P1DM9 - Data Mining Assignment

‘Absenteeism at work’

Authors

Arpit Gupta
Avinash Dewangan
Lovel Setia
Pravin Pawar
Kushali Alias Alkesh Esso Prabhu Dessai

Introduction

Absenteeism at work

According to Forbes, Absenteeism is an employee's intentional or habitual absence from work where excessive absences can equate to decreased productivity and can have a major effect on company finances, morale and other factors.

A list of reasons that can be a cause of an employee being absent can be -

1. Childcare
2. Illness
3. Harassment at workplace
4. Stress

There can be number of reasons for absenteeism depending upon various individuals. This can have adverse consequences like

1. Decreased productivity
2. Poor morale among co- workers
3. Increased labor costs
4. Poor customer service
5. High administration cost

Data Exploration

The Dataset consists of total 740 rows having 21 features each. These can be further classified into following types:

1. In Categorical Features
    - Reason for absence
    - Day of the week
    - Disciplinary failure
    - Social drinker
    - Month of absence
    - Seasons
    - Education
    - Social smoker
2. In Numerical Features
    - ID
    - Distance from residence to work
    - Age
    - Hit target
    - Pet
    - Height
    - Transportation expense
    - Service time
    - Work load Average/day
    - Son
    - Weight
    - Body mass index

Expected output: Absenteeism time in hours

Data Pre-Processing

To perform the data analysis and build a good model we should have a good knowledge of the data. In the given data set we can see:

1. We can see that the “month of absence” column consists of some invalid data i.e. month data 0. And from the given column definition we know that the number of months can be from 1-12. So, we can remove those rows.
   After removing those rows 737 rows are left.

2. If we assume that the disciplinary failure = 1 means that the person is necessarily absent due to some discipline issue. The count of such rows is 40.
   And if we count the no of rows where 'Absenteeism time in hours' is zero, it comes out to 41. So, we can consider that the value of that 1 row is inconsistent and is not correct. So, we will replace the value of that row with the mean of the 'Absenteeism time’ of that id.

3. We assumed that disciplinary failure = 1 means that the employees were asked to be absent. So, in such cases (40 rows) we can replace the value 0 with 8.

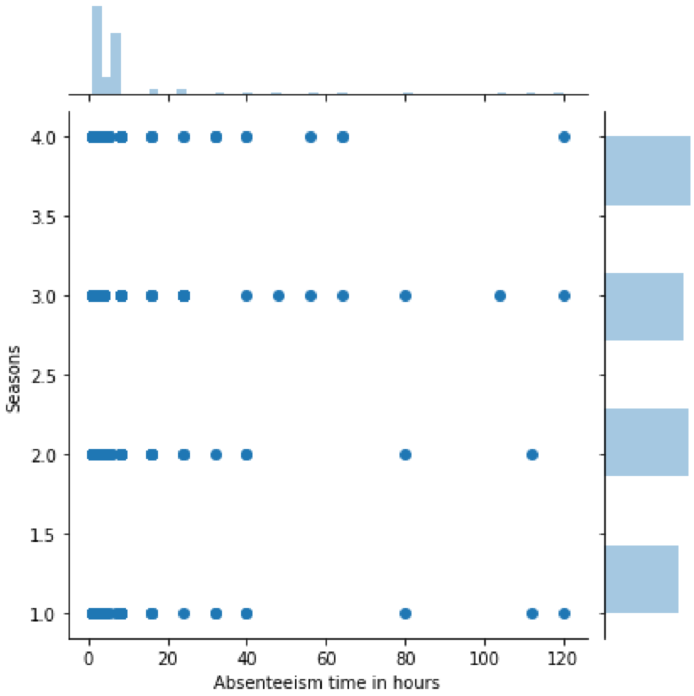
4. We know that the BMI is calculated on weight and height and since this is a derived column, we can remove this column from our dataset. ID column is also dropped since it does not contribute in prediction.

After doing the above pre-processing we have filtered our data and replaced all the inconsistent values. We are now ready for the for our data exploration and model building.

