Predicting Sales in Rossmann Stores

Blue Team

Kim Tricker *(Group Lead)*

Priyanka Sharma

Prabhakar Teja Seeda

Joshua Nielsen

ITCS 6162 Knowledge Discovery in Databases

University of North Carolina Charlotte

10 April 2018

Abstract

Rossmann, a drug store chain located in Europe, can turn a manually-executed weekly task into a data mining solution that will both improve accuracy and boost efficiency in the company. Store managers must predict store sales by gut-driven and handwritten calculations. The goal of our project is to revolutionize this process with data mining algorithms. We began with an exploratory data analysis, where we discovered that Type B stores represent the most sales while they have the fewest quantity of stores. We also found that holidays which are also school holidays average more sales than holidays that are not. We then performed linear regression, random forest, and CART models on the data. We found the best root mean squared error with the random forest model and had strong statistical evidence that the variances in the data could be explained by our models. Our models also confirmed what we found in the exploratory data analysis: holidays do not have a significant impact on sales.

Predicting Sales

As Germany's second largest drugstore chain, Rossmann serves thousands of customers in locations throughout Europe. With 28,000 employees, 3,000 stores, and 17,500 unique items stocked, the potential for data analytics to drive Rossmann's strategy cannot be understated. The first step towards integrating analytics into the company's business strategy is using it to solve pain points. Our objective will be to help Rossmann use algorithms and data mining methods to predict daily sales in a six-week window, a process required from store managers that is currently manual and gut-driven.

Predictive models like linear regressions, time series models or random forest algorithms use data to forecast sales while eliminating the inherent bias, subjectivity, and inconsistencies that occur when this process is done by managers with varying levels of experience. We will arrive at the appropriate predictive model by executing our project through the CRISP-DM process, ensuring that we understand both the business domain as well as the dataset and data mining procedures. Through this project, we expect to enhance our knowledge of predictive analytics by immersing ourselves in a real-world problem whose solution lies in data science techniques.

Dataset

Description and Purpose

This dataset contains historical sales data for 1,115 Rossmann stores. The dataset is partitioned into train and test sets. The purpose is to predict future sales over a six-week period using this historical data.

Source

Pradhan, Anshuman. (2015). Rossmann Store Sales. Location: Kaggle.

Related Work

Beam, D. & Schram, M. (2015). Rossmann Store Sales. SemanticsScholar. https://pdfs.semanticscholar.org/dec6/147288206499c0eec4778f7e0c704442d3ec.pdf.

Pavlyshenko, B.M. (2016). Linear, machine learning and probabilistic approaches for time series analysis. 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). https://ieeexplore.ieee.org/document/7583582/#full-text-section.

Business Research/Understanding

Project Objectives

Problem Domain: The value of predicting store sales is manifold: it can be used for scheduling staff, setting up promotions, or anticipating customer traffic. However, the manual process of making these predictions does not allow for the company to benefit from potentially valuable insights. Besides the inconsistencies involved in each individual manager making these predictions in the absence of a formal process, perhaps the biggest problem is accounting for the fluctuations in sales that come with promotions, competition, school and state holidays, seasonality, and locality.

Requirements: This project requires the Kaggle dataset provided by user Anshuman Pradhan. Additionally, Microsoft Excel and R are required tools for analysis.

Restrictions: Our project is restrained by time, resources, and available data. The limited project timeline restricts our time data exploration phase and prevents a thorough examination of adequate data mining algorithms. We also lack the advanced computing power that would allow us to attempt more sophisticated algorithms. Limited data availability is also a restriction to our analysis. More information such as competitor's promotion periods, weather patterns or data on political happenings would be beneficial to predicting sales in a given time period.

Data Mining Problem Definition: Our data mining problem will be to predict sales for the Rossmann stores through predictive modeling techniques.

Strategy: In this project, we will help Rossmann benefit from the value of their data by using predictive models to forecast company sales in a six-week period. We will begin by exploring the data using exploratory data practices to find hidden patterns or trends in the data. We will then prepare the data by transforming it from raw, messy data to clean data ready for the algorithms. We will then put the cleaned data through a linear regression model and a random forest algorithm and compare the results from each. After evaluating the results, we will choose the model that does the best job of making sales predictions and test the model with the partitioned test dataset.

Data Understanding

Exploratory Data Analysis

Description of the Data.

Document all features and attributes of all datasets.

- Store - a unique Id for each store
- Sales - the turnover for any given day (this is what you are predicting)
- Customers - the number of customers on a given day
- Open - an indicator for whether the store was open: 0 = closed, 1 = open
- 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
- 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
- CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
- Promo - indicates whether a store is running a promo on that day
- Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating

- Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2
- 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

Estimation, Data subset, Data Quality

Show results of EDA to include summaries, frequency distributions, box plots, regression analysis, correlation, etc. Identify patterns and trends that you find in the data.

From the given datasets, the data from store and train are merged. Also the test and store datasets are merged which are required in order to get the correlation among all the variables in the dataset to predict the necessary outcome.

Summary of train data

```
> summary(train)
     Store        StoreType  Assortment CompetitionDistance CompetitionOpenSinceMonth CompetitionOpenSinceYear     Promo2
 Min.   :   1.0   a:551627   a:537445   Min.   :   20       Min.   : 1.0              Min.   :1900             Min.   :0.0000
 1st Qu.: 280.0   b: 15830   b:  8294   1st Qu.:  710       1st Qu.: 4.0              1st Qu.:2006             1st Qu.:0.0000
 Median : 558.0   c:136840   c:471470   Median : 2330       Median : 8.0              Median :2010             Median :1.0000
 Mean   : 558.4   d:312912              Mean   : 5430       Mean   : 7.2              Mean   :2009             Mean   :0.5006
 3rd Qu.: 838.0                         3rd Qu.: 6890       3rd Qu.:10.0              3rd Qu.:2013             3rd Qu.:1.0000
 Max.   :1115.0                         Max.   :75860       Max.   :12.0              Max.   :2015             Max.   :1.0000
                                        NA's   :2642        NA's   :323348            NA's   :323348
 Promo2SinceWeek  Promo2SinceYear          PromoInterval      DayOfWeek          Date             Sales          Customers
 Min.   : 1.0     Min.   :2009                    :508031   Min.   :1.000   2013-01-02:   1115   Min.   :    0   Min.   :   0.0
 1st Qu.:13.0     1st Qu.:2011     Feb,May,Aug,Nov :118596   1st Qu.:2.000   2013-01-03:   1115   1st Qu.: 3727   1st Qu.: 405.0
 Median :22.0     Median :2012     Jan,Apr,Jul,Oct :293122   Median :4.000   2013-01-04:   1115   Median : 5744   Median : 609.0
 Mean   :23.3     Mean   :2012     Mar,Jun,Sept,Dec: 97460   Mean   :3.998   2013-01-05:   1115   Mean   : 5774   Mean   : 633.1
 3rd Qu.:37.0     3rd Qu.:2013                               3rd Qu.:6.000   2013-01-06:   1115   3rd Qu.: 7856   3rd Qu.: 837.0
 Max.   :50.0     Max.   :2015                               Max.   :7.000   2013-01-07:   1115   Max.   :41551   Max.   :7388.0
 NA's   :508031   NA's   :508031                                             (Other)   :1010519
      Open            Promo        StateHoliday SchoolHoliday    CompetitionOpenSince
 Min.   :0.0000   Min.   :0.0000   0:986159     Min.   :0.0000   Min.   :1900
 1st Qu.:1.0000   1st Qu.:0.0000   a: 20260     1st Qu.:0.0000   1st Qu.:2006
 Median :1.0000   Median :0.0000   b:  6690     Median :0.0000   Median :2010
 Mean   :0.8301   Mean   :0.3815   c:  4100     Mean   :0.1786   Mean   :2009
 3rd Qu.:1.0000   3rd Qu.:1.0000                3rd Qu.:0.0000   3rd Qu.:2013
 Max.   :1.0000   Max.   :1.0000                Max.   :1.0000   Max.   :2016
                                                                 NA's   :323348
>
```

