Employee Absenteeism

Merapureddy Ajay

2 May 2019

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:

**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

III 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) patient follow-up (22), medical consultation (23), blood

donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27),

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)

**Below shown the structure of data:**

> str(data)

'data.frame': 740 obs. of 21 variables:

$ ID : num 11 36 3 7 11 3 10 20 14 1 ...

$ Reason.for.absence : num 26 0 23 7 23 23 22 23 19 22 ...

$ Month.of.absence : num 7 7 7 7 7 7 7 7 7 7 ...

$ Day.of.the.week : num 3 3 4 5 5 6 6 6 2 2 ...

$ Seasons : num 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 : num 0 1 0 0 0 0 0 0 0 0 ...

$ Education : num 1 1 1 1 1 1 1 1 1 3 ...

$ Son : num 2 1 0 2 2 0 1 4 2 1 ...

$ Social.drinker : num 1 1 1 1 1 1 1 1 1 0 ...

$ Social.smoker : num 0 0 0 1 0 0 0 0 0 0 ...

$ Pet : num 1 0 0 0 1 0 4 0 0 1 ...

$ 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 : num 4 0 2 4 2 NA 8 4 40 8 ...

There are 20 independent variables and one dependent variable i.e. Absenteeism. time.in.hours which is a continuous variable.

2.1 Preprocessing:

Before modeling, we need to explore data.That process often called **Exploratory Data Analysis.** We need to explore the data ,clean the data and make visualizing data through graphs and plots .Most analysis like regression ,require data to be normally distributed. We can visualize that by looking at histogram of the data.

In the given data every independent variable is num .So, we manually change the data types of categorical variables.

The numeric variables in the given data set are

[1] "Transportation.expense"

[2] "Distance.from.Residence.to.Work" "Service.time"

[4] "Age" "Work.load.Average.day."

[6] "Hit.target" "Weight"

[8] "Height" "Body.mass.index"

[10] "Absenteeism.time.in.hours"

And the categorical variables are

[1] "Reason.for.absence" "Month.of.absence" "Day.of.the.week" "Seasons"

[5] "Disciplinary.failure" "Education" "Son" "Social.drinker"

[9] "Social.smoker" "Pet" "ID"

2.1.1: Missing Value Analysis:

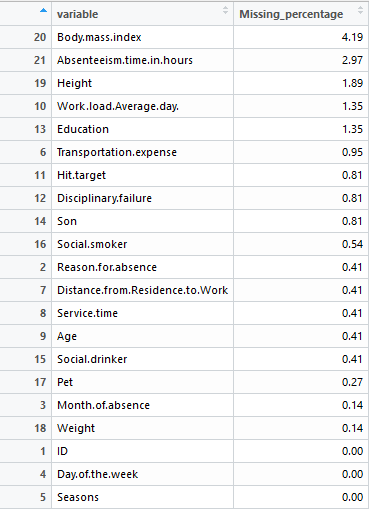
We need to check for missing values in the data. A table is shown below containing two columns named variable and missing\_percent.

As the missing percentage is less than 30 percent we can impute missing values.The method i chosed to impute missing values is KNN imputation.Knn imputation works good for this data set.

Before applying knn ,I manually imputed missing values with mode for Social.smoker ,Social.drinker and Education variables with use of ID variable.

The shape of data is (740,21).

After removing 0's in absenteeism.time.in.hours variable ,the shape of the data is (704,21).



2.1.2 Outlier Analysis:

Outliers are observations which are inconsistent with the rest of the data.It is one of the step of preprocessing ,before modeling data.We remove or impute outliers based on the count .

Outliers can be found using boxplot method.Boxplot is a graphical tool for data visualization.

There are 211 outliers in the given data set.We replace them with NA's and impute missing values using Knn method.

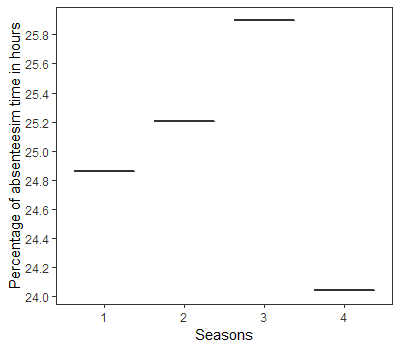
2.2 Data Visualization:

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

The above question can be answered with the use of data visualization .As we can extract valuable information from the insights.Let's start plotting.

