Employee Absenteeism at work

Mahesh Hiremath

30th July 2019

Contents

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. Since our target variable is continuous in nature, this is a regression problem.

Attribute 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 behavioural disorders
VI. Diseases of the nervous system
VII. Diseases of the eye and adnexa
VIII. VIII Diseases of the ear and mastoid process
IX. 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)

1.3 Exploratory Data Analysis

In the given data set there are 21 variables and data types of all variables are either float64 or int64. There are 740 observations and 21 columns along with missing values.

List of unique values present in each variable:

```
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
```

List of datatypes in each variable:

```
ID                                  int64
Reason for absence                float64
Month of absence                  float64
Day of the week                     int64
Seasons                             int64
Transportation expense            float64
Distance from Residence to Work   float64
Service time                      float64
Age                               float64
Work load Average/day             float64
Hit target                        float64
Disciplinary failure              float64
Education                         float64
Son                               float64
Social drinker                    float64
Social smoker                     float64
Pet                               float64
Weight                            float64
Height                            float64
Body mass index                   float64
Absenteeism time in hours         float64
dtype: object
```

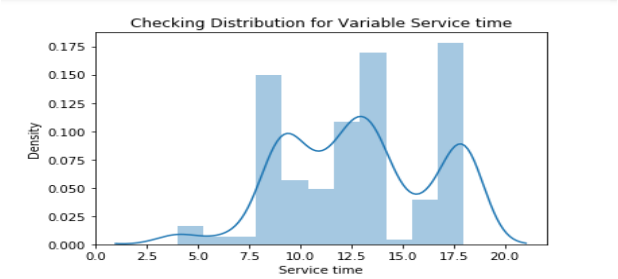
From exploratory data analysis we have concluded that there are 10 continuous variables and 11 categorical variables.