Data Analysis

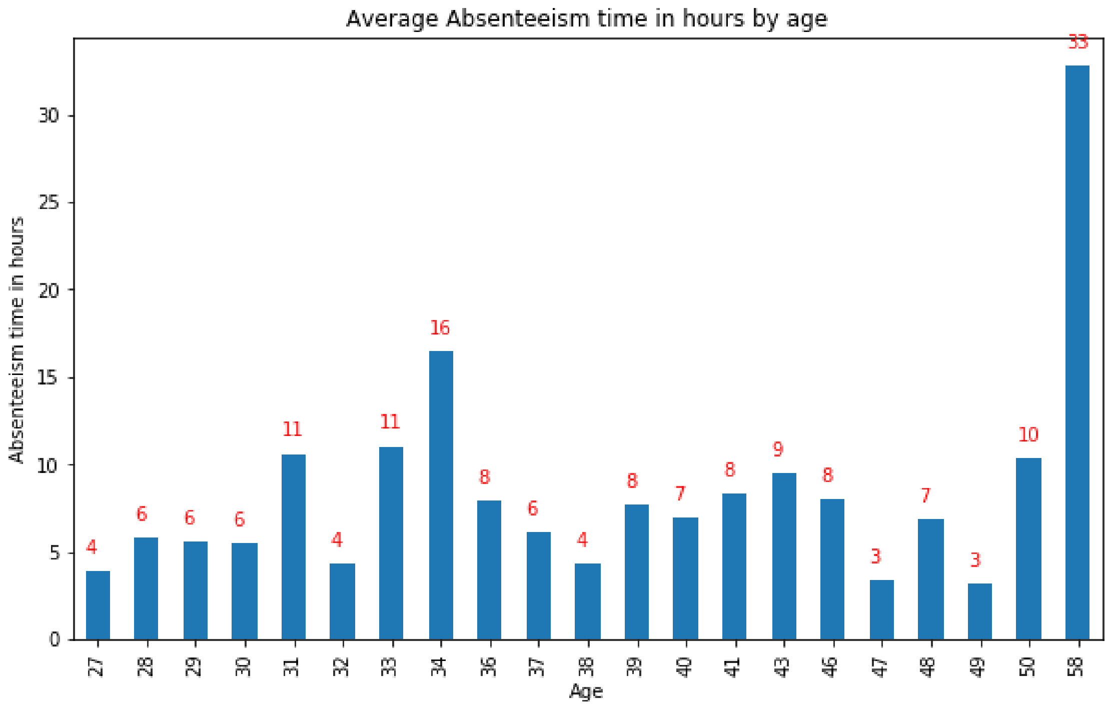
Let’s have a closer look for a few columns:

1. **Seasons:**
   All the employee data is present over 4 seasons. If we plot a joint-plot between 'Absenteeism time in hours' and ‘Seasons’, we can see that the data is evenly distributed over the all four seasons. So, we can’t conclude any results from this.



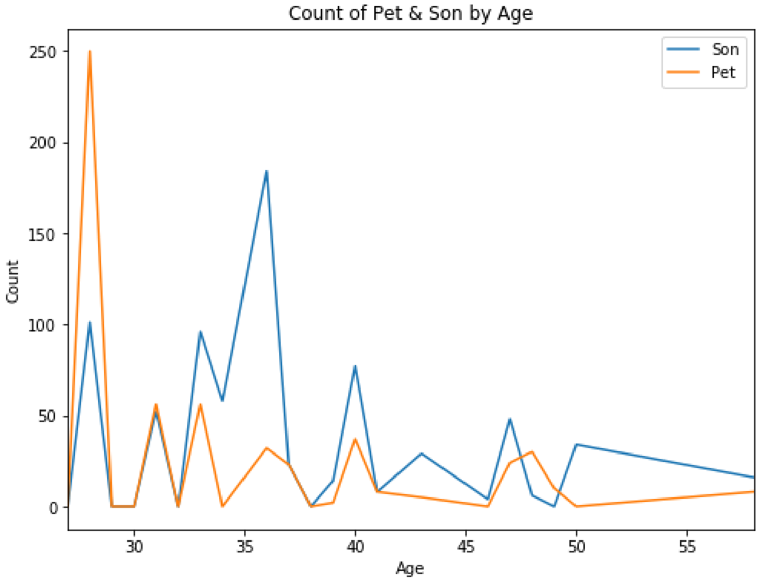
2. Age

For this column, I am trying to gather the type of employees who work at this company, is there a younger generation or is there a range?