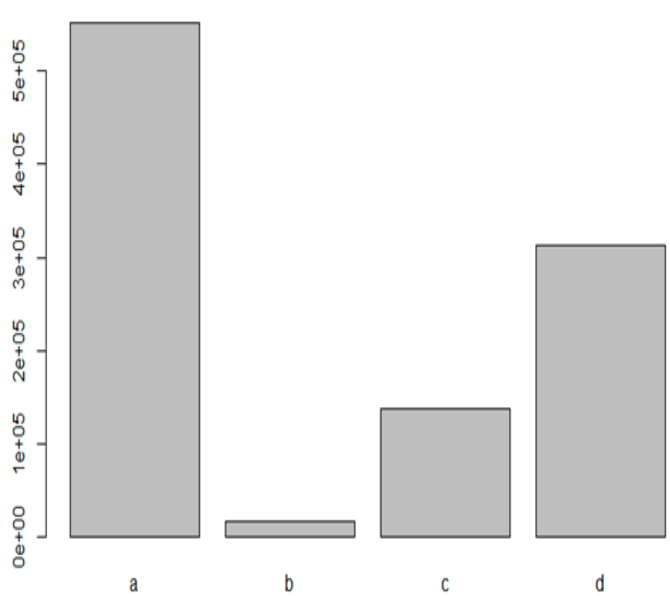
Summary of test data

```
                                                              NA's   :323348
> summary(test)
     Store        StoreType Assortment CompetitionDistance CompetitionOpenSinceMonth CompetitionOpenSinceYear     Promo2
 Min.   :   1.0   a:22128   a:20304    Min.   :   20       Min.   : 1.000            Min.   :1900             Min.   :0.0000
 1st Qu.: 279.8   b:  576   b:  432    1st Qu.:  720       1st Qu.: 4.000            1st Qu.:2006             1st Qu.:0.0000
 Median : 553.5   c: 4272   c:20352    Median : 2425       Median : 7.000            Median :2010             Median :1.0000
 Mean   : 555.9   d:14112              Mean   : 5089       Mean   : 7.035            Mean   :2009             Mean   :0.5806
 3rd Qu.: 832.2                        3rd Qu.: 6480       3rd Qu.: 9.000            3rd Qu.:2012             3rd Qu.:1.0000
 Max.   :1115.0                        Max.   :75860       Max.   :12.000            Max.   :2015             Max.   :1.0000
                                       NA's   :96          NA's   :15216             NA's   :15216
 Promo2SinceWeek Promo2SinceYear            PromoInterval         Id          DayOfWeek           Date            Open
 Min.   : 1.00   Min.   :2009                      :17232   Min.   :    1   Min.   :1.000   2015-08-01:  856   Min.   :0.0000
 1st Qu.:13.00   1st Qu.:2011   Feb,May,Aug,Nov  : 5712     1st Qu.:10273   1st Qu.:2.000   2015-08-02:  856   1st Qu.:1.0000
 Median :22.00   Median :2012   Jan,Apr,Jul,Oct  :13776     Median :20545   Median :4.000   2015-08-03:  856   Median :1.0000
 Mean   :24.43   Mean   :2012   Mar,Jun,Sept,Dec: 4368      Mean   :20545   Mean   :3.979   2015-08-04:  856   Mean   :0.8543
 3rd Qu.:37.00   3rd Qu.:2013                               3rd Qu.:30816   3rd Qu.:6.000   2015-08-05:  856   3rd Qu.:1.0000
 Max.   :49.00   Max.   :2015                               Max.   :41088   Max.   :7.000   2015-08-06:  856   Max.   :1.0000
 NA's   :17232   NA's   :17232                                                              (Other)   :35952   NA's   :11
     Promo        StateHoliday SchoolHoliday
 Min.   :0.0000   0:40908      Min.   :0.0000
 1st Qu.:0.0000   a:  180      1st Qu.:0.0000
 Median :0.0000                Median :0.0000
 Mean   :0.3958                Mean   :0.4435
 3rd Qu.:1.0000                3rd Qu.:1.0000
 Max.   :1.0000                Max.   :1.0000
```

EDA on Categorical Variables:

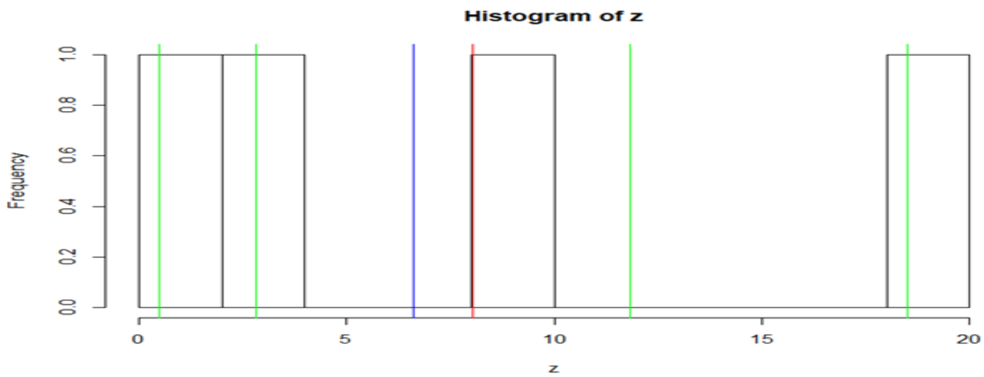
From the train and store merged set, the Data Analysis on categorical variables namely StoreType, Assortment SchoolHoliday and StateHoliday gives the following results.

