Employee Absenteeism

Vishal Mourya

Contents

Chapter 1
Introduction

1.1 Problem Statement

Employee Absenteeism is the absence of an employee from work. Its a major problem faced by almost all employers of today. Employees are absent from work and thus the work suffers. Absenteeism of employees from work leads to back logs, piling of work and thus work delay.

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

Our task is to build Regression models which will predict the absenteeism depending on multiple employee characteristics. Given below is a sample of the data set that we are using to predict the absenteeism of employee:

As you can see in the table below we have the following 11 variables, using which we have to correctly predict the quality of the wines:

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)

Chapter 2
Methodology

2.1 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 just 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**, followed by pre processing.

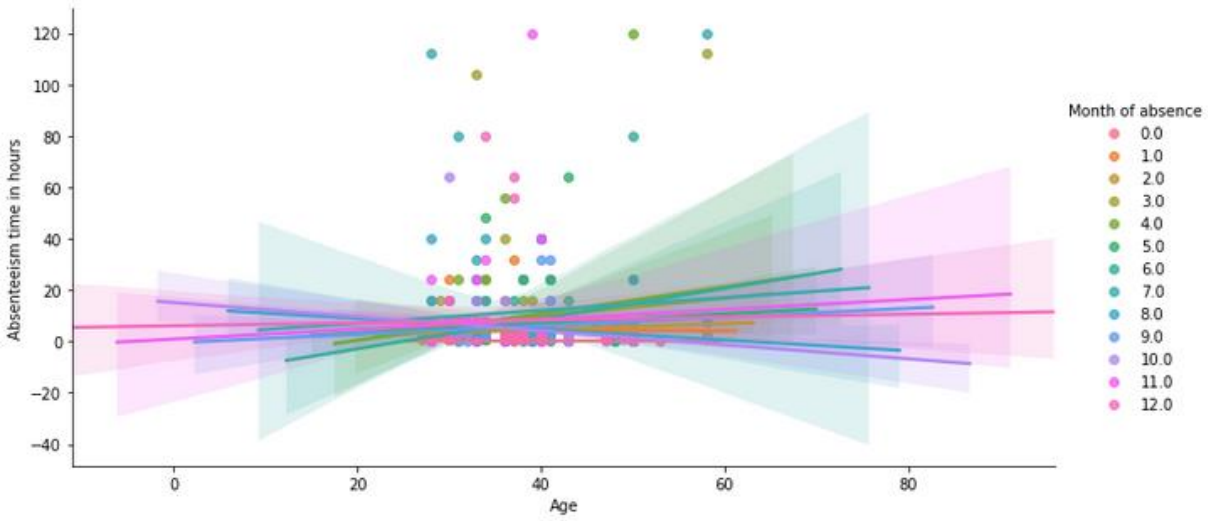
2.1.1 Univariate and Bivariate analysis:

1)Let us see if our hypothesis of reason of absence affecting absenteeism :-

```
23.0    119
28.0    102
27.0     60
13.0     51
0.0      34
19.0     33
26.0     30
25.0     30
22.0     29
11.0     24
10.0     22
1.0      16
14.0     16
18.0     16
7.0      15
12.0      8
6.0       6
8.0       5
21.0      5
5.0       3
24.0      3
16.0      3
9.0       3
4.0       2
15.0      2
3.0       1
2.0       1
17.0      0
Name: Reason for absence, dtype: int64
```

