Project Report

Employee Absenteeism

July 10, 2019

Contents

Chapter 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?

1.2 Data

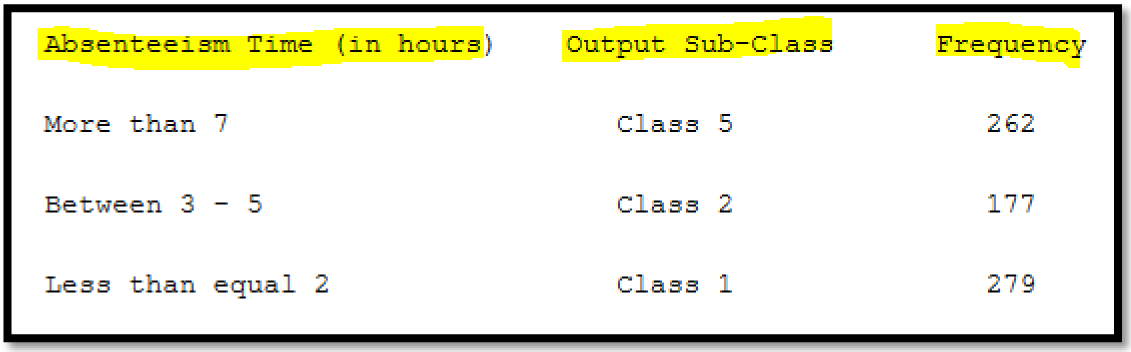
There are 21 variables in our data in which 20 are independent variables and 1 (Absenteeism time in hours) is dependent variable. Though our target variable is continuous in nature, we have classified the output set into sub-ranges from a possible 120 to 5 bins. Thus, we have converted this machine learning problem into a classification problem.

Hypothesis generation

These are some of the hypothesis which could affect the absenteeism time:

1. Certain medical condition will lead to higher absenteeism.

2. Higher transportation expense and 'distance from residence to work' will lead to higher absenteeism.



Absenteeism Time (in hours)	Output Sub-Class	Frequency
More than 7	Class 5	262
Between 3 - 5	Class 2	177
Less than equal 2	Class 1	279

Variables Information:

1. Individual identification (ID)

2. Reason for absence (ICD) -

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 behavioral 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).

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)

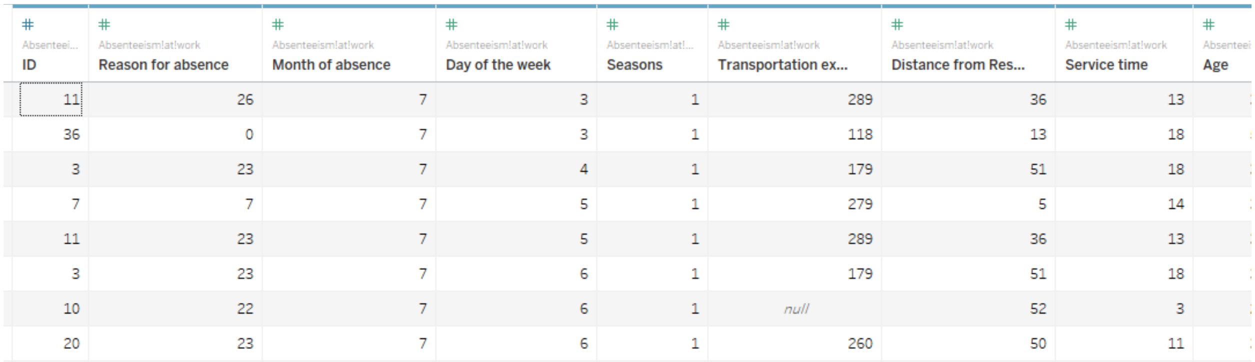
18. Weight

19. Height

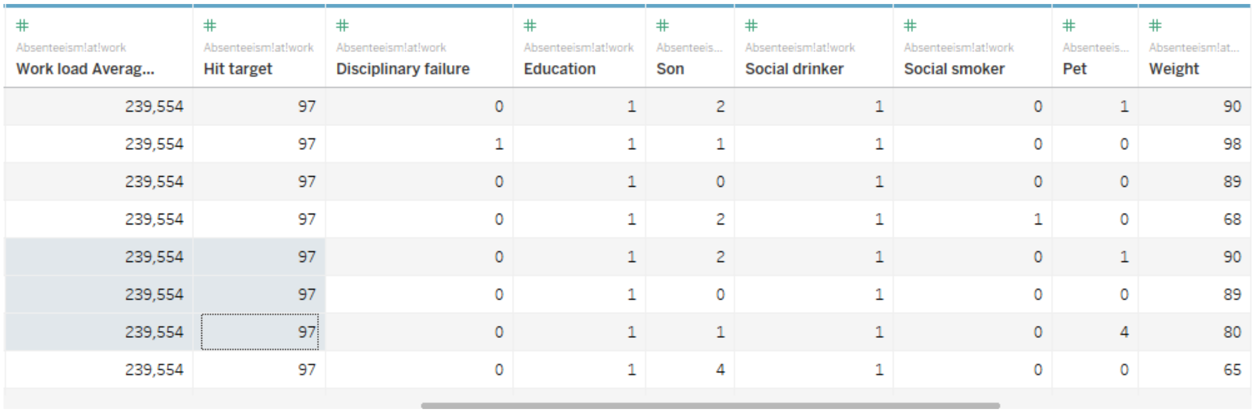
20. Body mass index

21. Absenteeism time in hours (target)

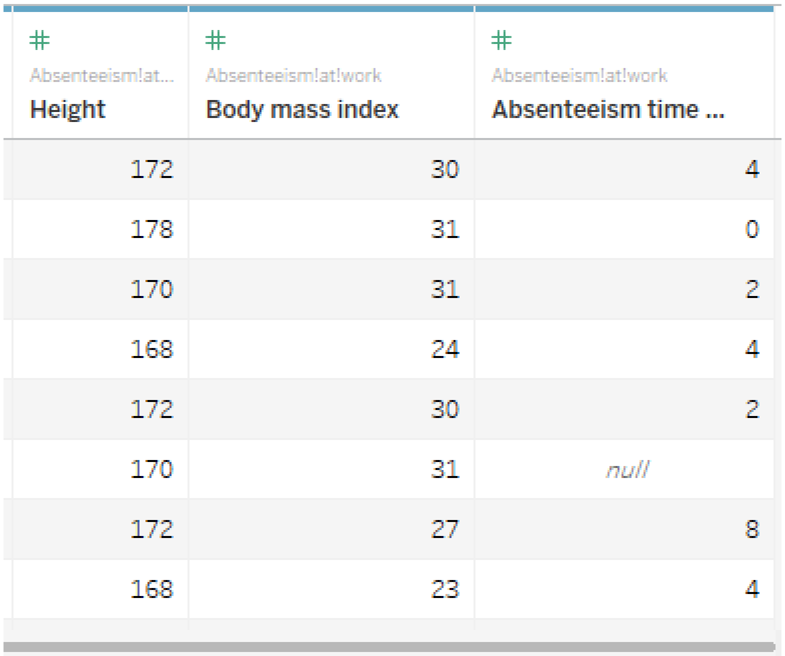
Sample of the dataset:



# Absenteei...	# Absenteeism!at!work	# Absenteeism!at!work	# Absenteeism!at!work	# Absenteeism!at!...	# Absenteeism!at!work	# Absenteeism!at!work	# Absenteeism!at!work	# Absenteei
ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation ex...	Distance from Res...	Service time	Age
11	26	7	3	1	289	36	13	
36	0	7	3	1	118	13	18	
3	23	7	4	1	179	51	18	
7	7	7	5	1	279	5	14	
11	23	7	5	1	289	36	13	
3	23	7	6	1	179	51	18	
10	22	7	6	1	*null*	52	3	
20	23	7	6	1	260	50	11	



# Absenteeism!at!work	# Absenteeism!at!work	# Absenteeism!at!work	# Absenteeism!at!work	# Absenteeis...	# Absenteeism!at!work	# Absenteeism!at!work	# Absenteeis...	# Absenteeism!at...
Work load Averag...	Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet	Weight
239,554	97	0	1	2	1	0	1	90
239,554	97	1	1	1	1	0	0	98
239,554	97	0	1	0	1	0	0	89
239,554	97	0	1	2	1	1	0	68
239,554	97	0	1	2	1	0	1	90
239,554	97	0	1	0	1	0	0	89
239,554	97	0	1	1	1	0	4	80
239,554	97	0	1	4	1	0	0	65



# Absenteeism!at... **Height**	# Absenteeism!at!work **Body mass index**	# Absenteeism!at!work **Absenteeism time ...**
172	30	4
178	31	0
170	31	2
168	24	4
172	30	2
170	31	*null*
172	27	8
168	23	4

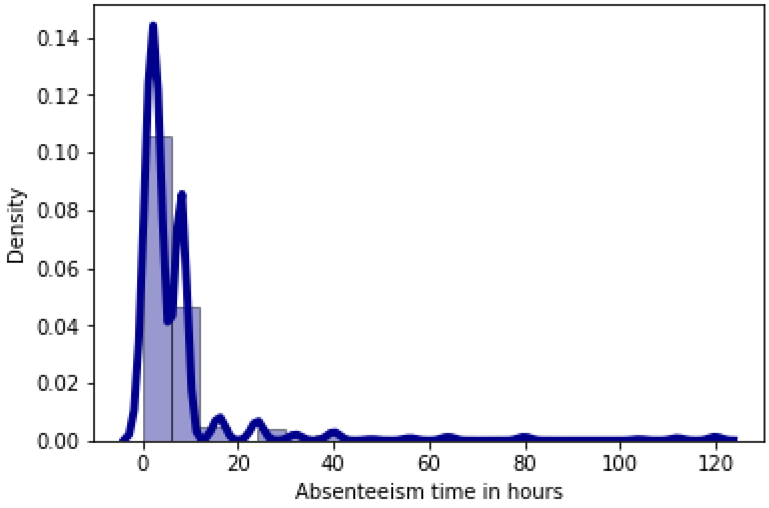
Chapter 2

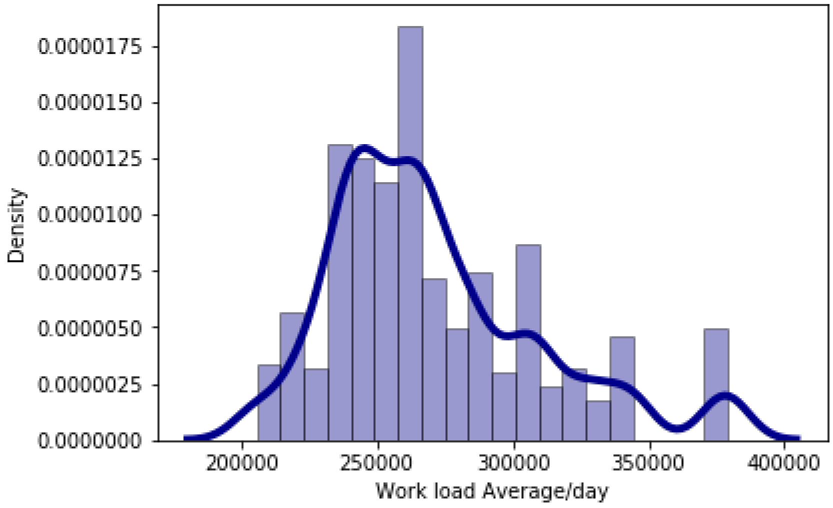
Methodology

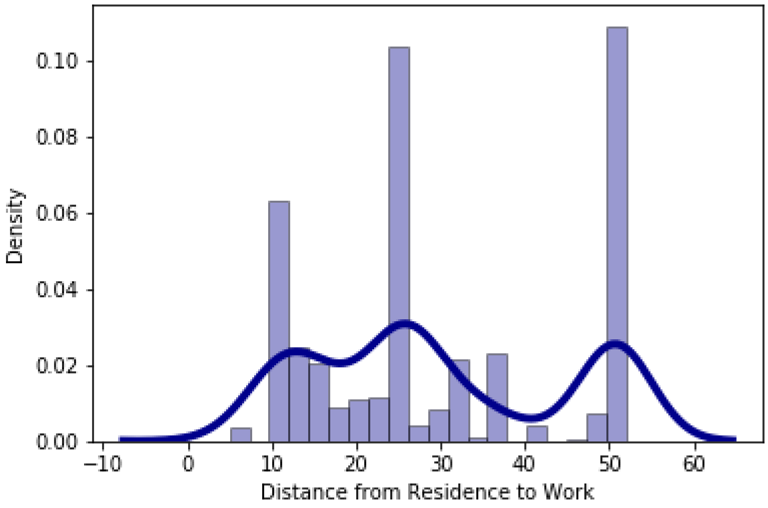
2.1 Pre-Processing

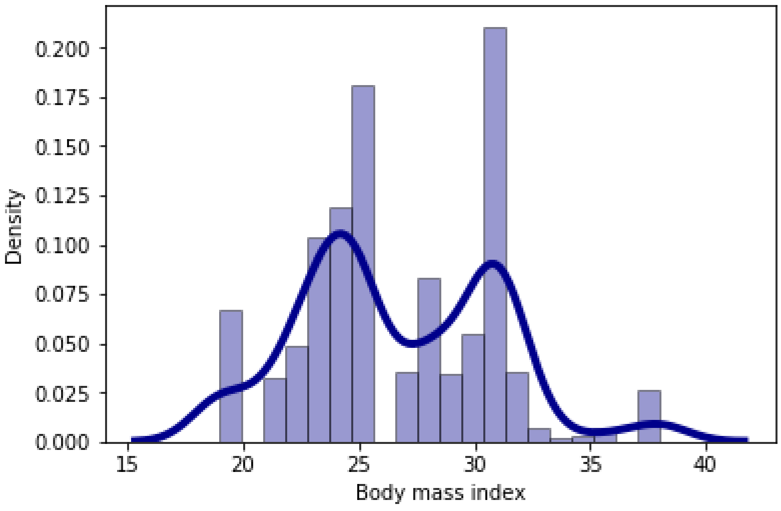
Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms *looking at data* refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process, we will first try and look at all the probability distributions of the variables. We can visualize that in a glance by looking at the probability distributions or probability density functions of the variable.

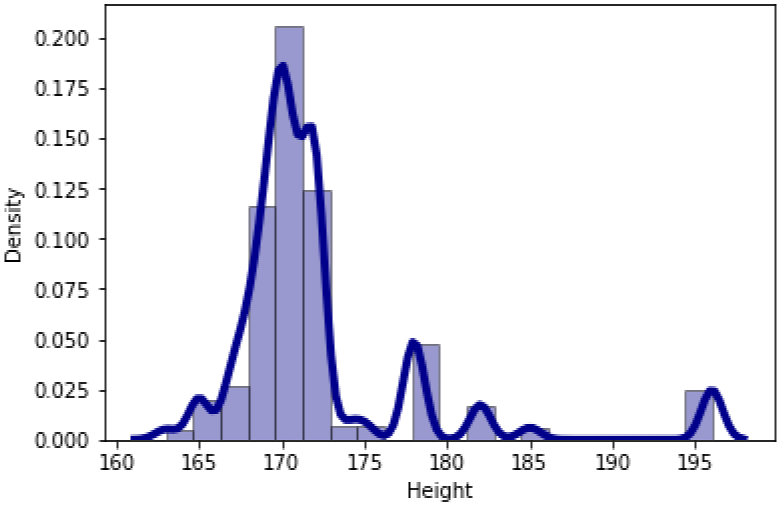
We have plotted the probability density functions of all the independent continuous variables in the data as well as the dependent Absenteeism variable. The blue lines indicate Kernel Density Estimations (KDE) of the variable. We can see all variables have skewed distribution.



