APPLICATION OF MACHINE LEARNING IN ABSENTEEISM ANALYSIS

A REPORT

By

Damaraparapu Vinayak Kumar

Doddi Yamini Durga Pradeep

Pemmaraju Srikari

Yogesh Venkatesan

Hariharan N

Submitted in partial fulfillment of the requirements for

UROP Project

In Bachelor of Technology

SRM University AP

2018 - 2022

SRM UNIVERSITY AP

Title

APPLICATION OF MACHINE LEARNING IN ABSENTEEISM ANALYSIS

By

Damaraparapu Vinayak Kumar

Doddi Yamini Durga Pradeep

Pemmaraju Srikari

Yogesh Venkatesan

Hariharan N

The UROP Committee certifies that this report complies with the regulations and meets the standard required for completion of
UROP Project

Report Advisor: *Dr. Sunil Chinnadurai*

Report Co-Advisor:

Committee Member:

Committee Member:

Date
14-05-2021

Head of the Department
Dr. Siva Sankar Yellampalli

Table of Contents

List of Tables

List of Figures

1. Abstract

This paper is a research to analyze absenteeism behavior in employees, identify various reasons and estimate how much each reason is contributing to absenteeism. We further discuss how this analysis can be used by the HR department in taking preventive measures to stop loss of, and to increase, productivity in organizations. Supervised learning is used to create Machine Learning Models which can be used to predict who may resort to absenteeism in the future. A dataset from UCI repository is used in this paper.

2. Introduction and Background

To do a prediction one needs to know the state of the object and the laws that govern it. This is true for all the physical objects. The more variables and the more complex the laws become, the more difficult it gets to predict the behavior of the object. Humans' brain is also made of the same physical matter and that matter is also governed by the same laws. But it gets complicated when the question arises whether the mind is a product of the brain or whether the brain is just a house for mind. This philosophical question can never be solved. And while predicting human behavior, there are huge numbers of variables that influence a person's mind. And we also don't know all the laws that govern the human mind. But with the advent of machine learning, this becomes a bit less complex. Because an intelligent machine doesn't need to know the laws that govern the data we provide it. These laws are represented as formulas in simple computations and as algorithms in complex ones. An intelligent computer can figure out the algorithms by itself once we provide correct data. Also, the more data we provide, the more accuracy the prediction gets. This is the difference between normal machine computations and machine learning or artificial intelligence. More or less, the machine develops an intuition towards what is happening in the data. And with the help of machine learning we have also discovered that humans are more predictable than we think. In future, this same machine learning can help us answer some of the world's most fundamental philosophical dilemmas and questions. But for now let us apply this in predicting absenteeism behavior in employees using machine learning.

People form organizations so that we can work more efficiently together towards a common goal. In this, we depend on each other by cooperating and doing team work. By being organized, we are becoming productive. And increasing this productivity is an important goal in organizations and companies.

When people take time away from work, sometimes it is planned and sometimes it is unplanned and unexpected. The latter one damages productivity on a significant scale. First, it affects individual productivity. And then as employees work mostly in groups, it also affects team productivity and in turn the value the company is providing to the marketplace. This also reduces the profit margins of the company.

The dataset we use contains a variety of data from the past absent days, health issues to some subtle factors like the educational qualification and whether the employee has a pet and other external factors. It is publicly available on Kaggle.

3. Scope of Work

This problem can be solved using the technology of Digital Analytics and Machine Learning. This project considers various reasons that are contributing to absenteeism and extracts a relationship between general information about the employee, the reasons for absence and the absenteeism behavior. By having these calculated insights into the contributing reasons of absenteeism Human Resource teams can take necessary action to take care of it. It can either by helping individuals or by making necessary changes in the work environment. We can also get to know if an employee might resort to absenteeism in the future using regression algorithms.

We used supervised learning in this project. In supervised learning, the machine learning algorithm takes both the inputs and corresponding outputs from the dataset and maps them together to form a relation or an equation or simply a formula to predict the outputs for the inputs that it hasn't yet seen. We will be using three different regressions to form machine learning models and then compare the accuracy, which are linear regression, ridge regression and lasso regression. We also perform parameter tuning, feature selection to improve accuracy and consistency over all used regressions.

The future scope of this work includes using the same methods on data from different fields of different types of organizations. In some organizations, we can take corrective measures and that may solve the problem in hand. But this same method can also be used by some organizations, which require the necessary performance of employees within the existing environment due to some particular reasons, to take the decision whether to hire a person or to change the role of an employee depending on their ability. An example for this type of application is in the Military. The work of the Military is so crucial and critical. A small mistake or error of any single person can lead to the sabotage of the mission and the security of a nation. That is why they already follow very strict protocols in selecting people. And unlike other companies, the environment cannot be changed much here. So if it can be predicted that a person cannot perform efficiently in a critical or stressful environment, it is good for all that person is not sent into that environment.

Thus the future development of this project helps in ergonomics - selecting right people for the right job.

4. Procedure and Methodology

This section includes the procedure of the project.

4.1 Dataset

The dataset we used was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil.

Data Description

ID: Individual identification (ID)

Reason for absence: 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).

