Bike Renting

Venkata Subramanian R

29 November 2018

**Table of Contents:**

1. **Introduction**
   1. **Problem Statement-------------------------------------------------------3**
   2. **Data----------------------------------------------------------------------------3**
2. **Methodology**
   1. **Pre-Processing-------------------------------------------------------------5**
      1. **Missing Value Analysis----------------------------------------------5**
      2. **Outlier Analysis--------------------------------------------------------6**
      3. **Correlation Plot--------------------------------------------------------8**
      4. **Feature Scaling--------------------------------------------------------8**
      5. **Feature Selection----------------------------------------------------10**
      6. **Modeling----------------------------------------------------------------11**
3. **Conclusion**
   1. **Model Evaluation--------------------------------------------------------15**
   2. **Model Selection----------------------------------------------------------16**
   3. **Solution---------------------------------------------------------------------16**
4. **Forecasting------------------------------------------------------------------------20**
5. **Appendix---------------------------------------------------------------------------21**

**Chapter 1**

**Introduction**

**1.1 Problem Statement**

The objective of this project is to predict the bike rental count on daily based on the environmental and seasonal settings. Analyze the data and develop a model to predict the behavioural changes with environmental and seasonal settings.

**1.2 Data**

Our task is to build a regression model which will find the Count of the daily users of the bike for rental purpose. Below are some sample data which will predict the Bike Rental Count.

Table 1.1: Bike Rental Sample Data (Columns 1-9)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| instant | dteday | season | yr | mnth | holiday | weekday | workingday | weathersit |
| 1 | 1/1/2011 | 1 | 0 | 1 | 0 | 6 | 0 | 2 |
| 2 | 1/2/2011 | 1 | 0 | 1 | 0 | 0 | 0 | 2 |
| 3 | 1/3/2011 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| 4 | 1/4/2011 | 1 | 0 | 1 | 0 | 2 | 1 | 1 |
| 5 | 1/5/2011 | 1 | 0 | 1 | 0 | 3 | 1 | 1 |
| 6 | 1/6/2011 | 1 | 0 | 1 | 0 | 4 | 1 | 1 |
| 7 | 1/7/2011 | 1 | 0 | 1 | 0 | 5 | 1 | 2 |
| 8 | 1/8/2011 | 1 | 0 | 1 | 0 | 6 | 0 | 2 |
| 9 | 1/9/2011 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

Table 1.2 Bike Rental Sample Data (Columns 10-16)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| temp | atemp | hum | windspeed | casual | registered | cnt |
| 0.344167 | 0.363625 | 0.805833 | 0.160446 | 331 | 654 | 985 |
| 0.363478 | 0.353739 | 0.696087 | 0.248539 | 131 | 670 | 801 |
| 0.196364 | 0.189405 | 0.437273 | 0.248309 | 120 | 1229 | 1349 |
| 0.2 | 0.212122 | 0.590435 | 0.160296 | 108 | 1454 | 1562 |
| 0.226957 | 0.22927 | 0.436957 | 0.1869 | 82 | 1518 | 1600 |
| 0.204348 | 0.233209 | 0.518261 | 0.089565 | 88 | 1518 | 1606 |
| 0.196522 | 0.208839 | 0.498696 | 0.168726 | 148 | 1362 | 1510 |
| 0.165 | 0.162254 | 0.535833 | 0.266804 | 68 | 891 | 959 |
| 0.138333 | 0.116175 | 0.434167 | 0.36195 | 54 | 768 | 822 |

**Chapter 2**

**Methodology**

**2.1** **Pre Processing**

Any predictive modeling requires that we look at the data before we start modeling. 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.**

To Start with, we will compute the Missing Values

**2.1.1 Missing Value Analysis**

In the given data set, there are no missing values found. So no need to perform an imputation methods.

**2.1.2 Outlier Analysis**

We have found that there are no missing values to impute. Now as the next step in Data Pre-Processing we need to detect the outliers that is present in the data set and remove it.

We can detect the outliers in the data set using Box Plot Method. In the below figure we have plotted the outliers in the employee absenteeism data set using Boxplot.

