Project Name

Employee Absenteeism

By : Nilanjan Singha Mahapatra

Date : 12 August 2018

Introduction :

The aim of the project is to find the cause of the issue of Absenteeism faced by the XYZ Courier Company and to forecast their losses if the issue of Absenteeism continues with the same trend.

DATA :

Data presented to us consist of both categorical as well as continuous variables while the target variable is continuous variable.

Below is the sample of our dataset.

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism
11	26	7	3	1	289	36	13	33	239,554	97	0	1	2	1	0	1	90	172	30	4
36	0	7	3	1	118	13	18	50	239,554	97	1	1	1	1	0	0	98	178	31	0
3	23	7	4	1	179	51	18	38	239,554	97	0	1	0	1	0	0	89	170	31	2
7	7	7	5	1	279	5	14	39	239,554	97	0	1	2	1	1	0	68	168	24	4

METHODOLOGY

1.Pre-processing the data.

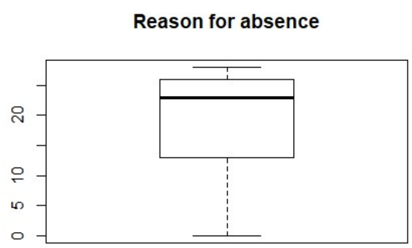
Any modeling requires a detailed look at the data before we start modeling. However in data mining terms looking into data refers to more than looking. Looking at the data refers to exploring the data, cleaning the data. This is often called Exploratory Analysis of the data.

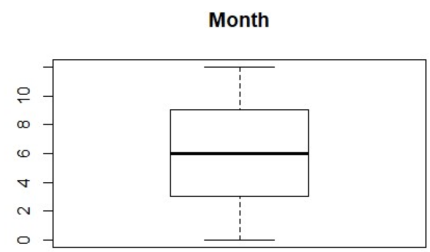
Missing Values:

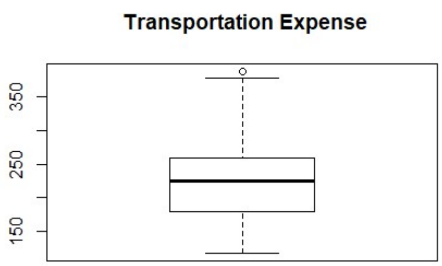
Missing values in data can change the result we expect from our models. Missing values can be different variables of the data . If Missing values is found in any observation of a particular variable we can either remove the variable if it is not that important to our model or we can input the Missing value of that observation by using different method like Mean and Imputation methods. In this case we have gone for Imputation Since the missing values were not MCAR in nature but MAR So we went for Imputation using **MICE** which stands for:**Multiple Imputation by Chained Equations.** (See Apendix for the code on Missing Value Analysis)

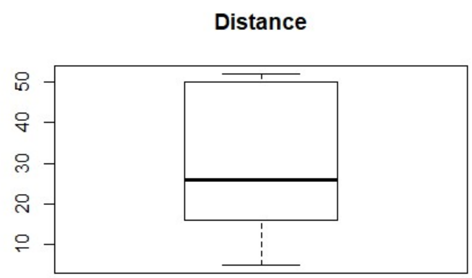
Outliers

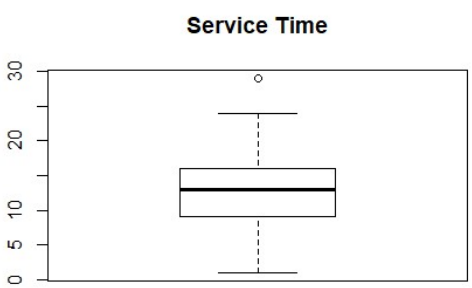
An outlier is an observation point that is at unusual distant from other observations.
An outlier may be due to variability in the measurement or it may indicate experimental error;
In case of experimental error we can exclude it from the data set.
Here we have used box plot to determine the outlier analysis of all predictor variables.
Below are the box plots of every variable for Outlier Analysis. (See Apendix for the code on Outlier Analysis)

