Absenteeism Time from Work

Marlyne Hakizimana

Introduction

Every work place can't help but have workers who, depending on various reasons, decide to be absent. Why is that? Is there a more recurring reason across various subjects?
According to Forbes, Absenteeism is an employee's intentional or habitual absence from work where excessive absences can equate to decreased productivity and can have a major effect on company finances, morale and other factors. There are different types of reasons that can lead to an absence at work such as Illness, Childcare, etc.

This project will be looking at a dataset from Brazil where records of absenteeism at work were collected from July 2007 to July 2010 at a courier company. The goal is to reduce worker absenteeism and evaluate the primary causes of absenteeism.
A few questions to consider along the way:

- Which areas of life affect Absenteeism(i.e: Work or Family..)?
- Is there an obvious relationship between reason for absence and absenteeism?

1. Data Exploration

This dataset consists of 740 observations and 21 feature**s** with:

- **8 Categorical features**: Reason for absence,Month of Absence, Day of the week, Seasons, Disciplinary failure,Education,Social drinker and Social smoker.
- **12 numerical features**: ID, Transportation Expense , Distance from Residence to Work,Service time, Age , Work load , Hit target , Son, Pet, Weight, Height, Body mass index.
- **Response variable:** Absenteeism time in hours.

The following is the full information on all features::

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
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

1.1 Pre Processing

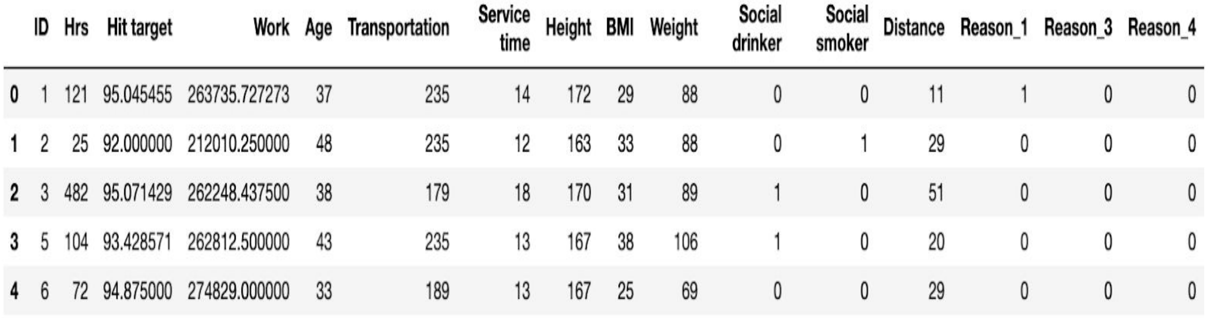
In order to build a good model, some understanding of the dataset is needed for better results. First, a general analysis with descriptive statistics is made in order to capture the shape or tendency of each column which results in a few observation points:

- The maximum value of Absenteeism time in hours is **9 standard deviations away from the mean!**
- The minimum values for **Reason for absence** and **Month of absence** are **zero!** These are considered unexpected values since 0 Month does not make sense.
- The **ID** column has 36 for the maximum value meaning that all these 740 observations have repeated observations among workers.

A closer look of **the Absenteeism time in hours and Reason for absence** columns reveals that all 0's in both **Absenteeism time in hours,Month of absence and Reason for absence columns** correspond to **Absenteeism time in hours=0**. If this conclusion were to represent every employee who didn't take any absent days, the Day of the week column shouldn't be populated. This is where I decide to categorize both 44 rows as outliers, a step that ends up filtering out all information on Disciplinary failure=1, another column within the dataset with binary information. This observation lets us know that the column **Discipline failure** won't be useful for the analysis.

Knowing that we have 33 unique ID's, the allabsents dataframe is rearranged from 696 rows to 33 rows by combining each repeated observation to each ID. In order to do it:

- The Reason for absence categorical values become 28 different columns where each one captures how many times a unique ID has been used a specific Reason for absence. The same process is used for Month of Absence, Seasons and Day of the week columns.
- The mean is extracted for all numerical columns except for the Absenteeism time in hours column
- The Absenteeism time in hours column becomes the sum of all missed hours for each ID.

