**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?

**Data**

Dataset Details:

Dataset Characteristics: Timeseries Multivariant

Number of Attributes: 21

Missing Values : Yes

Attribute Information:

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:

I Certain infectious and parasitic diseases

II Neoplasms

1. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioural disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system

XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities

XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX Injury, poisoning and certain other consequences of external causes XX External causes of morbidity and mortality

XXI Factors influencing health status and contact with health services.

And 7 categories without (CID)

(22)patient follow-up

1. medical consultation
2. blood donation
3. laboratory examination
4. unjustified absence
5. physiotherapy
6. dental consultation
   1. Month of absence
   2. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday

(6))

* 1. Seasons (summer (1), autumn (2), winter (3), spring (4))

1. Transportation expense
2. Distance from Residence to Work (kilometers)
3. Service time
   1. Age
4. Work load Average/day
5. Dataset having 740 rows and 21 columns.

**Methodology**

**Pre Processing**

Data preprocessing is a technique that involves transforming raw data into an

understandable format. Real-world data is often **incomplete**, **inconsistent**,

and/or **lacking** in certain **behaviors or trends**, and is likely to contain many **errors**.

Data preprocessing is a proven method of resolving such issues. Data

preprocessing **prepares raw data** for **further** processing*.* In these dataset we have id column which will not add any benefit to the model, so we dropped the the column id.

There is a value 0 in the absenteeism in month. This data might be incorrect so we have dropped the rows that have 0 month values. Then I convert absenteeism in hours to factor variable making it as 3 class variable i.e **“le2”, “bet 3 & 7” and “gt8”** which represents less than equal to 2,between 3 and 7 and greater than 8 respectively. Then I have converted some variables to factor variables as they were factor variables.

|  |
| --- |
| data.frame': 737 obs. of 20 variables:  $ reason\_for\_absence : Factor w/ 28 levels "0","1","2","3",..: 26 1 23 8 23 23 22 23 20 22 ...  $ month\_of\_absence : Factor w/ 12 levels "1","2","3","4",..: 7 7 7 7 7 7 7 7 7 7 ...  $ day\_of\_the\_week : Factor w/ 5 levels "2","3","4","5",..: 2 2 3 4 4 5 5 5 1 1 ...  $ seasons : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...  $ transportation\_expense : num 289 118 179 279 289 179 NA 260 155 235 ...  $ distance\_from\_residence\_to\_work: num 36 13 51 5 36 51 52 50 12 11 ...  $ service\_time : num 13 18 18 14 13 18 3 11 14 14 ...  $ age : num 33 50 38 39 33 38 28 36 34 37 ...  $ work\_load\_average\_day : num 239554 239554 239554 239554 239554 ...  $ hit\_target : num 97 97 97 97 97 97 97 97 97 97 ...  $ disciplinary\_failure : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...  $ education : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 3 ...  $ son : Factor w/ 5 levels "0","1","2","3",..: 3 2 1 3 3 1 2 5 3 2 ...  $ social\_drinker : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 1 ...  $ social\_smoker : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...  $ pet : Factor w/ 6 levels "0","1","2","4",..: 2 1 1 1 2 1 4 1 1 2 ...  $ weight : num 90 98 89 68 90 89 80 65 95 88 ...  $ height : num 172 178 170 168 172 170 172 168 196 172 ...  $ body\_mass\_index : num 30 31 31 24 30 31 27 23 25 29 ...  $ absenteeism\_time\_in\_hours : Factor w/ 3 levels "bet3&7","gt8",..: 1 3 3 1 3 NA 2 1 2 2 ...  **Summary of the absenteeism data** |
|  |
| |  | | --- | | reason\_for\_absence month\_of\_absence day\_of\_the\_week seasons transportation\_expense distance\_from\_residence\_to\_work service\_time  23 :148 3 : 87 2 :161 1 :169 Min. :118.0 Min. : 5.00 Min. : 1.00  28 :110 2 : 72 3 :153 2 :191 1st Qu.:179.0 1st Qu.:16.00 1st Qu.: 9.00  27 : 69 10 : 70 4 :155 3 :182 Median :225.0 Median :26.00 Median :13.00  13 : 55 7 : 67 5 :125 4 :194 Mean :221.3 Mean :29.63 Mean :12.55  0 : 40 5 : 64 6 :142 NA's: 1 3rd Qu.:260.0 3rd Qu.:50.00 3rd Qu.:16.00  (Other):311 (Other):376 NA's: 1 Max. :388.0 Max. :52.00 Max. :29.00  NA's : 4 NA's : 1 NA's :8 NA's :4 NA's :4  age work\_load\_average\_day hit\_target disciplinary\_failure education son social\_drinker social\_smoker  Min. :27.00 Min. :205917 Min. : 81.00 0 :691 1 :597 0 :294 0 :318 0 :678  1st Qu.:31.00 1st Qu.:244387 1st Qu.: 92.25 1 : 39 2 : 46 1 :226 1 :415 1 : 54  Median :37.00 Median :264249 Median : 95.00 NA's: 7 3 : 79 2 :154 NA's: 4 NA's: 5  Mean :36.42 Mean :271213 Mean : 94.59 4 : 4 3 : 15  3rd Qu.:40.00 3rd Qu.:284853 3rd Qu.: 97.00 NA's: 11 4 : 41  Max. :58.00 Max. :378884 Max. :100.00 NA's: 7  NA's :4 NA's :11 NA's :7  pet weight height body\_mass\_index absenteeism\_time\_in\_hours  0 :458 Min. : 56 Min. :163.0 Min. :19.00 bet3&7:177  1 :136 1st Qu.: 69 1st Qu.:169.0 1st Qu.:24.00 gt8 :261  2 : 95 Median : 83 Median :170.0 Median :25.00 le2 :276  4 : 32 Mean : 79 Mean :172.2 Mean :26.66 NA's : 23  5 : 6 3rd Qu.: 89 3rd Qu.:172.0 3rd Qu.:31.00  8 : 7 Max. :108 Max. :196.0 Max. :38.00  NA's: 3 NA's :2 NA's :15 NA's :32  **Graphs of the data**      The People who were suffering from puerperium, during pregnancy, child birth,  Congenital malformations, deformations and chromosomal abnormalities had  Absenteesim more than 8hrs.  The people who were suffering from certain infectious disease had absenteesim less than  2hrs. | |

