Rossmann Sales Prediction using store, promotion and Competitor information

Saleem Ahmad
saleemamd@gmail.com

Abstract

Prediction of number of customers for future business that will drive the sales is a common topic and perhaps the most important challenge the companies around the world are facing today.

Think of Volvo construction or H&M or Netflix or Amazon; these all share one common goal: **"Sales"**. If customer bought the product then how can companies learn the buying behaviour to predict the next sale or attract/retain the customer? This is also known as **churn rate**. In order to increase customer spending (thus sales) and overall revenue, reducing churn rate is crucial for the business success.

I was interested in solving the problem (even an old one) that would relate most to the real life and can be applied in data science world. After all my purpose with the DS course is to learn, prove my capability and solve a problem that can be related to diverse businesses in profound way and not become just another interesting study.

With the help of my mentor I was able to find the most suitable data set that encompassed all the issues related to customer's retention or prediction of possible revenue. This paper examines the previous data of a German store Rossmann and predicts the future sales in order to help the manager stay focused on staffing and increasing motivation.

Keywords

Prediction; Sales; Forecasting; Random Forest; Gradient Boost Model

Introduction

Rossmann, a German drug store has around 1115 stores across the country and has provided the past sales information of three 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 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.

"Store" is the only given related feature and there is no customer related data making it hard to recommend any prediction. Additionally, target is continuous, the problem could be a regression one based on both categorical feature (StoreType) and

continuous feature (Days). Some feature can be considered as categorical as well as continuous.

Next we will perform the selection and construction of features and how the past sales data will help to forecast the future sales.

Dataset and features

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

Limitations of Data and Data wrangling

The above mentioned three files didn't suffer from any serious data anomalies; few following data wrangling steps were performed in order to use the right features for our prediction task.

- "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 fields 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 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 to paying for the product
- **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.

Data Exploration

This section will try to analyse the datasets and make sense of the useful features that will be helpful to predict our sales. We will examine "train.csv" and then to make more meaningful picture we will look into "store.csv".

Store ID

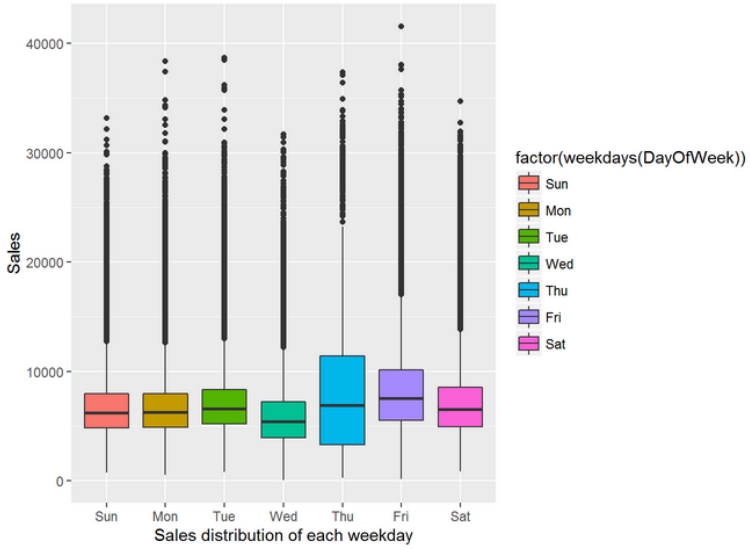
One can think that Store ID would be related to Sales. But we find that correlation is only 0.005126226.

```
with(train, cor.test(x=Store, y = Sales, metho
d = 'pearson'))
```

Day of Week

Every store has different sales on different days as people tend to shop on different days depending on their comfort. DayOfWeek has a large effect on Sales. The correlation coefficient is -0.46.

```
with(train, cor.test(x=DayOfWeek, y = Sales,
method = 'pearson'))
```



Open

Open with status "1" shows the stores as open or otherwise on a specific day. Because the sales must be zero if the store is closed; we excluded these stores in our analysis.

