EMPLOYEE ABSENTEEISM

Shivam Agarwal

25/04/2019

Contents

1. **Introduction 3**
   1. Problem Description 3
   2. Problem Statement 3
   3. Data 3
   4. Performance Metric 5
2. **Methodology 6**
   1. Exploratory Data Analysis 6
      1. Data Visualization 6
         1. Univariate Analysis 6
         2. Bivariate Analysis 10
      2. Data Preparing and Cleaning 13
         1. Missing Value Analysis 13
         2. Outlier Analysis 13
         3. Feature Selection 14
      3. Modelling 15
         1. Decision Tree 15
         2. Random Forest 15
3. **Conclusion 17**
   1. Model Evaluation 17
      1. Root Mean Square Value 17
   2. Model Selection 18
   3. Answer to asked question 19

**References 22**

Chapter 1

**Introduction**

* 1. **Problem Description**

Employee Absenteeism is the absence of an employee from work. It’s 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.

Absenteeism can be of two types:

• **Innocent absenteeism** - Is one in which the employee is absent from work due to genuine cause or reason. It may be due to his illness or personal family problem or any other real reason.

• **Culpable Absenteeism** - is one in which a person is absent from work without any genuine reason or cause. He may be pretending to be ill or just wanted a holiday and stay at home.

**1.2. 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?

**1.3. Data**

The data is a Time-Series data but instead we will approach it as Regression Problem. Our task is to build a regression model which will predict the absenteeism in hours per employee based on the employee attributes and information in their work place and general information available to the company about them.

Table 1.1: Employee Attribute (Columns 1-8)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Reason for Absence | Day of Month | Day of the week | Seasons | Transportation Expenses | Distance From Residence to work | Service Time |
| 11 | 26 | 7 | 3 | 1 | 289 | 36 | 13 |
| 36 | 0 | 7 | 3 | 1 | 118 | 13 | 18 |
| 3 | 23 | 7 | 4 | 1 | 179 | 51 | 18 |
| 7 | 7 | 7 | 5 | 1 | 279 | 5 | 14 |
| 11 | 27 | 7 | 5 | 1 | 289 | 36 | 13 |

Table 1.2: Employee Attribute (Columns 9-16)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Work load Average/day | Hit target | Disciplinary failure | Education | Son | Social drinker | Social smoker |
| 33 | 239554 | 97 | 0 | 1 | 2 | 1 | 0 |
| 50 | 239554 | 97 | 1 | 1 | 1 | 1 | 0 |
| 38 | 239554 | 97 | 0 | 1 | 0 | 1 | 0 |
| 39 | 239554 | 97 | 0 | 1 | 2 | 1 | 1 |
| 33 | 239554 | 97 | 0 | 1 | 2 | 1 | 0 |

Table 1.3: Employee Attribute (Columns 17-21)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pet | Weight | Height | Body mass Index | Absenteeism time in hours |
| 1 | 90 | 172 | 30 | 4 |
| 0 | 98 | 178 | 31 | 0 |
| 0 | 89 | 170 | 31 | 2 |
| 0 | 68 | 168 | 24 | 4 |
| 1 | 90 | 172 | 30 | 2 |

As we can see in the table above, we have we have the following 20 variables, using which we have to correctly predict the ‘Absenteeism time in hours ‘ for the employees.

Table 1.4: Predictor Variables

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

**1.4 Performance Metric**

RMSE : Root Mean Square Error (RMSE) 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.

Also, since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. So, RMSE becomes more useful when large errors are particularly undesirable. So, Roost Mean Square value seems like a perfect choice for our problem at hand.

Chapter 2

**Methodology**

**2.1 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is the first step in our data analysis process. We do this by taking a broad look at patterns, trends, outliers, unexpected results and so on in our existing data, using visual and quantitative methods to get a sense of the story this tells. To start with this process, we will first have a look at univariate analysis like plotting Box plot and whiskers for individual features, Histogram plots, Bar plots and Kernel Density Estimation for the same for the same. Then we will proceed to Multivariate analysis like Bar and Histogram and Bar plot using group-by function and Pivot table for the features with respect to the target variable.

