A Project Report on

**EMPLOYEE ABSENTEEISM**

Samrudhi Singh

16 March 2019

**Contents**

**1. Introduction**

1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

1.3 Exploratory Data Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. 5

**2. Methodology**

2.1 Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. 6

2.1.1 Missing Value Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .7

2.1.2 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

2.1.3 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

2.1.4 Feature Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.2.1 Decision Tree. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.2.2 Random Forest . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .11

2.2.3 Linear Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

**3. Conclusion**

3.1 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

3.2 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

3.3 Answers of asked questions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

**Appendix**

Extra Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17

**References**

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

Dataset Details:

Dataset Characteristics: Timeseries Multivariant

Number of Attributes: 21

Number of Observations: 740

Missing Values: Yes

**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)

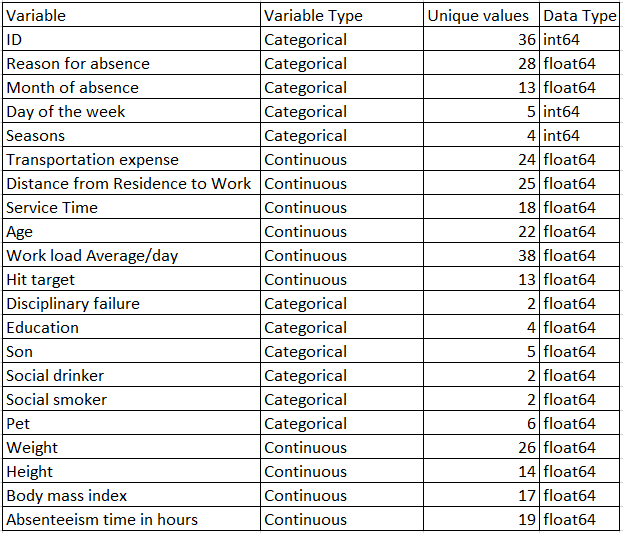
**18.** Weight

**19.** Height

**20.** Body mass index

**21**. Absenteeism time in hours (target)

The target variable for our case is Variable#21 - Absenteeism in hours. As it is given it is continuous in nature, we shall consider our problem to be of Regression Problem.



Chapter 2 Methodology

2.1 Pre Processing

Data preprocessing involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

At many times the input data is (a) Noisy: containing errors or outliers (b) Inconsistent: containing discrepancies in codes or names.

To first start for it we shall start with plotting histograms for the variables. This will help us to find whether our data is normally distributed or not. Regression analysis problems requires data to be normally distributed before feeding it to the model.

FIGURES

Missing Value Analysis:

The missing data may reduce the precision of calculated statistics because there is less information than originally planned. Hence missing value analysis is necessary.

Suppose the number of cases of missing values is extremely small; then, an expert researcher may drop or omit those values from the analysis.  In statistical language, if the number of the cases is less than 5% of the sample, then the researcher can drop them.

One the other hand if there is a larger number of missing values, i.e. 30% then it is better to drop those cases (rather than do imputation) and replace them.  On the other hand, in univariate analysis, imputation can decrease the amount of bias in the data, if the values are missing at random.

Since the data which we have received consists of missing values, we need to first filter our data and find these missing values. After that we need to apply the best method which is computing the missing values and giving the nearest possible value. However, we could have gone by deleting the missing values since the maximum is 4.189% missing in case of Body Mass Index column, but I did not do say since the data set was already very small.

For our project we have used KNN Imputation in R and Median method in Python to compute missing values.

FIGURE

Outlier Analysis

there exist data objects that do not comply with the general behaviour or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers. An outlier is an observation point that is distant from other observations.

We can make use of Boxplots to visualize the outliers.

For our case we have plotted Box plots for each of continuous variable against our target variable i.e. **Absenteeism time in hour**. Below boxplots have been helpful in deriving that there are many extreme values present in the variables. These extreme values may represent some anomaly or error in the data collection process.

FIGURES

From the box plots we can infer that almost all the variables have outliers except a few. For those data sets we have converted them first to NA and then computed them using the best method. We have used KNN Imputation in R and Median Method in Python for this purpose.

Feature Selection:

The process in which we select those features in our data that are most useful or most relevant for the problem being worked upon is a process called feature selection. Features in simple language are the variables present in the data set under consideration. Feature selection is also called variable selection or attribute selection. It mostly acts as a filter, muting out features that aren’t useful in addition to your existing features.

Feature selection is mainly done on the concept of multi collinearity. For 2 or more variables which are highly correlated to one another we can make use of only one variable and drop the rest of the variables which are just additionally adding to the unnecessary information in the data set.

For our project we have used ANOVA method for categorical variables and Correlation Analysis(with help of Corrgram Visualization).

FIGURE

Feature Scaling:

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidean distance between two data points in their computations, this is a problem. Suppose if we have a data set of weight having values like 5kg and 5000gms. In this case the features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To supress this effect, we need to bring all features to the same level of magnitudes.

There are majorly 2 methods followed for Feature Scaling process, namely Normalization and Standardization.

Normalization rescales the data in the range of features to scale it in [0, 1] or [−1, 1].

While Standardization transforms the data to have zero mean and a variance of 1.

For our project we have used Normalization method as the data was not uniformly distributed.

Modeling

After all the pre-processing steps are done, we will be using some data prediction models on our processed data to predict the target variable. We have built following models on our data set:

Decision Tree:

A Decision Tree is an algorithm used for supervised learning problems such as classification or regression. A decision tree is an effective machine learning modeling technique for regression and classification problems. To find solutions a decision tree makes sequential, hierarchical decision about the target variable based on the predictor data. The model is constructed based on the observed data.

Decision tree models where the target variable uses a discrete set of values are classified as Classification Trees.

A decision tree where the target variable takes a continuous value, usually numbers, are called Regression Trees. The two types are commonly referred to together at CART (Classification and Regression Tree).

Random Forest:

Random forests are a popular ensemble method that can be used to build [predictive models](https://www.datascience.com/resources/white-papers/executives-guide-to-predictive-data-modeling) for both classification and regression problems. In the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer. 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.

Linear Regression:

Linear regression is a basic and commonly used type of predictive analysis. Linear regression is used to describe relationship among variables. Linear Regression may be used for predicting one variable or it may predict multiple variables. These cases are known as Simple Linear Regression and Multiple Linear Regression respectively.

The one simple case is where a dependent variable may be related to independent or explanatory variable. The following equation can be used for depicting linear relationship:

y = b0+b1x

**Chapter 3**

**Conclusion**

After the development of our models, we are going to evaluate our models, select the best model for our dataset and try to get answers of the asked questions.

**3.1 Model Evaluation**

For evaluating our model, we shall take help of various error metrices, to support our model selection criteria.

Since, in our project the target variable prediction was that of a continuous variable type, we have made use of the following metrices:

**Root Mean Square Error** (RMSE):

Root Mean Square Error (RMSE) is the standard deviation of the [residuals](https://www.statisticshowto.datasciencecentral.com/residual/) (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.

**R-Squared:**

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

Lower values of **RMSE** and higher value of **R-Squared Value** indicate better fit.

**3.2 Model Selection**

On the basis of the RMSE and R- Squared value for the 3 models developed, we can decide which model to select for our project: