APRIL 22, 2019

EMPLOYEE ABSENTEEISM PREDICTION

DATA ANALYTICS

AKSHAY HIRPARA

Table of Contents

Project Description ____ 2

INTRODUCTION ____ 5

Data Explorer Library ____ 6

Data Pre-Processing ____ 7

Missing Value Analysis ____ 7

Outliers Analysis ____ 8

Feature Selection ____ 11

Feature Scaling ____ 14

Sampling Techniques ____ 17

Machine Learning ____ 19

Decision Tree ____ 20

Random Forest ____ 21

Linear Regression ____ 23

Evaluating Performance of Model ____ 25

MAE or MAD- Mean Absolute Error/Deviation ____ 25

MAPE- Mean Absolute Percentage Error ____ 25

Decision Tree Metrics ____ 27

Random Forest Metrics ____ 27

Linear Regression Metrics ____ 28

Visualization: ____ 31

Key Insights: ____ 31

Suggestions for Organisation-Category Wise ____ 31

Forecast for Work Loss and Absenteeism for coming 14 months ____ 31

New Concepts Used: ____ 36

Ludwig Algorithm: ____ 36

Data Explorer Library: ____ 38

Table of Figures:

Figure 1: Dimensions and Structure of Data Set ________ *5*

Figure 2: Unique Values ________ *6*

Figure 3: Missing Values ________ *8*

Figure 4: Box Plot for Outliers ________ *10*

Figure 5 :ANOVA Results ________ *12*

Figure 6:Corrgram Plot ________ *13*

Figure 7: Data Distribution Curves ________ *16*

Figure 8: Decision Tree Output ________ *21*

Figure 9: Random Forest ________ *22*

Figure 10:Summary of Linear Regression ________ *24*

Figure 11: Decision Tree Performance ________ *27*

Figure 12:Random Forest Performance ________ *28*

Figure 13:Linear Regression Performance ________ *29*

Figure 14: Performance Evaluation: R vs Python ________ *30*

Figure 15:Tableau Dashboard 1 ________ *33*

Figure 16:Tableau Dashboard 2 ________ *34*

Figure 17:Forecast Chart for next 14 months ________ *36*

Figure 18: Ludwig Syntax ________ *37*

Figure 19: Data Explorer Report Sample ________ *38*

Project Description:

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.

We have to help the company with the help of data analysis that what is the reason for absenteeism and what measures they have to take to reduce Absenteeism rate.

Attribute Information:

There is a variable in our data set which states Reason for Absence. It is divided in several categories. Below are they:

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the

immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioural disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system

XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities

XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX Injury, poisoning and certain other consequences of external causes

XX External causes of morbidity and mortality

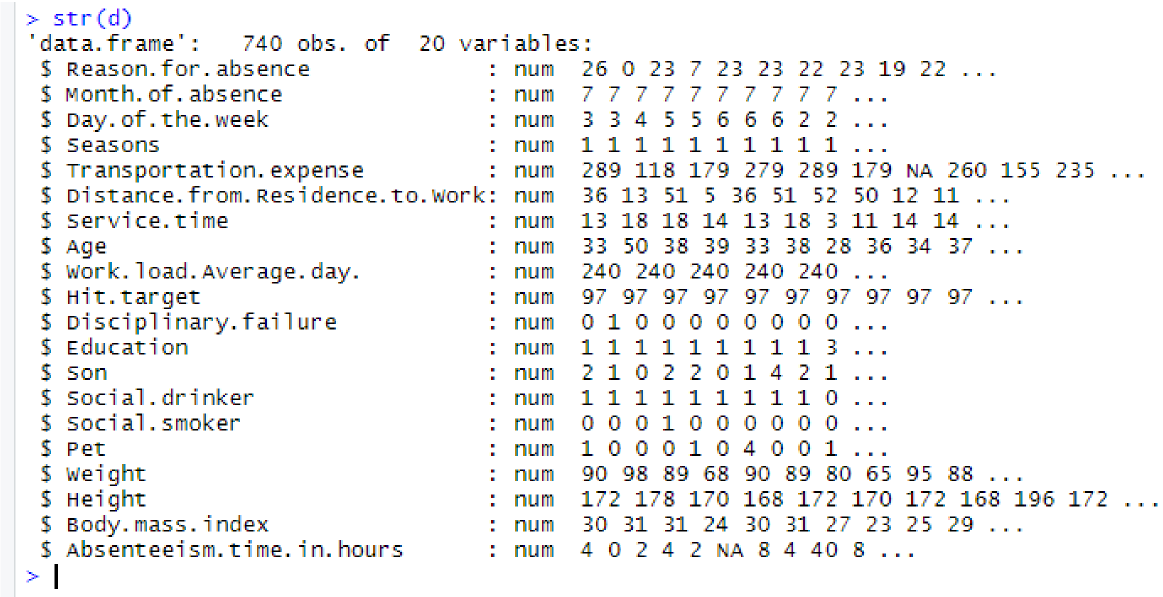
XXI Factors influencing health status and contact with health services.

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

Chapter-1

INTRODUCTION

The first step we always go for is knowing the data. We will check the data i.e. what are the number of rows and columns (Observations vs Variables), what is the data type of all the variables present and what are the variables associated with the data.



```
> str(d)
'data.frame':	740 obs. of  20 variables:
 $ Reason.for.absence             : num  26 0 23 7 23 23 22 23 19 22 ...
 $ Month.of.absence               : num  7 7 7 7 7 7 7 7 7 7 ...
 $ Day.of.the.week                : num  3 3 4 5 5 6 6 6 2 2 ...
 $ Seasons                        : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Transportation.expense         : num  289 118 179 279 289 179 NA 260 155 235 ...
 $ Distance.from.Residence.to.Work: num  36 13 51 5 36 51 52 50 12 11 ...
 $ Service.time                   : num  13 18 18 14 13 18 3 11 14 14 ...
 $ Age                            : num  33 50 38 39 33 38 28 36 34 37 ...
 $ Work.load.Average.day.         : num  240 240 240 240 240 ...
 $ Hit.target                     : num  97 97 97 97 97 97 97 97 97 97 ...
 $ Disciplinary.failure           : num  0 1 0 0 0 0 0 0 0 0 ...
 $ Education                      : num  1 1 1 1 1 1 1 1 1 3 ...
 $ Son                            : num  2 1 0 2 2 0 1 4 2 1 ...
 $ Social.drinker                 : num  1 1 1 1 1 1 1 1 1 0 ...
 $ Social.smoker                  : num  0 0 0 1 0 0 0 0 0 0 ...
 $ Pet                            : num  1 0 0 0 1 0 4 0 0 1 ...
 $ Weight                         : num  90 98 89 68 90 89 80 65 95 88 ...
 $ Height                         : num  172 178 170 168 172 170 172 168 196 172 ...
 $ Body.mass.index                : num  30 31 31 24 30 31 27 23 25 29 ...
 $ Absenteeism.time.in.hours      : num  4 0 2 4 2 NA 8 4 40 8 ...
> |
```

Figure 1: Dimensions and Structure of Data Set

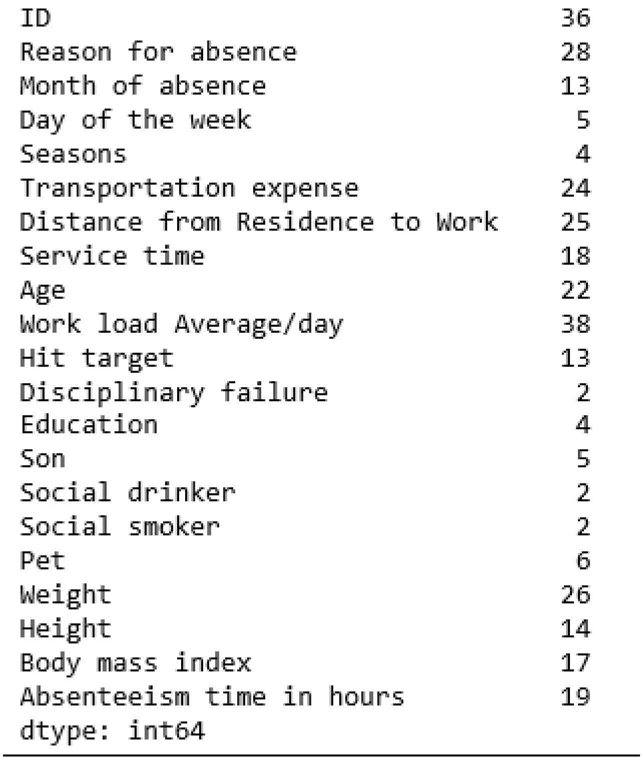
Unique Values:

Now we will check number of unique values present in each variable.

Checking unique values will give us more information about the data set and each variable.

For Example:

Social Drinker has 2 unique values. One is Yes and One is No.



```
ID                                 36
Reason for absence                 28
Month of absence                   13
Day of the week                     5
Seasons                             4
Transportation expense             24
Distance from Residence to Work    25
Service time                       18
Age                                22
Work load Average/day              38
Hit target                         13
Disciplinary failure                2
Education                           4
Son                                 5
Social drinker                      2
Social smoker                       2
Pet                                 6
Weight                             26
Height                             14
Body mass index                    17
Absenteeism time in hours          19
dtype: int64
```

Figure 2: Unique Values

Data Explorer Library:

“In Data Science, 80% of time spent is to prepare data, 20% of time is spent on complaining about need for preparing the data.”

This library helps to explore whole data with almost no code. The library helps to automatic generate a file which would contain:

Data Exploration

Missing Value Profiling

Data Distribution

Principal Component Analysis

Interactive Graphs on data distribution, missing values and principal component analysis.

Chapter-2

Data Pre-Processing

Now we will Clean, Analyse and Prepare the data for modelling.

Missing Value Analysis:

First, we will go for Missing Value Analysis. Missing Values are the values which are often forgotten to be filled during preparation of data. This might happen due to human error or even software error.

E.g. Sometimes while filling the google forms we might miss filling few questions as they won’t be compulsory.

These values are either imputed or deleted as per the industrial standards.

According to industrial standards We should impute missing values on data which contains ≤30% of missing values.

There are 3 Methods to impute Missing Values.

Central Statistic Method

Distance Formula- Knn Imputation

Prediction Method

Important Remark:

While using KNN Method for treating missing values we need to select K Value, we must make sure that selected K Value number is ODD i.e. 1,3,5,7……….

Because If K-Value is EVEN, Algorithm will get confused on which value to be selected as Majority.

In our Data Set, there are total 135 missing observations.