Fig 2.1

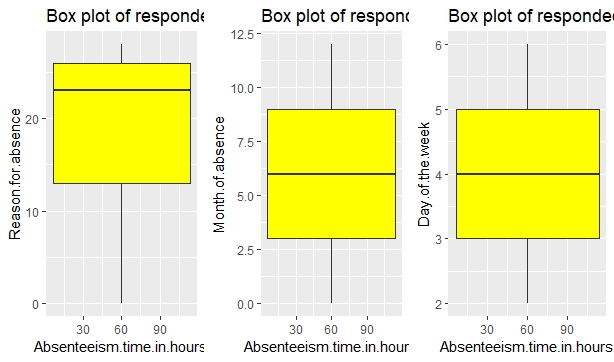
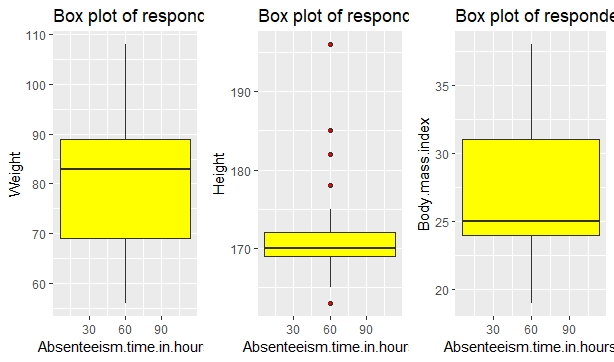


Fig 2.2



Fig 2.3



From the above plots we can infer that we have some outliers in Transportation Expense, Height, Age and Service Time Variables.

We will remove the outliers present in the data set and replace them with NA using the following R Code.

for (i in cnames) {

outliers\_value=employee\_sample[,i][employee\_sample[,i] %in% boxplot.stats(employee\_sample[,i])$out]

employee\_sample[,i][employee\_sample[,i] %in% outliers\_value]=NA

}

Once replaced with NA, we can impute the missing value using KNN Imputation Method which can be done using the following Code:

employee\_sample=knnImputation(employee\_sample,k=3)

Now with this Pre-Processing Technique, we have removed the outliers from the data set.

**2.1.3 Correlation Plot**

As the next step in Pre-Processing we need to find the multi- collinearity between the variables. If two variables contain the same information, then we can discard one variable while passing it to the model. So, with the help of Correlation Plot we can find the multi collinearity between the variables.

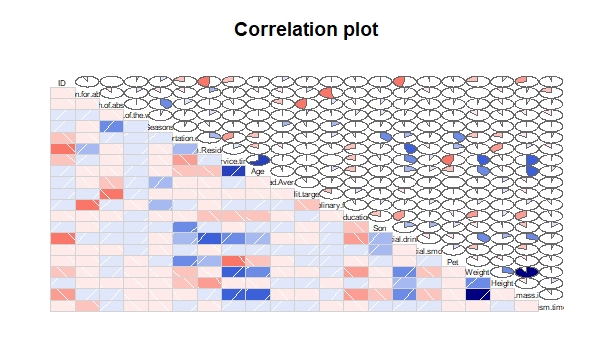


Fig 2.4

From the above image we can infer that there is no variable that has greater dependency on other variables. Each variable contains information and can be used to derive the target variable.

**2.1.4 Feature Scaling**

It is very important to scale the variables because the variables will have difference and variance in magnitude, units and range. Since most of the Machine Learning Algorithms use Euclidean distance between two data points in their computations, we need to take care of this dominance.

To Scale the feature, we have two techniques:

* **Normalization**
* **Standardization**

If the data is uniformly distributed, we can go for Standardization, on the other hand if the data is non-uniformly distributed, we can go for Normalization.

We can Infer the distribution of data from the below histograms of few variables.