StoreType:



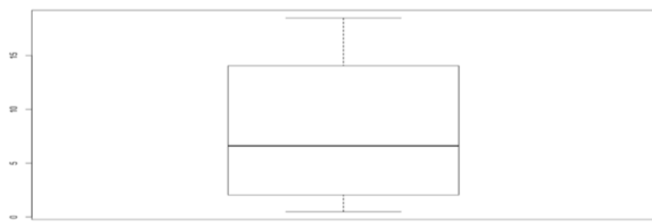
We plotted the data for store variable, we see store of type b are rare, so we adjusted the break using hist and smoothed the curve using density function. Since the mean and median are same, just for representation we took a square of plotted data.

We also conclude that most stores are of type a.



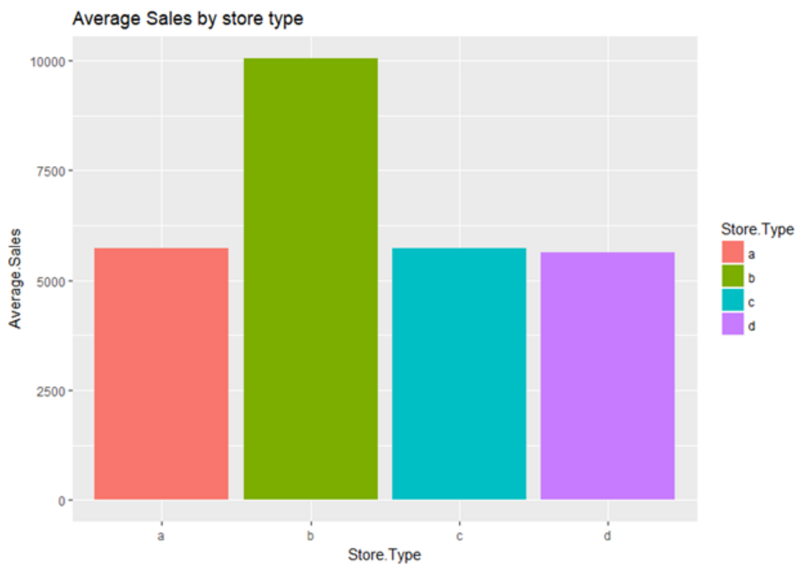
Green<-Shows quantiles,red<-represent mean,blue<-median

Next we did a box plot to check for outlier on the hist of plotted data



No outlier are detected.

Now we plotted the average sales for all store types.



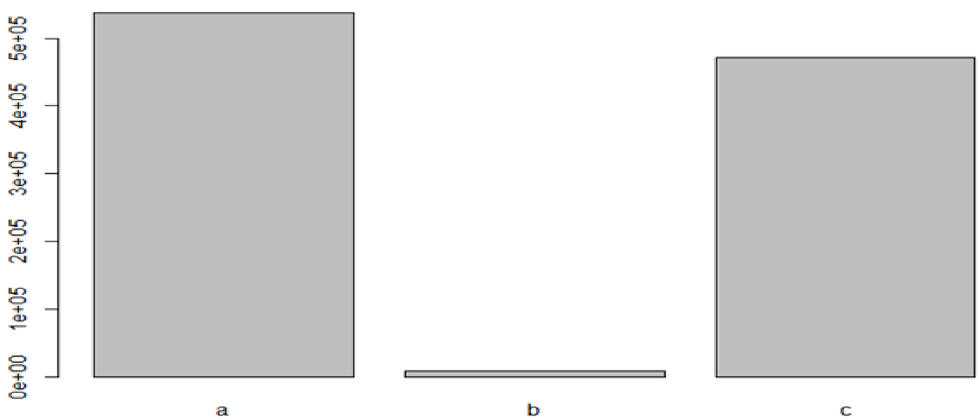
We see that although there are the fewest number of Type B stores, we still have maximum sales at stores of Type B.

Hence it was safe to not remove it.

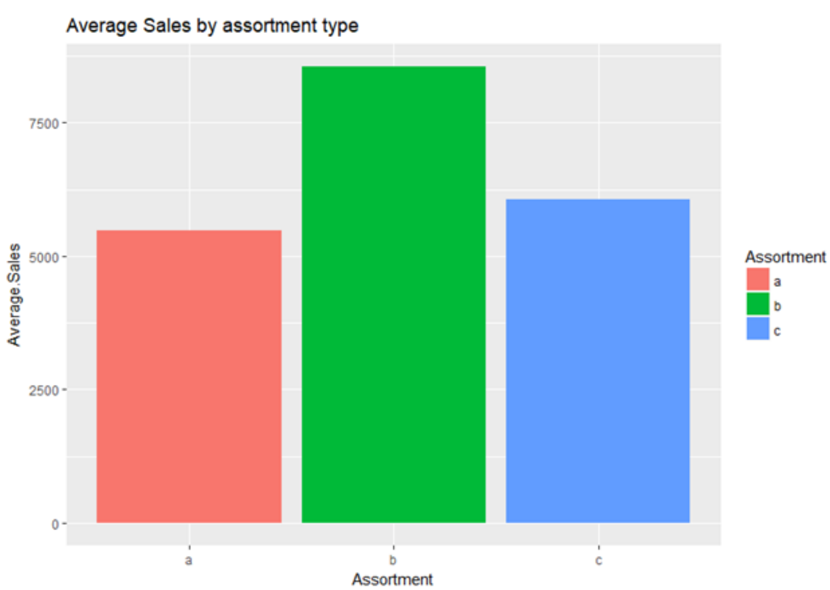
Assortment

We plotted the data for assortment variable, we see store of type b are rare, so we adjusted the break using hist and smoothed the curve using density function. Since the mean and median are same, just for representation we took a square of plotted data.

No outlier are detected.

