PREDICTING THE CAUSE OF ABSENTEEISM

&

LOSS PREDICTION OF A COURIER COMPANY

Praveen Kumar Singh

25-11-2019

**Contents**

**1 Introduction 2**

1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

**2 Methodology 4**

2.1 Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.1.1 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2.1.2 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.2.1 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.2.2 Multiple Linear Regression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.2.3 Regression Trees . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

2.2.4 Classification . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

**3 Conclusion 15**

3.1 Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

3.1.1 Mean Absolute Error (MAE) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

3.1.2 Mean Squared Error (MSE) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

3.2 Model Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

**Chapter 1**

**Introduction**

* 1. Problem Statement : XYZ is a courier company which is facing a genuine issue of absenteeism of the collection, transportation, delivery staffs. The company needs to overcome the following questions through the data provided by them :

1. What Changes company should bring to reduce the Cause of absenteeism?

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

absenteeism continues?

The aim of this project is to provide answers regarding above problems using machine learning techniques and data analysis.

* 1. DATA : The data provided by the company to solve the issue consists 21 attributes and 740 datapoints/observations.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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** |
| **0** | 11 | 26 | 7 | 3 | 1 | 289 | 36 | 13 | 33 |
| **1** | 36 | 0 | 7 | 3 | 1 | 118 | 13 | 18 | 50 |
| **2** | 3 | 23 | 7 | 4 | 1 | 179 | 51 | 18 | 38 |
| **3** | 7 | 7 | 7 | 5 | 1 | 279 | 5 | 14 | 39 |
| **4** | 11 | 23 | 7 | 5 | 1 | 289 | 36 | 13 | 33 |

Table 1.1 A brief view of the dataset with only six attributes.

**OVERVIEW OF THE DATA** : The given dataset has a total of 21 attributes in which, one will become the target variable and the rest 20 will become the predictor variable. The dataset has the following Attributes/columns :

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 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)

As the target variable (“Absenteeism time in hours”) is labeled and continuous(numeric) in nature, this problem can be stated as supervised problem of regression.

RangeIndex: 740 entries, 0 to 739

Data columns (total 21 columns):

ID 740 non-null int64

Reason for absence 737 non-null float64

Month of absence 739 non-null float64

Day of the week 740 non-null int64

Seasons 740 non-null int64

Transportation expense 733 non-null float64

Distance from Residence to Work 737 non-null float64

Service time 737 non-null float64

Age 737 non-null float64

Work load Average/day 730 non-null object

Hit target 734 non-null float64

Disciplinary failure 734 non-null float64

Education 730 non-null float64

Son 734 non-null float64

Social drinker 737 non-null float64

Social smoker 736 non-null float64

Pet 738 non-null float64

Weight 739 non-null float64

Height 726 non-null float64

Body mass index 709 non-null float64

Absenteeism time in hours 718 non-null float64

dtypes: float64(17), int64(3), object(1)

memory usage: 121.5+ KB

As we can see from the above table that our data has 740 rows and 21 columns having 20 numeric variables(int64 and float64) and 1 categorical variable(Object).

**Chapter 2**

**Methodology**

**2.1 Data Pre Processing :** Data pre-processing is the technique of preparing the data in a proper format for further analysis and modeling. Mainly pre-processing of the data comprises of the following processes :

1. Cleaning the data. e.g. Missing Value Analysis, Data Manipulation, etc.

2. Changing the variables in a proper format. e.g. Conversion of the variable into proper data type.

3. Merging and Sorting of the datasets.

2.1.1. **Missing Value Analysis** : We will try to look at the data if it consists of any unfilled observations. If it is, We will try to fill them up with suitable values.