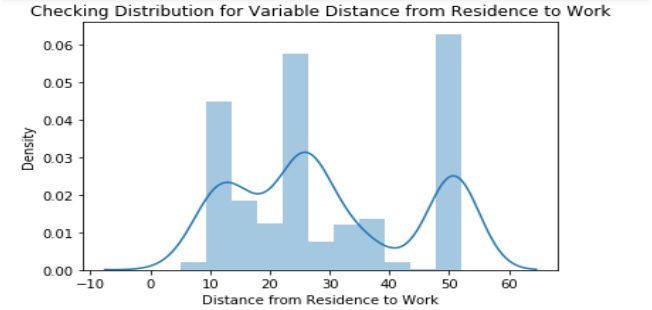
2. Methodology

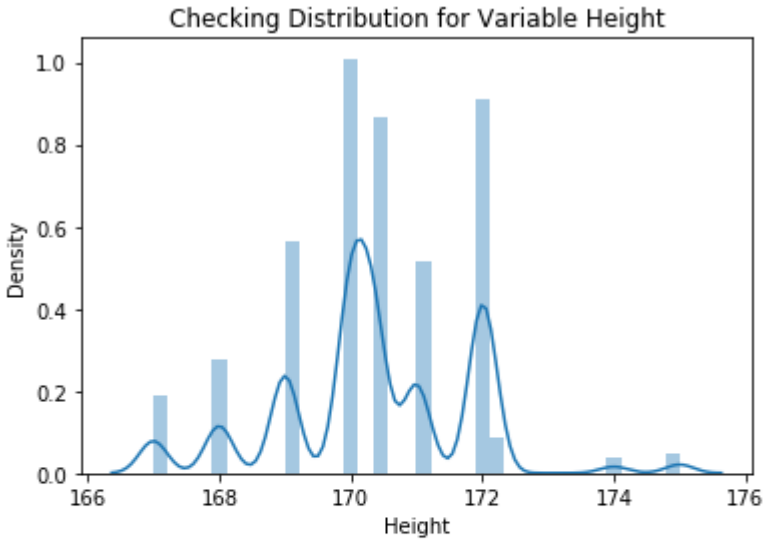
2.1 Pre Processing

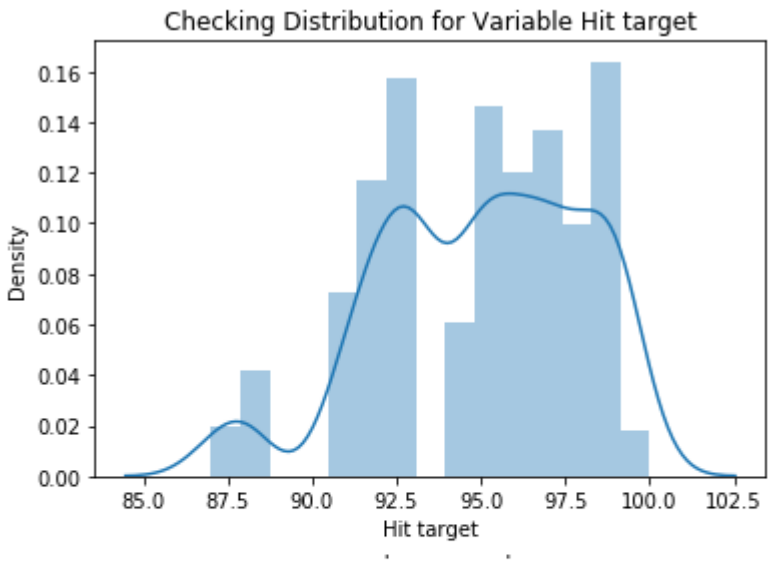
Data pre-processing is a stage where we have to look for clean/dirty data and to organise it accordingly as per the requirements of problem solving statement. In the process we do exploratory data analysis, it is used to explore the target and features so we know if we will need to transform or normalize some of the features based on their distribution, we might require to delete some because it might not give us any information in predicting future outcomes, or create some new features that might be useful for prediction.

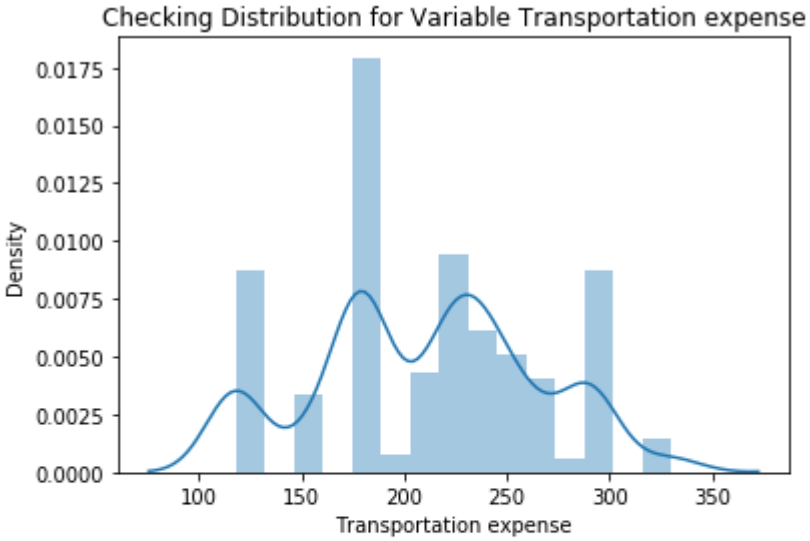
Following is the given data distribution-