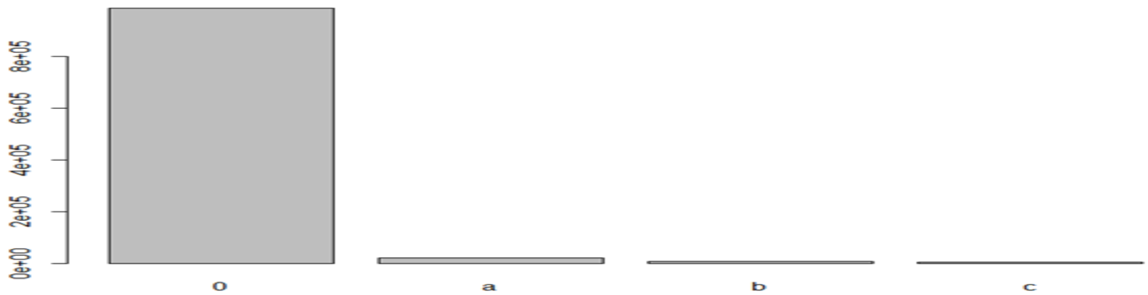


Now we plotted average sales for assortment type.



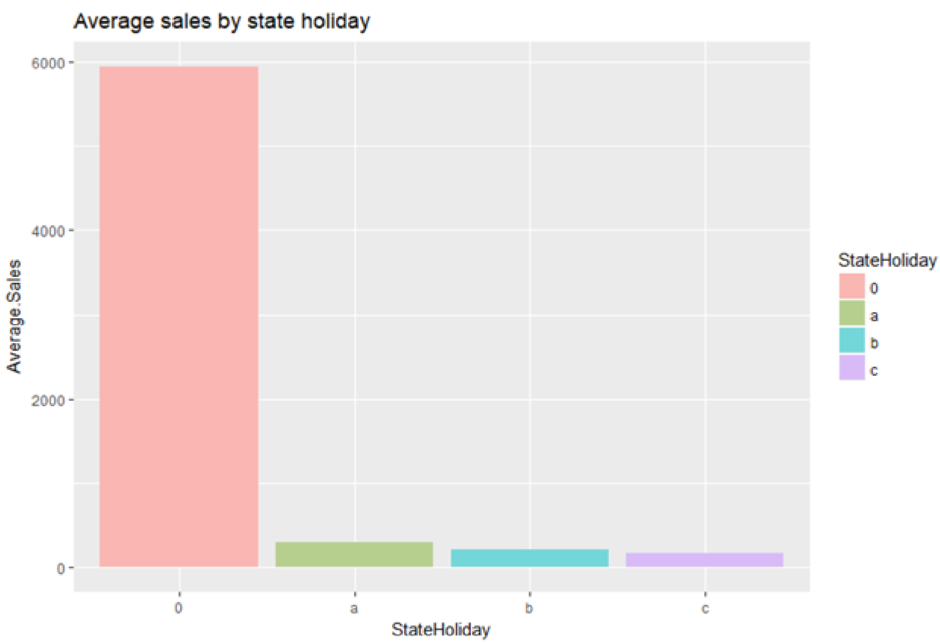
We see that even though stores with assortment type b are less, these are store with relatively more sales as compared to other assortment types.

StateHoliday:



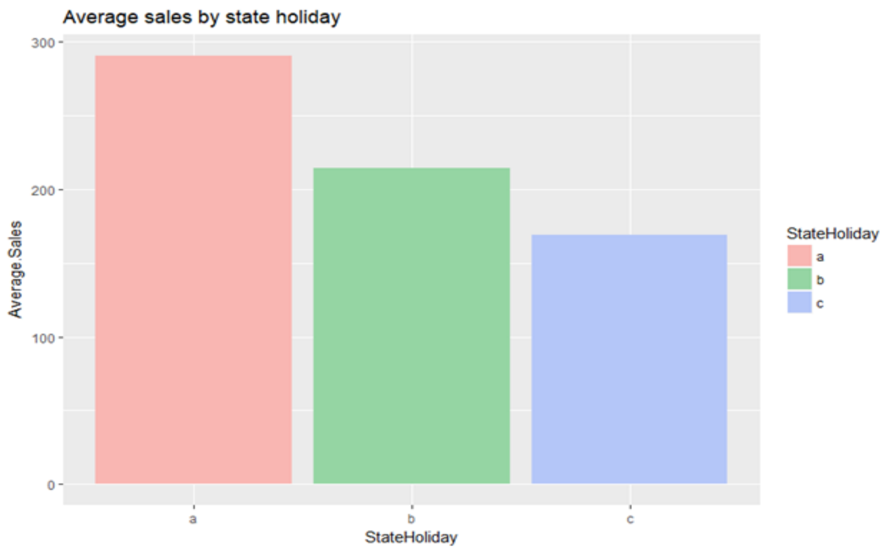
We plotted the data for state holiday variable. We see limited data for holidays.

We plotted the average of sales.



We conclude that even if it's a working day, sales are very high compared to state holidays.

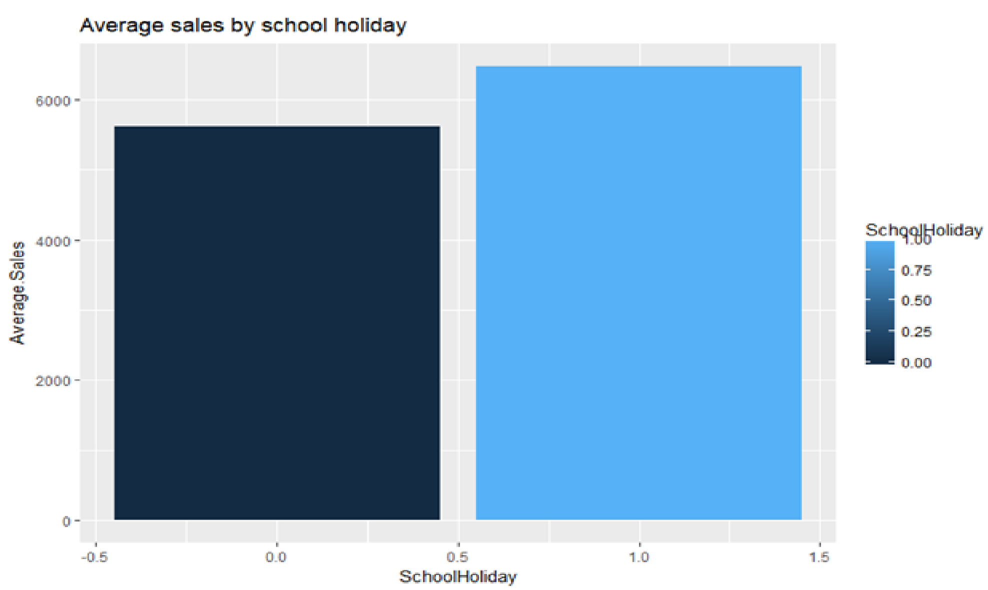
However, to see impact on sales due to holidays, we created a subset for which we removed values with no holiday (i.e., 0) which comprised most of the data. Presented below:



We see that sales are less relatively for holidays b and c.