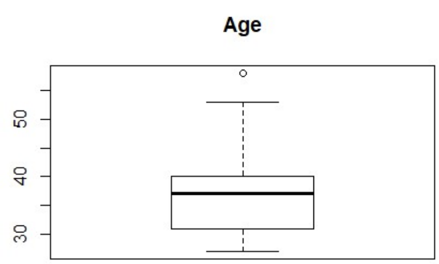

Reason for absence
0
5
10
20

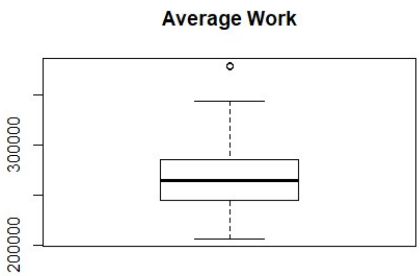

Month
0
2
4
6
8
10

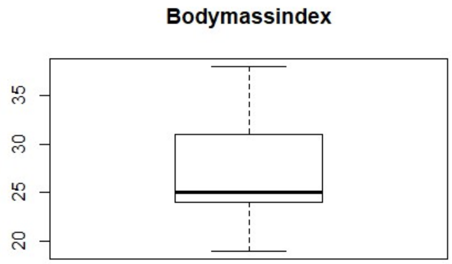

Transportation Expense
150
250
350

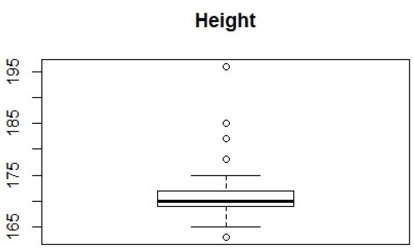

Distance
10
20
30
40
50

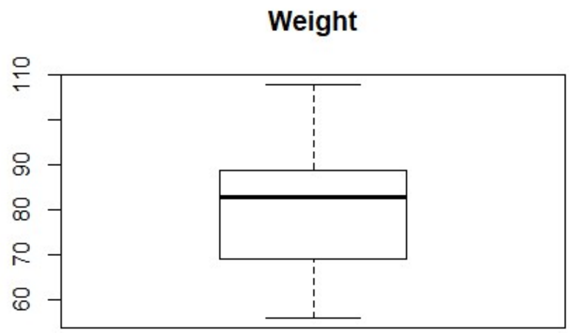

Service Time
0
5
10
20
30

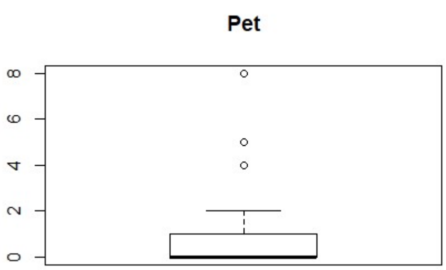

Age
30
40
50

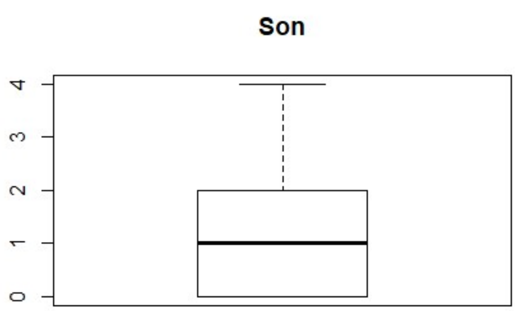

Average Work
200000
300000

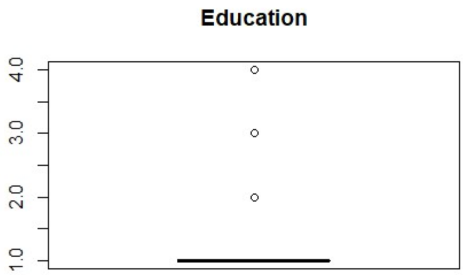

Bodymassindex
20
25
30
35

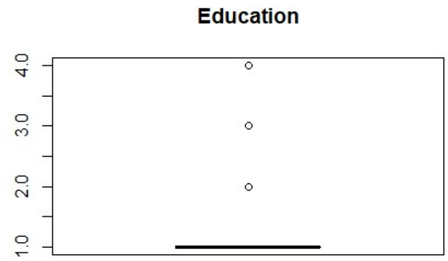

Height
165
175
185
195


Weight
60
70
80
90
110


Pet
0
2
4
6
8


Son
0
1
2
3
4





Outliers

Most of the variables had Outliers but since most of the data were important to the model development for example In the variable age : The Age '59' had issue but we can't remove that since a person can be 59 years old. But two variables Absenteeism and Service Time had improper Outliers . Service Time had more than 24 hrs as service for a particular day. Similarly the absenteeism had more than 24 hrs for each day which is also not proper. So we removed these observations from the data.