Below are the top five variables with highest missing values:

Body Mass Index- 4.04%

Height- 1.94%

Education- 1.39%

Work Load Average per Day- 1.11%

Transportation Expense- 0.83%

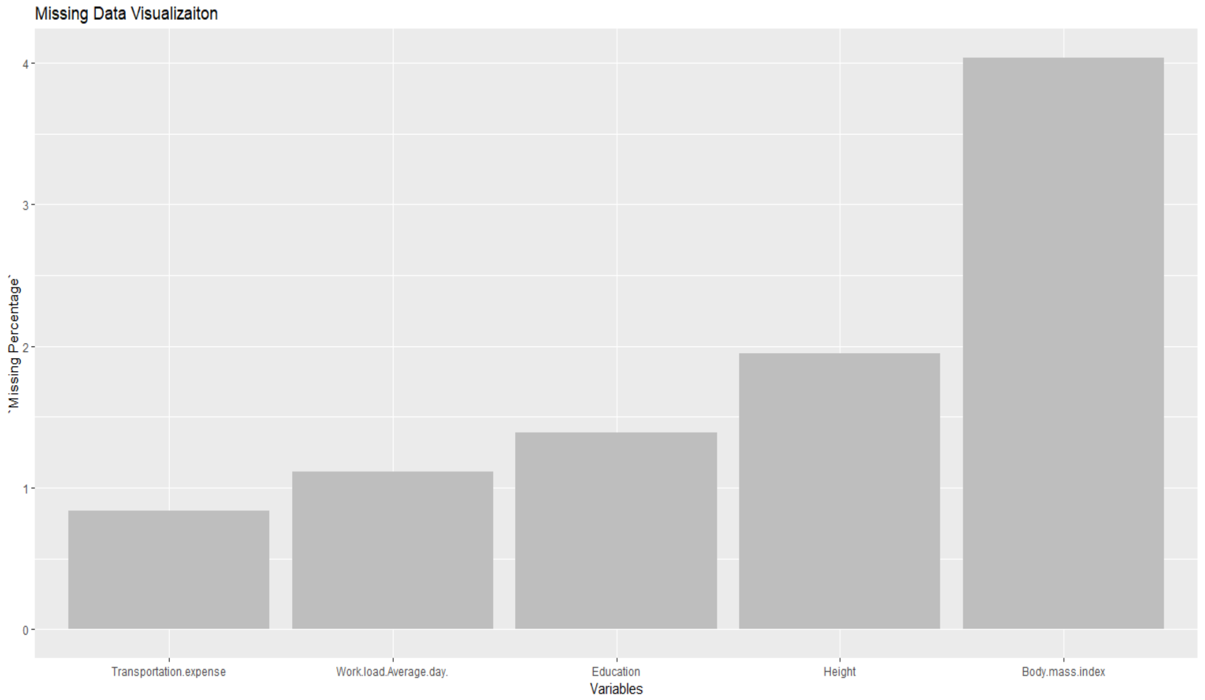


Figure 3: Missing Values

How do we Impute Missing Values?

For categorical variables, we have used Mode method which is a part of central tendency.

For continuous variables, we used three methods i.e. Mean, Median and KNN Imputation. KNN Imputation gave us best results hence we locked KNN method for treating missing values for continuous variables.

Outliers Analysis

Outliers are the data which are highly deviated from original mean.

e.g. I have a bank account. My average monthly transactions are 1.5 Lakh Dollars. Now suddenly, I transacted 10.5 Lakh Dollars. Now this is called outlier.

Outliers Analysis can be used in Fraud Detection, checking MRI Values etc.

What impact Outliers have in a data set?

Suppose we have a data set 1,3,5,7,9. The mean is 5. Now let's introduce outlier 1,3,5,7,9,14. Our mean would be 6.5. we can see that mean is deviated.

While dealing with large data sets it would create lot of problems and can have major impact.

How to detect outliers?

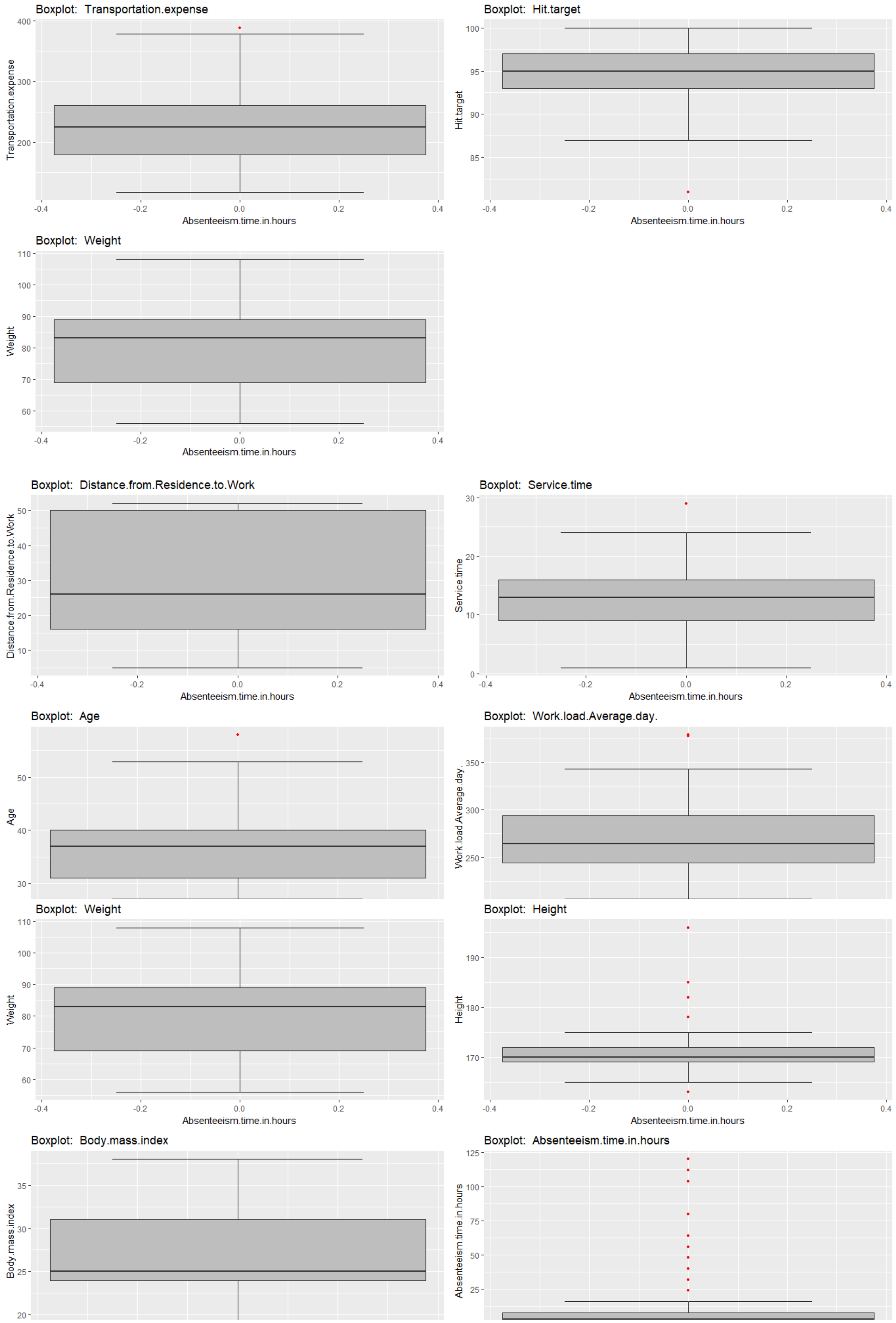
There are several methods at very high level to detect the outliers.

- Box Plot:
    - Data above the Upper Fence and below the Lower Fence would be considered as outliers.
- Statistical Technique- Grubbs Test
    - It has an assumption that data is normally distributed. Now in real cases, there are rare chances that data would be Normally distributed.
- R-Package Outlier
    - There is a package in R named Outliers. This uses concept of Mean to detect outliers.
    - Those data which are highly deviating from Mean would be considered as outliers.
- Replace with NA
    - In this method we calculate outliers. After calculating all outliers, we will replace them with NA.

Now all the NA's would be imputed with suitable missing value analysis method.

Remark:

While plotting distribution curves, we can also see that none of our variable is normally distributed. This might happen due to presence of outliers in our data set.



Boxplot: Transportation.expense

Boxplot: Hit.target

Boxplot: Weight

Boxplot: Distance.from.Residence.to.Work

Boxplot: Service.time

Boxplot: Age

Boxplot: Work.load.Average.day.

Boxplot: Weight

Boxplot: Height

Boxplot: Body.mass.index

Boxplot: Absenteeism.time.in.hours

Figure 4: Box Plot for Outliers

Explanation of Box Plots:

The red dots in the figure are the outliers. From the fig we can observe that, Height and Absenteeism in Hours have highest number of outliers.

Once outliers are detected we will replace them with NA.

After replacing outliers with NA, we will impute them with missing value analysis.

As per our Missing Value analysis observation, we will use KNN method to replace the NA's.

Feature Selection

This method is used to extract relevant and meaningful features out of data.

How does Feature Selection help to extract meaningful information?

This concept is used to reduce complexity of data by reducing variables which carries irrelevant information to explain target variable.

E.g. Suppose we received a data set with 10000 variables, now we cannot feed every variable to the model because not every variable can be used to extract meaningful information. Hence, we would use Feature Selection to reduce complexity of data and select only those variables which carries significant information.

Impact of irrelevant variables in a data set!

E.g. In text mining, suppose we extracted a paragraph with 28000 words. Now during analysis each word would act as variable. This means there would be 28000 variables *This would result in increase in size of irrelevant data. This is termed as Curse of Dimensionality.*

We would now go for two methods which are generally used in Feature Selection.

Correlation Analysis:

This method is used for Continuous or Numerical Variables.

$$Cor(X,Y) = \frac{CoV(X,Y)}{SD(X)SD(Y)}$$

Range of Correlation Values is between -1 to +1

Chi-Square Test

- This method is used for Categorical Variables.
- We develop Hypothesis on basis of Chi-Square Test.

Null Hypothesis- Two Variables are dependent

Alternate Hypothesis- Two variables are independent

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Ei=Expected Values

Oi=Original Values

Degree of Freedom= (number of rows-1) (no of columns -1)

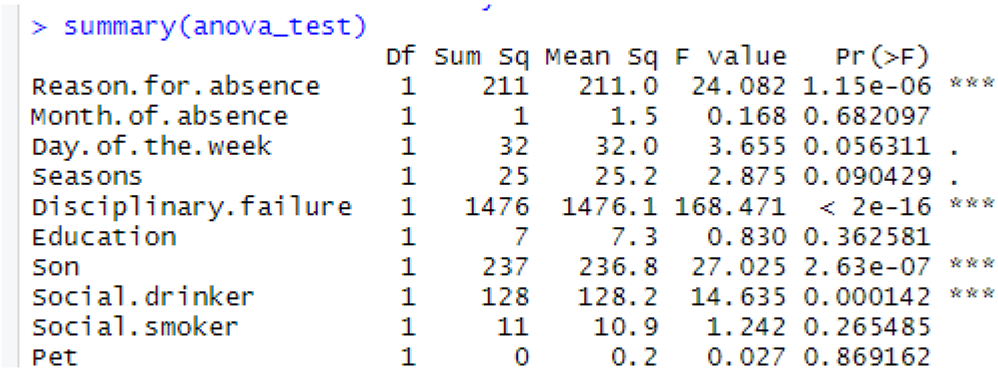
On basis of Chi Square value, we will calculate p-value.

If p-value is < 0.05, This would mean that Alternate Hypothesis is true that means both variables are independent of each other.

If p-value is > 0.05, This would mean that Null Hypothesis is True that mean both variables are dependent on each other.

