Rossmann sales Prediction

Saleem Ahmad
DS foundation course
Mentor: Matt Fornito

Content

- Introduction of the problem
- Benefits of solution
- First look at data
- Exploratory data analysis
- Model Selection
- Learning and recommendations

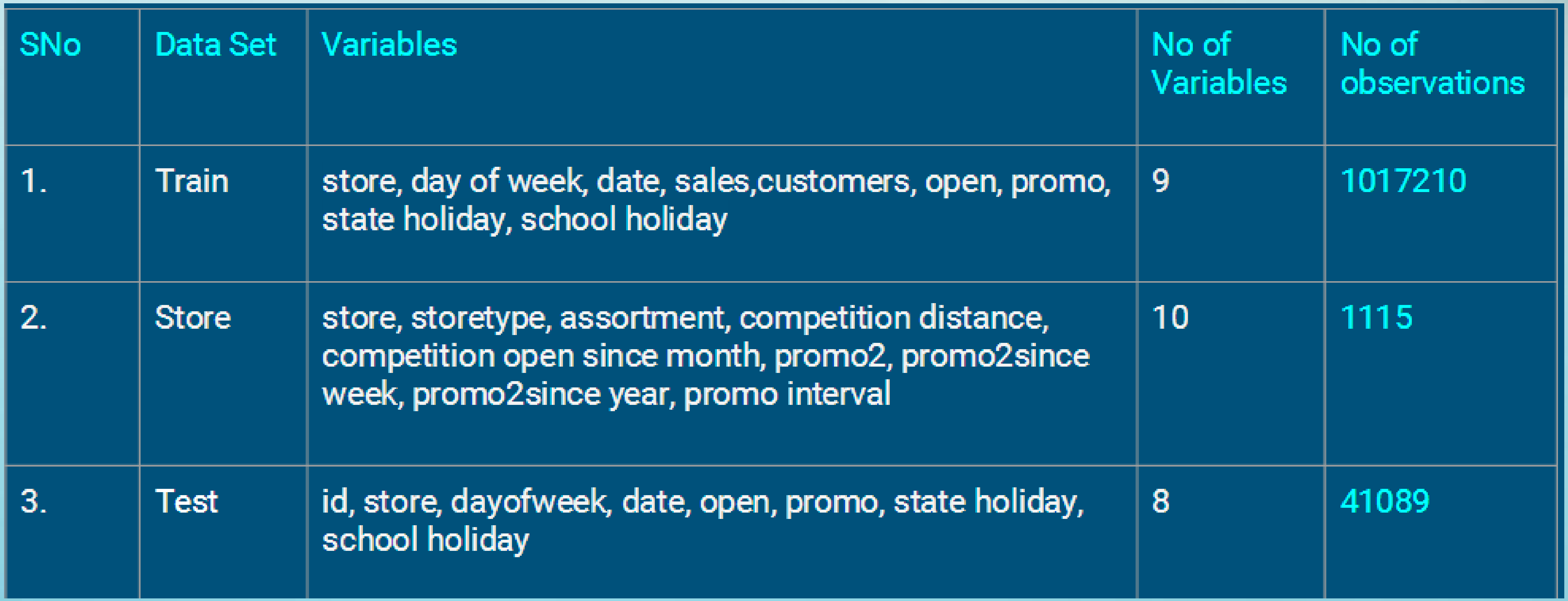
Introduction

- Rossmann is a German drug store with 1115 stores in Germany and more store in the EU.
- Task is to forecast the daily sale of 1115 different stores 6 weeks in advance.
- Historical data of 2 years 7 months is provided (Jan 2013 –Jul 2015)