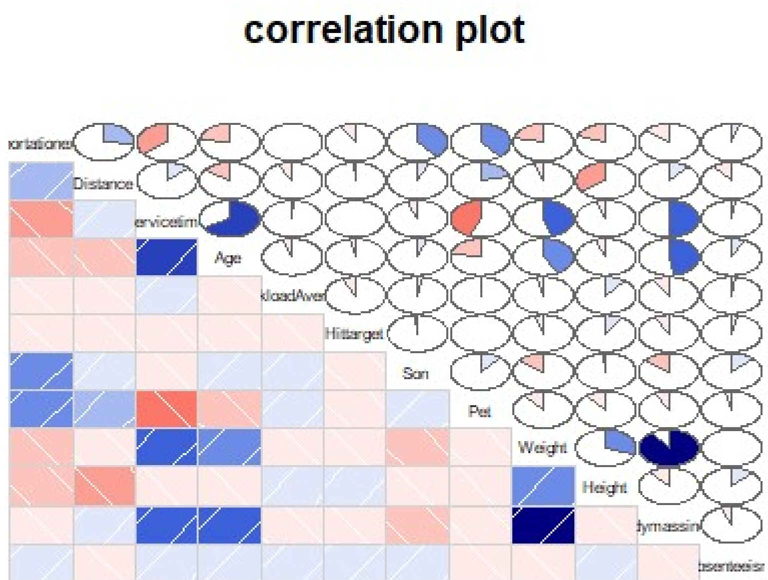
So, First we will go for Model Development with Outliers and then without Outliers just to see the difference.

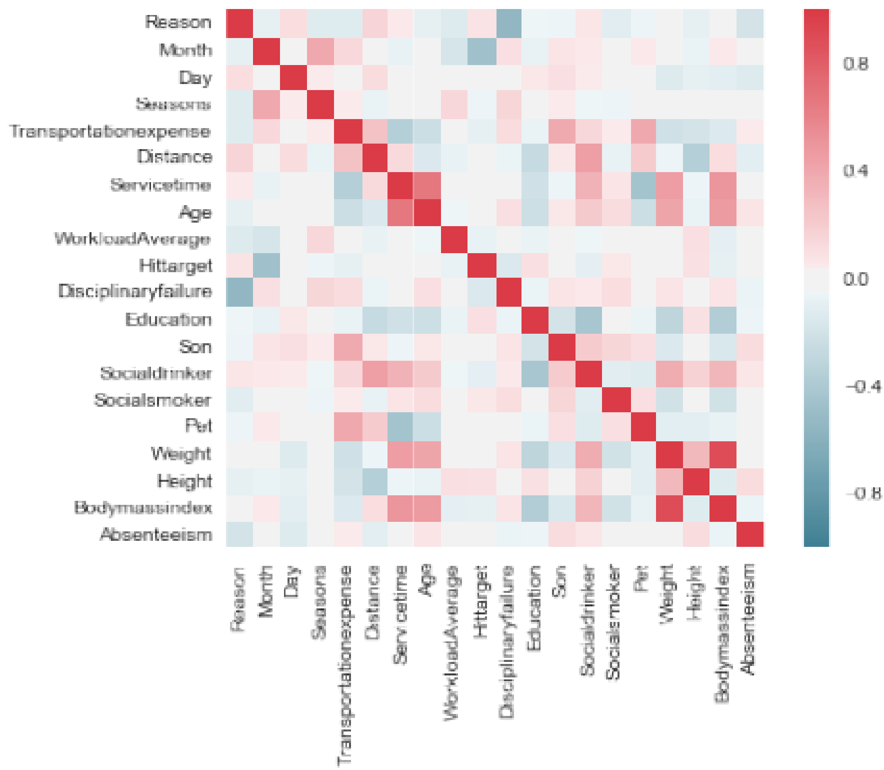
Feature Selection :

Feature Selection is a very important aspect of Data Pre-Processing. Here, We will check Whether there is multicolinearity between the predictor variables. To have the best data fitted Into the model there should be no correlation among independent variables. Ideally the Correlation should be zero between independent variables and high among independent and Dependent Variables.

We will use Correlation graph and heat map to check the Correlation among the variables. If there is high correlation between two independent variables we can remove any one of them.

