Employee Absenteeism

Apurv Kumar

04/08/2018

Chapter 1

Introduction

1.1 Problem Statement

According to the Oxford Dictionary, Absenteeism refers to the “act of staying away from the place of work without any good reason”. The issue of absenteeism is a critical problem for any organisation. For an organisation, big or small, it is essentially the people or the manpower that runs the company, they are the driving force of the company. Every employee working together as part of a bigger organisation, like cogs of a machine.
Every case of absenteeism leads to some loss for the company, because an employee absent for a day leads to some capital going to waste for that day.
Therefore if a company has a high case of absenteeism, it needs to acquire the reason for the same, so that it could change its policies or rules to better accumulate its employees, and ultimately bring down the rate of absenteeism.

In this project, we try to look at the features or reasons which lead to an organisation having an issue of absenteeism, and try to give solutions to the problems.

1.2 Data

Our task is to build a regression model to predict the case of absenteeism, and also to figure out the reason for the same, and try to give some deductions or solutions to solve the issue.
Given below is a sample of the data that we are using to create our regression model:

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
11	26.0	7.0	3	1	289.0	36.0	13.0	33.0	239,554
36	0.0	7.0	3	1	118.0	13.0	18.0	50.0	239,554
3	23.0	7.0	4	1	179.0	51.0	18.0	38.0	239,554
7	7.0	7.0	5	1	279.0	5.0	14.0	39.0	239,554
11	23.0	7.0	5	1	289.0	36.0	13.0	33.0	239,554

Table 1.1: Absenteeism Data (Columns: 1-10)

Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker
97.0	0.0	1.0	2.0	1.0	0.0
97.0	1.0	1.0	1.0	1.0	0.0
97.0	0.0	1.0	0.0	1.0	0.0
97.0	0.0	1.0	2.0	1.0	1.0
97.0	0.0	1.0	2.0	1.0	0.0

Table 1.2: Absenteeism Data (Columns: 11-16)

Pet	Weight	Height	Body mass index	Absenteeism time in hours
1.0	90.0	172.0	30.0	4.0
0.0	98.0	178.0	31.0	0.0
0.0	89.0	170.0	31.0	2.0
0.0	68.0	168.0	24.0	4.0
1.0	90.0	172.0	30.0	2.0

Table 1.3: Absenteeism Data (Columns: 17-21)

As you can see in the table below, we have 20 predictor variables, using which we have to make our analysis:

S. No	**Predictor**
1	ID
2	Reason for absence
3	Month of absence
4	Day of the week
5	Seasons
6	Transportation expense
7	Distance from Residence to Work
8	Service time
9	Age
10	Work load Average/day
11	Hit target
12	Disciplinary failure
13	Education
14	Son
15	Social drinker
16	Social smoker
17	Pet
18	Weight
19	Height
20	Body mass index

Table 1.4: Predictor variables

1.3 Data Description

Let us elaborate the predictors and the target variable to give a better description about them:

1. Individual identification (ID)
2. Reason for absence (ICD). Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:
   - 2.1. Certain infectious and parasitic diseases
   - 2.2. Neoplasms
   - 2.3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
   - 2.4. Endocrine, nutritional and metabolic diseases
   - 2.5. Mental and behavioural disorders
   - 2.6. Diseases of the nervous system
   - 2.7. Diseases of the eye and adnexa
   - 2.8. Diseases of the ear and mastoid process
   - 2.9. Diseases of the circulatory system
   - 2.10. Diseases of the respiratory system
   - 2.11. Diseases of the digestive system
   - 2.12. Diseases of the skin and subcutaneous tissue
   - 2.13. Diseases of the musculoskeletal system and connective tissue
   - 2.14. Diseases of the genitourinary system
   - 2.15. Pregnancy, childbirth and the puerperium
   - 2.16. Certain conditions originating in the perinatal period
   - 2.17. Congenital malformations, deformations and chromosomal abnormalities
   - 2.18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
   - 2.19. Injury, poisoning and certain other consequences of external causes
   - 2.20. External causes of morbidity and mortality
   - 2.21. Factors influencing health status and contact with health services.

   And 7 categories without (CID)
   - 2.22. patient follow-up (22),
   - 2.23. medical consultation (23),
   - 2.24. blood donation (24),
   - 2.25. laboratory examination (25),
   - 2.26. unjustified absence (26),
   - 2.27. physiotherapy (27),
   - 2.28. dental consultation (28).
3. Month of absence
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons (summer (1), autumn (2), winter (3), spring (4))
6. Transportation expense
7. Distance from Residence to Work (kilometers)
8. Service time
9. Age
10. Work load Average/day
11. Hit target
12. Disciplinary failure (yes=1; no=0)
13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
14. Son (number of children)
15. Social drinker (yes=1; no=0)
16. Social smoker (yes=1; no=0)
17. Pet (number of pet)
18. Weight
19. Height
20. Body mass index
21. Absenteeism time in hours (target)

Chapter 2

Methodology

2.1 Pre Processing

Pre processing is an essential part of any data analysis process. We explore the data, plot graphs to see its distribution and perform data cleaning and pruning in order to convert the input dataset into a proper format to be able to correctly do our analysis.

