Employee Absenteeism

Rishabh Ahuja

24th June 2018

CHAPTER 1

INTRODUCTION

1.1 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?

1.2 DATA

Our task is to determine the target variable (Absenteeism time in hours) which is a numerical variable. Thus , we will apply different regression models and choose the best out of them.

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expens	Distance from Residence to Wor	Service time	Age	Work load Average/day
11	26	7	3	1	289	36	13	33	239,554
36	0	7	3	1	118	13	18	50	239,554
3	23	7	4	1	179	51	18	38	239,554
7	7	7	5	1	279	5	14	39	239,554
11	23	7	5	1	289	36	13	33	239,554
3	23	7	6	1	179	51	18	38	239,554
10	22	7	6	1		52	3	28	239,554
20	23	7	6	1	260	50	11	36	239,554
14	19	7	2	1	155	12	14	34	239,554
1	22	7	2	1	235	11	14	37	239,554
20	1	7	2	1	260	50	11	36	239,554
20	1	7	3	1	260	50	11	36	239,554
20	11	7	4	1	260	50	11	36	239,554
3	11	7	4	1	179	51	18	38	239,554
3	23	7	4	1	179	51	18	38	239,554

Hit target	Disciplina	Education	Son	Social drir	Social smc	Pet	Weight	Height	Body mass inde	Absenteeism time in hours
97	0	1	2	1	0	1	90	172	30	4
97	1	1	1	1	0	0	98	178	31	0
97	0	1	0	1	0	0	89	170	31	2
97	0	1	2	1	1	0	68	168	24	4
97	0	1	2	1	0	1	90	172	30	2
97	0	1	0	1	0	0	89	170	31	
97	0	1	1	1	0	4	80	172	27	8
97	0	1	4	1	0	0	65	168	23	4
97	0	1	2	1	0	0	95	196	25	40
97	0	3	1	0	0	1	88	172	29	8
97	0	1	4	1	0	0	65	168	23	8
97	0	1	4	1	0	0	65	168	23	8
97	0	1	4	1	0	0	65	168	23	8
97	0	1	0	1	0	0	89	170	31	1
97	0	1	0	1	0	0	89	170		4

Table 1.1 and 1.2 contain all the variables for the Absenteeism at work dataset.

CHAPTER 2

Methodology

2.1 Pre Processing

1) Sorting – For the project to be better comprehensive, the dataset is first sorted in the order of "ID" and then in the order of "Reason for absence" i:e 2 level sorting is done.

2) Conversion – The variables are converted into suitable data types i:e either categorical or numerical. Out of the 21 variables, 10 are taken as categorical and the rest 11 as numerical.

Numerical Variables	ID	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Weight	Height	Body mass index	Absenteeism time in hours
Categorical Variables	Reason for absence	Month of absence	Day of the week	Seasons	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	------

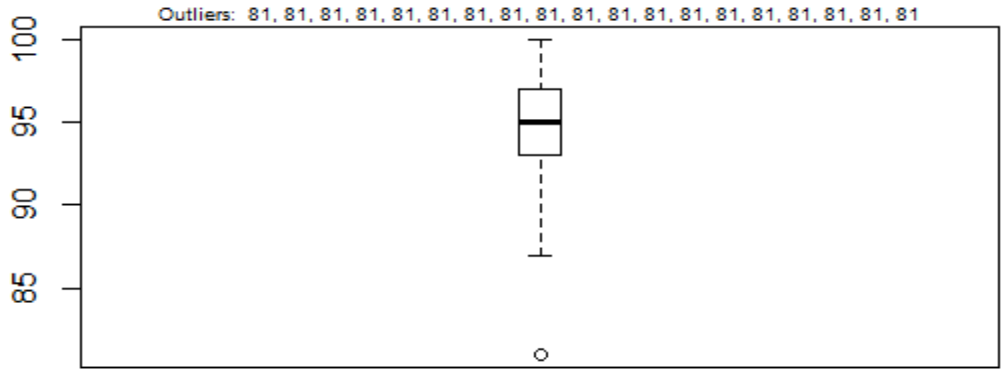
2) Missing Value Analysis – Next step is to impute the missing values. Out of the 3 ways of mean, median and KNN imputation, the KNN method is the most accurate. Thus it is adopted.