School Holiday:

The data consists of value 0/1 stating if there is a school holiday or not. We checked the impact of school holiday on sales.



There is not a significant impact on sales due to school holiday.

Multivariate analysis

We plan to bin data for sales as low, high, average sale and do multivariate analysis of all categorical variable and see regression.

EDA on Continuous Variables:

On checking the sales of the merged set when they are closed, we can conclude that there are no sales when the store is closed. So, the prediction for closed stores is trivial. So, all the values from the dataset when the dataset can be taken into a subset to see impact of sales for store. We however didn't delete it to see impact on other competition store because of a closed store.

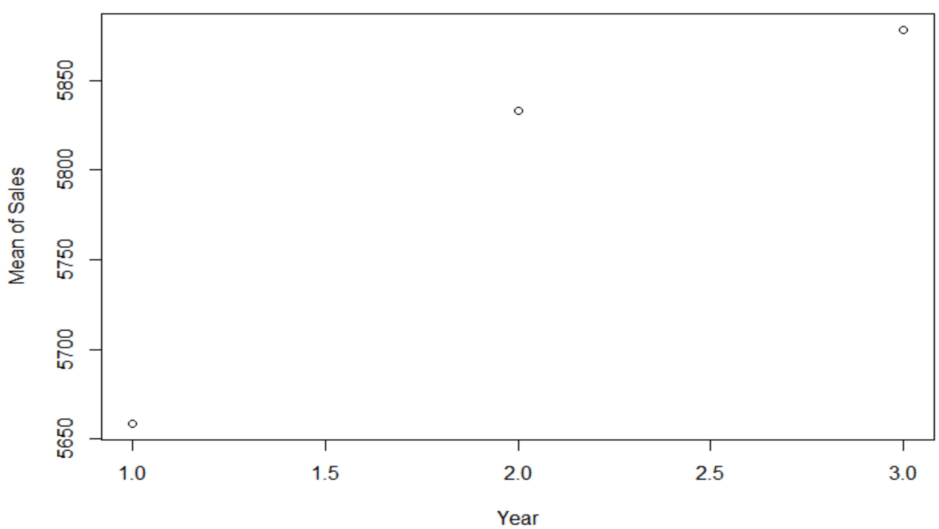
Date manipulation:

To read the date correctly, the date can be split into YearDate, MonthDate, DayDate,

In the test data, some of the stores are missing for which we are supposed to predict the sales for the dates. The number of unique Stores in the test dataset are 856. So, the stores which are not available in the test dataset can be removed from the merged train and store dataset.

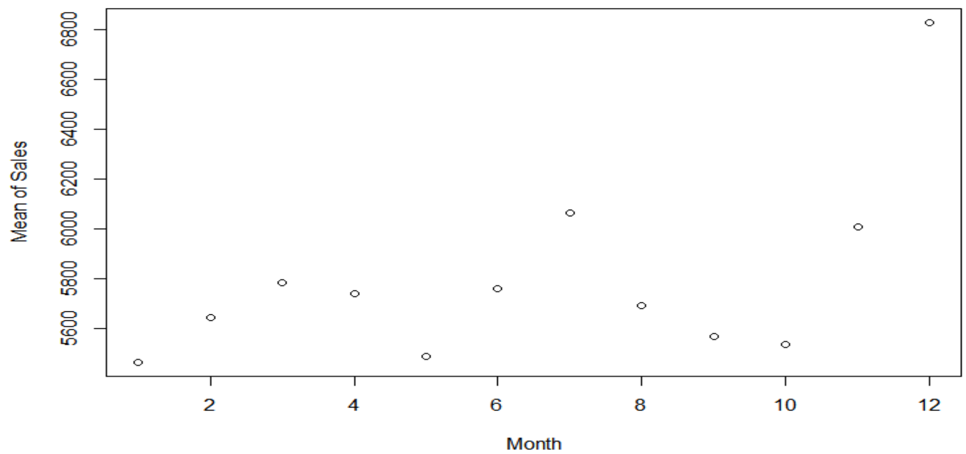
Approximately around 25 percent of the data can be removed from the dataset for the prediction because of this observation. This helps in faster prediction rate than before.

Yearly Sales:



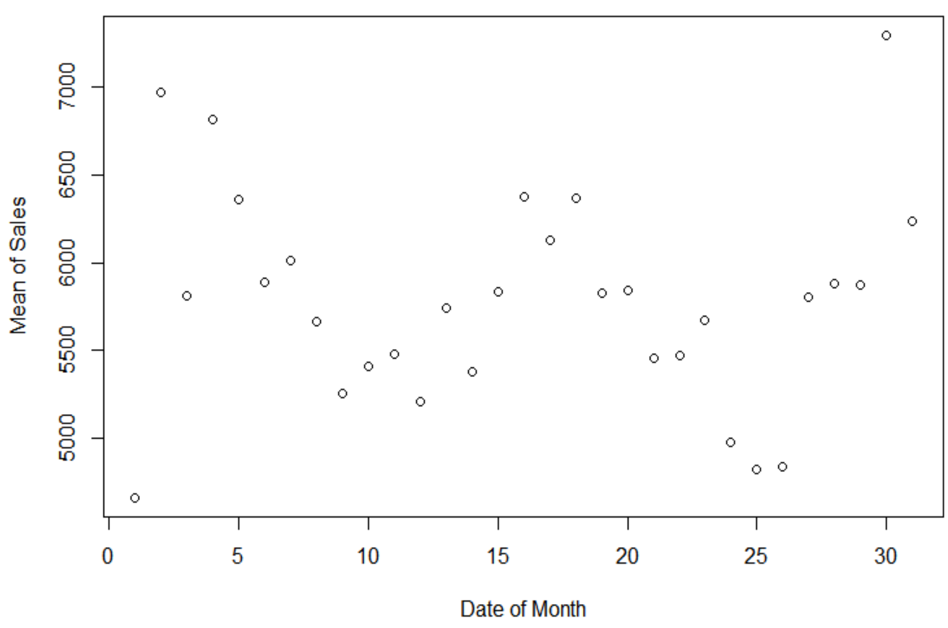
We observe the average sales increase over the year 13,14,15.