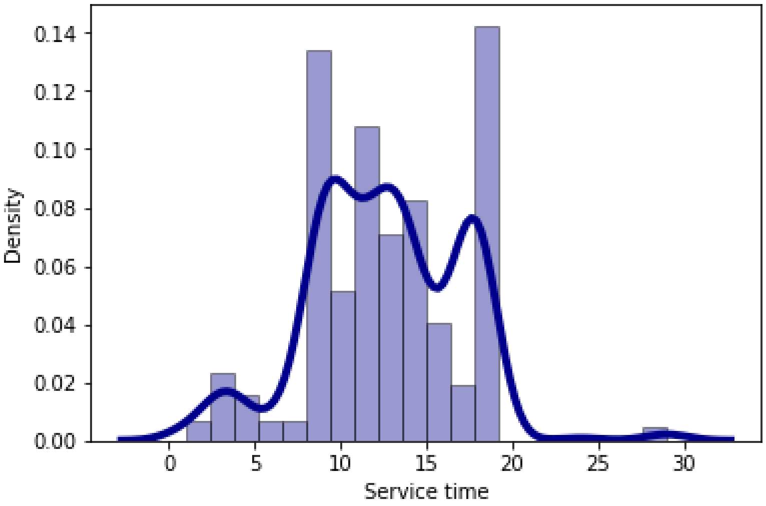

0.0000175
0.0000150
0.0000125
0.0000100
0.0000075
0.0000050
0.0000025
0.0000000
Density
200000
250000
300000
350000
400000
Work load Average/day


0.10
0.08
0.06
0.04
0.02
0.00
Density
−10
0
10
20
30
40
50
60
Distance from Residence to Work

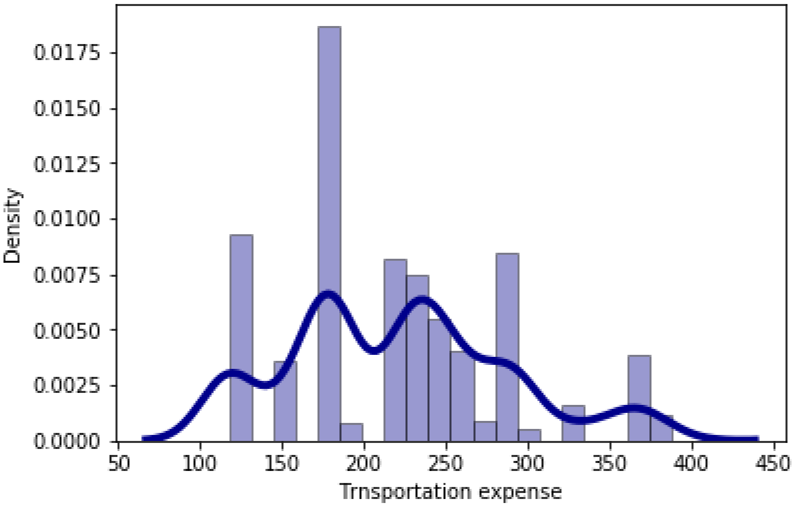

0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Density
15
20
25
30
35
40
Body mass index



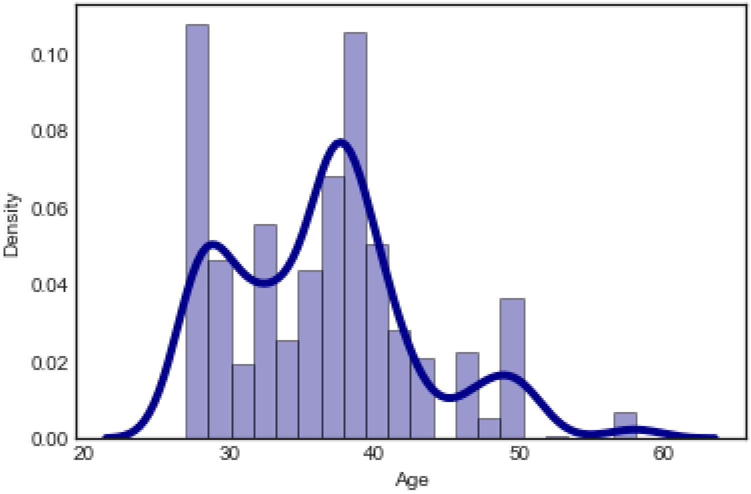
Density
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
160
165
170
175
180
185
190
195
Height

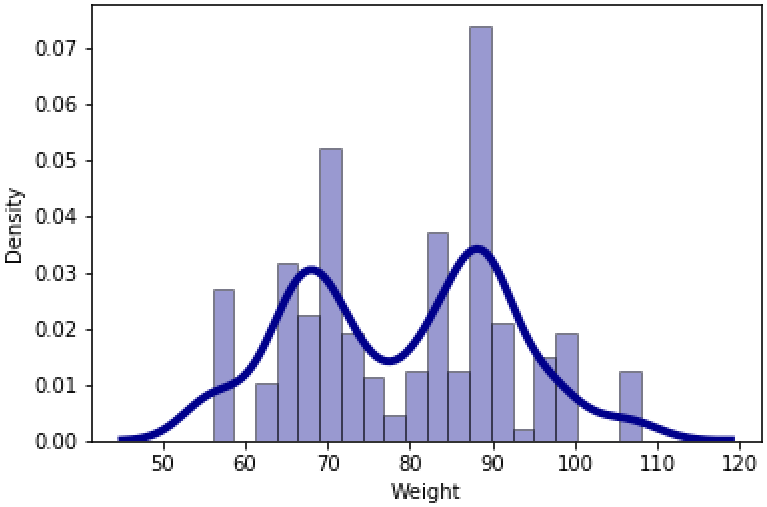


Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
25
30
Service time



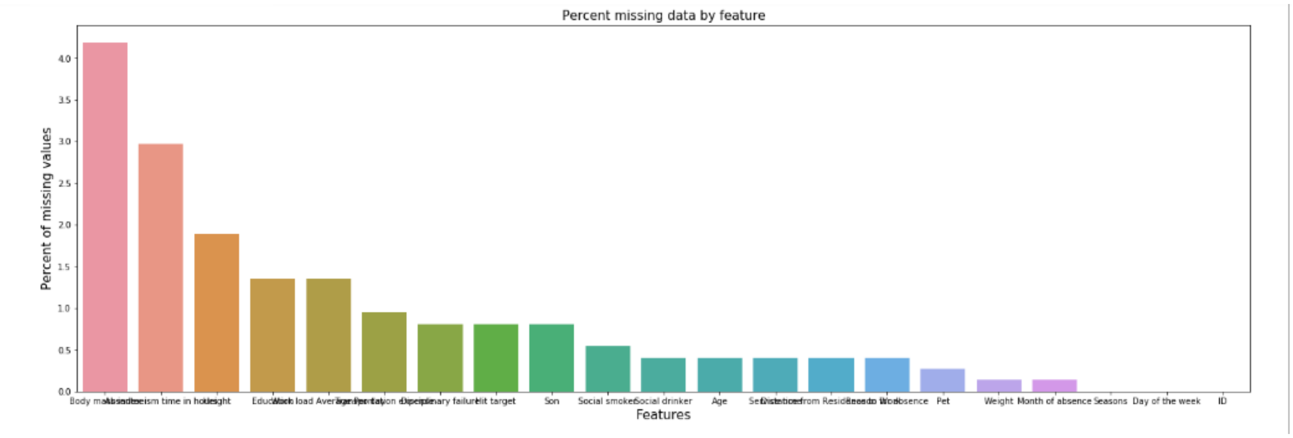
Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
50
100
150
200
250
300
350
400
450
Trnsportation expense


0.10
0.08
0.06
0.04
0.02
0.00
Density
20
30
40
50
60
Age


0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Density
50
60
70
80
90
100
110
120
Weight

Missing value Analysis:

As the data also contained missing values, missing values were imputed using KNN. Since the variables have skewed distributions, imputation with mean was ignored. We could have gone with median or KNN values. However, the imputed values were closer to KNN values.

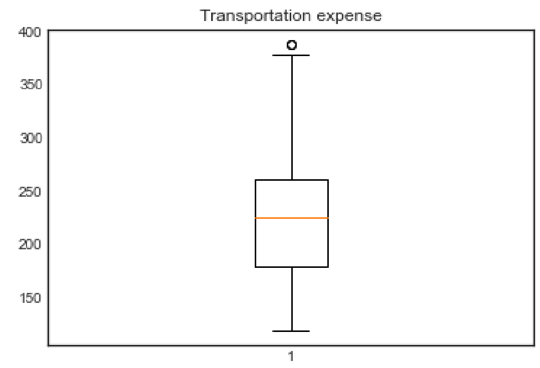


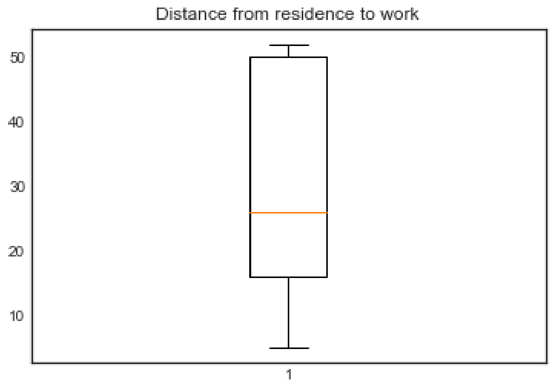
FEATURES	MISSING COUNT
Body mass index	31
Absenteeism time in hours	22
Height	14
Education	10
Work load Average Per day	10
Transportation expense	7
Disciplinary failure	6
Hit target	6
Son	6
Social smoker	4
Social drinker	3
Age	3
Service time	3
Distance from Residence to Work	3
Reason for absence	3
Pet	2
Weight	1
Month of absence	1
Seasons	0
Day of the week	0
ID	0

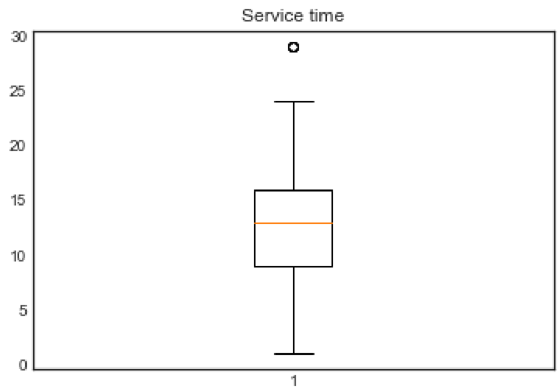
2.1.1 Outlier Analysis

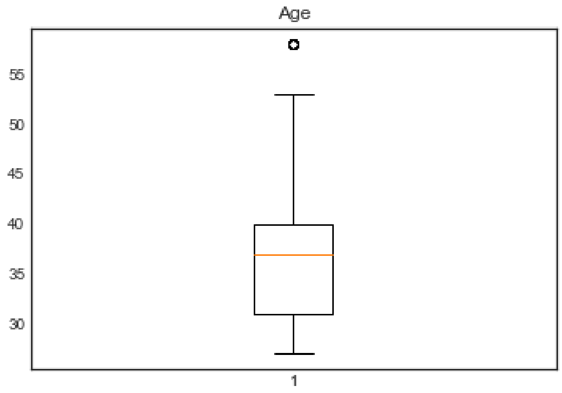
As seen above in pdf, most of the variables are skewed. The skew in these distributions can be most likely explained by the presence of outliers and extreme values in the data.
One of the other steps of **pre-processing** apart from checking for normality is the presence of outliers. In this case we replaced outliers with NaN and then imputed them with KNN algorithm. We visualize the outliers using *boxplots*.

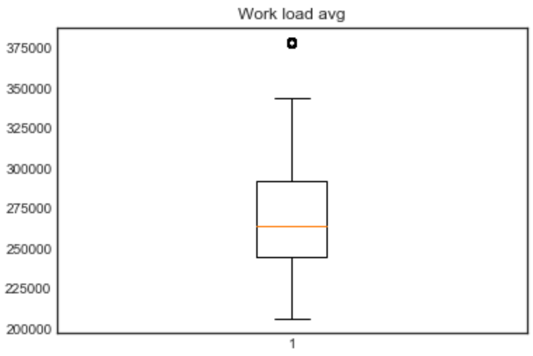
'Distance from work to home', 'weight' and 'Body mass index' have no outliers.

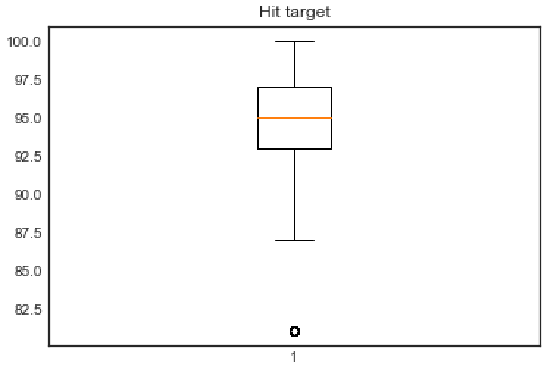


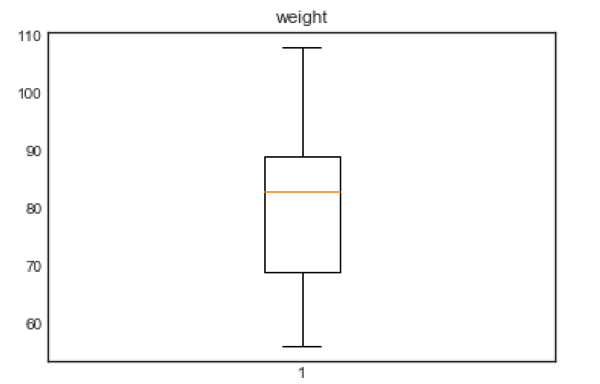


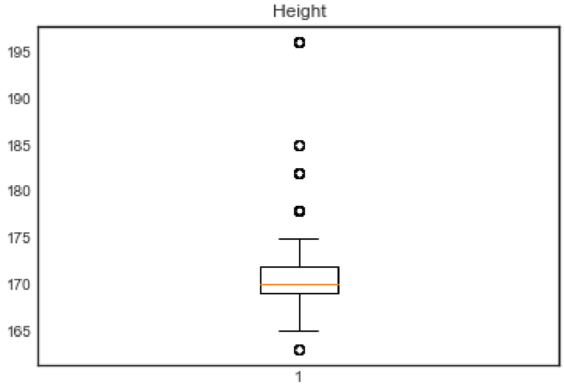


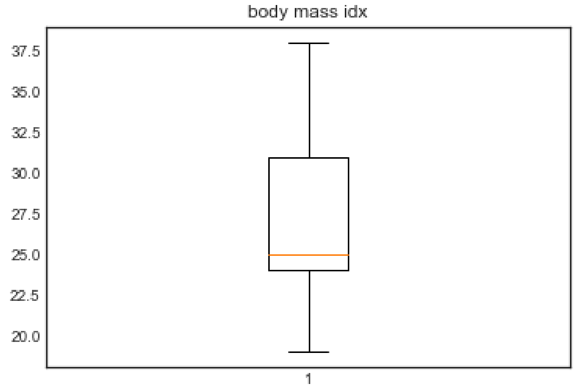


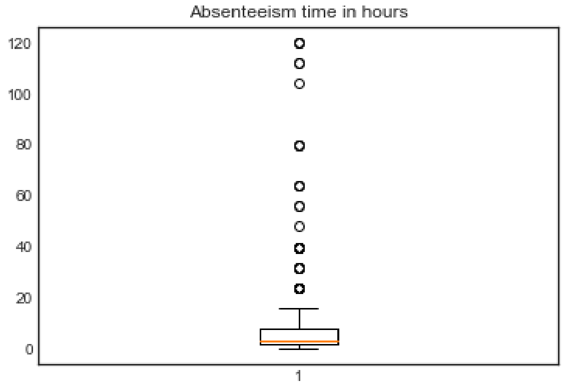










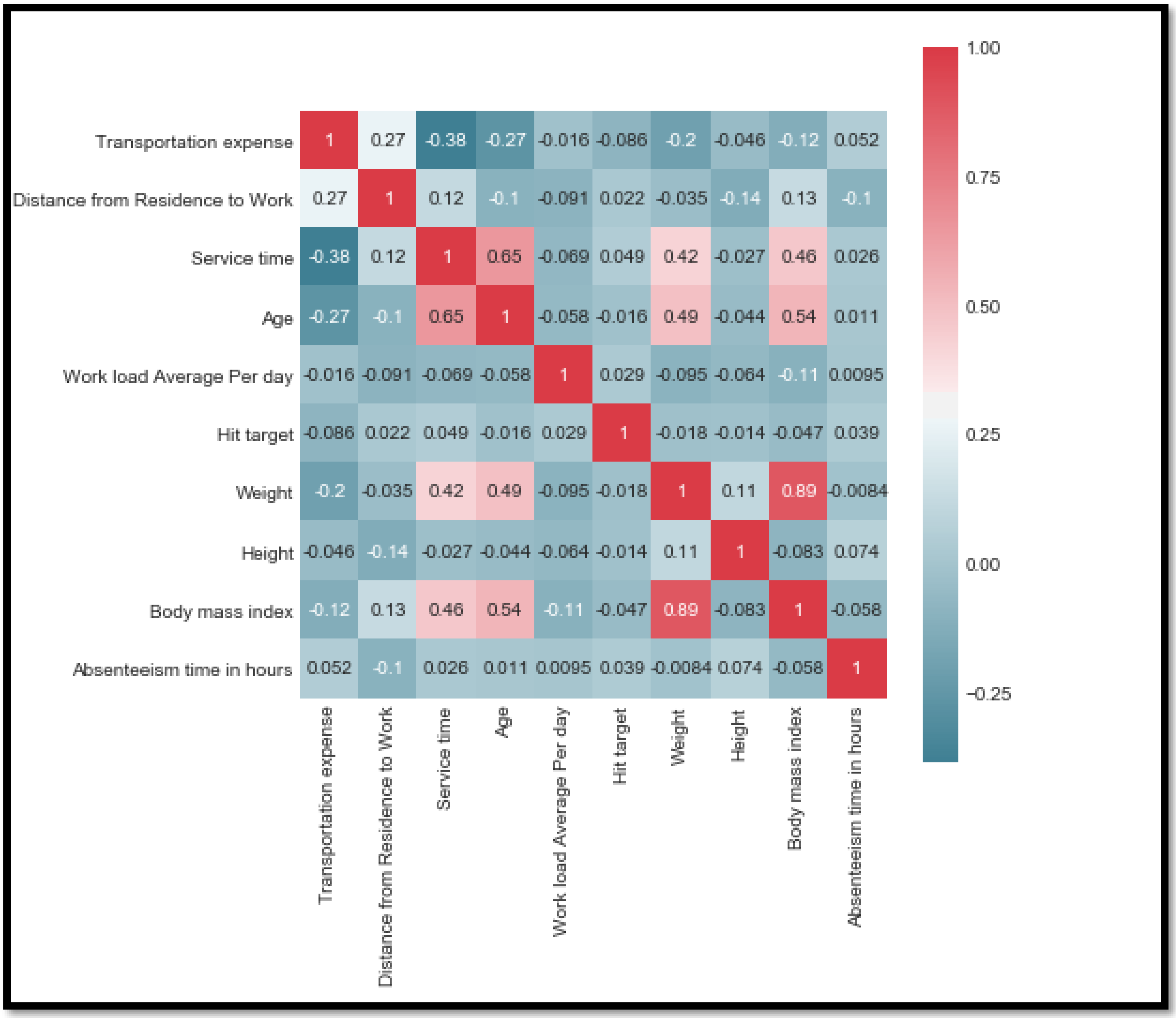


2.1.2 Feature Selection

Before performing any type of modeling, we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. There are several methods of doing that.