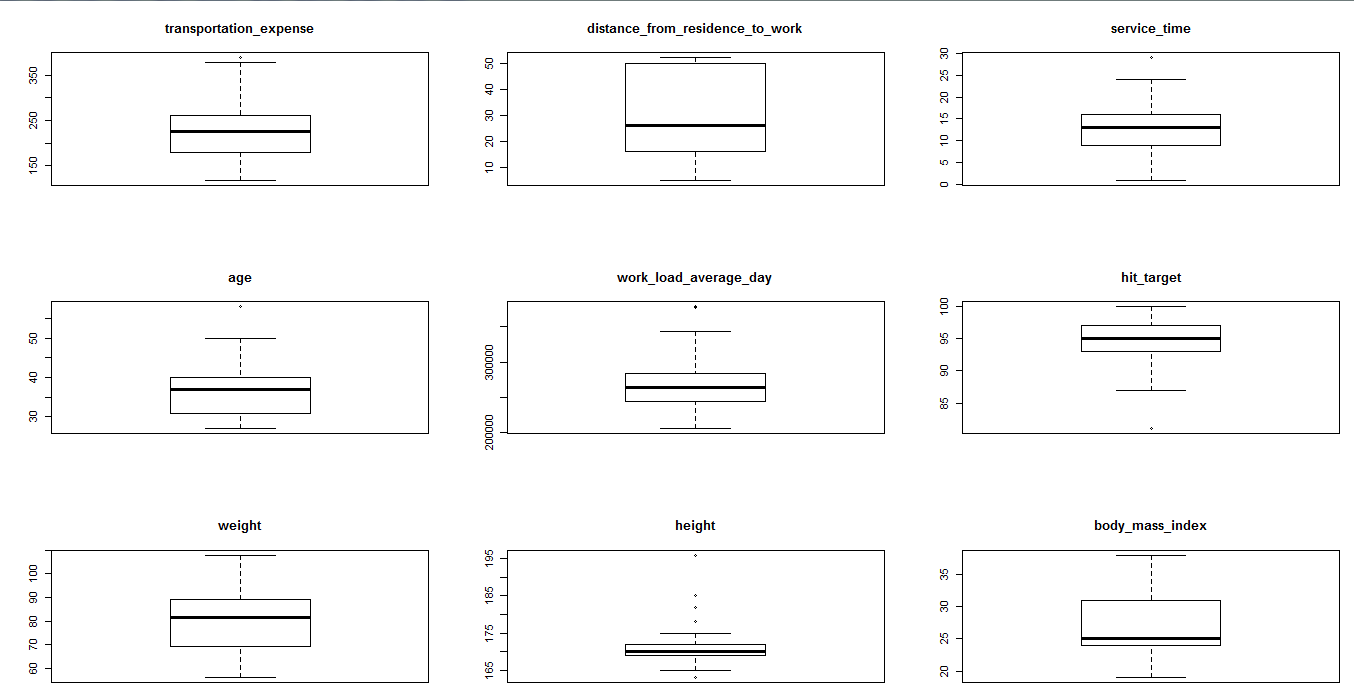
**Missing Value Analysis**

In my Absenteeism at work dataset I found missing values in every variable and the missing percentage is less than 30% of the data in every variable. So, I used three method (Mean, Median & KNN) to fill the value in missing place and out of these three method I found KNN Imputation works good in this dataset. So, I used KNNImputation method for missing value analysis and use K value as 3.