**2.1.1. Data Visualization**

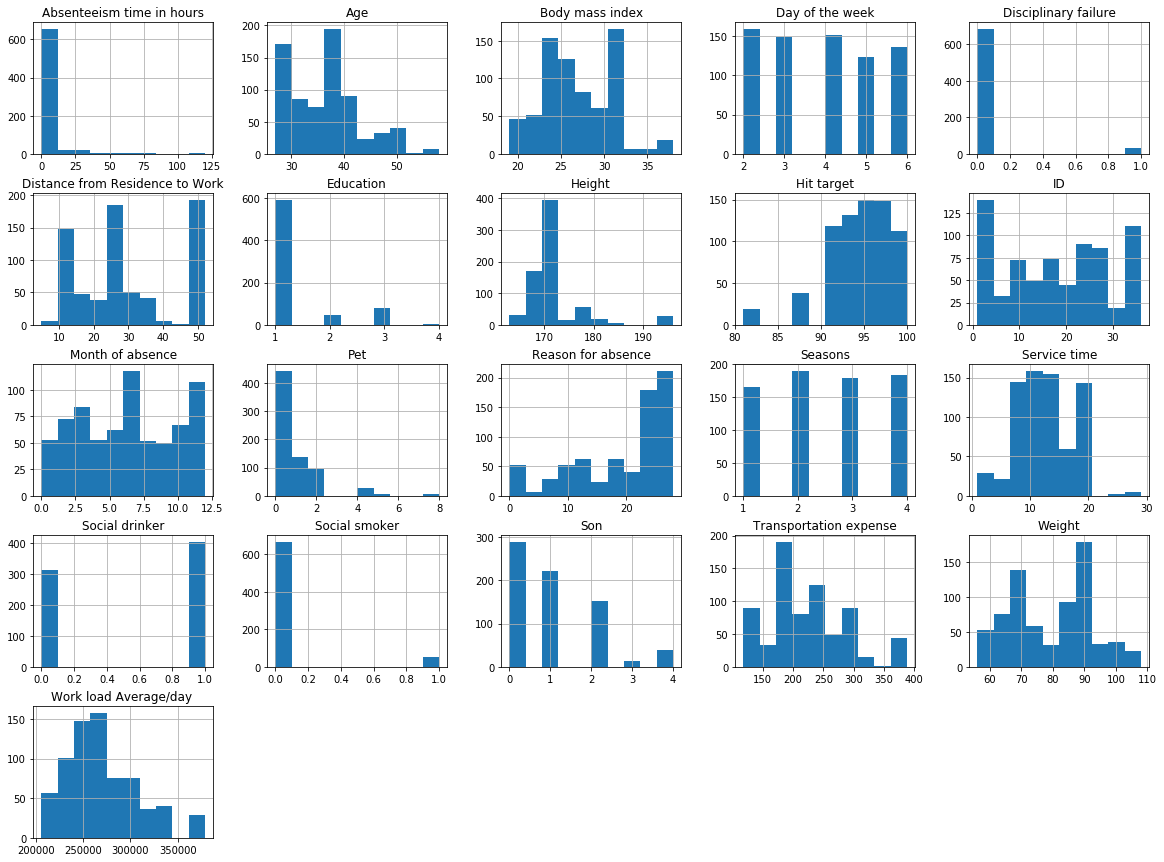
Data visualisation helps us to get better insights of the data. By visualising data, we can identify an area that need attention or improvement and also clarifies which factors influence target behaviour and how the resources are used by the target.

**2.1.1.1 Univariate Analysis**

Univariate analysis is the simplest form of data analysis where the data being analysed contains only one variable. Since it's a single variable it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

So, Let’s have a look at histogram plot, to identify the characteristic of the features and the data.

Figure 2.1 Histogram plot for distribution of features in the data



Histograms are constructed by binning the data and counting the number of observations in each bin. The objective of plotting Histogram plot is usually to visualise the shape of the distribution. The number of bins needs to be large enough to reveal interesting features and small enough not to be too noisy.

