**EMPLOYEE ABSENTEEISM**

**ASIF RAZA**

**13TH June, 2018**

**Table of Contents**

|  |  |
| --- | --- |
| **1. Introduction** | **3** |
| 2. Data | **3** |
| 2.1 Attribute Information | **3** |
| **3. Pre- processing** | **4** |
| 3.1 Exploration of Numerical variables | **4** |
| 3.2 Checking the distribution of the variables | **4** |
| 3.3 Missing Value analysis | **7** |
| 3.4 Missing value percentage | **9** |
| 3.5 Checking the outliers | **11** |
| 3.6 Boxplot of all numerical variable with dependent or target variables | **11** |
| 3.7 Outliers Treatment | **13** |
| 3.8 Feature Selection | **14** |
| 3.9 Correlation Plot | **14** |
| 3.10 Chi-Square test | **15** |
| 3.11 Feature Scaling | **16** |
| **4. Building Predictive model** | **16** |
| **4.1 Linear regression Model** | **16** |
| 4.1.1 Summary of the model | **19** |
| 4.1.2 Model Performance or Evaluation for linear regression model | **21** |
| **4.2 Decision Tree Regression** | **21** |
| 4.2.1 Performance of model | **22** |
| 4.2.2 Model Performance | **23** |
| **4.3 Random forest Algorithm** | **23** |
| 4.3.1 Performance of the model | **24** |
| **4.4 Model Selection** | **24** |
| **5. Measure for control Absenteeism** | **24** |
| **6. Suggestion & Recommendation** | **24** |
| **7. Conclusion** | **25** |
| **8. Complete R-Code** | **25** |

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. **Data**

Our Task is to build a regression model to reduce the employee absenteeism by multiple characteristics. Data contain dependent variable hour, so it is **time series** multi-variant problem statement.

As we have 1 data source which contain 21 variables and 740 observations. First we used these dataset for exploratory data analysis then used data for building models and then we analyzed to find the solution to reduce the number of absenteeism.

**2.1 Attribute Information**

1. Individual identification (ID)

2. Reason for absence (ICD).

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)

As you can see in the table, we have the following 21 variables, using this we can correctly predict the loss of revenue in every month if the same trend continues.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | | Reason for absence | | | Month of absence | | | Day of the week | | Seasons |
| 11 | | 26 | | | 7 | | | 3 | | 1 |
| 36 | | 0 | | | 7 | | | 3 | | 1 |
| 3 | | 23 | | | 7 | | | 4 | | 1 |
| 7 | | 7 | | | 7 | | | 5 | | 1 |
| 11 | | 23 | | | 7 | | | 5 | | 1 |
| 3 | | 23 | | | 7 | | | 6 | | 1 |
| Transportation expense | | **Distance from Residence to Work** | | | **Service time** | | | **Age** | | **Work load Average/day** |
| 289 | | 36 | | | 13 | | | 33 | | 239,554 |
| 118 | | 13 | | | 18 | | | 50 | | 239,554 |
| 179 | | 51 | | | 18 | | | 38 | | 239,554 |
| 279 | | 5 | | | 14 | | | 39 | | 239,554 |
| 289 | | 36 | | | 13 | | | 33 | | 239,554 |
| 179 | | 51 | | | 18 | | | 38 | | 239,554 |
|  | |  | | |  | | |  | |  |
| Hit target | | **Disciplinary failure** | | | **Education** | | | **Son** | | **Social drinker** |
| 97 | | 0 | | | 1 | | | 2 | | 1 |
| 97 | | 1 | | | 1 | | | 1 | | 1 |
| 97 | | 0 | | | 1 | | | 0 | | 1 |
| 97 | | 0 | | | 1 | | | 2 | | 1 |
| 97 | | 0 | | | 1 | | | 2 | | 1 |
| 97 | | 0 | | | 1 | | | 0 | | 1 |
|  | |  | | |  | | |  | |  |
| Social smoker | **Pet** | **Weight** | | **Height** | | | **Body mass index** | | **Absenteeism time in hours** | |
| 0 | 1 | 90 | 172 | | | 30 | | | 4 | |
| 0 | 0 | 98 | 178 | | | 31 | | | 0 | |
| 0 | 0 | 89 | 170 | | | 31 | | | 2 | |
| 1 | 0 | 68 | 168 | | | 24 | | | 4 | |
| 0 | 1 | 90 | 172 | | | 30 | | | 2 | |
| 0 | 0 | 89 | 170 | | | 31 | | |  | |

1. **Pre-Processing**

**3.1 Exploration of Numerical Variable**:

After getting the data altogether it has dimension of 21 variables and 740 observations. Out of 21 variables 7 were categorical variable and 14 were numerical variable, out of which one is dependent variable which is in the form of numerical value i.e. absenteeism time in hours.