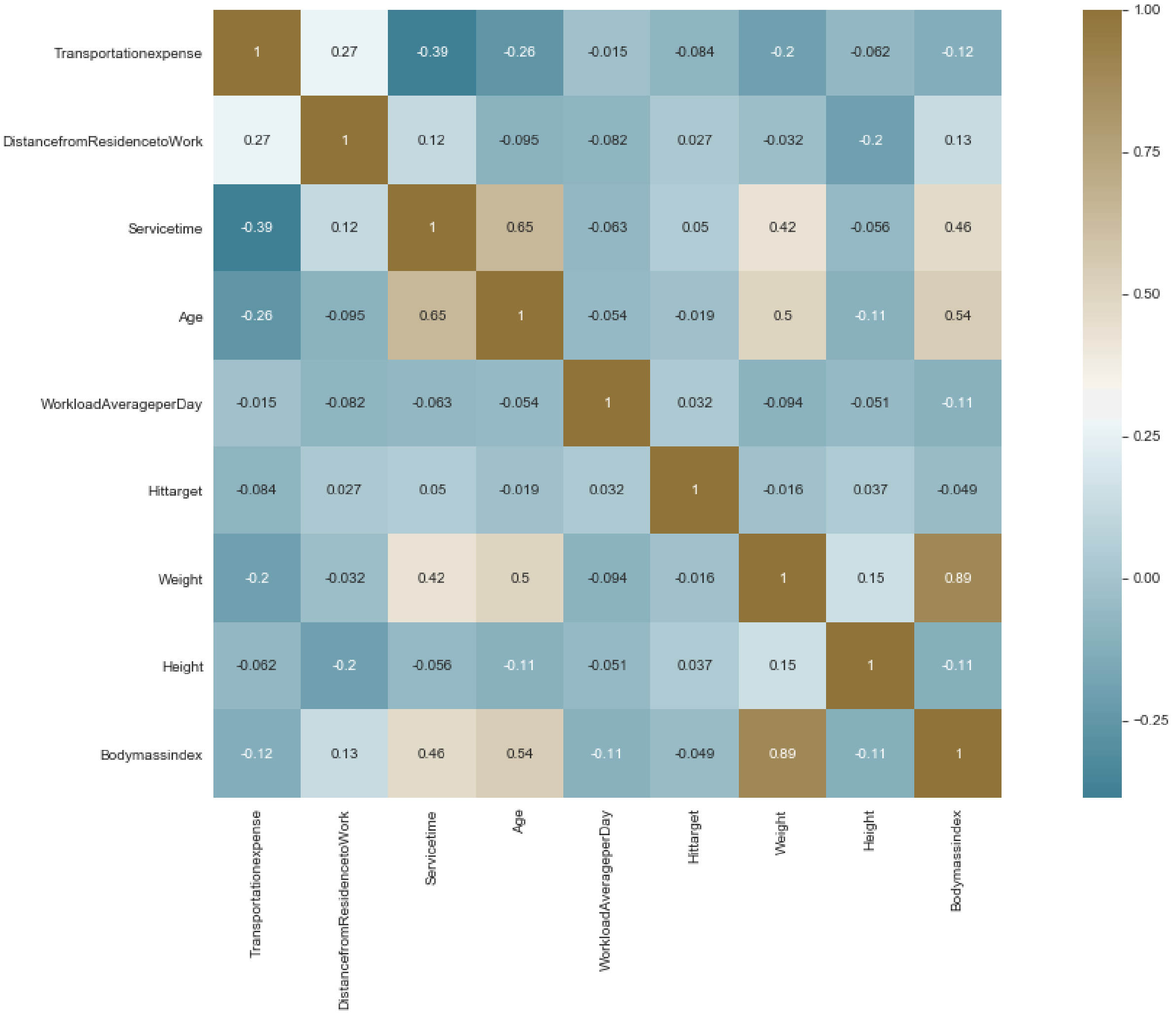
- ANOVA- Analysis of Variance:

ANOVA is a statistical test used to compare mean of two or more groups. It checks the impact of variable present in a data set.



```
> summary(anova_test)
                      Df Sum Sq Mean Sq F value   Pr(>F)    
Reason.for.absence     1    211   211.0  24.082 1.15e-06 ***
Month.of.absence       1      1     1.5   0.168 0.682097    
Day.of.the.week        1     32    32.0   3.655 0.056311 .  
Seasons                1     25    25.2   2.875 0.090429 .  
Disciplinary.failure   1   1476  1476.1 168.471  < 2e-16 ***
Education              1      7     7.3   0.830 0.362581    
Son                    1    237   236.8  27.025 2.63e-07 ***
Social.drinker         1    128   128.2  14.635 0.000142 ***
Social.smoker          1     11    10.9   1.242 0.265485    
Pet                    1      0     0.2   0.027 0.869162    
```

Figure 5 :ANOVA Results



	Transportationexpense	DistancefromResidencetoWork	Servicetime	Age	WorkloadAverageperDay	Hittarget	Weight	Height	Bodymassindex
Transportationexpense	1	0.27	-0.39	-0.26	-0.015	-0.084	-0.2	-0.062	-0.12
DistancefromResidencetoWork	0.27	1	0.12	-0.095	-0.082	0.027	-0.032	-0.2	0.13
Servicetime	-0.39	0.12	1	0.65	-0.063	0.05	0.42	-0.056	0.46
Age	-0.26	-0.095	0.65	1	-0.054	-0.019	0.5	-0.11	0.54
WorkloadAverageperDay	-0.015	-0.082	-0.063	-0.054	1	0.032	-0.094	-0.051	-0.11
Hittarget	-0.084	0.027	0.05	-0.019	0.032	1	-0.016	0.037	-0.049
Weight	-0.2	-0.032	0.42	0.5	-0.094	-0.016	1	0.15	0.89
Height	-0.062	-0.2	-0.056	-0.11	-0.051	0.037	0.15	1	-0.11
Bodymassindex	-0.12	0.13	0.46	0.54	-0.11	-0.049	0.89	-0.11	1

Figure 6:Corrgram Plot

E.g. we have two variables A and B. Now we will calculate mean of both A and B. If mean is same than it would mean that both variable carries same information, we can delete one.

Explanation:

If we check the ANOVA summary for Categorical Variables, The (*) and (.) i.e. Star mark and dot indicates variable importance. (.) means p-value is at the border of 0.05 and * means variable is good to be selected for modelling. Higher the Star mark higher the significance and p-value would be less than 0.05.

In corrgram plot, we can see the correlation of continuous variable with each other. Dark blue colour means positive correlation and Dark red colour means negative correlation.

On basis of above two methods we will select important variables:

- Selected Continuous/Numerical Variables:
    - Transportation expense
    - Distance from Residence to Work
    - Service time
    - Age
    - Work load Average/day
    - Hit target
    - Height
    - Body mass index
    - Absenteeism time in hours
- *#Remark: Body Mass Index itself is a ratio of Height and Weight.*
- Selected Categorical Variables:
    - Reason for absence
    - Day of the week
    - Disciplinary failure
    - Son
    - Social drinker

Feature Scaling-

This is also known as Feature Engineering and Variable Scaling.

E.g. We have a data set with variable named Salary and Age. Now in this case, salary can go to any range 1000$,10000$,50000$ while Age can go maximum till 100. Such type of cases can cause Anomaly in Data.

There are 2 methods named:

Normalization

Standardization

Normalization:

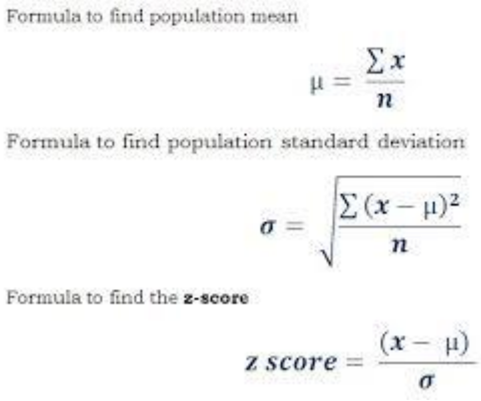
This is used to reduce unwanted variations. This convert our data range in between 0 to 1 and bring to common scale.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Normalization Formula

Standardization:

This will convert each datapoint to unit of Standard Deviation. This method is also knowna as Z-Score.



Formula to find population mean

$$\mu = \frac{\sum x}{n}$$

Formula to find population standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Formula to find the **z-score**

$$z\ score = \frac{(x - \mu)}{\sigma}$$

Standardization Formula

Now question is which method to use,

If most of our data is Normally distributed than we will use Standardization or Z-Score. If our data is not distributed Normally than we will use Normalization.

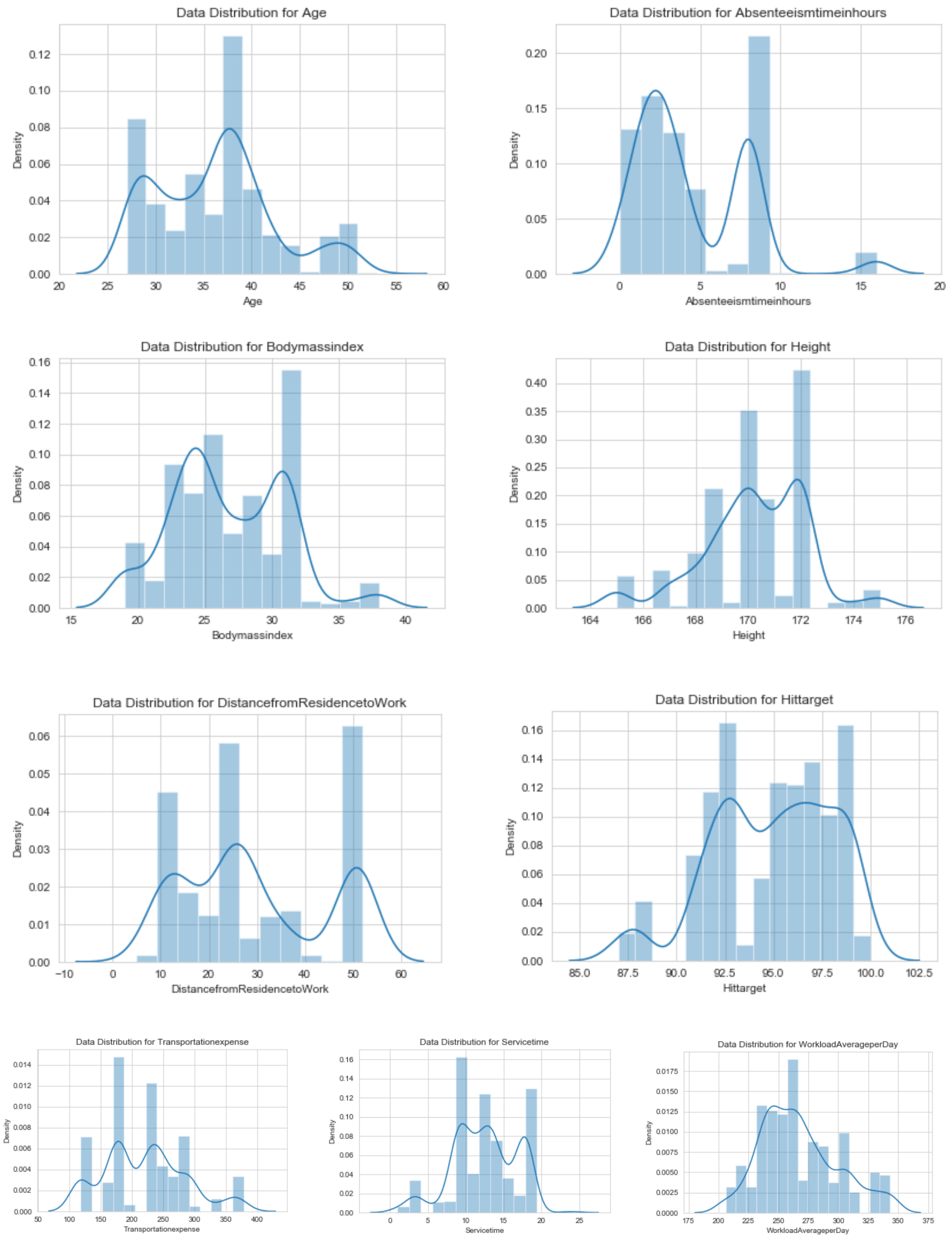


Figure 7: Data Distribution Curves

Explanation:

As we can see from the figure, our data is not at all normally distributed. Hence, we will use Normalization.

