Employee Absenteeism

Ratan Kumar

28th june 2018

Contents

Chapter 1

Introduction

Problem Statement

XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The company has shared it dataset and requested to have an answer on the following areas:

1. What changes company should bring to reduce the number of absenteeism?

2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

The Aim of this project is to resolve the above both issues.

1.2 Data

We need to build a model which can predict absenteeism time in hours of each employee based on the available data associated with each of them respectively.

Once we have got a model then we need to calculate loss of the company occurred due to absenteeism of employees.

ID	Reason fo	Month of	Day of the	Seasons	Transport	Distance f
11	26	7	3	1	289	36
36	0	7	3	1	118	13
3	23	7	4	1	179	51
7	7	7	5	1	279	5
11	23	7	5	1	289	36
3	23	7	6	1	179	51

Table 1.2.1: An overview of dataset

In the given dataset there are 21 variables and 740 observations. All of 21 variables are provided as numeric out of which 20 are predictors and the variable absenteeism time in hours is target variable.

Chapter 2

Methodology

2.1 Pre-processing

we need to make the data clean and transform it to a standard form before considering the type of model and problem statement. To start with, first we should analyze the variable type the dataset is having.
Here, all the variables are in numeric form which we can detect using below R code:

```
data = read.xlsx("Absenteeism_at_work_Project.xls", sheetIndex = 1)

str(data)
```

Just a quick look over the data and variables, one can find that the dataset is having categorical variables too rather than only numerical variables.

The data(plurals) of the variables Day of the week, Month of absence, etc indicates that they are representation of categories instead of numeric.

Thus for all such variables, we will convert these respective data types.

For this dataset I find below observations:-

Numerical variables	Categorical Variables
Distance From residence to Work	ID
Height	Day of the week
Service Time	Education
Hit Target	Social Smoker
Son	Reason for absence
Pet	Season
Body Mass Index	Month of absence
Transportation Expenses	Disciplinary Failure
Age	Social Smoker
Weight	
Work Load Average/day	
Absenteeism Time in Hours	

2.a.(i) Data Type Conversion

The data types can be converted using R code as below:-

data$ID = as.factor(as.character(data$ID))

data$Day.of.the.week = as.factor(as.character(data$Day.of.the.week))

data$Education = as.factor(as.character(data$Education))

and so on for the rest of variables.

2.a.(ii) Missing Value Analysis

The given dataset is having many values missing which we can detect using code below.

missing_value = data.frame(apply(data,2,function(x){sum(is.na(x))}))

Thus, here before proceeding we need to impute the missing values first using the best suited method of imputation.

After few hit and trials I find that KNN is slightly better than median and mean methods. However, median and KNN methods both are having the same imputed values in few cases.

I am finalizing KNN here to impute the missing values.

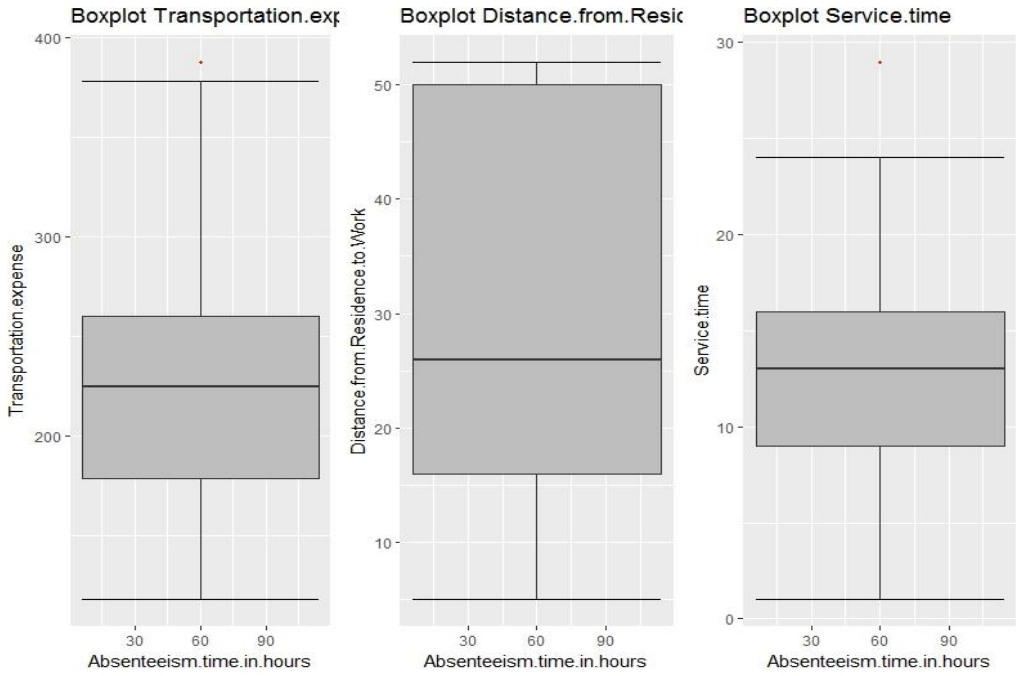
library(DMwR)