**3.2 Checking the distribution of the variables**:

Histogram was plotted to check the distribution of the variables.

|  |  |  |
| --- | --- | --- |
| age.jpeg | |  |
| body mass index.jpeg | distance from residence to work.jpeg | |
| height.jpeg | hit target.jpeg | |
| month of absence.jpeg | pet.jpeg | |
| service time.jpeg | son.jpeg | |
| Transportation expences.jpeg | weight.jpeg | |

All the variable are either left or right skewed. We can clearly observe from these probability distribution that almost **all the variables** are **skewed**. The skew in these distributions can be mostly explained by the presence of outliers and extreme value of the data.

**3.3 Missing value Analysis:-**

After doing proper analysis, we observed that there are121 values missing in the dataset which is shown below in the form of table where we have observed how many missing values are there against respective variable.

|  |  |
| --- | --- |
| **S.No** | **Missing Value** | **Variables** |
| **1** | 0 | ID |
| **2** | 3 | Reason.for.absence |
| **3** | 1 | Month.of.absence |
| **4** | 0 | Day.of.the.week |
| **5** | 0 | Seasons |
| **6** | 7 | Transportation.expense |
| **7** | 3 | Distance.from.Residence.to.Work |
| **8** | 3 | Service.time |
| **9** | 3 | Age |
| **10** | 0 | Work.load.Average.day |
| **11** | 6 | Hit.target |
| **12** | 6 | Disciplinary.failure |
| **13** | 10 | Education |
| **14** | 6 | Son |
| **15** | 3 | Social.drinker |
| **16** | 4 | Social.smoker |
| **17** | 2 | Pet |
| **18** | 1 | Weight |
| **19** | 14 | Height |
| **20** | 31 | Body.mass.index |
| **21** | 22 | Absenteeism.time.in.hours |

* 1. **Missing value percentage**:-

Here we convert the missing value in the form of missing percentage and observe that we have less than 30% missing value in all the variable then we can impute the value by mean, Median or Imputation method. Here we have shown in the table.

| **S.No** | **Missing-percentage** | **Columns** |
| --- | --- | --- |
|  |  |  |
| **1** | 0.0000000 | ID |
| **2** | 0.4054054 | Reason.for.absence |
| **3** | 0.1351351 | Month.of.absence |
| **4** | 0.0000000 | Day.of.the.week |
| **5** | 0.0000000 | Seasons |
| **6** | 0.9459459 | Transportation.expense |
| **7** | 0.4054054 | Distance.from.Residence.to.Work |
| **8** | 0.4054054 | Service.time |
| **9** | 0.4054054 | Age |
| **10** | 0.0000000 | Work.load.Average.day |
| **11** | 0.8108108 | Hit.target |
| **12** | 0.8108108 | Disciplinary.failure |
| **13** | 1.3513514 | Education |
| **14** | 0.8108108 | Son |
| **15** | 0.4054054 | Social.drinker |
| **16** | 0.5405405 | Social.smoker |
| **17** | 0.2702703 | Pet |
| **18** | 0.1351351 | Weight |
| **19** | 1.8918919 | Height |
| **20** | 4.1891892 | Body.mass.index |
| **21** | 2.9729730 | Absenteeism.time.in.hours |

Now our task is to impute missing value either by Mean, Median or KNN imputation method.

For selecting proper method for imputing value we do a small experiment that is we delete 20 observation from dataset and find its mean, median and KNN imputation and see which value is closest to the actual value.

After doing proper experiment, we foundthat KNN imputation is closest to the actual value so we fix KNN imputation for missing value analysis.

**3.5 Checking for outliers**:

We can clearly observe from above probability distribution(Histogram Plot) that most of the variable are skewed. The skew of these distributions can be most likely to be explained by the outliers and extreme value of the data.

Boxplot was used to check the presence of outliers. It is quite obvious from the box plot that **Transportation Expenses, Service Time, Age, Pet, hit target, Height** has outliers. This is clearly the effect of outliers and extreme values.

One of the other steps of pre processing apart from checking for normality is the presence of outliers. In this case we use a classic approach of removing outliers i.e. detect and delete the outliers. We visualize the outliers using boxplot.

We have plot the boxplots of the 12 variable with respect to dependent variable. A lot of useful interference can be made from the plot. First we have seen we have few outliers and extreme value in each of the data set. We have observe that in “ID, Month of absence, Distance from residence, son, weight, body mass index” these variable did not have outliers.

**3.6 Boxplot of all numerical variable with dependent or target variable**

|  |
| --- |
| box new 1.jpeg |
| box new 2.jpeg |
| hit target box.jpeg |
| box new 3.jpeg |

It can be seen from the box plot that **transportation expenses, service time, age, pet, height** has outliers but we are observe the outliers in between numerical variable and target value.

We will process these data and plot it again after outlier removal.

**3.7 Outlier Treatment:**

After performing outliers analysis and removing outliers using boxplot command line method or detect and delete the outliers. We plot a variable again to check the outliers.