Sampling Techniques:

Sampling is a method of selecting a small data from a huge collection of data.

We select sample in such a way that it represents characteristics of whole data.

E.g. Suppose we got a data worth 7 Crore observation. Now this would consume lot of memory. so, we create a subset which would represent entire data

Let’s go through few methods,

Simple Random Sampling-
This is the simplest method.

E.g. Pick any 100 observations out of 26000 observations.

In this method, each observation would have equal chance of getting picked.

Systematic Sampling/Kth Name
Select every Kth observation as sample

K=N/n

N= Number of observations

n= Desired Sample

Suppose N=10000 and n=5000

K=2, This means select every 2nd observation.

There are several disadvantages associated with this method.

Suppose K=2, now in our data every 2nd observation is of same category.

E.g. We have a variable named designation. In designation there are 2 categories Director and Group HR. Now suppose K=2 and in our data every 2nd observation is Group HR. so we would only get those data whose designation is Group HR and hence there won’t be any data with designation as Director. This would lead to our data being biased.

Stratified Sampling

In this method Stratus would be created, Stratums is a subset of population which have same characteristics.

E.g. Suppose we have 100 observations, in these 100 observations there are 30-President,30-Vice President and 10-Directors. Now we want only 10 data out of 100 as sample. This method would create sample with equal proportion such as 3 data of President, 3 data of Vice-President and 1 data of Director.

Now question is which categorical data should be selected to create stratums?

For determining suitable categorical variable, feature selection would be used as we need to select that variable which carries relevant information.

Now by performing Data-Pre-processing, our data is now ready to be fed to Machine Learning algorithm.

Chapter-5

Machine Learning

There are three types of Machine Learning algorithms:

- **Supervised Machine Learning**
- **Un-Supervised Machine Learning**
- **Recommendation System**

Supervised Machine Learning-

- We have target variable in this type of machine learning technique.
- Regression and Classification are types of supervised machine learning.

Unsupervised Machine Learning-

- We do not have any target variable in this type of machine learning technique.
- Clustering and Text Mining comes under Un-supervised machine learning.

Recommendation System-

- When we open Amazon Website or application and select a product to buy. It will show 'similar products.'
- When you are using Instagram, it will show you certain advertisements. That advertisements would be related to past searches on browser history.
- When you are using Twitter, you like certain tweets. Now your news feed would be customised in accordance with tweets you like, retweets etc

Above all are examples of Recommendation Systems.

As per our business problem, we need to forecast 'Absenteeism Time'. The data provided contains a Target Variable Absenteeism Time in Hours which is to be forecasted. Hence, we will use Regression models.

We will use Decision Tree, Random Forest and Linear Regression.

Decision Tree:

- The concept behind decision tree is to generate tree like graphs and obtain rules.
- Decision Tree can be used for both Classification and Regression.
- Rpart in R and Decision Tree Regressor in Python library would be used to build Decision Tree model.
- It provides output in form of rules which becomes very easy for a user to understand the model.
    - For e.g. If a person is consultant then he has an iPhone.
    - If a person is consultant then he is working in KPMG.
    - If a phone has apple behind the model than it is iPhone.

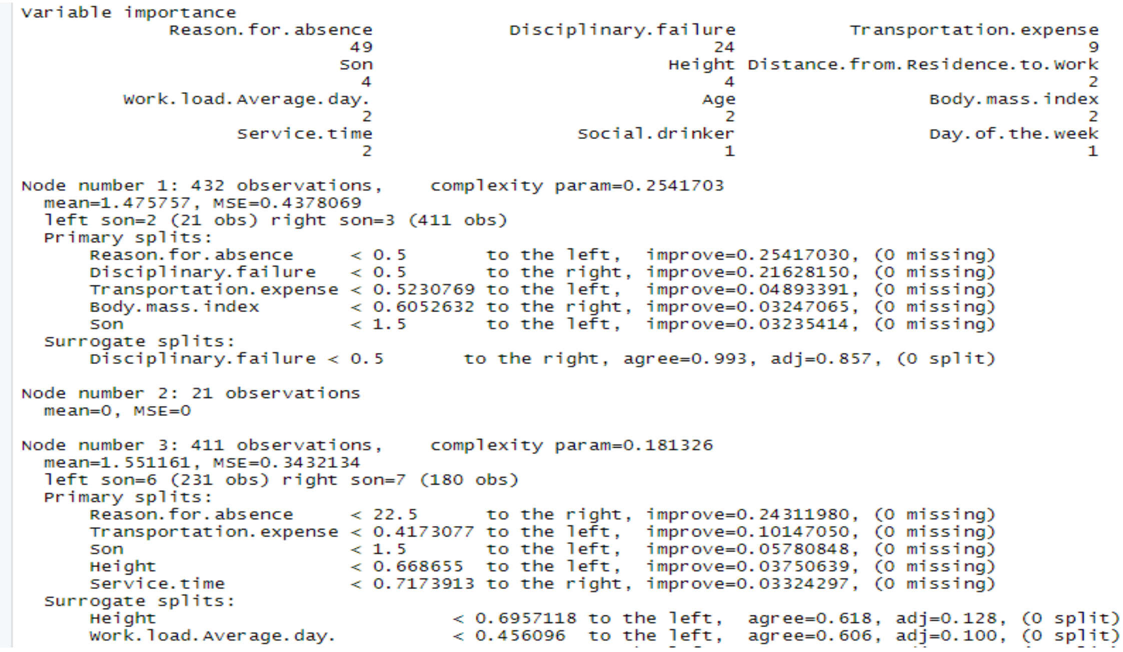
Now when we will solve any business problem, we would get 1000 such rules or even lakhs of rules. so, the question is how to validate those rules and select them.

There are 3 terms which would help validate those rules:

Support: Tells how frequently the item has appeared. Support should be greater than 20%

Confidence (LHS=RHS): Tells how often a rule is True. It should be greater than 80% as per industrial standards.

Lift: Confidence divided by Support gives lift. It should be greater than 1.



```
Variable importance
           Reason.for.absence          Disciplinary.failure        Transportation.expense
                           49                            24                             9
                          Son                        Height Distance.from.Residence.to.Work
                            4                             4                             2
        Work.load.Average.day.                           Age               Body.mass.index
                            2                             2                             2
                 Service.time                Social.drinker               Day.of.the.week
                            2                             1                             1

Node number 1: 432 observations,    complexity param=0.2541703
  mean=1.475757, MSE=0.4378069
  left son=2 (21 obs) right son=3 (411 obs)
  Primary splits:
      Reason.for.absence     < 0.5       to the left,  improve=0.25417030, (0 missing)
      Disciplinary.failure   < 0.5       to the right, improve=0.21628150, (0 missing)
      Transportation.expense < 0.5230769 to the left,  improve=0.04893391, (0 missing)
      Body.mass.index        < 0.6052632 to the right, improve=0.03247065, (0 missing)
      Son                    < 1.5       to the left,  improve=0.03235414, (0 missing)
  Surrogate splits:
      Disciplinary.failure < 0.5       to the right, agree=0.993, adj=0.857, (0 split)

Node number 2: 21 observations
  mean=0, MSE=0

Node number 3: 411 observations,    complexity param=0.181326
  mean=1.551161, MSE=0.3432134
  left son=6 (231 obs) right son=7 (180 obs)
  Primary splits:
      Reason.for.absence     < 22.5      to the right, improve=0.24311980, (0 missing)
      Transportation.expense < 0.4173077 to the left,  improve=0.10147050, (0 missing)
      Son                    < 1.5       to the left,  improve=0.05780848, (0 missing)
      Height                 < 0.668655  to the left,  improve=0.03750639, (0 missing)
      Service.time           < 0.7173913 to the right, improve=0.03324297, (0 missing)
  Surrogate splits:
      Height                          < 0.6957118 to the left,  agree=0.618, adj=0.128, (0 split)
      Work.load.Average.day.          < 0.456096  to the left,  agree=0.606, adj=0.100, (0 split)
```

Figure 8: Decision Tree Output

Random Forest:

The concept behind Random Forest is to generate ‘n’ number of trees to make our model accurate.

Why Random Forest is superior to Decision Tree?

Suppose we have a data with crores of observations. Now a single decision tree will not be able to cover entire data to get proper meaning out of it. In such cases we will use Random Forest.

Random Forest works on 2 algorithms:

Biermann’s Bagging Idea- Once we will build a decision tree there would be some errors. Now when next tree is built, we will feed that error to next tree to get meaningful decision out on it.

Random Selection of Features- It is random selection of data for building a decision tree.

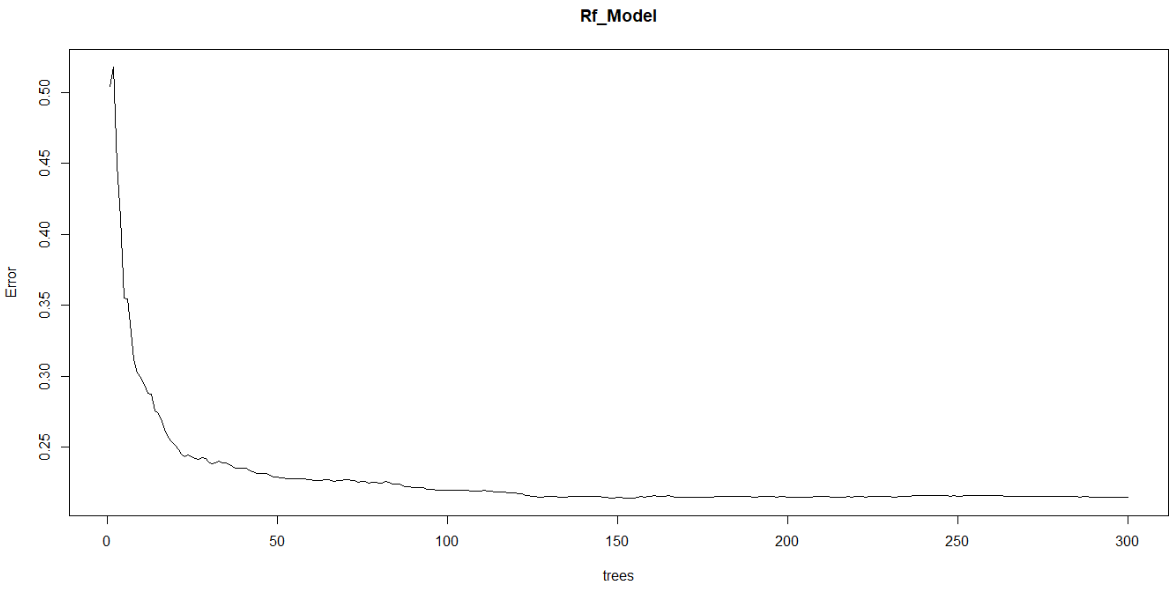


Figure 9: Random Forest

Explanation:

As we can see from the figure, As the number of trees increases error rate decreases. After building 130 trees, error rate is constant.

