Group Number 278: Rossmann Store Sale Analysis

First Name	Last Name	Email (hawk.iit.edu)	Student ID
Vishnu	Pajjuri	vpajjuri1@hawk.iit.edu	A20450928
Amrutha	Gowda	alakshmanegowda@hawk.iit.edu	A20452896
Aditya	Dubey	adubey6@hawk.iit.edu	A20432876
Gauri	Khatri	gkhatri@hawk.iit.edu	A20443345

Table of Contents

1. Introduction

Rossmann Store Sales data describes various features related to Store, Sales, Customers, StoreType, Open, StateHoliday, SchoolHoliday, Assortment, CompetitionDistance, Promo. Store Sales are influenced by many factors. We are tasked with predicting their daily sales for up to six weeks in advance. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity.

2. Data

The dataset hosted by Rossmann company.

This data has been collected from kaggle.com(https://www.kaggle.com/c/rossmann-store-sales/data) and is available in csv format (7 MB). There are 10,17,209 instances(records) and 15 attributes(columns) in the dataset.

Attributes in this dataset are as follows:

- Id - an Id that represents a (Store, Date) duple within the test set
- 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

3. Problems to be Solved

1. How sales are affected based on -School Holiday, Promo, Promo Interval, state holidays.

2. Analyze effect of promotion on sales.

3. Predicting the future sales based on the transactions performed.

4. Solutions

Make sure your solutions can solve the problems in part 3 one by one

If you are going to build predictive models, clearly indicate the dependent and independent variables

1. We will use Multiple linear regression to predict the sales on given day based on the above-mentioned factors.

2. ANOVA is performed on stores participating in promotion has high number of sales or not

3. From the models build accuracy is tested based on which we predict the future sales, dependent variable is Sales and independent variables were School Holiday, Promo, Promo Interval, State holidays

5. Experiments and Results

5.1. Methods and Process

Solve the problems your proposed one by one

1. Data Preprocessing

Missing values in the data set are replaced with mean

Creating N-1 variables for categorical values like StoreType

Converting nominal to numerical variables: Yes/No to 0/1 for Promo, State Holiday, School Holiday

Remodelled the date to MM/DD/YYYY format using date function to maintain symmetric pattern throughout.

```
> head(final_train_new)
  Store DayOfWeek_1 DayOfWeek_2 DayOfWeek_3 DayOfWeek_4 DayOfWeek_5 DayOfWeek_6 DayOfWeek_7 Sales Open Promo
1     1           0           0           0           0           1           0           0  5263    1     1
2     1           0           0           0           0           0           1           0  4952    1     0
3     1           0           0           0           0           1           0           0  4190    1     0
4     1           0           0           1           0           0           0           0  6454    1     1
5     1           0           0           1           0           0           0           0  3310    1     0
6     1           0           0           0           0           0           0           1     0    0     0
  StateHoliday SchoolHoliday StoreType_1 StoreType_2 StoreType_3 StoreType_4 Assortment_1 Assortment_2 Assortment_3
1            1             1           0           0           1           0            1            0            0
2            1             0           0           0           1           0            1            0            0
3            1             1           0           0           1           0            1            0            0
4            1             0           0           0           1           0            1            0            0
5            1             0           0           0           1           0            1            0            0
6            1             0           0           0           1           0            1            0            0
  CompetitionDistance CompetitionOpenSinceMonth_1 CompetitionOpenSinceMonth_2 CompetitionOpenSinceMonth_3
1                1270                           0                           0                           0
2                1270                           0                           0                           0
3                1270                           0                           0                           0
4                1270                           0                           0                           0
5                1270                           0                           0                           0
6                1270                           0                           0                           0
```

2. ANOVA:

Null Hypothesis: Average sales with promo and without promo are same
Alternate Hypothesis: Average sales with promo and without are not same