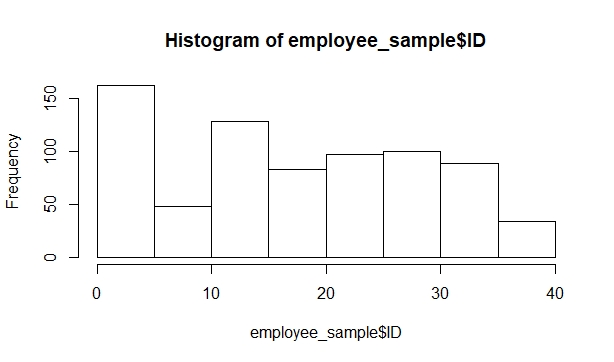


Fig 2.5

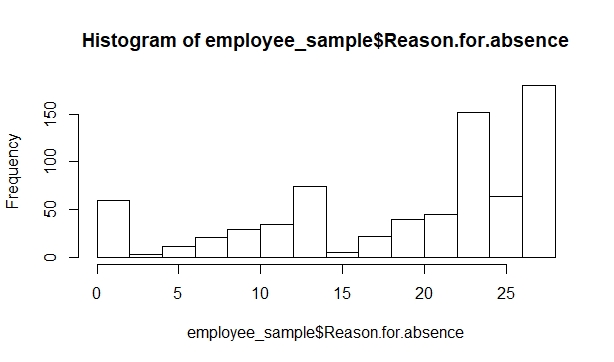


Fig 2.6

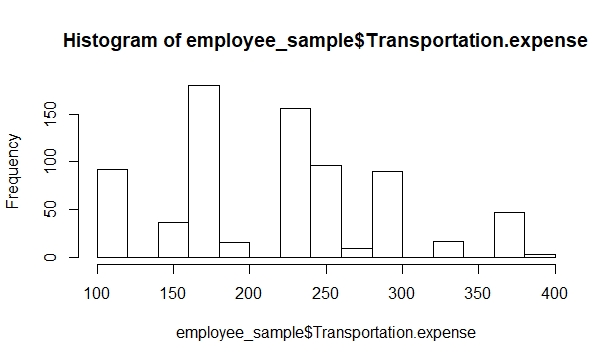


Fig 2.7

Based on the above histograms we can infer that the data is not normally distributed. So, we can go for **Normalization** approach to scale the data.

Normalization can be done only on the Continuous Variables. So, we will remove all the categorical variables and apply normalization only to the continuous variables.

**2.1.5 Feature Selection**

Before Performing any type of modeling, we need to access the importance of each predictor variables in our data sets. There is every chance that all variables in our analysis set may not be as important as all in the problem of class prediction. Below we have used Random Forest to Perform Feature Selection.

RF\_Model=randomForest(Absenteeism.time.in.hours ~.,data = employee\_sample,importance=TRUE,ntree=500)

importance(RF\_Model,type = 1)

%IncMSE

ID 19.247260

Reason.for.absence 59.944346

Month.of.absence 9.697707

Day.of.the.week 4.839629

Seasons 6.377380

Transportation.expense 21.164961

Distance.from.Residence.to.Work 10.882331

Service.time 13.401875

Age 15.732723

Work.load.Average.day. 10.529339

Hit.target 7.228418

Disciplinary.failure 23.521568

Education 3.294699

Son 16.234532

Social.drinker 5.772173

Social.smoker 3.801180

Pet 9.193516

Weight 11.245046

Height 12.868441

Body.mass.index 12.244614

From the above data we can infer that Reason for Absence is the most important predictor in our data set.

**2.6 Modeling**

**2.2.1 Model Selection**

We need to Predict the Absenteeism time in hours which is a continuous variable. For dependent variable we have 4 categories:

* Nominal
* Ordinal
* Interval
* Ratio

If Nominal it will come under classification model. If Interval or Ratio it will come under regression model. Since our variable is a continuous one it will come under Regression Model.

So, we will start from the basic model and try to predict the Target variable.