(See Apendix for the code on Feature Selection)





As per the above figures we can see that there is very high correlation between Weight and Body Mass index and also no correlation between Absenteeism(target) with weight. So We can drop the feature Weight from our data . So we select all features except weight.

Model Development

After Data Pre-processing comes the Model Development part. Since we have a continuous variable as our target variable so we are going for Random Forest and Linear Regression.

Random Forest Regressor

The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning.

Since our target variable is Continuous so we will go for random forest Regressor.

In Random Forest Model first we use the data without outliers, We split the data into training and test data and the no of decision tree is taken as 500 and no of samples for each split is based on:

- ➔ floor(sqrt(ncol(train) - 1)) where train is the training set.

Which in this case comes to be 4.

Below is the model summary for Random Forest with Outliers.

```
Call:
 randomForest(formula = Absenteeism ~ ., data = train, importance = TRUE,
ntree = 500, mtry = floor(sqrt(ncol(train) - 1)))
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 206.2692
                    % Var explained: 4.8
```

The % var is very low to 4.8 %

Now let's fit the data with no Outliers and check the Random Forest model summary

```
Call:
 randomForest(formula = Absenteeism ~ ., data = train, importance = TRUE,
ntree = 500, mtry = floor(sqrt(ncol(train) - 1)))
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 15.66559
                    % Var explained: 25.53
```

The % var increases to 25.33 . Hence we are going to predict the test data with this model.

Multiple Linear Regression

In Multiple Linear Regression, We will divide the pre-processed data into two parts i.e train data and test data and then fit the train data into the Multiple Linear Regression . We then check the various factors from the model to get details about the model. First we will build the model with Data having outliers like we did in random Forest Regressor and then check the model summary.

```
Call:
lm(formula = Absenteeism ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.824  -4.054  -1.540   1.669 103.020 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -4.429e+01  2.798e+01  -1.583   0.1141    
Reason                -4.487e-01  7.913e-02  -5.670 2.42e-08 ***
Month                 -6.996e-02  2.056e-01  -0.340   0.7337    
Day                   -9.189e-01  3.857e-01  -2.383   0.0176 *  
Seasons                2.632e-01  5.695e-01   0.462   0.6442    
Transportationexpense  3.867e-03  1.062e-02   0.364   0.7159    
Distance              -5.151e-02  5.494e-02  -0.938   0.3488    
Servicetime            1.784e-01  2.292e-01   0.778   0.4367    
Age                    1.373e-01  1.337e-01   1.027   0.3050    
WorkloadAverage       -1.580e-05  1.550e-05  -1.019   0.3087    
Hittarget              2.102e-01  1.698e-01   1.238   0.2163    
Disciplinaryfailure   -1.544e+01  2.891e+00  -5.340 1.41e-07 ***
Education             -1.702e+00  1.006e+00  -1.691   0.0915 .  
Son                    1.166e+00  5.521e-01   2.112   0.0352 *  
Socialdrinker          1.326e+00  1.768e+00   0.750   0.4536    
Socialsmoker          -3.015e+00  2.226e+00  -1.354   0.1762    
Pet                    9.983e-02  5.031e-01   0.198   0.8428    
Height                 2.835e-01  1.138e-01   2.491   0.0131 *  
Bodymassindex         -2.473e-01  1.770e-01  -1.397   0.1629    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.02 on 499 degrees of freedom
Multiple R-squared:  0.1542,	Adjusted R-squared:  0.1237 
F-statistic: 5.053 on 18 and 499 DF,  p-value: 1.189e-10
```

As you can see the model has very low Adjusted R square value with only 12.37 % .