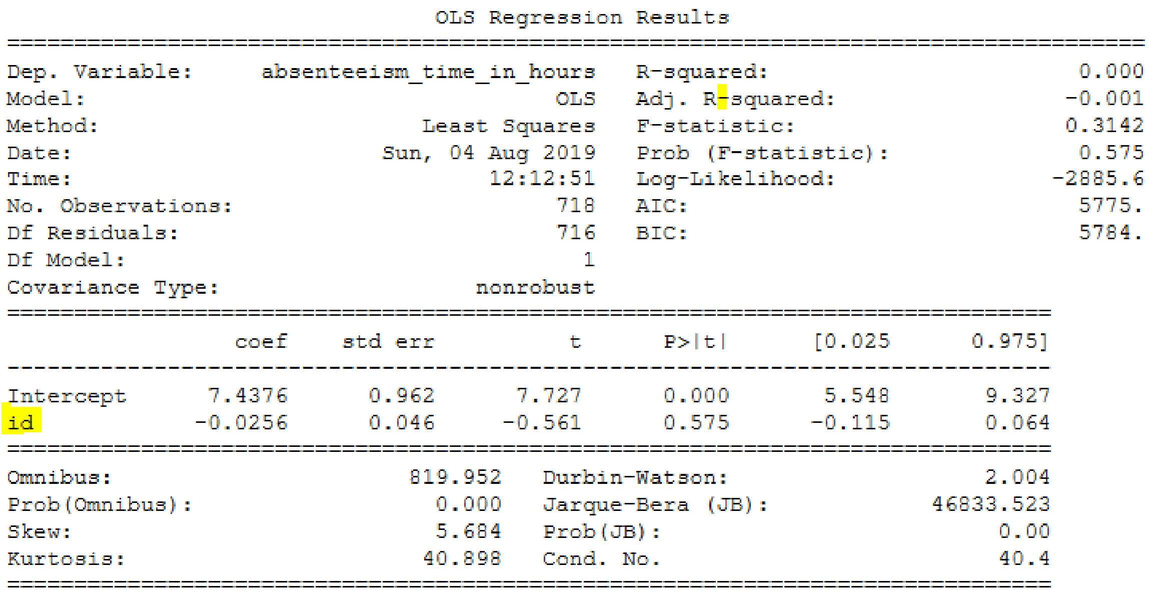
A very simple way of looking at correlations in the data is shown below through the correlation matrix:

Clearly, Only Weight and Body Mass Index have high correlation (>0.8). So, we can drop Body Mass Index from the feature selection.

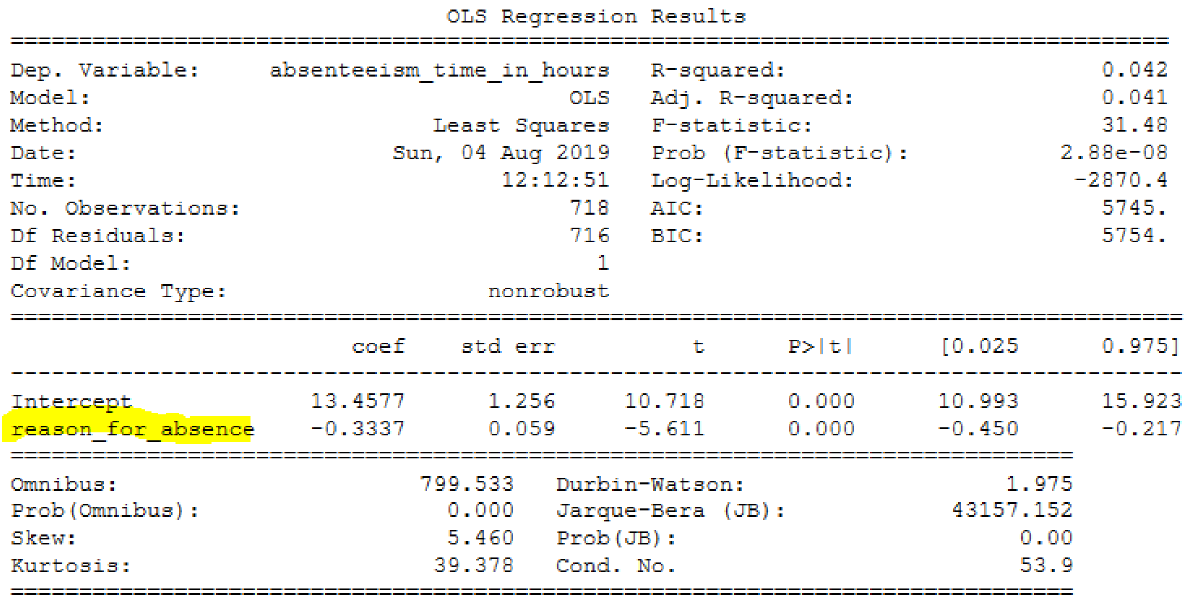


	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average Per day	Hit target	Weight	Height	Body mass index	Absenteeism time in hours
Transportation expense	1	0.27	-0.38	-0.27	-0.016	-0.086	-0.2	-0.046	-0.12	0.052
Distance from Residence to Work	0.27	1	0.12	-0.1	-0.091	0.022	-0.035	-0.14	0.13	-0.1
Service time	-0.38	0.12	1	0.65	-0.069	0.049	0.42	-0.027	0.46	0.026
Age	-0.27	-0.1	0.65	1	-0.058	-0.016	0.49	-0.044	0.54	0.011
Work load Average Per day	-0.016	-0.091	-0.069	-0.058	1	0.029	-0.095	-0.064	-0.11	0.0095
Hit target	-0.086	0.022	0.049	-0.016	0.029	1	-0.018	-0.014	-0.047	0.039
Weight	-0.2	-0.035	0.42	0.49	-0.095	-0.018	1	0.11	0.89	-0.0084
Height	-0.046	-0.14	-0.027	-0.044	-0.064	-0.014	0.11	1	-0.083	0.074
Body mass index	-0.12	0.13	0.46	0.54	-0.11	-0.047	0.89	-0.083	1	-0.058
Absenteeism time in hours	0.052	-0.1	0.026	0.011	0.0095	0.039	-0.0084	0.074	-0.058	1

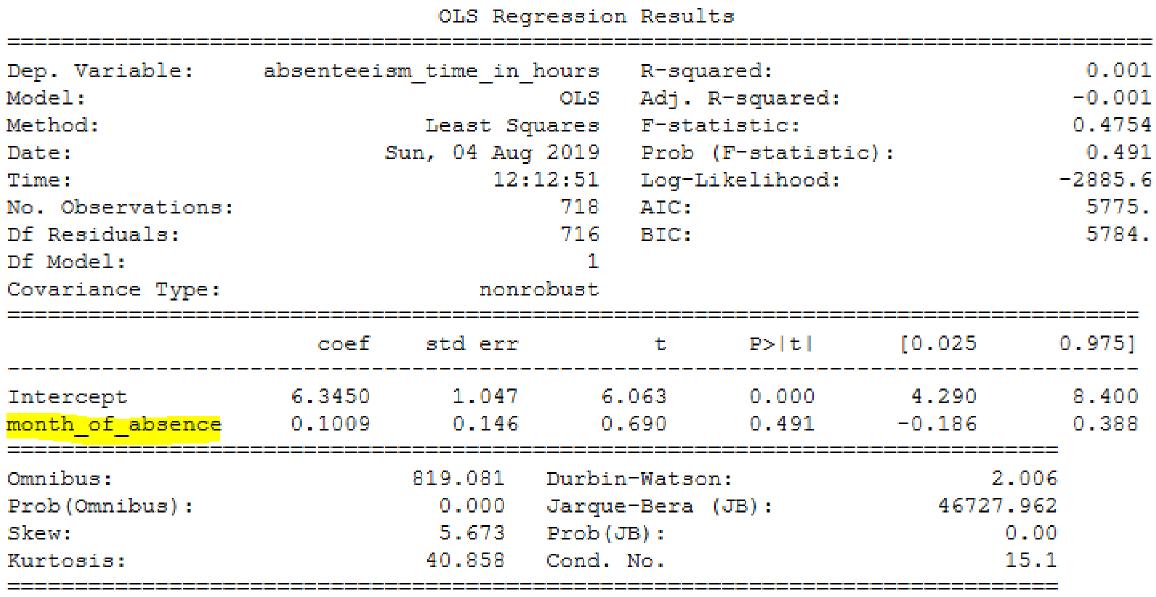
We have used ANOVA test to select categorical variables for model development.



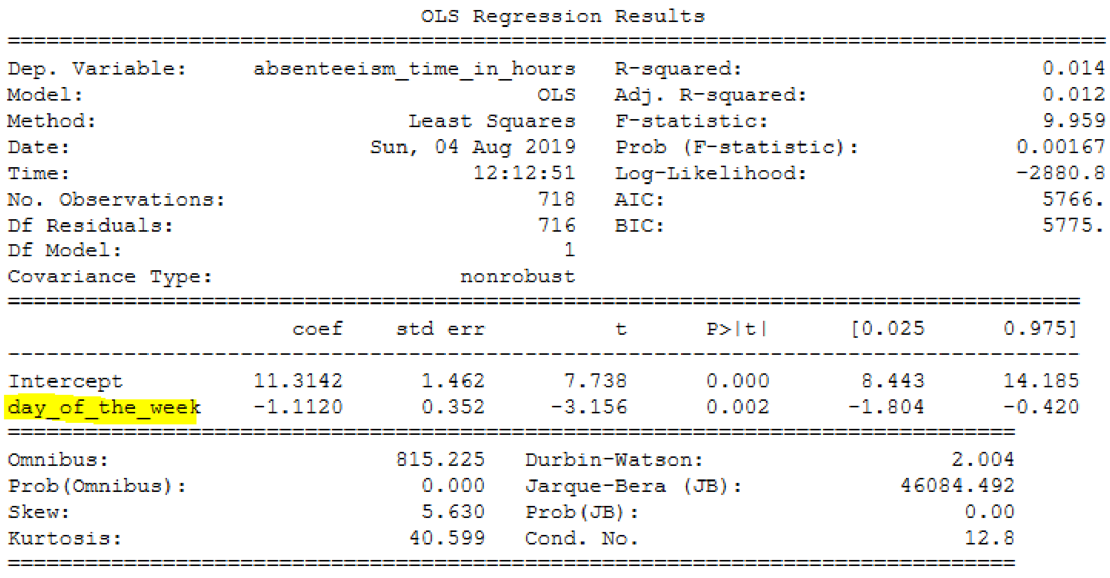
```
                            OLS Regression Results
=====================================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                       0.000
Model:                                   OLS   Adj. R-squared:                 -0.001
Method:                        Least Squares   F-statistic:                    0.3142
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):              0.575
Time:                               12:12:51   Log-Likelihood:                -2885.6
No. Observations:                        718   AIC:                             5775.
Df Residuals:                            716   BIC:                             5784.
Df Model:                                  1
Covariance Type:                   nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.4376      0.962      7.727      0.000       5.548       9.327
id            -0.0256      0.046     -0.561      0.575      -0.115       0.064
==============================================================================
Omnibus:                      819.952   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46833.523
Skew:                           5.684   Prob(JB):                         0.00
Kurtosis:                      40.898   Cond. No.                         40.4
==============================================================================
```



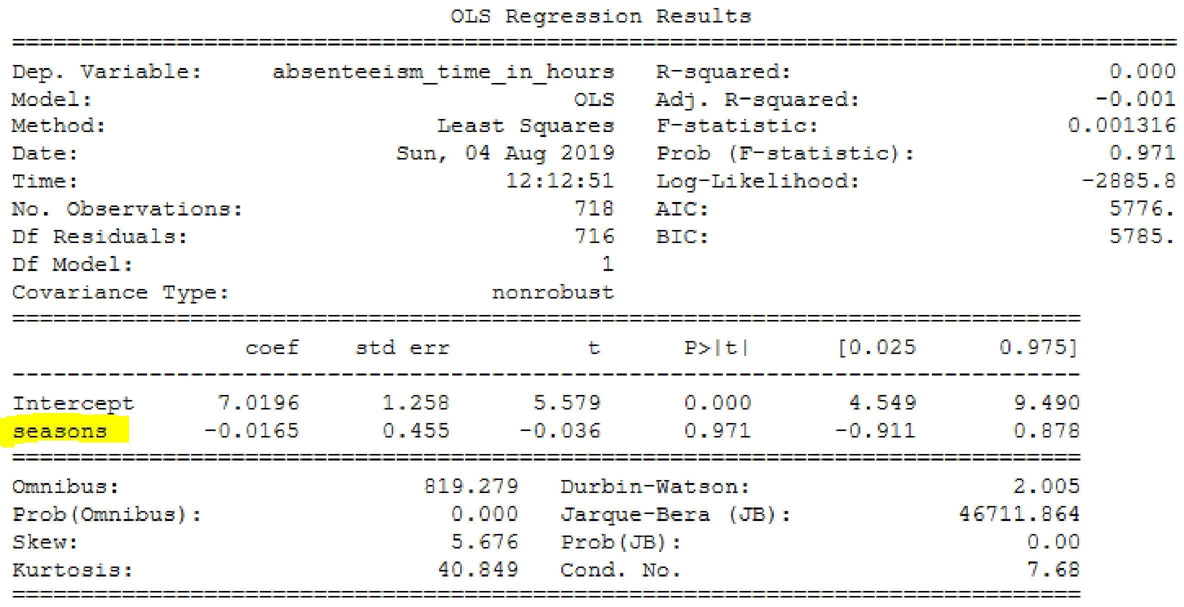
```
                            OLS Regression Results
=====================================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                       0.042
Model:                                   OLS   Adj. R-squared:                  0.041
Method:                        Least Squares   F-statistic:                     31.48
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):           2.88e-08
Time:                               12:12:51   Log-Likelihood:                -2870.4
No. Observations:                        718   AIC:                             5745.
Df Residuals:                            716   BIC:                             5754.
Df Model:                                  1
Covariance Type:                   nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept             13.4577      1.256     10.718      0.000      10.993      15.923
reason_for_absence    -0.3337      0.059     -5.611      0.000      -0.450      -0.217
==============================================================================
Omnibus:                      799.533   Durbin-Watson:                   1.975
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            43157.152
Skew:                           5.460   Prob(JB):                         0.00
Kurtosis:                      39.378   Cond. No.                         53.9
==============================================================================
```