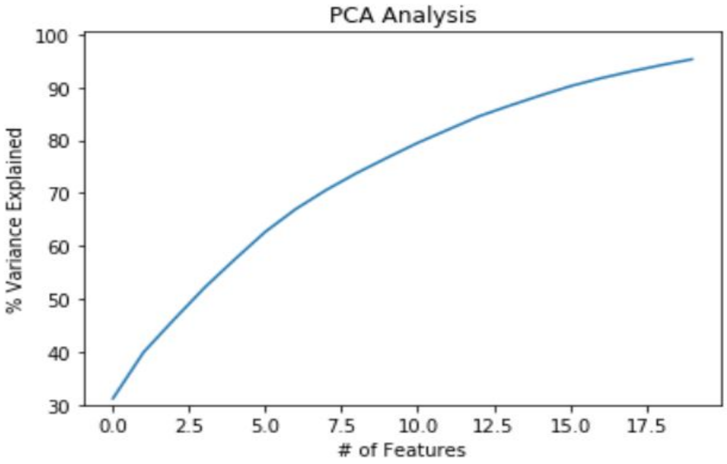


	ID	Hrs	Hit target	Work	Age	Transportation	Service time	Height	BMI	Weight	Social drinker	Social smoker	Distance	Reason_1	Reason_3	Reason_4
0	1	121	95.045455	263735.727273	37	235	14	172	29	88	0	0	11	1	0	0
1	2	25	92.000000	212010.250000	48	235	12	163	33	88	0	1	29	0	0	0
2	3	482	95.071429	262248.437500	38	179	18	170	31	89	1	0	51	0	0	0
3	5	104	93.428571	262812.500000	43	235	13	167	38	106	1	0	20	0	0	0
4	6	72	94.875000	274829.000000	33	189	13	167	25	69	0	0	29	0	0	0

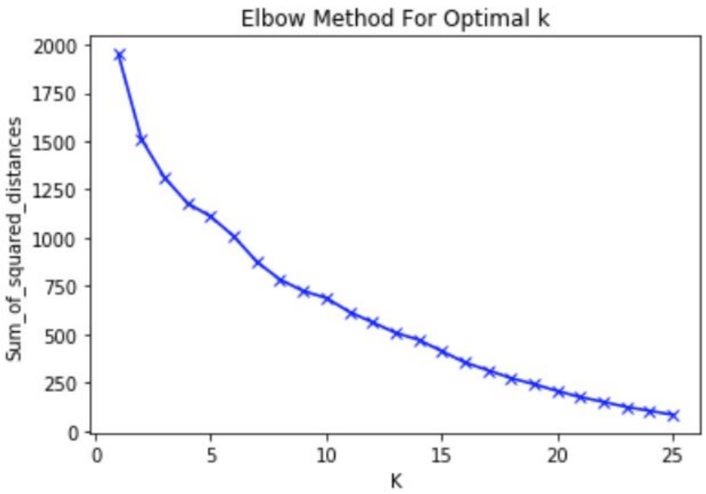
1.1.0 Clustering

With 33 unique representations, Clustering and PCA are best used in order to determine any insights between our employees.

PCA, otherwise known as **Principal Component Analysis**, is a statistical procedure that uses an orthogonal transformation to convert a set of observations into linearly uncorrelated variables called principal components. In our case, there exists 62 columns, using PCA before clustering ensures the reduction of variables and the use of uncorrelated components. The proportion of variance explained method is used for each feature where it looks at how many principal components will be needed to explain 95% of the total variance in the model. The figure below shows that 20 principal components explain 95% of the total variance.



After that, we proceed to the K-means Clustering method itself by first determining how many K-means will be needed for dividing our data. The **Elbow Method** is used on the newly transformed dataset with 20 principal components to determine the optimal number of clusters with k-means ranging from 0 to 23.