**Outlier Analysis**

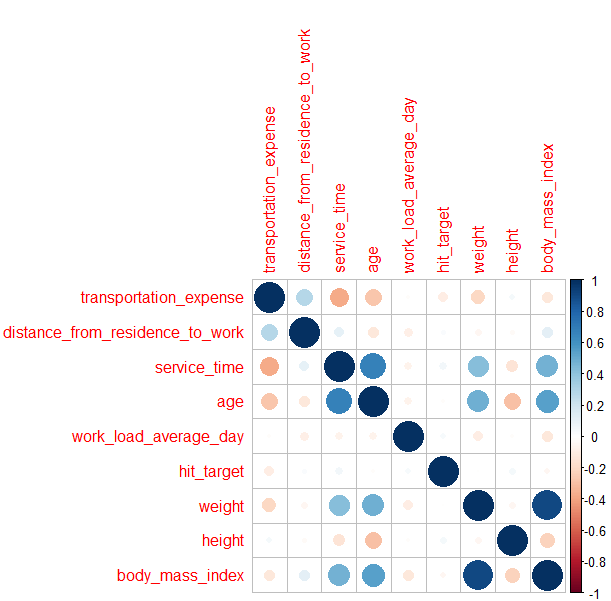
I observed from the dataset that most of the variables are skewed like; transportation expense, height, age, service time, etc. The skew in these variables can be most likely explained by the presence of outliers and extreme values in the data.

So first I replace the outliers with NA and then apply the KNNImputation method to fill these place with some proper value.



**Correlation Check and chisquare test**

* In feature selection method I checked the association between two variables.
* For continuous variables I used correlation analysis which tells the direction and strength of the linear relationship between two quantitative variables.
* I found Height and bmi are highly correlated and also weight and bmi are highly correlated. So I removed body mass index.
* For categorical variable I used chi-square test of independence to compare two variables in a contingency table to see if they are related.
* I reject two variable season and education as the p-value is greater than 0.05.



[1] "reason\_for\_absence"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 415.26, df = 54, p-value < 2.2e-16

[1] "month\_of\_absence"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 37.623, df = 22, p-value = 0.02019

[1] "day\_of\_the\_week"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 19.396, df = 8, p-value = 0.01288

[1] "seasons"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 7.1087, df = 6, p-value = 0.3109

[1] "disciplinary\_failure"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 42.727, df = 2, p-value = 5.272e-10

[1] "education"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 10.26, df = 6, p-value = 0.1141

[1] "son"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 70.717, df = 8, p-value = 3.538e-12

[1] "social\_drinker"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 11.228, df = 2, p-value = 0.003647

[1] "social\_smoker"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 6.7548, df = 2, p-value = 0.03414

[1] "pet"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 31.522, df = 10, p-value = 0.0004808