From the above histogram plot, we can clearly observe that none of the features in our data are actually skewed. Although feature like ‘work load average/day’ seems like it is right skewed a little. Also if observed properly, It is worth noting the following points:

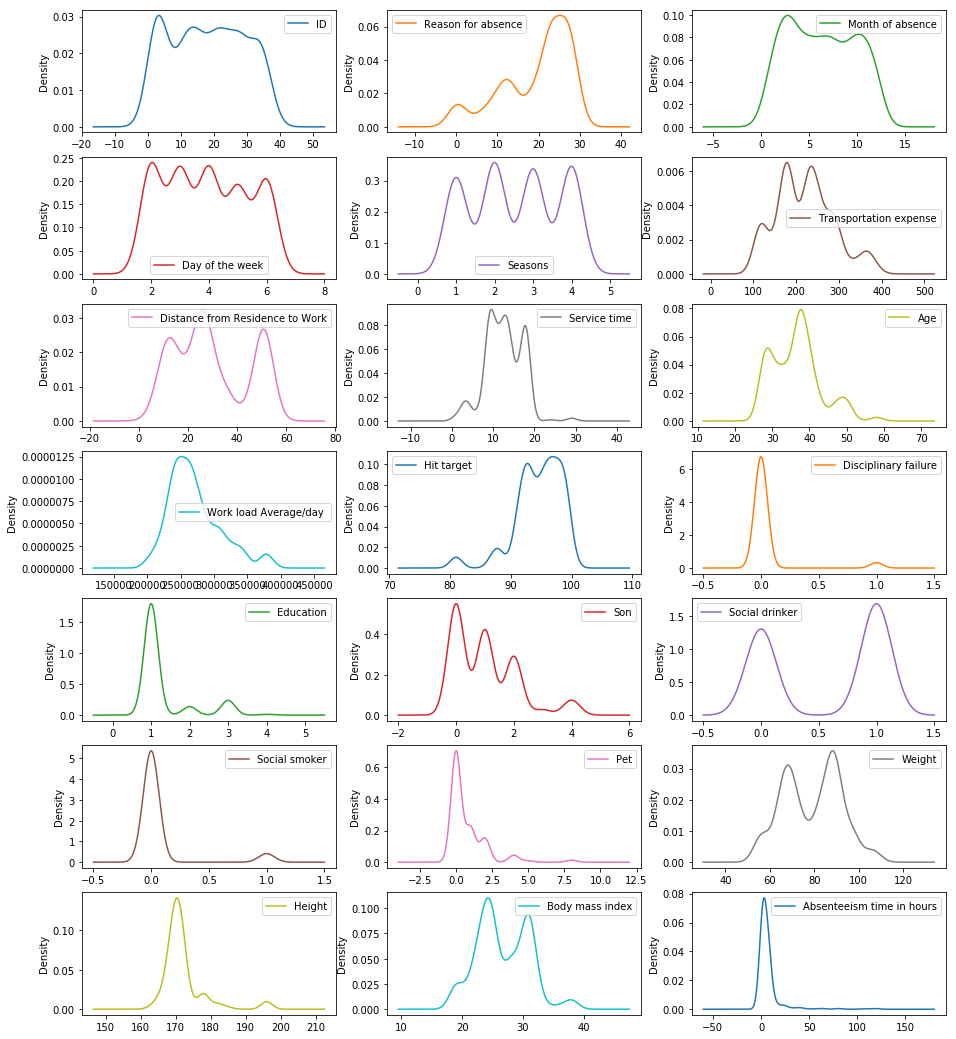
1. Majority of the employees working in the company have age below 40 years.