A dataset can have various issues like missing values, exceptional data, incorrect columns name, invalid characters, etc. Therefore it is essential that we perform the step of data pre processing. Let us start:

2.1.1 Missing Value Analysis

A dataset in real world would always have some values missing for some of the cells. It is our role to be able to efficiently handle it. A model trained before handling the missing values would yield incorrect predictions.

In our analysis I look into 3 ways to handle the missing values:

1. Imputation by mean: We replace every missing value by the mean of the corresponding column.
2. Imputation by median: We replace every missing value by the median of the corresponding column.
3. Imputation by KNN: We use the K Nearest Neighbour algorithm to impute the missing values.

Let us look at the missing values list for our dataset:

	Missing value count
ID	0
Reason for absence	3
Month of absence	1
Day of the week	0
Seasons	0
Transportation expense	7
Distance from Residence to Work	3
Service time	3
Age	3
Workload	10
Hit target	6
Disciplinary failure	6
Education	10
Son	6
Social drinker	3
Social smoker	4
Pet	2
Weight	1
Height	14
Body mass index	31
Absenteeism time in hours	22

Table 2.1: Missing value count

We can get a better understanding of the count of missing values if we plot the missing percentage, as shown in the table below.

	Predictors	Missing_percentage
0	Body mass index	4.189189
1	Absenteeism time in hours	2.972973
2	Height	1.891892
3	Workload	1.351351
4	Education	1.351351
5	Transportation expense	0.945946
6	Son	0.810811
7	Disciplinary failure	0.810811
8	Hit target	0.810811
9	Social smoker	0.540541
10	Age	0.405405
11	Reason for absence	0.405405
12	Service time	0.405405
13	Distance from Residence to Work	0.405405
14	Social drinker	0.405405
15	Pet	0.270270
16	Weight	0.135135
17	Month of absence	0.135135
18	Seasons	0.000000
19	Day of the week	0.000000
20	ID	0.000000

Table 2.2: Missing value percentage

We have already stated the 3 ways to handle missing values. For each of the three ways we perform the following steps to select the best among the three:

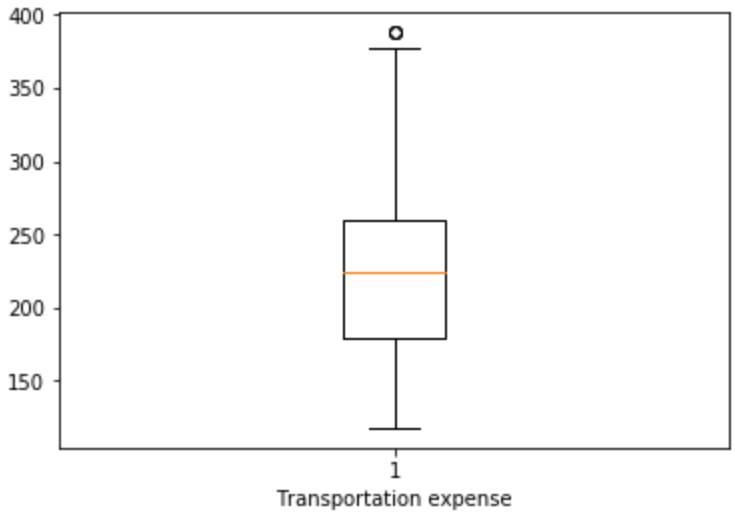
1. We replace one non-missing numerical value with NaN.
2. We perform the imputation method (mean, median or KNN).
3. Through step 2, we decide upon which methods imputed the NaN value with a value which is closest to the original value, and then use that method as default.

In our analysis, we find that the KNN imputation generates best results among the three. Therefore we apply KNN imputation on the entire dataset.

2.1.2 Outlier Analysis

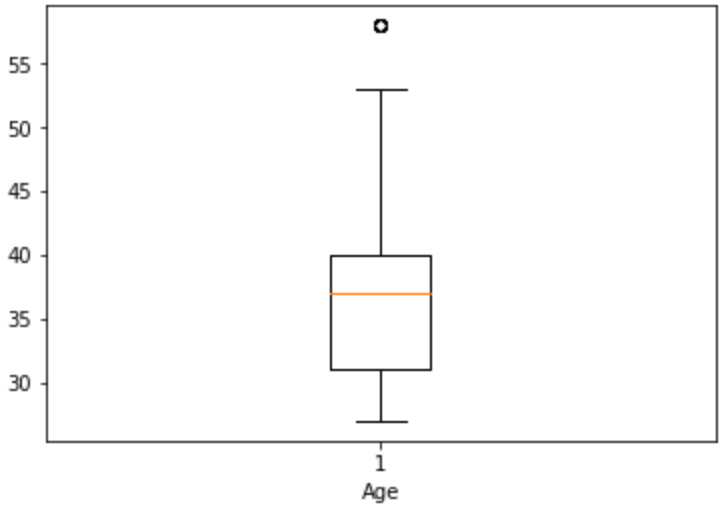
Once the missing value analysis is complete, we move to the outlier analysis.
Outlier refers to data points which have very extreme values and do not usually denote the general population. Failing to handle the outliers will result in our model producing incorrect and biased predictions.