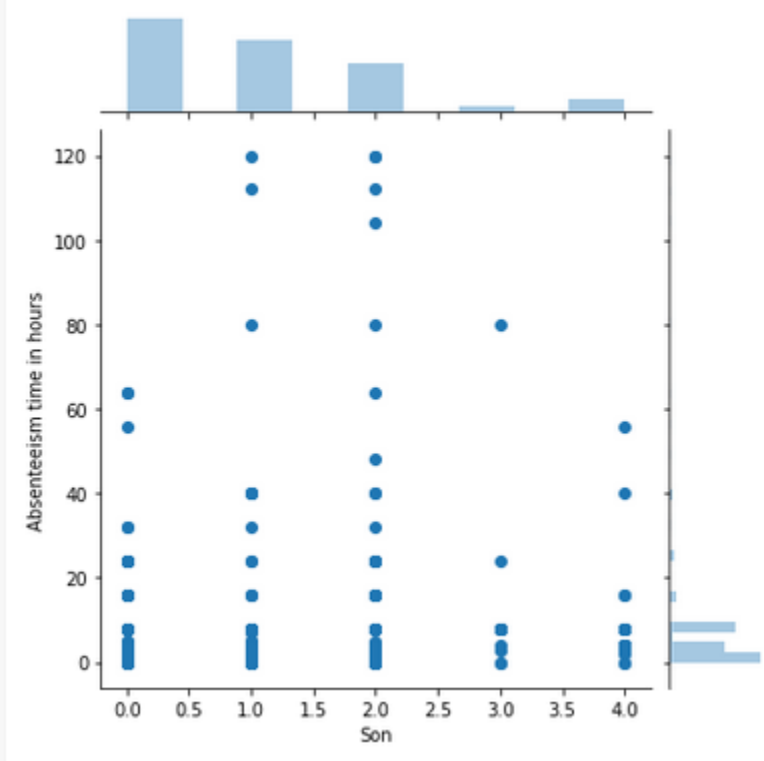
From the above fig we can observe that, main reason for absenteeism is **13(Diseases of the musculoskeletal system and connective tissue)** which means that the employee might have to do heavy work which results in connective tissues or bone issues. So, there should be a bone check up and first aid maintained for the employees. Another reason is **19(Injury, poisoning and certain other consequences of external causes)** which means same that because of work the employee are suffering and there are not enough medication or aid maintained to heal them. So, the employee has to take external treatment which results in absenteeism.
So, the workload should be reduced and also a team should be to look after the health if the employee gets any injury and should be treated.

2)Let us see if month causes any effect on absenteeism :-



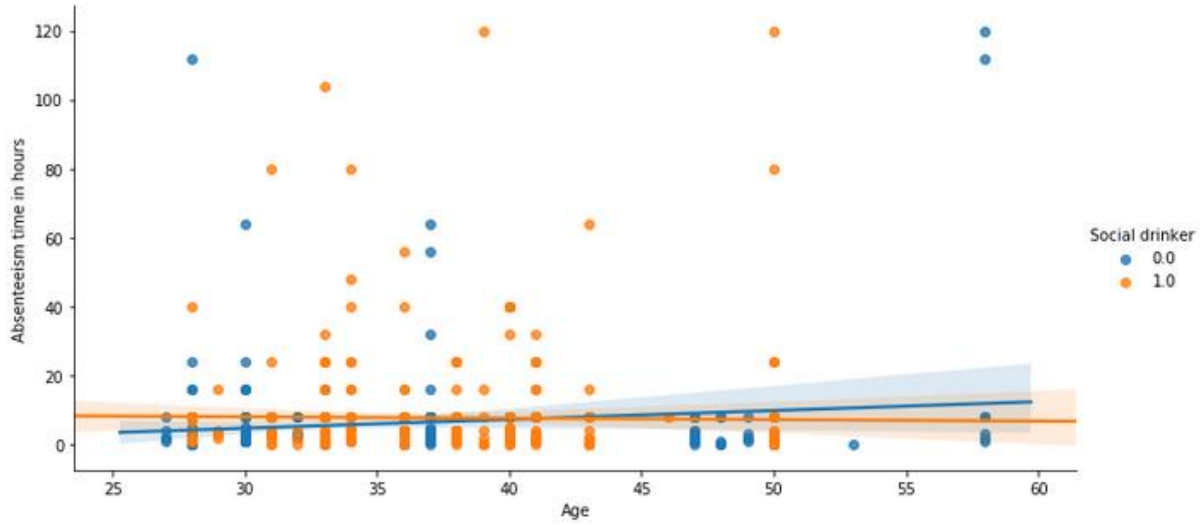
From the above fig we can say that most of the employees are absent in the season of winter. Most fall ill during winter season.

3)Let us see if no. of children causes any effect on absenteeism :-



From the above fig we can observe that employees having 1 or 2 children stays absent for long as they might have to look after them.

3)Let us see if Social Drinking causes any effect on absenteeism :-



From the above fig we can say that from age 33-45 has more absenteeism and also they are social drinker so the reason may be they cannot come most of the day because of the hangover.