Benefits of Solution

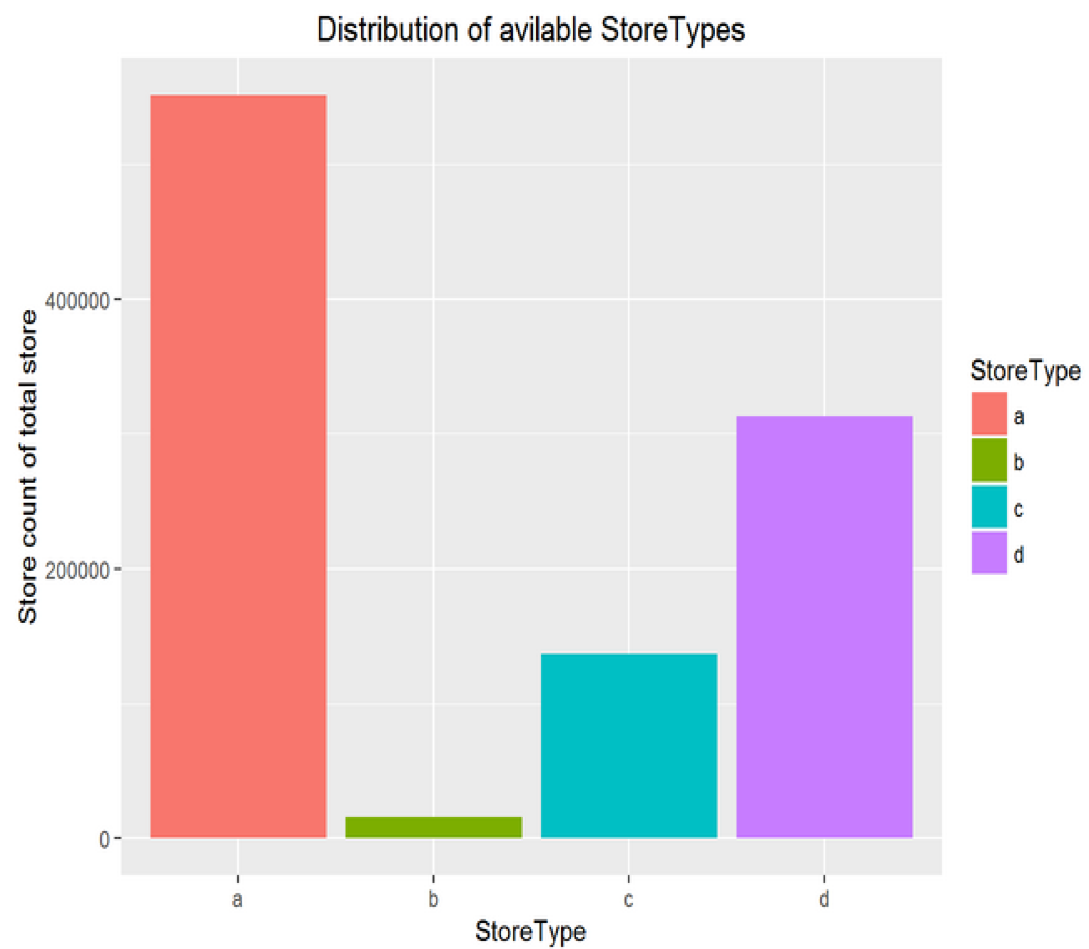
- Mangers can better be prepared for inventory i.e not have too many and not run out the goods either
- Managers can devote their valuable time on staff scheduling
- It will provide enough time for managers to work with what is important i.e customers and colleagues.

