales information of 3 years. Currently stores manager face the problem to predict their daily sales for up to six weeks in advance. Sales are influenced by many factors, including promotions, competition, school, state holidays and location. With thousands of individual managers predicting the sales based on their own individual's situation, the accuracy can be significantly varied. Reliable forecasts enable these managers to create the better scheduling of staff, improve supply chain and increase productivity in general. With the given past data, the task is to predict future sales and help managers to focus on the other important issues.

2 Datasets

Datasets provided by Rossmann through Kaggle are three files 1) train.csv 2) test.csv and 3) store.csv.

"train.csv" has 1017209 rows of data representing the daily sales of different stores from year 2013 to year 2015. "store.csv" has additional features (41088 lines because there are 41088 observations of "test.csv" 1115 stores) that are complementary to both "train.csv" and "test.csv". "test.csv" has 41088 lines of daily sales of 1115 stores over the period of three months. Our task is to predict the sales value of "test.csv" by using the "train.csv" and "store.csv".

Dataset statistic



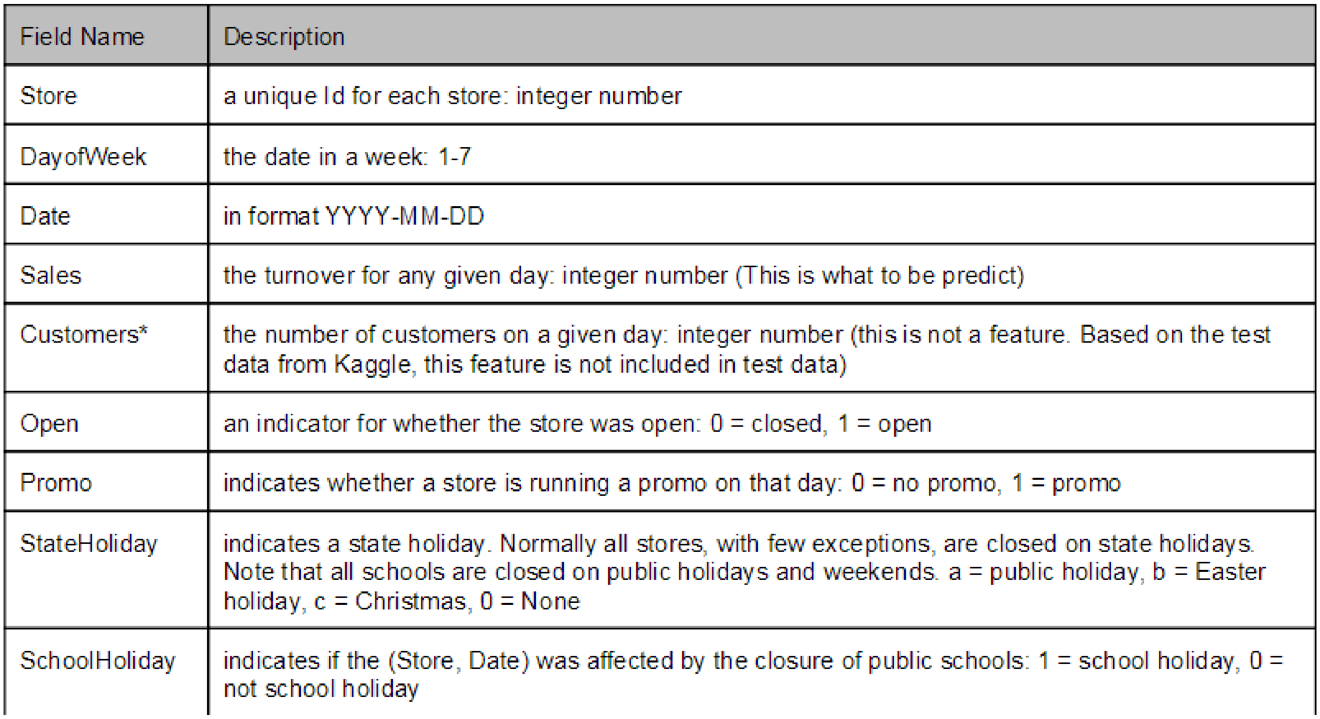
STATISTICS	NUMBERS
Dataset size	1017209
Testing data size	41088
Total stores number	1115
Training data Time ranges	2013-01-01 to 2015-07-31
Testing data Time ranges	2015-08-01 to 2015-09-17