2.1.2 Missing value Analysis and Outlier Analysis

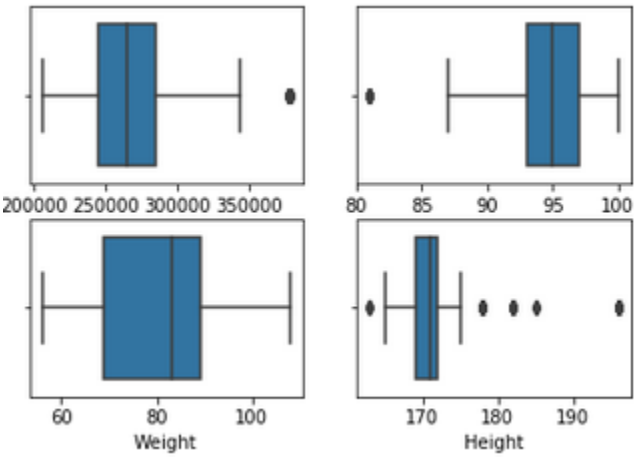
Missing values occur when no data value is stored for the variable in an observation. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Missing values are a common occurrence, and you need to have a strategy for treating them. Typically, ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values. We check for missing values in our data. We saw that the target variable has most number of missing value which seems to be less than 30% of the total data. And since it the target data we need to impute data. Here, we use KNN imputation and impute the data.

Now, we plotted box plot to see if there are any outliers in the data. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. There are numerous impacts of outliers in the data set. It increases the error variance and reduces the power of statistical tests.

If the outliers are non-randomly distributed, they can decrease normality. They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions. After plotting box plot, we got the figures as below:-

From the figure 1,2,3,4 we can see that height has most outliers followed by some other variables.

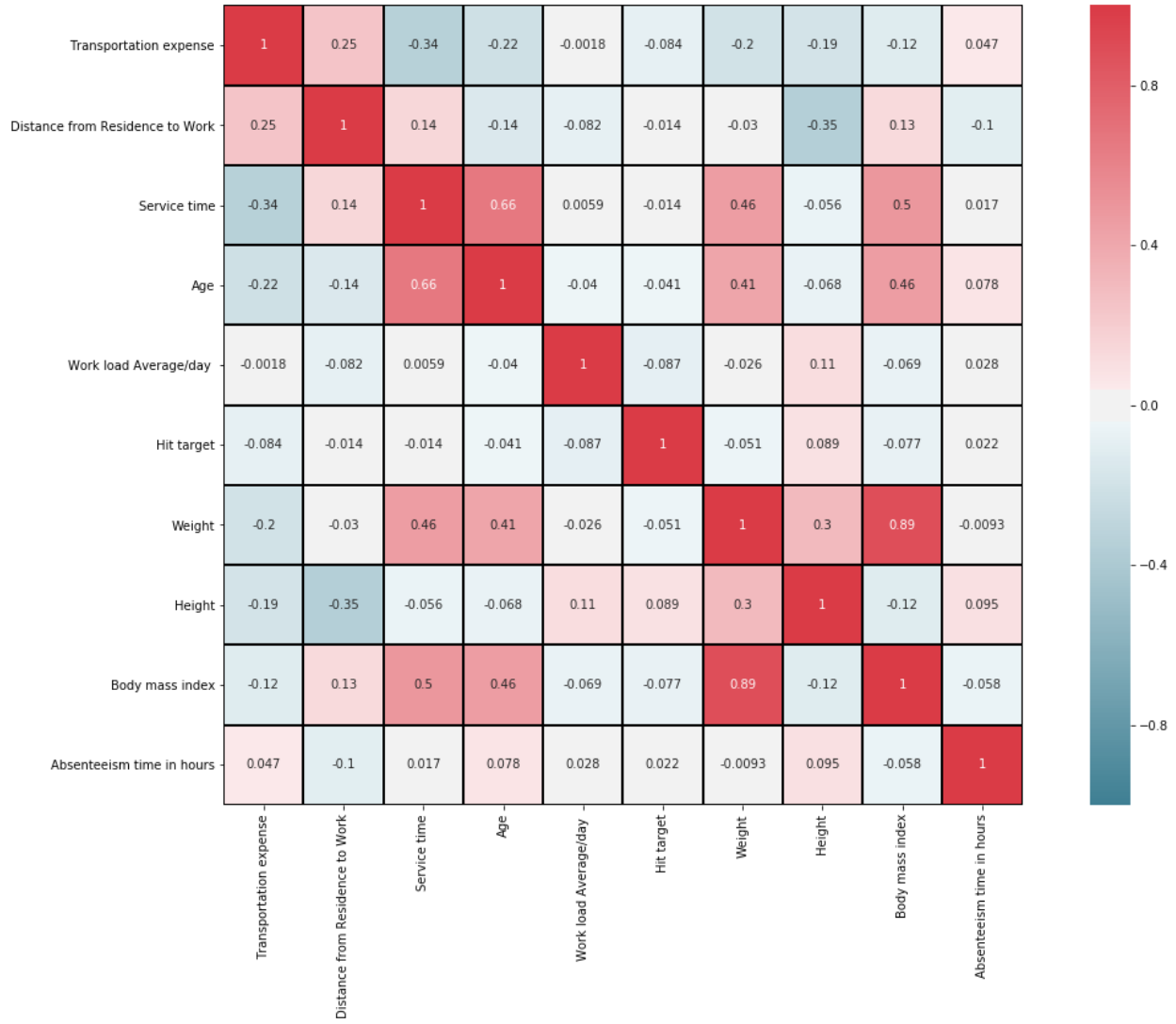
So, we replace them by NA and later impute this missing values using KNN imputation.



