Project Report 1 Edwisor

Shrikrishna Padalkar

October 2018

1 Splitting The Data

We have split the data into training and test set, where 80% observations lie in training set and the rest in testing set.

2 Handling Missing Values

There are 3 main types of Missing Values, they are as follows:-

- Missing Completely At Random(MCAR)
- Missing At Random(MAR)
- Missing Not At Random(MNAR)

1. Missing Completely At Random(MCAR):-

   We call missing values as MCAR if there is no any relationship between the missing values and the non missing values. These are missing at random.

2. Missing At Random(MAR):-

   MAR means that the missing values have some kind of association with the non missing entries. Viz. Women are less likely to enter their weights than men. Similarly, men are less likely to enter their salary details than their women counterparts.

3. Missing Not At Random(MNAR):-

   MNAR suggests that missing values are neither MAR nor MCAR.

3 Evaluate the Model Performance

Evaluation of the model performance is carried to check the model performance on new, unseen data. We have used "K-fold Cross Validation" in order to test the model. Following are the steps being carried out for evaluation:-

1. The **Training data** is divided into K(here 10) mutually exclusive (non-overlapping) folds.
2. The first is kept hidden from the rest of the K-1 folds. Call it test fold.
3. Training is performed on k-1 folds combined as one single set. Model is fit on this aggregation of K-1 folds.
4. The model's performance is evaluated on the test fold.
5. The steps 2-4 are repeated for all the remaining K-1 (here 9) folds.

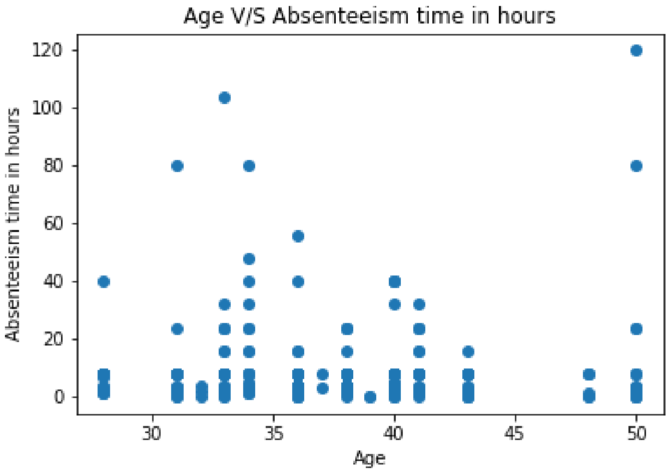
Error is calculated after fitting the model for each fold. The net error is the average of the errors obtained from each fold.

4 Observations

Following are a few observations based on the relationship between the variables and the trends among particular categories.

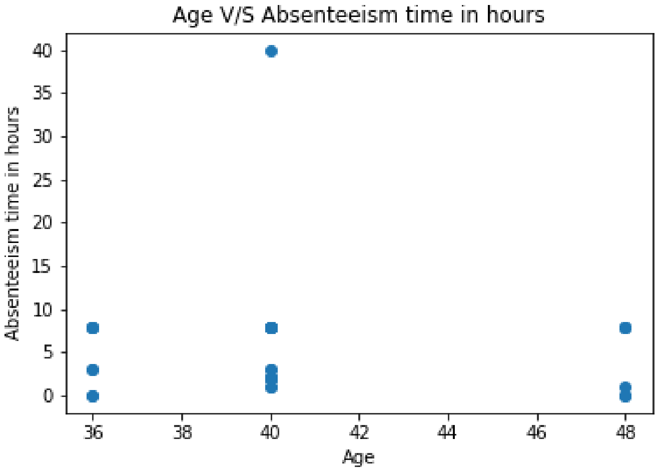
- Analysis of Smokers, Drinkers and Non Addict
  - Those who smoke or drink remain absent for more than those who are non addict. The below graph explains the same.

or Drinkers Versus Absenteeism Hours.png or Drinkers Versus Absen-

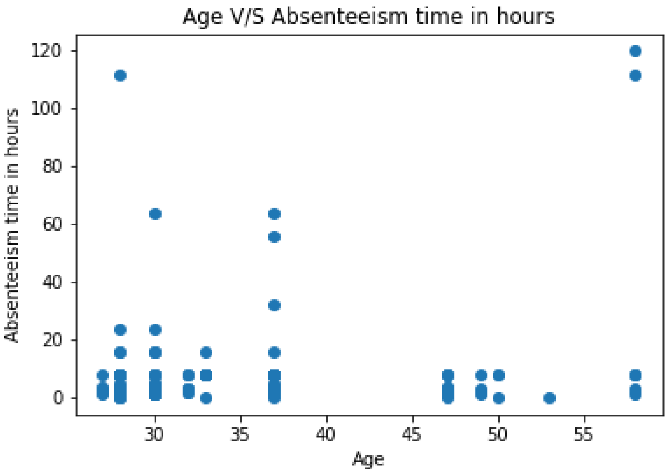


teeism Hours.png

- Those who are exclusive smokers(not drinkers) remain less absent than those who are exclusive drinkers(not smokers). Smokers Versus Absnteeism Hours.png Smokers Versus Absnteeism Hours.png



- The most frequent reason for absence among drinkers is dental consultation (28) followed by medical consultation (23) then physiotherapy (27).
- Refraining from Alcohol will help more in reducing the Absenteeism as compared to preventing smoking. Hours for Non Addict.png Hours



for Non Addict.png

- Analyzing Reason for Absence

  Most of the absentees are below 10 hours. Maximum hours of absenteeism occur due Medical Consultation.