Monthly Sales:



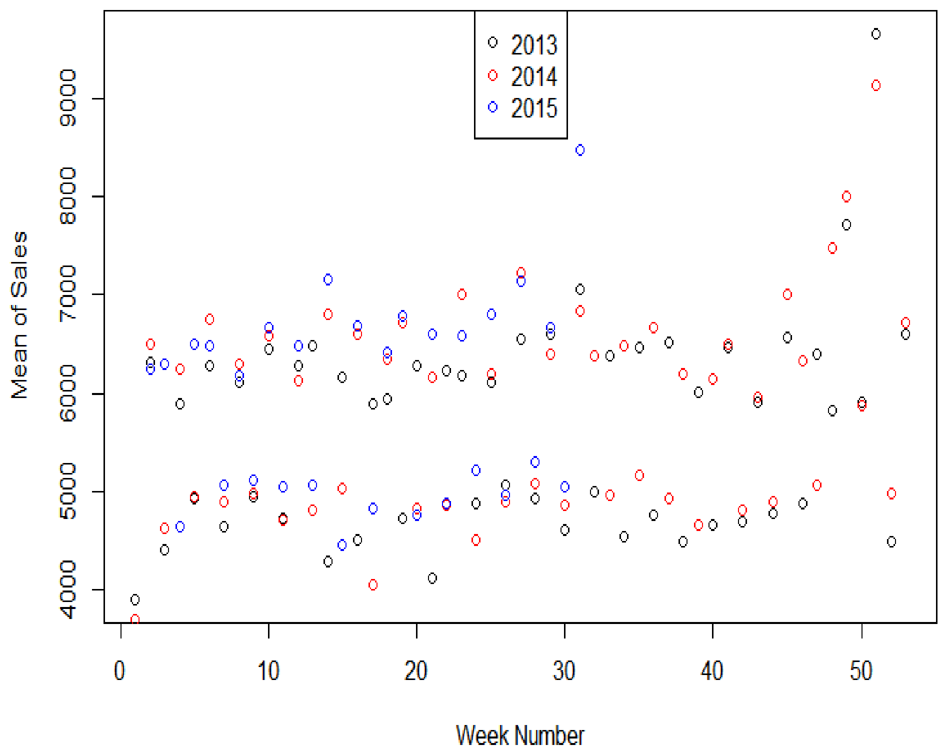
We observe increased sales in holiday season November and December.

Distribution of Sales over a month:



We observe that sales are high in beginning or end of month as most people get salary.

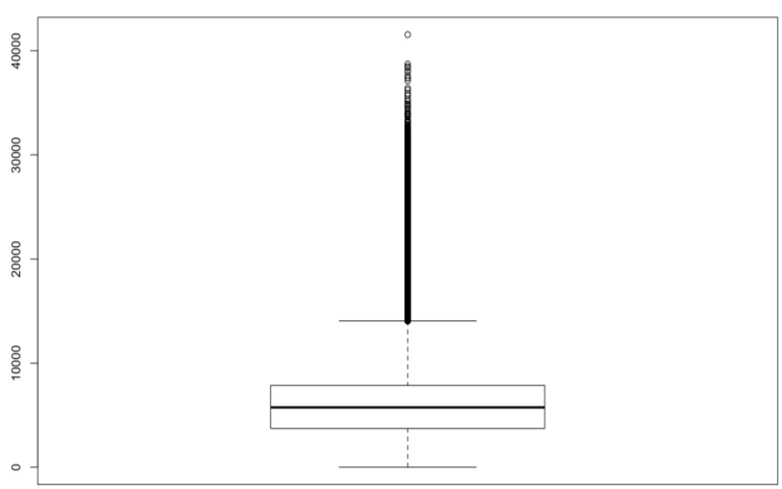
Weekly sales over all the years:



The week of Christmas most probably has higher mean of Sales, we see lesser data of 2015 as there is no data for 5 months.

Sales:

For the Sales column in the merged dataset, after plotting the box plot, it can be observed that sales around or greater than 18000 are listed as Outliers.



By Checking the summary of the data for sales greater than 18000, some of the stores have high Sales values compared to others. It can imply that the data is correct rather than populated with the outliers.

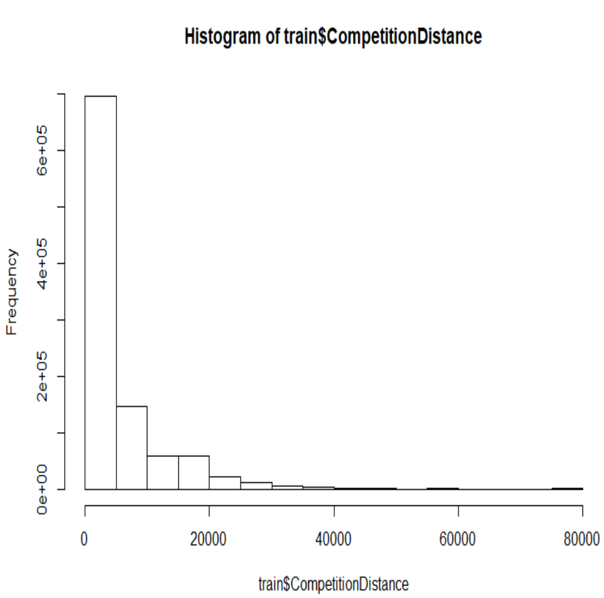
Similarly, on checking the data available for each store gives a length of 942 for store 1 to 6 and the minimum is 758. There are little to none outliers in the dataset and we have enough data in the dataset which would help us in prediction part.

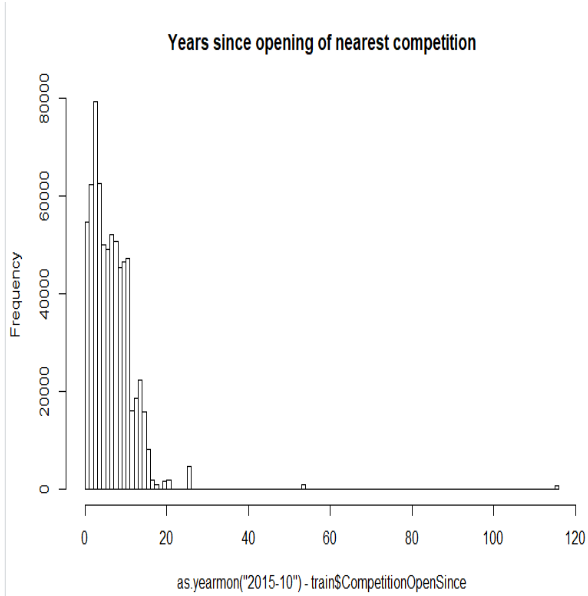
Relation between Promo and Sales:

On checking the correlation value between Sales and Promo, we get the value as 0.45 which is high. Thus, we can conclude that Sales and Promo are strongly related. Similarly, the StateHoliday and Sales are also strongly related as the Sales are high on StateHoliday.