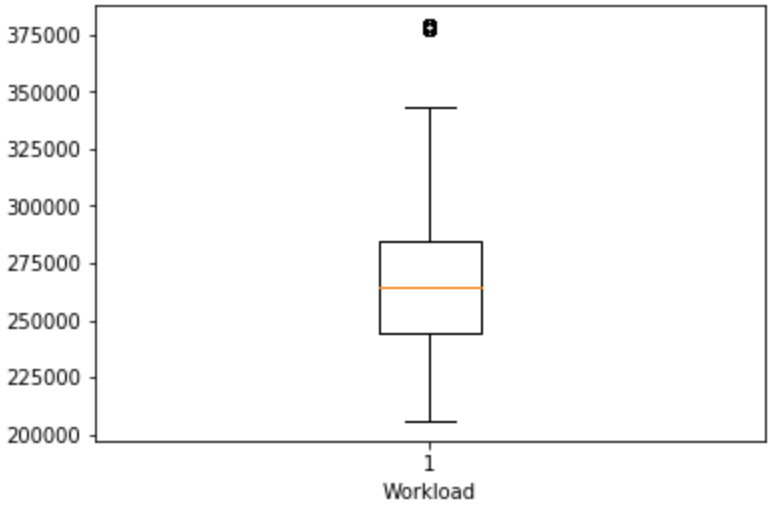
One of the efficient ways to visualise the presence of outliers in any dataset is by using the Boxplot method to plot the outliers. Boxplot works for numerical data points, not categorical. Shown below is the plot for each of the numerical columns:

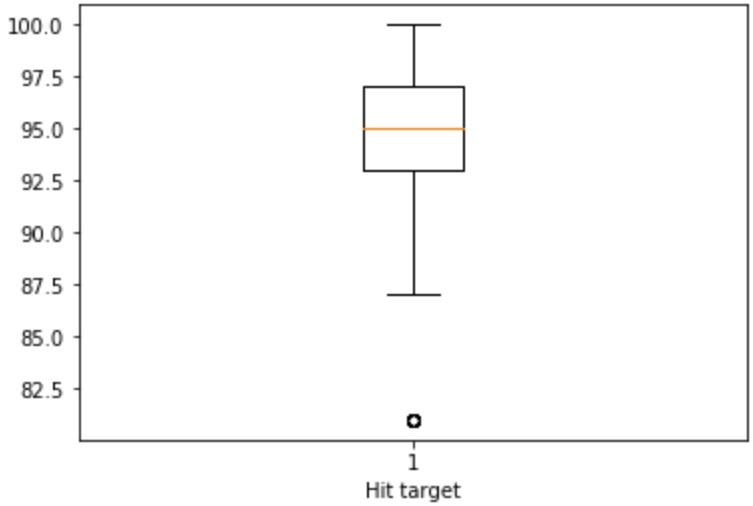


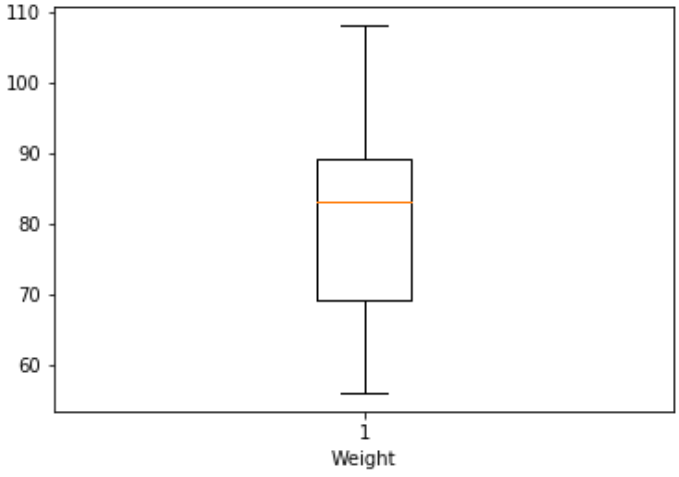

50
40
30
20
10
1
Distance from Residence to Work

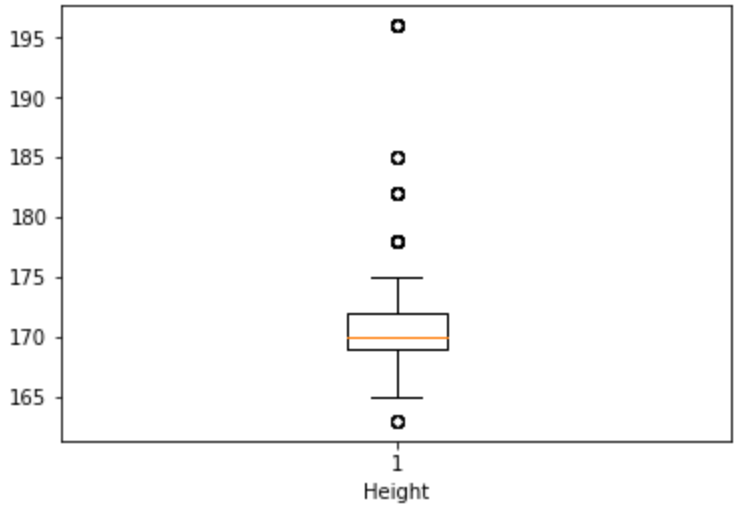

30
25
20
15
10
5
0
1
Service time

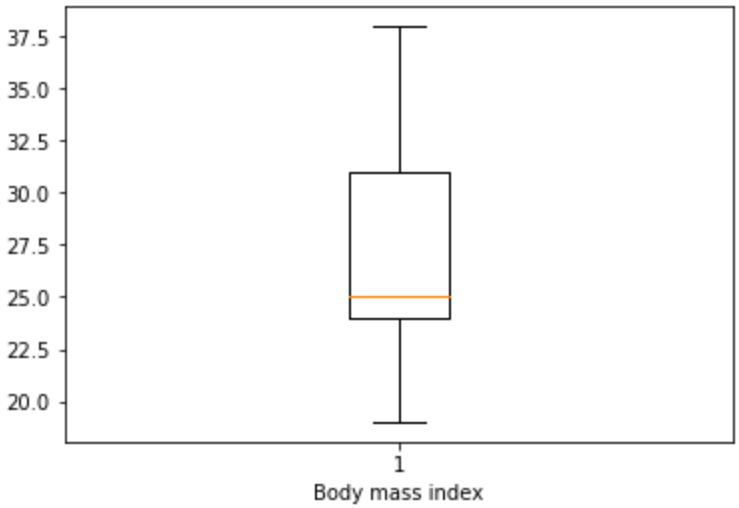

55
50
45
40
35
30
1
Age


375000
350000
325000
300000
275000
250000
225000
200000
1
Workload


100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
1
Hit target


110
100
90
80
70
60
1
Weight





Looking at the above plots, we can efficiently visualise that some of the predictors have outlier points present in them. We drop such points so that our dataset now reflects a generalized dataset, which would prove efficient in training our model.

2.1.3 Feature Selection

Our dataset has 20 predictors, and it is possible that not all of the predictors are necessary for our regression model. Training a model with 20 models might make the model a little bit complex, and could make the model overfit. Therefore we shall make some analysis, and choose only those features which are required.

The parameters to choose the importance of a feature is to calculate the correlation among all the numerical predictors. Suppose A and B have high correlation among them, then it means that they are producing the same information, therefore we can discard one of them.

One of the convenient ways to visualize correlation is to plot a heatmap, which makes it easy to see which predictors are depended on each other, and therefore we can remove the redundant predictors not required for our model. Find below the heatmap:

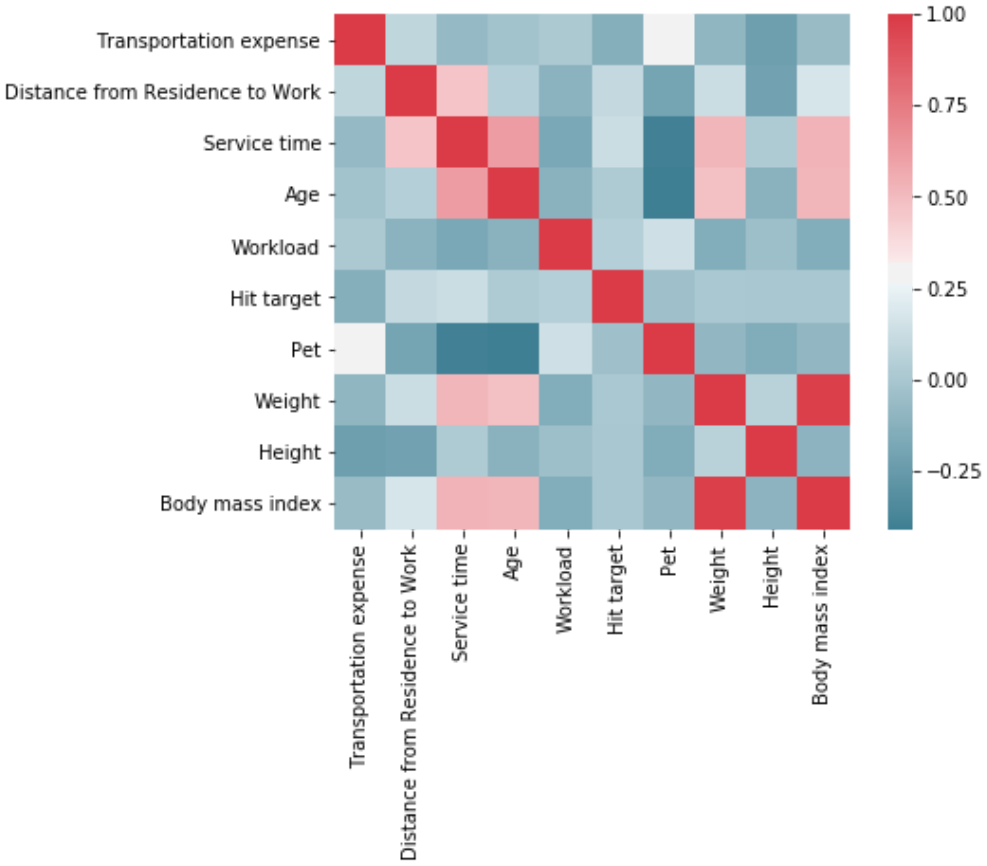


Table 2.3: Heatmap to visualize correlation among numerical predictors

From the above heatmap, we can clearly visualize that predictors like service time, age and weight have high correlation with other predictors, therefore we remove them from our selection of predictors as they contribute redundant information.