2.1 Limitations and data wrangling

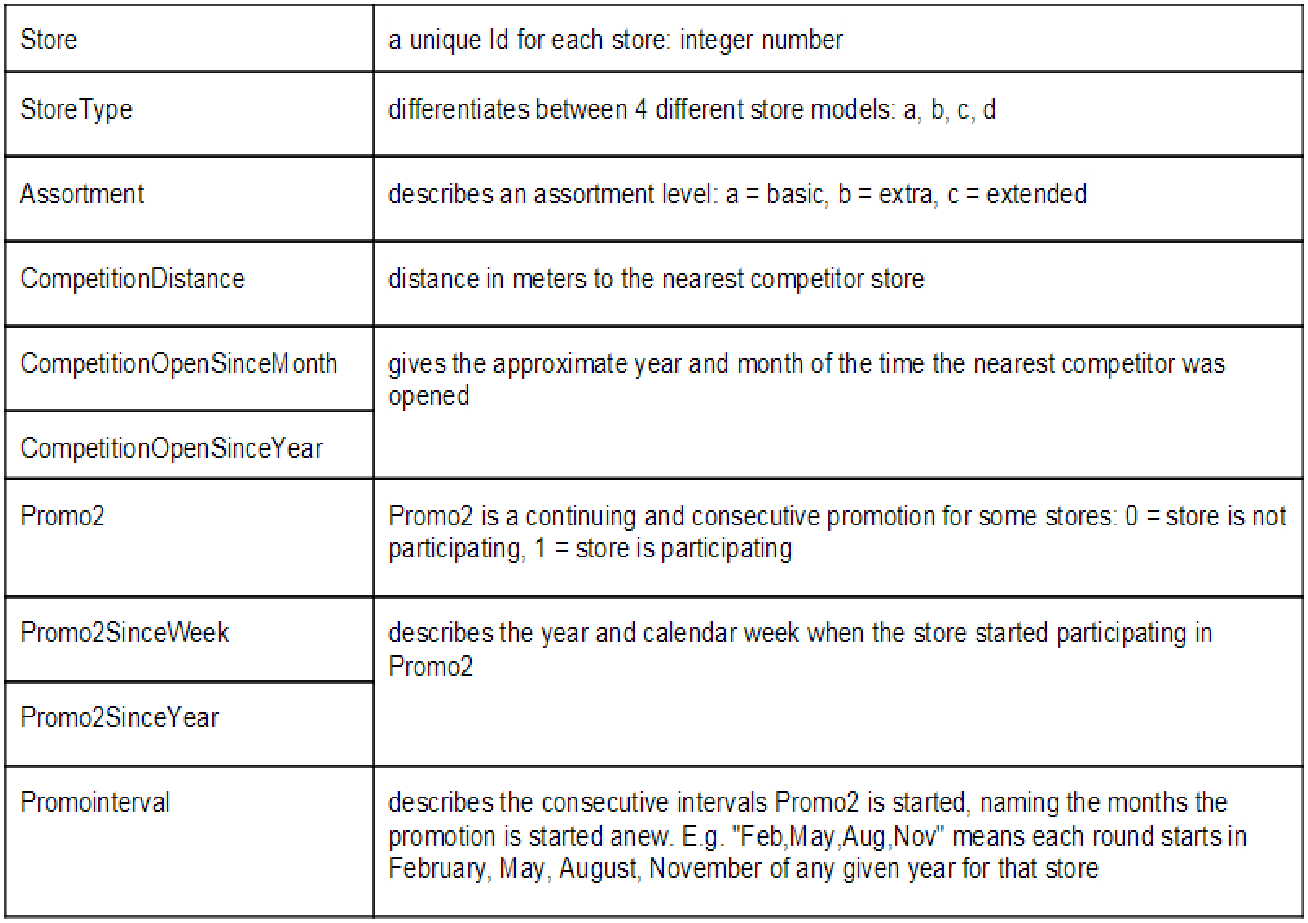
- "train.csv" doesn't have the complementary fields that necessary to predict the sales and these fields are in "store.csv" therefore one has to merge the two files.
- In order to predict the sales of future days, we needed the "day", "year", "month" columns. These field were not given and were extracted from the "Date" column.
- Store status is "0" i.e the stores is closed and therefore no sales were made. We removed these stores from 'train" data as there was no point to have these stores to predict the future sales.
- Store number "622" has status as "closed in "test" data and its status was changed to "open" so as to predict the sales correctly.
- The whole "customer" column is unavailable in "test.csv" that was fixed by merging "test.csv" and "store.csv".
- The data doesn't give any information about weather that could be very important feature affecting sales.
- The data doesn't give any information about demographic statistics making it difficult to predict the trend i.e young people tend to buy more often than their counterpart.

