Employee Absenteeism

Akshay Kumar Chaturvedi

5 July 2018

Contents

Chapter 1

Introduction

1.1 Problem Statement

Courier is used by almost everyone nowadays. The courier companies heavily rely on human capital for collection, transportation and delivery. A courier company is facing an issue of employee absenteeism, this project deals with predicting employee absenteeism so that company can take measures to reduce the employee absenteeism.

1.2 Data

Data was divided in two sets, 'train' set and 'test' set, 'train' set was used to train the models whereas 'test' set was used to test the models. There are total 21 columns in the data; dataset contains 740 observations. Please find below a sample of data in tables 1.1, 1.2 and 1.3.

Table 1.1: Sample Data (Columns: 1-7)

ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work
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

Table 1.2: Sample Data (Columns: 8-14)

Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Education	Son
13	33	239,554	97	0	1	2
18	50	239,554	97	1	1	1
18	38	239,554	97	0	1	0
14	39	239,554	97	0	1	2
13	33	239,554	97	0	1	2

Table 1.3: Sample Data (Columns: 15-21)

Social drinker	Social smoker	Pet	Weight	Height	Body mass index	Absenteeism time in hours
1	0	1	90	172	30	4
1	0	0	98	178	31	0
1	0	0	89	170	31	2
1	1	0	68	168	24	4
1	0	1	90	172	30	2

Out of 21 variables (or columns), 9 are categorical namely 'ID', 'Reason for absence', 'Month of absence', 'Day of the week', 'Seasons', 'Disciplinary failure', 'Education', 'Social drinker' and 'Social smoker'. 'Absenteeism time in hours' is the target variable, it is a numerical variable so our problem is a regression problem. Apart from these 10 variables, there are 11 numerical variables; their correlation plot can be seen below in figure 1.1.

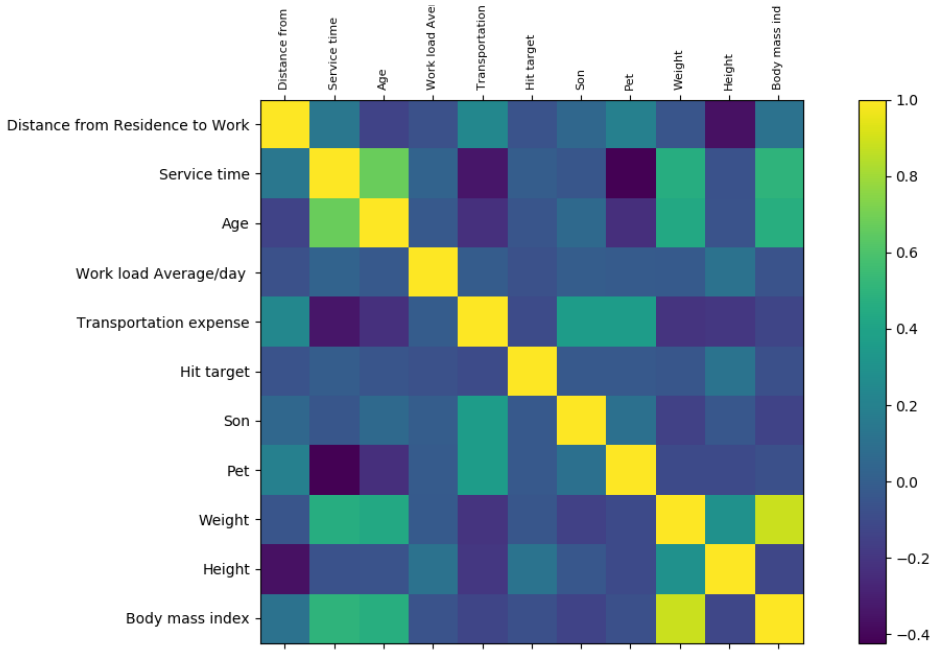

Figure 1.1: Correlation Plot of Numerical Variables

From the correlation plot in figure 1.1, we can observe that 'Weight' and 'Body mass index' are pretty strongly correlated, 'Age' and 'Service time' are also correlated, rest are very weakly correlated.

Distribution of these numerical variables can be seen below in figure 1.2.