Just like heatmap works for the numerical predictors, we use a chi-square test for categorical problems to find out the importance of the categorical predictors with respect to the target variable.

Predictors	**Chi square test value**
Reason for absence	1.2415113115449792e-62
Month of absence	0.0001398127694953616 7
Day of the week	0.1165073435130744
Seasons	7.905112809814826e-05
Disciplinary failure	1.362786558952351e-64
Education	1.0
Son	4.0026627157919955e-06
Social drinker	0.5105602652535947
Social smoker	0.06889337951144041
Pet	0.04882036309194913

Table 2.4: P value between the categorical and the target variable

We choose a null hypothesis : The 2 categorical variables are independent.
Chi square test returns a p value. If the p value is more than 0.05, we reject the null hypothesis (and therefore we drop that predictor), otherwise we accept it.
Based on the above observations we drop the following observations: Day of week, Education, Social smoker, Social Drinker, Pet and ID (we drop ID on the basis that it definitely has no relation with any predictions, it is just a unique number)

From the feature selection step, we end up with a concise list of features which might play an important role in the regression model. Let us move to the next step of feature scaling.

2.1.4 Feature Scaling

Each of the features that we have selected till now have a different scale of values, and just feeding them directly into a model will result in biased output being generated. Therefore feature scaling refers to the process of converting the range of the values if the features into a concise range.
We have used the process of normalization on the values to get them in the range of 0-1.

FInd below a snapshot of the values:

	Transportation expense	Distance from Residence to Work	Workload	Hit target	Height	Body mass index
0	0.657692	0.659574	0.244925	0.769231	0.7	0.572146
1	0.234615	0.978723	0.244925	0.769231	0.5	0.631579
2	0.619231	0.000000	0.244925	0.769231	0.3	0.263158
3	0.657692	0.659574	0.244925	0.769231	0.7	0.578947
4	0.234615	0.978723	0.244925	0.769231	0.5	0.631579

Table 2.5 : Values after normalization

2.2 Modelling

2.2.1 Model Selection

There are broadly 2 types of models: Regression and classification.

1. If the target variable is nominal (aka categorical), then we perform a classification.
2. If the target variable can be any number, then it is a regression problem.

In our case, the target variable is “Absenteeism time in hours” is a number, which could take any value and does not denote any category, therefore we shall be developing a regression model.
We always start with making a simple model first, and them moving on something more complex. Therefore let us start with the first model, a Decision Tree Regressor.

2.2.2 Decision Tree Regressor

Before beginning the creation of the model, it is important for us to divide our dataset into training and test set, so that we have a way to determining how efficient the model built actually is.

```
X_train, X_test, y_train, y_test = train_test_split( X, Y, test_size = 0.2)
```

The above code takes 80% of the dataset as training and the remaining 20% as test.

Now let us continue to the creation of the Decision Tree Regressor, as shown below:

```
DT_model = tree.DecisionTreeRegressor().fit(X_train, y_train)
```

```
DT_Predictions = DT_model.predict(X_test)
```

```
print(mean_squared_error(DT_Predictions,y_test))
```

```
456.48383044747135
```

```
DT_model.feature_importances_
```

```
array([0.28856306, 0.02369419, 0.18932244, 0.03091781, 0.14706064,
       0.15207684, 0.07854706, 0.        , 0.03324875, 0.03852775,
       0.01804144])
```

One of the efficient ways to check if the model built is performing well is to check the mean squared error (MSE) value. Mean squared error is the average squared difference between the predicted values and the actual value.
For the above Decision Tree the MSE is 456.483

Another important parameter to look for is the feature importance, and how much percentage is given to each feature. Based on the Decision Tree, the most important feature in the regression model is “Reason for Absence”, which makes sense.

Now we can always try for a complex model to check if we can get a lesser MSE. Let us take the case of a Random Forest Regressor, which is basically a collection of decision trees, and it averages out the decision, therefore in theory it should produce a better result.

2.2.3 Random Forest Regressor

```
RF_model = RandomForestRegressor(n_estimators = 500).fit(X_train,y_train)
```

```
RF_Predictions = RF_model.predict(X_test)
```

```
RF_model.feature_importances_
```

```
array([0.26639474, 0.05498919, 0.08604026, 0.08708316, 0.07296306,
       0.20414631, 0.06658182, 0.006185  , 0.04194506, 0.06244042,
       0.05123099])
```

```
print(mean_squared_error(RF_Predictions,y_test))
```

```
241.0856743722133
```

The MSE in case of a Random Forest is 241.085 which is much less when compared to a Decision Tree. We have taken the number of trees in the forest to be 500 (as shown in the parameter n_estimators = 500).
The Random Forest also gives the most importance to the predictor “Reason for Absence”.
Since the Random Forest produces a better result than a decision tree, we shall choose it as our default to answer the questions posed to us in this project.

Chapter 3

Analysis and Conclusion

3.1 Questions to be answered

We have been given a couple of questions to be answered through our analysis. Let us try and attempt them.