- **Accessibility**: There is no information about how people can access stores. Generally if it stores are right next to train and bus station people prefer them over the other stores and this affects the sales
- **Number of POS**: How many POS do the stores have that reduces the time from buying cha paying the product’s
- **Available staff**: No information is provided about the staff that are able to help customers in guiding, locating and selecting the products. This feature attracts customers and hence increases the sales.
- **Online purchase/lead delivery time/ pick up location:** This information is also missing. People tend to buy more from Ikea as their lead delivery time is less. Goods are delivered right to their doors as compared to picking up from local post office.
- **In-Stores Café and kids play area:** No information is provided about these two important feature that tend to attract customers and increase sales.

2.2 Dataset Information

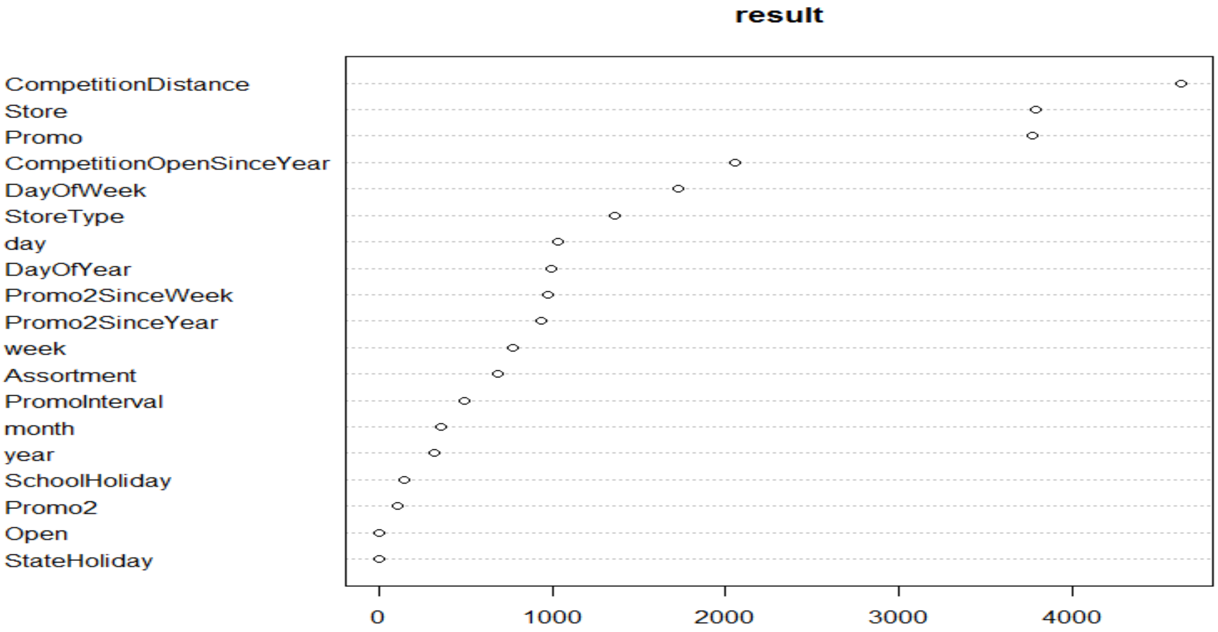


Field Name	Description
Store	a unique Id for each store: integer number
DayofWeek	the date in a week: 1-7
Date	in format YYYY-MM-DD
Sales	the turnover for any given day: integer number (This is what to be predict)
Customers*	the number of customers on a given day: integer number (this is not a feature. Based on the test data from Kaggle, this feature is not included in test data)
Open	an indicator for whether the store was open: 0 = closed, 1 = open
Promo	indicates whether a store is running a promo on that day: 0 = no promo, 1 = promo
StateHoliday	indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = public holiday, b = Easter holiday, c = Christmas, 0 = None
SchoolHoliday	indicates if the (Store, Date) was affected by the closure of public schools: 1 = school holiday, 0 = not school holiday



Store	a unique Id for each store: integer number
StoreType	differentiates between 4 different store models: a, b, c, d
Assortment	describes an assortment level: a = basic, b = extra, c = extended
CompetitionDistance	distance in meters to the nearest competitor store
CompetitionOpenSinceMonth	gives the approximate year and month of the time the nearest competitor was opened
CompetitionOpenSinceYear	
Promo2	Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
Promo2SinceWeek	describes the year and calendar week when the store started participating in Promo2
Promo2SinceYear	
Promointerval	describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

2.3 Important Dataset features