We will first use Linear Regression Model to Predict the value of Target Variable

**RCode:**

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

summary(lm\_model)

Residuals:

Min 1Q Median 3Q Max

-0.54816 -0.09269 -0.02304 0.09227 0.77039

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.5278215 0.0914312 5.773 1.28e-08 \*\*\*

ID -0.1096016 0.0413845 -2.648 0.00831 \*\*

Reason.for.absence -0.3094956 0.0295216 -10.484 < 2e-16 \*\*\*

Month.of.absence 0.0081263 0.0328976 0.247 0.80498

Day.of.the.week -0.0089486 0.0210527 -0.425 0.67095

Seasons -0.0007535 0.0225944 -0.033 0.97341

Transportation.expense 0.1153092 0.0407810 2.828 0.00486 \*\*

Distance.from.Residence.to.Work -0.0960658 0.0350438 -2.741 0.00631 \*\*

Service.time -0.0903438 0.0788349 -1.146 0.25228

Age -0.0176833 0.0590960 -0.299 0.76487

Work.load.Average.day. 0.0115782 0.0331564 0.349 0.72707

Hit.target 0.0293842 0.0352500 0.834 0.40486

Disciplinary.failure -0.5308144 0.0397276 -13.361 < 2e-16 \*\*\*

Education -0.0068938 0.0147225 -0.468 0.63978

Son 0.0778096 0.0325046 2.394 0.01700 \*

Social.drinker 0.0516498 0.0254886 2.026 0.04319 \*

Social.smoker 0.0444737 0.0343718 1.294 0.19622

Pet -0.0153247 0.0072126 -2.125 0.03404 \*

Weight -0.0410761 0.0927778 -0.443 0.65812

Height 0.0765864 0.0461009 1.661 0.09721 .

Body.mass.index 0.0442695 0.1106558 0.400 0.68926

---

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

Residual standard error: 0.1752 on 571 degrees of freedom

Multiple R-squared: 0.3745, Adjusted R-squared: 0.3512

F-statistic: 17.35 on 20 and 571 DF,p-value: < 2.2e-16

From the above observation on seeing the value of *Adjusted R-Squared,* we can see that the model can explain only 35% of the data. From the above model summary, we can find the variables which carry more information to derive the target variable.

Reason for absence and Disciplinary Failure contain high information to derive the target variable. Transportation Expense, Distance from residence to work and Social Drinker contains weightage in deriving the target variable.

When executed using **Optimum Least Square in Python**, the model shows the output as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | Absenteeism.time.in.hours | **R-squared:** | 0.722 |
| **Model:** | OLS | **Adj. R-squared:** | 0.712 |
| **Method:** | Least Squares | **F-statistic:** | 74.33 |
| **Date:** | Mon, 10 Dec 2018 | **Prob (F-statistic):** | 1.91e-144 |
| **Time:** | 08:48:29 | **Log-Likelihood:** | 186.54 |
| **No. Observations:** | 592 | **AIC:** | -333.1 |
| **Df Residuals:** | 572 | **BIC:** | -245.4 |
| **Df Model:** | 20 |  |  |
| **Covariance Type:** | nonrobust |  |  |