First look at data

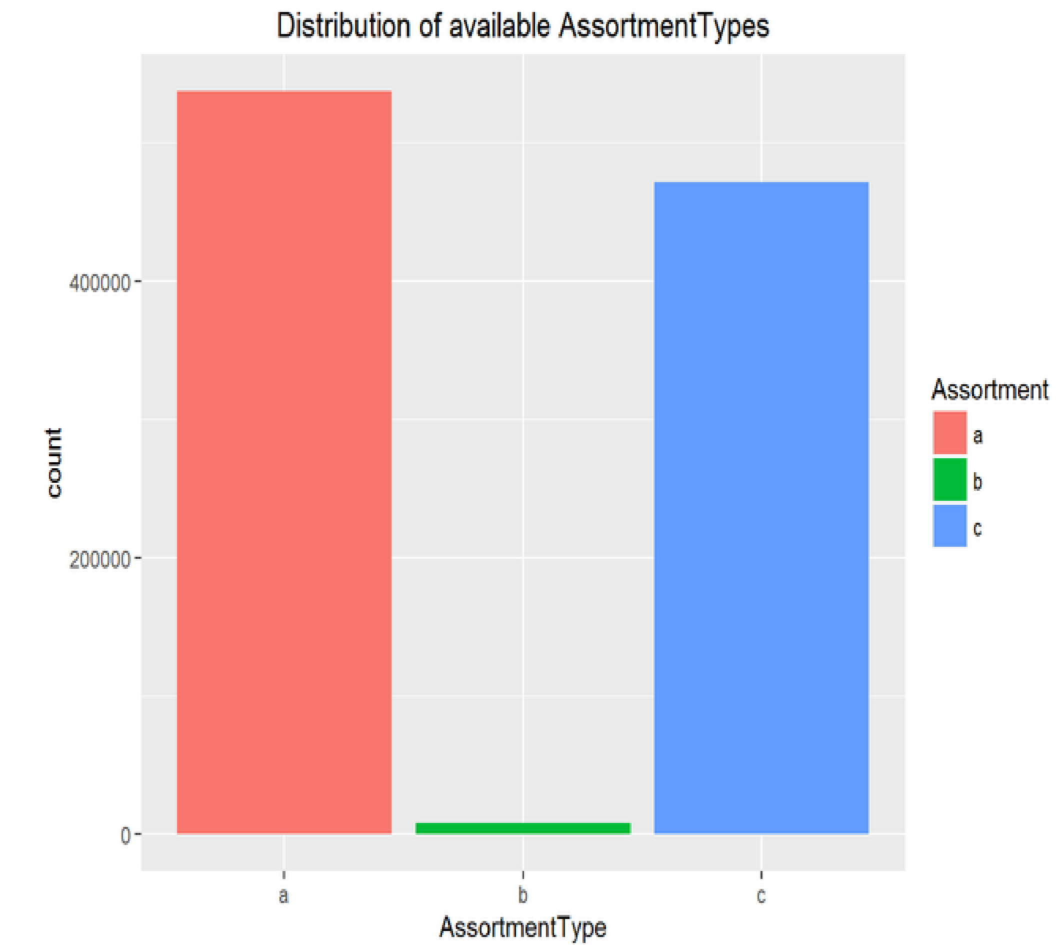


SNo	Data Set	Variables	No of Variables	No of observations
1.	Train	store, day of week, date, sales,customers, open, promo, state holiday, school holiday	9	1017210
2.	Store	store, storetype, assortment, competition distance, competition open since month, promo2, promo2since week, promo2since year, promo interval	10	1115
3.	Test	id, store, dayofweek, date, open, promo, state holiday, school holiday	8	41089

