Group 278: Rossman Store Sale Analysis

Output Snapshots:

1. Data Pre-processing

```
> 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

```
> 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
```

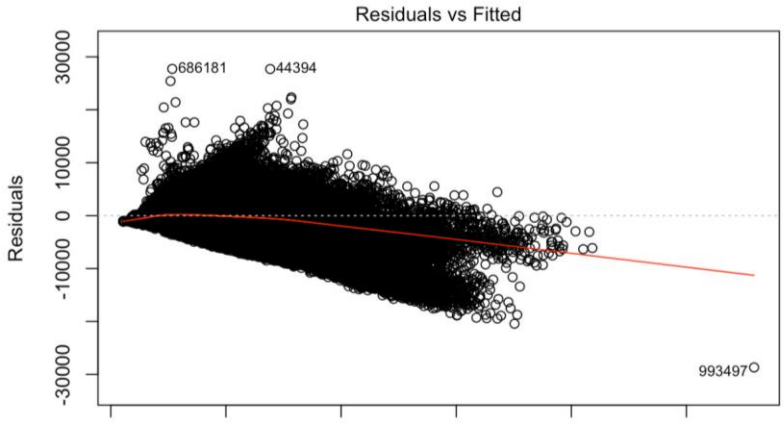
3. Multiple Linear Regression

```
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
```




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

Pre-processing

	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

```
data <-read.csv("C:/Users/dubey/OneDrive/Desktop/IIT COURSES/DATA-ANY
head(data)

#adding a label
View(data)
unique(data$Sales)
for(i in 1 :nrow(data)){
  if(data$Sales[i]>= 6959){data$SalesStatus[i]<- 'yes'}
  else{ data$SalesStatus[i]<- 'no'}

}


str(data)
```

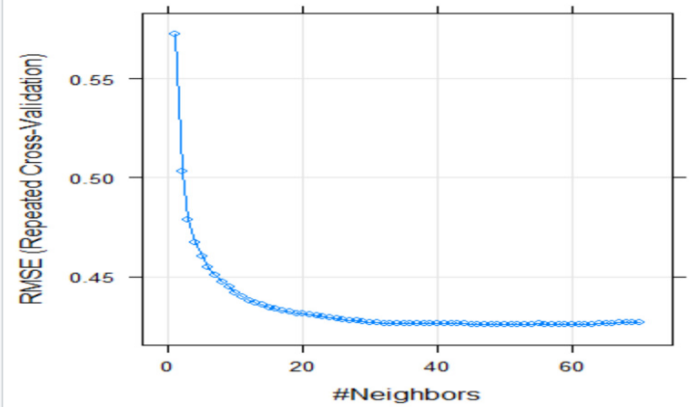
4. K- Nearest Neighbour

```
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
```




```
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
```

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)
```

53:1 (Top Level)

Console Terminal Jobs

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

7. Random Forest

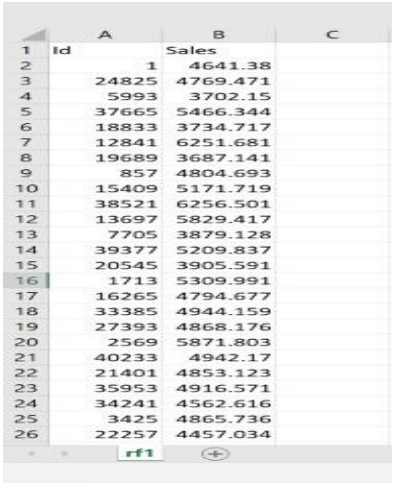
```
Call:
 randomForest(x = train[, feature.names], y = log(train$Sales + 1), ntree = 50, mtry = 5, sampsize = 1e+05, do.trace = TRUE)
               Type of random forest: regression
                     Number of trees: 50
No. of variables tried at each split: 5

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

```
Model Summary:

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

6 weeks prediction



	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