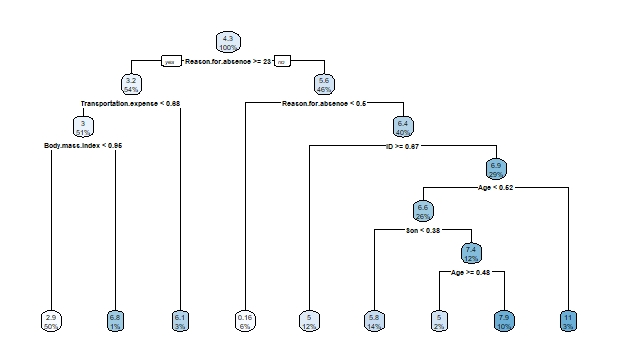
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **ID** | 0.0652 | 0.030 | 2.198 | 0.028 | 0.007 | 0.123 |
| **Reason.for.absence** | -0.2294 | 0.029 | -7.799 | 0.000 | -0.287 | -0.172 |
| **Month.of.absence** | 0.0787 | 0.032 | 2.475 | 0.014 | 0.016 | 0.141 |
| **Day.of.the.week** | -0.0283 | 0.022 | -1.290 | 0.198 | -0.071 | 0.015 |
| **Seasons** | -0.0314 | 0.023 | -1.363 | 0.173 | -0.077 | 0.014 |
| **Transportation.expense** | 0.1765 | 0.040 | 4.366 | 0.000 | 0.097 | 0.256 |
| **Distance.from.Residence.to.Work** | -0.0179 | 0.036 | -0.503 | 0.615 | -0.088 | 0.052 |
| **Service.time** | 0.1138 | 0.070 | 1.635 | 0.103 | -0.023 | 0.250 |
| **Age** | -0.0727 | 0.059 | -1.232 | 0.218 | -0.189 | 0.043 |
| **Work.load.Average.day.** | 0.0861 | 0.032 | 2.658 | 0.008 | 0.022 | 0.150 |
| **Hit.target** | 0.1125 | 0.035 | 3.243 | 0.001 | 0.044 | 0.181 |
| **Disciplinary.failure** | -0.4488 | 0.040 | -11.119 | 0.000 | -0.528 | -0.370 |
| **Education** | 0.0255 | 0.013 | 1.941 | 0.053 | -0.000 | 0.051 |
| **Son** | 0.1004 | 0.034 | 2.962 | 0.003 | 0.034 | 0.167 |
| **Social.drinker** | 0.0574 | 0.026 | 2.215 | 0.027 | 0.006 | 0.108 |
| **Social.smoker** | 0.0300 | 0.033 | 0.903 | 0.367 | -0.035 | 0.095 |
| **Pet** | -0.0100 | 0.008 | -1.329 | 0.184 | -0.025 | 0.005 |
| **Weight** | -0.0835 | 0.094 | -0.886 | 0.376 | -0.268 | 0.102 |
| **Height** | 0.1597 | 0.045 | 3.528 | 0.000 | 0.071 | 0.249 |
| **Body.mass.index** | 0.1710 | 0.113 | 1.517 | 0.130 | -0.050 | 0.392 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 140.791 | **Durbin-Watson:** | 1.940 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 367.621 |
| **Skew:** | 1.189 | **Prob(JB):** | 1.49e-80 |
| **Kurtosis:** | 6.041 | **Cond. No.** | 50.5 |

We can see from the above Inference that when executed in Python, the Adjusted R Square Value is much Improved. It varies from 71 to 81%, when executed with different samples.

**Decision Tree Plot**

Fig 2.9



**Conclusion**

**3.1 Model Evaluation**

Now that we have developed a few models for predicting our target variable and we have to decide which one to choose. We will evaluate some error metrics and decide which one to choose.

We can evaluate the Predictive Performance of the model with real values of the target variables and calculating some average error measures.

**3.1.1 Mean Absolute Error (MAE)**

regr.eval(test\_data[,20],predictions\_model,stats = 'mae')

MAE for the above Model is **0.12650535**

regr.eval(test\_data[,20],test\_model,stats = 'mae')

MAE for the above Model is **0.10777599**

**3.1.2 Root Mean Standard Error (RMSE)**

regr.eval(test\_data[,20],predictions\_model,stats = 'rmse')

RMSE for the above Model is **0.16186749**

regr.eval(test\_data[,20],test\_model,stats = 'rmse')

RMSE for the above Model is **0.15265838**

**3.1.3 Mean Standard Error (MSE)**

regr.eval(test\_data[,21],predictions\_model,stats ='mse')

MSE for the above model is **0.02620109**

regr.eval(test\_data[,21],test\_model,stats = 'mse')

MSE for the above model is **0.02330458**

**3.2 Model Selection**

Based on the Above Observations, it is good to choose Decision Tree Algorithm for the above model. But Linear Regression Model doesn’t have much Error value. The difference between the two models in terms of error metrics is very less. But Since Decision tree is efficient in terms of error metrics, we can choose decision tree model for the above solution.

**3.3 Solution**

From VIF Test, we found the variables which are important in deriving the target variable.