State Holiday

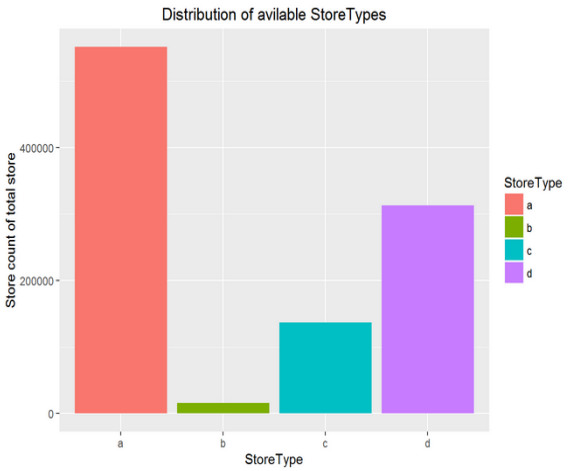
Most stores are closed on State holidays and people tend to stock drugs before and after the holiday.

School Holiday

School holiday has correlation of only 0.09

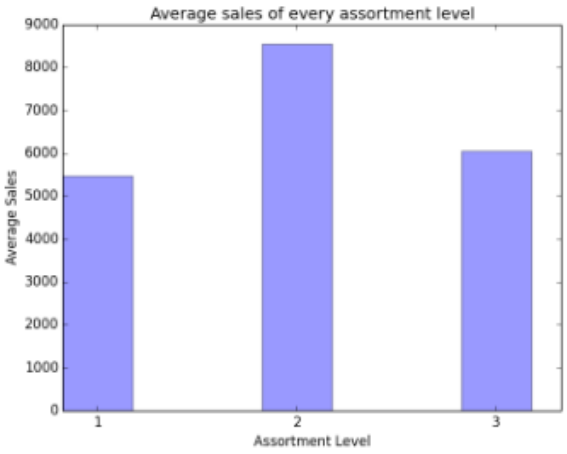
Store Type

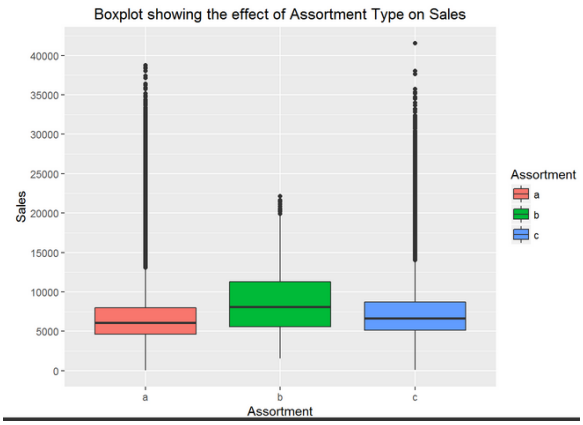
There are four different store models a, b, c and d.



Assortment

Different stores have different assortment level and thus making assortment also as a feature. The average sale for each Assortment is as following.





Year

The sales might change in different years. This feature was extracted from Date column of train.csv

Month

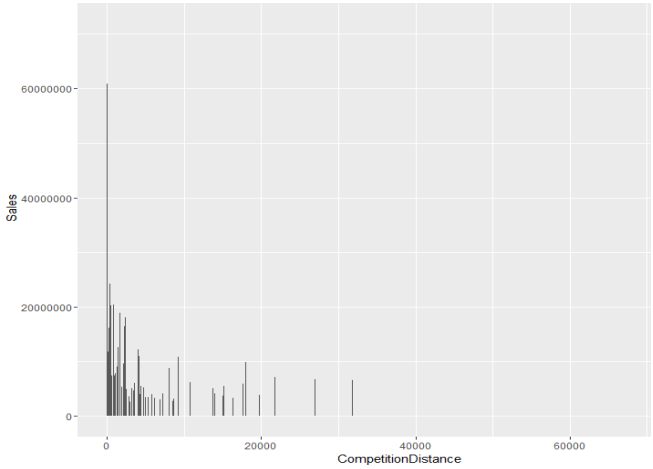
People have different needs depending on weathers during different months, so it would affect sales as well. This feature was extracted from Date column of train.csv.

Day

Individual day affect the sales as well. People will spend more depending on which day they get paid.