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

From our model development, we learnt that the parameter “Reason for absence” was given the most importance in our regression model. In order for us to better understand, let us plot the histogram related to it, as shown below:

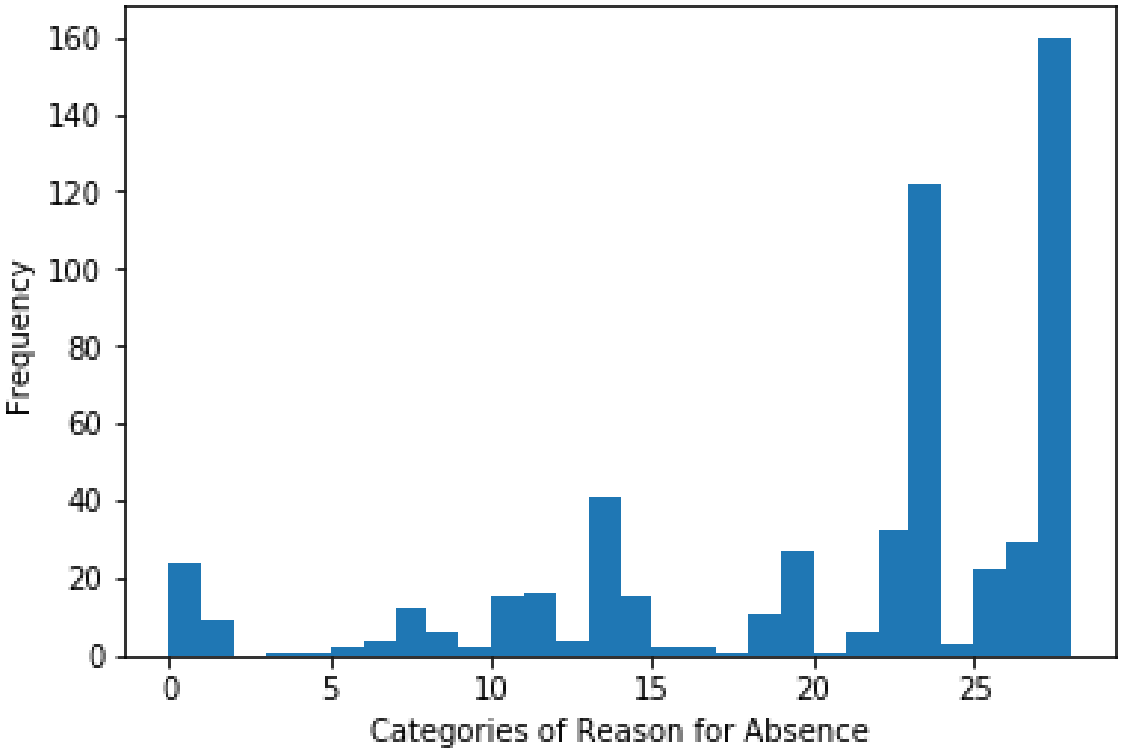


Table 3.1: Histogram denoting the frequency of the categories of “Reason for Absence”

From the above histogram, we see that the majority of the reason correspond to number 28 (dental consultation) and number 24 (blood donation).
One of the solutions could be to have occasional blood donation camps within the company, so that employees do not have to take a leave. With respect to the dental consultation, if there is an in-house dentist sitting in the campus, that would also drastically reduce the case of absenteeism (because dental consultation is a big chunk of the reason for absence).

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

We shall assume that every single case of absenteeism corresponds to 1 unit of losses to the company. Taking the assumption that the current trend of absenteeism repeats, we shall have the losses for every month as follows:

```
Month:  1
Losses to company:  225.34104945285452
Month:  2
Losses to company:  238.50615066992566
Month:  3
Losses to company:  765.6643858794639
Month:  4
Losses to company:  326.2946945723323
Month:  5
Losses to company:  201.49163054563752
Month:  6
Losses to company:  173.65674977182934
Month:  7
Losses to company:  544.9502259658784
Month:  8
Losses to company:  321.08729656571927
Month:  9
Losses to company:  142.44934786629142
Month:  10
Losses to company:  375.71334434191834
Month:  11
Losses to company:  307.01660123389684
Month:  12
Losses to company:  302.0102627596973
```

Table 3.2: Losses projected every month for upcoming year

3.2 Conclusion

There was a significant difference in the MSE for Random Forest and Decision Tree regressor models. Therefore we suggest to use the Random Forest Regressor.

3.3 Limitations

It is important for us to state the limitations of our work, and also suggest some improvements.

1. We are not sure if the Reason for Absence is genuine or not. If it's not, the predictions will be a little inaccurate.
2. For a better predictions, it is important that the dataset collected comes from all the branches of the office, not just one. If it is collected from a single source, it will lead to biased predictions.
3. There are a much more complex, and possibly accurate ways to handle the above scenario (example, a time series method).
4. The size of the data (around 700 observations) might not always be enough to accurately train our model, because we need adequate dataset for training, testing as well as validation.

THANK YOU.