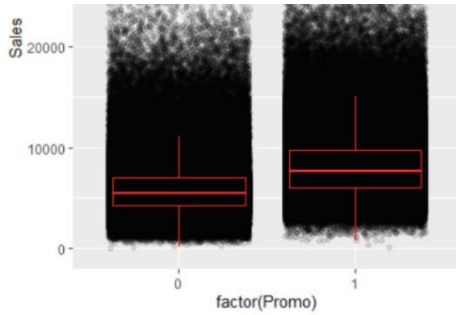
```
> anov=lm(sales~promo)
> summary(anov)

Call:
lm(formula = sales ~ promo)

Residuals:
   Min     1Q Median     3Q    Max 
 -7991  -2278    -30   1852  37145 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4406.051      4.329  1017.8   <2e-16 ***
promo       3585.101      7.008   511.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3434 on 1017207 degrees of freedom
Multiple R-squared:  0.2046,  Adjusted R-squared:  0.2046 
F-statistic: 2.617e+05 on 1 and 1017207 DF,  p-value: < 2.2e-16
```



As p-value is less than 0.05 we reject Null hypothesis and accept alternate Hypothesis

3. Multiple Linear Regression

At 95% confidence level, P value is less than 0.05. There is **60.11**% variation in Sales can be explained by the variations in dependent variables

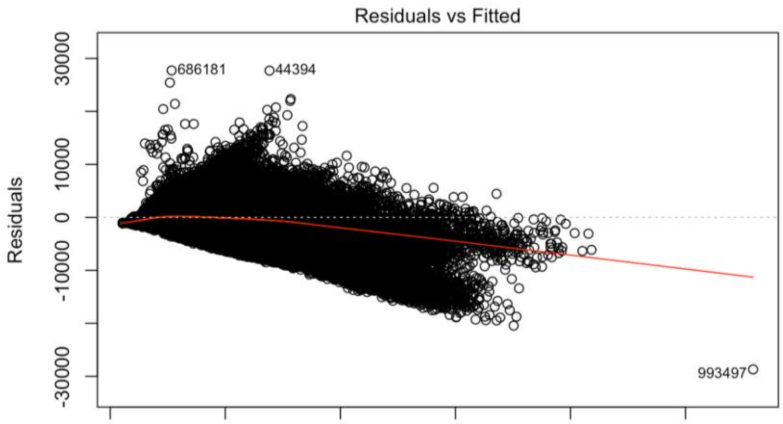
```
Call:
lm(formula = Sales ~ ., data = new_train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7496 -0.3799 -0.0616  0.2595  8.9750 

Coefficients: (8 not defined because of singularities)
                              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                 -9.978e-15  6.262e-04    0.000 1.000000    
DayOfWeek_1                  1.650e-02  2.254e-03    7.318 2.52e-13 ***
DayOfWeek_2                 -8.048e-02  2.268e-03  -35.480  < 2e-16 ***
DayOfWeek_3                 -1.100e-01  2.275e-03  -48.367  < 2e-16 ***
DayOfWeek_4                 -1.055e-01  2.242e-03  -47.044  < 2e-16 ***
DayOfWeek_5                 -7.838e-02  2.277e-03  -34.430  < 2e-16 ***
DayOfWeek_6                 -8.096e-02  2.259e-03  -35.830  < 2e-16 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6316 on 1017113 degrees of freedom
Multiple R-squared:  0.6011,    Adjusted R-squared:  0.6011
F-statistic: 1.614e+04 on 95 and 1017113 DF,  p-value: < 2.2e-16
```



4. K- Nearest Neighbor

PreProcessing

We have added one more column sale status by calucalting the mean and checking all the sale values which are above mean or not.