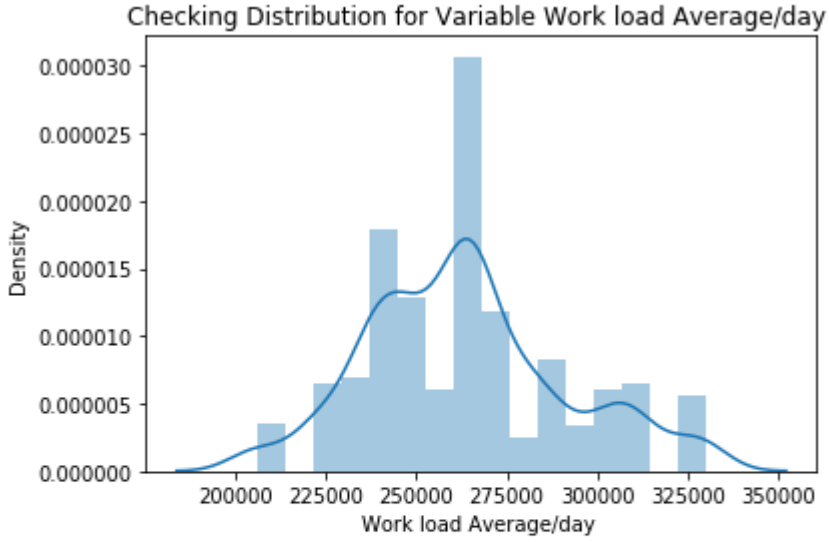


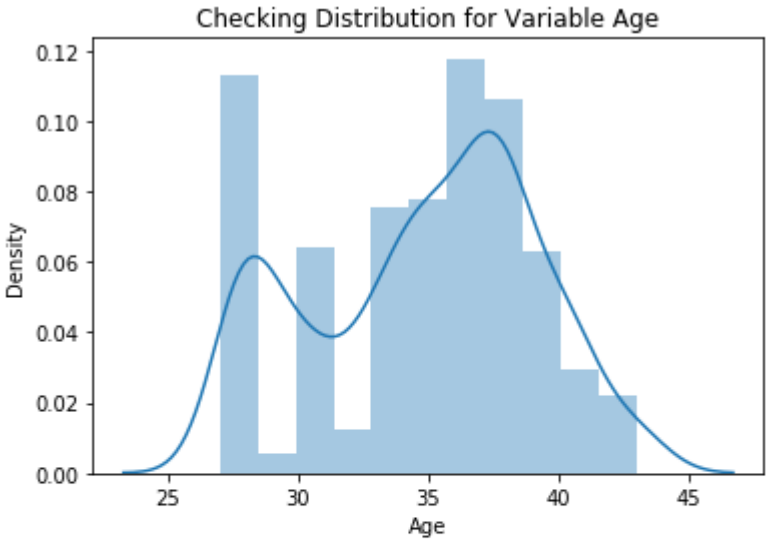


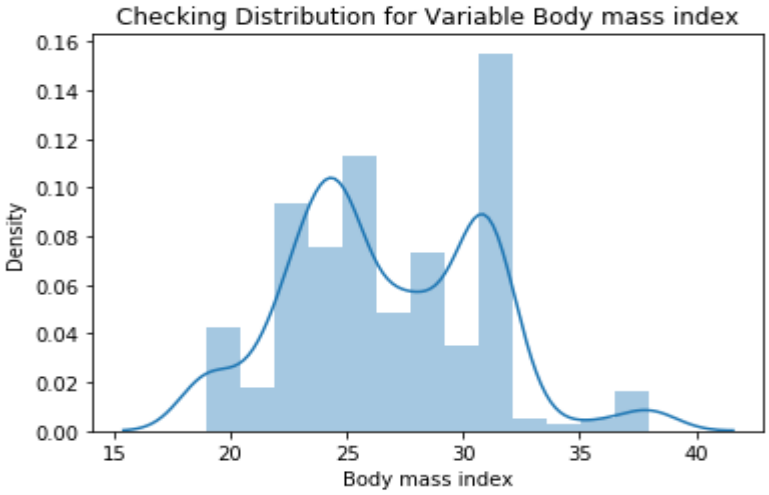

Checking Distribution for Variable Height
Density
Height
1.0
0.8
0.6
0.4
0.2
0.0
166
168
170
172
174
176


Checking Distribution for Variable Hit target
Density
Hit target
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
85.0
87.5
90.0
92.5
95.0
97.5
100.0
102.5


Checking Distribution for Variable Transportation expense
Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
100
150
200
250
300
350
Transportation expense


Checking Distribution for Variable Work load Average/day
Density
0.000030
0.000025
0.000020
0.000015
0.000010
0.000005
0.000000
200000
225000
250000
275000
300000
325000
350000
Work load Average/day


Checking Distribution for Variable Age
Density
0.12
0.10
0.08
0.06
0.04
0.02
0.00
25
30
35
40
45
Age