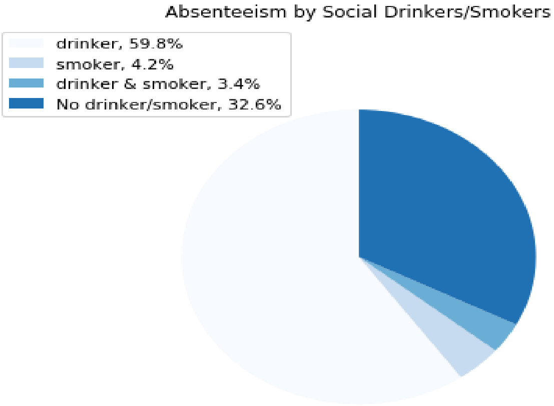
From the plot we can see that most workers are in between their late 30's and early 60's.

3. Son and Pet



When we plotted a chart between the count of ‘Son and Pet’ vs the age. We get an interesting analysis that most workers till the age of 40 have a higher pet and son count.

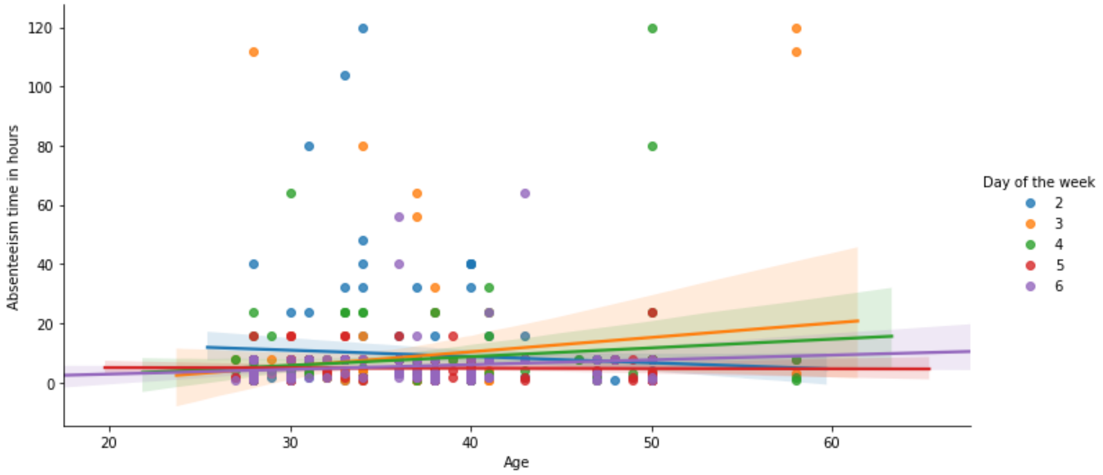
4. Social drinking and smoking



If we plot a chart for the social drinking and smoking. We see that the percentage of workers who drink more are more likely to be absent.

5. Day of the week

-



From the above plot we can see that most of the workers are absent on the second day of the week and are in the age group of 30-40.

6. Reason for absence

Another column of importance is Reason for absence with its 28 levels where each number represents a reason given by an employee for their absence.
Most common reasons are the categories outside of the ICD (International Code of Diseases) meaning that the most reasons were non serious diseases such as:

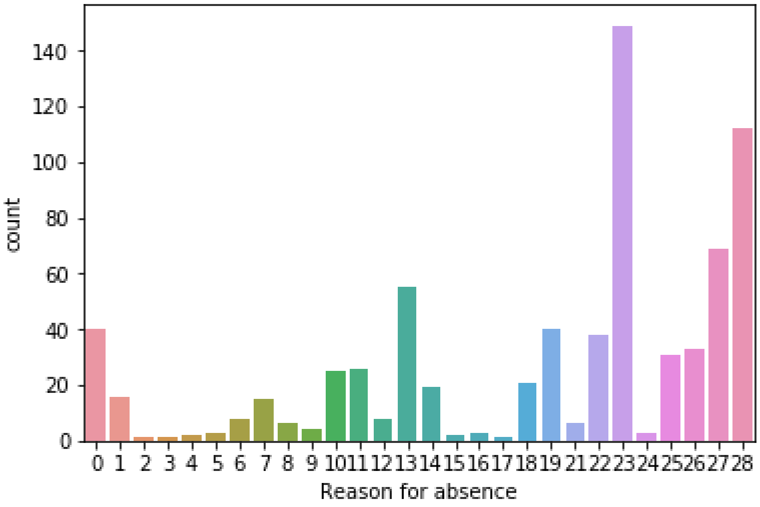
- Medical consultation (23) - 149
- Dental consultation (28) - 112
- Physiotherapy (27) - 69

