Absenteeism Prediction

Kartik Kumar
dept. of Computer Science and Engineering
National Institute Of Technology, Delhi
181210026@nitdelhi.ac.in

Abstract—

The dictionary definition of absenteeism is 'the practice of regularly staying away from work or school without good reason'. Management is an array of different concepts which involves specialization and specification of issues as well as their solutions, especially related to that of an employee. Various theories have been put forth (namely Maslow's hierarchy of needs, Herzberg's theory, etc.) to understand the growing needs of motivation and retention. Despite this intensive research, most organizations face the problem of employees remaining absent from work, popularly known as absenteeism. This paper lays emphasis on the authenticity and genuine reasons of an employee to stay away from work. Hence, absenteeism is calculated on mathematical grounds as well as study based on questionnaire has been carried out in order to find out the reasons pertaining to increase in absenteeism in recent times. Research has been carried out on a leading pre-publishing service providing company in Indore. Pertaining to the privacy policy of the company, its name has been concealed and is referred to as 'Y company'.

Introduction

Absence is the failure of worker to report for work when he is scheduled to the work. A work is to be treated as absent for the purpose of this absenteeism statistics even when he does not turn up for a week after obtaining prior permission. K.G. Desai classified absenteeism in to two types viz, authorized absenteeism and unauthorized absenteeism. Authorized absenteeism is permitted absenteeism i.e., taking leave prior permission of an employer. Unauthorized absenteeism means taking leave without prior permission of an employer. Absence of worker on account of strike or lockout or layoff i.e., involuntary absent is not considered as absence for the purpose of absenteeism study. Absenteeism rate is the percentage of man days lost due to voluntary absent (both authorized and unauthorized) to the corresponding total man days schedule to work. It can be expressed as under:

Man days lost (both authorized and unauthorized)

--- X 100

Man days scheduled to work

Absenteeism is of two types –

1. Innocent absenteeism - Is one in which the employee is absent from work due to genuine cause or reason. It may be due to his illness or personal family problem or any other real reason

2. Culpable Absenteeism - is one in which a person is absent from work without any genuine reason or cause. He may be pretending to be ill or just wanted a holiday and stay at home. Many employees will, on occasions, need a few days off work because of illness, however, when absences become more frequent or long term and reach an unacceptable level, action by management is necessary. Absence from work can be expensive in both monetary and human terms.

The costs incurred when an employee is absent from work may include:

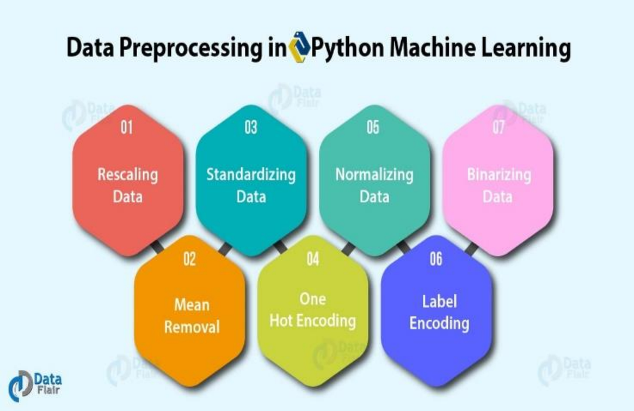
- Replacing the employee or requiring other staff to cover the absence
- Inability to provide services, or achieve section and departmental objectives
- Low morale and general dissatisfaction from other staff, particularly if the absence is perceived as unwarranted

DATA USED

The datasets used for this project were drawn from Kaggle The training dataset has about 700 rows of data from various articles on the internet. We had to do quite a bit of pre-processing of the data which includes both categorical as well as numerical data , a full training dataset has the following attributes:

- Reason for Absence
- Date
- Transportation Expenses
- Distance to work
- Age
- Daily work load average
- Body Mass Index
- Education
- Children
- Pets
- Absenteeism in Hours

DATA PREPROCESING



The dataset needs to under some processing as to make the dataset useful and easy to get a though what features are important and what are not.

This is the half part of the project:-

- One hot encoding the nominal variable reason for absence
- Grouping of these categories to boost the model
- Dealing with date and extract day of the week and month and remove the redundant part year
- Converting education into categorical value
- Converting the absenteeism in hours to categorical as our goal is to predict absence of employee

MODELS

Since our problem is to predict whether the employee will be absent in future or not, understanding the crux of problem this is a multivariate binary classification problem.