2.2 Outlier Analysis

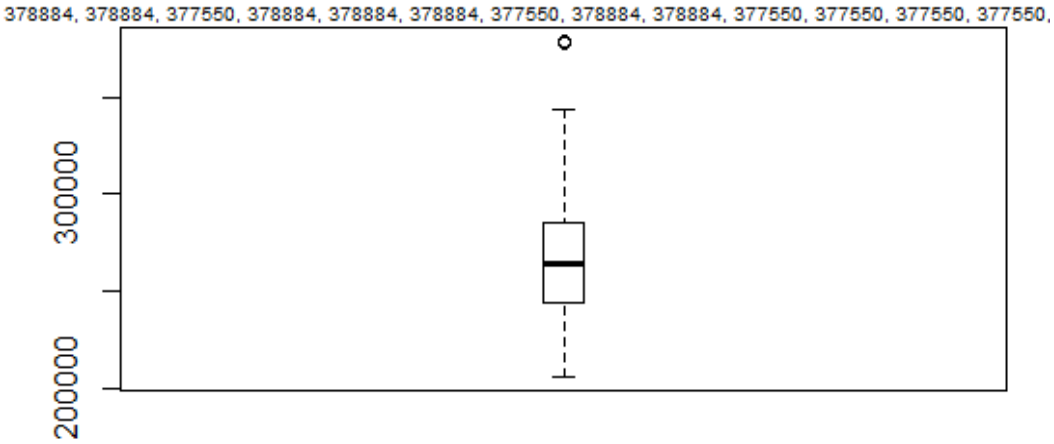
Outlier analysis is used to remove the messy data points which are beyond the maximum and the minimum values of the variable. For determining the outliers, boxplots come in handy. On analysis the given dataset, the outlier analysis should be applied to 3 numerical variables which are "Work load Average/day", "Hit target" and "Absenteeism time in hours". Rest all numerical variables have uniform values according to the "ID" and thus any outlier in those numerical variables would occur merely due to the different frequencies of different ID numbers. On the other hand, the 3 variables listed above hold different values according to different IDs.

Once the outliers in the 3 variables are determined, they are then replaced with NA. Then those NA values are imputed using KNN imputation.

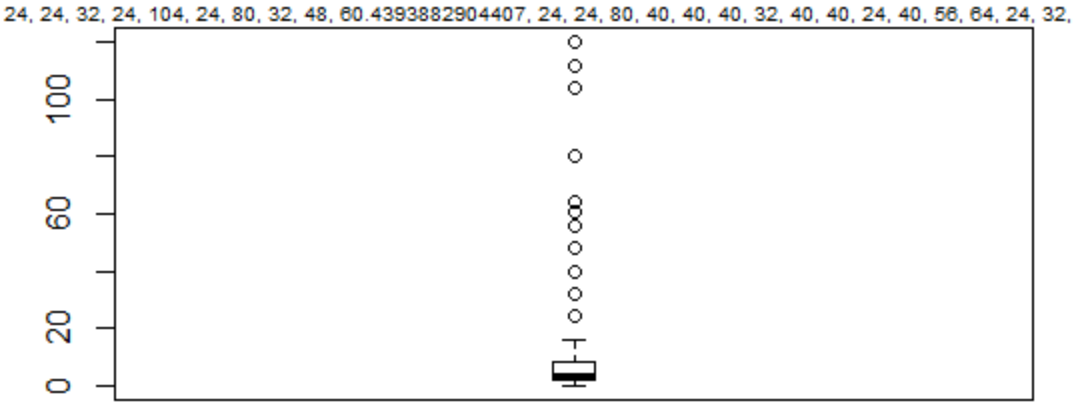
Hit Target



Work load Average/day

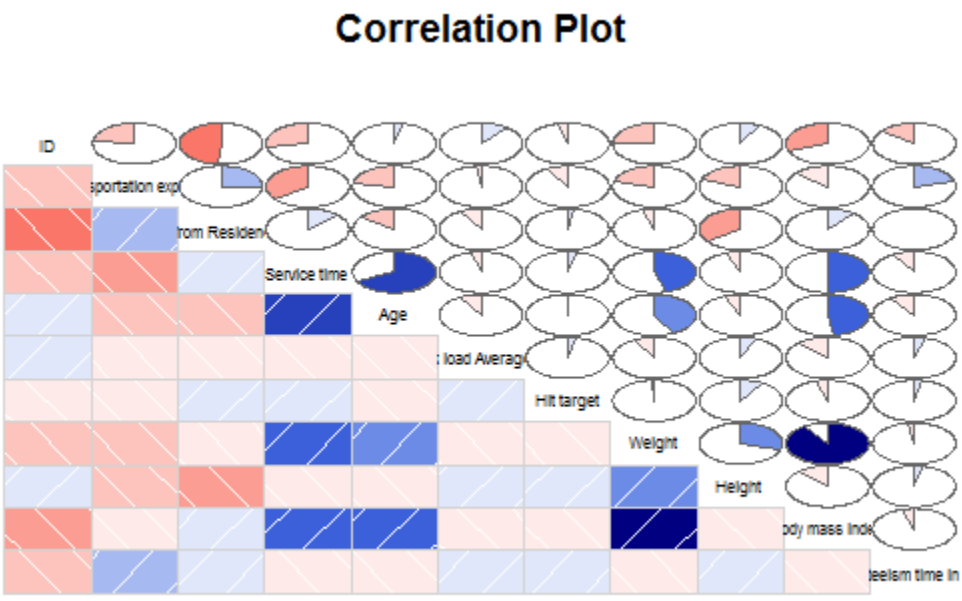


Absenteeism time in hours



2.3 Feature Selection

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a probability that many variables in our analysis are not important at all. We have used Correlation plots to eliminate the features.



The “weight” and “Body mass Index” variables have high correlation. Thus, out of them, “Body mass Index” is removed.

2.4 Feature Scaling

Feature scaling is implemented for all the numerical variables to be in the same range and thus the magnitude of any particular variable doesn’t hamper the model results. Feature scaling in the code is done on all numerical variables except “ID” as it is used as a reference variable.

CHAPTER 3

Modeling

I tried using decision tree and regression models. Both the models gave fairly high accuracy(>85%) . So, on the basis of adjusted R-square value linear regression method was used after removing multicollinear elements with VIF >8.

On the basis of the model developed, following are the answers of the problems given in the dataset :-

1) Company needs to make several changes to bring down the absenteeism cases. These changes are made on the basis of the degree of dependence of the absenteeism variable on other variables. Some of them include :-
    a) ID – It has high dependence on absenteeism. Thus, IDs with high frequency of leaves need to be removed from thee organization. The threshold for same is chosen as 40days. Thus, on that basis ID numbers 3,20,22,28,34 need to be removed.
    b) Reasons for absence – Some reasons impact the absenteeism highly while others to moderate level. For e:g, absenteeism rate shows high dependence on Reason 9 (circulatory system) and thus these need to be taken care of by providing proper healthcare to the employees. Other such reason of absences are Reason 19 and Reason 22.
    c) Distance from Residence to Work – It also affects the absenteeism rate highly. The threshold for the same is taken as 25km and thus for the employees having distance from work >25km suitable arrangements need to be made.
    d) Apart from them Service time and Age are also important factors but they have been removed due to multicollinearity issues.
2) For predicting the absenteeism time every month, the model was sorted according to the “month of absence” variable. Then the sum for absenteeism time for that month was calculated. As, the dataset is normalized, it was then converted back to it’s original range(0,16), by multiplying the sum of normalized hours by 16.

Month of absence	Absenteeism time in hours
1	193.04
2	258.83
3	402.16
4	231.89
5	283.77
6	244.83
7	379.4
8	232.03
9	181.15
10	287.76
11	280.8
12	208.59