

Employee Absenteeism

Pranav Tewari | 24 February 2019

Chapter 1

Introduction

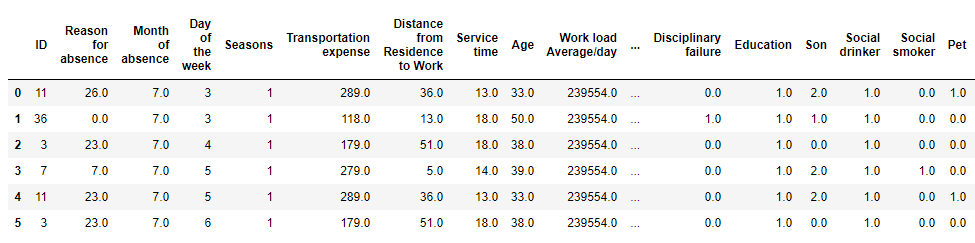
* 1. Problem Statement

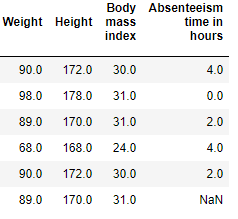
XYZ is a courier company. As we appreciate that human capital plays an important role in collection, transportation and delivery. The company is passing through genuine issue of Absenteeism. The company needs to know the answer of following questions:

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. Data

Below is the sample of data set which we are using to predict:

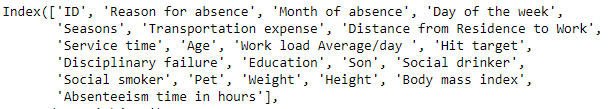
Table 1.1: Sample Data (Column: 1-6)





The table below shows the columns of the data set using which we have to predict the values.

Table 1.2: Columns of Data



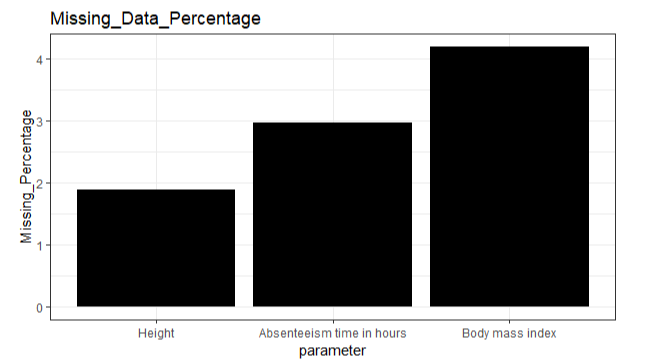
Chapter 2

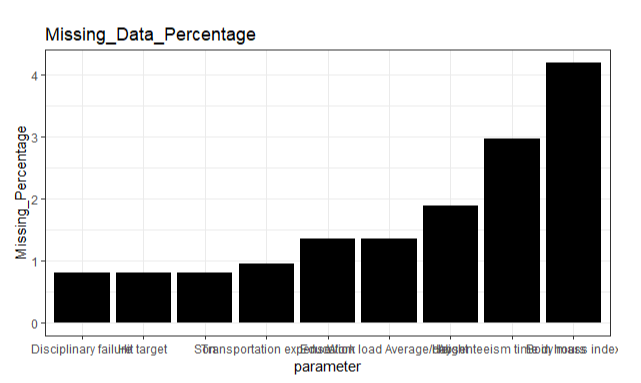
Methodology

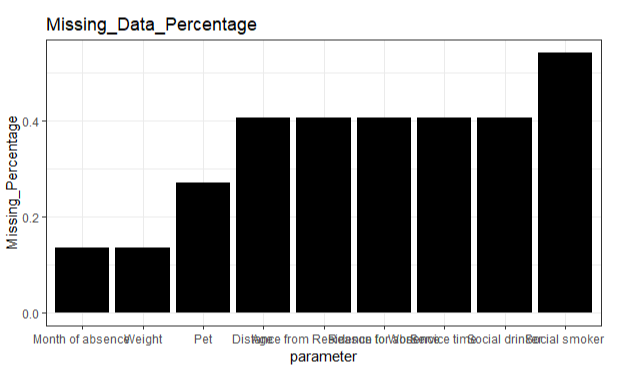
* 1. Pre Processing

Any predictive modelling requires that we look at the data before we start modelling. However, in data mining terms looking at data refers to so much more than looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process we will first try to look at all the probability distributions of the variables. Most analysis like regression, requires data to be normally distributed. We can visualize that in a glance by looking at the probability distribution or probability density functions of the variable.