These 7 categories are used by **45% of employees**.

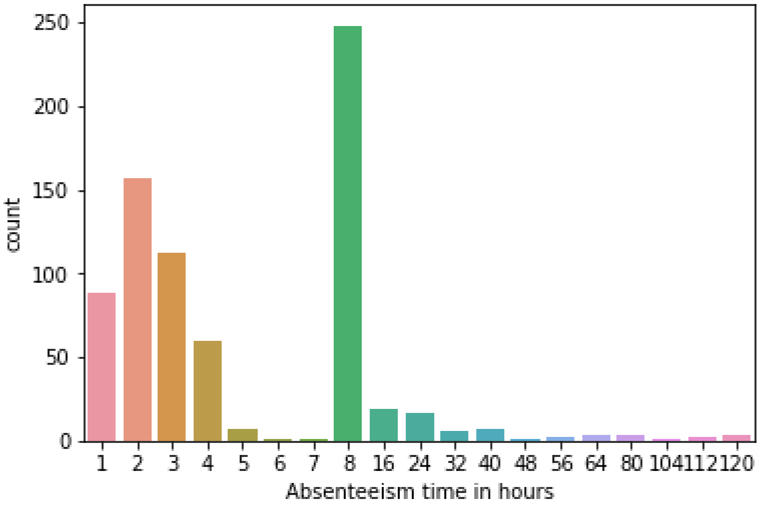
For the ICD diseases or serious disease, the main 2 categories with relatively high numbers Are

- Diseases of the musculoskeletal system and connective tissue (13) – 55
- Injury, poisoning and certain other consequences of external causes (19) – 40

These two seem like they could be related to injuries one gets from this type of industry. For example, someone who is always doing deliveries for a long period of time is more likely to have musculoskeletal issues. It is surprising how only a few absences are related to birthing or children related.



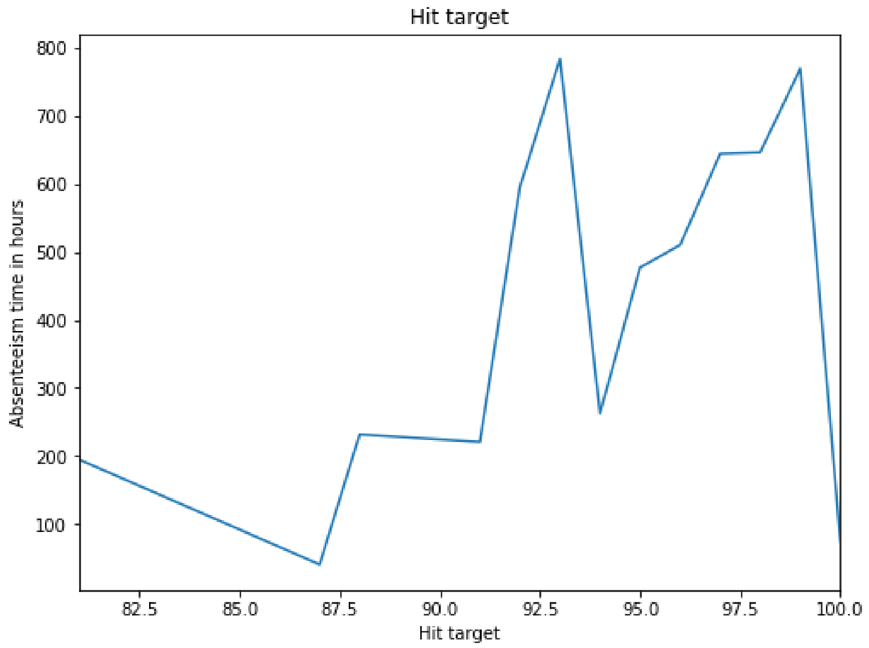
7. Absenteeism time in hours



From the above chart we can clearly see that most of the employees have taken only one day leave. Few of them are also the case of disciplinary failure.

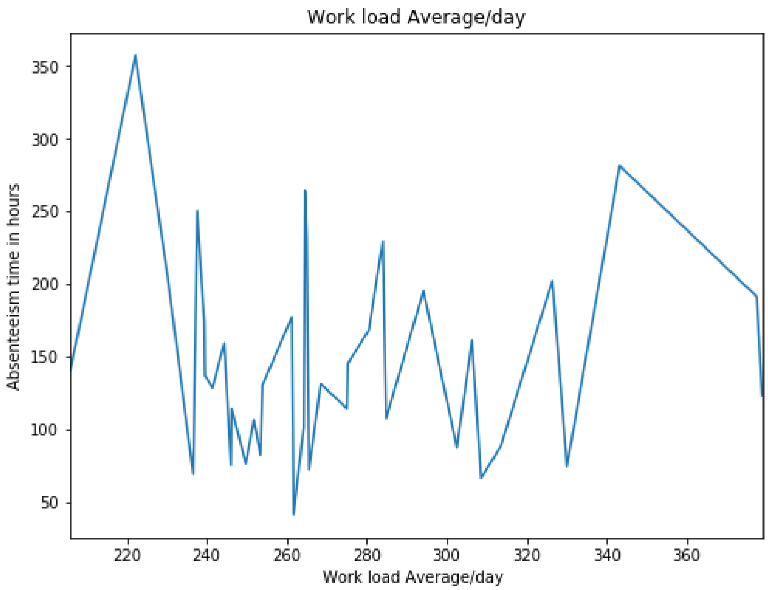
Other than these workers, most frequently absent in the bin of 1-4 hrs. and these cannot be considered as absenteeism.

8. Hit Target



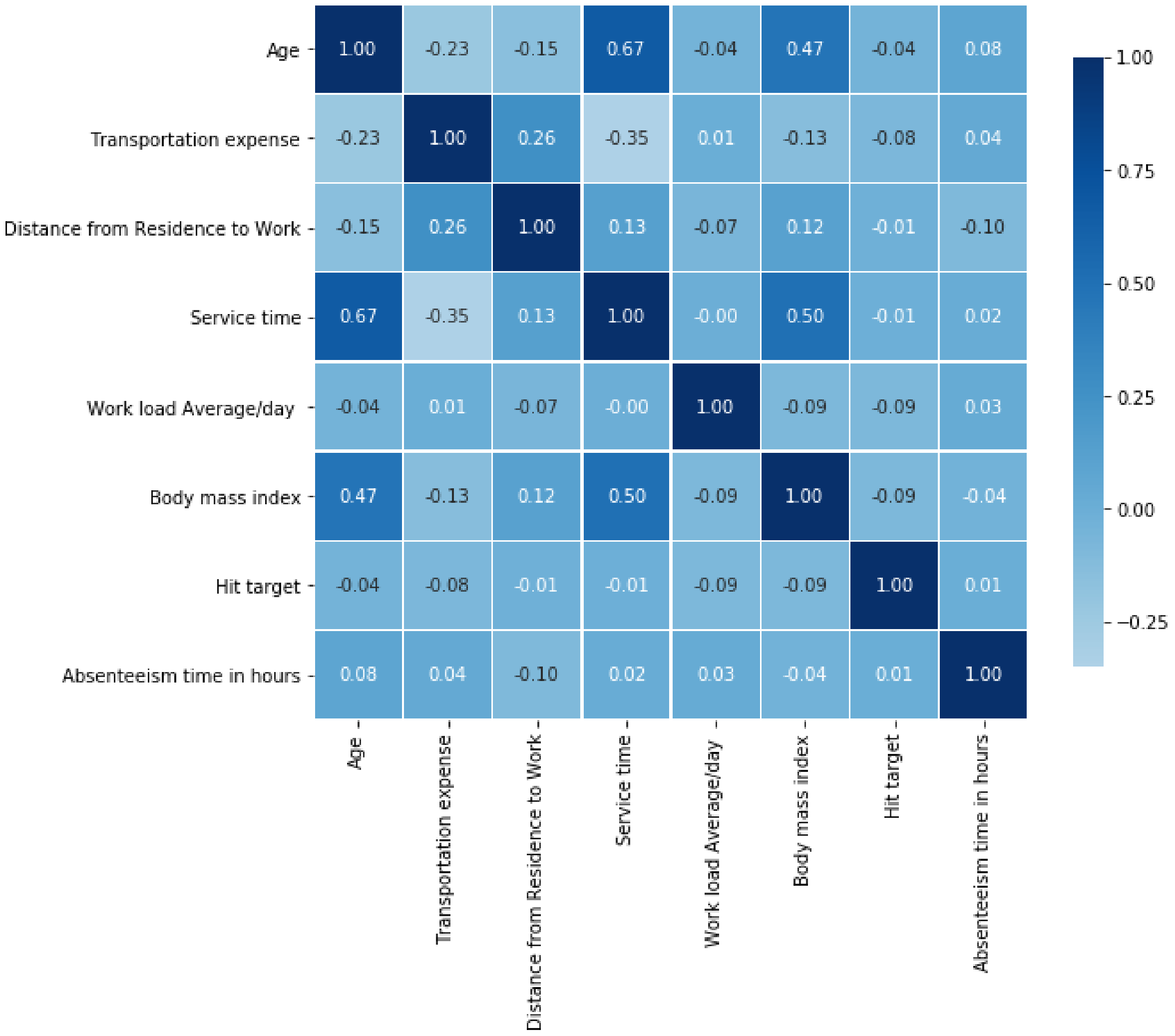
Above plot indicates Higher the hit target higher the absenteeism hours