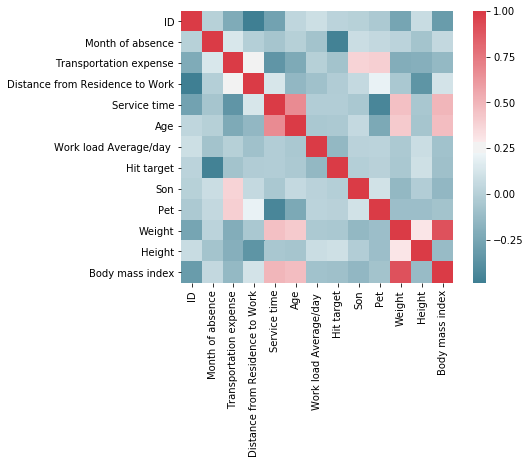
|  |  |  |
| --- | --- | --- |
| box new af.jpeg | | |
| box new af 2.jpeg | | |
| Rplot b.jpeg | Rplot01 a.jpeg | Rplot c.jpeg |
| box new 4.jpeg | | |

**3.8 Feature selection:**

Before performing any modeling we need to assess the importance of each independent variable. For that we do another preprocessing technique to check the correlation analysis for continuous variable and then we checked the Chi-Square test of independent.

**3.9 Correlation Plot:**

Here we have plot a correlation plot of all continuous variable. Here we have 13 variables and observe the correlation analysis between independent and dependent variables.

****

**Color coding:**

= Highly Positively Correlated with two continuous variable.

= Highly Negatively Correlated with two continuous variable.

From the above plot, we analyzed that all independent variable are correlated with dependent variable, except Weight is correlated with body mass index.  
It means weight and body mass index carry same value, for further analysis we could drop the weight variable from the dataset to decrease the complexity of the model. Body mass index and weight both define the target variable.

**3.10 Chi-Square Test**

In this test, we check the dependency between the two categorical variable. Here we have 7 categorical variable which are as follows:

|  |  |
| --- | --- |
|  | **P- Value** |
| Reason for absence | 5.6929209125065616e-70 |
| Day of the week | 0.0335239917836647 |
| Seasons | 5.84210399047505e-05 |
| Disciplinary failure | 4.375044921283115e-114 |
| Education | 1.0 |
| Social drinker | 2.128642459000793e-08 |
| Social smoker | 0.018742041906335 |

**3.11 Feature Scaling method:**

**Normalization Method**

Here our task is to scale the numerical variable in the range of 0 – 1.We have total 12 numerical variable and to scale so that it’s easy to make model.

1. **Building Predictive Model**

The data is first divided into training and testing set in the proportion of **80:20** for developing a model.

**4.1 Linear Regression Model**

Provided problem statement is falls under **Regression category.**

Linear regression model used for the continuous target variable as we have target variable as a continuous.

We need to find ways to reduce the absenteeism of the employee depending on the multiple parameters provided so that we can reduce the employee absenteeism. We have to split the data into the train and test data.

Now we have divided the data into train and test data. 80% of data is falling under train data category and rest 20% is test data.

Before building linear regression model, we need to check the multi co linearity effect on the data or not.

When two independent variable are highly correlated to each other, then multi collinear effect occurs in the model. It will inflict the variance of different or strong repressor or predictors.

Before building the model, first we need to check the multi collinear effect if the effect is 0 then we feed that model to build the linear regression model.

We have few test which identify the multi collinear effect in the data and one of the test is **Variance Inflation Factor(VIF).**

VIF= 1/ (1-r^2)

where r^2= correlation coefficient

vif(data[,-19])

|  |
| --- |
| Variables VIF |
| 1 ID 1.847359 |
| 2 Reason.for.absence 1.535850 |
| 3 Month.of.absence 1.600687 |
| 4 Day.of.the.week 1.074263 |
| 5 Seasons 1.301002 |
| 6 Transportation.expense 1.863481 |
| 7 Distance.from.Residence.to.Work 2.474856 |
| 8 Service.time 3.507007 |
| 9 Age 2.719056 |
| 10 Hit.target 1.364049 |
| 11 Disciplinary.failure 1.524748 |
| 12 Son 1.360044 |
| 13 Social.drinker 2.405254 |
| 14 Social.smoker 1.255234 |
| 15 Pet 1.676604 |
| 16 Height 1.676830 |
| 17 Body.mass.index 1.949958 |

Here we have a library **USDM**, which help us to influence the VIF factor and implement the VIF factor on all the independent variable. VIF is a function used to check the multi co linearity or any variable which is highly correlated to each other. As we are running this test only for the independent variable and exclude the target variable.

Here we provide a particular value for each of the independent variable by the help of VIF. First VIF take one of the independent variable like **ID** then calculate the correlation coefficient with all other independent variable by iterating one by one, then it will calculate the particular value of independent value based on the correlation value with all other variable.

No variable from the 17 input variables has collinearity problem.

