OPIM 5671: DATA MINING AND BUSINESS INTELLIGENCE

‘FORECASTING SALES USING STORE, PROMOTION, AND COMPETITOR DATA’

MAY 4, 2017

FINAL PROJECT REPORT

TEAM #8

SUBMITTED BY:
QIUIYI JIA
KUNJA DUTTA
SUSHANT GANDHI
RIKDEV BHATTACHARYA

UNIVERSITY OF CONNECTICUT, SCHOOL OF BUSINESS
SPRING’ 17

CONTENTS

I. EXECUTIVE SUMMARY:

Through this project report, we aim to build, compare, and evaluate models which best predicts the sales of 1115 Rossmann stores. The dataset, sourced from Kaggle.com is a combination of time series data along with a wide variety of features. Due to the presence of diverse market conditions along multiple store and assortment types we would have to build over 50,000 unique models to correctly predict individual store sales. To avoid this, we have decided to cluster the stores based on K-means clustering and thereby use ARIMA based Time Series models on those clusters to come up with forecasts. These forecasts were then distributed among the stores based on a rolling weighted average of the previous six months of store sales.

- To tackle the huge data volume, we binned the data based on monthly sales instead of daily sales.
- We had initially observed that the data had bi-weekly trend which was detrended based on our binning into monthly figures.
- We could not find any seasonality in the data, apart from seasonal peaks in December.
- We excluded 180 stores with more than six months of missing data so as to avoid any bias that might creep into the models.
- We transformed some key information regarding the competition so that it can be used as one of the key regressors in the model.
- We observed that time series forecasting models produced robust predictions for each of the 4 clusters.
- External Regressors such as whether the store was open and whether any promotion was running in the store was also helpful in predicting the store sales.
- We used a rolling weighted average method to allocate the forecasts among the member stores among each of the 4 clusters.

II. BACKGROUND AND BUSINESS OBJECTIVE

Rossmann operates more than 3,000 drug stores in over 33 European and Asian countries. On a daily-basis more than a million customers visit each of their 3000 odd drug stores in diverse market settings with different demands and consumer behavior. Sales in each of these stores are thus influenced by many factors, including promotions, competition, state and school holidays, seasonality, and locality. Robust and reliable sales forecasts enable store managers to create effective staff schedules, decrease operational wastages and losses and increase profitability by reducing the demand supply gap.

Our objective through this project is to forecast the sales of a select 1115 stores spread across different parts of Germany using a time series data and related variables. We aim to create models in SAS using time series forecasting techniques. The sales forecast would be for a monthly schedule and be made based upon the results from the best chosen ARIMA models for the 4 chosen store clusters. Accurate prediction of sales in these stores will help reduce any operational wastage and meet the demand ahead of any of its competitors.

III. DATASET DEFINITION

The dataset contains historical time series sales data for 1,115 stores of Rossmann ranging over a period from 1st January 2013 to 31st July 2015 with close to 1,017,209 observations. Below is the descriptions for the fields in dataset sourced from Kaggle-

#	Variable Name	Description
1	Id	an Id that represents a (Store, Date) tuple within the test set
2	Store	a unique Id for each store
3	Sales	the turnover for any given day (this is what you are predicting)
4	Customers	the number of customers on a given day
5	Open	an indicator for whether the store was open: 0 = closed, 1 = open
6	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
7	SchoolHoliday	indicates if the (Store, Date) was affected by the closure of public schools
8	StoreType	differentiates between 4 different store models: a, b, c, d