The above is result of GBM variable importance selection. As can be seen “CompetitionDistance” has highest effect sales and then “Store”, “Promo”,”DayOfWeek” respectively.

3 Preliminary Exploration

- When we compare train and test data, we find all test stores are available in train data but 259 stores are not in test data.
- Sales are highest on Sunday and Monday and almost even on other days.
- There is strong +ve correlation between Sales and Promotion
- Few stores were open but didn’t have sales that might be due to the reason: some customers just do window shopping and not actually buy anything.
- Strong correlation also exist between number of customers and sales as one can imagine.
- Type B stores never closed with possible exception of renovation.
- Assortment level ‘b’ is only offered at StoreType ‘b’
- All type B stores have comparatively higher sales mostly on weekends. That is also plausible as people tend to do more shopping on weekends.
- Store also show high sales on days before and after closing days. It is also understandable as people tend to stock the goods in anticipation of any disruption that in this case is stores status “closed”.

4 Our approach

As mentioned earlier, our task is to predict the sales for “test” data based on given “train” data. We will use the 1) Random Forest Model and 2) Gradient Boost Model

1) **Random Forest Model:** Random Forest Tress tries to construct a multitude of decision trees. Then it classifies the data into the decision tree node and for each node it calculate the mean value and use this value for prediction. Random Forest tree uses random amount of data for training. With this randomized data it is hard for random forest tree to overfit and that makes it more easier to tune the data compared to Gradient Boosting Tree.
2) **H20 Gradient Boosting Model:** A GBM is an ensemble of either regression or classification tree models. Both are forward-learning ensemble methods that obtain predictive results using gradually improved estimations. Boosting is a flexible nonlinear regression procedure that helps improve the accuracy of trees. Weak classification algorithms are sequentially applied to the incrementally changed data to create a series of decision trees, producing an ensemble of weak prediction models. While boosting trees increases their accuracy, it also decreases speed and user interpretability. The gradient boosting method generalizes tree boosting to minimize these drawbacks.