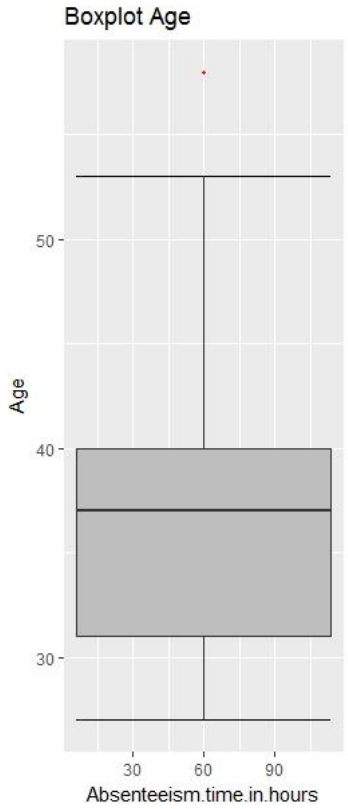
data = knnImputation(data,k=3)

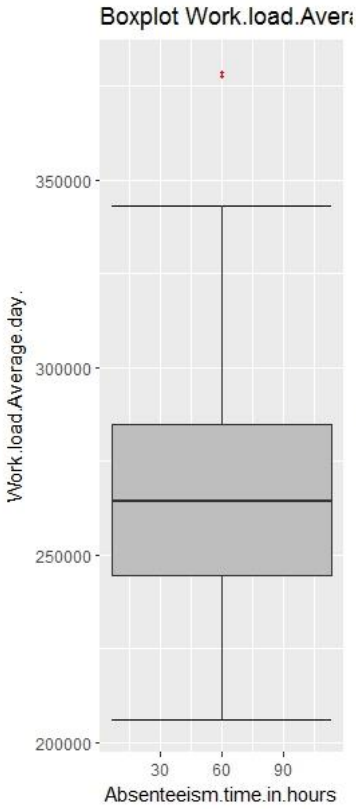
2.a.(iii) Outlier Analysis

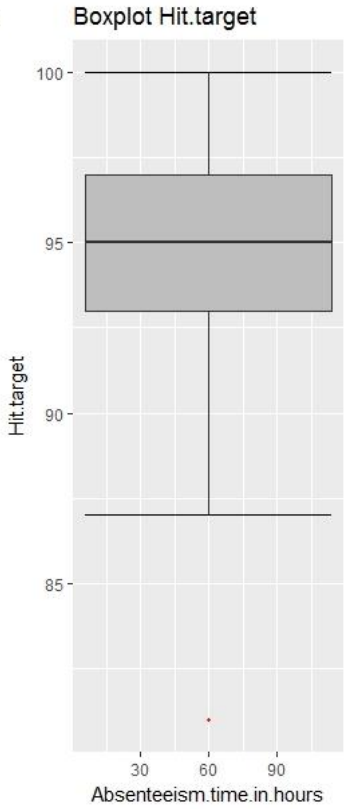
Outliers are the unwanted abnormal values that may get generated due to rough handling of data or an out of the range value in which most of the point lies.

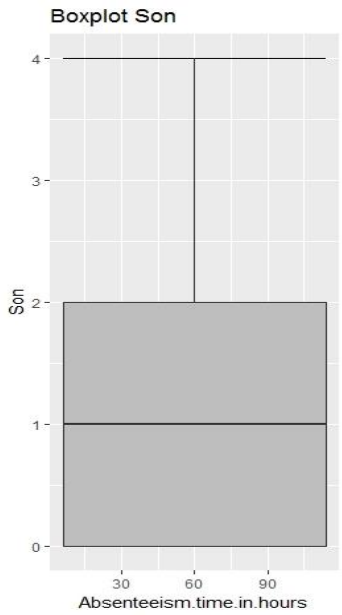
For this dataset I am using Boxplot method to get if the outliers are present.

