**Employee Absenteeism**

Aditya Kanungo  
15th July 2018

**Contents**

**1 Introduction**  
 1.1 Problem Statement   
 1.2 Data

**2 Methodology**

2.1 Pre-Processing

2.1.1 Missing value-analysis  
 2.1.2 Dropping un-necessary variables (Non-predictors)

2.1.3 Replacing some variables and converting into categorical  
 2.1.4 Visualization of dataset  
 2.1.5 Chi-Square test for correlation analysis of categorical variables  
 2.1.6 Outlier Analysis  
 2.1.7 Feature Scaling

2.2 Modeling  
 2.2.1 Model Selection  
 2.2.2 Binary Classification  
 2.2.3 Logistic regression  
 2.2.4 Gaussian Naive Bayes Classifier  
 2.2.5 k-Nearest Neighbors

**3 Conclusion**

3.1 Model Evaluation

**Appendix A – Python Code  
Appendix B – R Code  
References**

**Chapter 1**

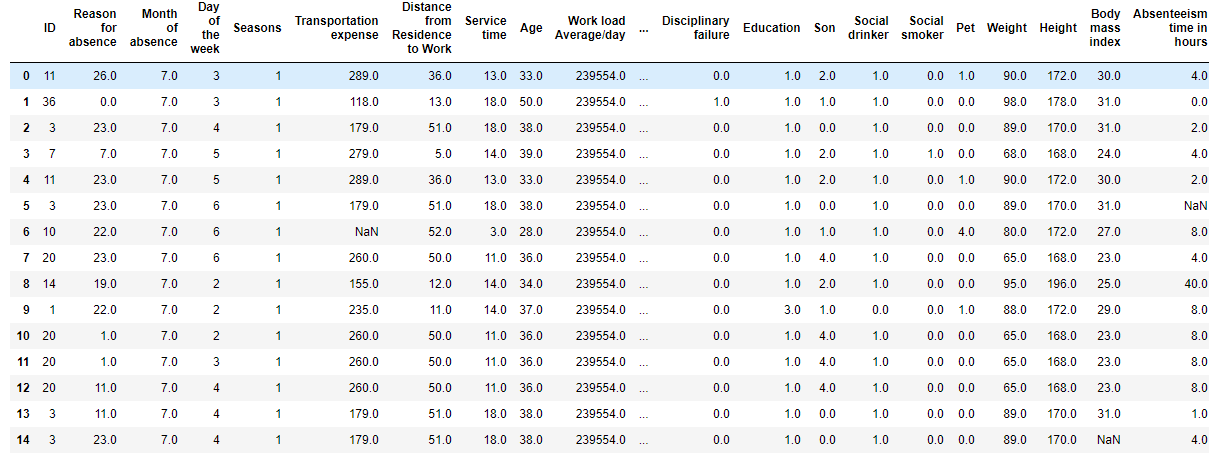
**Introduction**

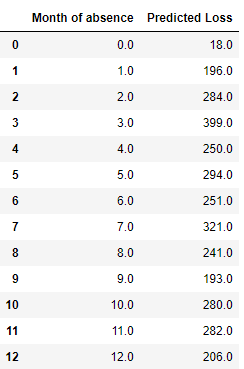
* 1. **Problem Statement**

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. We are expected to develop an algorithm to predict the loss in absenteeism based on employee details. The company has shared it dataset and requested to have an answer on the following areas:

1. What changes company should bring to reduce the number of absenteeism?
2. How much losses every month can we project in 2011 if same trend of

absenteeism continues?

* 1. **Data**
     1. **Employee Absenteeism Dataset:**
* (ID, Reason for absence, Month of absence, Day of the week, Seasons, Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Hit target, Disciplinary failure, Education, Son, Social drinker, Social smoker, Pet, Weight, Height, Body mass index, Absenteeism time in hours])
* **(**740 rows × 21 columns)
* 
  + 1. **Final submission:**
* Set of rules or changes to implemented in order to reduce absenteeism
* Monthly loss for next year when previous trends continues.



* + 1. **Predictors:**
* Using these we have to correctly predict the monthly absenteeism.

**Chapter 2**

**Methodology**

**2.1 Pre-Processing**

Before implementing any Machine Learning model on our training data, we are required to look at the data before we start modeling. However, in data science 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.

In our case the raw data that we will be working on employee absenteeism dataset that has employee absenteeism details along with their work-related details.

**2.1.1 Missing value-analysis**

We will first check for any missing values in the data to check if any variable is null or empty. During which we found that there are few missing values in our dataset.

Missing values in Numerical data: In numeric data we can treat missing values by imputing it with mean value.

Missing values in Category data: In category data we can treat missing value by copying the value of categorical variable adjacent to the same value of another highly correlated variable.