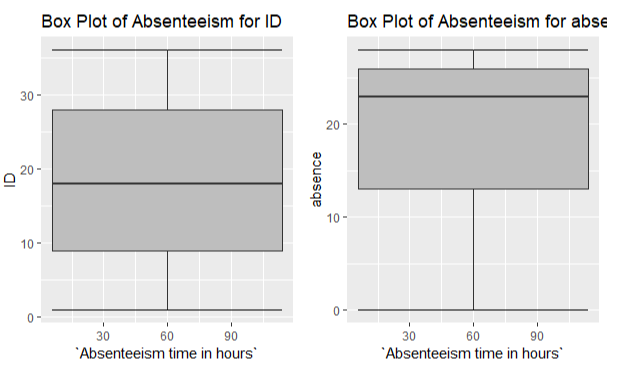
Below figures are Bar Plot of the Missing Values in the Data set.

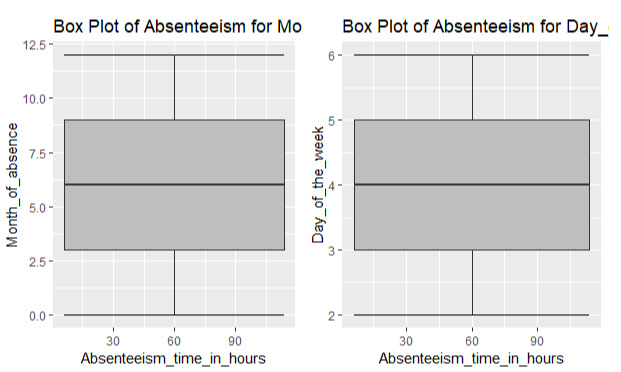


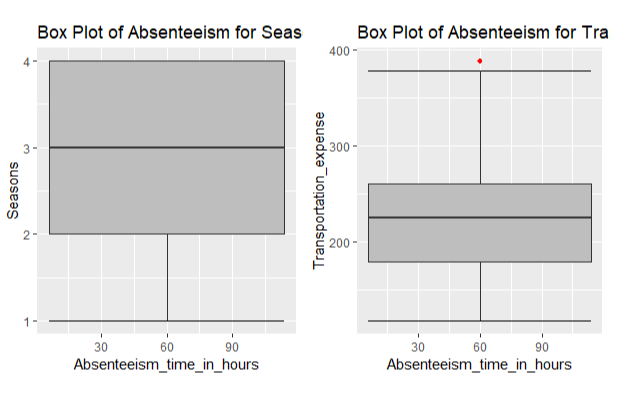


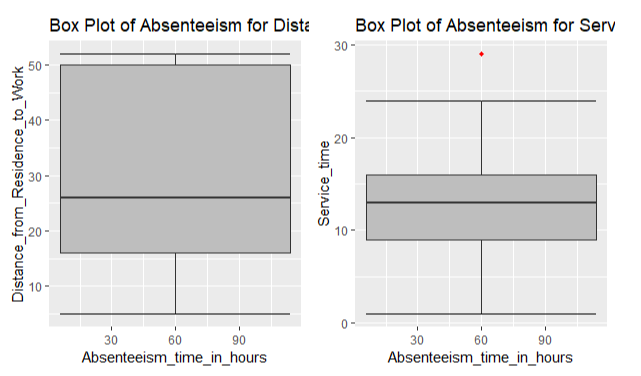


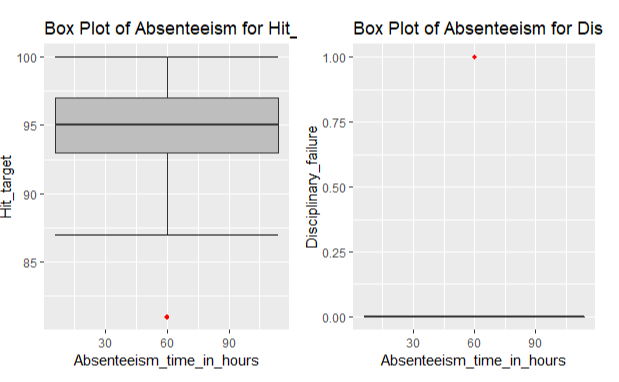
Below figures are box Plot of the Variables against the count.