Key Points:

Why should we use Random Forest?

Higher the number of trees, higher the accuracy and parameters of our model. We can build 100,300,500 trees.

Random Forest can be used for a larger data set with 1000 independent variables without any variable getting deleted.

This model can be used for both Classification as well as Regression.

Random Forest works on concept of Gini Index. Gini Index measures impurity of data. We will calculate Gini Index and select that variable as parent node whose Gini Index is Lowest.

Linear Regression:

Linear Regression is a type of statistical model. There is high difference between Machine Learning models and Statistical models. In ML models we extract rules and patterns while in SM we deal with coefficients or weights of each variable.

E.g. We have 10 Independent variables; we will calculate weight of each independent variable to see how it explains target variable.

Terms Used in Linear Regression:

R Square: It explains the overall measure of association. E.g. If R-Square value is 86%, that would mean that our Target variables is explained 86% by all Independent Variables combined.

Adjusted R Square: It is derived from R-Square.

Why should we go for Adjusted R-Square?

Now at production stage if we keep feeding our models with more and more Independent variables. Multicollinearity would increase. This would result in Over-Fitting. Now Adjusted R-Square will help nullify effect of Overfitting.

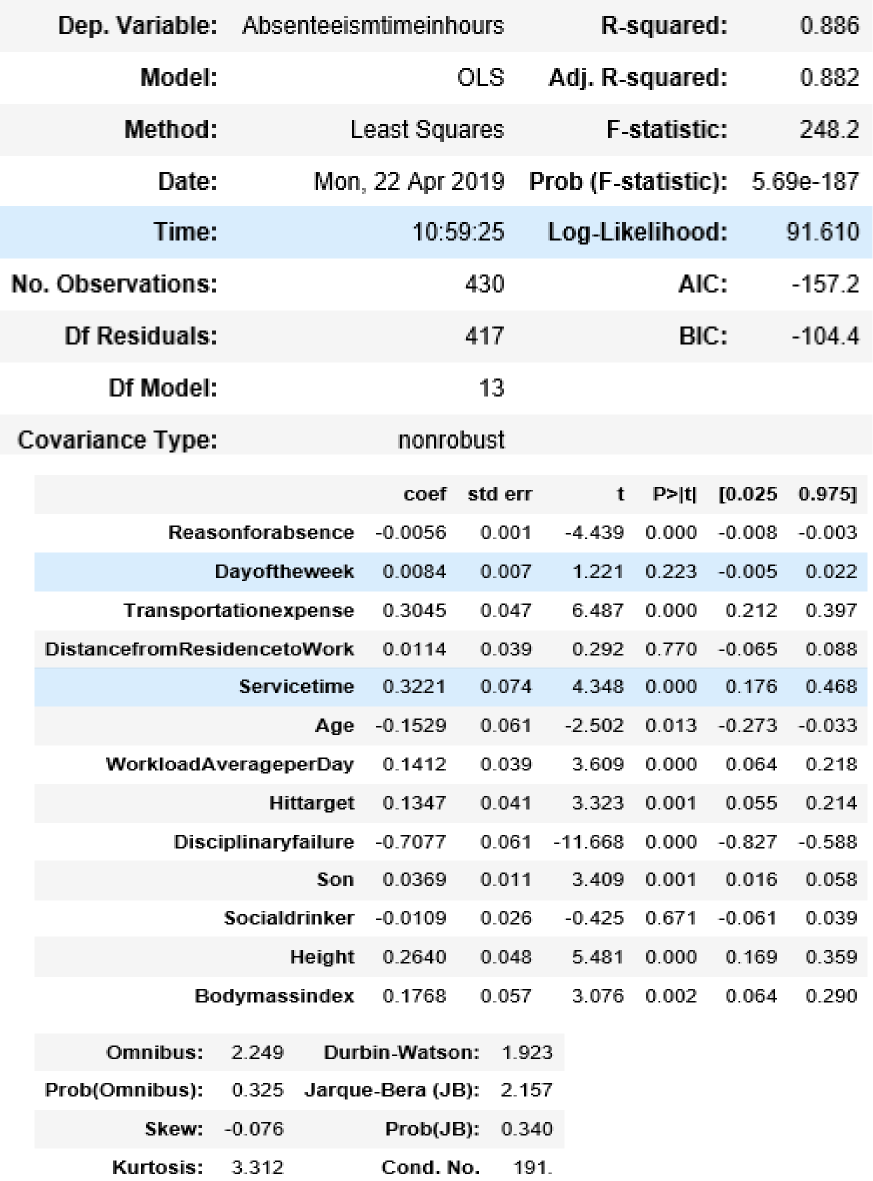
Assumptions of Linear Regression:

Linear Relationship: There is Linear Relationship existing between Independent Variable and Target Variable.

Normality: Data is Normally distributed.

No Multicollinearity: There is no multicollinearity effect.

No Autocorrelation: All the errors generated while developing a model does not follow same pattern. They are unique.



Dep. Variable:	Absenteeismtimeinhours	**R-squared:**	0.886
Model:	OLS	**Adj. R-squared:**	0.882
Method:	Least Squares	**F-statistic:**	248.2
Date:	Mon, 22 Apr 2019	**Prob (F-statistic):**	5.69e-187
Time:	10:59:25	**Log-Likelihood:**	91.610
No. Observations:	430	**AIC:**	-157.2
Df Residuals:	417	**BIC:**	-104.4
Df Model:	13		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Reasonforabsence	-0.0056	0.001	-4.439	0.000	-0.008	-0.003
Dayoftheweek	0.0084	0.007	1.221	0.223	-0.005	0.022
Transportationexpense	0.3045	0.047	6.487	0.000	0.212	0.397
DistancefromResidencetoWork	0.0114	0.039	0.292	0.770	-0.065	0.088
Servicetime	0.3221	0.074	4.348	0.000	0.176	0.468
Age	-0.1529	0.061	-2.502	0.013	-0.273	-0.033
WorkloadAverageperDay	0.1412	0.039	3.609	0.000	0.064	0.218
Hittarget	0.1347	0.041	3.323	0.001	0.055	0.214
Disciplinaryfailure	-0.7077	0.061	-11.668	0.000	-0.827	-0.588
Son	0.0369	0.011	3.409	0.001	0.016	0.058
Socialdrinker	-0.0109	0.026	-0.425	0.671	-0.061	0.039
Height	0.2640	0.048	5.481	0.000	0.169	0.359
Bodymassindex	0.1768	0.057	3.076	0.002	0.064	0.290

Omnibus:	2.249	**Durbin-Watson:**	1.923
Prob(Omnibus):	0.325	**Jarque-Bera (JB):**	2.157
Skew:	-0.076	**Prob(JB):**	0.340
Kurtosis:	3.312	**Cond. No.**	191.

Figure 10:Summary of Linear Regression

Chapter-6

Evaluating Performance of Model

Once we build several models, it would be the time to evaluate them. In our previous project which was a Classification problem we used Confusion Matrix.

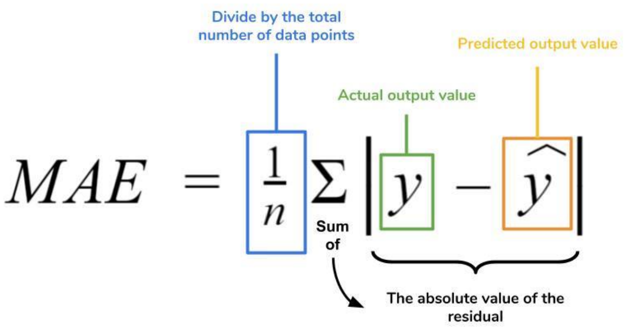
In Regression, Confusion Matrix won’t be used. There are other concepts.

Below are the concepts used for evaluating performance of Regression model:

MAE or MAD- Mean Absolute Error/Deviation

It is average of absolute errors.

Below is the formula of MAE:


Divide by the total number of data points
Predicted output value
Actual output value
$MAE = \frac{1}{n}\sum|y - \hat{y}|$
Sum of
The absolute value of the residual

Formula: Mean Absolute Error

MAPE- Mean Absolute Percentage Error

In MAPE we would get actual percentage of error.

It would be easy for our client to understand as the number would be percentage.

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

where A_t is the actual value and F_t is the forecast value.

Formula: Mean Absolute Percentage Error

- RMSE/RMSD- Root Mean Square Error/Deviation

It squares the error finds their average and take square-root of that average value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Formula: Root Mean Square Error

Now question is which method is the best:

If we have time series data, we should go for Root Mean Square Error (RMSE).

If we want error in percentage, we would go for MAPE.

If we want result in deviation of error, we would use MAE.

At very high level, MAPE and RMSE is used.

Now I will enlist Performance of Model.

I have divided performance in several types for brief research in Data Science. Both R and Python would have two types of models. One with Dummies in categorical variables and one model without dummies in categorical variables.