The linear correlation coefficients ranges between:

min correlation ( Month.of.absence ~ ID ): -0.0004764915

max correlation ( Age ~ Service.time ): 0.6708817

---------- VIFs of the remained variables --------

|  |
| --- |
| Variables VIF |
| 1 ID 1.847359 |
| 2 Reason.for.absence 1.535850 |
| 3 Month.of.absence 1.600687 |
| 4 Day.of.the.week 1.074263 |
| 5 Seasons 1.301002 |
| 6 Transportation.expense 1.863481 |
| 7 Distance.from.Residence.to.Work 2.474856 |
| 8 Service.time 3.507007 |
| 9 Age 2.719056 |
| 10 Hit.target 1.364049 |
| 11 Disciplinary.failure 1.524748 |
| 12 Son 1.360044 |
| 13 Social.drinker 2.405254 |
| 14 Social.smoker 1.255234 |
| 15 Pet 1.676604 |
| 16 Height 1.676830 |
| 17 Body.mass.index 1.949958 |

We have another metrics which helps us to identity whether to keep the variable or delete the variable from the data set which can be VIFCOR( variance influence factor correlation ), it will take data as first argument, second argument is threshold as we know that correlation ranges from -1 to +1

We used Library USDM for checking the multi collinear effect in the dataset. Correlation factor range from -1 to 1, if it is -1, then it is highly negatively correlated and if it is +1 then it is highly positively correlated to each other.

In either case we are not accepting the highly positively or negatively correlated. In this scenario we need to assign a threshold value for which output data is acceptable. If the correlation value is till the 90% either it is positively correlated or negatively correlated, it is acceptable if the correlation between the two variable is beyond the 0.9 whether that variable is positive or negative we just remove that variable.

First it finds the pair of variable which has maximum linear correlation, it basically find linear correlation which is actually greater than threshold value and exclude one of them which have greater VIF and remove one of them VIF if high.

VIFCOR repeat this iteration till no variable is highly correlation with other variables. If VIF > 10 then we have co linearity problem in our variables but here in this case all the variables have VIF < 10 so here there is no correlation effect in dataset.

After running the collinear test, then Minimum correlation between independent variables (**month of absence ~ ID=-0.0004)** which is close to 0< 0.9 and maximum correlation between (**Age ~ Service.time** = 0.67) which is less than threshold value. There is no highly positively or negatively correlated variable in our input data.

**Now let us consider these input variables to feed in the model. Now we develop a linear regression model on the top of it.**

Residuals:

Min 1Q Median 3Q Max

-24.526 -4.865 -1.494 1.463 105.551

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |  |  |  |  |
| Estimate Std. Error t value Pr(>|t|) | | | | | | | | |
| (Intercept) 19.34846 5.13457 3.768 0.000181 \*\*\* | | | | | | | | |
| ID -2.19325 2.30244 -0.953 0.341207 | | | | | | | | |
| Reason.for.absence -0.50696 0.07896 -6.421 2.85e-10 \*\*\* | | | | | | | | |
| Month.of.absence 1.62060 2.35942 0.687 0.492445 | | | | | | | | |
| Day.of.the.week -0.87337 0.38491 -2.269 0.023636 \* | | | | | | | | |
| Seasons -0.18697 0.56415 -0.331 0.740445 | | | | | | | | |
| Transportation.expense 1.48884 2.92055 0.510 0.610400 | | | | | | | | |
| Distance.from.Residence.to.Work -3.59896 2.65170 -1.357 0.175243 | | | | | | | | |
| Service.time -2.22927 6.31198 -0.353 0.724083 | | | | | | | | |
| Age 8.66299 4.07042 2.128 0.033739 \* | | | | | | | | |
| Hit.target 2.23019 3.18631 0.700 0.484257 | | | | | | | | |
| Disciplinary.failure -17.71694 2.84126 -6.236 8.73e-10 \*\*\* | | | | | | | | |
| Son 4.13313 2.23301 1.851 0.064694 . | | | | | | | | |
| Social.drinker 2.69646 1.65811 1.626 0.104449 | | | | | | | | |
| Social.smoker -1.82676 2.21878 -0.823 0.410669 | | | | | | | | |
| Pet -2.00087 4.07822 -0.491 0.623881 | | | | | | | | |
| Height 2.49538 3.79075 0.658 0.510622 | | | | | | | | |
| Body.mass.index -6.20973 3.23594 -1.919 0.055481 | | | | | | | | |

---

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

Residual standard error: 12.91 on 574 degrees of freedom

Multiple R-squared: 0.1439, Adjusted R-squared: 0.1185

F-statistic: 5.675 on 17 and 574 DF, p-value: 5.32e-12

**4.1.1 Summary of model:**

Summary can also help you to extract the correlation coefficient of employee absenteeism or coefficient of independent variable and also get the r^2 and adjusted r^2. This linear regression model is a statistical model which will calculate the regression coefficient of employee absenteeism which will help us to extract the amount of information of each independent variable to reduce the employee absenteeism.

Here we get the summary statistics of error which is called residuals where we got maximum erroras 105.55 and minimum errors -24.52.

**Regression Coefficient:**

**Statistical model** does not save the pattern in the memory, we havel calculate the regression coefficient which helps us to extract the amount of information each independent variable has contributed.

Estimate is nothing but coefficient of each independent variable. Here we have some estimator value of all the independent variable.

Here we consider Age of the employee, age in terms of years, if one unit increases in age, then 10.811 times decrease in the employee absenteeism.