	Id	Store	DayOfWeek	Date	Open	Promo	StateHoliday	SchoolHoliday	Sales	SalesStatus
1	1	1	4	9/17/2015	1	1	0	0	5263	no
2	2	3	4	9/17/2015	1	1	0	0	6064	no
3	3	7	4	9/17/2015	1	1	0	0	8314	yes
4	4	8	4	9/17/2015	1	1	0	0	13995	yes
5	5	9	4	9/17/2015	1	1	0	0	4822	no
6	6	10	4	9/17/2015	1	1	0	0	5651	no
7	7	11	4	9/17/2015	1	1	0	0	15344	yes
8	8	12	4	9/17/2015	1	1	0	0	8492	yes
9	9	13	4	9/17/2015	1	1	0	0	8565	yes
10	10	14	4	9/17/2015	1	1	0	0	7185	yes
11	11	15	4	9/17/2015	1	1	0	0	10457	yes
12	12	16	4	9/17/2015	1	1	0	0	8959	yes
13	13	19	4	9/17/2015	1	1	0	0	8821	yes
14	14	20	4	9/17/2015	1	1	0	0	6544	no
15	15	21	4	9/17/2015	1	1	0	0	9191	yes

By performing KNN we get the impact of variables on sales.

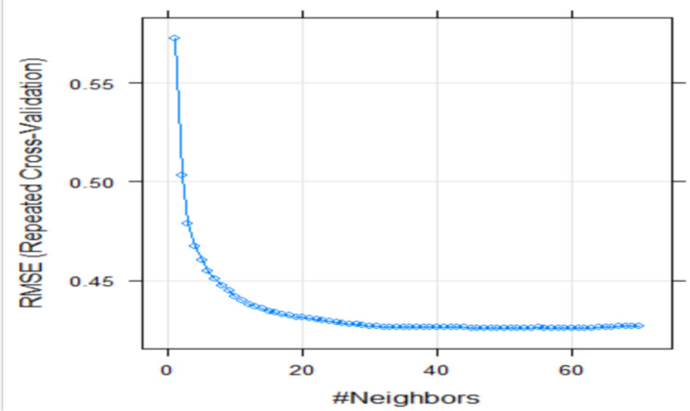
School Holiday and Promo has major impact on the Sales.

```
k-Nearest Neighbors

5143 samples
   6 predictor
   2 classes: 'no', 'yes'

Pre-processing: centered (6), scaled (6)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 4628, 4628, 4629, 4629, 4629, 4629, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.6504620  0.2969388
   2  0.6502060  0.2956611
   3  0.6675120  0.3264492
   4  0.6720454  0.3350319
   5  0.6789808  0.3467064
   6  0.6846204  0.3572214
   7  0.6928504  0.3730133
   8  0.6947313  0.3765735
   9  0.6992042  0.3850284
  10  0.7003688  0.3871562
  11  0.7052961  0.3962581
  12  0.7040011  0.3934900
  13  0.7091859  0.4037098
  14  0.7102869  0.4061333
  15  0.7098974  0.4048302
  16  0.7117118  0.4085577
  17  0.7140456  0.4130137
  18  0.7165731  0.4179507
  19  0.7181954  0.4209264
  20  0.7176106  0.4197171
```



5. Naïve Bayes

```
train<- train[,-c(1,4,9)]
#Naive bayes
head(train)
model <- naiveBayes(SalesStatus ~., data=train)
model
plot(model)


train %>%
  filter(SalesStatus == "1")   %>%
  summarise(mean(Promo), sd(Promo))
plot(model)
```

```
> model <- naiveBayes(SalesStatus ~., data=train)
> model

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.4631068 0.5368932 

Conditional probabilities:
   Store
Y       [,1]     [,2]
  0 554.7908 320.5263
  1 523.5302 317.7189

   DayOfWeek
Y       [,1]     [,2]
  0 3.763941 1.645884
  1 4.076311 1.940961

   Open
Y        [,1]      [,2]
```