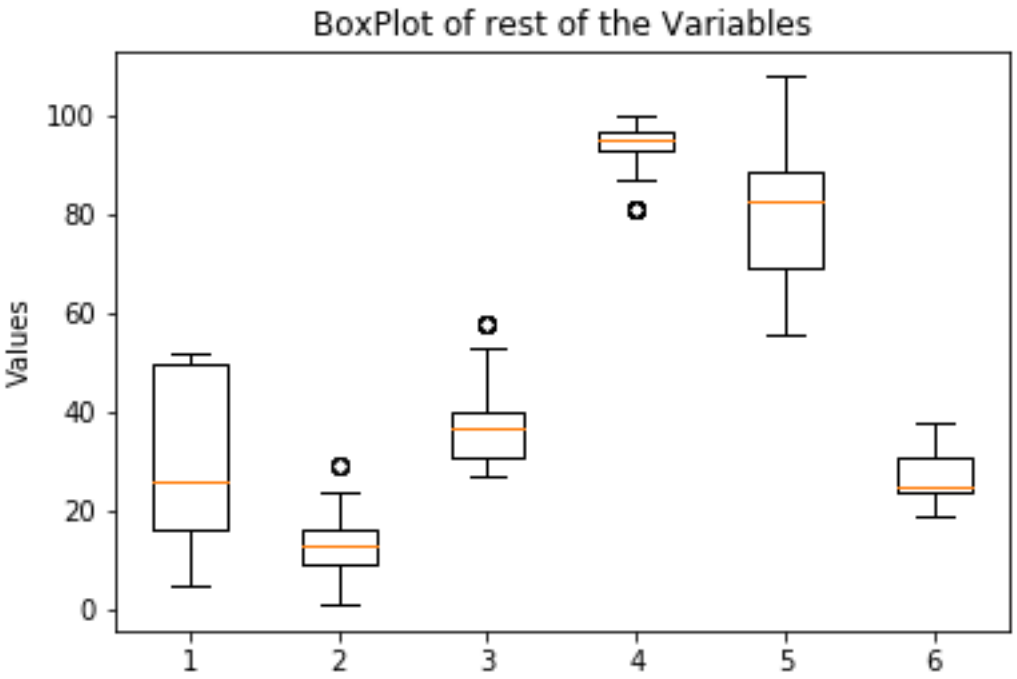

Checking Distribution for Variable Body mass index
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
15
20
25
30
35
40
Body mass index

2.2.1 Missing Value Analysis

In most of the data science projects almost half of the time is occupied to clean the data. Missing values or missing data occurs due to many reasons like failed to fill the redundant information, improper data transfer or when inappropriate information is stored. So if a column has more than 30% of data as missing values in such cases we ignore such columns only if its not the targeted variable. In the given data maximum percentage of missing values are present in column “Body mass index” i.e. 4.189%

2.1.2 Outlier Analysis

We can clearly observe from these probability distributions that 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 use an approach of removing outliers by visualizing the outliers using boxplots.



From the above boxplot we can observe that variables service time, age and hit target have outliers and variables distance from residence to work, weight and body mass index has no outliers.