2. A very large portion of the population has only passed ‘High School’.

3. More than half of the employees in the company are ‘social drinker’.

4. Only a very few portion of the employees in the company are ‘social smoker’.

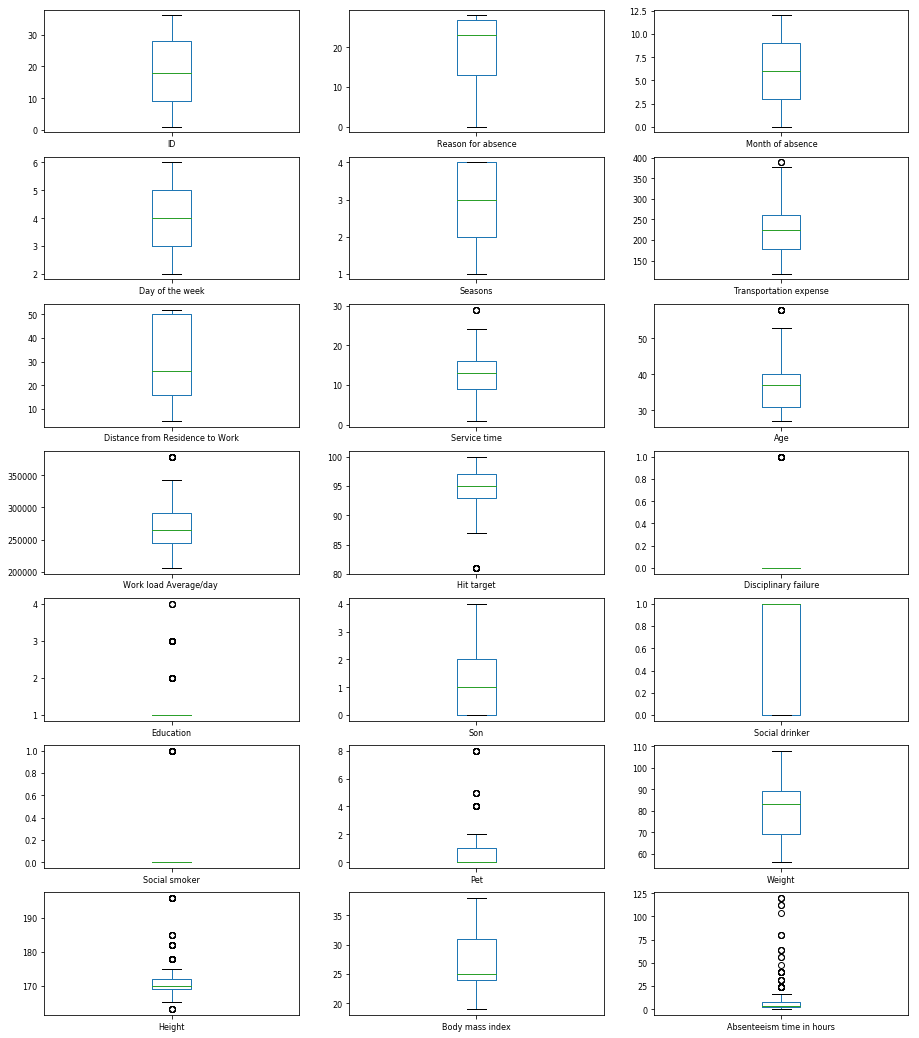
Figure 2.2: KDE plot for distribution of features in the data.



A Density Plot visualises the distribution of data over a continuous interval or time period. Density plots can be thought of as plots of smoothed histograms. An advantage Density Plots have over Histograms is that they're better at determining the distribution shape because they're not affected by the number of bins used.

So, looking at the above density plot, we can observe that none of the features follow Gaussian distribution. Few of the features like ‘Disciplinary failure’, ‘Social smoker’, ‘Work load average/day’ seems to follow Gaussian distribution at first sight but they either have long tail at the left or right or they are either jiggered at the end.

Figure 2.3: Box plot of features in the data.



From the above Box and whisker plots, we can observe that not all the features contain outliers. Continuous features like ‘Weight’, ‘Distance from residence to work’ does not contain any outliers at all. Few features like ‘work load average/day’, ‘Hit target’ and ‘Height’ have a very few outliers.