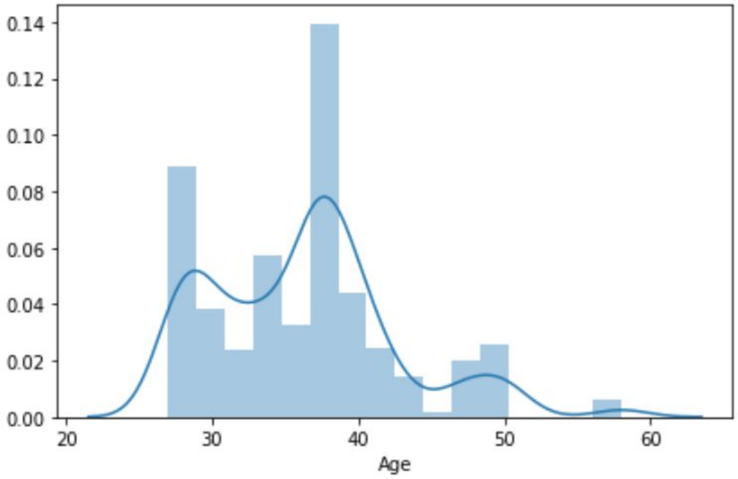
We can see that, from this above result, we would need to use between 15 to 20 clusters. Since we only have 33 customer IDs, we can see how the clustering method will not be able to find distinct groups employees.

Even though, this method is inconclusive by looking at each unique ID, having more data with more employees would have given us an in depth insight. Rather than concluding my analysis, I proceeded to assume that our original 696 rows represent 696 unique employee ID's instead of the 33. Thanks to this assumption, other prediction tools are used in this analysis.

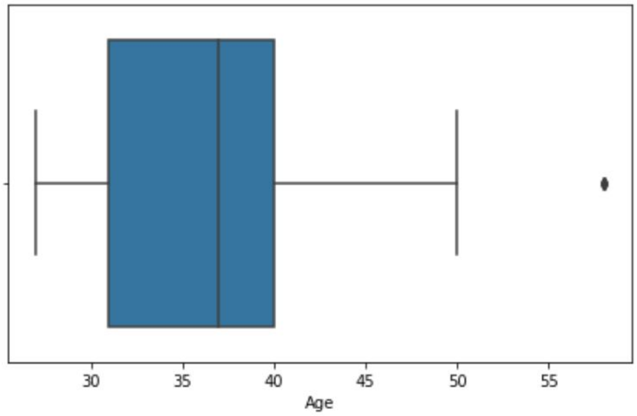
With this new path,I decide to a closer look at a few columns.

1.1.1 Age

For this column, I am trying to gather the type of employees who work at this company, is there a more younger generation or is there a range?

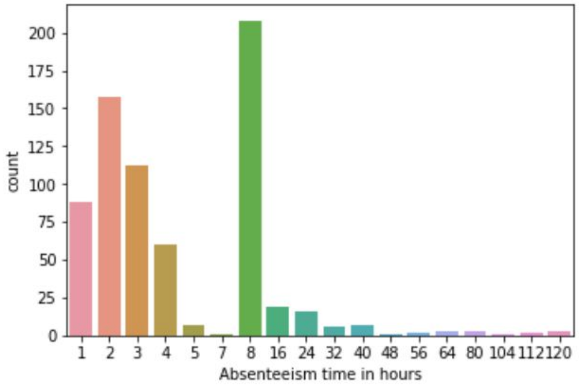


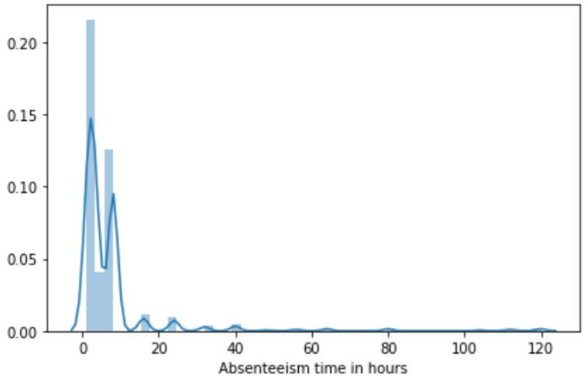
Most workers are in between their late 20's and early 40's.



Since the outlier age is higher than 55 yrs, it will be taken out for a better analysis.