Now let’s build the same model without the outliers.

```
Call:
lm(formula = Absenteeism ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.9939  -2.0564  -0.7656   1.0939  20.0888 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            4.036e+00  9.870e+00   0.409   0.6828    
Reason                -2.151e-01  2.850e-02  -7.545 2.32e-13 ***
Month                 -6.611e-02  7.026e-02  -0.941   0.3472    
Day                   -2.331e-01  1.340e-01  -1.739   0.0827 .  
Seasons                5.420e-02  1.940e-01   0.279   0.7801    
```

```
Transportationexpense  1.561e-02  3.925e-03   3.977 8.06e-05 ***
Distance              -9.128e-03  1.927e-02  -0.474   0.6359    
Servicetime            8.235e-02  8.167e-02   1.008   0.3138    
Age                   -7.246e-02  4.591e-02  -1.578   0.1152    
WorkloadAverage        9.468e-06  5.214e-06   1.816   0.0700 .  
Hittarget             -8.452e-02  5.623e-02  -1.503   0.1334    
Disciplinaryfailure   -9.901e+00  1.014e+00  -9.763  < 2e-16 ***
Education              2.622e-01  3.215e-01   0.816   0.4151    
Son                    4.157e-01  2.001e-01   2.077   0.0383 *  
Socialdrinker          1.100e+00  6.316e-01   1.742   0.0821 .  
Socialsmoker           3.802e-01  7.849e-01   0.484   0.6283    
Pet                   -3.485e-01  1.805e-01  -1.930   0.0542 .  
Height                 4.257e-02  4.271e-02   0.997   0.3194    
Bodymassindex          8.274e-02  6.116e-02   1.353   0.1767    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.006 on 475 degrees of freedom
Multiple R-squared:  0.2575,	Adjusted R-squared:  0.2294 
F-statistic: 9.151 on 18 and 475 DF,  p-value: < 2.2e-16
```

As you can see above the number of significant variables also increased along with the Adjusted R squared value to 23 % which is not good but better than the previous. Hence we will predict the test data based on this model.

MODEL EVALUATION

For Model Evaluation we will calculate MSE (Mean Square Error) for both the models:

Random Forest Regressor :

For Random Forest Regressor MSE is : 13.65

Multiple Linear Regression MSE is : 15.37

CONCLUSION

As the Random Forest Regressor is having the least MSE out of the two . So we would select Random Forest Regressor over the two.

Ans 2: If the following trend (which is positive between month and absenteeism) continues the forecasted loss per month would be around 1,81,291 Work Load . That means 1,81,294 amount of work would not be carried out and hence would go to loss.

Ans 1: As per the decision tree significance levels in descending order following are the list

```
Reason                30.1525957
Month                  6.6911800
Day                    0.4136174
Seasons                5.5033282
Transportationexpense 11.6583630
Distance               7.4935318
Servicetime            8.7222257
Age                    9.6414504
WorkloadAverage        2.4542190
Hittarget              5.1507549
Disciplinaryfailure   17.5994945
Education             -0.1384012
Son                    8.5078184
Socialdrinker          5.2758807
Socialsmoker           0.8685886
Pet                    7.2149781
Height                 8.3270568
Bodymassindex          7.7316648
```

As you can see the decision tree has reason as the most significant variable and then transportation expense, Disciplinary failure with education as the least significant. So the firm should reduce the transportation expense , discipline failure, service time, reduce Age (Hire young people), less social drinkers and reduce the Workload Average.

APPENDIX

Rcode:

Missing Value Analysis using MICE

```
install.packages("mice")
library(mice)
tempData <- mice(data_new,m=5,maxit=50,seed=500)
completeData <- complete(tempData,2)
class(completeData)
```

Outliers

```
cnames = colnames(completeData)
```

```
# to detect outliers in the data
for(i in cnames)
{
  print(i)
  val = completeData[,i][completeData[,i] %in% boxplot.stats(completeData[,i],coef=1.5)$out]

  print(val)
}
```

Removing Outliers

```
for(i in (1:740)) {

  if(completeData_final[i,19] > 39) {
    rows_not_keep <- i
    completeData_final <- completeData_final[-(i),]
    print(i)
  }

}

for(i in (1:711)) {

  if(completeData_final[i,7] > 24) {
    rows_not_keep <- i
    completeData_final <- completeData_final[-(i),]
    print(i)
  }

}
```

Feature Selection

```
data_numeric = completeData[,(5:20)]
data_numeric = data_numeric[,-(7:8)]
data_numeric = data_numeric[,-(8:9)]


install.packages("corrgram")
library(corrgram)
```

```
corrgram(data_numeric, order=F ,upper.panel = panel.pie, text.panel = panel.txt,
main='correlation plot')

install.packages("caret")
library(caret)
corelationmatrix <- cor(data_numeric)

print(corelationmatrix)

# find attributes taht are highly correlated (ideally >0.75)

highlycorrelated <- findCorrelation(corelationmatrix, cutoff = 0.7)

print(highlycorrelated)
```