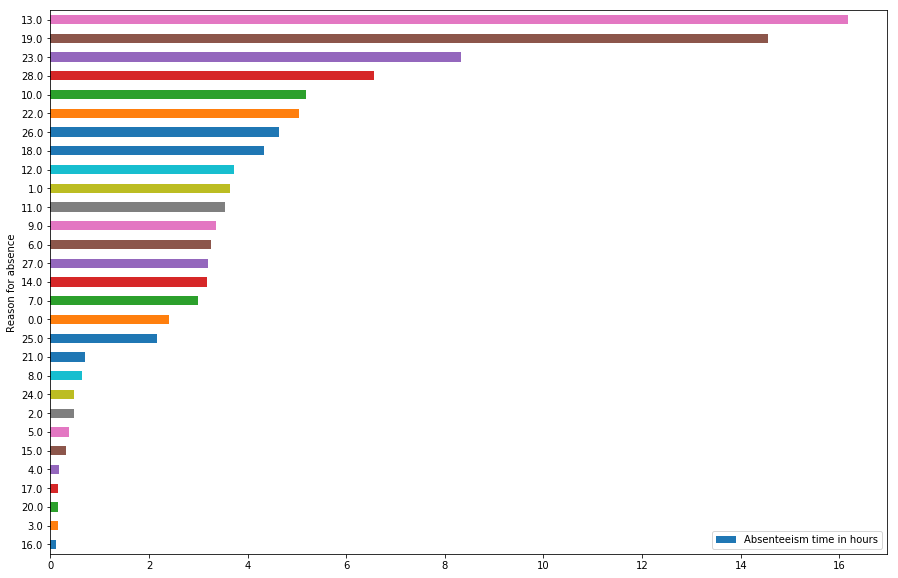
It is also evident from the above plot that none of the features are symmetric to the median and it can easily be interpreted that none of the features follow symmetric distribution. Also, it can also be observed that Median of the feature ‘Body mass index’ is very close to 25th percentile value which means median of this feature is almost equal to 25th percentile.

**2.1.1.2 Bivariate Analysis**

Bivariate analysis refers to the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves one predictor variable and one target variable.

So, let’s have a look at the Histogram and Bar Plots to understand the Employee behaviour better.

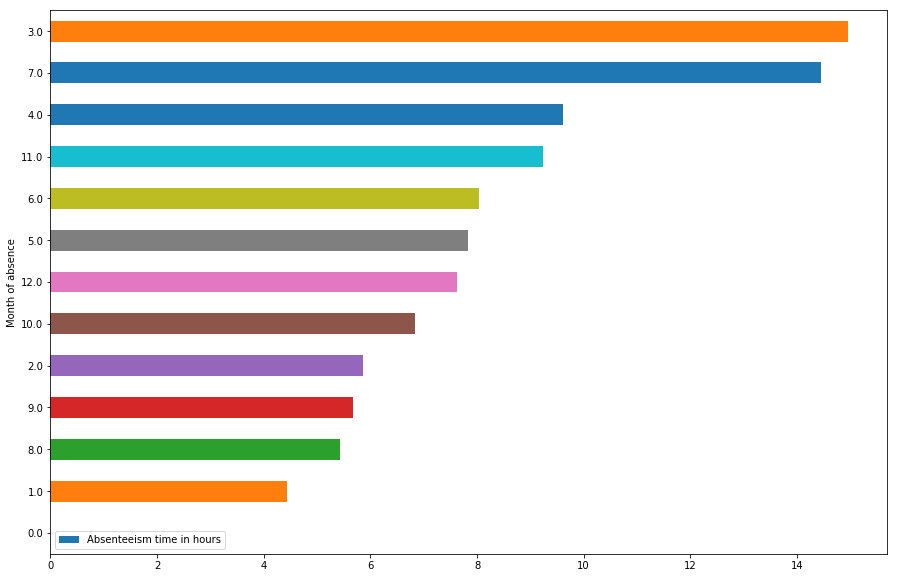
Figure 2.4: Bar graph showing relation between the Reason for absence and Absenteeism time

****

Top 3 categories in order of Absenteeism time are:

1. Category 13: Diseases of the musculoskeletal system and connective tissue - 16.67 % of total time.
2. Category 19: Injury, poisoning and certain other consequences of external causes - 14.55 % of total time.
3. Category 23: Medical consultation - 8.32 % of total time.
4. Category 28: Dental consultation - 6.56 % 0f total time.

Figure 2.5: Bar graph showing Employees absenteeism in different month



Top 3 months in order of Absenteeism time are:

1. Month 3:March - 14.95 % of total time
2. Month 7:July - 14.45 % of total time
3. Month 4:April - 9.62 % of total time

Figure 2.6: Bar graph showing Total absenteeism in different days of the week

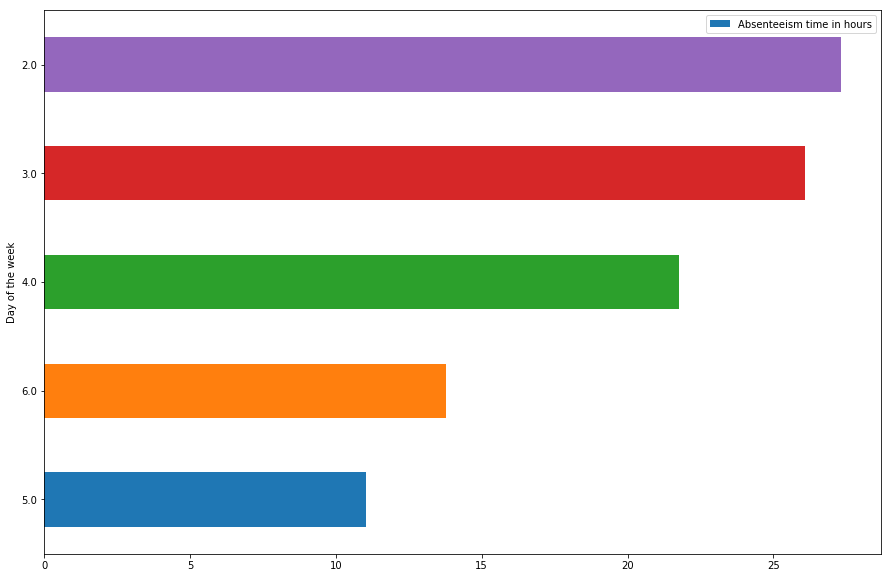


Figure 2.7: Bar graph showing Relation between education and Absenteeism time



Employees having qualification till high school (1) are mostly absent from their work.

**2.1.2 Data Preparation and Cleaning**

**2.1.2.1 Missing Value Analysis**

One of the most common problems I have faced in Data Cleaning/Exploratory Data Analysis is handling the missing values. Firstly, there is no good way to deal with missing data. But still missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

So, in our data, there are plenty of missing values available in different variables. So, after computing the percentage of missing data that is available to us in the dataset, it accounts to around 12% of the data. It is also important to note that, the missing value has been calculated after removing the missing values within the target variable. Also, as we have very less data available to us, we impute the missing values in other columns using KNN imputation, because that fits the best after trying various other imputation techniques like Mean, Median and Random value imputation.

**2.1.2.2 Outlier Analysis**

In statistics, an outlier is an observation point that is distant from other observations. In layman terms; we can say that an outlier is something which is separated from the crowd. Also, Outlier analysis is very important because they affect the mean and median which in turn affects the error (absolute and mean) in any data set. When we plot the error we might get big deviations if outliers are in the data set.

In Box plots analysis of individual features, we can clearly observe from these box plots that, not every feature contains outliers and many of them even have very few outliers. Also, given the constraint that, we have only 649 data-points and after removing the outliers, the data gets decreased by almost 25%. So, dropping the outliers is probably not the best idea.

Instead we will try to visualise and find out the outliers using box plots and will fill them with NA that means we have created ‘missing values’ in place of outliers within the data. Now, we can treat these outliers like missing values and impute them using standard imputation techniques. In our case, we use KNN imputation to impute these missing values.

**2.1.2.3 Feature Selection**

We have two types of variables, continuous and categorical. To select a significant continuous variable we observe the correlation them and discard the highly correlated feature.