Decision Tree Metrics:

Model	R (DA)RMSE	R (DN)(RMSE)	Python (DA)(RMSE)	Python (DN)(RMSE)
Decision Tree	0.481932668	0.4973766	0.2015483	0.207643684

Model	R (DA)(Rsq)	R (DN)((Rsq))	Python (DA)((Rsq))	Python (DN)(Rsq)
Decision Tree	0.459862289	0.487924049	0.160311503	0.258133449

DA= Dummies Applied, DN=Dummies Not Applied

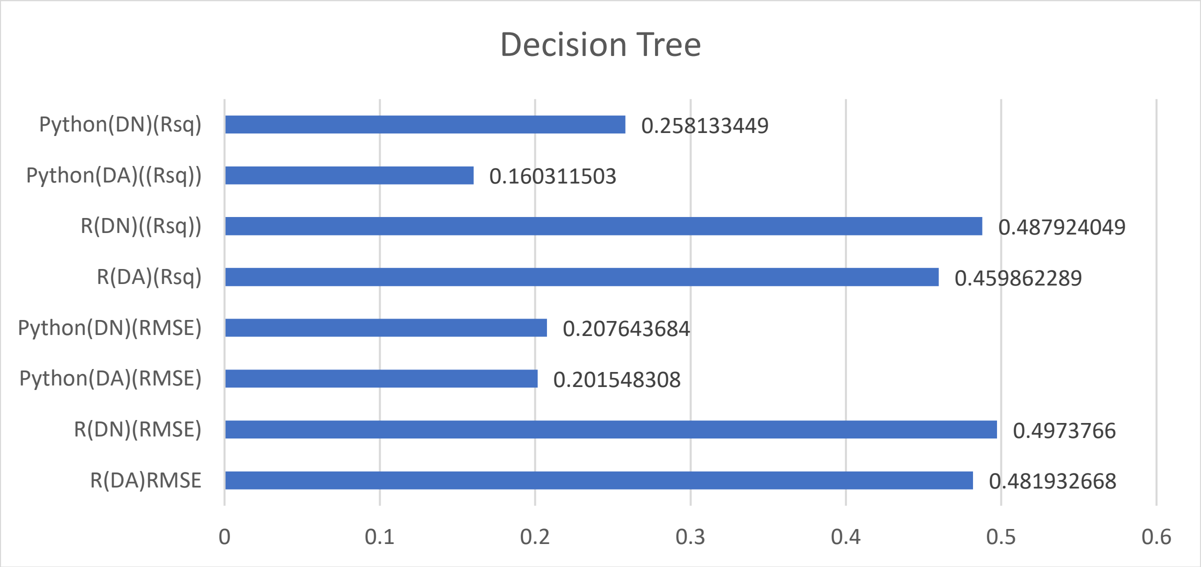


Figure 11: Decision Tree Performance

Random Forest Metrics:

Model	R (DA)RMSE	R (DN)(RMSE)	Python (DA)(RMSE)	Python (DN)(RMSE)
Random Forest	0.481932668	0.4710576	0.150765292	0.16571349

Model	R (DA)(Rsq)	R (DN)((Rsq))	Python (DA)((Rsq))	Python (DN)(Rsq)
Random Forest	0.530532877	0.539493746	0.530146319	0.527497547

DA= Dummies Applied, DN=Dummies Not Applied

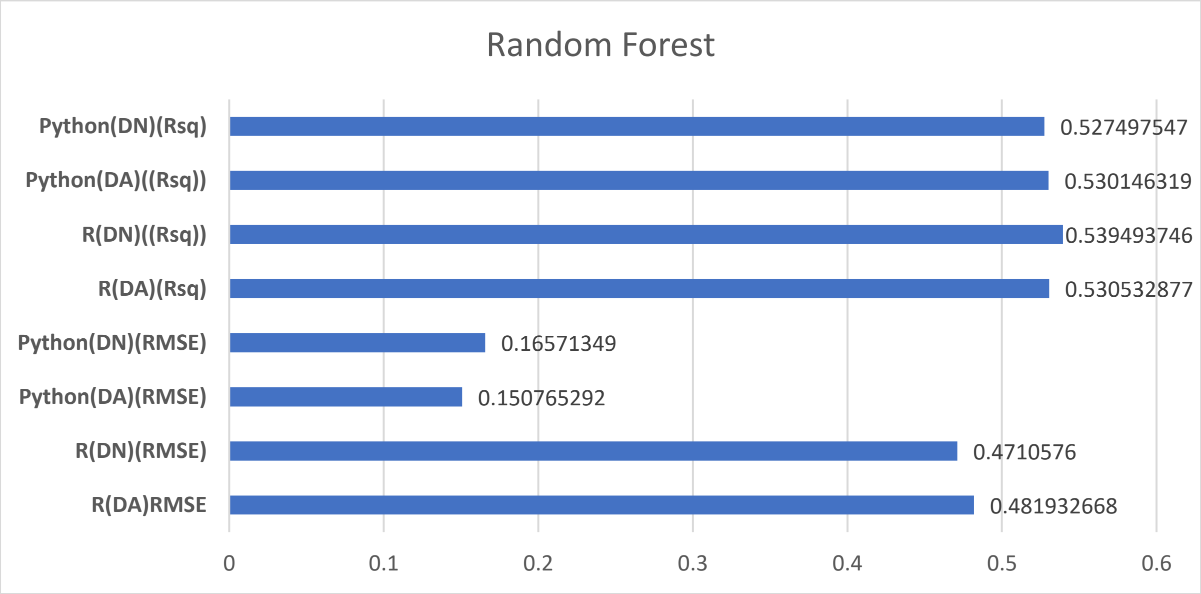


Figure 12:Random Forest Performance

Linear Regression Metrics:

Model	R (DA)RMSE	R (DN)(RMSE)	Python (DA)(RMSE)	Python (DN)(RMSE)
Linear Regression	0.491481033	0.4762331	0.1533122	0.203513892

Model	R (DN)((Rsq))	R (DA)(Rsq)	Python (DA)((Rsq))	Python (DN)(Rsq)
Linear Regression	0.481706827	0.434077176	0.514137506	0.287349718

DA= Dummies Applied, DN=Dummies Not Applied

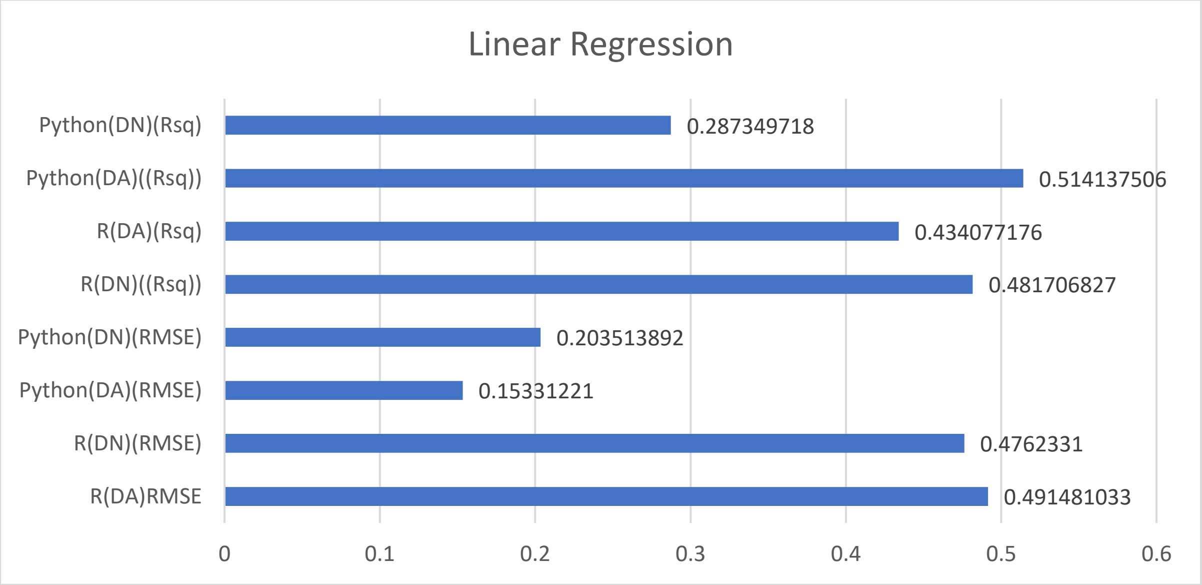


Figure 13:Linear Regression Performance

After evaluating all the models, we would choose Linear Regression (without Dummies Applied). As the RMSE and R-Square value are decent and Adjusted R-Square value is 88%.

As per current scenario, we should focus on Adjusted R-Square as we do not know what future data holds.

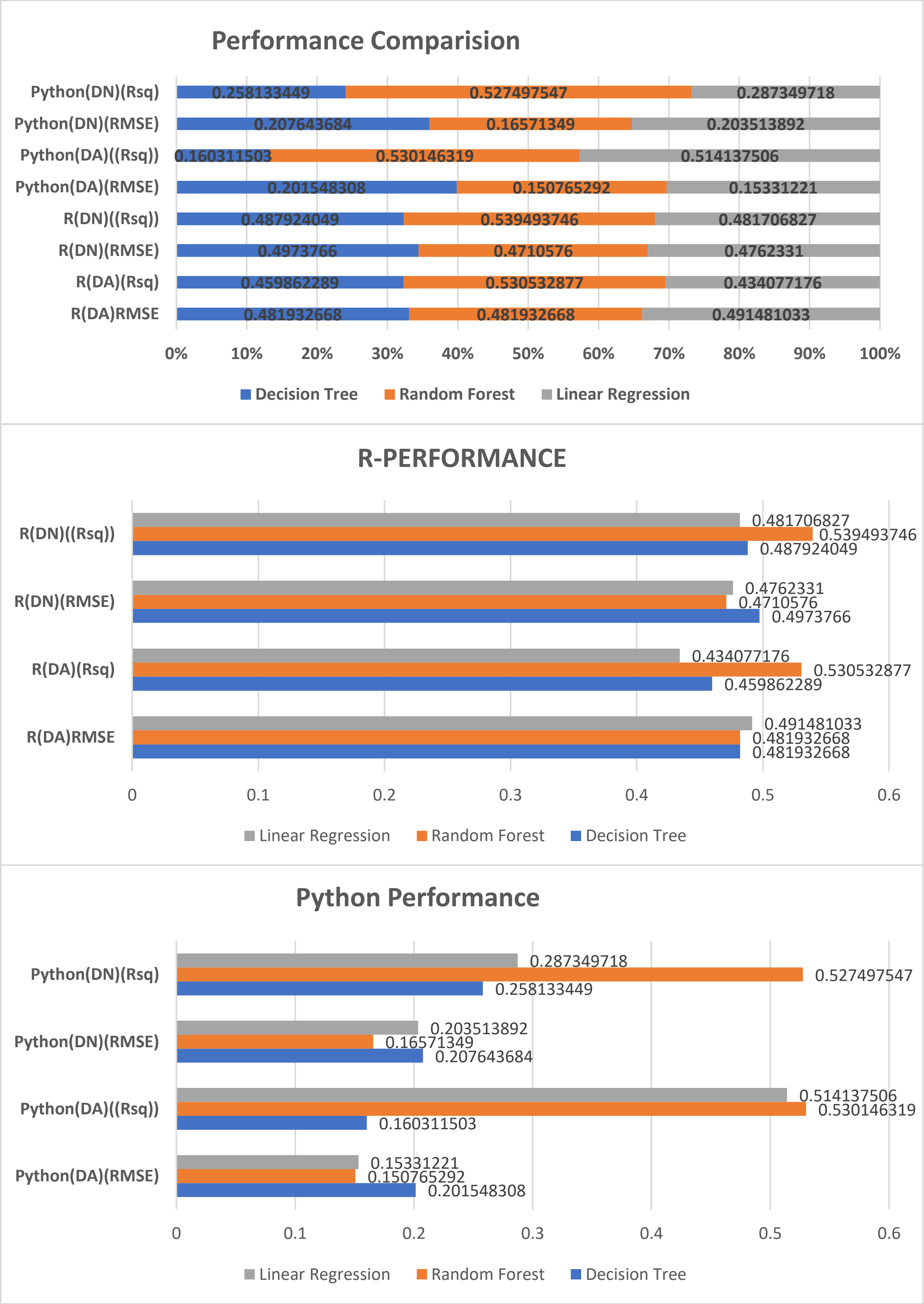


Figure 14: Performance Evaluation: R vs Python

Chapter 7

Visualization:

Data Visualization is one of the most integral part of any analysis project. Visualization is used to represent data in such a way that even a non-catalyst person can also understand.

Tableau is one of the fasted BI (Business Intelligence) tool. It is very easy to learn and has advanced concepts associated with it.