Chapter 4

R Extra Plots

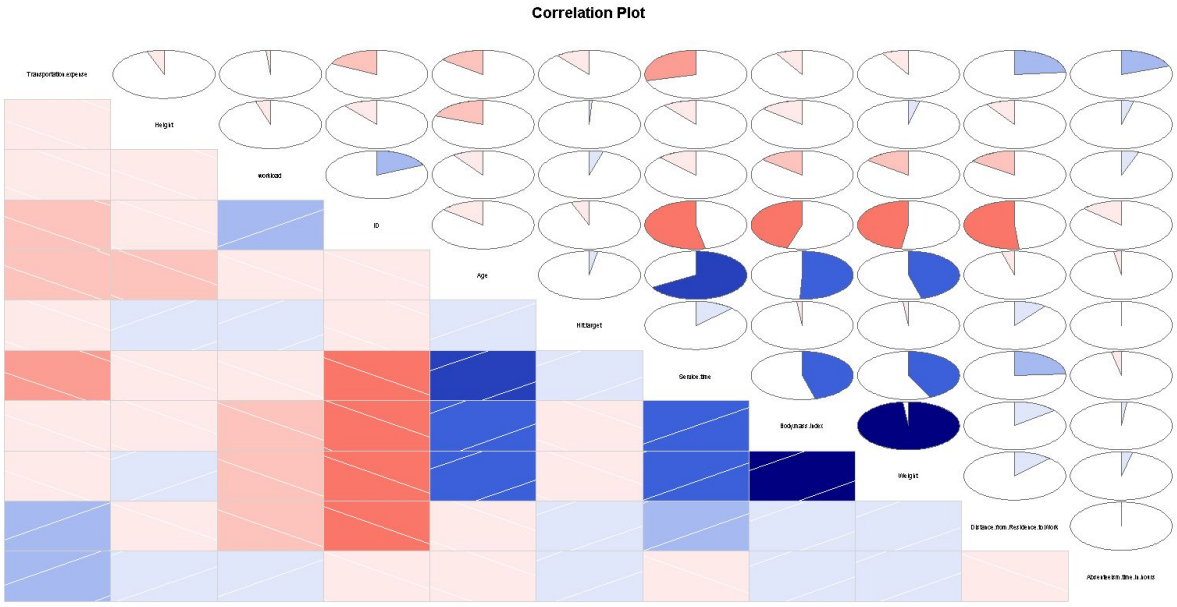


Fig 4.1: Correlation heatmap plot from R

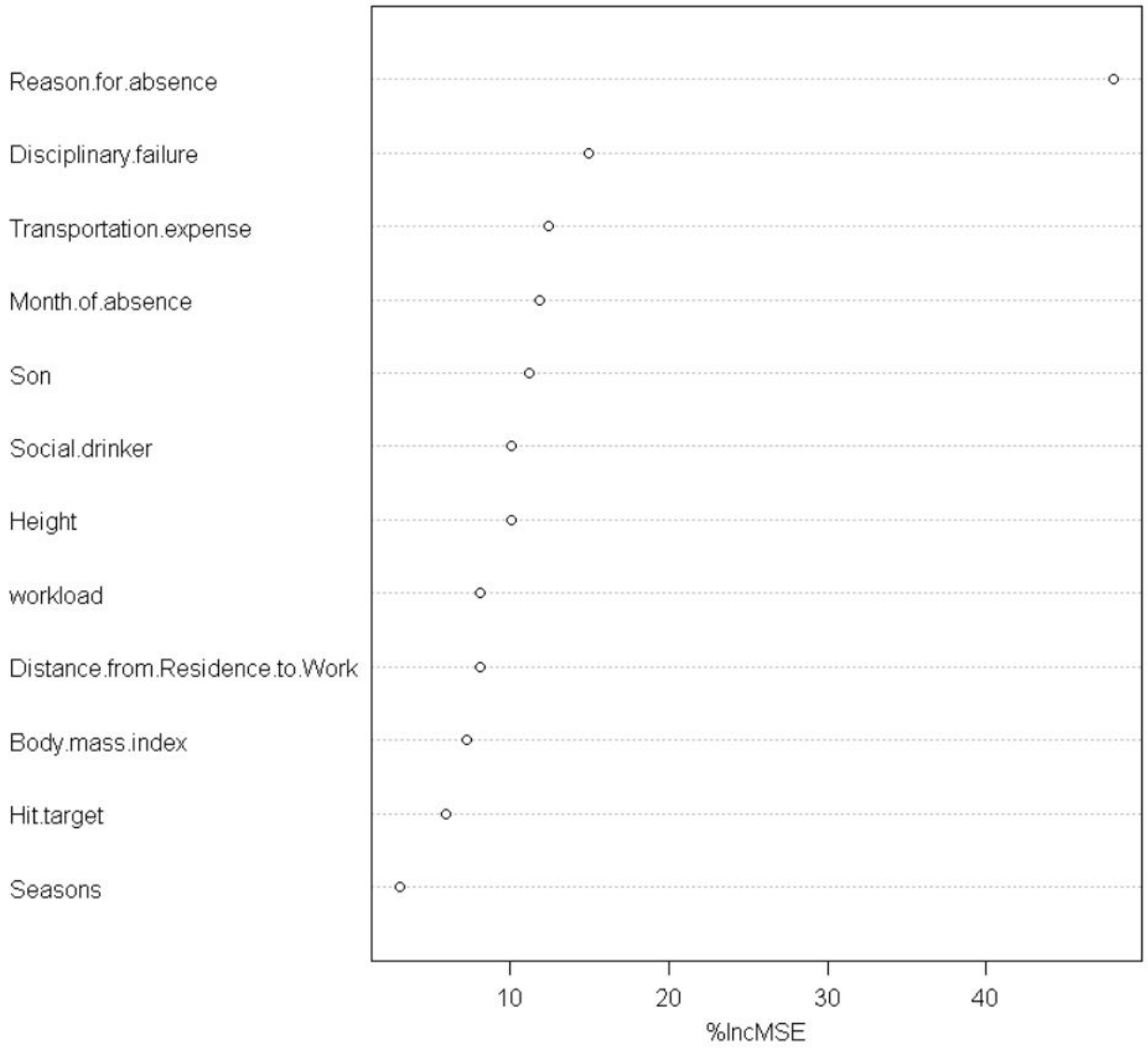


Figure 4.2: VarImpPlot from R to show the importance