If one unit increases in the distance from residence to work variable which leads to 3.59 times decrease in the employee absenteeism.

This is how we have how we interpret the coefficient.

Here minus ( - )symbol means decrease in the employee absenteeism and plus ( + ) means increases the employee absenteeism.

**So from above linear regression model we observe that we reduce the employee absenteeism by the help of these variables whose estimator is minus which is**

* ID
* Reason of absence
* Day of week
* Seasons
* Distance from residence to work
* Transportation Expenses
* Service time
* Disciplinary Failure
* Social Smoker
* Pet
* Body Mass index

Here we have standard error is 12.91 which is also called standard deviation error it measures the average amount that coefficient estimators varies from actual average value of our response variable.

This standard error t-value have contribute to calculate the p-value..

T-value measure how many standard deviation coefficients which are away from 0. Using the t- value we have calculate the p-value

In our model summary, we have get the **p – value** and those variable whose p- value < 0.05 we accept the alternative hypothesis or reject the **NULL Hypothesis** saying that these variable is significant to us.

Here in these model, Variables whose **p-value < 0.05 is**

* **Reason of absence(\*\*\*)**
* **Distance from residence to work (\*)**
* **Age(\*\*)**
* **Disciplinary failure (\*\*\*)**

This above variable is significant to reduce the employee absenteeism but **reason of absence & disciplinary failure** are highly significant to explain the target variable and  **Distance from residence to work & Age** is least significant to explain the target variable.

The remaining other variable has p-value>0.05 which means we accept the null hypothesis saying that this independent variables not significant to explain the employee absenteeism or target variable.

\*\*\* (3 star)- Highly Significant variable to explain employee absenteeism

\*\* (2 Star )- Moderate Significant variable to explain dependent variable

\* ( 1 Star )- Least Significant variable explain dependent variable

**Adjusted R^2= 0.1439 or 14.39%**

**R^2= 0.1185 or 11.85%**

Here, 14.39% of the independent variable has explained the target variable which is not at all acceptable at least we should have r^2 above 80%.

Here F – statistics > 1 then it is good relation between predictor and respond or target variable, here we get F- statistics is 5.675 which is reasonably good statistics.

Overall P-Value< 0.05 which is get from F- Statistics model is reasonably good as we get limited number of observation.

**4.1.2 Model Performance or Evaluation for Linear Regression Model**

As this is time series problem, so we can take the root mean squared error (RMSE) for predicting the model performance and also get the accuracy of the model.

|  |  |  |  |
| --- | --- | --- | --- |
| **MAE** | **RMSE** | **MAPE** | **MSE** |
| 5.54964 | 11.44698 | Inf | 131.0333 |

**RMSE= 11.44%**

**Accuracy is 88.56%** which has good accuracy of the model to reduce the employee absenteeism.

**4.2 Decision Tree Regression:-**

Here internally decision tree works by nodes and branches. First important partis we need to select the parent node, that can be done by the information gain .Information gain is the expected amount of information that would be needed.

Out of all the independent variable, the variable which is considered as parent node which is contributing nearly all information of remaining variable. Those variable whose Information gain is high that will be contributing much information compared to all other independent variable.

IG= ( Entropy of the system before Split – Entropy of the system after split)

Entropy ( H ) = - ∑iPi log2Pi

Here we find the entropy of the system before split (we keep both target variable and independent variable together to calculate the impurity) & then calculate the entropy after split and then take difference to get the information Gain.

Now the independent variable whose information gain is high is considered as a **Parent Node**. We have total 740 observation and 19 variables. After that we build a model on the train data and then apply that model on the test data to know the accuracy of the model.

Decision tree helps us to reduce the employee absenteeism, what we predicting that when the new employee come what is the chances of absent in the company when new employee join the company. Here we predict the employee absenteeism of the new employee which is about to join.

Here we have 18 independent variable after doing proper data exploration, what we do we want to predict the Target variable to know the Employee absenteeism when new employee join the company.

For that we built a decision tree on the top of the data, 1st we need to convert the data into train and test data.

For building a decision tree regression model, we load the library RPART, we are using ANOVA method for building a regression type problem statement.