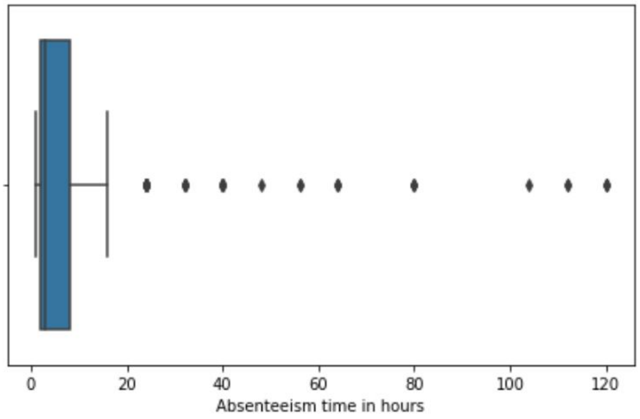
1.1.2 Absenteeism time in hours(response variable)





The above plots show us how the data is more skewed to the right with **most values less than 80 hrs of absence**. In fact, **61% of employees have been absent for less than 8hrs and 29% of them absent for exactly 8hrs**. Hence, **91% of Absenteeism time in hours is explained by only 8 or less hours of absenteeism**.

They also tell us that these values are not continuous but rather **categorical** since most employees missed specific hours. This insight helps us determine the type of modelling we will be using: **CLASSIFICATION**



The boxplot shows the overall data range with most outliers being **16hrs or more of absenteeism**.

1.1.3 Reason for absence

Another column of importance is **Reason for absence** with its 28 levels where each number represents a reason given by an employee for their absence. The column originally had categorical input that was converted into numerical input.

Most common reasons are the categories outside of the ICD (International Code of Diseases) meaning that the most reasons were **non serious diseases** such as :

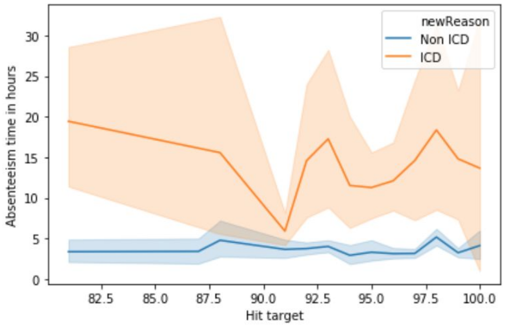
- patient follow-up (22),
- medical consultation (23),
- blood donation (24),
- laboratory examination (25),
- unjustified absence (26),
- physiotherapy (27),
- dental consultation (28).

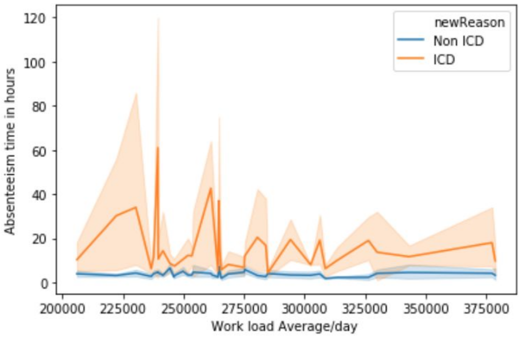
These 7 categories are used by **62% of employees** meaning that the ICD diseases are represented by **28% of employees**.
For the ICD diseases or **serious disease**, the main 2 categories with relatively high numbers are

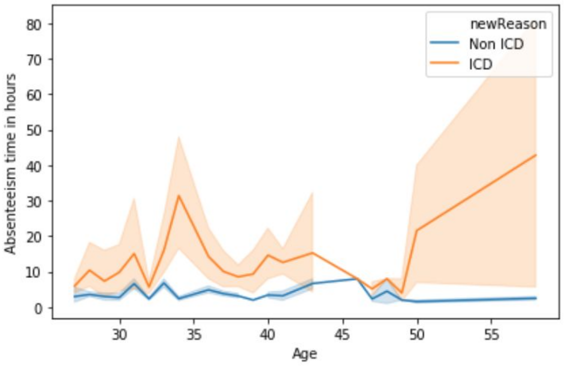
- **13:Diseases of the musculoskeletal system and connective tissue**
- **19:Injury, poisoning and certain other consequences of external causes**.