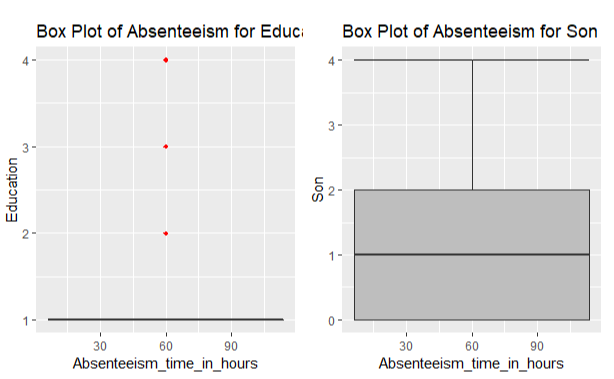


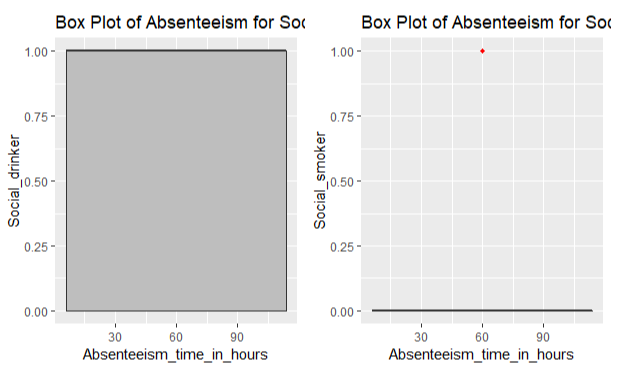


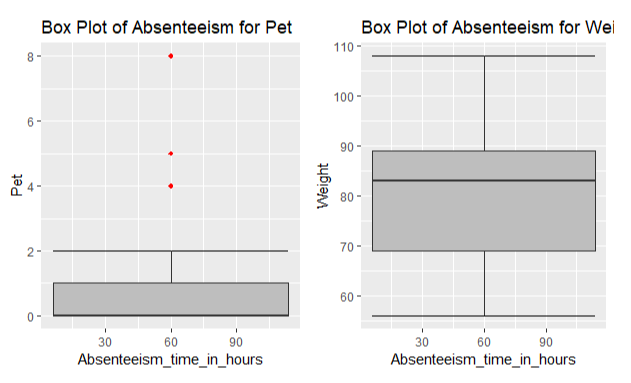


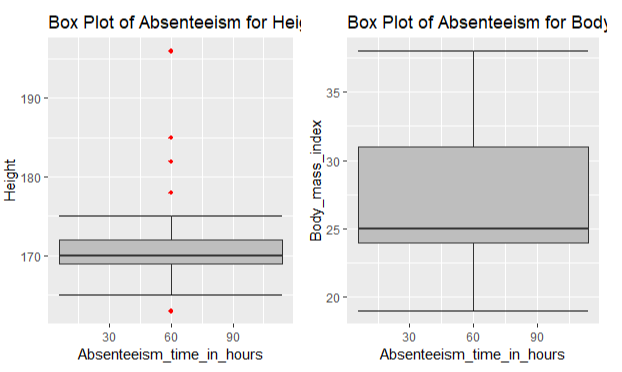


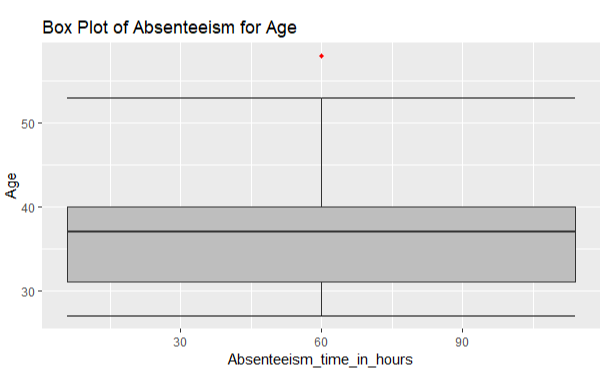






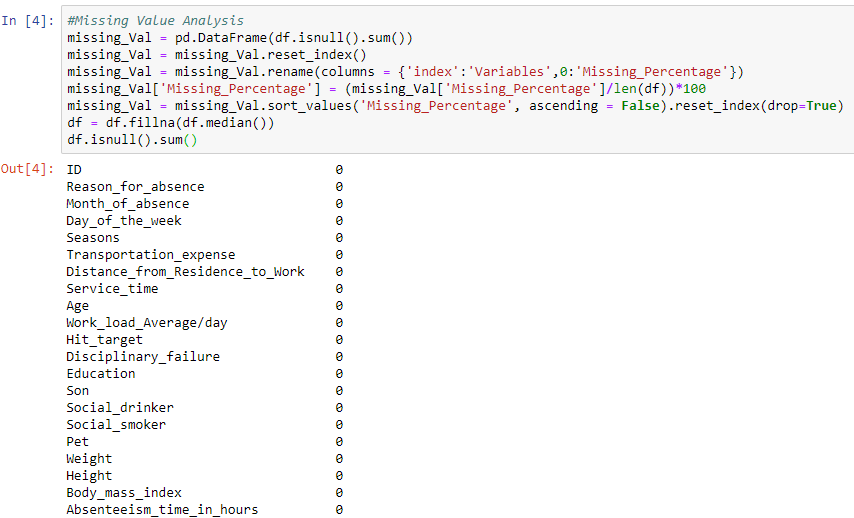






* + 1. Null Value Analysis

It is one of the Pre Processing steps in which the missing values in the variables are imputed with either of the methods i.e., (Mean, Median, Mode, KNN). The values are imputed only if they are missing up to 30% of the total data in the variable else it is left out as the variable does not carry meaningful details to predict the outcome.

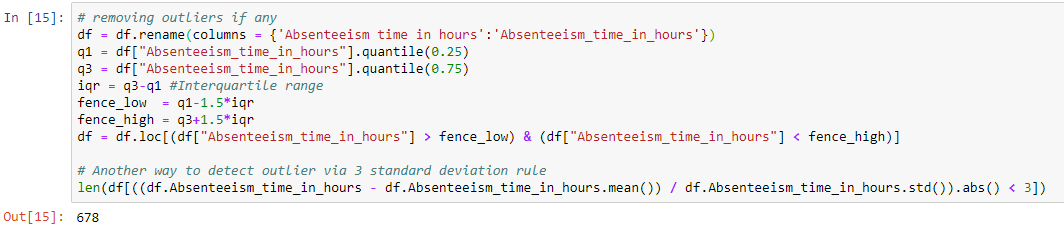


* + 1. Outlier Analysis

One of the step of Pre Processing is outlier analysis where we can see the variables having values which are varying from other values very widely.

These values are those which can be rejected or substituted using KNN.