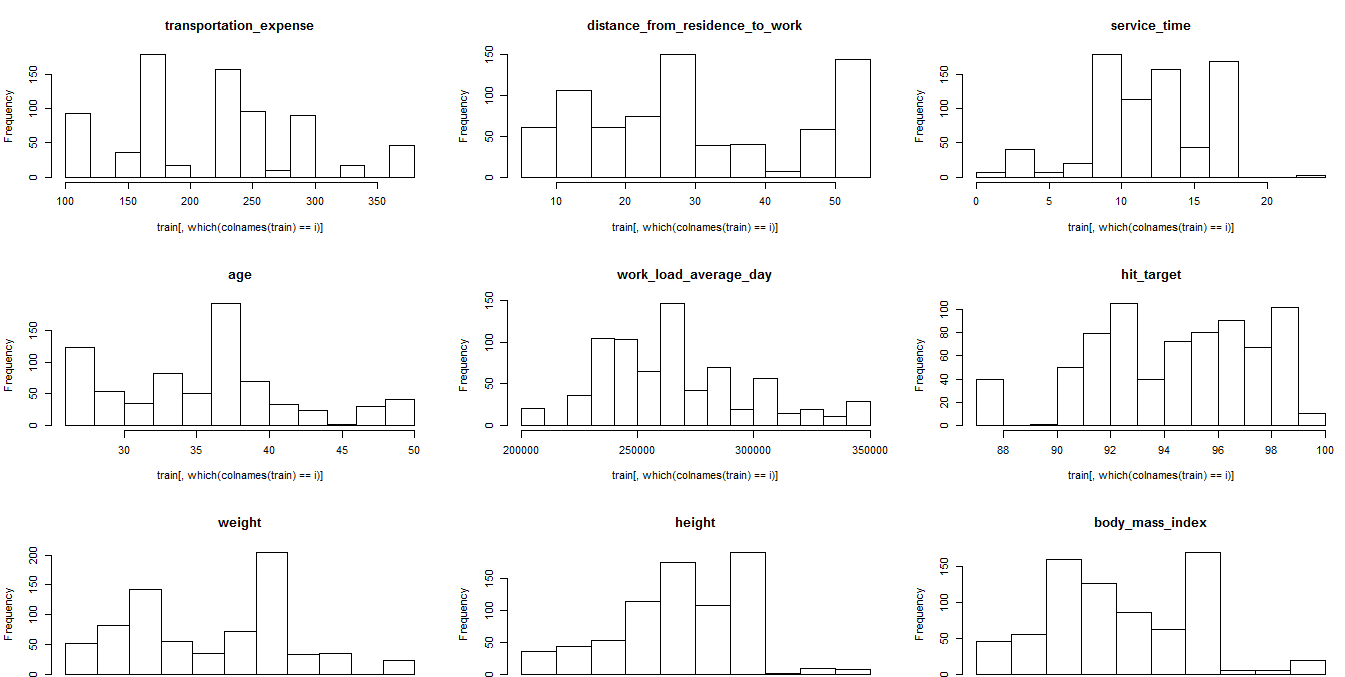
[1] "absenteeism\_time\_in\_hours"

Pearson's Chi-squared test

data: train$absenteeism\_time\_in\_hours and train[, which(colnames(train) == i)]

X-squared = 1474, df = 4, p-value < 2.2e-16

**Histogram of variables**



Decission Tree

**"For class between 3 and 7"**

[1] "Recall is 0.193548387096774"

[1] "Specificity is 0.824742268041237"

[1] "Precission is 0.260869565217391"

[1] "Accuracy is 0.671875"

**"for class greater than 8"**

[1] "Recall is 0.866666666666667"

[1] "Specificity is 0.734375"

[1] "Precission is 0.696428571428571"

[1] "Accuracy is 0.788990825688073"

**"For class less than 2"**

[1] "Recall is 0.650793650793651"

[1] "Specificity is 0.703125"

[1] "Precission is 0.683333333333333"

[1] "Accuracy is 0.677165354330709"

[1] "**Total accuracy** is 0.712677060006261"

KNN

**Random Forest**

predict

bet3&7 gt8 le2

bet3&7 27 0 0

gt8 2 37 0

le2 0 0 55

**"For class between 3 and 7"**

[1] "Recall is 1"

[1] "Specificity is 0.978723404255319"

[1] "Precission is 0.931034482758621"

[1] "Accuracy is 0.983471074380165"

**"for class greater than 8"**

[1] "Recall is 0.948717948717949"

[1] "Specificity is 1"

[1] "Precission is 1"

[1] "Accuracy is 0.983471074380165"

**"For class less than 2"**

[1] "Recall is 1"

[1] "Specificity is 1"

[1] "Precission is 1"

[1] "Accuracy is 1"

[1] "**Total accuracy** is 0.988980716253444"

Naïve Bayes

**"For class between 3 and 7"**

[1] "Recall is 0.419354838709677"

[1] "Specificity is 0.663157894736842"

[1] "Precission is 0.288888888888889"

[1] "Accuracy is 0.603174603174603"

**"for class greater than 8"**

[1] "Recall is 0.622222222222222"

[1] "Specificity is 0.75"

[1] "Precission is 0.636363636363636"

[1] "Accuracy is 0.697247706422018"

**"For class less than 2"**

[1] "Recall is 0.555555555555556"

[1] "Specificity is 0.732142857142857"

[1] "Precission is 0.7"

[1] "Accuracy is 0.638655462184874"

[1] "**Total accuracy** is 0.646359257260498"

Best Model

Random Forest is the best model for this dataset. When we apply our model in the on the test dataset and it give an accuracy of 99%. The recall and specificity is also very good. We have absenteeism more for medical consultation, dental consultation . So company should take some steps to reduce the consulatation.