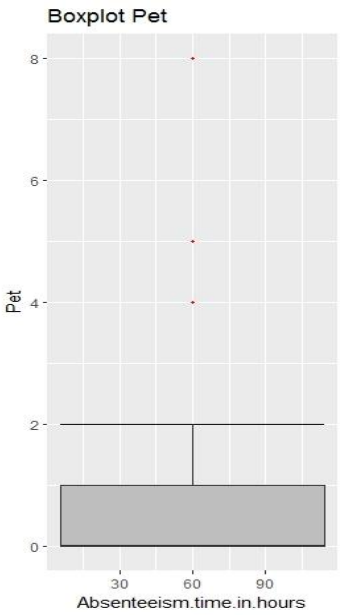


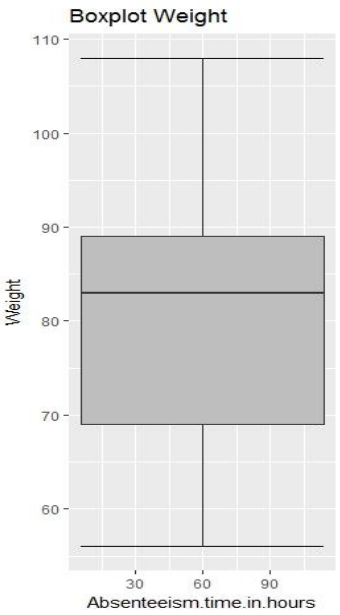

Boxplot Age
Age
50
40
30
30
60
90
Absenteeism.time.in.hours


Boxplot Work.load.Aver
Work.load.Average.day.
350000
300000
250000
200000
30
60
90
Absenteeism.time.in.hours


Boxplot Hit.target
Hit.target
100
95
90
85
30
60
90
Absenteeism.time.in.hours


Boxplot Son
Son
4
3
2
1
0
30
60
90
Absenteeism.time.in.hours


Boxplot Pet
Pet
8
6
4
2
0
30
60
90
Absenteeism.time.in.hours


Boxplot Weight
Weight
110
100
90
80
70
60
30
60
90
Absenteeism.time.in.hours

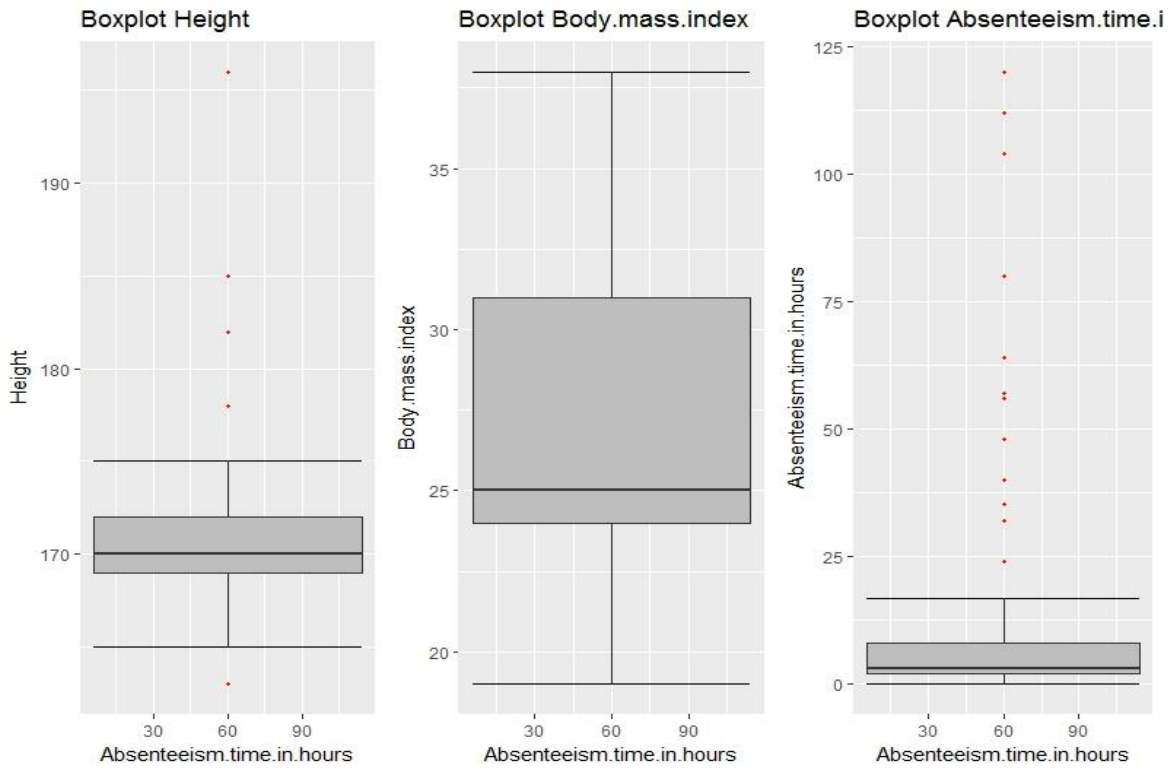


Fig 2.1 Boxplot for variables to detect outliers or anomalies

Here the box plot displays outliers of each variable. Now we have to remove these outliers.