Exploratory Data Analysis

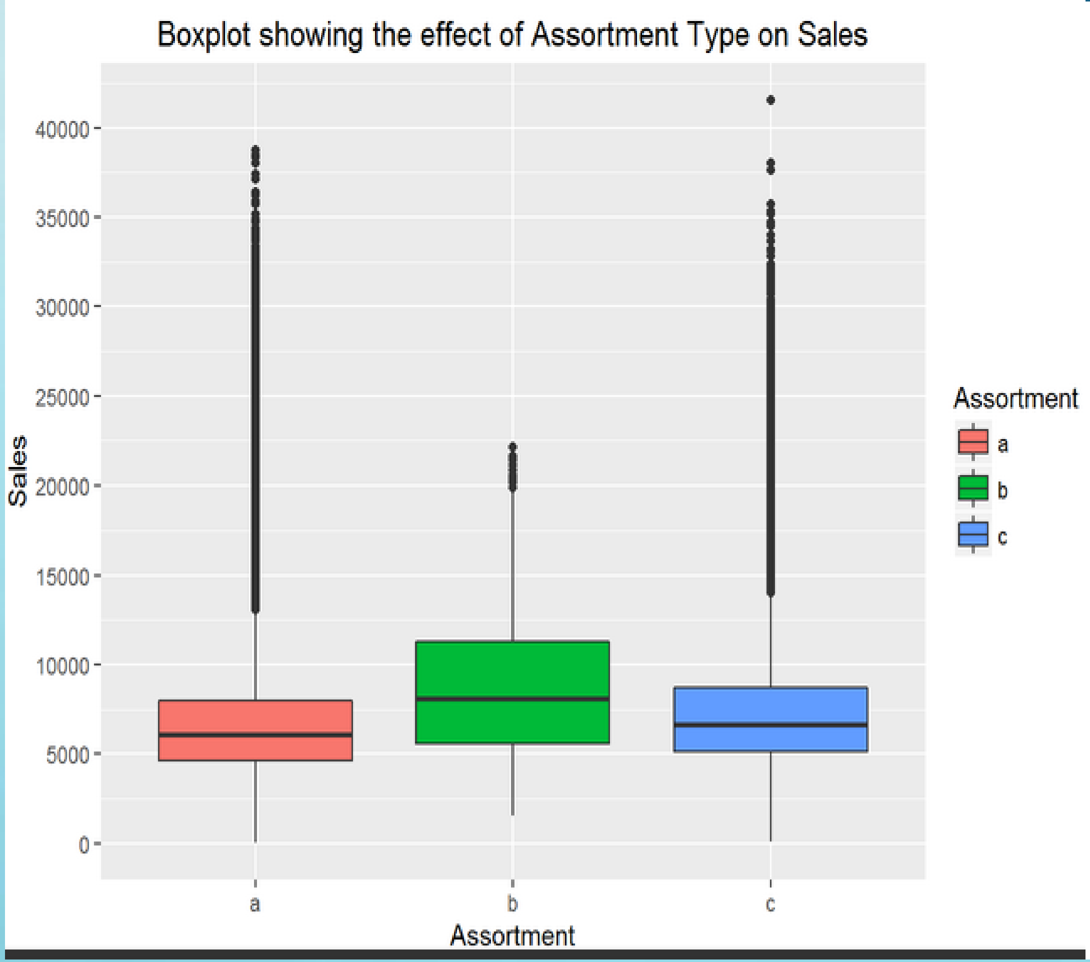


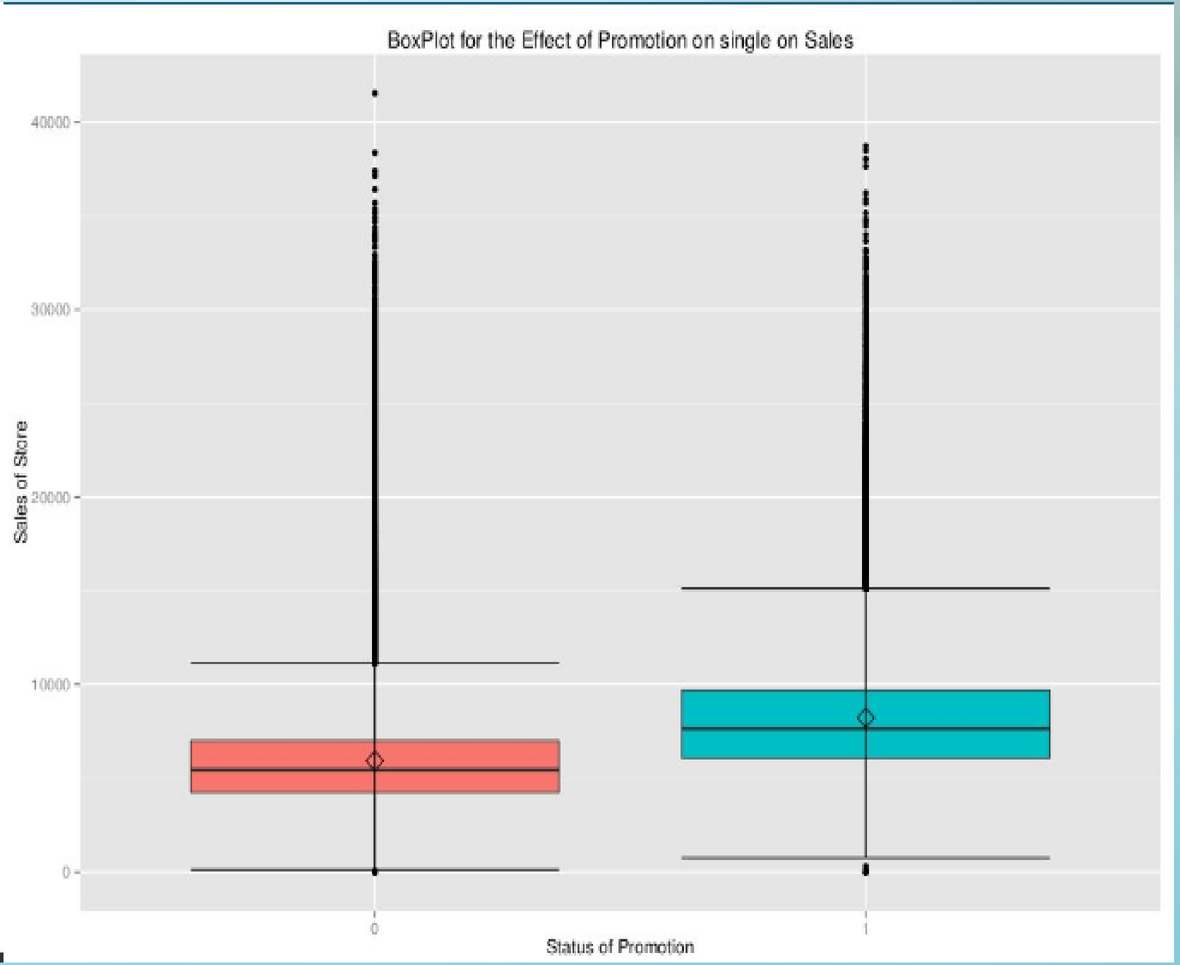