Logistic Regression

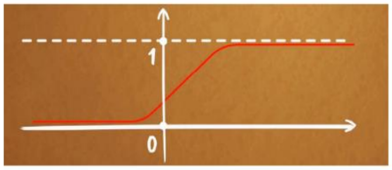
- A regression technique in which possible outcomes are categorical rather than numerical
- It predicts the probability of occurring an event

Linear regression model:

$$Y=\beta 0+\beta 1X1+\ldots+\beta kXk+\varepsilon$$

Logistic regression model:

$$p(X) = \frac{e\,(\beta 0+\beta 1X1+\ldots+\beta kXk\,)}{1+e\,(\beta 0+\beta 1X1+\ldots+\beta kXk\,)}$$



Visual representation of a logistic function

The logistic regression model is not very useful in itself. The right-hand side of the model is an exponent which is very computationally inefficient and generally hard to grasp.

Logit regression model

When we talk about a ‘logistic regression’ what we usually mean is ‘logit’ regression – a variation of the model where we have taken the log of both sides

$log (pX/(1-pX)) =log(e (\beta0+\beta1x+\cdots\beta kxk))$

$log (pX/(1-pX)) = \beta0 + \beta1x + \cdots \beta kxk$

$log (odds) = \boldsymbol{\beta0} + \boldsymbol{\beta1x} + \cdots \boldsymbol{\beta kxk}$

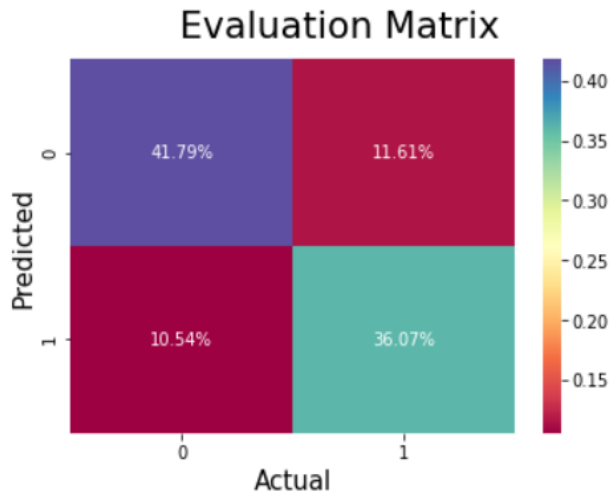
Evaluation Matrix and Accuracy

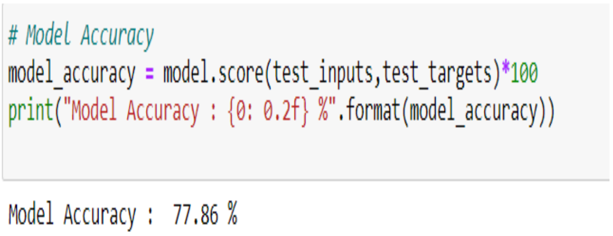
To evaluate the performance of classification problem, confusion metrics have been the best .

- **True Positive (TP):** when predicted fake news pieces are actually annotated as fake news
- **True Negative (TN):** when predicted true news pieces are actually annotated as true news
- **False Negative (FN):** when predicted true news pieces are actually annotated as fake news
- **False Positive (FP):** when predicted fake news pieces are actually annotated as true news

By formulating this as a classification problem, we can define following metrics,

$$\begin{aligned}
Precision &= \frac{|TP|}{|TP|+|FP|} \\
Recall &= \frac{|TP|}{|TP|+|FN|} \\
F1 &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \\
Accuracy &= \frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}
\end{aligned}$$



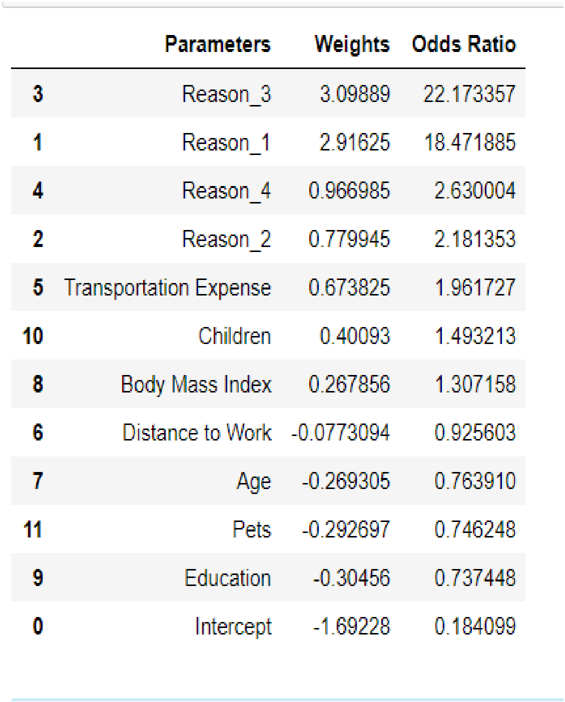


```
# Model Accuracy
model_accuracy = model.score(test_inputs,test_targets)*100
print("Model Accuracy : {0: 0.2f} %".format(model_accuracy))
```

```
Model Accuracy :  77.86 %
```