- **Relationship Between Distance of work from Residence and Hours of Absenteeism**

  There data comprises of employees staying as close as less than 10 Km to as far as move 50 Kms. We created bins of employees. The bins are **Below 10KM**, **Between 10 and 20**, **Between 20 and 40**, **Between 40 and 50**, **Above 50**.
  The average hours of absenteeism for each of these groups was calculated. The results show an approximately negative relationship between the distance from residence and the absenteeism. Those staying near to the workplace remained absent the more than those who stay far. The average hours of Absenteeism for the groups is as follows:-
  Below 10KM —- 25.0hrs.
  Between 10 and 20 —- 9.2hrs.
  Between 20 and 40 —- 6.49hrs.
  Between 40 and 50 —- 4.21hrs.
  Above 50 —- 5.23hrs.

- **Relationship Between Education and Absenteeism**
  People highly educated remain absent the least whereas those who are mere high school pass outs remain absent the most. Highly educated people take better care of themselves.

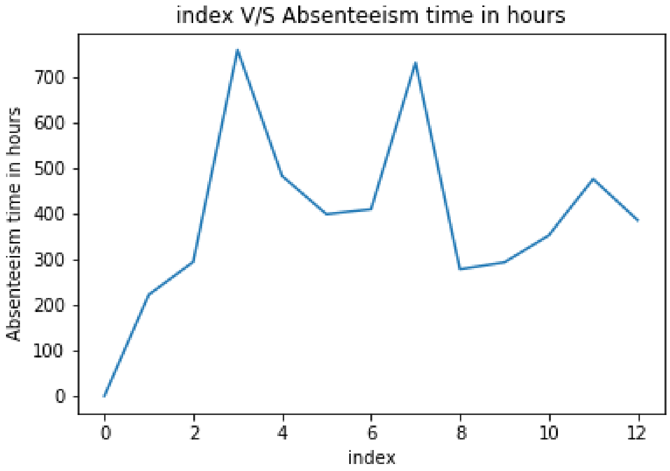
5 Time Series Analysis

Dickey Fuller Test:-

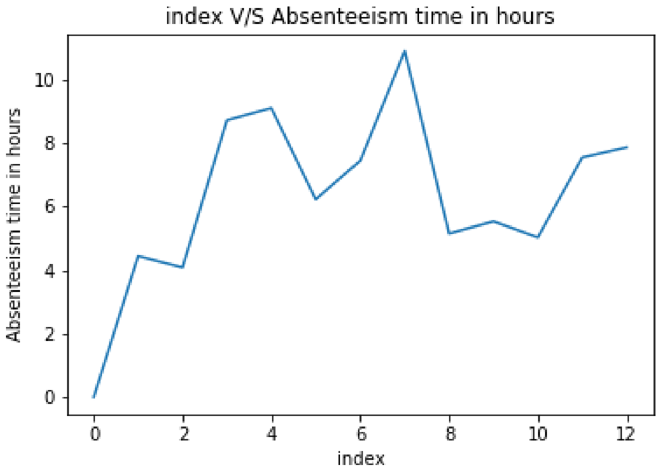
y_t = β_0+β_1.y$_($$t-1$)+u$_t$
y_t - $y_($$t-1$)

We have converted the data in a time series format. Two data frames are created, one Month V/S Average Hours of Absenteeism and the other Month V/S Total Hours of Absenteeism.
The following are the observations from the data:-

- Maximum **total** hours of absenteeism were recorded in the month of March and the minimum in the month of January.
  Hours of Absenteeism Per Month.png Hours of Absenteeism Per Month.png



- Maximum **average** hours of absenteeism were recorded in the month of July and the minimum in the month of January.
  Hours of Absenteeism Per Month.png Hours of Absenteeism Per Month.png



If the same trend continues in 2011 then we can project a loss of 423 hrs. per month.

6 Multiple Regression Techniques

A few regression techniques have been tried on this data. We have split the data into training and testing sets with 80% observations in the training set and rest 20% in the testing set. We have used two methods for splitting the data, one is using the inbuilt split function with the random seed set to 42. While the other is splitting manually, top 20% observations placed in the test set and the rest in the training set. We have applied the various supervised Machine Learning algorithms on the data.

Algorithm	Train RMSE	Train R-Squared Error	Test RMSE	Test R-Squared Error
Ridge	13.21	0.12	9.33	-0.01
Lasso	13.63	0.06	9.18	0.02
Linear	13.21	0.12	9.33	-0.010
SVR	14.09	0.00	9.54	-0.05
Random Forest	6.99	0.75	11.80	-0.61

Table 1: Splitting Data Manually