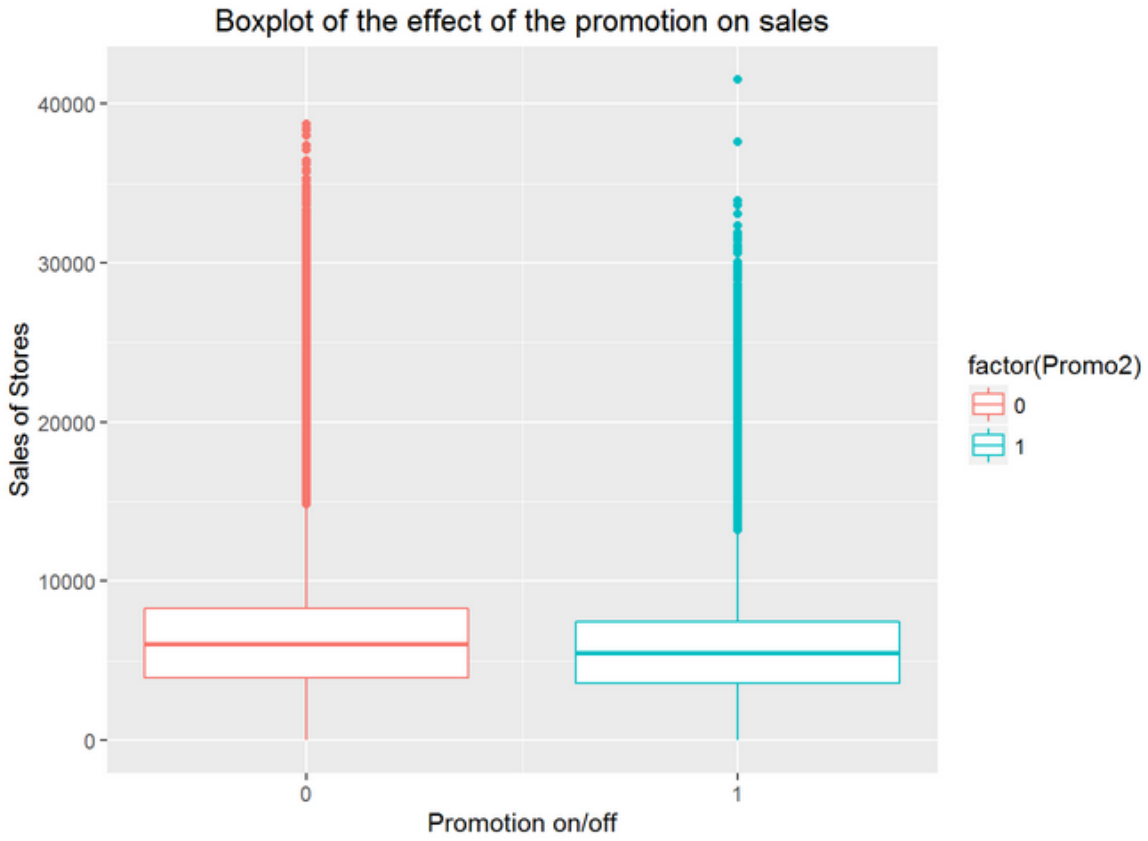
Competition Distance

It is natural that if a competitor is located near to the store that must affect store's sales