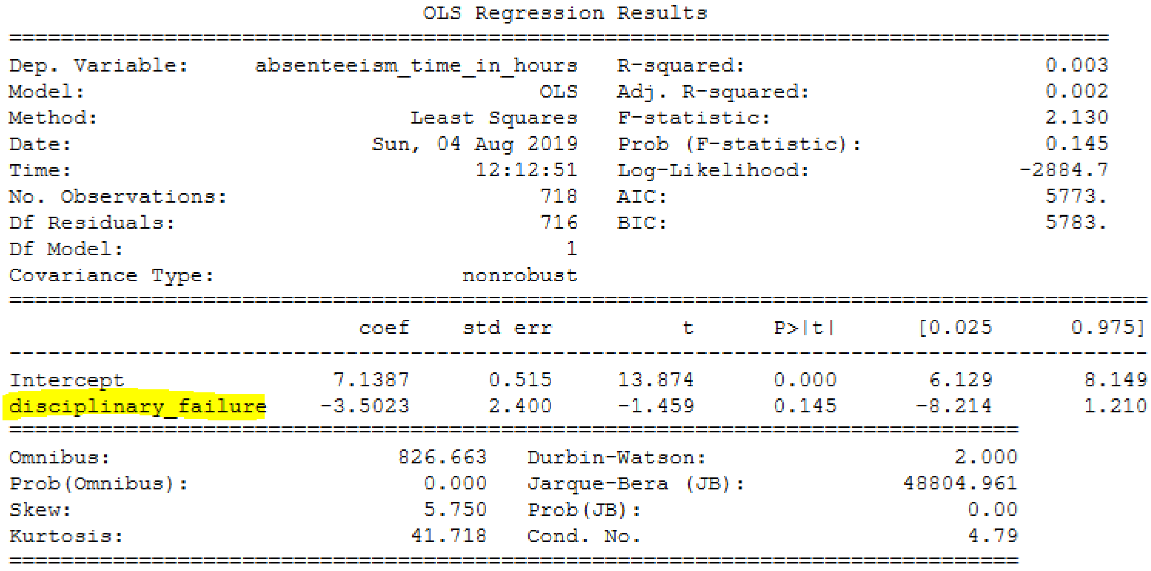
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                 0.001
Model:                               OLS   Adj. R-squared:               -0.001
Method:                    Least Squares   F-statistic:                  0.4754
Date:                   Sun, 04 Aug 2019   Prob (F-statistic):            0.491
Time:                           12:12:51   Log-Likelihood:              -2885.6
No. Observations:                    718   AIC:                           5775.
Df Residuals:                        716   BIC:                           5784.
Df Model:                              1
Covariance Type:               nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept             6.3450      1.047      6.063      0.000       4.290       8.400
month_of_absence      0.1009      0.146      0.690      0.491      -0.186       0.388
==============================================================================
Omnibus:                      819.081   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46727.962
Skew:                           5.673   Prob(JB):                         0.00
Kurtosis:                      40.858   Cond. No.                         15.1
==============================================================================
```



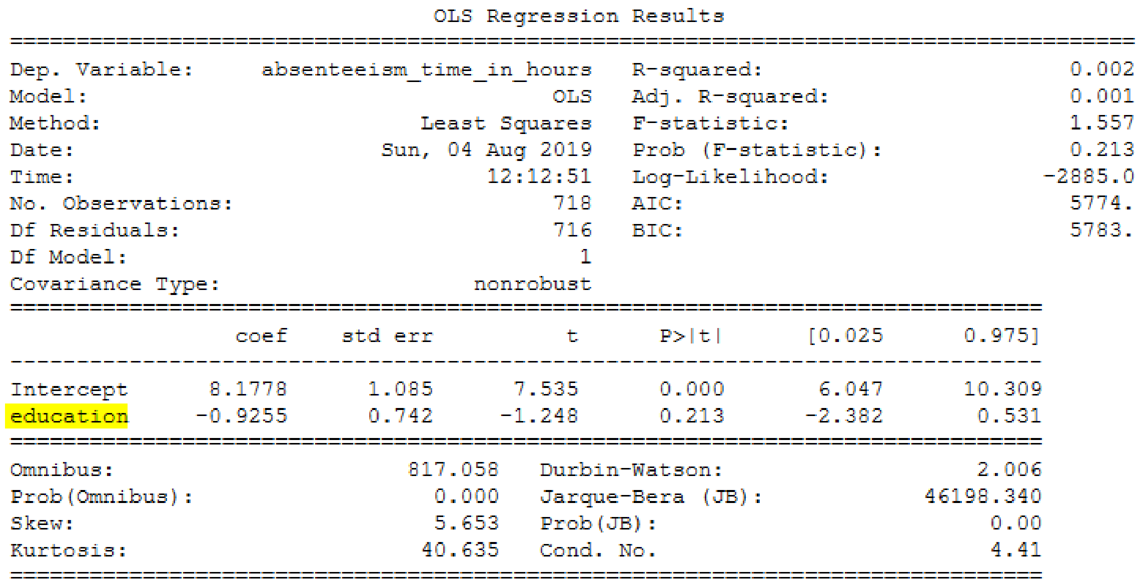
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                 0.014
Model:                               OLS   Adj. R-squared:                0.012
Method:                    Least Squares   F-statistic:                   9.959
Date:                   Sun, 04 Aug 2019   Prob (F-statistic):          0.00167
Time:                           12:12:51   Log-Likelihood:              -2880.8
No. Observations:                    718   AIC:                           5766.
Df Residuals:                        716   BIC:                           5775.
Df Model:                              1
Covariance Type:               nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept          11.3142      1.462      7.738      0.000       8.443      14.185
day_of_the_week    -1.1120      0.352     -3.156      0.002      -1.804      -0.420
==============================================================================
Omnibus:                      815.225   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46084.492
Skew:                           5.630   Prob(JB):                         0.00
Kurtosis:                      40.599   Cond. No.                         12.8
==============================================================================
```



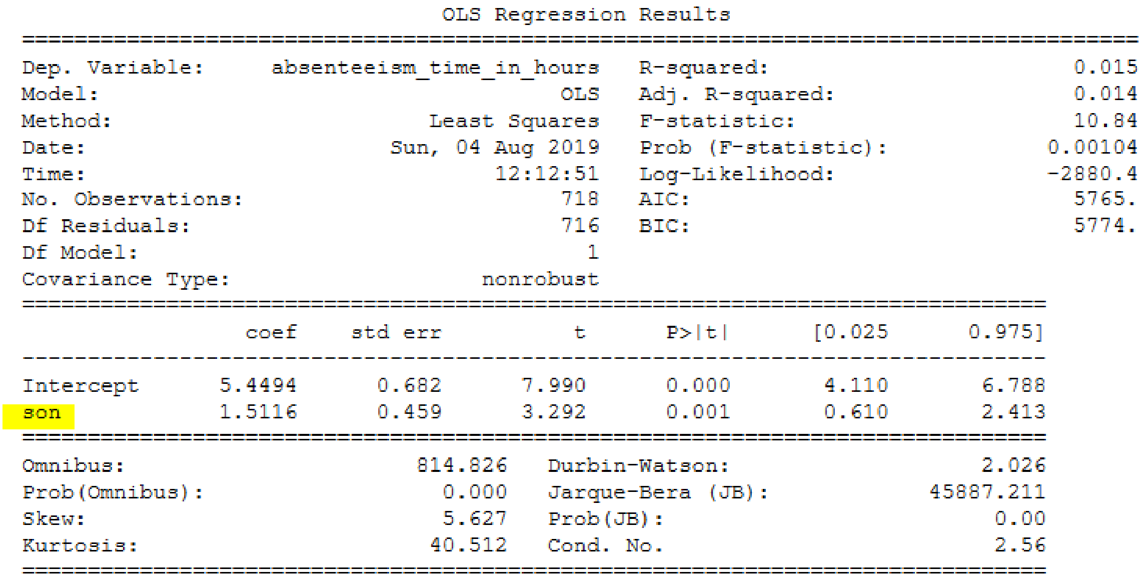
```
                                OLS Regression Results
=====================================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                       0.000
Model:                                   OLS   Adj. R-squared:                 -0.001
Method:                        Least Squares   F-statistic:                  0.001316
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):              0.971
Time:                               12:12:51   Log-Likelihood:                -2885.8
No. Observations:                        718   AIC:                             5776.
Df Residuals:                            716   BIC:                             5785.
Df Model:                                  1
Covariance Type:                   nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.0196      1.258      5.579      0.000       4.549       9.490
seasons       -0.0165      0.455     -0.036      0.971      -0.911       0.878
==============================================================================
Omnibus:                      819.279   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46711.864
Skew:                           5.676   Prob(JB):                         0.00
Kurtosis:                      40.849   Cond. No.                         7.68
==============================================================================
```



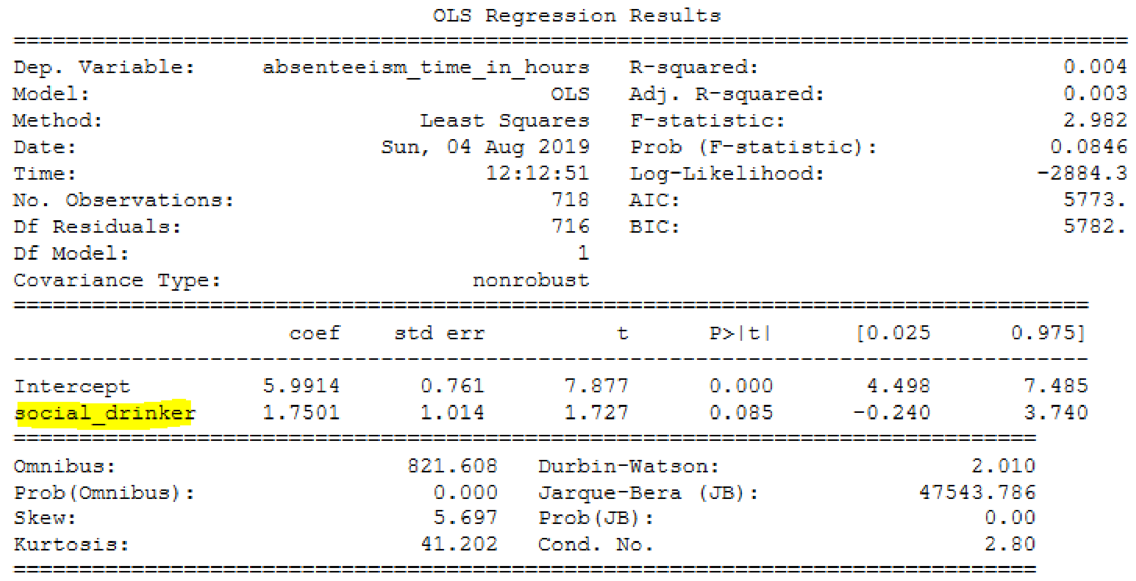
```
                                OLS Regression Results
=====================================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                       0.003
Model:                                   OLS   Adj. R-squared:                  0.002
Method:                        Least Squares   F-statistic:                     2.130
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):              0.145
Time:                               12:12:51   Log-Likelihood:                -2884.7
No. Observations:                        718   AIC:                             5773.
Df Residuals:                            716   BIC:                             5783.
Df Model:                                  1
Covariance Type:                   nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept                7.1387      0.515     13.874      0.000       6.129       8.149
disciplinary_failure    -3.5023      2.400     -1.459      0.145      -8.214       1.210
==============================================================================
Omnibus:                      826.663   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            48804.961
Skew:                           5.750   Prob(JB):                         0.00
Kurtosis:                      41.718   Cond. No.                         4.79
==============================================================================
```



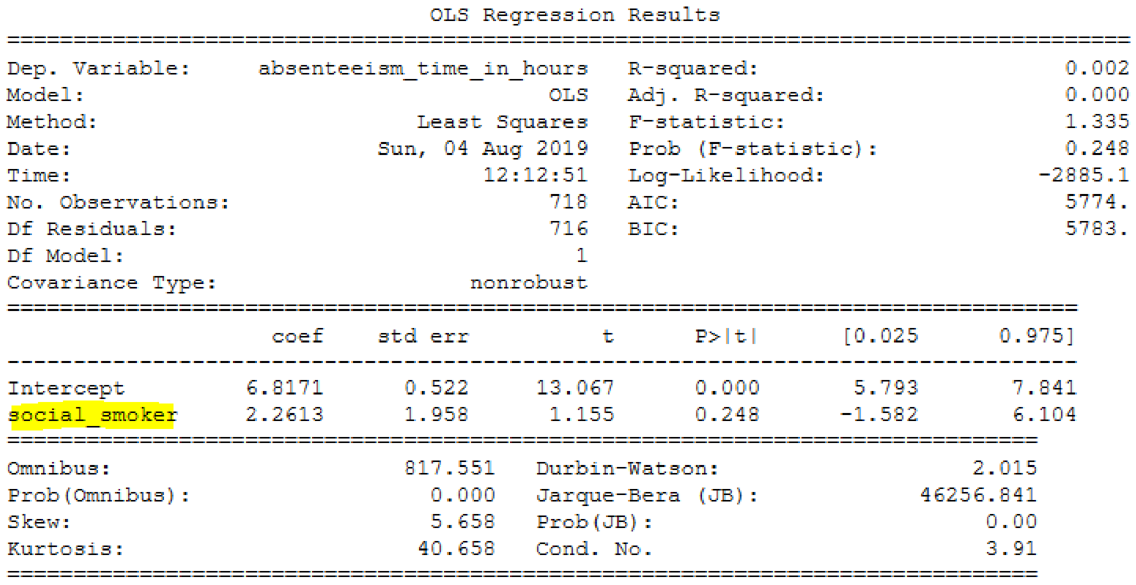
```
                              OLS Regression Results                              
==================================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                    0.002
Model:                                   OLS   Adj. R-squared:               0.001
Method:                        Least Squares   F-statistic:                  1.557
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):           0.213
Time:                               12:12:51   Log-Likelihood:             -2885.0
No. Observations:                        718   AIC:                          5774.
Df Residuals:                            716   BIC:                          5783.
Df Model:                                  1                                      
Covariance Type:                   nonrobust                                      
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.1778      1.085      7.535      0.000       6.047      10.309
education     -0.9255      0.742     -1.248      0.213      -2.382       0.531
==============================================================================
Omnibus:                      817.058   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46198.340
Skew:                           5.653   Prob(JB):                         0.00
Kurtosis:                      40.635   Cond. No.                         4.41
==============================================================================
```



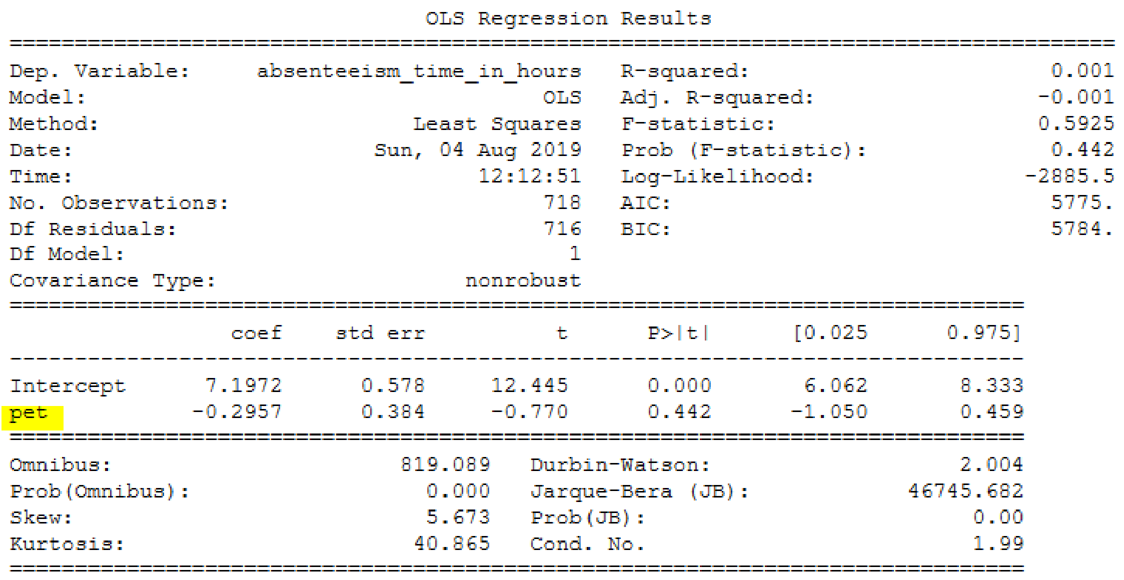
```
                              OLS Regression Results                              
==================================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                    0.015
Model:                                   OLS   Adj. R-squared:               0.014
Method:                        Least Squares   F-statistic:                  10.84
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):         0.00104
Time:                               12:12:51   Log-Likelihood:             -2880.4
No. Observations:                        718   AIC:                          5765.
Df Residuals:                            716   BIC:                          5774.
Df Model:                                  1                                      
Covariance Type:                   nonrobust                                      
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      5.4494      0.682      7.990      0.000       4.110       6.788
son            1.5116      0.459      3.292      0.001       0.610       2.413
==============================================================================
Omnibus:                      814.826   Durbin-Watson:                   2.026
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            45887.211
Skew:                           5.627   Prob(JB):                         0.00
Kurtosis:                      40.512   Cond. No.                         2.56
==============================================================================
```



```
                                OLS Regression Results
=====================================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                       0.004
Model:                                   OLS   Adj. R-squared:                  0.003
Method:                        Least Squares   F-statistic:                     2.982
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):             0.0846
Time:                               12:12:51   Log-Likelihood:                -2884.3
No. Observations:                        718   AIC:                             5773.
Df Residuals:                            716   BIC:                             5782.
Df Model:                                  1
Covariance Type:                   nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          5.9914      0.761      7.877      0.000       4.498       7.485
social_drinker     1.7501      1.014      1.727      0.085      -0.240       3.740
==============================================================================
Omnibus:                      821.608   Durbin-Watson:                   2.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            47543.786
Skew:                           5.697   Prob(JB):                         0.00
Kurtosis:                      41.202   Cond. No.                         2.80
==============================================================================
```



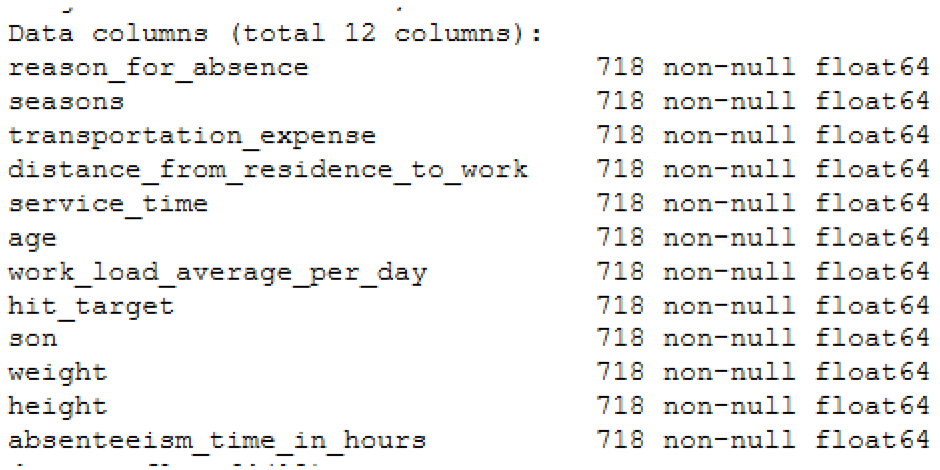
```
                                OLS Regression Results
=====================================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                       0.002
Model:                                   OLS   Adj. R-squared:                  0.000
Method:                        Least Squares   F-statistic:                     1.335
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):              0.248
Time:                               12:12:51   Log-Likelihood:                -2885.1
No. Observations:                        718   AIC:                             5774.
Df Residuals:                            716   BIC:                             5783.
Df Model:                                  1
Covariance Type:                   nonrobust
==================================================================================
                    coef    std err          t      P>|t|       [0.025      0.975]
----------------------------------------------------------------------------------
Intercept         6.8171      0.522     13.067      0.000        5.793       7.841
social_smoker     2.2613      1.958      1.155      0.248       -1.582       6.104
==============================================================================
Omnibus:                      817.551   Durbin-Watson:                   2.015
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46256.841
Skew:                           5.658   Prob(JB):                         0.00
Kurtosis:                      40.658   Cond. No.                         3.91
==============================================================================
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     absenteeism_time_in_hours   R-squared:                       0.001
Model:                                   OLS   Adj. R-squared:                 -0.001
Method:                        Least Squares   F-statistic:                    0.5925
Date:                       Sun, 04 Aug 2019   Prob (F-statistic):              0.442
Time:                               12:12:51   Log-Likelihood:                -2885.5
No. Observations:                        718   AIC:                             5775.
Df Residuals:                            716   BIC:                             5784.
Df Model:                                  1
Covariance Type:                   nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.1972      0.578     12.445      0.000       6.062       8.333
pet           -0.2957      0.384     -0.770      0.442      -1.050       0.459
==============================================================================
Omnibus:                      819.089   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46745.682
Skew:                           5.673   Prob(JB):                         0.00
Kurtosis:                      40.865   Cond. No.                         1.99
==============================================================================
```

For P-value < 0.05, we reject the null hypothesis.