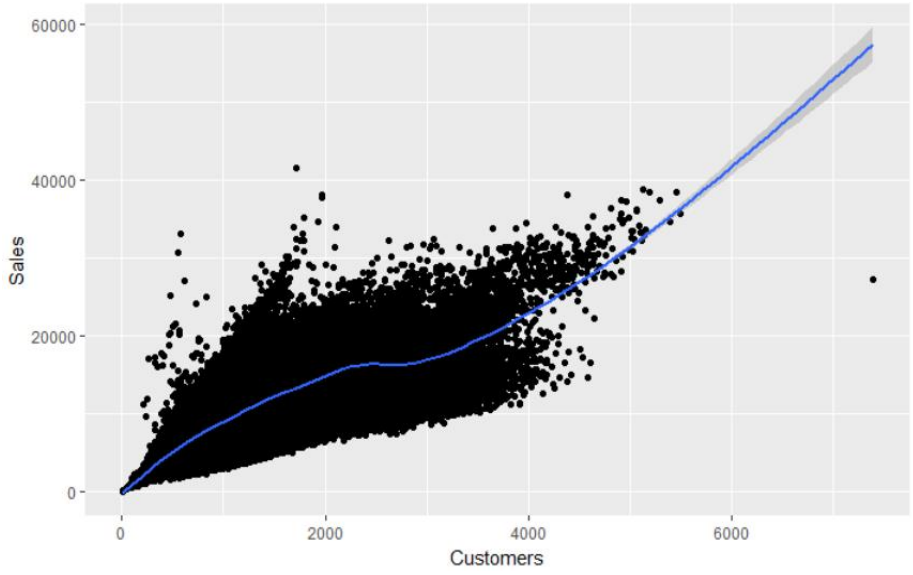
9	Assortment	describes an assortment level: a = basic, b = extra, c = extended
10	CompetitionDistance	distance in meters to the nearest competitor store
11	CompetitionOpenSince[Month/Year]	gives the approximate year and month of the time the nearest competitor was opened
12	Promo	indicates whether a store is running a promo on that day
13	Promo2	Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
14	Promo2Since[Year/Week]	describes the year and the calendar week when the store started participating in Promo2
15	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

IV. DATA EXPLORATION AND VISUALIZATION

The purpose of exploratory data analysis is to explore relationship among variables and to get business insights. Team used data.table and ggplot2 packages in RStudio to explore and visualize relationship or trends among variables.

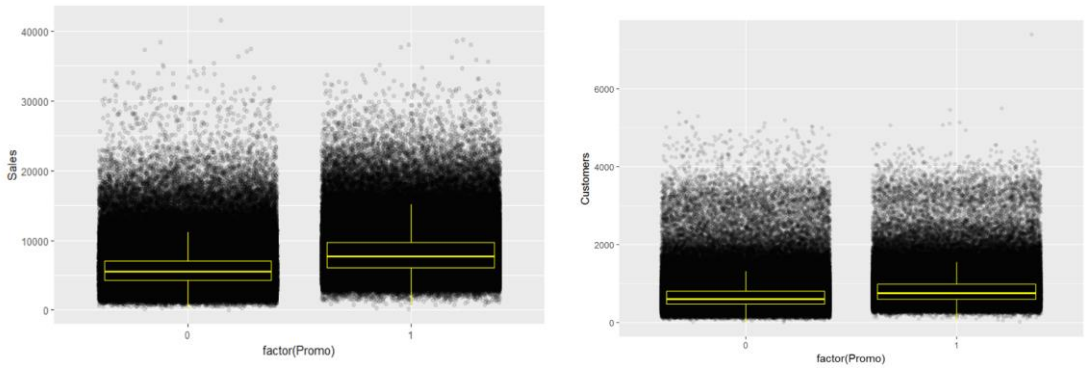
1. Relationship between Sales and Customers

Sales and Customers are two unique variables in train.csv. Sales is as expected strongly correlated with the number of customers. Sales increases as number of Customers increases with a diminishing trend. See Exhibit 1.