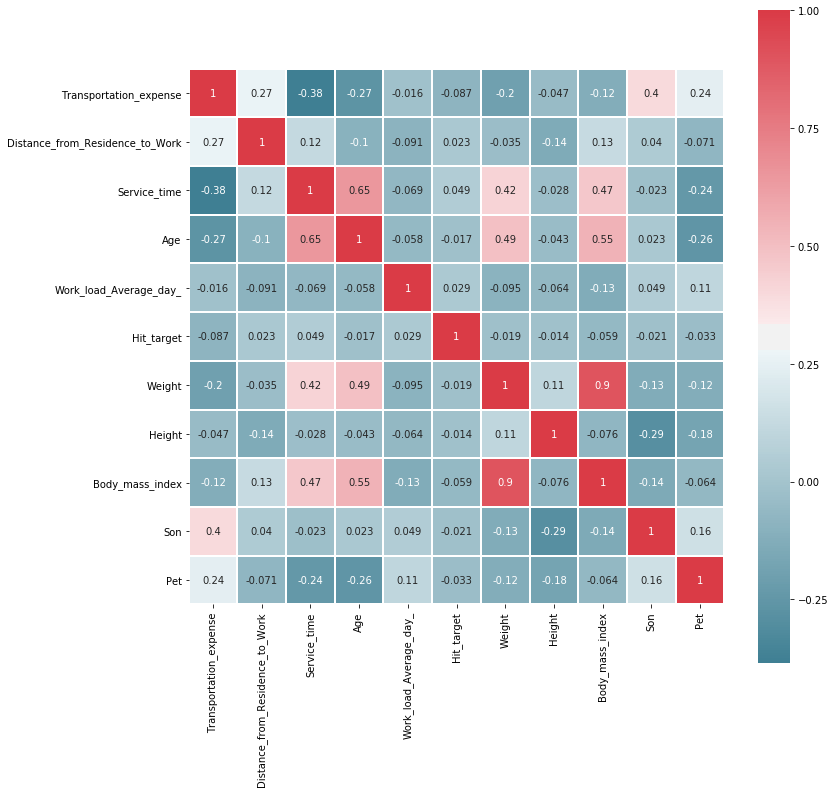


Figure 2.8: Heat map showing the correlation between different continuous variables.

From above figure we observe that “Weight” and “Body\_mass\_Index” are highly correlated so we drop one of the features before model development.

Now for categorical variable we used ANOVA test and check the p-value of each variable should be less than 0.05 to be significant.

Table2.1: Shows the result of ANOVA test for categorical variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **sum\_sq** | **df** | **F** | **PR(>F)** |
| **Seasons** | 0.605689 | 1.0 | 0.003654 | 9.518160e-01 |
| **Month\_of\_absence** | 46.330299 | 1.0 | 0.279496 | 5.971970e-01 |
| **Social\_drinker** | 883.489078 | 1.0 | 5.329816 | 2.125037e-02 |
| **Reason\_for\_absence** | 9210.794529 | 1.0 | 55.565868 | 2.641196e-13 |
| **Day\_of\_the\_week** | 906.673436 | 1.0 | 5.469680 | 1.962609e-02 |
| **Disciplinary\_failure** | 4932.016177 | 1.0 | 29.753325 | 6.783889e-08 |
| **Education** | 121.682941 | 1.0 | 0.734075 | 3.918548e-01 |
| **Social\_smoker** | 255.650456 | 1.0 | 1.542260 | 2.146924e-01 |
| **Residual** | 117526.343920 | 709.0 | NaN | NaN |

From the above table, we observe the variable ('Seasons', 'Month\_of\_absence', 'Social\_smoker', 'Education', 'Day\_of\_the\_week') are insignificant.

Finally after dropping insignificant variable we left with 14 independent variables to train a model.

**2.1.3 Modelling**

**2.1.3.1 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.

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

.So, after we implement Decision Tree Regression on our data, we get the following results:

**Decision Tree RMSE: 0.10140833228255668**

**(RMSE, when max\_depth=1)**

**2.1.3.2 Random Forest**

Random Forest Regression or Regression Trees are known to be very unstable, in other words, a small change in our data may drastically change your model. The Random Forest uses this instability as an advantage through bagging resulting on a very stable model. So, after we implement Random Forest Regression or Regression Trees on our data, we get the following results:

**depth - 100 -- n\_estimators - 2 RMSE : 0.10106265300370718**

Chapter 3

**Conclusion**

**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 any of the following criteria:

1. Predictive Performance

2. Interpretability

3. Computational Efficiency

In our case of Employee Absenteeism, the latter two, Interpretability and Computation Efficiency, do not hold much significance. Therefore we will use Predictive performance as the criteria to compare and evaluate models.

Predictive performance can be measured by comparing Predictions of the models with real values of the target variables, and calculating some average error measure.

**3.1.1 Root Mean Square Error**

Root Mean Square Error (RMSE) 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.

Also, Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. So, RMSE becomes more useful when large errors are particularly undesirable. So, Roost Mean Square value seems like a perfect choice for our problem at hand.

**3.2 Model Selection**

We saw both the model Decision tree and Random Forest perform comparatively on RMSE (Root Mean Square Error) ,Although Decision Tree gives the best result.