Variables selected for model development:



```
Data columns (total 12 columns):
reason_for_absence              718 non-null float64
seasons                         718 non-null float64
transportation_expense          718 non-null float64
distance_from_residence_to_work 718 non-null float64
service_time                    718 non-null float64
age                             718 non-null float64
work_load_average_per_day       718 non-null float64
hit_target                      718 non-null float64
son                             718 non-null float64
weight                          718 non-null float64
height                          718 non-null float64
absenteeism_time_in_hours       718 non-null float64
```

Feature Scaling

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
Also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula for a min-max of [0, 1] is given as:

$$\boxed{x' = \frac{x - \min(x)}{\max(x) - \min(x)}}$$

where is an original value, is the normalized value.

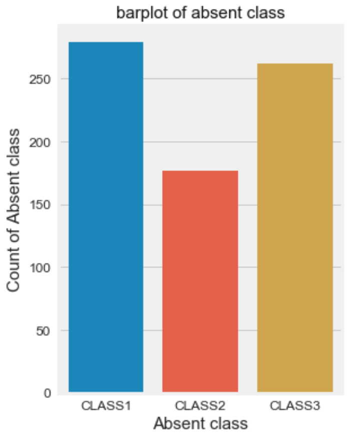
2.2 Modeling
2.2.1 Model Selection

The dependent variable can fall in either of the four categories:

1. Nominal
2. Ordinal
3. Interval
4. Ratio

If the dependent variable is Nominal the only predictive analysis that we can perform is **Classification**, and if the dependent variable is Interval or Ratio the normal method is to do a **Regression** analysis, or **classification after binning.**

As our target variable (Absenteeism time in hours (target)) is Numerical, but we have classified it into sub-ranges. We will use classification.



Evaluation Metrics for classification problems

accuracy_score: Accuracy classification score.

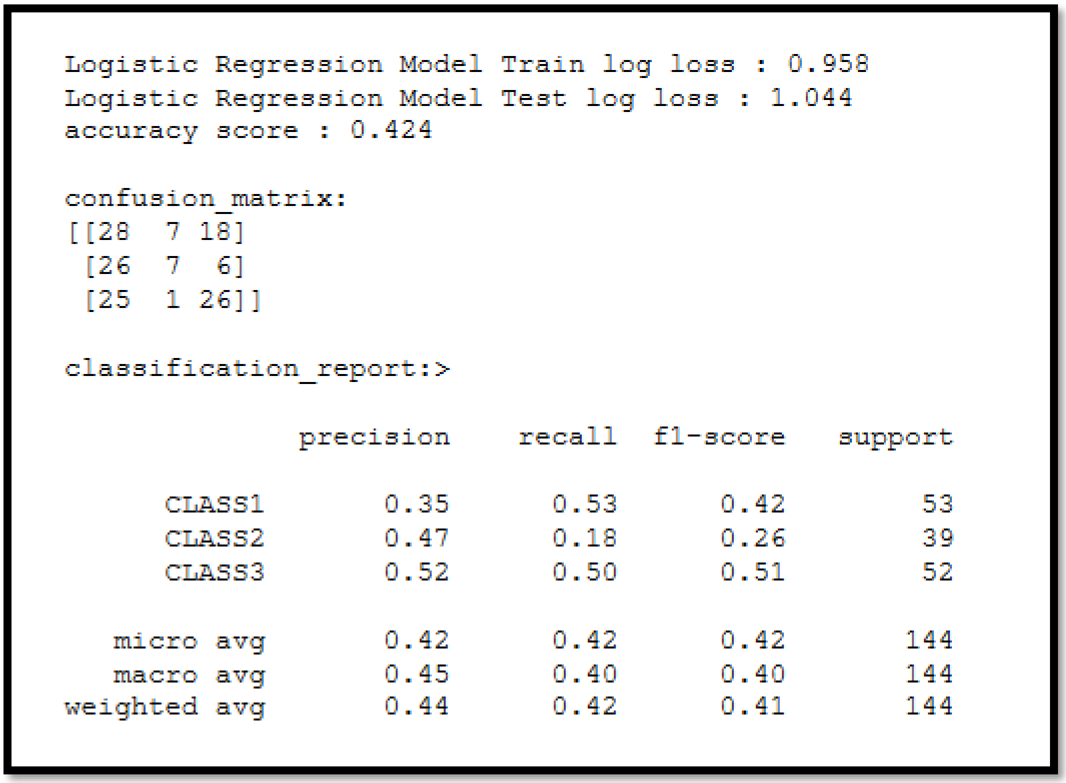
roc_curve: Compute Receiver operating characteristic (ROC)

Log-Loss : Logarithmic loss
It measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value. A perfect model would have a log loss of 0.

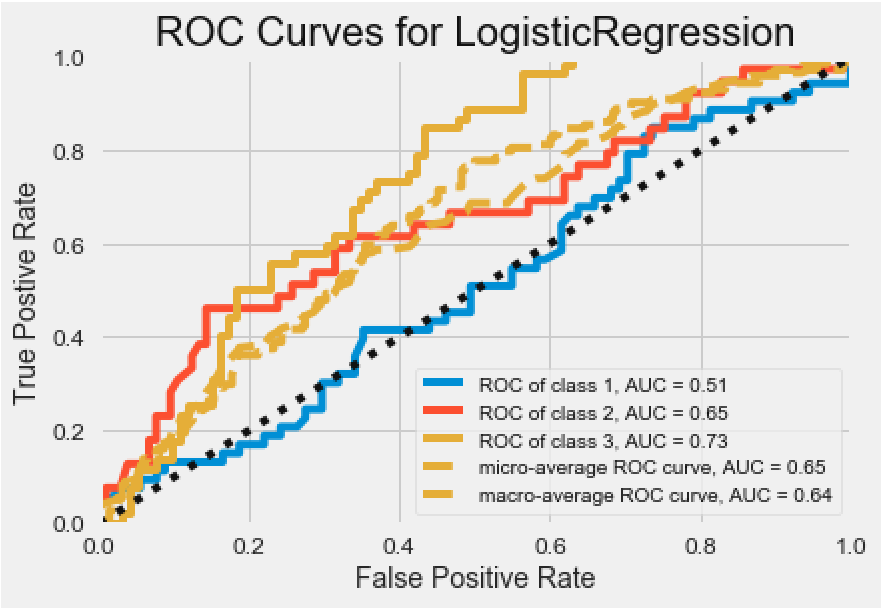
We always start your model building from the simplest to more complex.

Logistic Regression:

- Logistic regression is an estimation of Logit function. Logit function is simply a log of odds in favor of the event.
- This function creates a s-shaped curve with the probability estimate, which is very similar to the required step wise function



```
Logistic Regression Model Train log loss : 0.958
Logistic Regression Model Test log loss : 1.044
accuracy score : 0.424

confusion_matrix:
[[28  7 18]
 [26  7  6]
 [25  1 26]]

classification_report:>

             precision    recall  f1-score   support

      CLASS1       0.35      0.53      0.42        53
      CLASS2       0.47      0.18      0.26        39
      CLASS3       0.52      0.50      0.51        52

   micro avg       0.42      0.42      0.42       144
   macro avg       0.45      0.40      0.40       144
weighted avg       0.44      0.42      0.41       144
```



Result: log_loss is very high, and AUC is low. Also accuracy of the model is very low.Let's explore another model.

Naïve Bayes:

Naïve Bayes is simple yet powerful classification algorithm. It is suitable for both binary and multiclass problems.

The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal weight/importance

contribution to the outcome.

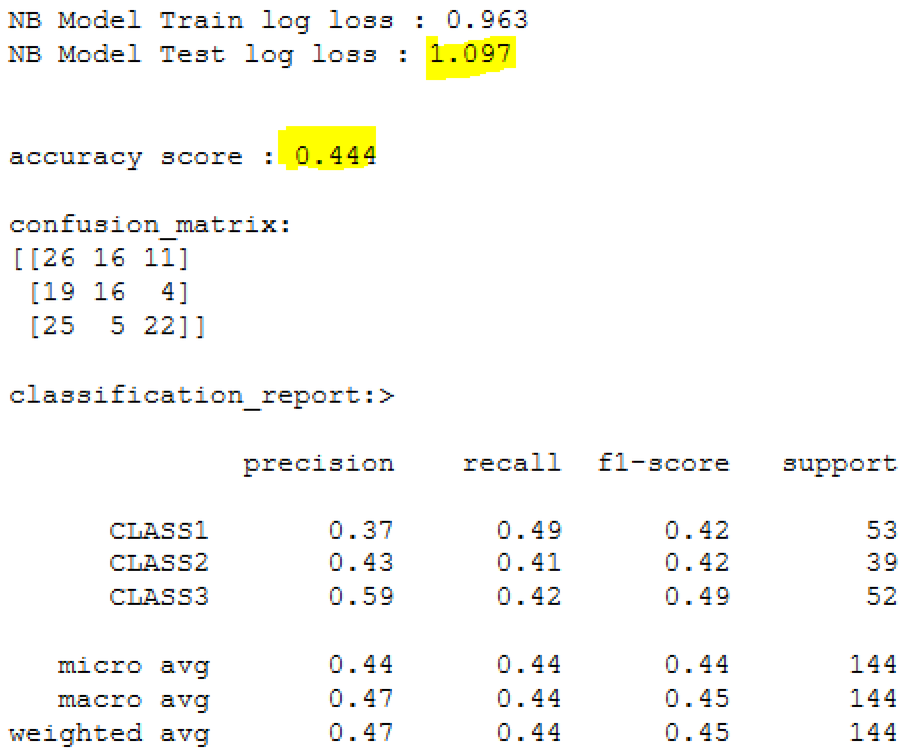
Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

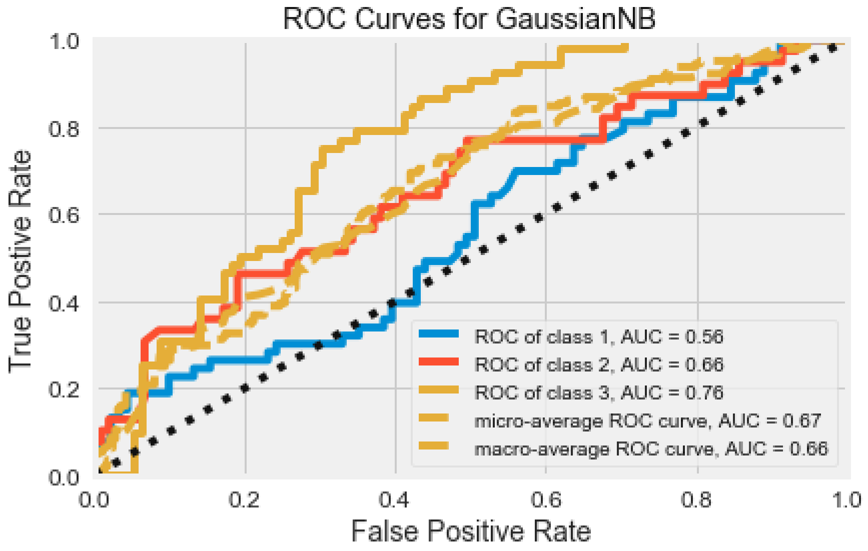
where A and B are events and P(B) is not 0.

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.
- P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).

- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.



```
NB Model Train log loss : 0.963
NB Model Test log loss : 1.097


accuracy score : 0.444

confusion_matrix:
[[26 16 11]
 [19 16  4]
 [25  5 22]]

classification_report:>

              precision    recall  f1-score   support

      CLASS1       0.37      0.49      0.42        53
      CLASS2       0.43      0.41      0.42        39
      CLASS3       0.59      0.42      0.49        52

   micro avg       0.44      0.44      0.44       144
   macro avg       0.47      0.44      0.45       144
weighted avg       0.47      0.44      0.45       144
```



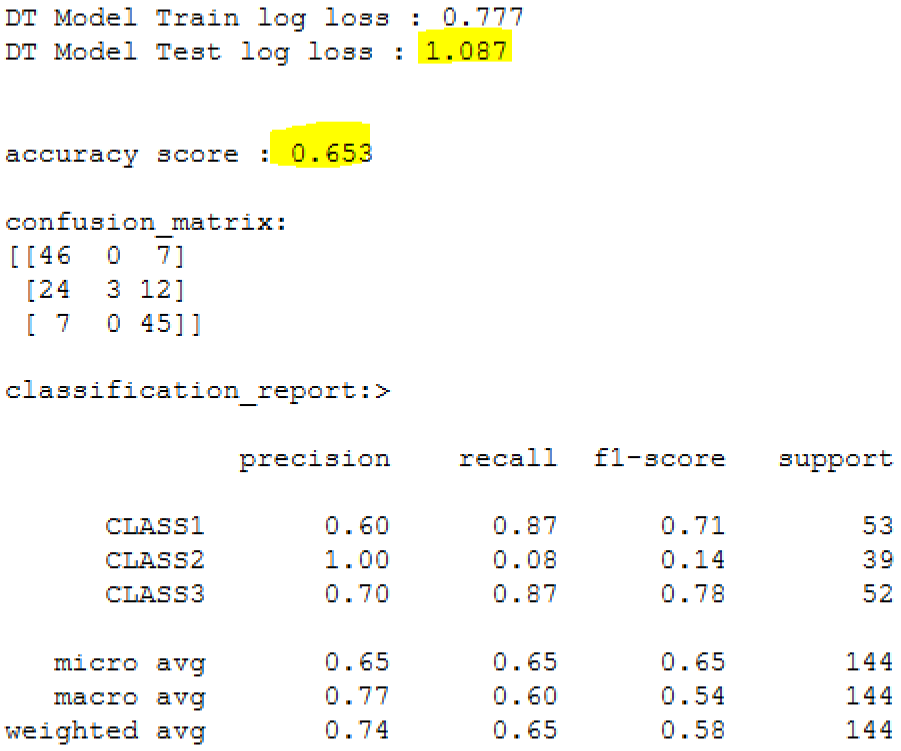
Result: Again, the Logloss is very high, and AUC is low. Also, accuracy of the model is very low.