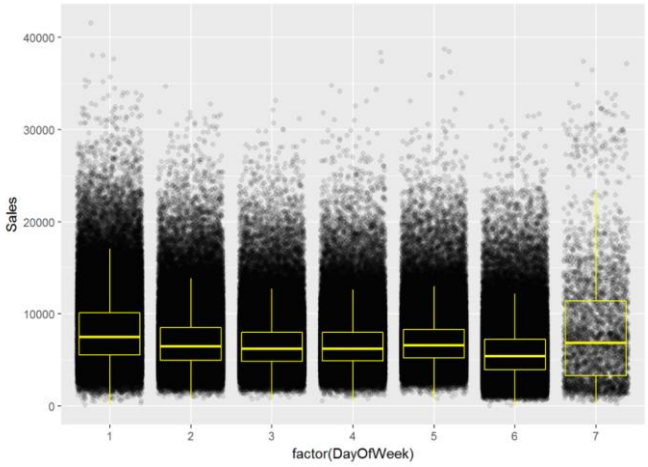
2. Effect of Promos on Sales and Customers

With promotion, amount of sales and number of customers both increase. But it looks like the boxplots of customers overlap a little more than the boxplots of sales, that means the promotion may not mainly attract more customers but make current customers spend more. See Exhibit 2.



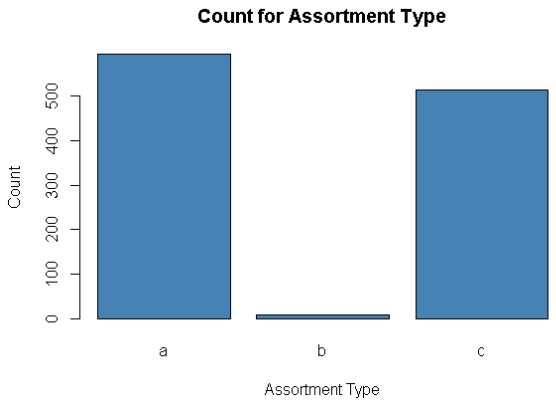
3. Relationship between Sales and DayofWeek

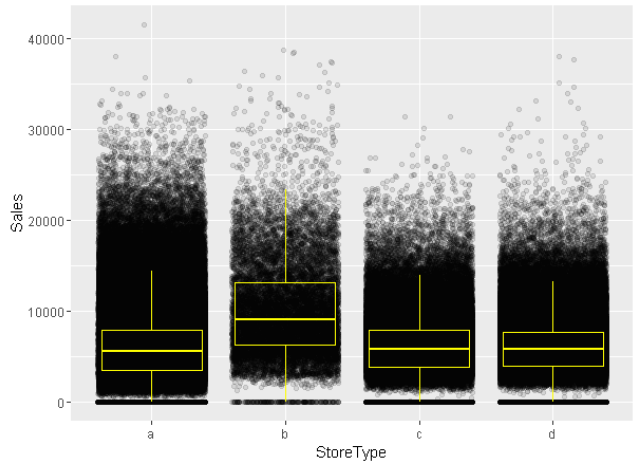
On weekdays, the factor DayofWeek do not vary much on sales. But the variability of sales on Sundays when the stores are open is quite high. See Exhibit 3.

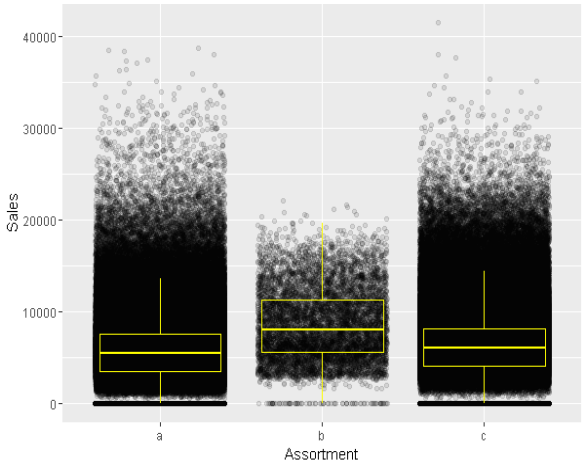


4. Store type and Assortment Type Glance

In store.csv, store type and assortment type are two important regressors that may affect prediction analysis. Based on bar plots, number of store type b and assortment type b are relatively low. Surprisingly, their median total sales are larger compared with other types. See Exhibit 4.