6. Decision Tree:

```
> summary(dt_regressor_1)
Call:
rpart(formula = Sales ~ ., data = new_train, control = rpart.control(minsplit = 1))
  n= 1017209

    CP nsplit rel error xerror xstd
1 0.01      0         1      0    0

Node number 1: 1017209 observations
  mean=5773.819, MSE=1.482192e+07

> head(dt_pred_1)
[1] 5773.819 5773.819 5773.819 5773.819 5773.819 5773.819
```

7. Random Forest:

```
Call:
 randomForest(x = train[, feature.names], y = log(train$Sales + 1), ntree = 50, mtry = 5, sampsize = 1e+05, do.trace = TRUE)
               Type of random forest: regression
                     Number of trees: 50
No. of variables tried at each split: 5

          Mean of squared residuals: 0.02559512
                    % Var explained: 86.22
```

Give the necessary codes, snapshots and explanations

5.2. Evaluations and Results

Given a same problem, you may have several solutions or build several models

1. MLR

```
> # MLR Model Prediction Summary
> MAE(predictions,new_test$Sales)
[1] 1711.394
> RMSE(predictions,new_test$Sales)
[1] 2431.425
```

2. K- Nearest Neighbor

```
RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 60.
> plot(fit)
> varImp(fit)
loess r-squared variable importance

              Overall
SchoolHoliday  100.00
Promo           99.62
DayOfWeek       31.86
Open            27.46
```

3. Decision Tree

```
> # Decision Tree Prediction Summary
> MAE(dt_pred_1,new_test$Sales)
[1] 2887.725
> RMSE(dt_pred_1,new_test$Sales)
[1] 3849.924
```

4. Random Forest

Adjusted R square value is **89.55**% from which we can say this is the best model so far.

```
Model Summary:

MSE:  0.01888739
RMSE:  0.1374314
MAE:  0.09929556
RMSLE:  0.01420444
Mean Residual Deviance :  0.01888739
Adj.R^2 :  0.895532
```

	A	B	C
1	Id	Sales	
2	1	4641.38	
3	24825	4769.471	
4	5993	3702.15	
5	37665	5466.344	
6	18833	3734.717	
7	12841	6251.681	
8	19689	3687.141	
9	857	4804.693	
10	15409	5171.719	
11	38521	6256.501	
12	13697	5829.417	
13	7705	3879.128	
14	39377	5209.837	
15	20545	3905.591	
16	1713	5309.991	
17	16265	4794.677	
18	33385	4944.159	
19	27393	4868.176	
20	2569	5871.803	
21	40233	4942.17	
22	21401	4853.123	
23	35953	4916.571	
24	34241	4562.616	
25	3425	4865.736	
26	22257	4457.034	

rf1

Evaluate your solutions based on selected metrics and compare them

5.3. Findings

- Provide the summary of your findings, explanations, conclusions
- Sales depends on Promo, store with promotion have highest number of sales.
- School Holidays does not impact much on sales.
- Among all the models p value got for Random Forest is highest hence we are predicting future sales with this model.

6. Conclusions and Future Work

6.1. Conclusions

- A short summary of your whole project and conclusions, such as what you want to, why you want to do so, which solutions you use, and which findings or final results you get finally.
- One can make a simple machine learning model that predicts the sales of the Rossmann stores with a 13.7% error.
- Model only uses a fraction of the features provided by Rossmann. The model can therefore be implemented in a simple app and easily accessed by store managers to accurately forecast sales.

6.2. Limitations

- Applied only three algorithms i.e. MLR, KNN, Naïve Bayes, random forest. So, there are scope for applying more algorithms like time series linear models, XGBoost, Unobserved Component Model, Principal Component Regression,
- By taking the regression of all the models for all the sales data may predict the sales better. We would have weighted average of two or more models we would have got better result.

6.3. Potential Improvements or Future Work

- We believe the sales number of a day is also related to the sales number before that day. Adding time series to the model can improve accuracy.
- We will try adding time series to the feature vector to see what we can achieve.
- Different machine learning algorithm such as GLM Poisson model can also be interesting to explore.