Decision Tree:

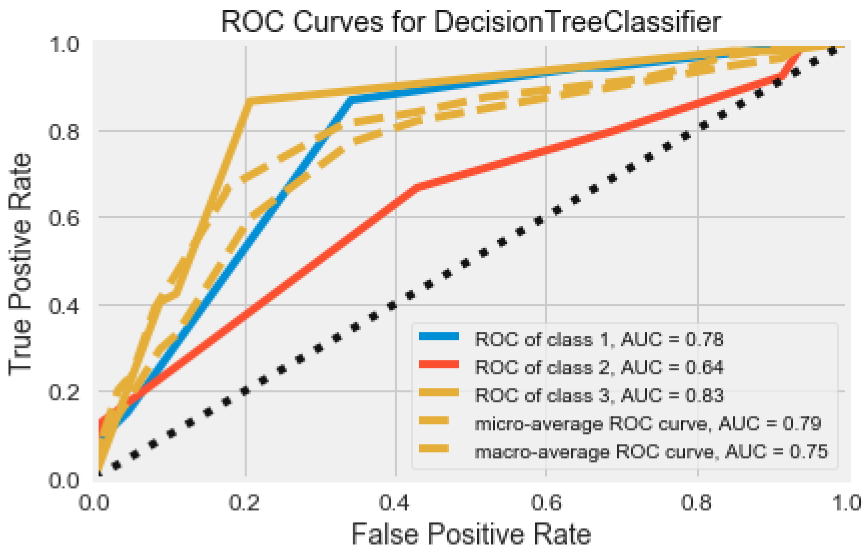
Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning.

Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. Information gain is used to decide which feature to split on at each step in building the tree.

Tree models where the target variable can take a discrete set of values are called **classification trees**.



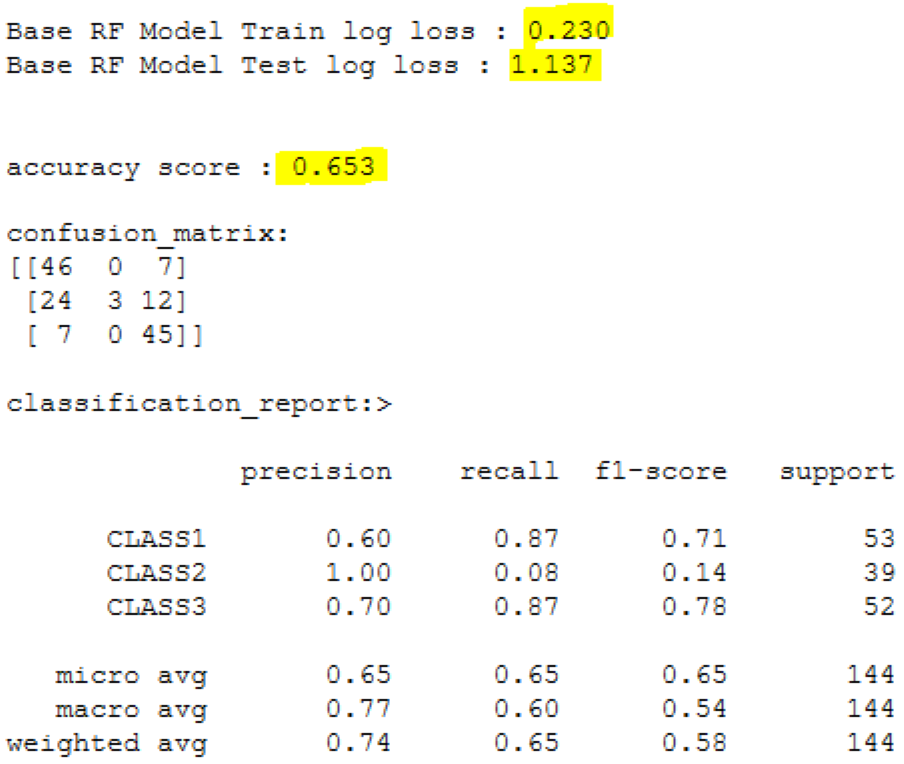
```
DT Model Train log loss : 0.777
DT Model Test log loss : 1.087


accuracy score :  0.653

confusion_matrix:
[[46  0  7]
 [24  3 12]
 [ 7  0 45]]

classification_report:>

              precision    recall  f1-score   support

      CLASS1       0.60      0.87      0.71        53
      CLASS2       1.00      0.08      0.14        39
      CLASS3       0.70      0.87      0.78        52

   micro avg       0.65      0.65      0.65       144
   macro avg       0.77      0.60      0.54       144
weighted avg       0.74      0.65      0.58       144
```



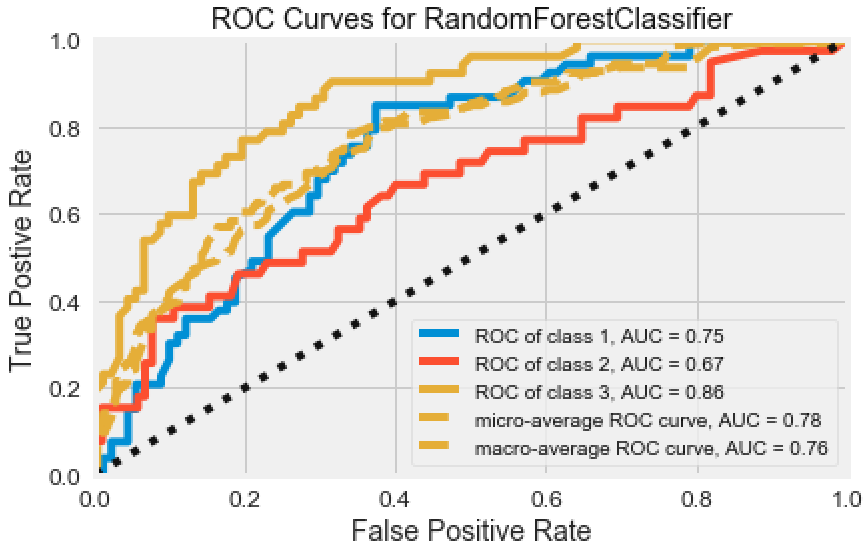
Result: Accuracy score and AUC is higher as compared to naïve bayes but still prediction capability is poor.

Random Forest:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an **ensemble**. It is called random forest as we are building n no. of trees randomly. In other words, to build the decision trees it selects randomly n no of variables and n no of observations to build each decision tree.



```
Base RF Model Train log loss : 0.230
Base RF Model Test log loss : 1.137


accuracy score : 0.653

confusion_matrix:
[[46  0  7]
 [24  3 12]
 [ 7  0 45]]

classification_report:>

              precision    recall  f1-score   support

      CLASS1       0.60      0.87      0.71        53
      CLASS2       1.00      0.08      0.14        39
      CLASS3       0.70      0.87      0.78        52

   micro avg       0.65      0.65      0.65       144
   macro avg       0.77      0.60      0.54       144
weighted avg       0.74      0.65      0.58       144
```



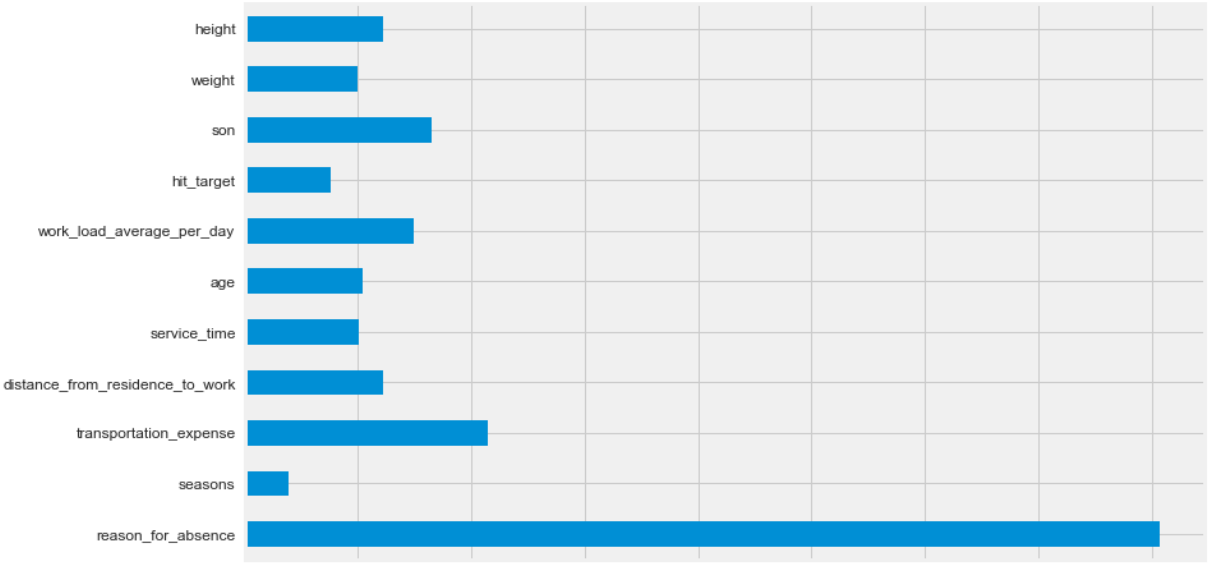


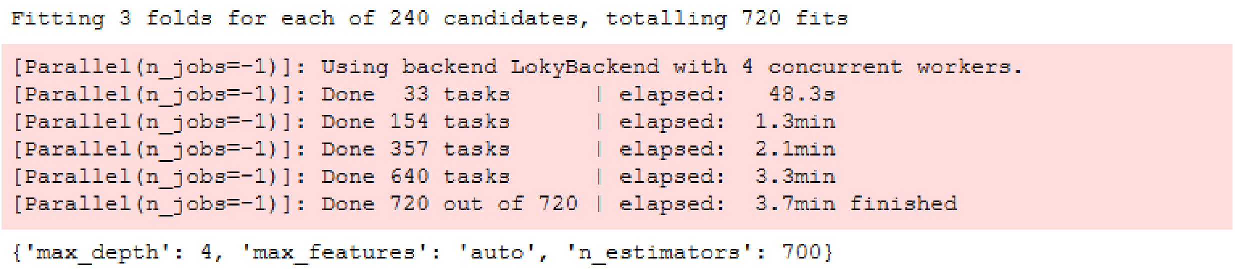
Fig showing feature importance using Random Forest.

Result : Overfitting issue seen here. Lets fine tune the hyperparameter to see if there is any improvement.

Hyperparameter Tuning using GridSearchCV

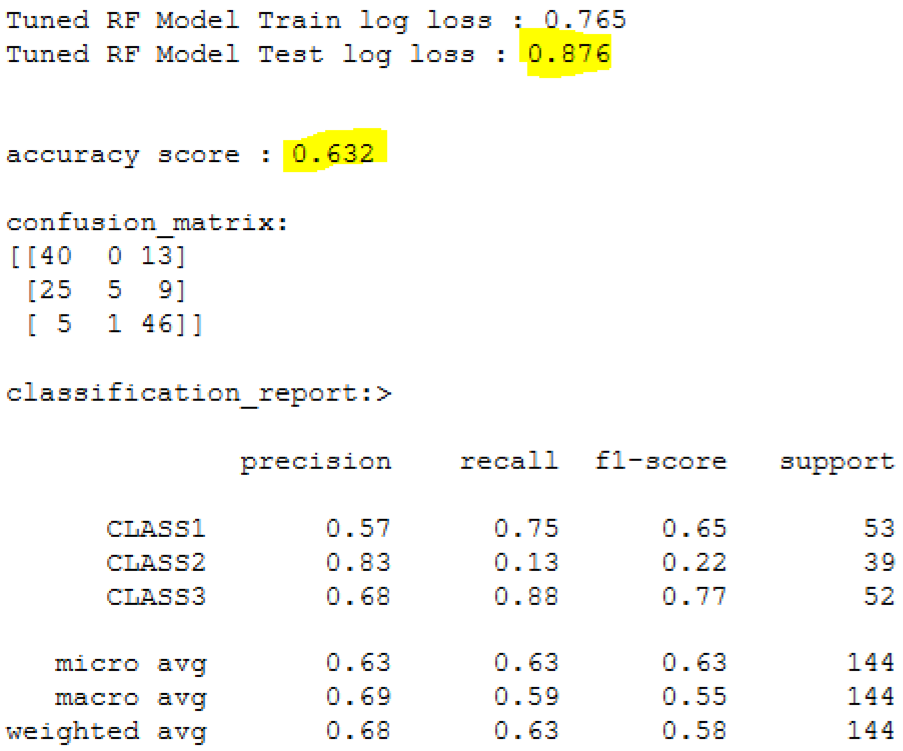
Hyperparameter tuning using GridSearchCV is performed to get a set of optimal hyperparameters.

```
{'max_depth': [1, 2, 3, 4, 5, None],
 'max_features': ['auto', 'sqrt'],
 'n_estimators': [50,
                  100,
                  150,
                  200,
                  250,
                  300,
                  350,
                  400,
                  450,
                  500,
                  550,
                  600,
                  650,
                  700,
                  750,
                  800,
                  850,
                  900,
                  950,
                  1000]}
Fitting 3 folds for each of 240 candidates, totalling 720 fits
```

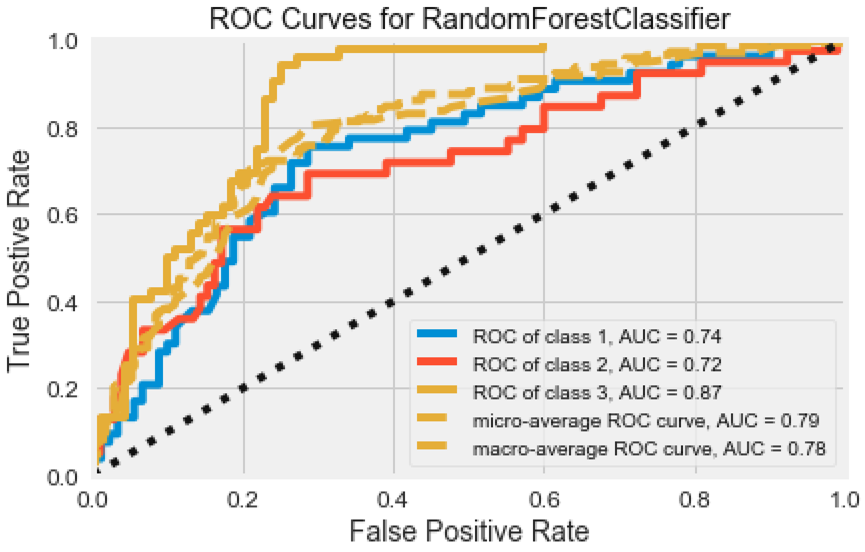


```
Fitting 3 folds for each of 240 candidates, totalling 720 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  33 tasks      | elapsed:   48.3s
[Parallel(n_jobs=-1)]: Done 154 tasks      | elapsed:  1.3min
[Parallel(n_jobs=-1)]: Done 357 tasks      | elapsed:  2.1min
[Parallel(n_jobs=-1)]: Done 640 tasks      | elapsed:  3.3min
[Parallel(n_jobs=-1)]: Done 720 out of 720 | elapsed:  3.7min finished
{'max_depth': 4, 'max_features': 'auto', 'n_estimators': 700}
```

Post Hyperparameter tuning, we get the error metric as below:



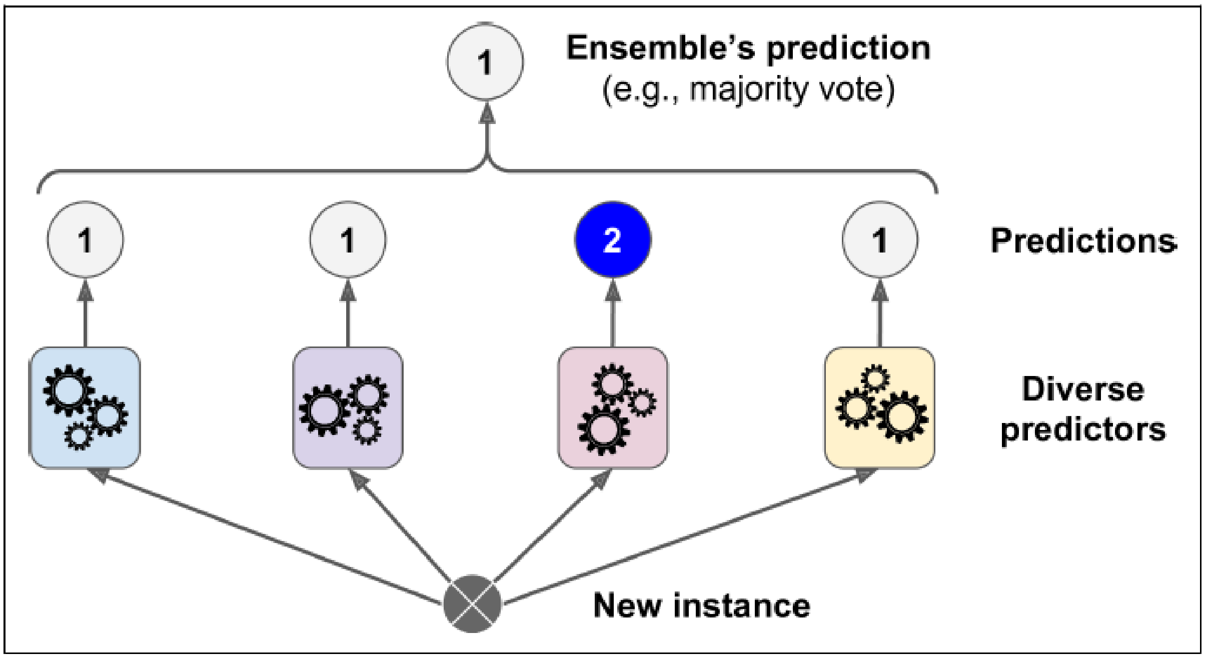
```
Tuned RF Model Train log loss : 0.765
Tuned RF Model Test log loss : 0.876


accuracy score : 0.632

confusion_matrix:
[[40  0 13]
 [25  5  9]
 [ 5  1 46]]

classification_report:>

              precision    recall  f1-score   support

      CLASS1       0.57      0.75      0.65        53
      CLASS2       0.83      0.13      0.22        39
      CLASS3       0.68      0.88      0.77        52

   micro avg       0.63      0.63      0.63       144
   macro avg       0.69      0.59      0.55       144
weighted avg       0.68      0.63      0.58       144
```

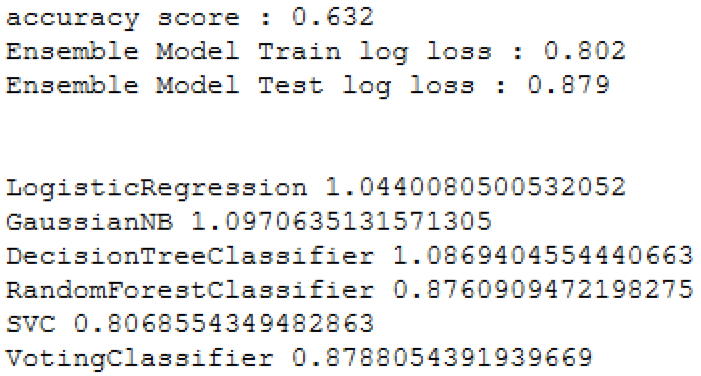


Result : Logloss on the test data is reduced, hence overfitting is reduced.