Distribution of avilable StoreTypes



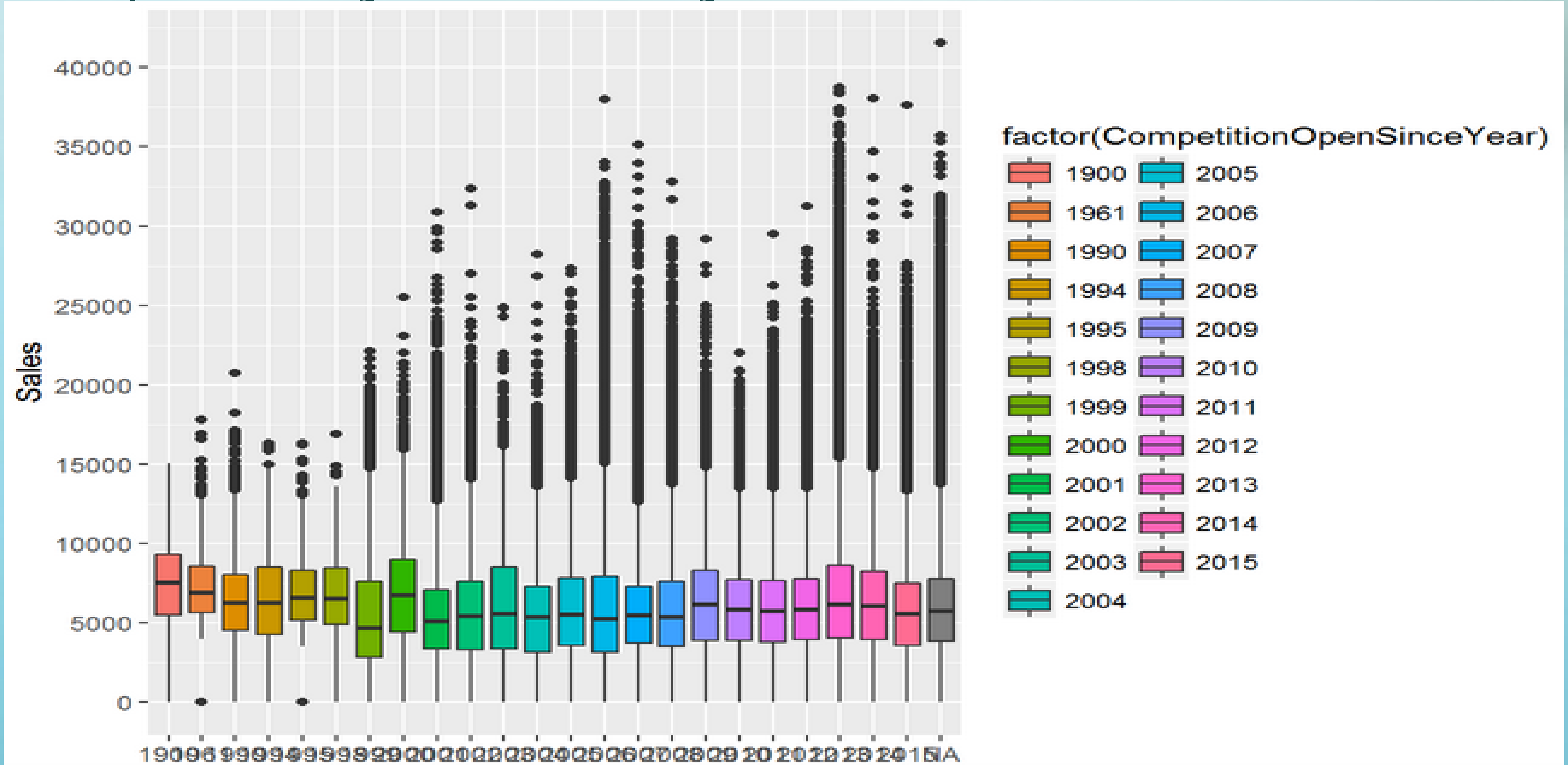
Distribution of available AssortmentTypes