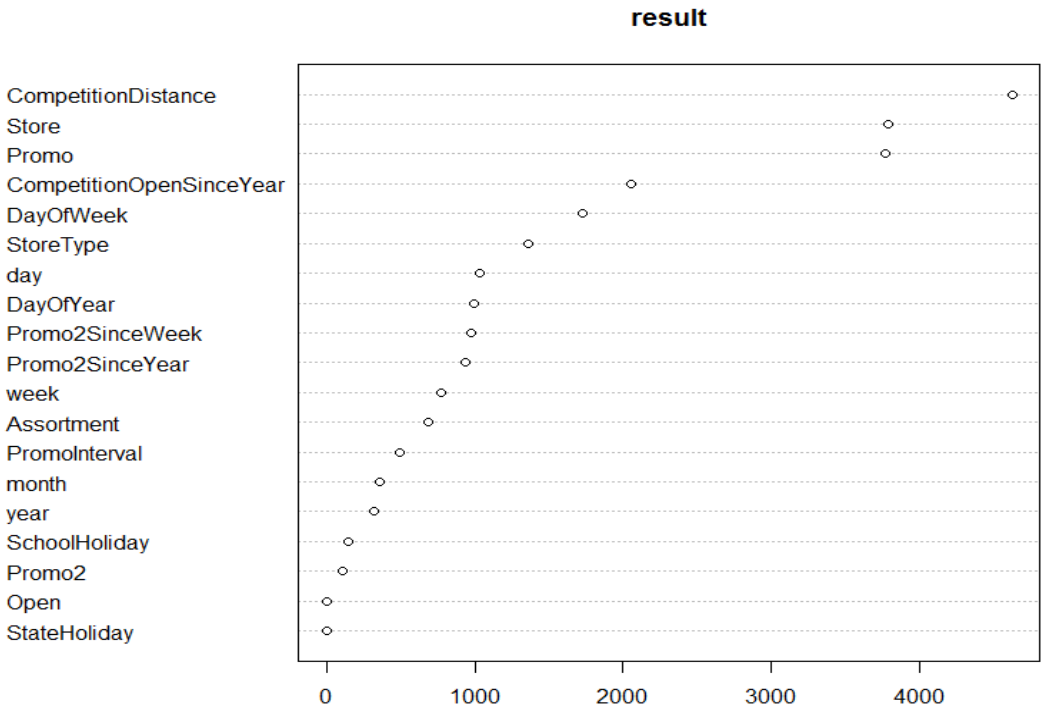


Promo

Promo will be used to attract the customers that in turn will affect sales and thus we have taken it as a feature.



Features Importance



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

EDA Exploration Findings

- 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 are never closed with possible exception of renovation.
- All type B stores have comparatively higher sales mostly on weekends. That is also plausible as people tend to do more shopping on weekends
- Competition distance has a huge effect on sales.

Our approach

As mentioned earlier, our task is to predict the sales for "test" data based on given "train" data. We skipped other prediction method like linear regression, logistics regression, SVC regression and rather directly forwarded to ensemble learning methods, which are proven to be the better methods and were also great from our learning point of view. Two of the most respected ensemble methods are Random Forest tree and Gradient Boost tree.

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 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.

Results

Sales Prediction by using Random Forest

```
result <-
randomForest(train1[,variable.names],
             log(train1$Sales+1),
             mtry=5,
             ntree=150,
             max_depth = 50,
             sampsize=150000,
             do.trace=TRUE)
```

Id	Sales
1	5492.377
2	5696.619
3	4265.786
4	6274.903
5	4225.764
6	6407.893
7	4253.402
8	5594.105
9	4809.631
10	7446.457
11	6395.057
12	4440.037
13	4983.975
14	4322.891
15	5924.284
16	4671.969
17	4698.053
18	4645.445
19	6730.357
20	4977.543
21	4642.034
22	5837.102

Sales Prediction by using H20.GBM

```
resultGbm <- h2o.gbm(x=variable,
             y="logSales",
             training_frame=trainGbm,
model_id="introGBM",
             nbins_cats=1115,
             sample_rate = 0.5,
             col_sample_rate = 0.5,
             max_depth = 50,
             learn_rate=0.05,
             ntrees = 150 )
```

Id	Sales
1	4148.529
2	4370.353
3	4828.661
4	5128.957
5	0.749573
6	4022.979
7	3615.046
8	3483.728
9	3515.348
10	3526.567
11	4301.487
12	0.749574
13	4225.029
14	4723.789
15	4780.204
16	5714.716
17	5582.889
18	7055.229
19	0.749904
20	4221.043
21	3988.575
22	3337.31

Future Recommendations

- We find from the "Feature Importance" graph that both StateHoliday and SchoolHoliday are not very important features as we thought. Just from a "Google search" it is interesting to find out that different states in Germany has different public holidays. So having that information and incorporating them into the above model would be a good idea.
- People tend to stock before and after the closing day. So creating these two features and using them into our model would also be interesting to see the effect they will create on sales.
- As pointed out earlier weather data would also make a significant impact on drug sales as cold weather causes t flu and dry weather creates allergy and skin problems thus affecting the sales.

References:

1) Data Source:www.kaggle.com/c /rossmann-store-sales
2) https://cran.r-project.org/
3) http://stackoverflow.com
4) http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html#Data%20Science%20Algorithms
5) www.cyclismo.org/tutorial/R/tables.html