n= 592

node), split, n, deviance, yval

\* denotes terminal node

1) root 592 111678.2000 7.0499730

2) Reason.for.absence>=20 339 2891.2490 3.5735600 \*

3) Reason.for.absence< 20 253 99200.4100 11.7080900

6) Reason.for.absence< 0.5 38 547.4438 0.9140786 \*

7) Reason.for.absence>=0.5 215 93443.0400 13.6158700

14) Age< 0.7258065 200 55428.3700 12.0720600

28) Height< 0.8333333 190 48653.0200 11.3179600

56) Reason.for.absence< 18.5 159 28620.1300 9.8642280

112) Hit.target>=0.8157895 53 652.0667 6.1206610 \*

113) Hit.target< 0.8157895 106 26853.9200 11.7360100

226) Body.mass.index>=0.3684211 46 1671.8260 7.7826090 \*

227) Body.mass.index< 0.3684211 60 23911.9500 14.7669500

454) ID>=0.4142857 50 17161.6200 12.4403400

908) Reason.for.absence>=9.5 37 3412.7020 9.2707350 \*

909) Reason.for.absence< 9.5 13 12319.2300 21.4615400 \*

455) ID< 0.4142857 10 5126.4000 26.4000000 \*

57) Reason.for.absence>=18.5 31 17973.4200 18.7741900

114) Hit.target< 0.7631579 14 182.9286 6.9285710 \*

115) Hit.target>=0.7631579 17 14208.2400 28.5294100 \*

29) Height>=0.8333333 10 4614.4000 26.4000000 \*

15) Age>=0.7258065 15 31182.4000 34.2000000 \*

These models are in the form of tabular but its gets splits. Here we have pattern value, here we have all the decision tree pattern value or rules for the decision tree.

Now we apply decision tree regression model on the test data and find get predicted value of the test data we have actual value then compare with the actual value of the test data and get performance of the model.

.

**4.2.1 Performance of Model**

**We have four type of error which help us to evaluate the performance of these regression model.**  We have 4 error matrices i.e.

* RMSE ( root Mean Square Error)
* MAE (Mean absolute error)
* MAPE ( Mean absolute percentage error)
* MSE ( Mean square error )

Here we have used MSE and RMSE for finding the error matrices because our data is transaction and time series analysis.

MSE will give the error in the form of number and RMSE we get the error in the form of percentage so we calculate the RMSE to calculate the error rate.

Here we get the predicted value and compare with the actual value of the target value and got the error which we get through calculating RMSE and get accuracy of the model.

|  |  |  |  |
| --- | --- | --- | --- |
| MAE | RMSE | MAPE | MSE |
| 5.141665 | 12.4052 | INF | 153.8911 |

Here, we have regr.eval function which is available in library ( DMwR), which helps to calculate the regression error matrix which include MAE(mean absolute error), RMSE(root mean square error), MAPE(mean absolute percentage error ), MSE(mean square error).

**4.2.2Model Performace:-**

Here we get the RMSE= 12.40, accuracy of the model is 87.60% which is quite good which we get from the decision tree algorithm to explain the employee absenteeism.

* 1. **Random Forest:-**

Now we develop a model by random forest. Now, we run a random forest algorithm. Here we use three fold cross validation in this model due the computational cost. Random forest Packages used for building this model. In this model we get built a model on the train data and apply on the test data.

Here in this case we develop a model and extract to read the 2 rules of the tree.

exec=extractRules(treelist, train[,-18])

3557 rules (length<=6) were extracted from the first 100 trees.

> exec[1:2,]

[1] "X[,2]<=22.3814439197402 & X[,2]<=0.5 & X[,4]<=0 & X[,6]<=0.453703703703704 & X[,16]<=0.270462946485137"

[2] "X[,2]<=22.3814439197402 & X[,2]<=0.5 & X[,4]>0 & X[,6]<=0.453703703703704 & X[,14]<=0 & X[,16]<=0.270462946485137"

len freq err

[1,] "6" "0.015" "0"

[2,] "6" "0.003" "0"

condition

[1,] "X[,2]<=22.3814439197402 & X[,2]<=0.5 & X[,4]>0 & X[,6]<=0.453703703703704 & X[,14]<=0 & X[,16]<=0.270462946485137"

[2,] "X[,2]<=22.3814439197402 & X[,2]<=0.5 & X[,4]>0 & X[,6]<=0.453703703703704 & X[,14]>0 & X[,16]<=0.270462946485137"

pred

[1,] "0"

[2,] "0"

Here, length tells us no of variable value taken in the condition, it means we have 6 variable is included, frequency tells us percentage of data satisfied the condition which is 0.015% in whole data satisfied the condition. Error means amount of percentage error is 0 which doesn’t satisfy the data.

**4.3.1Performace of the Model:-**

Here we get the RMSE= 12.61, accuracy of the model is 87.39% which is quite good which we get from the Random Forest algorithm to explain the employee absenteeism. Here we draw random forest with 100 trees and after increasing in the tree still we get slight decrease in the RMSE, so we can consider the random forest with ntrees=100.

|  |  |  |  |
| --- | --- | --- | --- |
| MAE | RMSE | MAPE | MSE |
| 5.1052 | 12.61327 | INF | 159.0944 |

**4.4 Model Selection:-**

We can see that both the model comparatively on average and we select either of the two models without any loss of information.