9. Work Load Average



Above plot indicates that higher the absenteeism time in hours occurs for both extremes of work load (lowest and highest).

10. Co-relation between columns



From the above co-relation analysis, we cannot find any pattern which can lead to the absenteeism.

The above analysis shows how both Age, hit target and Work load Average/day have high absenteeism time in hours.

Model Building and Evaluation

1. Using Regression Model

We need to predict the value of absenteeism in hrs. We first tried solving this problem using regression.

We used regression models:

a) RandomForestRegressor
b) Linear Regression
c) Linear Regression with further preprocessing
d) Linear Regression with Feature reduction

We split the dataset into two parts:
1. Training Dataset - Containing 70 % of total records.
2. Test Dataset – Containing 30 % of total records.

a) RandomForestRegressor

We first used RandomForestRegressor to predict the values of Absenteeism.
When we verified the output of the test data following was the output for various methods:

Test data R-2 score: 0.125
Test set of RMSE: 10.89

b) Linear regression

Next we tried linear regression with as is data set.

With this we got following R-2 score and RMSE:
Test R-2 score: 0.147
Test RMSE : 10.76

c) Linear regression with further preprocessing

Next we tried improving linear regression with further preprocessing. Following preprocessing was done on the data set before running this model.

1. Using box plot we found that absenteeism hours > 24 hours were outliers. We removed these records and found that the RMSE value improved significantly. However R-2 score is still pretty low.
2. We scaled the all the attributes 'StandardScaler' library
3. We converted the categorical attributes into encoded attributes.

With this we got following R-2 score and RMSE:
Test R-2 score: 0.375
Test RMSE : 3.32

d) Linear regression with feature reduction

Next we used Recursive Feature Elimination library (RFE) to reduce number of features used in building model without impacting R-2 score and RMSE.

Below table shows how R-2 score and RMSE vary with different number of features.