These two seem like they could be related to injuries one gets from this type of industry. For example, someone who is always doing deliveries for a long period of time is more likely to have musculoskeletal issues. It is surprising how only a few absences are related to birthing or children related.

With this information, I wanted to analyze how depending on the reason, insights can be found from other features :Is there a specific category of reasons with less absenteeism?
I decided to create **two categories of Reason for absence** : ICD and non ICD categories.







The above plots show how both Age, Hit target and Work load Average/day have high absenteeism time in hours for ICD group and low absenteeism time in hours for non ICD group.

2. Model Evaluation

After exploration, I wanted to have a general idea of the statistical significance of all columns. In order to so, a few changes are made:

- Outliers from Age, Service time and response variable Absenteeism time in hours were removed.
- Qualitative variable are converted into binary information using pandas.get_dummies

2.1 Classification

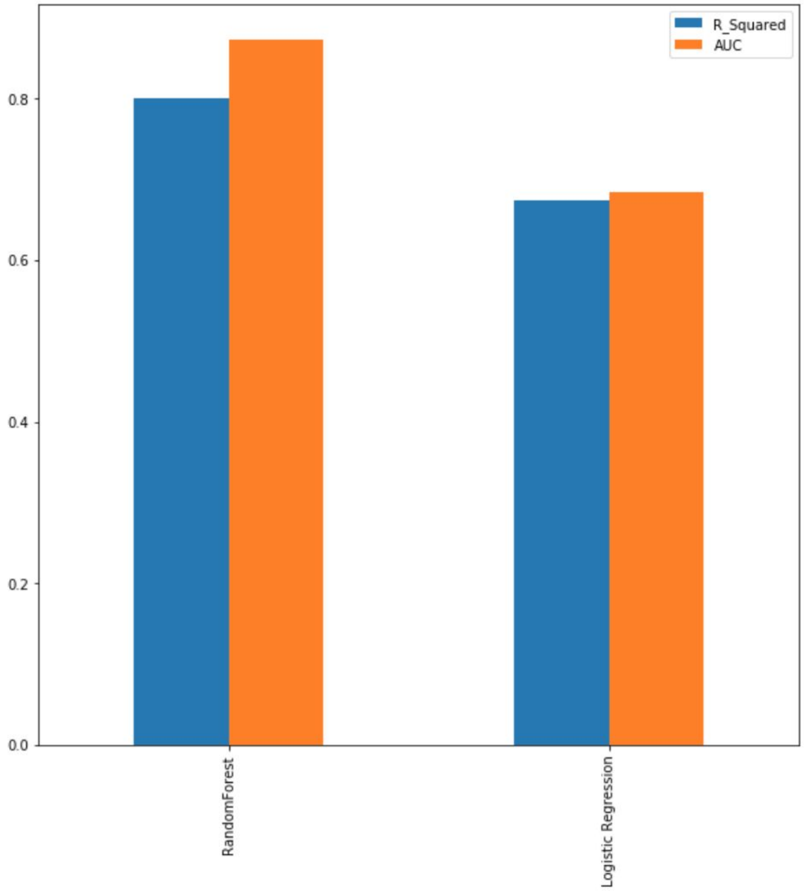
I decided to binarize our response variable to predict if an employee was absent for less than a day (<8hrs) or not (<=8hrs).
In other words, our response variable is given:

- 0: for absence hrs less than 8 hours
- 1:for absence hours more or equal to 8 hours