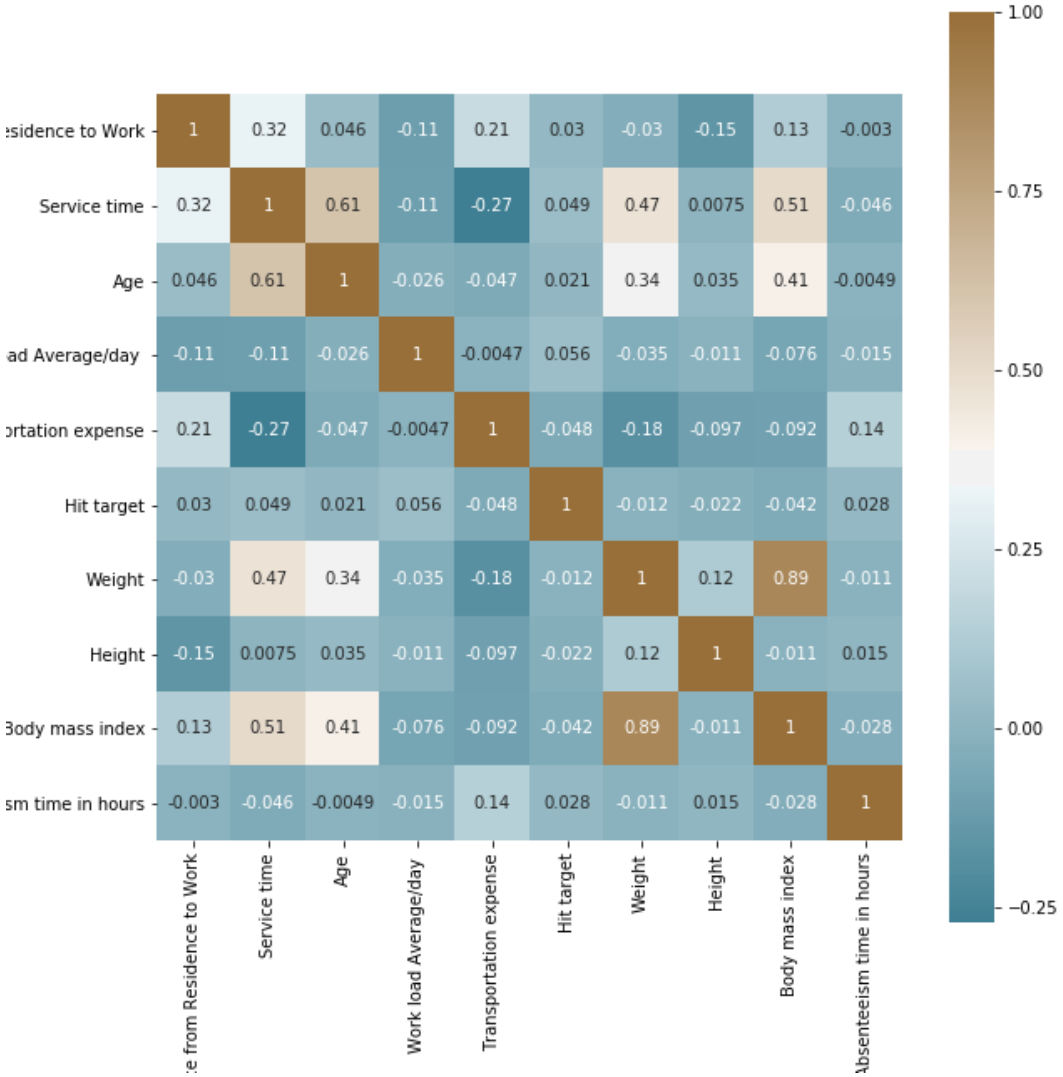
Therefore we have converted the outliers into NAs and imputed them with mean.

2.1.3 Feature Selection

Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model. The data features that we use to train our machine learning models have a huge influence on the performance.

Feature selection is the process where we can select those features which contribute most to our prediction variable and output. This helps to avoid the problem of multi-collinearity. In this project we have selected **"**Correlation Analysis" for numerical variable and "ANOVA (Analysis of variance) for categorical variable".

Following is the heatmap of correlation analysis-



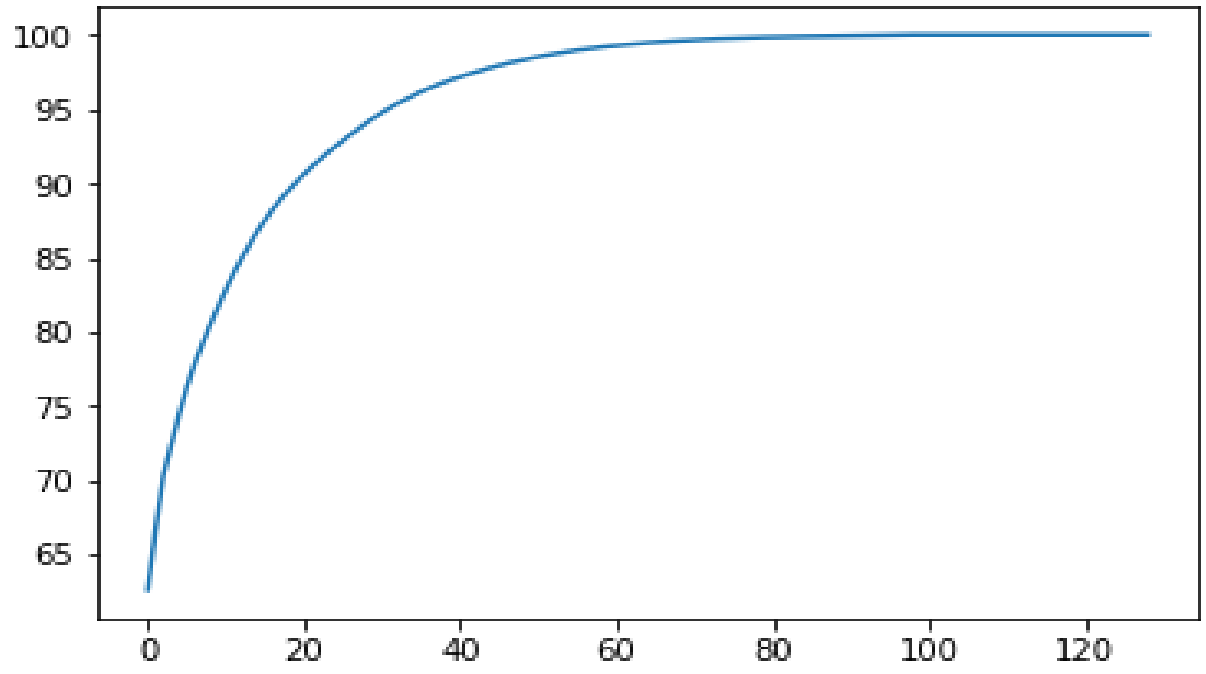
From the above heatmap it is clear that variables "Weight" and "Body mass index" have high correlation i.e. 0.89 which is greater than 0.7, therefore we have excluded the weight column.