Number of features	R-2 score	RMSE
5	0.001	4.2
10	0.03	4.14
15	0.06	4.07
25	0.05	4.09
30	0.02	4.16
31	0.13	3.9
35	**0.33**	**3.42**
38	**0.39**	**3.28**
40	0.37	3.33
45	0.36	3.36
50	0.37	3.31
55	0.37	3.31
60	0.36	3.36

We found that with **38 features** we have an optimal balance of best score with list number of features:

Test R-2 score: 0.392

Test RMSE : 3.2815

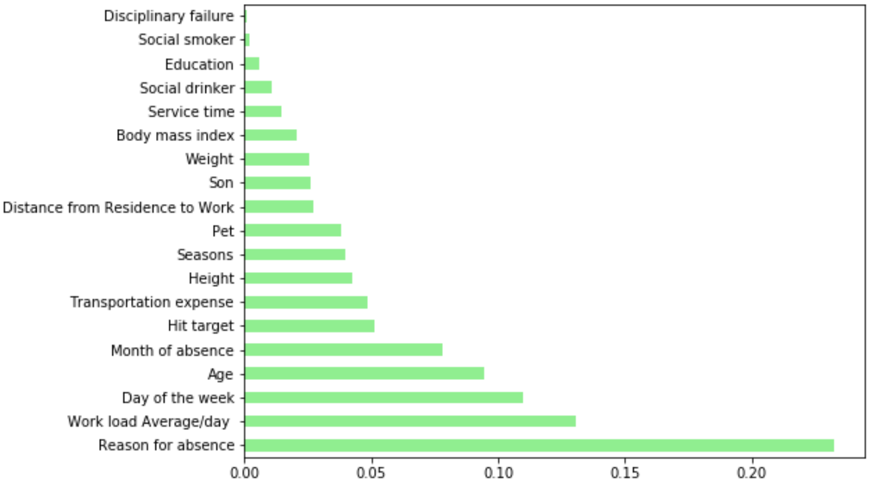
Below table gives details of features which were selected along with their score:

Feature	Feature Ranking	Selected
Disciplinary failure	1	TRUE
Reason for absence_2	1	TRUE
Reason for absence_11	1	TRUE
Reason for absence_23	1	TRUE
Reason for absence_25	1	TRUE
Reason for absence_27	1	TRUE
Reason for absence_28	1	TRUE
Month of absence_0	1	TRUE
Month of absence_1	1	TRUE
Month of absence_2	1	TRUE
Month of absence_3	1	TRUE

Feature	Feature Ranking	Selected
Month of absence_4	1	TRUE
Month of absence_5	1	TRUE
Month of absence_6	1	TRUE
Month of absence_7	1	TRUE
Month of absence_8	1	TRUE
Month of absence_9	1	TRUE
Month of absence_10	1	TRUE
Month of absence_11	1	TRUE
Month of absence_12	1	TRUE
Day of the week_2	1	TRUE
Day of the week_3	1	TRUE
Day of the week_4	1	TRUE
Day of the week_5	1	TRUE
Day of the week_6	1	TRUE
Seasons_1	1	TRUE
Seasons_2	1	TRUE
Seasons_3	1	TRUE
Seasons_4	1	TRUE
Education_1	1	TRUE
Education_2	1	TRUE
Education_3	1	TRUE
Education_4	1	TRUE
Son_0	1	TRUE
Son_1	1	TRUE
Son_2	1	TRUE
Son_3	1	TRUE
Son_4	1	TRUE
Reason for absence_16	2	FALSE
Reason for absence_9	3	FALSE
Reason for absence_12	4	FALSE
Reason for absence_10	5	FALSE
Reason for absence_14	6	FALSE
Reason for absence_13	7	FALSE
Reason for absence_8	8	FALSE
Reason for absence_22	9	FALSE
Reason for absence_6	10	FALSE
Reason for absence_26	11	FALSE
Reason for absence_4	12	FALSE
Reason for absence_7	13	FALSE
Reason for absence_15	14	FALSE
Reason for absence_18	15	FALSE

Feature	Feature Ranking	Selected
Reason for absence_5	16	FALSE
Reason for absence_19	17	FALSE
Reason for absence_1	18	FALSE
Reason for absence_21	19	FALSE
Reason for absence_24	20	FALSE
Reason for absence_3	21	FALSE
Reason for absence_17	22	FALSE
Reason for absence_0	23	FALSE
Pet_2	24	FALSE
Pet_1	25	FALSE
Pet_0	26	FALSE
Pet_4	27	FALSE
Pet_8	28	FALSE
Pet_5	29	FALSE
Transportation expense	30	FALSE
Age	31	FALSE
Service time	32	FALSE
Distance from Residence to	33	FALSE
Social drinker	34	FALSE
Work load Average/day	35	FALSE
Social smoker	36	FALSE
Body mass index	37	FALSE
Hit target	38	FALSE

Even though we have been able to improve score using various techniques, the R-2 score of regression models is not up to the mark. And on checking the feature importance of each feature we got to know that it is giving more importance to “Reason of absence” than any other feature. This behavior is not helping as it is giving higher priority to one and neglecting all other features.