Exploratory Data Analysis

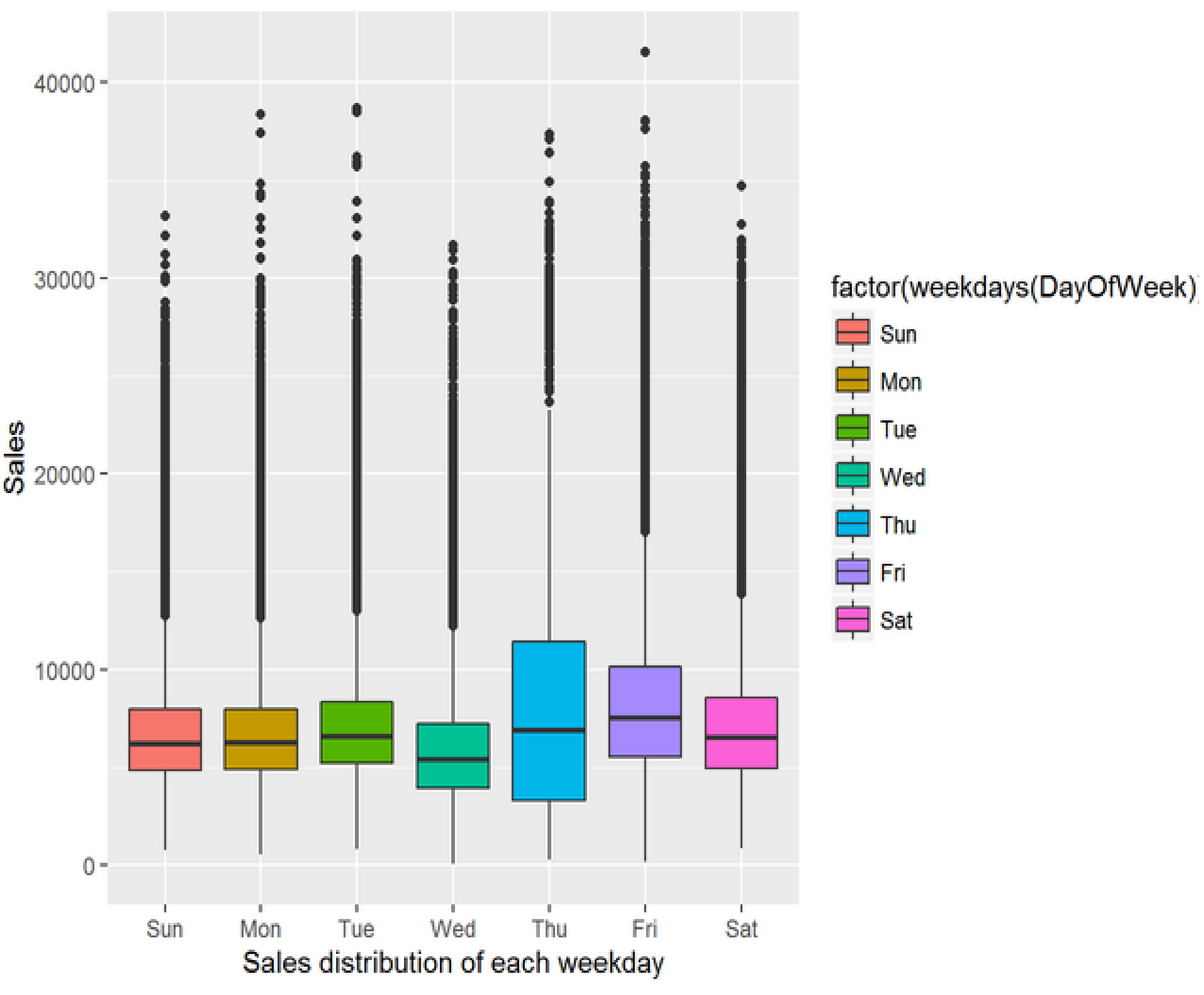


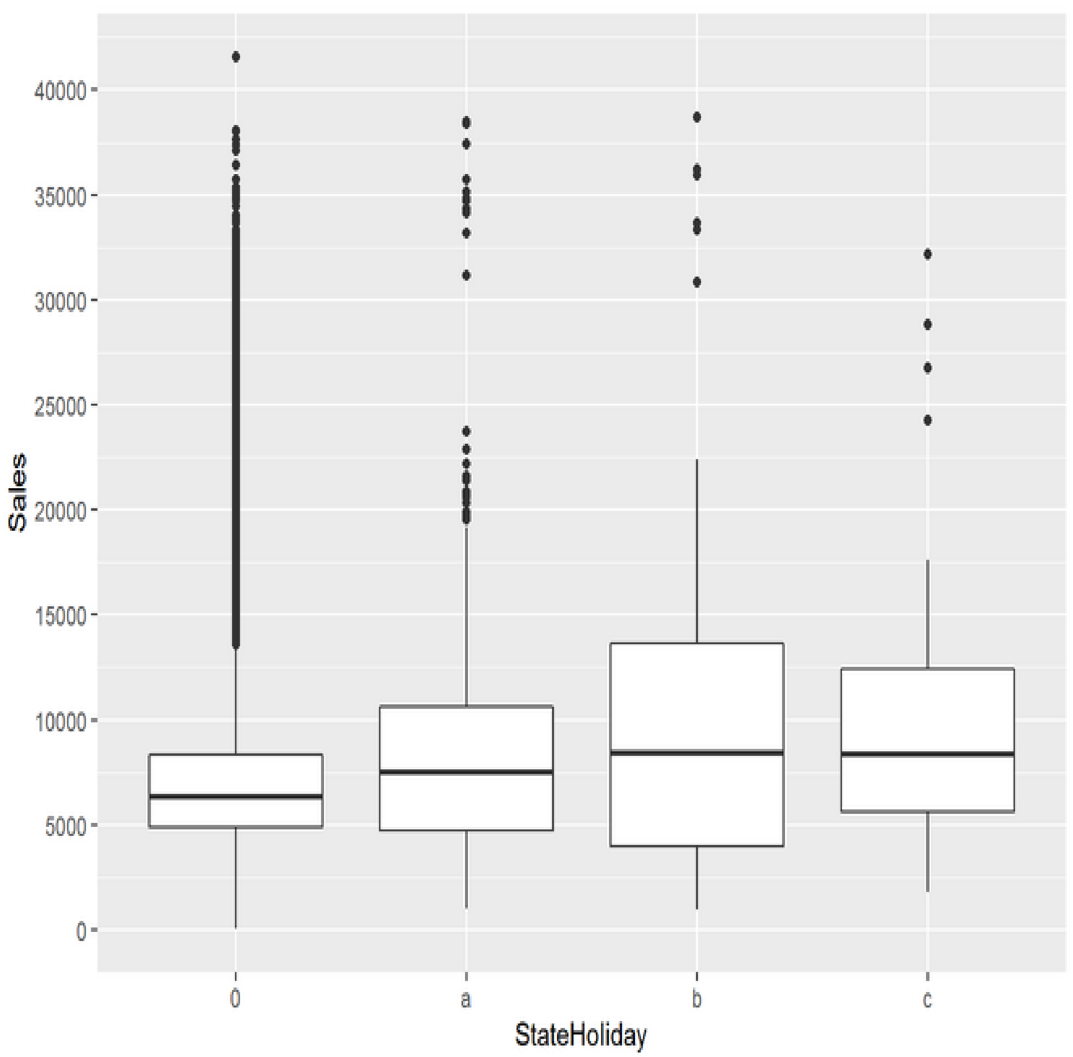


Exploratory Data Analysis

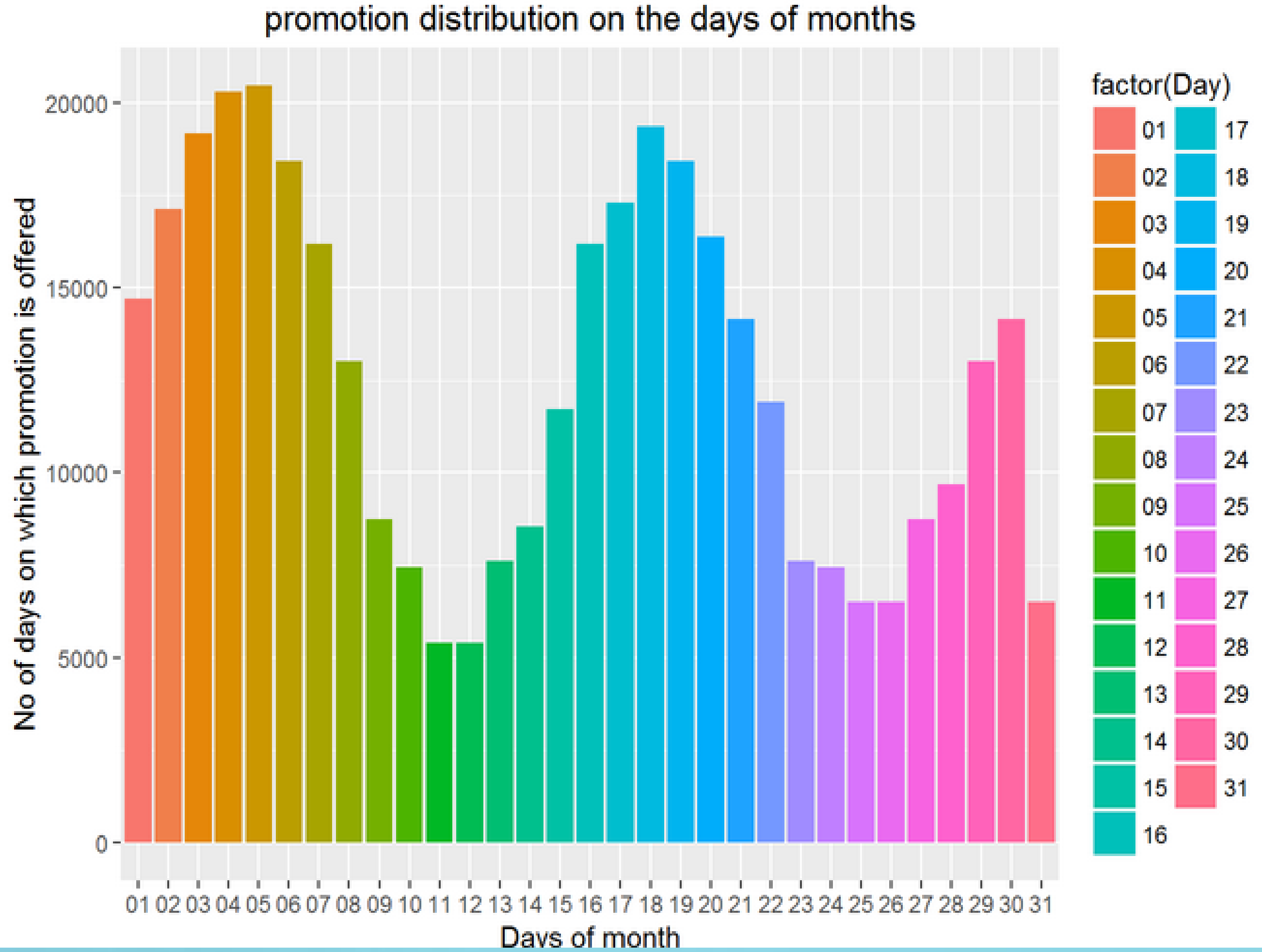


Exploratory Data Analysis

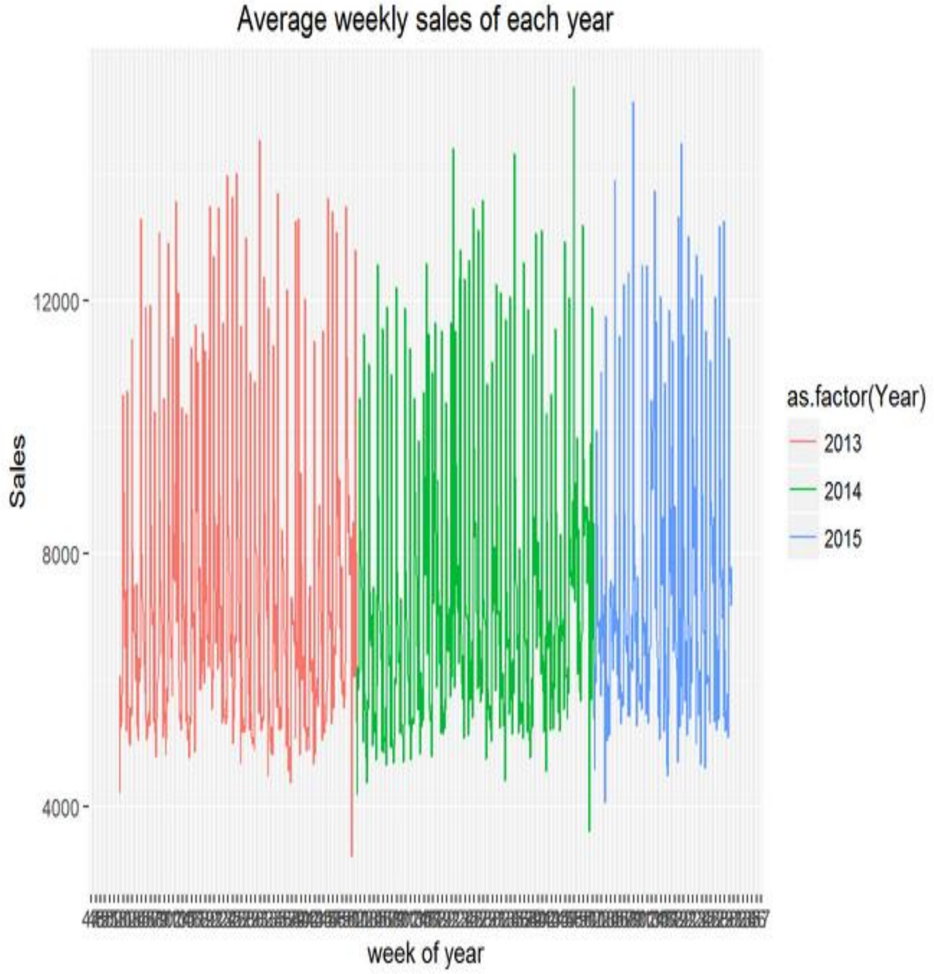




Exploratory Data Analysis



promotion distribution on the days of months



Average weekly sales of each year

Learning from EDA

- 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.

Learning from EDA

- 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.

Model Selection (Things to consider)

- Which model to use
- Should we apply same model to each store or separate model on different store. Using a different model for each store would be a cumbersome challenge.
- Given sales data might be time-series data.
- Only the data that has sales>0 is considered

Random Forest Model