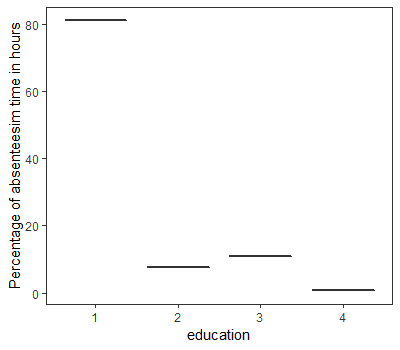
To answer the above question ,we need to get the plots of absenteeism. time.in.hours with other variables.When comparing with the categorical data we aggregate the absenteeism.time..in.hours and then make visualizations.

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

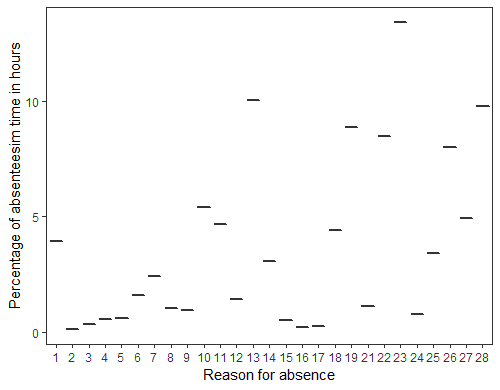


Employees tend to take leave in winter season more than 25 % of time.

# Education (high school (1), graduate (2), postgraduate (3), master and doctor (4)



Employees whose education background is high school take 80 percent of absenteeism time in hours.



23 => medical consultation (Employees with this reason are absent for more than 12.5% of total time)

13 => Diseases of the musculoskeletal system and connective tissue (Employees with this reason are absent for 10% of total time)

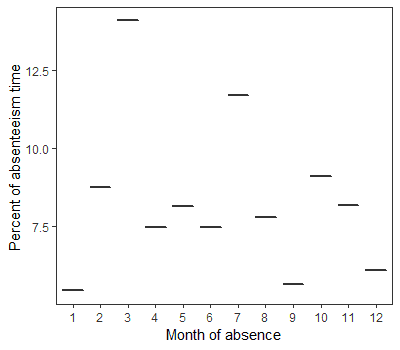
28 => dental consultation (28) (Employees with this reason are absent for around 9% of total time)

19 =>Injury, poisoning and certain other consequences of external causes

22 => 7 categories without (CID) patient follow-up (Employees with this reason are absent for around 8% of total time)

26 => unjustified absence (26) (Employees with this reason are absent for around 8% of total time)

27,10,11,18,1,14,25 => between 2.5 to 5 percent.



3 => march => employee in month of march tend to absent of about 13% of time.

7=> july => employee in month of march tend to absent of about 11% of time.

Absenteeism.time.in.hours can be reduced by considering following points:

1.Company should provide hygienic working conditions (Providing safety and health measures).

2.Service time should be around 12 hours or less than that. (High workload leads to high absenteeism time).

3. Providing leave facility based on employee's needs and organizational requirement.

4.Events can be organised in month of march and july .So there is a good chance to reduce absenteeism.

5.To reduce absenteeism it's better not to hire people who have diseases of musculoskelatal system and connective tissue complaints.

2.3 Feature selection :

In machine learning and statistics, feature selection, also known as variable selection.Based on the type of data we choose different methods to select variables.In our case dependent variable is Absenteeism.time.in.hours which is continuous , we choose anova method to select independent categorical variables and corrgram plot to select independent continuous variables.