1. **MEASURES FOR CONTROL ABSENTEEISM**

* Adoption of a well defined recruitment procedure
* Cordial relationship between supervisors and workers
* Provision of reasonable wages and allowances and job security for workers
* Motivation of workers and social measures
* Improved communication and prompt redressed of grievances
* Liberal grant of leave
* Safety and accident prevention
* Provision of healthy and hygienic working conditions
* Development of workers education

1. **SUGGESTIONS AND RECOMMENDATIONS**

* The best and simplest way to reduce absenteeism is providing counseling to those employees who take leave unnecessarily and making them aware of the problems of absenteeism and their importance at the work place.
* The management's strict attitude in granting leave even when the need is genuine tempts the workers to go on leave though on loss of play. Hence an effective way of dealing with absenteeism is to liberalize leave rules.
* The rules and regulation relating to attendance must be explained to workers. In order to reduce work load, must appoint sufficient employees. Only them the existing employees can work better without any stress or strain and by this absenteeism can be reduced
* Giving employees incentives for reduced absenteeism is not the same as rewarding or giving employees bonuses for reduced absenteeism. An incentive provides an employee with a boost to their motivation to avoid unnecessary absenteeism.
* Periodical medical camps for free check-Ups can improve the health of employees. By this absenteeism can be reduced.
* Proper counseling regarding religion and caste must be endeavored
* Improving welfare measures considerably reduces absenteeism.
* Strict disciplinary measures to reduce absenteeism must be configured.

1. **CONCLUSION**

This study analysis the issue of employees absenteeism and explores in detail preventative and corrective actions. Absenteeism has a negative impact on a company's employee morale. There are a number of programs that can be implemented individually or collectively to reduce employee absenteeism. Absenteeism is a serious and costly problem faced by companies throughout the world. This problem requires that all employees understand the consequences of such behavior from a company's standpoint as well as a personal standpoint. All companies must approach this problem from a proactive position with employee prevention programs and progressive discipline programs.

1. **Complete R-Code**

rm(list=ls())

setwd("C:/users/user")

getwd()

#load data

data= read.csv("RF\_Absenteeism\_at\_work\_Project.csv")

#LOAD Libraries

x=c("ggplt2", "corrgram", "DMwR", "Caret", "RandomForest", "unbalance", "C50", "dummies", "e1071", "information", "MASS", "rpart", "gbm", "ROSE")

lapply(x, require, character.only=TRUE)

names(data)

#library for ploting the graph

library(scales)

library(psych)

library(gplots)

library(ggplot2)

#explore the data

str(data)

#missing value analysis

missing\_val= data.frame(apply(data, 2, function(x)(sum(is.na(x)))))

View(missing\_val)

#calculate how much missing value in particular variables

sum(is.na(data$ID))

sum(is.na(data$Reason.for.absence))

#convert the data into a proper shape

missing\_val$columns=row.names(missing\_val)

row.names(missing\_val)= NULL

#LET us rename the 1st variable

names(missing\_val)[1]="missing-percentage"

#calcualte the percentage of missing value

missing\_val$`missing-percentage`=(missing\_val$`missing-percentage`/nrow(data))\*100

#convert into descending order

missing\_val= missing\_val[order(-missing\_val$`missing-percentage`),]

View(missing\_val)

#reaaranging the columns

missing\_val=missing\_val[, c(2,1)]

#Plot the missing percentage in the histogram

ggplot(data = missing\_val[1:21,], aes(x= reorder(columns, -`missing-percentage`), y= `missing-percentage`))+

geom\_bar(stat="identity", fill= "grey")+xlab("parameter")+

ggtitle("missing\_data percentage (train)") + theme\_bw()

#Method of imputing the missing value

# mean method

data[71,1]

data$ID[is.na(data$ID)]= mean(data$ID, na.rm = T)

data[71,1]=NA

data[71,1]

#median method

data$ID[is.na(data$ID)]=median(data$ID, na.rm = T)

#KNN Imputation method

data=knnImputation(data, k=5)

View(data)

#histogram plot

ggplot(data, aes(x=data$Month.of.absence))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("month of absence")

ggplot(data, aes(x=data$Transportation.expense))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("transportation expences")

ggplot(data, aes(x=data$Distance.from.Residence.to.Work))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("distance from residence to work")

ggplot(data, aes(x=data$Service.time))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("service time")

ggplot(data, aes(x=data$Age))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("age")

ggplot(data, aes(x=data$Hit.target))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("hit target")

ggplot(data, aes(x=data$Son))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("son")

ggplot(data, aes(x=data$Pet))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("pet")

ggplot(data, aes(x=data$Weight))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("weight")

ggplot(data, aes(x=data$Height))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("height")

ggplot(data, aes(x=data$Absenteeism.time.in.hours))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("absenteeism time in hours")

ggplot(data, aes(x=data$Body.mass.index))+geom\_histogram(fill="DarkSlateBlue", colour="black")+ggtitle("body mass index")

#boxplot distribution and outlier analysis

numeric\_index= sapply(data, is.numeric)

numeric\_index

numeric\_data= data[, numeric\_index]

cnames= colnames(numeric\_data)

cnames

library(ggplot2)

for (i in 1:length(cnames)) {

assign(paste0("gn", i), ggplot(aes\_string((cnames[i]),x= "Absenteeism.time.in.hours"), data= subset(data))+

stat\_boxplot(geom= "errorbar", width=0.5)+

geom\_boxplot(outlier.colour="red", fill="grey", outlier.shape=18,

outlier.size=1, notch=FALSE)+

theme(legend.position="bottom")+

labs(y=cnames[i], x="Absenteeism employee")+

ggtitle(paste("box plot for", cnames[i])))