* Reason for Absence
* Distance from Residence to work
* Transportation Expense
* Disciplinary Failure
* Social Drinker
* Social Smoker

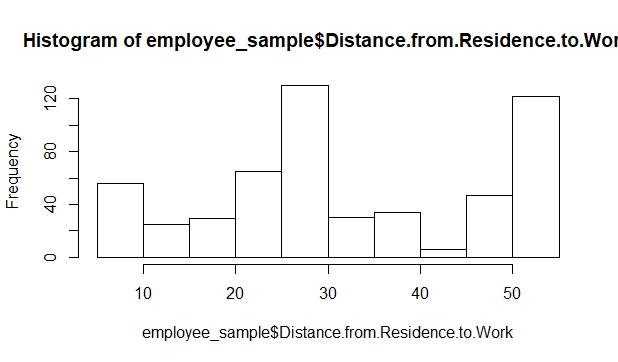
**Scenario 1:** In the Histogram plot of Reason for Absence Variable we see that most of the employees fall under the category 22 to 28. This includes some regular activities like medical consultation, blood donation, physiotherapy etc.

**Solution 1:** This can be lowered by providing free hospitality inside companies for employees to check their health regularly on a weekly or monthly basis.

The company can also conduct Blood Donation Camp monthly once or 2 months once to make their employees actively participate in it which can lower the absenteeism rate

**Scenario 2**: Distance from residence to work and Transportation Expenses

Fig 3.1



Based on the above histogram, we can infer that around 30-35% of the people travel more than 50kms from their residence to work. This will cause strain in their body and may lead to affect in health. So, there is every chance they may get health issues and take leaves. This will also increase their transportation charges in travelling to office daily.

**Solution 2:** This Situation can be avoided by arranging Sophisticated Cabs for Pick up and drop of employees from their Residence. Another way out is giving Work from Home Option at least 4-5 days in a month.

**Scenario 3:** Disciplinary Failure, Social Smoker and Social Drinker

Fig 3.2

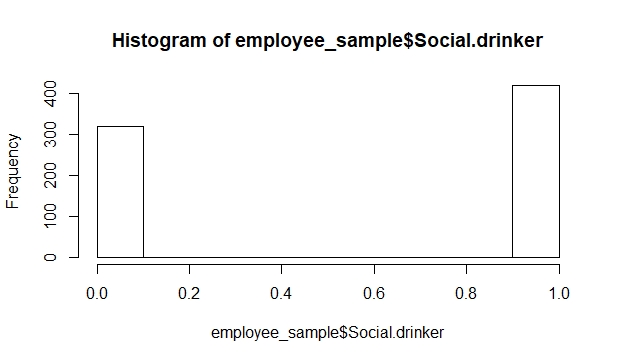


Fig 3.3

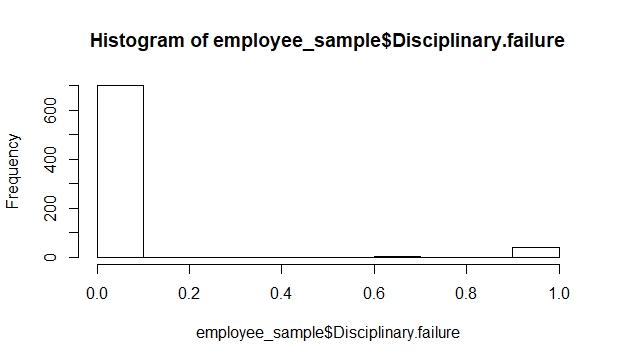
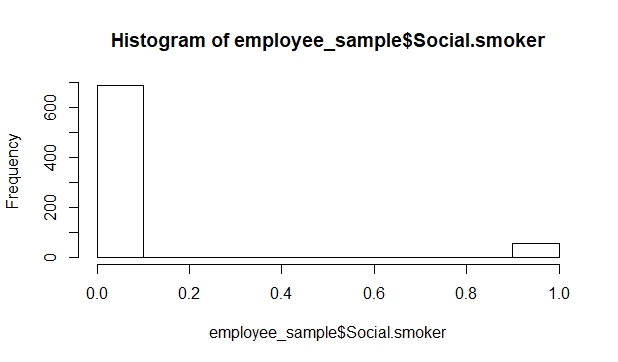