Relation between Competition:

We plot histogram for CompetitionDistance and then convert *CompetitionOpenSince* variables to one Date variable we plot the sales for competition from open date





We thus infer that the effect of the distance to the next competitor shows lower distance to the next competitor implies slightly higher sales.

Data Preparation

Merge Data

We have merged data based on store in both test and train data-set.

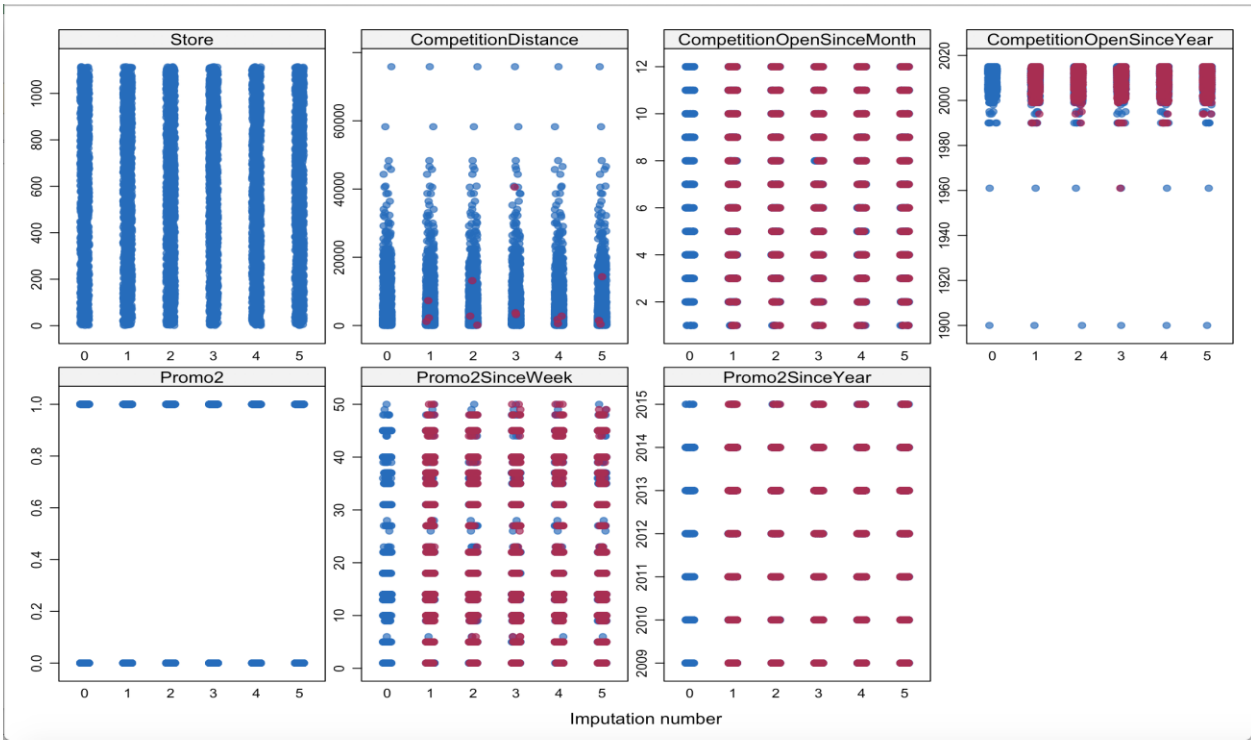
Cleansing

Removal of anamolous attributes, features not useful to research, etc.

For the cleansing, first considered whether there were any duplicates available in the datasets namely store, train and test. All the datasets returned false when checked for any duplicates. Secondly, there are no features which could be removed as all the features seem to be dependent on each other in order to produce the final output from the datasets.

Missing Values.

The store dataset consisted of large amounts of missing data in the columns like CompetitionOpenSinceMonth, CompetitionOpenSinceYear in the dataset. So, the missing data in the dataset were populated using the Mice package. The Mice Package consists of 5 types of methods. Predictive Mean Matching method is being used for populating the missing values in the store dataset. This method can generate any number of predictive datasets as per the parameter mentioned and any one dataset can be selected from the generated datasets. The below image shows the data values generated for the columns for each imputation as shown in the figure.



Renovated store, missing Open Field

In test data Open is missing, we filled these values considering if a promo is there we set store as open

	Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek
25689	622	a	c	NA	NA	NA	0	NA
25690	622	a	c	NA	NA	NA	0	NA
25693	622	a	c	NA	NA	NA	0	NA
25703	622	a	c	NA	NA	NA	0	NA
25708	622	a	c	NA	NA	NA	0	NA
25710	622	a	c	NA	NA	NA	0	NA
25717	622	a	c	NA	NA	NA	0	NA
25718	622	a	c	NA	NA	NA	0	NA
25722	622	a	c	NA	NA	NA	0	NA
25726	622	a	c	NA	NA	NA	0	NA
25727	622	a	c	NA	NA	NA	0	NA

	Promo2SinceYear	PromoInterval	Id	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday
25689	NA		3048	1	2015-09-14	NA	1	0	0
25690	NA		9040	1	2015-09-07	NA	0	0	0
25693	NA		10752	6	2015-09-05	NA	0	0	0
25703	NA		7328	3	2015-09-09	NA	0	0	0
25708	NA		8184	2	2015-09-08	NA	0	0	0
25710	NA		4760	6	2015-09-12	NA	0	0	0
25717	NA		2192	2	2015-09-15	NA	1	0	0
25718	NA		1336	3	2015-09-16	NA	1	0	0
25722	NA		6472	4	2015-09-10	NA	0	0	0
25726	NA		480	4	2015-09-17	NA	1	0	0
25727	NA		5616	5	2015-09-11	NA	0	0	0

On few dates we see there is promo so we filled value as open<-1 instead of NA.

Also we see for consecutive days the store has missing data,except for sunday we can consider that store closed due to maintenance but as its open on monday of each week we and fill it at Open<-1.

Normalization of Numeric Variables

In the train dataset, we have the components Sales and Customers. But, the fields are dependent on particular store and hence we do not require any normalization for those fields. The other fields in the datasets from all the datasets are small and would not impact much as the range of those values is less.