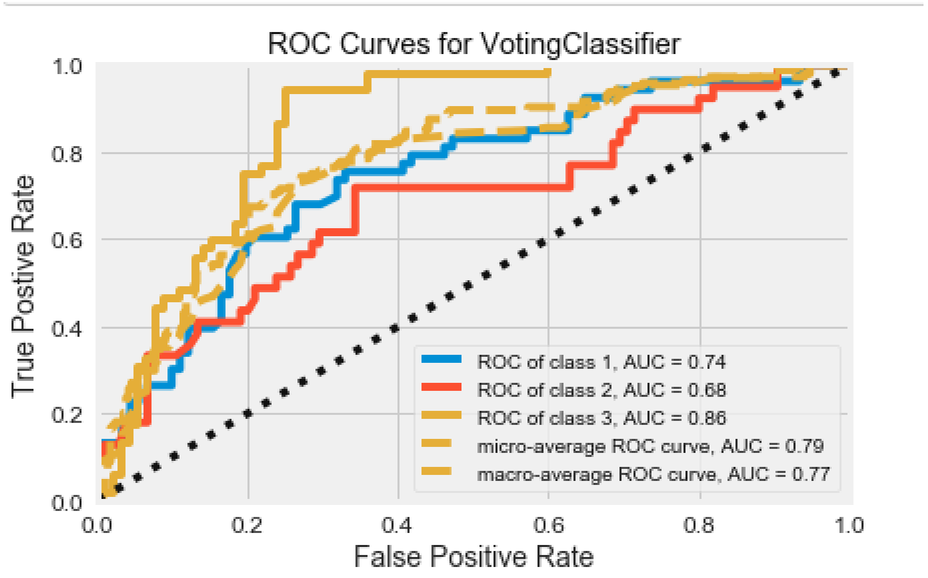
Ensemble method:

This is in simple terms **wisdom of the crowd**. if we aggregate the predictions of a group of predictors (such as classifiers or regressors), we will often get better predictions than with the best individual predictor. A group of predictors is called an ensemble; thus, this technique is called Ensemble Learning, and an Ensemble Learning algorithm is called an Ensemble method.





```
accuracy score : 0.632
Ensemble Model Train log loss : 0.802
Ensemble Model Test log loss : 0.879


LogisticRegression 1.0440080500532052
GaussianNB 1.0970635131571305
DecisionTreeClassifier 1.0869404554440663
RandomForestClassifier 0.8760909472198275
SVC 0.8068554349482863
VotingClassifier 0.8788054391939669
```



Result: Log loss, Accuracy score and AUC is similar to Random Forest.

XGBoost

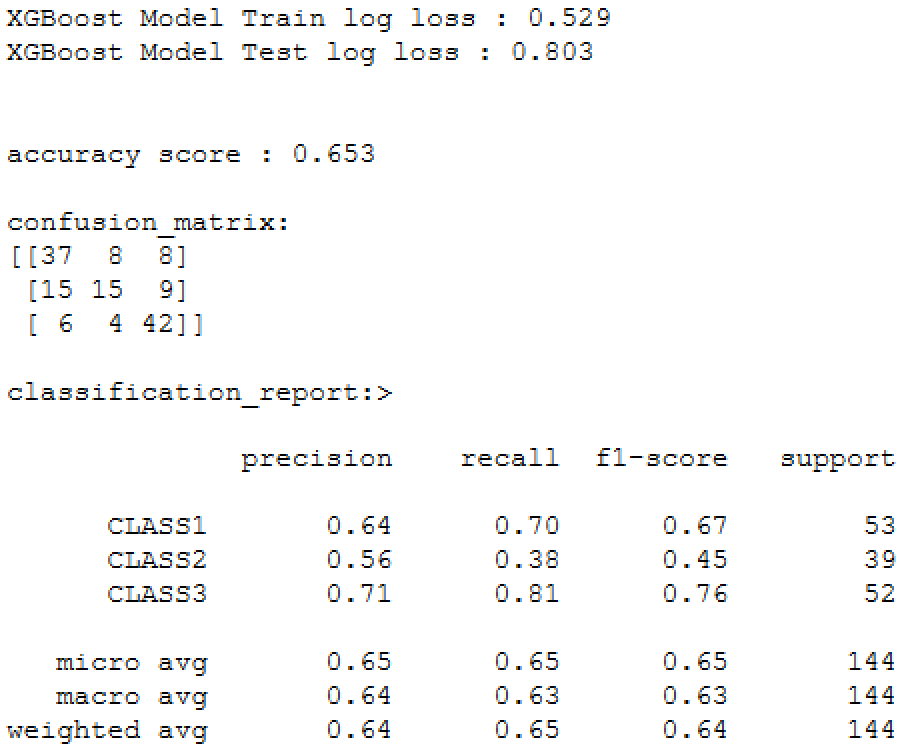
XGBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms and uses the gradient boosting (GBM) framework at its core.

But what makes XGBoost so popular?

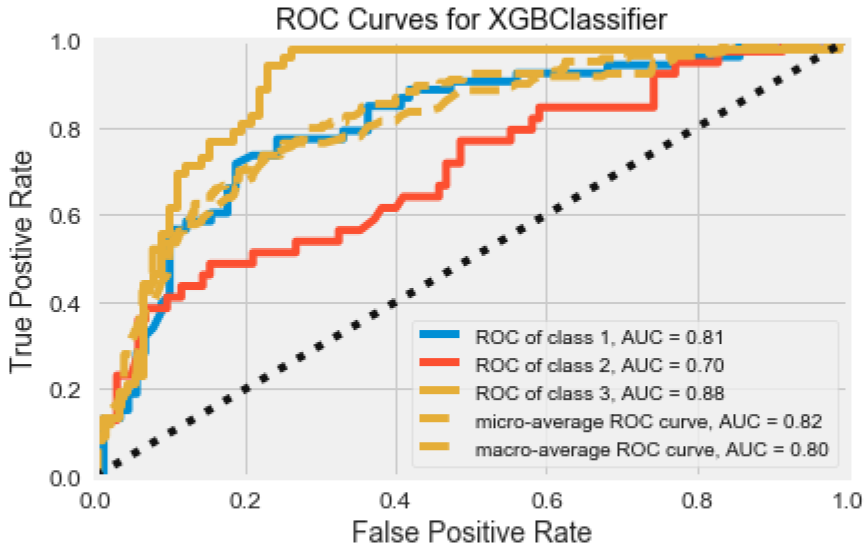
Speed and performance: Originally written in C++, it is comparatively faster than other ensemble classifiers.

Core algorithm is parallelizable: Because the core XGBoost algorithm is parallelizable it can harness the power of multi-core computers. It is also parallelizable onto GPU's and across networks of computers making it feasible to train on very large datasets as well.

Consistently outperforms other algorithm methods: It has shown better performance on a variety of machine learning benchmark datasets.



```
XGBoost Model Train log loss : 0.529
XGBoost Model Test log loss : 0.803


accuracy score : 0.653

confusion_matrix:
[[37  8  8]
 [15 15  9]
 [ 6  4 42]]

classification_report:>

              precision    recall  f1-score   support

      CLASS1       0.64      0.70      0.67        53
      CLASS2       0.56      0.38      0.45        39
      CLASS3       0.71      0.81      0.76        52

   micro avg       0.65      0.65      0.65       144
   macro avg       0.64      0.63      0.63       144
weighted avg       0.64      0.65      0.64       144
```



Result: Logloss is lowest for XGBoost. Also AUC is high, accuracy is good.

Chapter 3

Conclusion

3.1 Model Evaluation

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Predictive Performance
2. Interpretability
3. Computational Efficiency

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

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

Most commonly used metrics for multi-classes are F1 score, Average Accuracy, Log-loss.

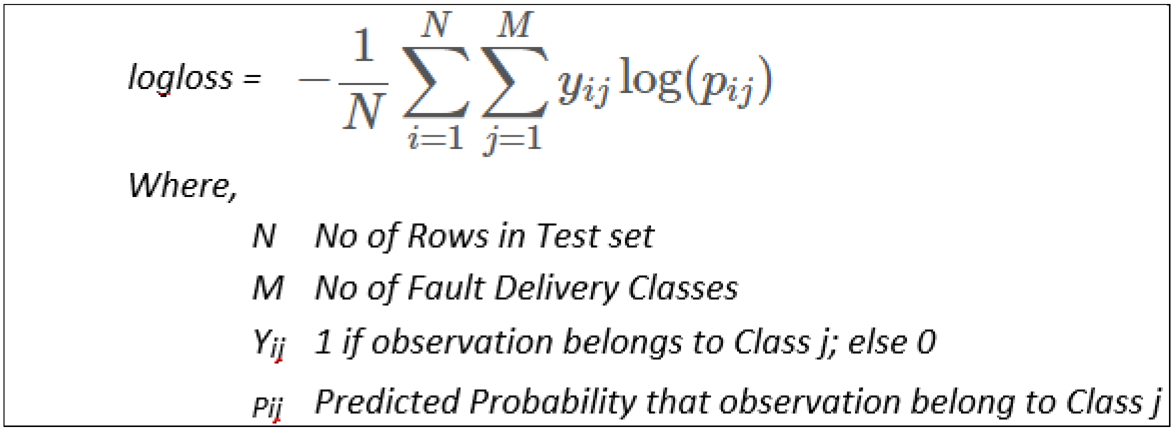
We have used log loss for our model.

3.1.1 Log-Loss

Logarithmic loss measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value. A perfect model would have a log loss of 0. Log loss increases as the predicted probability diverges from the actual label.

The logarithm used is the natural logarithm (base-e).

Log-loss for multi-class is defined as:



logloss = $$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

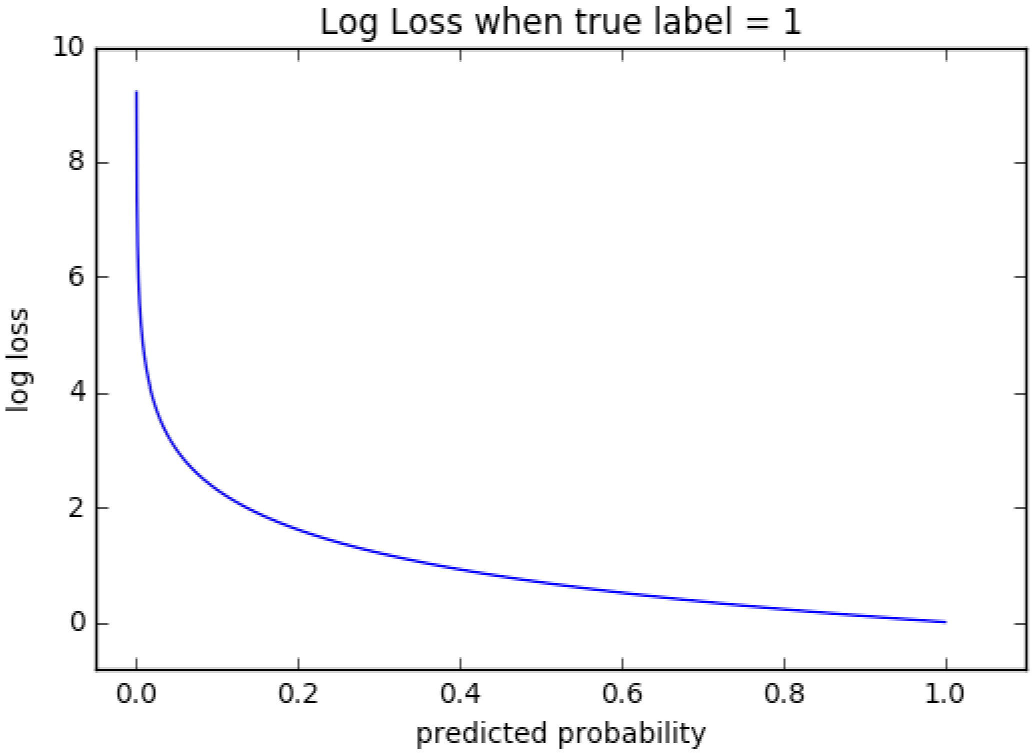
Where,

N *No of Rows in Test set*

M *No of Fault Delivery Classes*

Y_{ij} *1 if observation belongs to Class j; else 0*

P_{ij} *Predicted Probability that observation belong to Class j*



3.2 Model Selection

Logistic Regression	log loss	accuracy_score	ROCAUC
train	0.958		
test	1.044	0.424	0.65

Naïve Bayes	log loss	accuracy_score	AUC
train	0.963		
test	1.097	0.444	0.66

Decision Tree	log loss	accuracy_score	AUC
train	0.777		
test	1.087	0.653	0.75

Random Forest	log loss	accuracy_score	AUC
train	0.23		
test	1.137	0.653	0.76

Hypertuned Random Forest	log loss	accuracy_score	AUC
train	0.765		
test	0.876	0.632	0.78

Ensemble	log loss	accuracy_score	AUC
train	0.802		
test	0.879	0.632	0.77

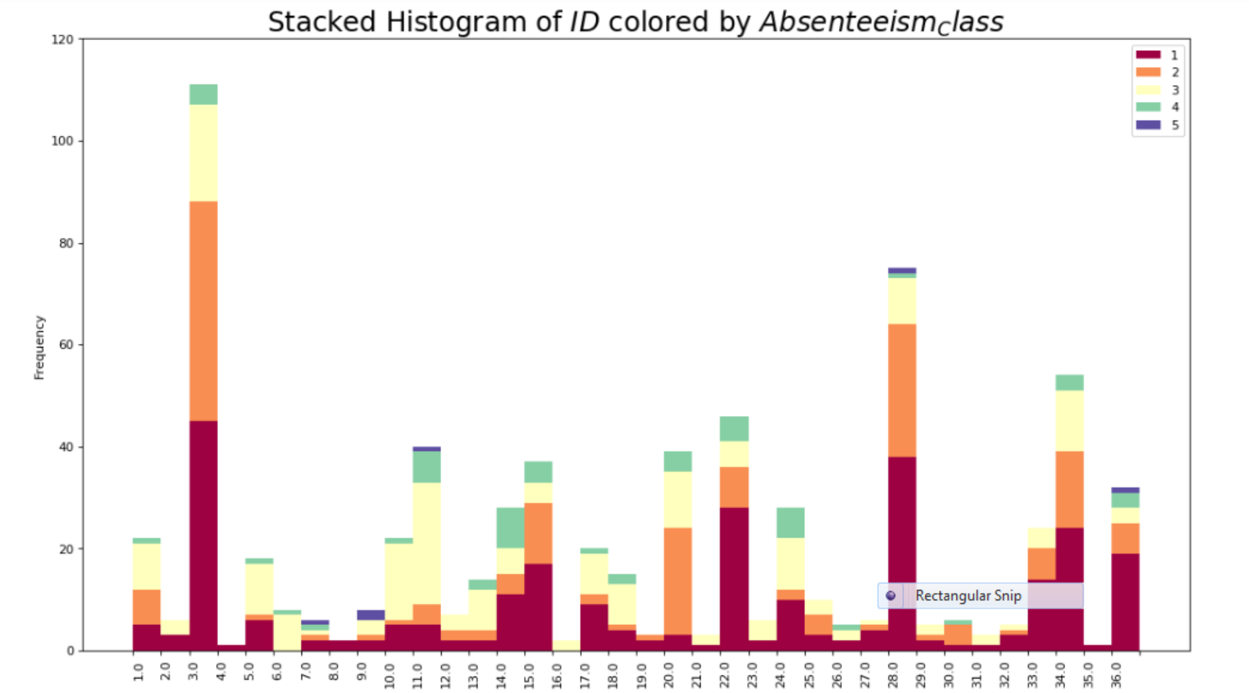
XGBoost	log loss	accuracy_score	AUC
train	0.529		
test	0.803	0.653	0.8

Conclusion : Logloss for test dataset is lowest for XGBoost. It means it is better at predicting on unseen data. Also AUC is higher comparatively other models and accuracy is also good.