Summary Table

In this we will see summary table for attributes and corresponding odds ratios to see which are the most promising explanatory variables



	Parameters	Weights	Odds Ratio
3	Reason_3	3.09889	22.173357
1	Reason_1	2.91625	18.471885
4	Reason_4	0.966985	2.630004
2	Reason_2	0.779945	2.181353
5	Transportation Expense	0.673825	1.961727
10	Children	0.40093	1.493213
8	Body Mass Index	0.267856	1.307158
6	Distance to Work	-0.0773094	0.925603
7	Age	-0.269305	0.763910
11	Pets	-0.292697	0.746248
9	Education	-0.30456	0.737448
0	Intercept	-1.69228	0.184099

Intuitions

In this we will observe the behaviour of attributes and their strength of explanation of dataset.

- Intercept is just to reduce the error terms and increase accuracy , no any kind of interpretability is there for intercept

- Reason for Absence is the largest explanatory variable. For ex- If a person has given a Reason_3 , odds of absence increases 22 times nearly

- Month Value,Day of the Week and other parameters whose weight is nearly zero has no effect on model so dropping those columns simplify our model

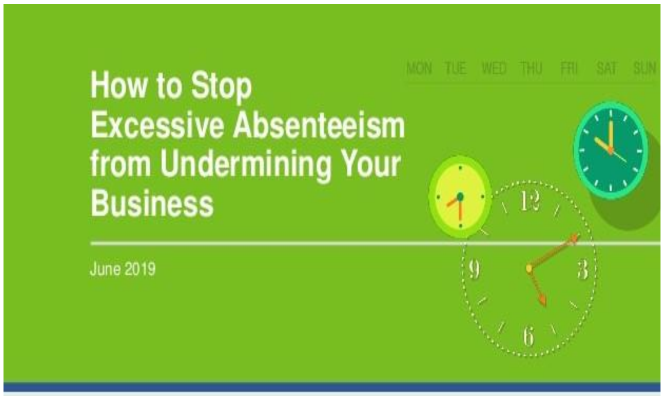
- For **Children** the odds of absence increases when a standarized unit of children increases(Can be interpreted as It is more likely to absent when number of child increases, since they are absent more likely to take care of their children)

- For **Pets**,Education and other parameters with negative weights , odds of absence decreases(Can be interpreted as It is less likely to absent when Pets increases, since pets play among each others and stay healthy or there is someone to take care

Conclusions

Since every company is different, it will require various levels of analyses to identify the factors that impact absenteeism for a specific employer. If absenteeism is identified as a significant problem, the company will need to take a hard look at the cause of the problem and begin to consider strategies to recapture lost revenues. Furthermore, as the economy tightens and the related financial stress increases for most employees, it is very likely that employers may see an increase in absenteeism due to stress related issues. The more aware a company is of issues related to employee absenteeism, the more successful they will be in implementing strategies to reduce the related cost and increase productivity. One study cited evidence that options to work from home, reduced workweeks and standard weekday work hours were helpful in reducing absenteeism. Shift work and compressed work schedules, however, led to increased absenteeism. Other studies show that allowing workers to have more input into decisions that affect their jobs and increasing their responsibilities, when appropriate, makes jobs more interesting with improved productivity. Thus, job satisfaction was shown to be an important factor in decreased absences.

Measures to reduce absenteeism


How to Stop Excessive Absenteeism from Undermining Your Business
June 2019
MON TUE WED THU FRI SAT SUN
12
9
3
6

- *DISCIPLINARY PROCEEDINGS/ABSENCE MANAGEMENT PROGRAM*
- *CREATION OF POSITIVE COMPANY CULTURE*
- *CHILDCARE AND FLEXIBLE SCHEDULING*
- *INCENTIVES*
- *HIRING EMPLOYEES FOR FUTURE SCOPE, WHO ARE INTERESTED AND SATISFIED WITH THE JOB*

References

- *https://www.researchgate.net/publication/301796227_Absenteeism_Problems_And_Costs_Causes_Effects_And_Cures*
- *https://www.Kaggle.com*
- *https://www.isbr.in/journals/33-41-f-1.pdf*