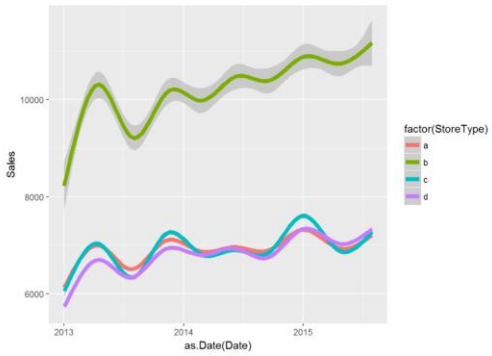


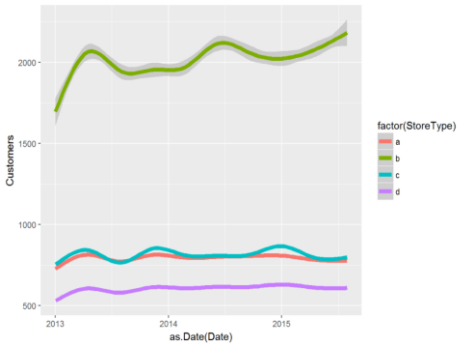


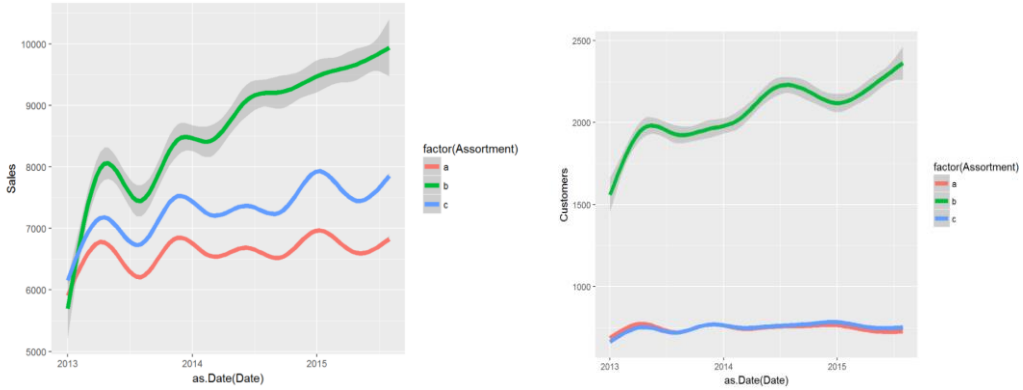


5. Trend of Sales and Customers based on Store & Assortment Type

Different store types and assortment types imply different overall levels of sales and number of Customers. Most types implies an increasing trend on Sales and Customer numbers. Moreover, one key finding is that store type b and assortment type b have relatively much steeper slope of trends. See Exhibit 5.