Fig 3.4

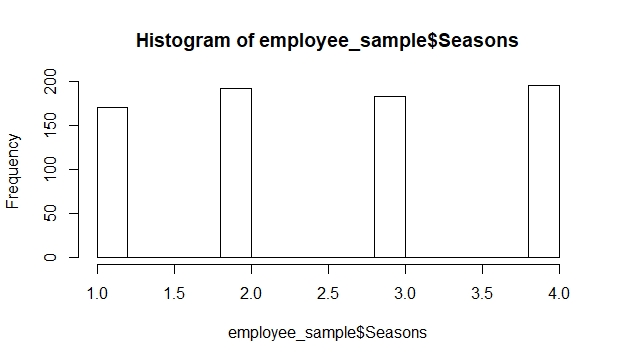


Based on the above observations, we can see that disciplinary failure and Social Smokers are very less. Drinkers are comparatively more in numbers.

**Solution 3**: These Variables doesn’t have much impact in Absenteeism of employees. One possible way is to get medical certificate from employees on a regular basis.

**Scenario 4:** Based on the Season

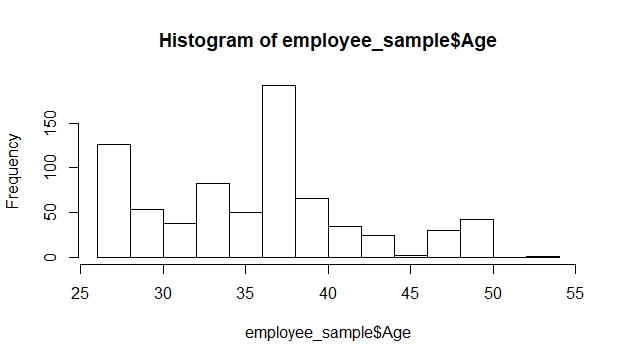
Fig 3.5



Based on the above histogram, around 55-60% of the leaves were taken in Autumn and Spring Season. **Solution 4:** Autumn Season is one where disease spread frequently and people suffer from many viral infections. Frequent Health check ups in company and awareness programs can be conducted to employees to take care of their health and daily pick ups and drops in cabs during heavy rains will help them to take care of their health and bring down absenteeism rate for the company.

**Scenario 5:** Based on Age

Fig 3.6



From the above diagram we can infer that, we can predict that most people are aged from 36-39. At the age of 39, they may not be able to reach their targets

**Solution 5:** Employing people between the age 23-28 more will improve the business and bring down absenteeism rate. People aged above 35 can be employed inside office and can be provided works like tracking the shipment and servicing customer calls. Young age people will be able to achieve more target and will be able to bring down absenteeism rate.

**Forecasting:**

When the same trend of absenteeism continues, we need to predict the loss every month the company will face

Absenteeism time in hours is forecasted for the year 2011 using **Linear Regression Model**

forecast\_LRM=ts(data=predictions\_model,start=c(2011,1),end=c(2011,12),frequency = 12)

|  |  |
| --- | --- |
| Jan | 5.35073 |
| Feb | 4.043169 |
| Mar | 5.826125 |
| Apr | 5.267901 |
| May | 9.258001 |
| Jun | 9.213145 |
| Jul | 6.076942 |
| Aug | 4.471843 |
| Sep | 6.377731 |
| Oct | 4.996535 |
| Nov | 6.731908 |
| Dec | 4.931555 |

The below forecast is done for Decision Tree Model

forecast\_mdel=ts(data = test\_model,start = c(2011,1),end = c(2011,12),frequency = 12)

|  |  |
| --- | --- |
| Jan | 7.9375 |
| Feb | 2.642857 |
| Mar | 2.642857 |
| Apr | 7.725817 |
| May | 5.992416 |
| Jun | 7.725817 |
| Jul | 3.1 |
| Aug | 5.992416 |
| Sep | 2.642857 |
| Oct | 2.736527 |
| Nov | 10.05394 |
| Dec | 0.158448 |

Absenteeism time in hours is forecasted for the year 2011 using Decision Tree Algorithm.

**Appendix A - Extra Figures**