Firstly, let's check for categorical features which are significant

> summary(aov(formula = Absenteeism.time.in.hours~., data[,c(cat\_index,21)]))

Df Sum Sq Mean Sq F value Pr(>F)

ID 32 1573 49.17 7.175 < 2e-16 \*\*\*

Reason.for.absence 26 1539 59.18 8.636 < 2e-16 \*\*\*

Month.of.absence 11 85 7.71 1.125 0.338

Day.of.the.week 4 17 4.21 0.615 0.652

Seasons 3 24 8.06 1.177 0.318

Disciplinary.failure 1 112 112.08 16.357 5.9e-05 \*\*\*

Son 2 2 0.93 0.135 0.873

Pet 1 1 1.44 0.211 0.646

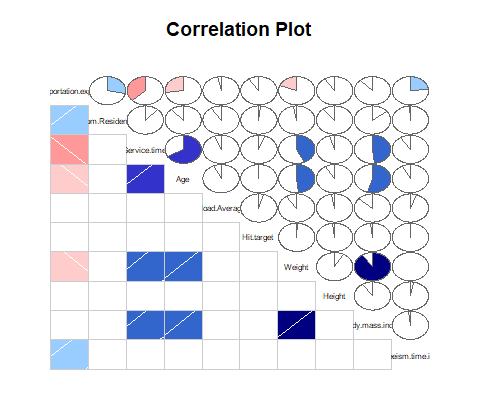
Residuals 623 4269 6.85

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> # Reason.for.absence ,disciplinary.failure,Son are significant.So, we remove other categorical features.

Below shown correlation plot.we can see there is strong linear relationship between weight and body.mass.index .So, we remove weight variable.



One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.We use one hot encoding to convert the categorical data in the dataset.The size of data after one hot encoding of categorical variables is (704,71).

2.4 Modeling:

Now we have the cleaned data which can be used to bulid models.

We start building models from simple to complex and check the prediction performance of each model .The first step before model buliding is fixing the size of train and test data.

2.4.1 Starting of with **Multiple Linear Regression:**

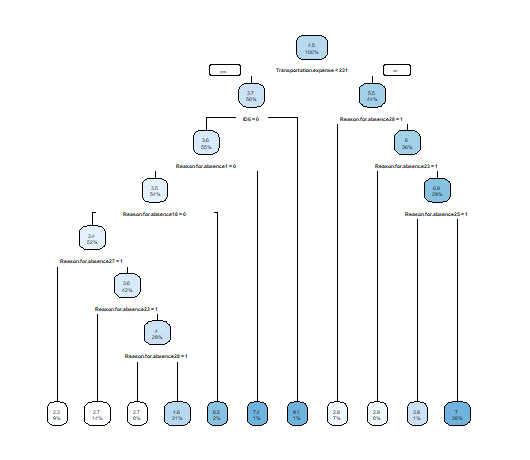
After passing train data to lm() function ,it calculates and assign weights to each predictor variable .

We choose rmse as our performance measure.Performance measure is used to choose optimal model.

The rmse for this model is 2.774819.

2.4.2 **Regression Trees:**

Decision trees is one of the predictive modeling approaches used in machine learning.The regression tree is a type of decision tree.After inputting train data to decision tree algo we get a decision tree



The above figure shows the decision tree of the model build based on the input data.Rmse value is 2.731049.

**2.4.3 Random Forest:**

Random Forest is one of the most effective machine learning models for predictive analytics.Random forests or random decision forests are an ensemble learning method for classification, regression.

Train data is inputted to the random forest function then we have a model.Now, we use test data to predict and then evaluate.

Performance measure : rmse value is 2.909031

As input data as 71 variables we try PCA (Principle component analysis) to reduce dimension of the data then use that data to bulid models and select the best model based on performance measure.

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.

So, after apllying pca to the data we get principle components and we use those PC's to predict Absenteeism.time.in.hours .

After inputting the data containing 40 components to each method above , we have the models and we use test data to predict Absenteeism.time.in.hours and evaluate the models.Below shown the performance measures of each ML model after applying PCA to the input data.

**Multiple Linear Regression:**

Performance Measure : rmse value is 2.719461

**Regression Trees:**

Performance Measure : rmse value is 2.997098

**Random Forest:**

Performance Measure : rmse value is 2.930581

2.5 Model Selection:

We use linear regression model which is built using train data after applying PCA , to predict Absenteeism.time.hours for new input data .

We choosed this model based on the performance measure.