}

gridExtra::grid.arrange(gn1, gn3, gn6, ncol=3)

gridExtra::grid.arrange(gn7,gn8, gn9, ncol=3)

gridExtra::grid.arrange(gn10,gn12, ncol=2)

gridExtra::grid.arrange(gn13,gn16, gn17, ncol=3)

gridExtra::grid.arrange(gn18,gn19, ncol=2)

#delete those observation which contain outliers

val=data$area.code[data$area.code %in% boxplot.stats(data$area.code)$out]

data=data[which(data$area.code %in% val),]

data=data[which(!data$area.code %in% val),]

data=df

#detect and delete the outliners from all numerical variable by iterating the loop

for (i in cnames){

print(i)

val=data[,i][data[,i] %in% boxplot.stats(data[,i])$out]

print(length(val))

data[,i][data[,i] %in% val] =NA

}

sum(is.na(data))

View(data)

data=knnImputation(data, k=3)

#boxplot analysis after detect and impute outliers

for (i in 1:length(cnames)) {

assign(paste0("gn", i), ggplot(aes\_string((cnames[i]),x= "Absenteeism.time.in.hours"), data= subset(data))+

stat\_boxplot(geom= "errorbar", width=0.5)+

geom\_boxplot(outlier.colour="red", fill="grey", outlier.shape=18,

outlier.size=1, notch=FALSE)+

theme(legend.position="bottom")+

labs(y=cnames[i], x="Absenteeism employee")+

ggtitle(paste("box plot for", cnames[i])))

}

gridExtra::grid.arrange(gn1, gn3, gn6, ncol=3)

gridExtra::grid.arrange(gn7,gn8, gn9, ncol=3)

gridExtra::grid.arrange(gn10,gn12, ncol=2)

gridExtra::grid.arrange(gn13,gn16, gn17, ncol=3)

gridExtra::grid.arrange(gn18,gn19, ncol=2)

#chec shape of data

dim(data)

#Split numerical and categorical variable

catnames= subset(data, select= -c(Reason.for.absence, Seasons, Day.of.the.week, Disciplinary.failure,

Seasons,Education, Social.drinker, Social.smoker))

#drop variables

data= subset(data, select= -c(Education, Weight, Work.load.Average.day))

#correlation plot

corrgram(data[,numeric\_index], order=F,

upper.panel=panel.pie, text.panel=panel.txt, main="correlation plot")

numeric\_index= sapply(catnames, is.numeric)

numeric\_index

#chi sqaure of independence and selecting only categorical variable

factor\_index=sapply(data, is.factor)

factor\_data= data[, factor\_index]

View(factor\_data)

factor\_index=subset[factor\_index]

#normalizaion method

cnames1=c("ID", "Month.of.absence", "Transportation.expense", "Distance.from.Residence.to.Work",

"Service.time", "Age", "Hit.target", "Son",

"Pet", "Height", "Body.mass.index")

for (i in cnames1){

print(i)

data[,i]= (data[,i]-min(data[,i]))/(max(data[,i]-min(data[,i])))

}

#Build a model

#decision tree alogorithm(regression problm)

library(rpart)

library(MASS)

library(randomForest)

#sampling technique

train\_index= sample(1:nrow(data), 0.8\*nrow(data))

train= data[train\_index,]

test=data[-train\_index,]

#rpart for regression

fit= rpart(Absenteeism.time.in.hours~., data=train, method="anova")

fit

#predict for new test case

prediction\_DT= predict(fit, test[,-18])

prediction\_DT

#calculate MAPE

MAPE= function(y, yhat){

mean(abs((y-yhat)/y))\*100

}

MAPE(test[,18], prediction\_DT )

#alternative method

regr.eval(test[,18], prediction\_DT, stats = c("mae", "rmse", 'mape', 'mse'))

#linear regression method

#check multicollinearity

library(usdm)

vif(data[,-18])

vifcor(data[,-18], th= 0.9)

#Run regression model

lm\_model= lm(Absenteeism.time.in.hours~., data = train)

summary(lm\_model)

#Predict the target variable with absence of actual target variable

prediction\_LR=predict(lm\_model, test[,1:17])

MAPE(test[,18], prediction\_LR)

prediction\_LR

#Performance of model

regr.eval(test[,18], prediction\_LR, stats = c("mae", "rmse", 'mape', 'mse'))

#random forest algorithm

RF\_model= randomForest(Absenteeism.time.in.hours ~., train, importance=TRUE, ntree=500)

#extract rules from random forest

library(inTrees)

treelist= RF2List(RF\_model)

exec=extractRules(treelist, train[,-18])

exec[1:2,]

#make rules more readable

readablerules= presentRules(exec, colnames(train))

readablerules[1:2,]

rulemetrix=getRuleMetric(exec, train[,-18], train$Absenteeism.time.in.hours)

rulemetrix[1:2,]

RF\_prediction= predict(RF\_model, test[, -18])

RF\_prediction

#Performance of model

regr.eval(test[,18], RF\_prediction, stats = c("mae", "rmse", 'mape', 'mse'))