Table 3.1: Model Performance table

|  | **Model** | **RMSE** |
| --- | --- | --- |
| **0** | Decision Tree Regression | 0.10140 |
| **1** | Random Forest Regression | 0.10106 |

So , it is obvious from above model performance table, both Decision Tree and Random Forest perform good but we go with Decision Tree.

**3.3 Answer to asked question**

**Q1. What changes company should bring to reduce the number of absenteeism?**

**A1.** Looking at the Exploratory data analysis of the features, we observe and make following conclusion :

1. The rate of Absenteeism is maximum in Season 3 : Winter followed by Season 1 : Summer , Season 4 : Spring, Season 2 : Autumn.

2. Also, We can say that the ‘Absenteeism rate’ is maximum in Month 7 : July followed by Month 4 : April , Month 3 : March, Month 12 : December, Month 11 : November , Month 6 : June , Month 5 : May etc.

3. Looking at the Bar plot of ‘Absenteeism rate’ Vs ‘Day of the week’, it can clearly be observed that the ‘Absenteeism rate’ is maximum on the third day of the week i.e Day 3 : Tuesday followed by Day 2 : Monday, Day 4 : Wednesday. Also, the ‘absenteeism rate’ is lowest on Day 6 : Saturday followed by Day 5 : Friday.

4. From the Bar plot of ‘Absenteeism rate’ Vs ‘Reason of absence’ we can observe that ‘9 : Diseases of the circulatory system’ is the most frequent reason for the absence of the employees. The second most frequent reason given by the employees for their absence is ‘2 : Neoplasms’ followed by ‘6 : Diseases of the nervous system’, ‘12: Diseases of the skin and subcutaneous tissue’, ’19 : Injury, poisoning and certain other consequences of external causes’ etc.

5. Looking at the Bar plot of ‘Absenteeism rate’ Vs ‘ID’ It can be observed that the absence rate is maximum for employee with ID : 9, followed by employees with ID : 7,26,14,13, 36, 11 and 6. Also, it can be observed that Employee with employee ID : 4,8 and 35 never absents and are very much regular to work.

So, Now that we have understood the behaviour of Employee attributes against their Absenteeism rate in the company, we can introduce following changes to reduce the number of Absenteeism :

1. Firstly, we can start by Increasing the employee morale, engagement, and commitment to the organisation.

2. We can set a certain threshold for minimum number of absence for employees during the workdays, and employees not meeting the criteria can be questioned.

3. As Absenteeism rate is maximum in Season of ‘Winter’ and in month of July, April and March respectively, We can issue special notices regarding the Absenteeism scenario around the company. 4. A Health care facility can be introduced in the company, so that the employees can have regular Medical check-ups to keep them fit and Working. Also, It would help with the company’s reputation to have taken the responsibility of their Employees.

5. Also, we can introduce person to person phone calls if off sick and return to work interviews. This way, Employee would feel responsible for their action towards the company goal and achievements. 6. We can also we can come up with other ideas like : An incentive or conversion scheme for unused sick days.

7. Also, Strict action could be taken towards the employee with high absence rate in the workplace without any valid reason for the absence and Employees with no absence or a minimum absence can be rewarded with perks.

8. Lastly , we can apply performance policies to act at the root of the problem. In some cases, absence rate might be reduced by clear specification of employees’ responsibilities and targets.

**Q2. How much losses every month can we project in 2011 if same trend of absenteeism continues?**

**A2.** Employee absence, whether caused by sickness or pressures outside of the workplace, can cost employers a large amount of money if not properly managed.

Lowering absence levels across a business not only leads to a reduction in money being lost by the business, but also a happier, productive and content workforce.

To calculate Loss per month, We introduce the following formula :

Loss = (Work load average/day \*Absenteeism time in hours )/Service Time

So, below chart represents loss per month and more likely, the same trend could follow in 2011 :

|  |  |
| --- | --- |
| Month | Loss Per Month |
| 0 | 0 |
| January | 6367727 |
| February | 8268542 |
| March | 15707449 |
| April | 10999489 |
| May | 9326392 |
| June | 14362241 |
| July | 19015383 |
| August | 8791557 |
| September | 6482816 |
| October | 8896045 |
| November | 12255373 |
| December | 12349895 |

Looking at the above results, we can observe that most likely, the company would incur most of loss in the month of 'July', followed by 'March' and 'June'.

**References**