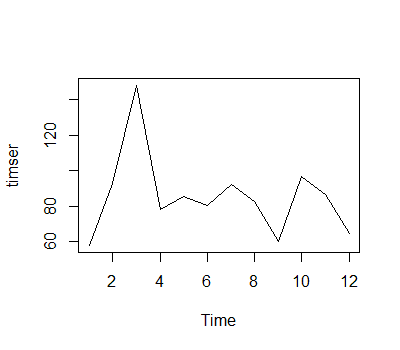
Time Series Analysis:

Now, Applying Time Series Analysis to predict losses every month that can we project in 2011 if same trend of absenteeism continues.

A time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data.

We use Month of Absence variable and aggregate Absenteeism.time.in.hours variable and get sum .As the data that we have is recorded from july 2007 to july 2010 , we divide each observation of aggregated Absenteeism.time.in.hours by 3 and july month (7) by 4 to get absenteeism.time per month.

Now, we have a plot of that absenteeism.time per month whose index is Month.of.absence



We use ARIMA model to predict losses .ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series.

But to apply Arima model the data should be stationary . We can check if the data is stationary or not using Augmented Dickey-Fuller Test.

###############################################

# Augmented Dickey-Fuller Test Unit Root Test #

###############################################

Residual standard error: 34.33 on 8 degrees of freedom

Multiple R-squared: 0.1278, Adjusted R-squared: -0.0903

F-statistic: 0.5859 on 2 and 8 DF**, p-value: 0.5788**

Value of test-statistic is: -0.5411

#p>0.05 that means data is not stationary.

We apply log to make variance uniform over time.And we differentiate to make mean constant over time.

Now,let's perform the Augumented Dickey-fuller test to check if data is stationary or not.Below shown is the result of the test.

Residual standard error: 0.3387 on 9 degrees of freedom

Multiple R-squared: 0.6513, Adjusted R-squared: 0.6125

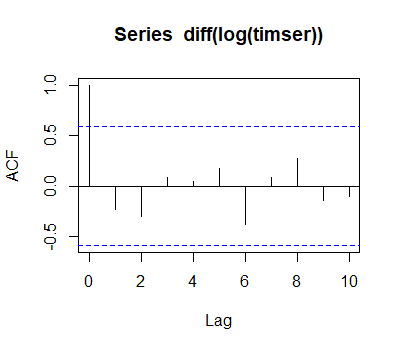
F-statistic: 16.81 on 1 and 9 DF, p-value: 0.002677

# p< 0.05 that means we can reject null hypothesis .

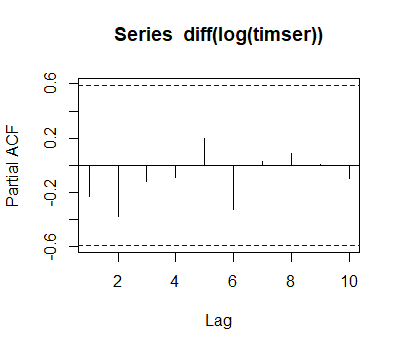
#Therefore the diff(log(timser)) data is stationary

Now, we find p,q,d parameters in arima using acf and pacf plots.

d=1 as we differntiated data once.



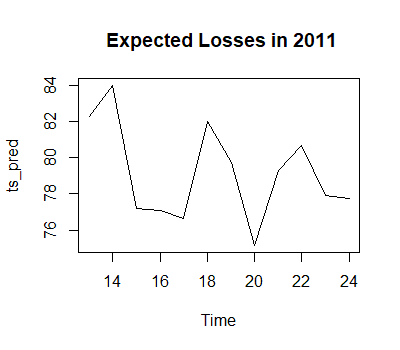
We see the ith value i.e. the upward line below blue dashed line which is just before inverted line.



For the acf and pacf plot we get values q=0 and p=4 .Now we bulid arima model by passing p,d,q values as order=c(4,1,0) and log of absenteeism.time per month.

AIC value is 8.91 .

Now, we use the model to predict losses in 2011 .Below shown the plot of predicted absenteeism.time.in.hours per month.



|  |  |
| --- | --- |
| Month ofabsence | Absenteeism time in hours |
| january | 82 |
| feburary | 84 |
| march | 77 |
| april | 77 |
| may | 76 |
| june | 82 |
| july | 79 |
| august | 75 |
| september | 79 |
| october | 80 |
| november | 77 |
| december | 77 |

Above shown the table which shows the Absenteeism time in hours for every month in 2011 .