Storytelling and Dashboards are the salient features of Tableau.

Key Insights:

Suggestions for Organisation-Category Wise

Forecast for Work Loss and Absenteeism for coming 14 months

- **Education vs Absenteeism**
  - *84.43% of Absenteeism is from Category 1 i.e. High School people.*

- **Reason of Absence vs Absenteeism**
  - *Highest Absenteeism rate was for Category 13 i.e. 17.45%, Category 19 on second position at 16.69%.*
  - Category 13 is diseases of the musculoskeletal system and connective tissue majorly happens due to physical workload, smoking, obesity and poor nutrition. At the further end, it also results in morbidity.
  - Category 19 is External causes of morbidity and mortality which is condition of being unhealthy due to some illness.
- **Complex Relations and Insights**

Education vs Reason of Absenteeism vs Work Load

Education Category: High School

Workload shared by them is 81.30%

Absenteeism percentage shared is 84.33%

What organisation should do for High school category Employees?

- Issue of musculoskeletal system and connective tissue was observed which might have resulted in morbidity, injury, poisoning and other external diesases .
- Due to these reasons, they might have to consult for medical or dental consolation.
- **Adding all these categories would be 49.59%.**
- ***Organisation should divide Work Load properly so that a particular person won't get prune to various diseases.***
- ***A medical doctor at the organisation who would provide check-up's at regular time interval would be beneficial.***

Education Category: Graduate

Workload shared by them is 6.35%

Absenteeism percentage shared is 5.71%

What organisation should do for Graduate category employees?

- Issue of Digestive System, Abnormal clinical findings and Neoplasms are found which requires laboratory tests which requires several visits.
- A medical consultant might be helpful.
- ***If organisation would provide flexible work hours. It would be helpful for employees to get their tests done and also provide work.***

Education Category: Post Graduates

Workload shared by them is 10.53%

Absenteeism percentage shared is 8.3%

What organisation should do for Graduate category employees?

- Major reason for absence found was Infectious & Parasitic Disease, musculoskeletal system and connective tissue & Diseases of the skin and subcutaneous tissue (Energy Reserve- Tissue mostly found in upper-arm, thigh, abdomen & buttocks) which leads to Medical Consultation.
- Combined these major reason absenteeism percentage reaches to 61.54%

- One common thing found is this category shares 2nd place when it comes to workload. So as seen in High-School category students issue of musculoskeletal system and connective tissue is found here.
- ***As discussed before Medical Consultation is required in the organisation itself and proper distribution of workload is required.***

Education Category: Master/Doctors

Workload shared by them is 0.49%

Absenteeism percentage shared is 0.42%

This category shares minimal workload and minimum absenteeism. Moreover, number of employees of this category as also minimal.

What organisation should do for Graduate category employees?

- This category has disease related to genitourinary system hence they require follow ups. Combined absenteeism percentage due to these reasons is 76.2%.
- Other two major reasons include issues of Injury and Poisoning as well as Dental Consultation.
- ***As discussed, a medical and dental consultant is very much required in this organisation***
- ***90% of their issues can be solved with the help of Medical and Dental consultants inside the organisation.***

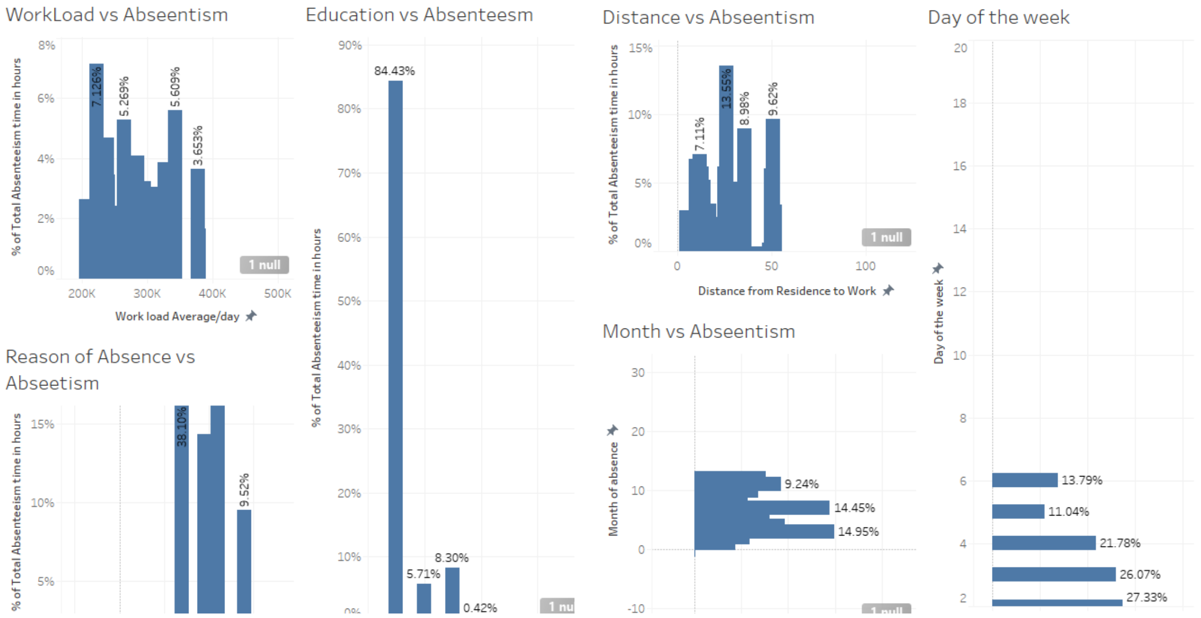
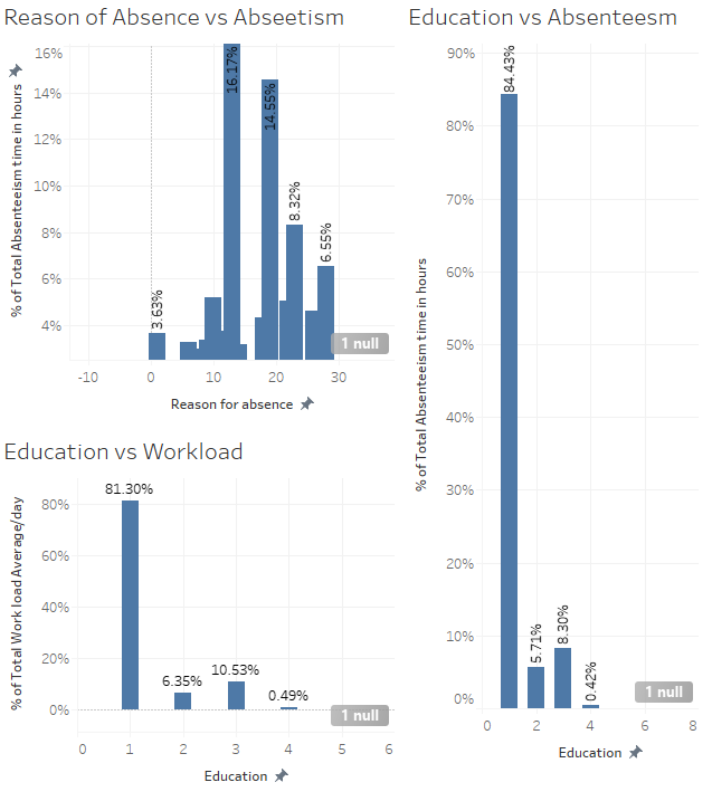


Figure 15:Tableau Dashboard 1


Reason of Absence vs Abseetism

Education vs Absenteesm

Education vs Workload

Figure 16:Tableau Dashboard 2

How much of work loss can be predicted for coming year?

- I've used excel Forecasting formula to detect work loss in coming time.
- Lets first go through current scenario.

Month	Sum of Absenteeism.time.in.hours	Sum of Work.load.Average.day.	Sum of Work Loss
1	171.4590381	15746.35276	2257.655825
2	279.4983086	19454.925	3168.956879
3	440.7239972	23663.868	5166.325374
4	255.7509108	14567.117	2919.720731
5	260.0849507	15411.33774	2721.608401
6	242.3733478	14007.03494	2663.284462
7	364.2530002	16258.458	3838.244763
8	236.4222543	12380.75	2326.62139
9	180.6778073	13319.239	2040.446016
10	281.4240948	17827.70605	3161.337906
11	247.2886958	16764.71102	2943.487932
12	198.0990289	12353.96566	2139.465477
Grand Total	**3158.055435**	**191755.4652**	**35347.15516**

Title	Value
Maximum Work Load	23663.87
Maximum Absenteeism	440.724
Max Work Loss	5166.325

- As march month is financial year end, Work Load was at its peak. In fact, absenteeism and work loss is also at peak. So, when the time was to deliver the best this organisation saw highest absenteeism rate and work loss.

- **Let's see what would happen in next 14 months.**

Month.of.absence	Work Loss	Forecast(Work Loss)	Lower Confidence Bound(Work Loss)	Upper Confidence Bound(Work Loss)
1	45.1531165			
2	44.01328998			
3	58.70824289			
4	55.08907039			
5	43.89690969			
6	50.25065023			
7	59.04991944			
8	44.74271903			
9	40.80892033			
10	47.89905919			
11	49.88962596			
12	44.57219744	44.57219744	44.57	44.57
13		36.10211341	29.08	43.13
14		39.26441369	32.24	46.29
15		49.66028992	42.64	56.68
16		42.78612402	35.76	49.81
17		32.93916126	25.85	40.02
18		36.10146154	29.02	43.19
19		46.49733777	39.41	53.58
20		39.62317188	32.54	46.71
21		29.77620911	22.63	36.92
22		32.93850939	25.79	40.09
23		43.33438562	36.19	50.48
24		36.46021973	29.31	43.61

Chart-Forecast Chart for next 14 Months

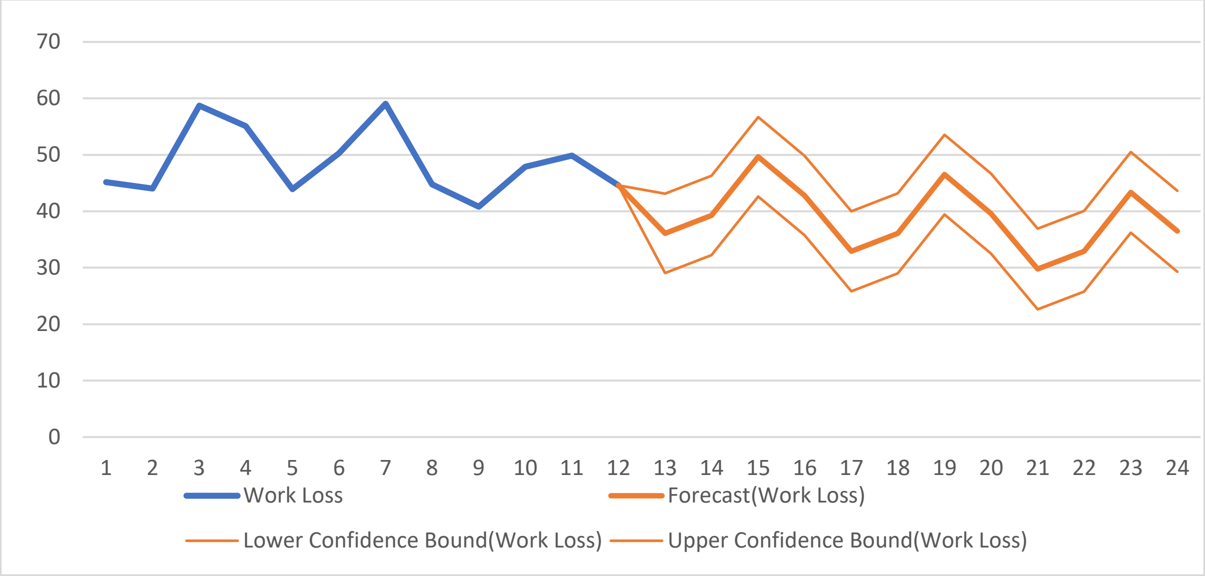


Figure 17:Forecast Chart for next 14 months

New Concepts Used:

Ludwig Algorithm:

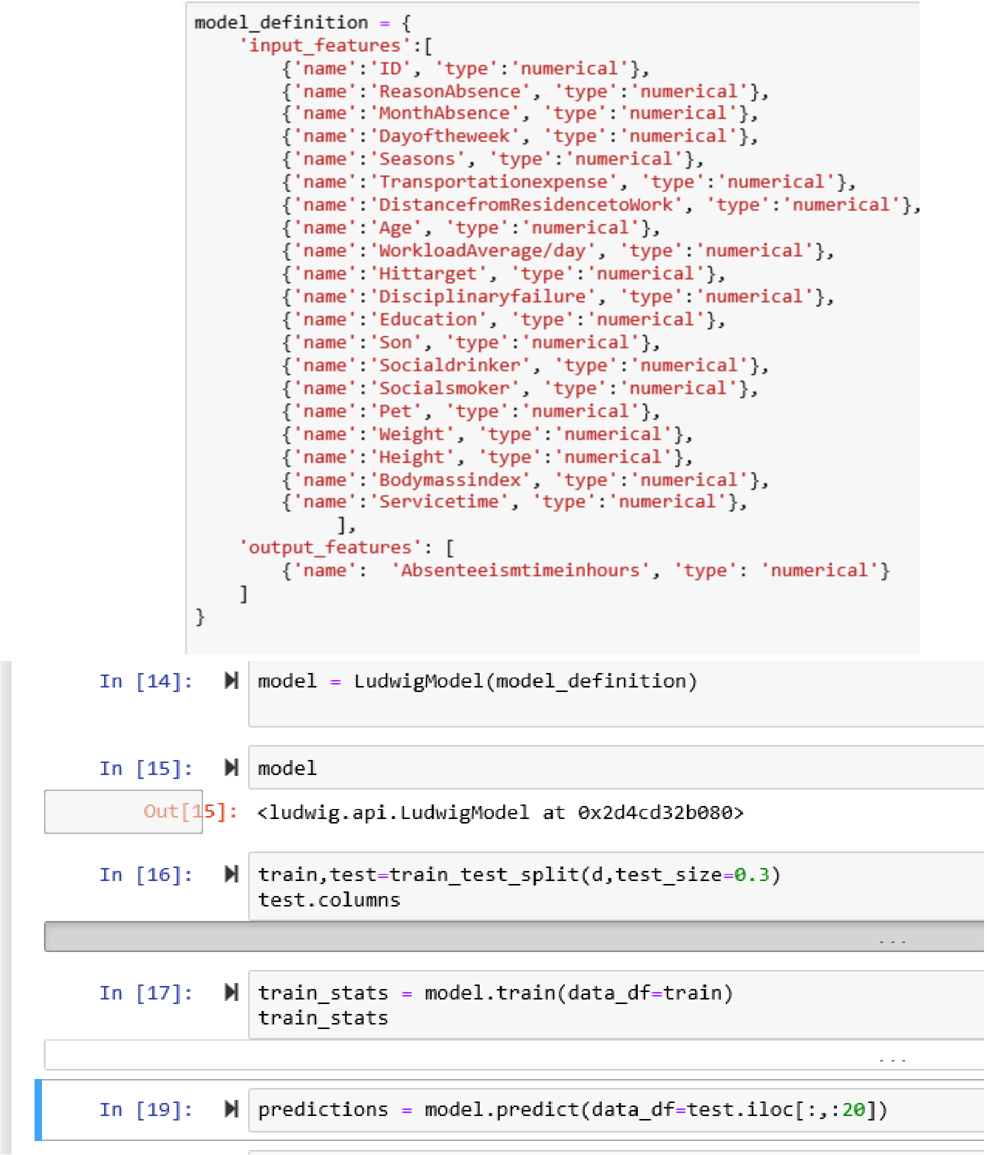
- The algorithm is launched by UBER- Artificial Intelligence team. They are using this algorithm since past many years but just before few months they made this algorithm public in GitHub.
- I tried to apply on our data set and was successfully able to do that after several failed efforts.
- The Error Metrics I got was very high RMSE was around 2.34 in numbers which is not at all acceptable.

- **Description:**
  - Ludwig is a TensorFlow based toolbox that allows to train and test deep learning models without the need to write code.

- **Properties:**
  - Requires Minimal Coding.
  - WE can perform Sentiment Analysis, Natural Language Processing, Classification, Regression and what not!

- We can also perform Deep Learning models.
- Currently visualizations are not available for Jupyter Environment.
- This algorithm is going to make disruptive change in coming time.

Syntax:



```
model_definition = {
    'input_features':[
        {'name':'ID', 'type':'numerical'},
        {'name':'ReasonAbsence', 'type':'numerical'},
        {'name':'MonthAbsence', 'type':'numerical'},
        {'name':'Dayoftheweek', 'type':'numerical'},
        {'name':'Seasons', 'type':'numerical'},
        {'name':'Transportationexpense', 'type':'numerical'},
        {'name':'DistancefromResidencetoWork', 'type':'numerical'},
        {'name':'Age', 'type':'numerical'},
        {'name':'WorkloadAverage/day', 'type':'numerical'},
        {'name':'Hittarget', 'type':'numerical'},
        {'name':'Disciplinaryfailure', 'type':'numerical'},
        {'name':'Education', 'type':'numerical'},
        {'name':'Son', 'type':'numerical'},
        {'name':'Socialdrinker', 'type':'numerical'},
        {'name':'Socialsmoker', 'type':'numerical'},
        {'name':'Pet', 'type':'numerical'},
        {'name':'Weight', 'type':'numerical'},
        {'name':'Height', 'type':'numerical'},
        {'name':'Bodymassindex', 'type':'numerical'},
        {'name':'Servicetime', 'type':'numerical'},
             ],
    'output_features': [
        {'name':  'Absenteeismtimeinhours', 'type': 'numerical'}
    ]
}
```

```
In [14]:   model = LudwigModel(model_definition)

In [15]:   model
Out[15]:   <ludwig.api.LudwigModel at 0x2d4cd32b080>

In [16]:   train,test=train_test_split(d,test_size=0.3)
           test.columns

In [17]:   train_stats = model.train(data_df=train)
           train_stats

In [19]:   predictions = model.predict(data_df=test.iloc[:,:20])
```

Figure 18: Ludwig Syntax

- Performance was not up to the mark. As RMSE value obtained is 2.42 and R Square Score is -105.66.

Data Explorer Library:

This library helps to create basic analysis report with interactive charts. This library generates a report which contains Missing Value calculation, Data Distribution, Data Analysing report.

Here is how the report looks like:

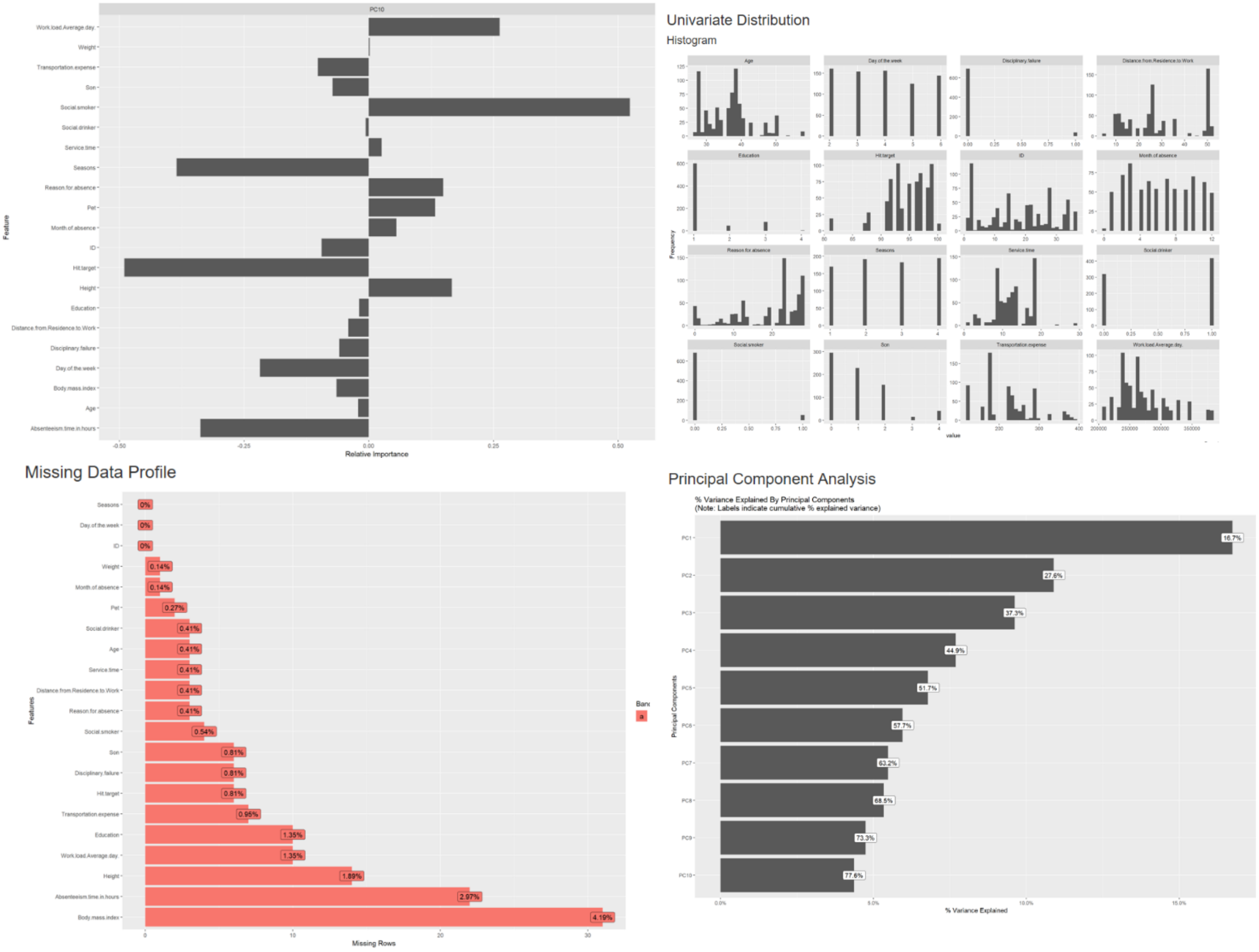


Figure 19: Data Explorer Report Sample