2.1.3 Feature Selection

Before performing any type of modelling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. There are several methods of doing that. We are using correlation matrix to identify most correlated numeric variables and anova test for categorical variables.

2.1.3.1 Correlation Matrix



From the above figure we can see that weight is highly correlated to body mass index. Also,service time is slightly correlated to age. So, we drop weight and service time.

2.1.4 Feature Scaling

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing steps. If training an algorithm using different features and some of them are off the scale in their magnitude, then the results might be dominated by them. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. We use normalization here for feature scaling.

Normalization brings all of the variables into proportion with one another. It transforms data into a range between 0 and 1. We have to see the variables that are scattered highly and apply normalization. We normalize the following variables in our data so that we can process to the modelling phase. Normality check for variables is in appendix

Formulae used for normalization is

$$Value_{new} = \frac{Value - minValue}{maxValue - minValue}$$

2.2 Modelling

As we know that our model is regression model, we first divide the data into train and test and then apply the train data on the following models, which gives us the metrics.

2.2.1 Decision tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

2.2.2 Random Forest

The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning.

2.2.3 Linear Regression

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

3.1 Model Evaluation

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using RMSE value.
RMSE: Root Mean Square Error is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit Lower the value of RMSE the better the model.

1) Decision tree:

After implementing the model we get the rmse value as #RMSE=0.0755528

2) Random Forest

After implementing random forest model we get the value as #RMSE=0.10004058

3) Linear Regression

After implementing linear regression model we get the value as #RMSE=0.18

3.2 Model Selection

We can see that Decision tree has the lowest value of RMSE i.e. 0.0755528

3.3 Answers to the problem statement

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

---> As we saw from the analysis that the main reason for absence is , main reason for absenteeism is **13(Diseases of the musculoskeletal system and connective tissue)** which means that the employee might have to do heavy work which results in connective tissues or bone issues. So, there should be a bone check-up and first aid maintained for the employers Another reason is **19(Injury, poisoning and certain other consequences of external causes)** which means same that because of work the employee are suffering and there are not enough medication or aid maintained to heal them. So, the employee has to take external treatment which results in absenteeism. So, the workload should be reduced and also a team should be to look after the health if the employee gets any injury and should be treated.

So, there should a system assigning a work once the employee finishes its assigned work. So, that he doesn't have to be absent thinking that he has no target to complete.

There are also employees which are heavy social drinker, they can be informed and tell them to consume drinks in appropriate amount so that the do not remain absent the next day.

There are some employees who remain absent the most. The company can act or give them warning to reduce the hours of absenteeism or strict action will be taken.