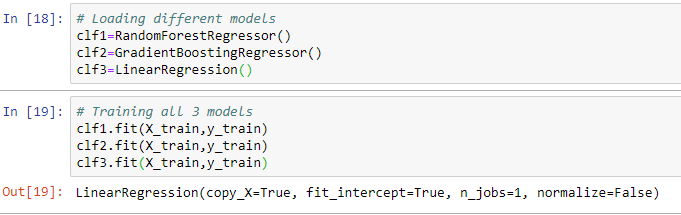
But herein case we have removed the outlier data from our data set to make it normally distributed.



* 1. Modeling
     1. Model selection

Model selection can depend upon the independent variable, in our case its **Absenteeism Hours.** Here we can try and test multiple models and the one which best suits or gives the best result on train model.

You can start your model with the most simplest to the most complex and can try all these models simultaneously.



If the dependent variable is nominal the only predictive analysis that we can perform is **Classification,** and if the dependent variable is interval or ratio, the normal way is to do **regression** analysis, or classification after binning, but here in our case we do both classification and Linear Regression.

* + 1. Classification

Classification for the prediction of count cannot be done as count of anything is not a categorical value. Though the dependent variables are categorical but predicting count may not give the result expected and may have to categorize the count as well depending upon the dependent variables selected which will lead to a very hectic task.

Chapter 3

Conclusion

* 1. Model Evaluation

Now that we have few models for predicting the target variable, we need to decide which one to choose.

There are several criteria that exits for evaluating and comparing models. We can compare the models using any of the following criteria:

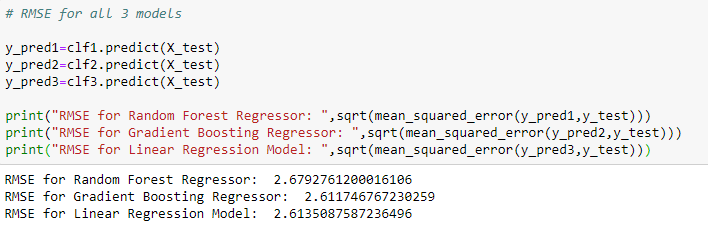
1. Predictive Performance
2. Interpretability
3. Computational Efficiency

In our case i.e., predicting Absenteeism Hours, the latter 2, Interpretability and computational Efficiency, do not hold much significance. Therefore we will use predictive performance as criteria to compare and evaluate models.

Predictive performance can be measured by comparing predictions of the models with real values of the target variables, and calculating some average error measures.

* + 1. Root Mean Square Error (RMSE)

RMSE is one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our model.



* 1. Model Selection

We can see that Gradient Boosting is comparatively performing better than linear regression and Random Forrest hence we will use that model to avoid any loss of information.

* 1. Answer to Problem Statement

1. Here we have found that the maximum number of Absenteeism almost every month in the company is because of **Blood Donation by employees is the reason which leads to approx. 119 hours of Absenteeism yearly followed by Dental Consultation and Laboratory Examination leading to approx. 42 and 43 hours of Absenteeism** and so on.

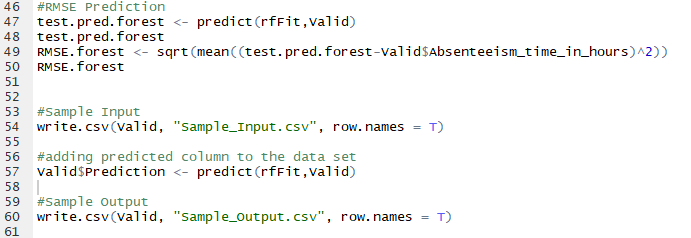
To Avoid Absenteeism due to such reasons **company should arrange for a Medical camp at their location** so that employees can donate blood at one place and at one time not leading to Absenteeism every month.

Apart from above the **medical camp should include ENT and Physio Therapy as well** and the medical camp **should be arranged on a non-working day i.e., Sunday and should be for Employees Family members as well.** This would reduce the hours of Absenteeism drastically.

1. If the same trend of Absenteeism Follows next year as well the company is predicted to loose almost **1080 targets hits** will be missed every month.

Appendix A – R code





Appendix B – Python Code