In order to remove the outliers first we need to replace each of the outlier values with NA and then impute using already selected most suitable method of imputation KNN.

Having a look at this data I feel that outliers available in Absenteeism time in hours must not be treated. This is because it may be possible that a single person remain absent throughout the month and considering the abruptally high absent time as outlier may make the model less efficient.

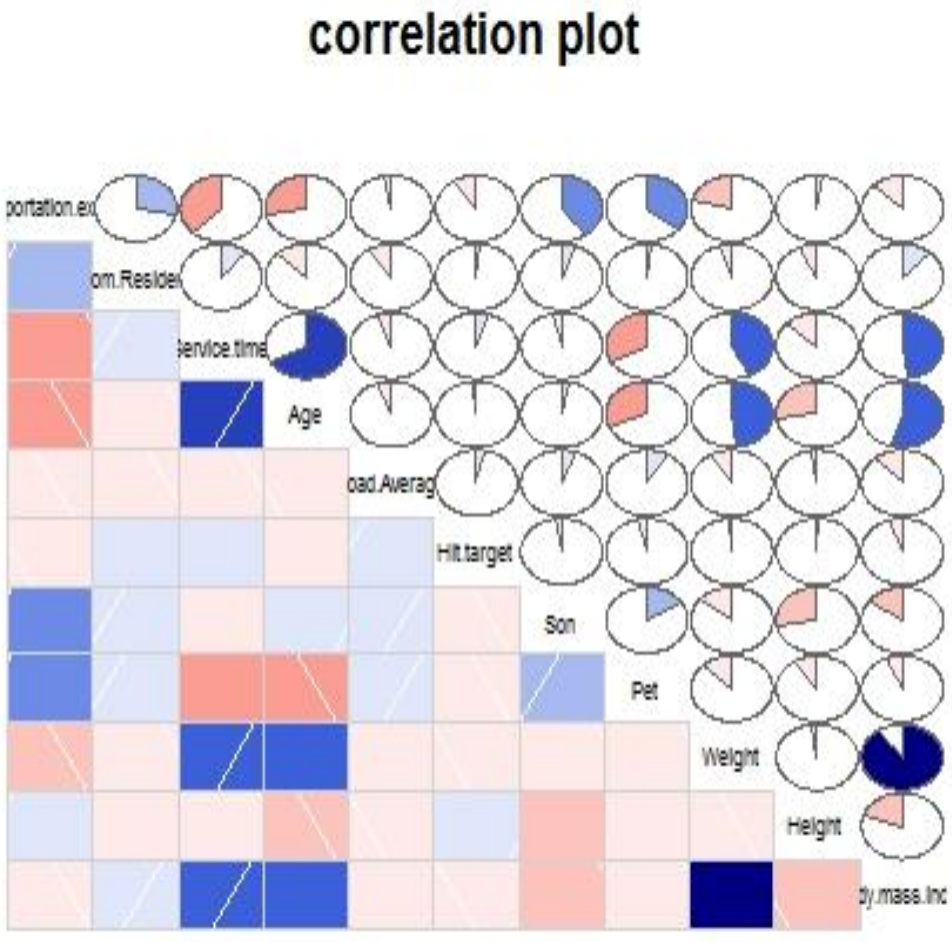
So, I only resolved the outliers of rest of the variables and not the target variable.

2.a.(iv) Feature selection

In this step of data pre-processing we will select the most relevant features from the dataset.

Here are two types of features available one is numerical and other one is categorical variables.

To get the relevant number of numerical features contributing to determine target variable I have used Corrgram plot of the variables.



The correlation plot shows how the variables are correlated. Here we can observe that Weight and BMI are highly positively correlated and thus they can incur Multicollinearity in the model if both the variables are feed into the model.

Thus it is required to remove one of the variable. I am deleting BMI here and keeping weight as relevant feature because weight is a basic property whereas BMI is a derived value based on weight and height.

Now we are done with the feature selection of numerical variable.

Coming to the selection of Categorical variable I have used ANOVA test over the features.

```
library("lsr")

anova_test = aov(Absenteeism.time.in.hours ~ ID + Day.of.the.week + Education + Social.smoker + Social.drinker + Reason.for.absence + Seasons + Month.of.absence + Disciplinary.failure, data = data)

summary(anova_test)
```

Anova uses one categorical and one numerical variable to calculate the relevancy of that particular variable.