2. Using Classification Model

Since overall score with regression turned out to low even after using various techniques to optimize the model we decided to convert this problem into a classification one.

First, we divided the absenteeism time into bins. For that we will create 4 bins:

Bin 0: <= 4 hrs
Bin 1: > 4 and <=8 hrs
Bin 2: >8 and <=24 hrs
Bin 3: > 24 hrs

We have used these custom partitions for forming beans because:

- We found that, majority of the hours are less than 24. Hence we have 3 bins below 24 and just one bin above 24
- From business perspective absenteeism can be meaningfully classified as half a day or less, almost full day, 2-3 days and more than 3 days.

We tried the following 3 models for classification:

a) Random Forest Classifier
b) Naive Bayes
c) Decision tree using GINI index

Following is a comparison of the 3 classification models in terms of accuracy and f1-score

	Random Forest Classifier	Naive Bayes	Decision Tree
Overall accuracy	0.73	0.61	0.77
<= 4 hrs f1-score	0.83	0.79	0.85
> 4 and <=8 hrs f1-score	0.64	0.29	0.72
>8 and <=24 hrs f1-score	0.15	0.22	0.00
> 24 hrs f1-score	0.40	0.15	0.00

The best accuracy we have achieved is using decision tree classifier (77%). Hence we concluded that decision tree classifier model can be used for predicting absenteeism of employees.

Conclusion

This analysis was done to sort out a method to reduce workers' absenteeism time and reduce impact of absenteeism on business and customer satisfaction.

Main factors causing absenteeism at work and few observations:

Below we have called out some of the main factors which cause absenteeism at work and called out some of the observations we have based on exploratory data analysis done:

1. Medical consultation
2. Dental consultation
3. Physiotherapy
4. Diseases of the musculoskeletal system and connective tissue
5. Injury, poisoning and certain other consequences of external causes
6. Work load – employees having very low workload and very high workload and more prone to be absent for higher number of hours. Although it is not significantly higher.
7. Hit target – those who are close to achieving their hit target are more prone to be absent then those who are farther away from their hit target.
8. According to the observed pattern, it was noticed that workload, hit target and doctor's consultations (both medical and dental) were the main reasons of remaining absent.
9. People who drink are more likely to be absent. Also there is more absenteeism on Monday because of people who drink.
10. Most of the absenteeism hours belong to category of ‘half day or less’ or full day. These 2 should be targeted mainly to improve situation on absenteeism.

Model that can help determine absenteeism hours for employees:

We have come up with GINI index based Decision Tree based classifier that helps predict whether an employee is likely to be absent for one of following 3 category of absenteeism hours with 0.77 R-2 score:

1. Half a day or less (<=4 hours)
2. Almost full day (>4 and <=8 hours)
3. 2 to 3 days (>8 and <=24 hours)
4. More than 3 days

Alternately we can use regression model to directly predict the number of absenteeism hours with 38 features with R-2 score of 0.39

Following are some of the suggestions for improving absenteeism at work:

1. Flexible schedule:
   Since medical, dental consultations are high, giving an option for a flexible schedule where the few hours lost can be compensated by either coming in early or leaving late, depending on the employee's role.
2. Employee Wellness program:

For example, the amount of Work load can lead to absenteeism due to stress. Relaxing activities during lunch breaks targeting specific muscles for a delivery employee not only reduce stress, potential health problems but also increase employee morale.
We have inferred this based on high absenteeism occurring because of Physiotherapy, Diseases of the musculoskeletal system and connective tissue and Injury, poisoning and certain other consequences of external causes

3. Hit Target:
   Employees who are close to achieving their hit target seem to be have more absenteeism hours. We can come up with alternate incentive programs so that employees who are close to achieving their hit target have some additional motivations to continue stick with their work schedules.