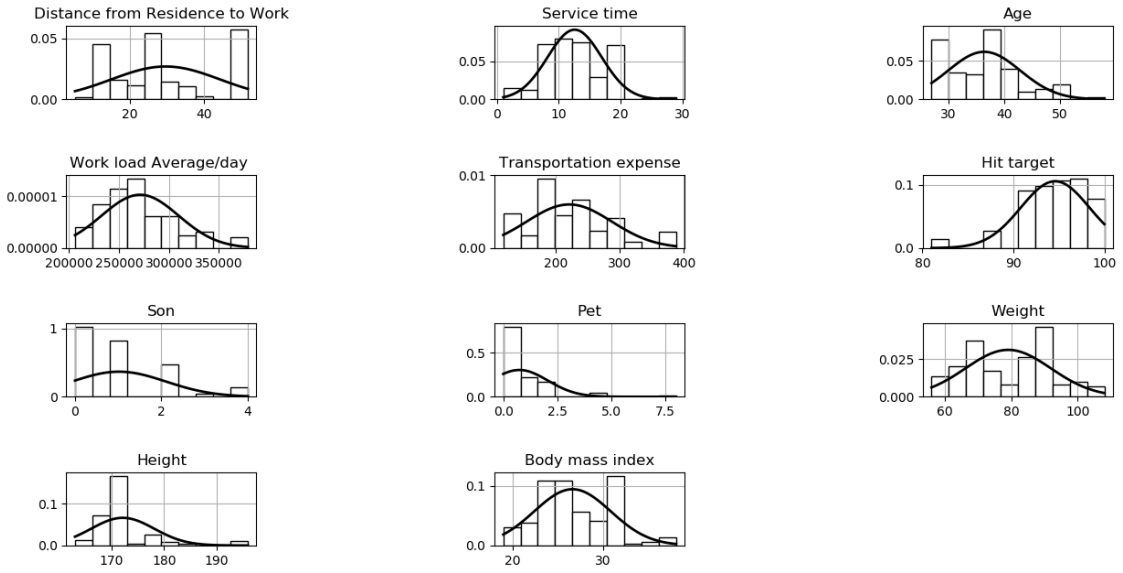

Figure 1.2: Distribution of Numerical Variables

From figure 1.2 we can observe that none of the numerical variables follow a normal distribution. Boxplots were also created for these variables to visualize the outliers if any, they can be seen below in figure 1.3.

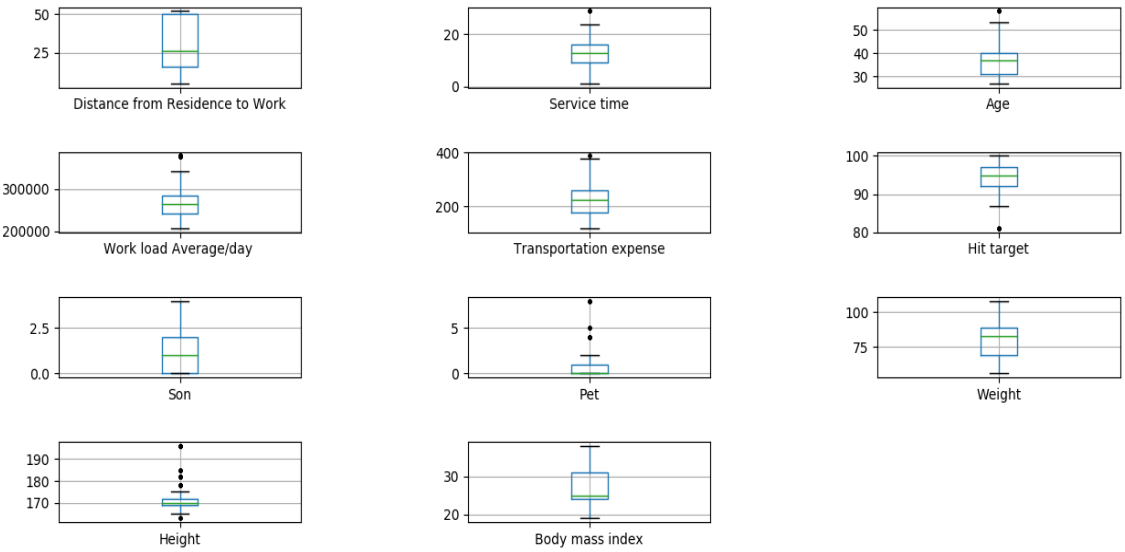

Figure 1.3: Boxplots of Numerical Variables

Boxplots show us that seven numerical variables contain outliers, but they are very few.

From above information we can conclude that:

- We do not have highly correlated variables so it cannot help us in feature selection.
- As none of our numerical variables have approximately normal distribution, so we cannot standardize them to have mean 0 and standard deviation 1.
- Our data has very few outliers, so we can ignore them.

Chapter 2

Methodology

2.1 Pre-processing

We have categorical variables with more than two levels, so we have created dummy variables for all of them.

We have continued to build our models with all the numerical variables as only two variables are highly correlated as can be seen in figure 1.1.

2.2 Modeling

Data was divided into 80:20 ratio, 80% of data was used as 'train' set and rest of the 20% was used as 'test' set. All the models were trained on 'train' set and tested on 'test' set. Five models were used in experiments namely 'Linear Regression'-LR, 'Support Vector Machine'-SVM, 'Gradient Boosted Tree'-GBT, 'Decision Tree'-DT and 'Random Forest'-RF. For evaluation, root mean square error (RMSE) has been used. Performance metrics below are from Python programming language.

Below table 2.1 contains results.

Table 2.1: Results

Model	Train_RMSE	Test_RMSE
LR	1.24e-11	1.44e-11
SVM	0.099	0.2518
GBT	0.00029	0.000358
DT	0	0
RF	0	0

Some key points on models used in the experiments above:

- Support Vector Machine has been used with radial basis function as kernel, it was chosen as linear kernel performed very poorly.
- Random Forest uses 500 trees.
- Gradient Boosted Tree uses 100 trees and maximum depth of tree is 2.

Chapter 3

Conclusion

In conclusion, out of all the models shown here and others not shown here but tried Decision Tree(DT) emerges as the best model according to performance metrics and also as it takes less time than Random Forest(RF), giving an overall test set error of 0.