Using the *pr* probability value generated by ANOVA test we can select those variables which are having p value less than 0.05.

Here one point must be considered. We have to do our calculation according to month in which employees are making absence.

Thus I am not deleting the Month of absence variable despite having p value greater than 0.05.

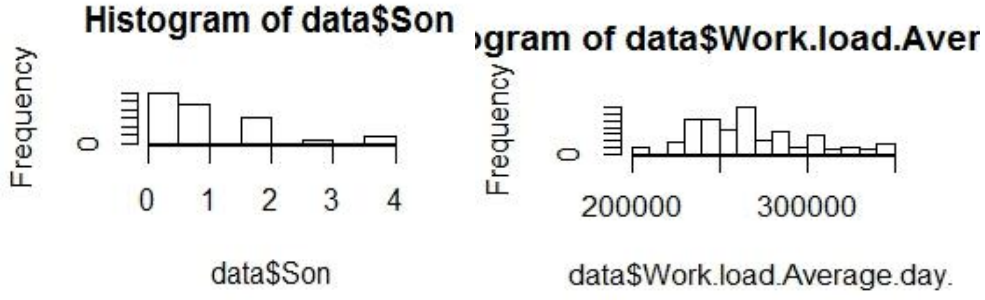
2.a.(v) Feature Scaling

The dataset which we have contains data of the variables having quite different range of values. For instance age lies between 30 to 60 whereas 280000 to 300000. These range values must be treated otherwise they would make the model bias and inefficient.

In order to scale the features there are usually two methods:-

1. **Standardisation**
2. **Normalization**

The method of standardization works only on the data which is normally distributed.
The Normalization method can work for any kind of distribution either positive skewed data or be it negative skewed data.



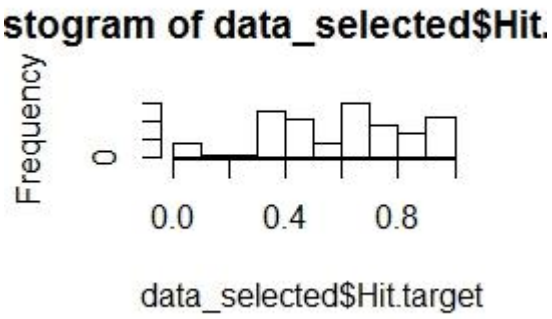


Fig. Histogram plot of some random variables

The above three randomly chosen variable's histogram plot is indicating that the data distributed is not normal.

In this case we have to use Normalization to scale our features.

```
num_names =
c("Transportation.expense","Distance.from.Residence.to.Work","Serv
ice.time","Age","Work.load.Average.day.","Hit.target","Son","Pet","
Weight","Height")


for (i in num_names) {
   print(i)
   data_selected[,i] = ((data_selected[,i] - min(data_selected[,i])) /
                (max(data_selected[,i]) - min(data_selected[,i])))

}
```

2.b Modeling

Once we are done with the data cleaning process now we are ready to apply various models that we have.

The given problem is of predictive analysis where the data to be predicted is a numerical value. These sorts of problem falls under domain of Regression.

Thus we will apply various Regression models available and will calculate the performance of each model using suitable Error metrics.

Before application of any model we need to divide our dataset into two parts:-

(i) Training data
(ii) Test data

Training data is the subset of whole population having 80% of the observation. It is used to train the model using the respective algorithm which the predictive modeling is using.

Test data is the dataset which we use to check our prediction and evaluate the model using error metrics to find the accuracy of the model.

R code

```
library("rpart")

train_index = sample(1:nrow(data_selected), 0.8*
nrow(data_selected))

train = data_selected[train_index,]

test = data_selected[-train_index,]
```

After applying Linear Regression, Random Forest and Decision Tree I found that Random Forest with number of trees = 100 is giving the best accuracy of the model.

library("randomForest")

RF_model = randomForest(Absenteeism.time.in.hours~. , train, ntree=100)

RF_prediction = predict(RF_model,test[,-14])

Model Evaluation

regr.eval(test[,13], RF_prediction, stats = 'rmse')

The error rate of the random Forest model is 8.78 which means the accuracy turn out to be

100-8.78 = 91.22%

Loss Prediction

I feel that loss is the total amount of unit of work which is not done or remains pending. This pending work might have been easily achieved if the employees were regular.

Formula that I used for loss calculation

Loss = absenteeism time * work load average/day