2.2.4 Feature Scaling

Feature scaling is a method used to standardize 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. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since our data is not uniformly distributed we will use “Normalization” as Feature Scaling Method.

2.2.5 Principal Component Analysis

Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data. After creating dummy variable of categorical variables the shape of our data became (718,123) this high number of columns leads to bad accuracy.



We have applied PCA algorithm on our data and from the above graph we have concluded that 45 variables out of 107 explains more than 95% of data. So we have selected only those 45 variables to feed our models.

2.2 Modeling

After cleaning and preprocessing the data we would be using models to train on this data and to find the required information. We will be using regression models to predict the target variable.

Followings are the selected models-

2.2.1 Decision Tree

It is a supervised machine learning algorithm. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Each branch connects nodes with “and” and multiple branches are connected by “or”. It can be used for classification and regression. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users.

The RMSE value and R^2 value for our project in Python –

Decision Tree	PYTHON
RMSE Train	0.566
RMSE Test	0.546
R^2 Test	0.972

2.2.2 Random Forest

Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset. 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. It means to build each decision tree on random forest we are not going to use the same data. The RMSE value and R^2 value for our project in Python-

Random Forest	PYTHON
RMSE Train	0.018
RMSE Test	0.009
R^2 Test	0.999

2.2.3 Linear Regression

Linear Regression is one of the statistical methods of prediction. It is applicable only on continuous data. To build any model we have some assumptions to put on data and model. Here are the assumptions to the linear regression model.

The RMSE value and R^2 value for our project in Python-

Linear Regression	PYTHON
RMSE Train	3.941
RMSE Test	0.011
R^2 Test	0.999

2.2.4 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Gradient Boosting	PYTHON
RMSE Train	0.0003
RMSE Test	0.0002
R^2 Test	0.999

Conclusion

Here we are going to select the best model suited for our project and to find the solutions required.

3.1 Model Evaluation

The “Root Mean Square Error (RMSE)” is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are, RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Whereas R-squared is a relative measure of fit, RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE and higher value of R-Squared Value indicate better fit.

3.2 Model Selection

From the observation of all RMSE Value and R-Squared Value we have concluded that Linear Regression Model has minimum value of RMSE and it’s R-Squared Value is also maximum (i.e. 1). The RMSE value of Testing data and Training does not differs a lot this implies that it is not the case of overfitting.

3.3 Answers of asked questions

1. Distance from residence to work: Based on the data the employees having shorter distance from home to work has BMI issues i.e. they are not so healthy as others.
2. Service Time: Employees having higher service time has higher BMI i.e. positively correlated so may suffer from various diseases or lack of fitness.
3. Age: Employees having more work experience tends to work more than others and has higher weight and BMI i.e. they are more productive and obese than compared to other employees.
4. Work load Average/day: Employees having more work load average/day has lower BMI and more hit targets.
5. Transportation Expenses: Employees with more transportation expenses remains more absent at work and are more distant to work from residence.

References

1. For Data Cleaning and Model Development - https://edwisor.com/career-data-scientist , https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b

2. For PCA - https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/