**2.1.2 Dropping un-necessary variables (Non-predictors)**

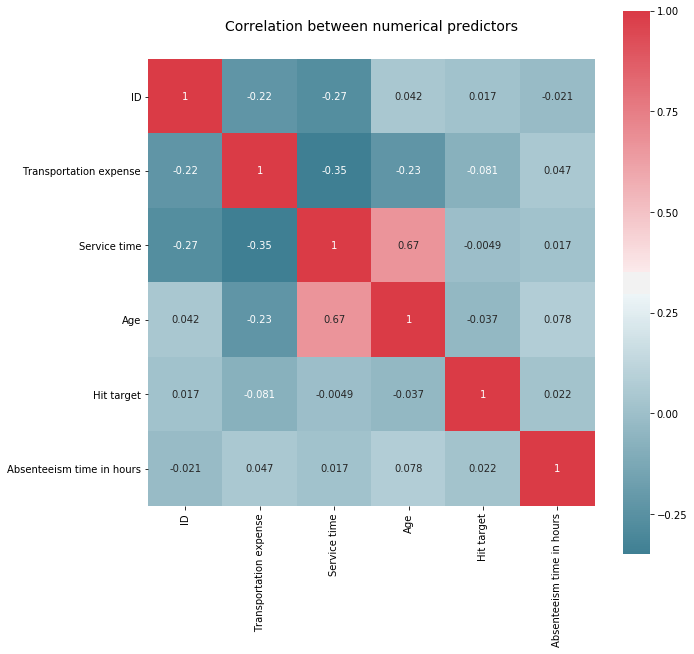
Dropping those variables from the train and test dataset which are not predictors and won’t be of any significance in the analysis to predict the Churn Score.

* Drop ‘ID’ from our dataset.
  + 1. **Replacing some variables and converting into categorical**

In our dataset we have many category variables with more than 2 unique values. We will convert the datatype of these variables into category or factor.

**2.1.4 Visualization of dataset**

2.1.3.3 Correlation analysis using heatmap



**From the above heat-map we made in Python we can infer the following:**

- None of the numeric variables are highly correlated.

**2.1.4 Outlier Analysis**

In [statistics](https://en.wikipedia.org/wiki/Statistics), an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the [data set](https://en.wikipedia.org/wiki/Data_set). An outlier can cause serious problems in statistical analyses.

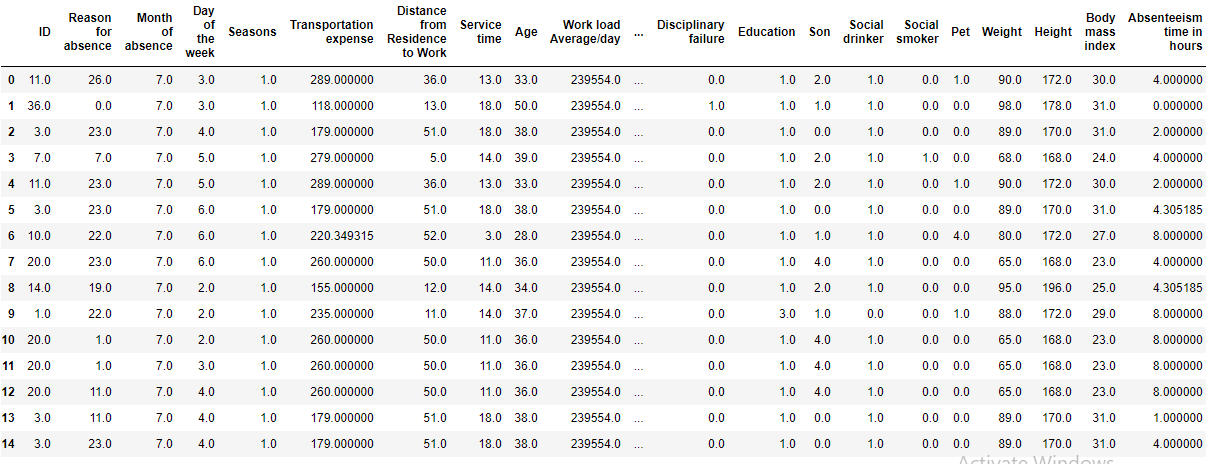
Initial Box-plot analysis confirms significant outliers in the numeric data. So, we will use boxplot method and inter quartile range (IQR) to find outliers in our dataset and replace it with null value or NA. Then we can impute the NA or the replaced outlier values using mean, median or knn-imputation method.

**2.1.4 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](https://en.wikipedia.org/wiki/Machine_learning) algorithms, objective functions will not work properly without [normalization](https://en.wikipedia.org/wiki/Normalization_(statistics)). Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

In order to feed the data to our model we need to scale our data so that we can compare and analyze different variables on the same ground. We will use standardization method to scale our numeric data for further analysis.

2.1.4.1 Dataset after data preprocessing



**2.3 Modeling**

**2.3.1 Model Selection**

In early stage of our analysis process we have come to understand that Our objective is to predict how much losses every month can we project in 2011 if same trend of absenteeism continues, which we can predict in terms of number of house of absenteeism i.e. how much hours of absenteeism the company faces each month. Also, we have to deduce what changes the company should bring to reduce the number of absenteeism, for which we can use exploratory data analysis and different visualization techniques to understand the absenteeism trends.

We can apply different regression models such as Linear regression and decision tree regressor on our dataset and choose the model with the best accuracy. As our target variable is continuous.

**2.3.2 Regression**

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

**2.3.2 Linear regression.**

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

**2.3.2 Decision Tree regressor**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

**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 of employee absenteeism, the latter two, Interpretability and Computation Efficiency, do not hold much significance. Here, we will use Mean Squared Error MSE 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 accuracy or some average error measure.

In our case of Ads classification, we have applied four models namely linear regression and Decision tree regressor.

**References**

[**https://www.analyticsvidhya.com/**](https://www.analyticsvidhya.com/)

[**https://machinelearningmastery.com**](https://machinelearningmastery.com)

[**https://docs.python.org/3/**](https://docs.python.org/3/)