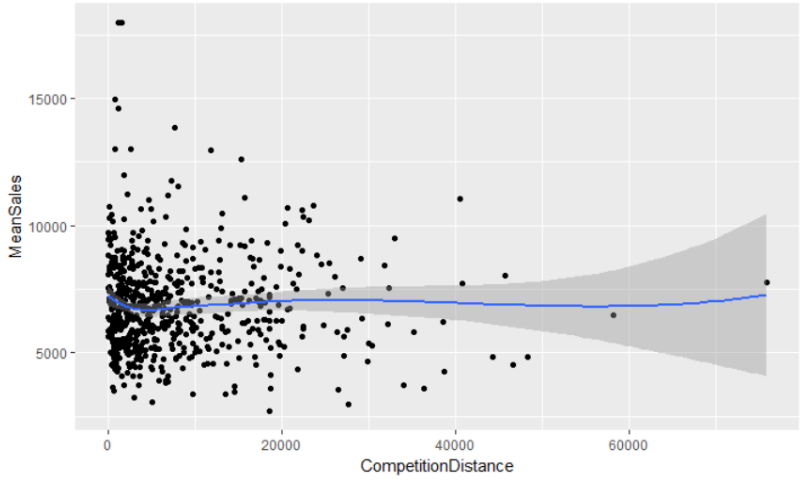






6. Effect of Competitor Distance

Pattern implied by the plot shows that lower distance to the next competitor implies slightly higher sales. This may indicate that competition motivates customers purchase behavior. Another assumption maybe stores with many competitors are located in inner cities with crowded population and thus generate high sales level. See Exhibit 6.



V. DATA MANIPULATION

The following steps were involved in the data manipulation process:

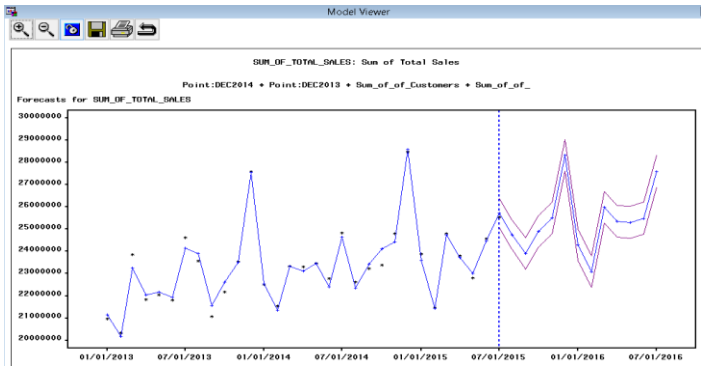
1. Merging Train.csv and store.csv datasets based on store id#
2. Pivot data to convert daily sales data to monthly sales data
3. Removing stores data with missing values, as the values are missing for 6 months' timeframe.

4. Combining CompetitorSinceMonth and CompetitorSinceYear columns into CompetitorMonths, which gives the number of months since competitor opened.
5. Loading data in R and normalizing variables.
6. K means clustering with 4 optimal clusters.
7. Once all stores have a cluster number, then again pivot data to give sales numbers for each cluster.
8. Time Series Modelling in SAS TSFS using ARIMA models.
9. Once we have the sales forecast for all the clusters, we give the weighted sales forecasts to each store in that cluster. We do this by first looking at the sales proportion of each store in that cluster for last 6 months. Then dividing the forecasted sales in, the same proportion to get a better store sales prediction.

VI. DATA MODELING AND FORECASTING

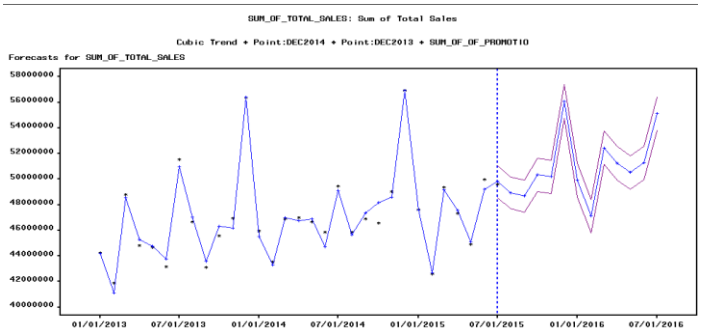
The 4 clusters are modeled using SAS Enterprise Guide. Discussions on each are as follows.