New Variables

The datasets have clean data and variables have direct relation among each other and do not require any new variables to be created. In order to visualize the data during the Exploratory Data Analysis, the date column from both the train and test datasets have been utilized. New Variables namely DateYear, DateMonth, DateDay, DateWeek have been added to the datasets.

Other Transformations

The store and train datasets as well as store and test datasets could be joined in order to make the predictions because the train and test data do not contain the entire data related to the store which would have an impact on the final prediction.

Modeling

The primary goal of the kaggle competition was to forecast sales for a six-week period. This will also be the primary focus of our study. Since the test data included in the competition has no actual sales we separate out the last month of data in the training set to use for validation.

Prediction

A linear model is the first choice for the sales predictions. Many other algorithms could help with classification tasks but predicting dollar value sales limits the algorithms that we can use. Since our dependent variable is continuous a linear regression seems most appropriate. We will also explore CART and random forest to determine which algorithm provides best results for this dataset. Our dependent variable will be sales and our independent variables will include store, weekday, promo, open, state and school holiday, and additional date fields created during EDA and prep. All the variables are either binary or categorical and should be straightforward to use in fitting a linear model. The amount of data in this set is quite large for evaluating on a standard home computer. Many other tutorials on this challenge only used a subset of models or selected a few stores and created models for those individually. In our study we very quickly realized that we would need to model by store rather than using store as an independent variable. However, we also wanted to be able to model across the entire dataset since although they all follow basic retail trend, the sales and trends between stores can vary significantly in shape and magnitude. We used the dplyr package to vectorize the fitting of the linear models. This avoids iteration and splitting of the data which are more computational and storage intensive. The method worked quite well for the linear models but did not extend to work well with the random forest and rpart packages. For these algorithms we chose a few stores at random to model and evaluate the results between algorithms.

Evaluation

Results

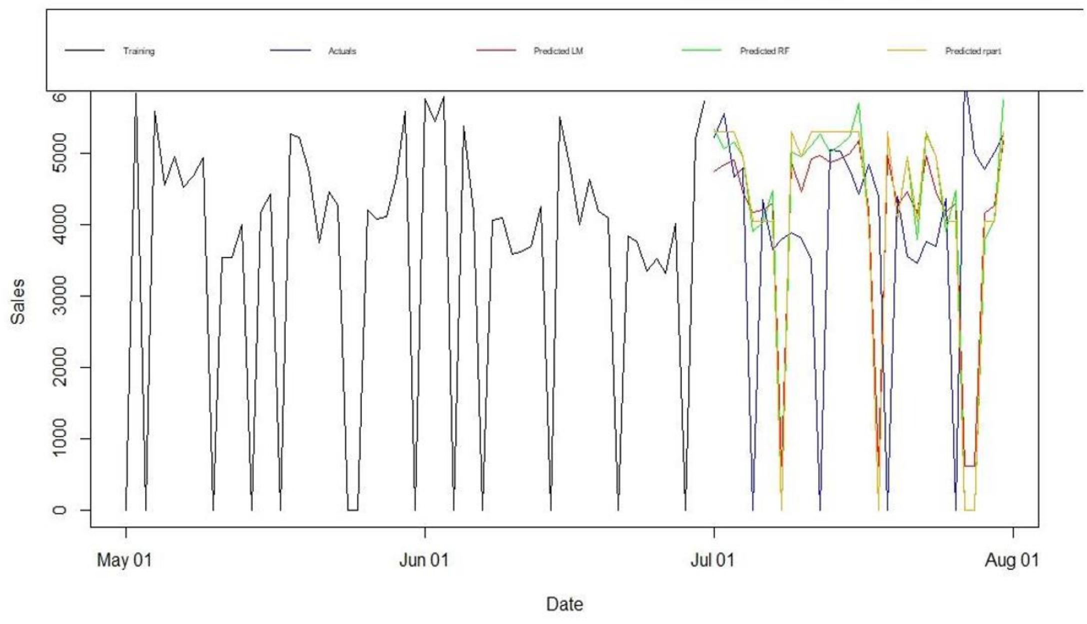
Considering the duration of time, we had for this project we ended up with pretty good results from the modeling. The Kaggle competition used root mean square percent error to determine the winner, so we will use that metric as well. In addition, we also computed the standard root mean square error. For the linear models we created across all stores we achieved a RMSPE of 0.14674 and a RMSE of 1011.24. The tables below show the comparison of results for each of the algorithms for 4 stores. Except for store 271, the random forest seems to perform quite well. Linear regression comes very close to matching the results of the random forest and CART comes in last nearly every time.

Left: Root mean square percent error by store and algorithm

Store	Linear Regression	CART	Random Forest
1	0.1138958	0.124554	0.1075935
271	0.1907402	0.1667515	0.3155605
527	0.1462616	0.1447844	0.1183477
821	0.1189555	0.1222399	0.152646

Right: Root mean square error by store and algorithm

Store	Linear Regression	CART	Random Forest
1	459.103	509.9863	442.1688
271	1300.056	1000.801	1739.977
527	1376.934	1458.406	1260.711
821	934.0313	981.2091	1102.329



Care must be taken when looking at the root mean square

error since those values cannot be used to compare accuracy between stores since each store has a different magnitude of sales. The last chart shows the actual and predicted sales by date. The tree-based models appear to have a fewer number of distinct predicted values than the linear model.

Results

All the models trained exhibit extremely low p values, so we can be confident in the variation the model explains. R squared values were typically above 0.85 with adjusted R squared not varying too much indicating that we are not likely overfitting or using variables with collinearity. Among all algorithms we consistently see that whether the store is open or not has the greatest impact followed by the day of week and then promotion. Holidays always had the lesser impact. These should be expected since a closed store intuitively results in the greatest variation of sales. Even during a promotion, it is not odd to see a significant difference between weekday and weekend sales.

Report of Results

Knowledge discovered. (overall summary of results)

Predictive Capabilities. (how can the model be used for future cases, depending on the goals of the research)

Limitations. The greatest limitation was lack of time and relative inexperience with the R language. Although we were familiar enough with R to be able to perform the necessary functions, the complexity of vectorizing to model across subgroupings of the dataset requires a greater familiarity.

Future Work. Given the positive results of the random forest algorithm we would like to see how the overall results across all models would compare with the linear regression. It would also be beneficial to do more work using the competition data to determine what impact a close competitor has versus one further away. Holidays didn’t show up as being a strong predictor in the model and exploring the reason for that would be beneficial to store managers.

References

Rossmann GmbH. Signavio. <https://www.signavio.com/customers/dirk-rossmann-gmbh/>. Accessed 7 April 2018.