of the different predictors for the Random Forest in R

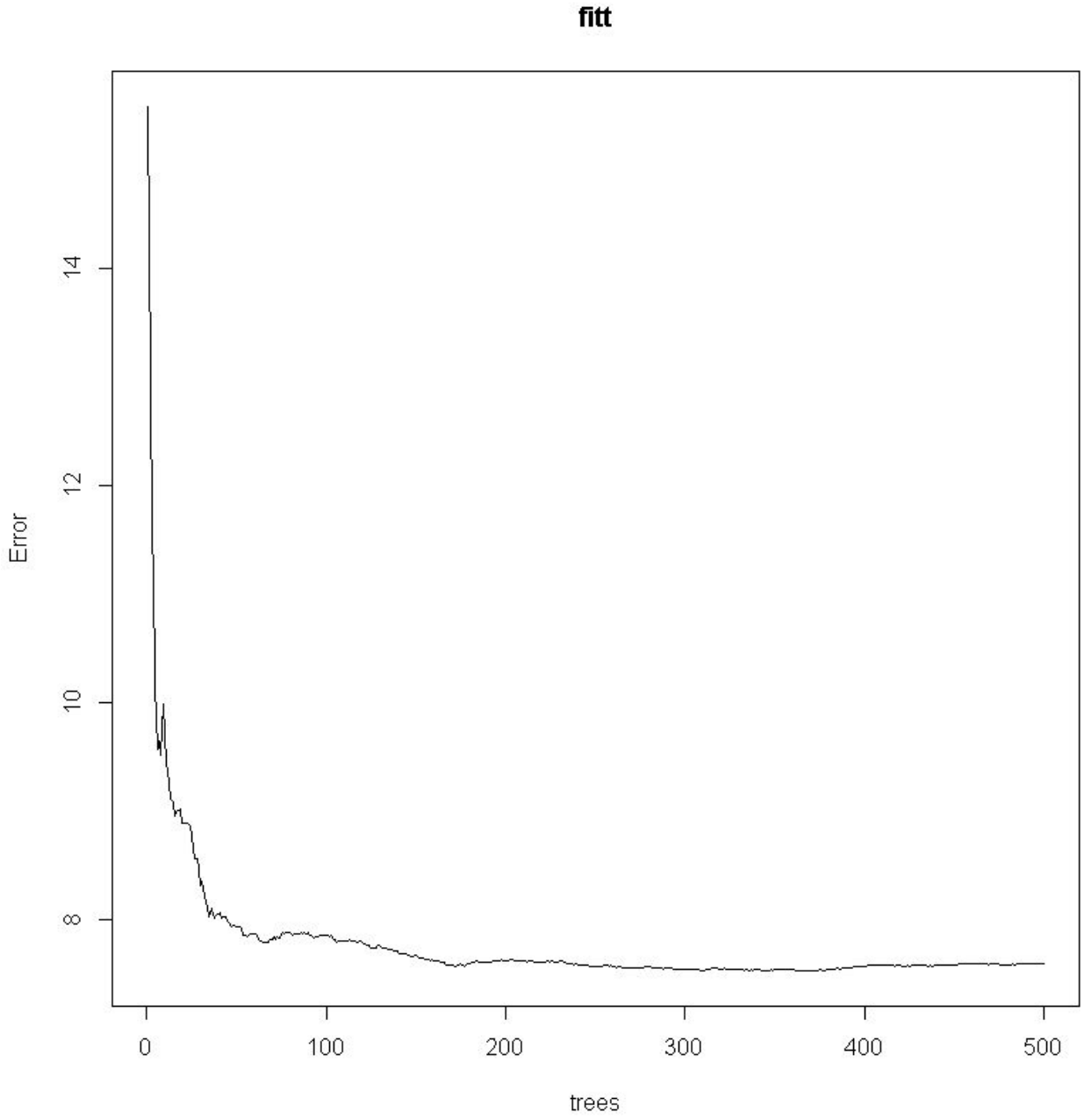


Fig 4.3: Showing the decreasing error as the number of trees in the forest increases. (Code written in R)