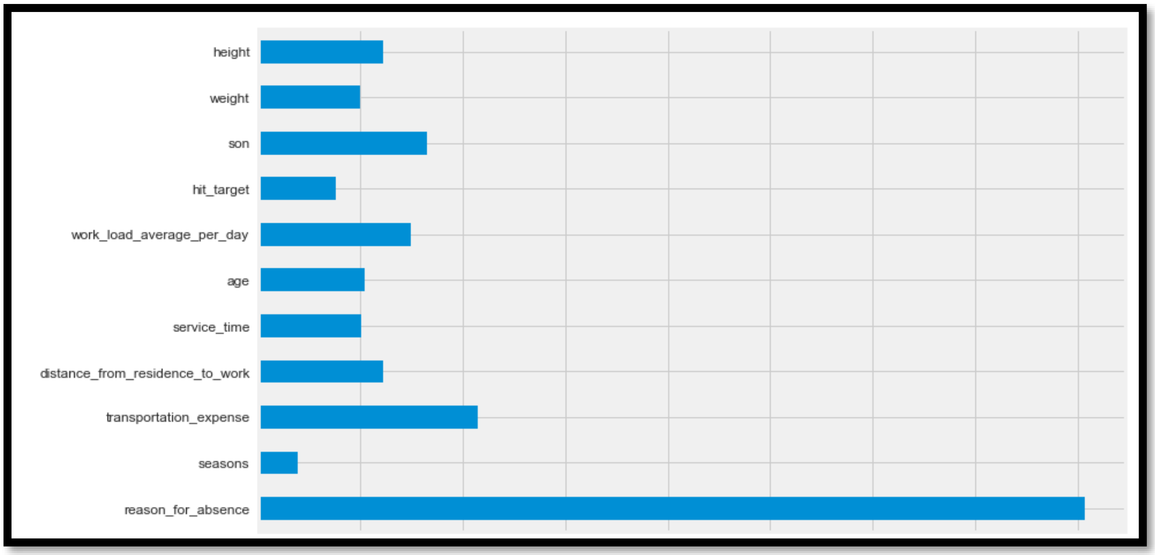
1. What changes company should bring to reduce the number of absenteeism?

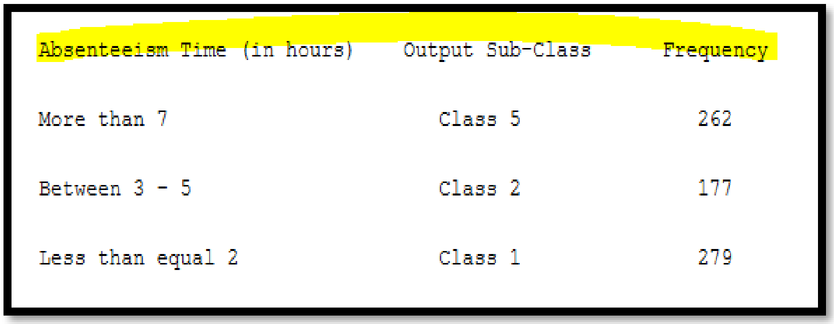
ANSWER: Employers can take below actions on some of these reasons to improve the productivity of their employees.

a) Based on the stacked histogram of emp ID by absenteesism class, we see IDs 7,9,11,28,36 have been absent for prolonged duration.We will now anlyse further for these ID to find the reason for their absence.

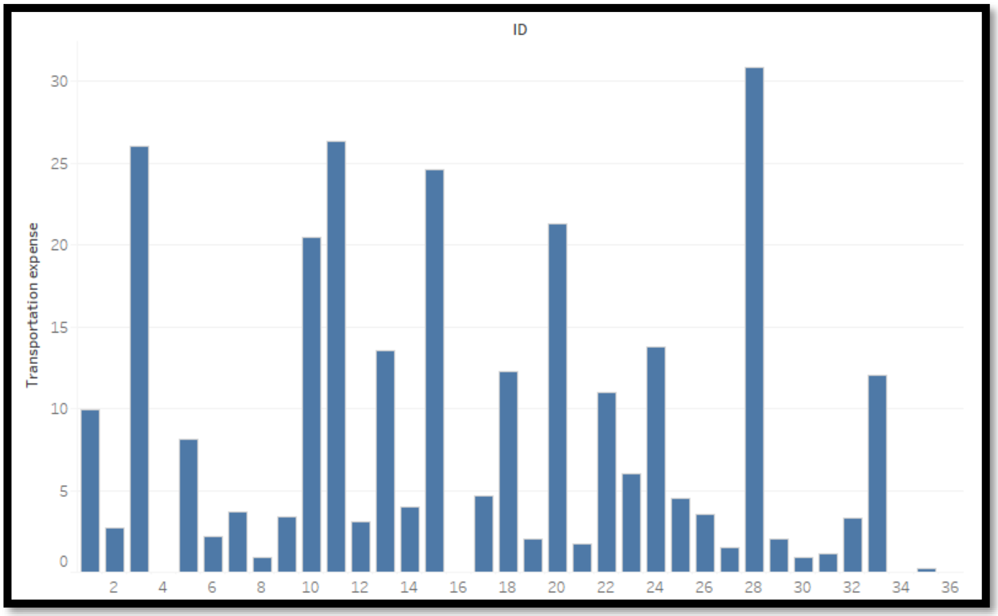


b) **Transportation expense**: This is second largest factor affecting absenteeism. and can be reduced either by granting a travel allowance to employees residing far from their workplace or twice a week “work from home” option. We see IDs 3,10,11,15,20,28 have high transportation expense.

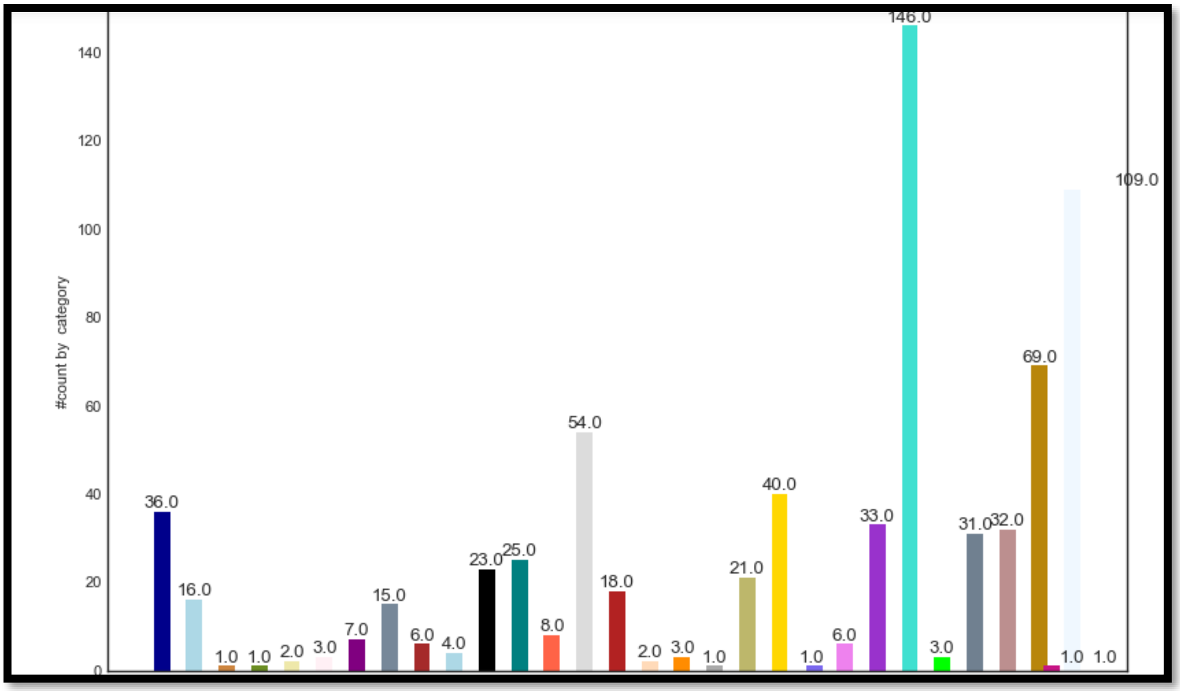




Absenteeism Time (in hours)	Output Sub-Class	Frequency
More than 7	Class 5	262
Between 3 - 5	Class 2	177
Less than equal 2	Class 1	279

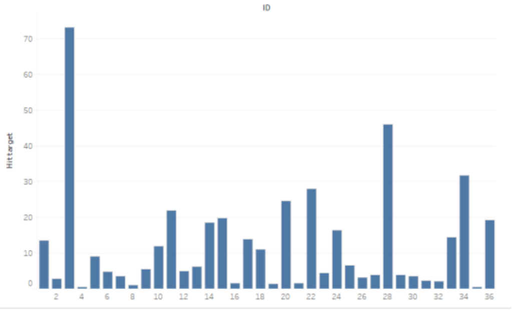


c) **Reason for absence (ICD)**: This is the single largest factor affecting absenteeism. We see Medical consultation (ICD=23) & dental consultation (ICD=28) are most frequent. Hence, Monthly medical and dental checkups can be planned in the office itself.

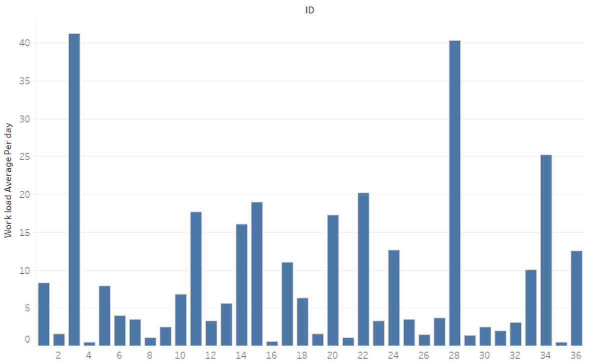


d) Work load as well as Hit target is comparatively higher for ID 3 and 28. This may be due to their prolonged absence. Hence, these employees need to be monitored for their absence

Histogram of Hit target vs ID



Bar plot of Work load average per day with ID

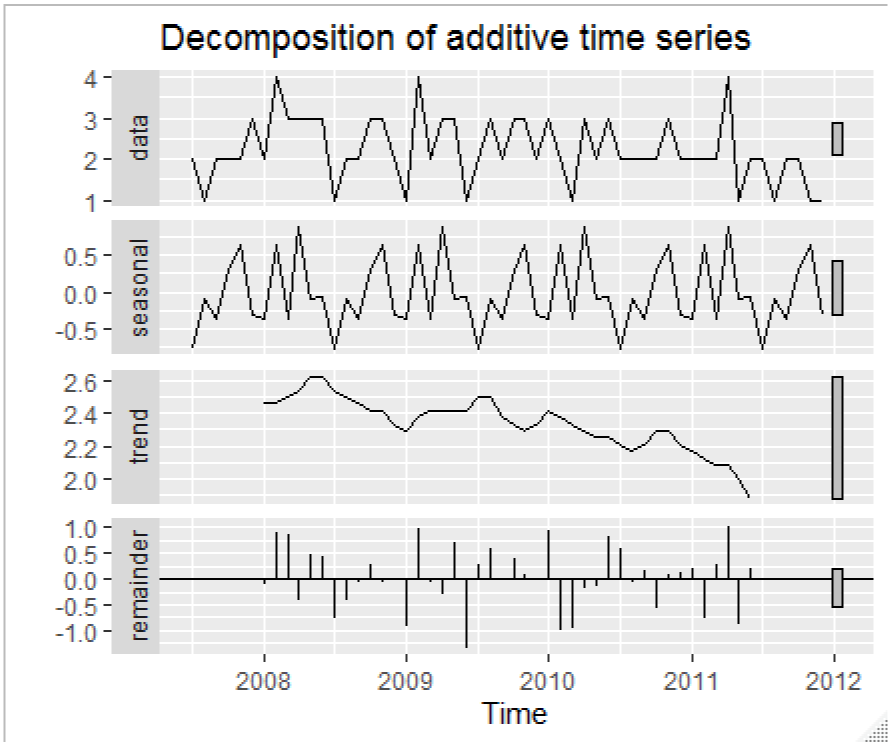


2. How much losses every month can we project in 2011 if same trend of absenteeism continues?

Note : I have used time series forecast for this question and is not related to the model chosen. I need to work on time-series problems.

ANSWER: The database was created with records of absenteeism at work from July 2007 to July 2010.

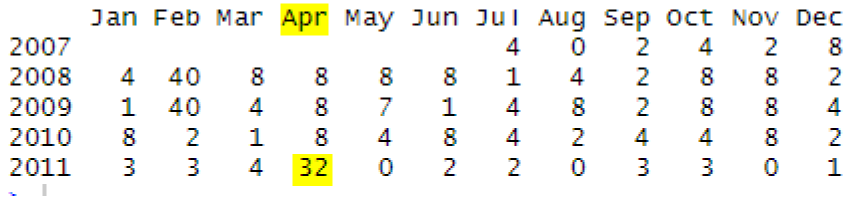
We can observe that the trend is on decline for 2011. Max absenteeism is in the month of April (32 hours).



The seasonal variation looks constant; it doesn't change when the time series value increases. Therefore, we have used the additive model.

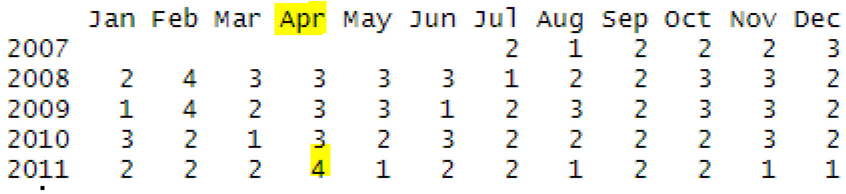
Time series = Seasonal + Trend + Random

Absenteeism (in hours) prediction for year 2011



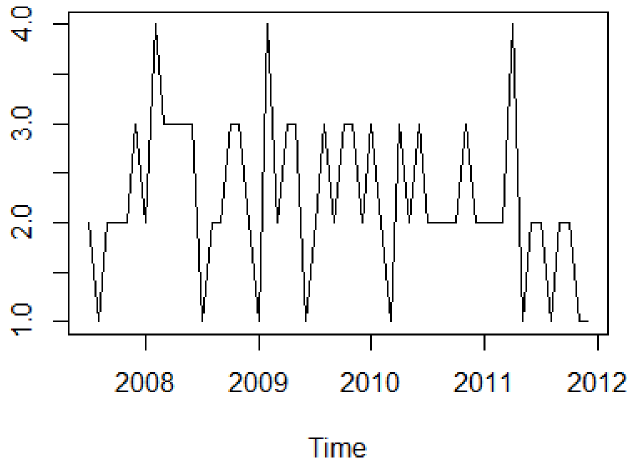
```
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2007                           4   0   2   4   2   8
2008   4  40   8   8   8   8   1   4   2   8   8   2
2009   1  40   4   8   7   1   4   8   2   8   8   4
2010   8   2   1   8   4   8   4   2   4   4   8   2
2011   3   3   4  32   0   2   2   0   3   3   0   1
```

Absenteeism class prediction for year 2011

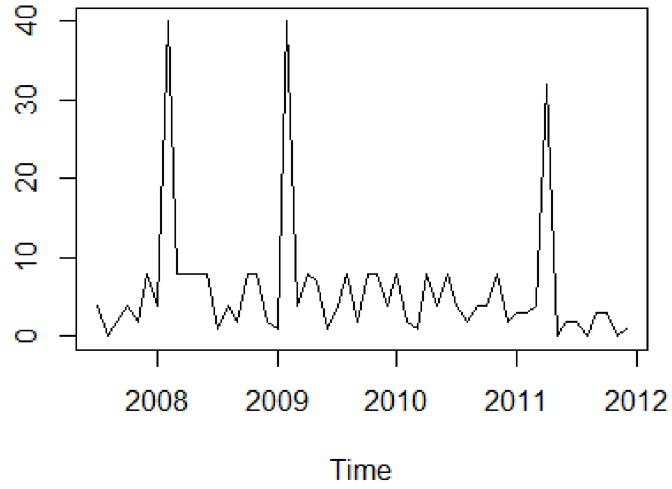


```
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2007                           2   1   2   2   2   3
2008   2   4   3   3   3   3   1   2   2   3   3   2
2009   1   4   2   3   3   1   2   3   2   3   3   2
2010   3   2   1   3   2   3   2   2   2   2   3   2
2011   2   2   2   4   1   2   2   1   2   2   1   1
```

Plot of Absenteeism class vs year



Plot of Absenteeism (time in hours) vs year


0
10
20
30
40
2008
2009
2010
2011
2012
Time

References:

https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html

https://stackoverflow.com/questions/736514/r-random-forests-variable-importance

https://www.rdocumentation.org/packages/MLmetrics/versions/1.1.1/topics/MultiLogLoss

https://en.wikipedia.org/wiki/Multiclass_classification