Cluster 1: There was no trend, no seasonality, events at Dec 2013 and Dec 2014, significant regressors were Sum of Customers and Sum of School Holidays. Based on these the model selected is AR(2) + Cubic Trend + Intervention : point (Dec 2013)+Intervention:point(Dec 2014)+ Regressor (Sum of Customers) +Regressor (Sum of Promotion Days). The forecast graph is below

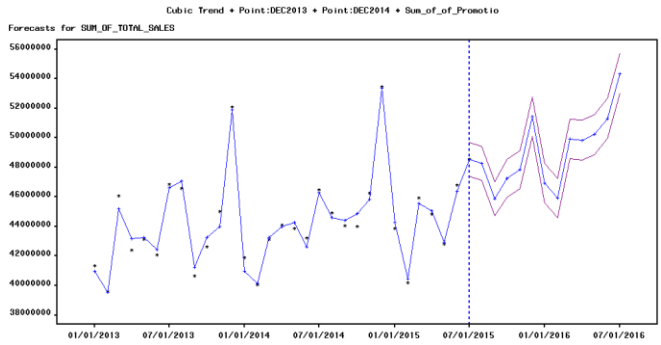


Cluster 2: There was no trend, no seasonality, events at Dec 2013 and Dec 2014, significant regressors were Sum of Customers and Promotion Days. Based on these the model selected is AR(2) + Cubic Trend +Intervention :

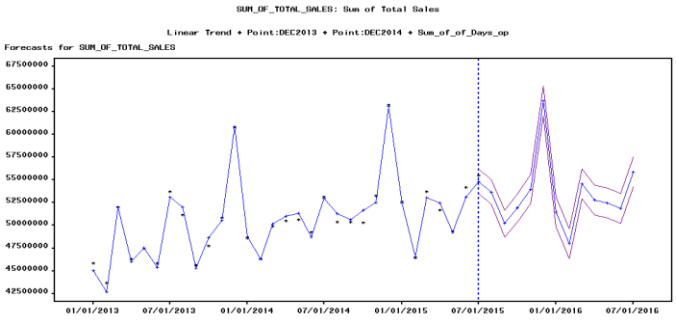
point (Dec 2013)+Intervention:point(Dec 2014)+ Regressor(Sum of Customers)+Regressor (Sum of Promotion Days). The forecast graph is below



Cluster 3: There was no trend, no seasonality, events at Dec 2013 and Dec 2014, significant regressors were Sum of Customers and Promotion Days. Based on these the model selected is AR(2) + Cubic Trend + Intervention : point (Dec 2013)+Intervention:point(Dec 2014)+Regressor (Sum of Customers) + Regressor (Sum of Promotion Day). The forecast graph is below



Cluster 4: There was no trend, no seasonality, events at Dec 2013 and Dec 2014, significant regressors were Sum of Customers and Sum of Days Open. Based on these the model selected is AR(2) + Linear Trend + Intervention : point (Dec 2013)+Intervention:point(Dec 2014)+Regressor (Sum of Customers) + Regressor (Sum of Days Open). The forecast graph is below



The following observations were made based on the forecast for each cluster.

Cluster No	No of Stores	Average Monthly Sales	Model RMSE	Accuracy Rank	Variations in Forecast Boundary
1	137	$23,368,651	283528	1	2.65%
2	295	$27,081,813	526390	3	2.66%
3	210	$44,411,689	478732	2	1.96%
4	297	$50,647,950	603921	4	3.12%

VII. CONCLUSION, KEY LEARNINGS AND RECCOMENDATION

Based on the vital observations made for the data set, the clusters, and its modeled forecasts, the following recommendations can be made.

1. Events around the month of December are uniform across all clusters. This indicates that the stores should have a good stock of items before December to avoid any demand-supply gap.
2. Promotion days influence the sales significantly. Running innovative offers from time to time will be particularly beneficial for Rossmann.
3. Cluster 3 generates good sales; its modeling provides less error and the boundary variations for its forecast is less. Setting up more stores with similar characteristics as that of cluster 3 will be profitable.
4. Store type b and Assortment type b have higher sales. Rossmann should focus on these store and assortment types. Increasing marketing efforts on these types of stores will maximize sales.

VIII. APPENDIX

REFERENCES

https://www.kaggle.com/c/rossmann-store-sales/data&Sons, Ltd.