- Rows with zeros in train data are removed and stores in test data are input as “1” i.e open.
- Date feature of train data and test date is split in to day, month, year, DayOfYear and week
- Train and store data is merged
- Features used in Random forest are (Store, DayOfWeek, Open, Promo, StateHoliday, SchoolHoliday, StoreType, Assortment, CompetitionDistance, Competitionopensinceyear, promo2, promo2sinceweek, promot2sinceyear, promotinterval, month, year, day, DayOfYear, Week)
- Sales is predicted

H20.GBM Model

- Rows with zeros in train data are removed and stores in test data are input as “1” i.e open.
- Date feature of train data and test date is split in to day, month, year, DayOfYear and week
- Train and store data is merged
- Log transformation of sales is used to avoid the sensitivity to high sales
- Features used in Random forest are (Store, DayOfWeek, Open, Promo, StateHoliday, SchoolHoliday, StoreType, Assortment, CompetitionDistance, Competitionopensinceyear, promo2, promo2sinceweek, promot2sinceyear, promotinterval, month, year, day, DayOfYear, Week)
- Sales is predicted

Learning and recommendations

- Although single model has been used to predict the sales. Stores can be categorised into some cluster and then use different models on different cluster to compare the results.
- A combination of different models might produce better results.
- Having the data of different states about the SchoolHoliday and StateHoliday will be helpful for better predictions.
- Weather data that has huge impact on drug purchase and will help to predict better sales.
- The variables about the sales just before and after the closing day and using them into model would be interesting to see.

Thank You