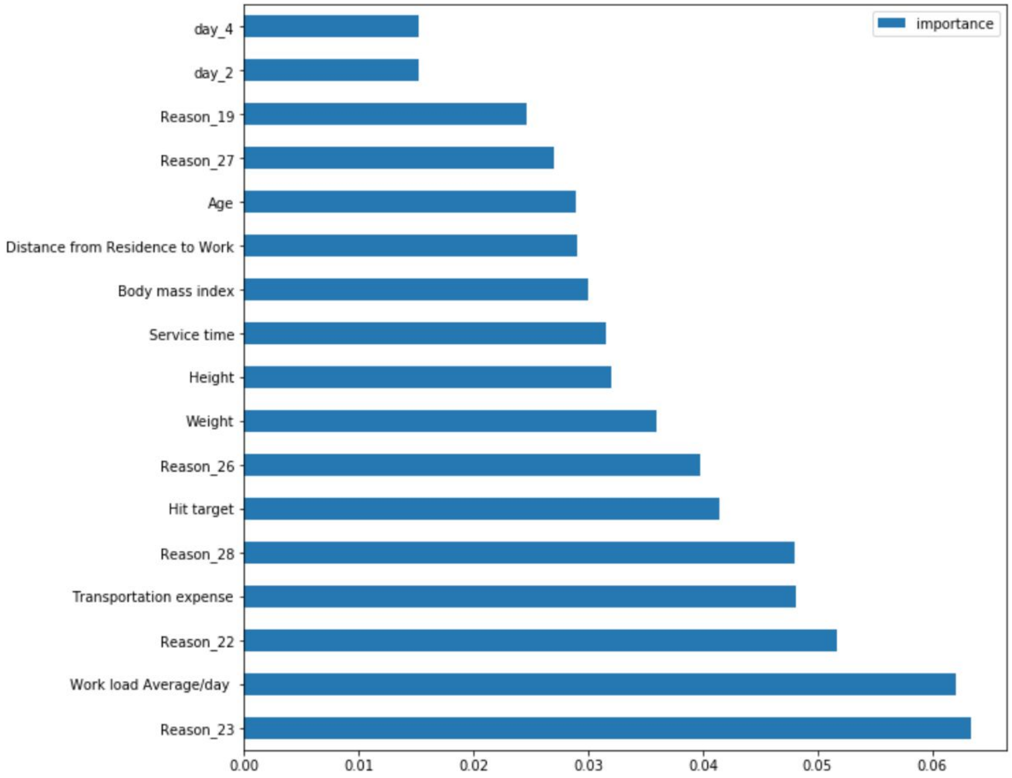
Classification models:

- **Random Forest Classifier:** It is a meta-estimator that fits a number of decision trees on various samples of the dataset and uses average to improve the predictive accuracy of the model.
- **Logistic Regression**: With a binary response variable, a logistic function is ideal since it is used to model probabilities describing the possible outcomes.

For both models, RMSE, R^2 and AUC score were used as evaluation metrics.



Random Forest performs the best with 78% accuracy score and a quick look at which features performed better showed that **Work Load Average/day, Transport and Doctor consultations** were the biggest predictors.



3. CONCLUSION

This project's goal was to find ways of reducing workers absenteeism time and it was accomplished by two types of modeling:Unsupervised and Supervised. The first part of the project used PCA and Clustering by accounting for the uniqueness of the Employee IDs and expected to find insights between different types of clusters. However, due to the inconclusiveness of it and since the data only had 33 observations, a second part of the project was created. By assuming that each row represents a unique observations, two classification models: Random forest and Logistic regression were used to predict whether or not an employee was absent for less that 8 hours.

Based on the model performance metrics, the recommended model was the Random Forest Classifier with an R-square value of 0.82. Even if there is no an easy way to see how a Random Forest works, due to its use of a lot of decision trees, it is able to calculate the best path while still accounting outliers or variance.

3.1 Applicability

Given that the main reasons of absenteeism are Work related reasons such as Workload, Hit target and doctor consultations, there are a few suggestions that could help reduce absenteeism at the workplace:

- Flexible schedule. Since medical consultations are high, giving an option for a flexible schedule where the few hours lost can be compensated by either coming in early or leaving late, depending on the employee's role.
- Employee Wellness program. For example, the amount of Work load can lead to absenteeism due to stress. Relaxing activities during lunch breaks targeting specific muscles for a delivery employee not only reduce stress, potential health problems but also increase employee morale.

Note:
Here are a few variations made to the models that did not make it to the final analysis:

- Categorized Reason for absence into 2 classes, where the first one had ICD reasons and the second one were non ICD ones
- Categorized Reason for absence into putting similar categories together such as respiratory and digestive.

In both cases, the model performance decreased instead of increasing .