- Month of absence - ranges from 1 to 12
- Day of the week (Saturday(0), Sunday(1), Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
- Seasons (summer (1), autumn (2), winter (3), spring (4))
- Transportation expense - the expense for a person to travel from his residence to the office
- Distance from Residence to Work (kilometers)
- Service time - the months a person has been working in the company
- Age
- Work load Average/day - work already done/doing now
- Hit target - work target given for near future
- Disciplinary failure (yes=1; no=0)
- Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
- Son (number of children)
- Social drinker (yes=1; no=0)
- Social smoker (yes=1; no=0)
- Pet (number of pet)
- Weight
- Height
- Body mass index
- Absenteeism time in hours (target)

Data type/structure

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 740 entries, 0 to 739
Data columns (total 21 columns):
ID                                740 non-null int64
Reason for absence                740 non-null int64
Month of absence                  740 non-null int64
Day of the week                   740 non-null int64
Seasons                           740 non-null int64
Transportation expense            740 non-null int64
Distance from Residence to Work   740 non-null int64
Service time                      740 non-null int64
Age                               740 non-null int64
Work load Average/day             740 non-null int64
Hit target                        740 non-null int64
Disciplinary failure              740 non-null int64
Education                         740 non-null int64
Son                               740 non-null int64
Social drinker                    740 non-null int64
Social smoker                     740 non-null int64
Pet                               740 non-null int64
Weight                            740 non-null int64
Height                            740 non-null int64
Body mass index                   740 non-null int64
Absenteeism time in hours         740 non-null int64
dtypes: int64(21)
memory usage: 121.5 KB
```

Figure 4.1

4.2. Data Cleaning

- Missing Values - If missing values are allowed in the data frame, it leads to inaccurate model building. Missing data makes the results invalid by reducing the power of statistical analysis. It can lead to the conclusion of null hypothesis when it is not or the other way around. It can also cause bias in estimation of parameters. Using the Pandas library, we created a data frame for the taken dataset. Then the data frame is checked for missing values or otherwise called NULL values. If any missing values are found, we can either delete the rows containing the missing values or we can impute values for variables using different imputation algorithms. It is found that the dataset does not contain any missing values and so no need to take any action

regarding this.

- Outlier Analysis - This is done on continuous variables. Outliers are the values that deviate too much from the mean value and they may result in incorrect analysis. This may be caused due to experimental errors, variability in measurement or novelty. These values deviate the mean of the data too much towards their value. This creates a false perception about the data. If there is a very large value, the mean of that data can be greater than the majority of values and leads to wrong conclusions that all the values are generally higher than what they really are.

Outlier analysis is performed only on continuous variables because there are high or low in categorical variables.

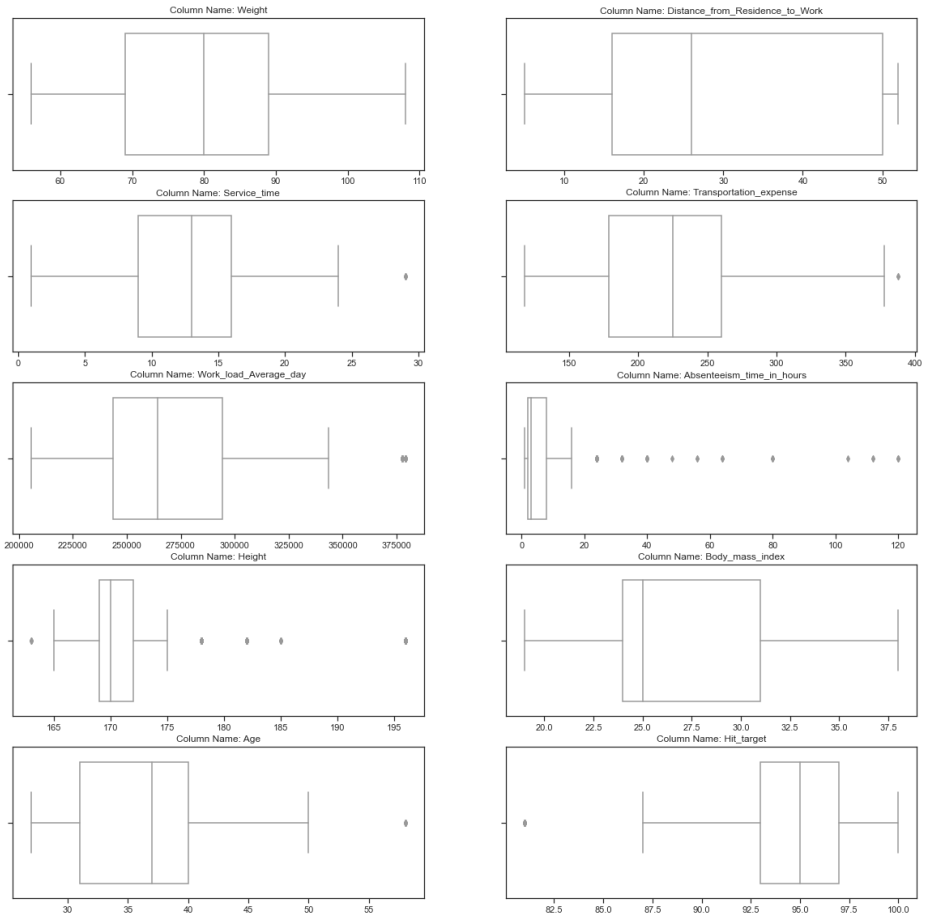

Figure 4.2

4.3. Data Preprocessing

Data Preprocessing is the process of transforming the given data into usable or meaningful format for the model. There are many tasks in data preprocessing and they are used according to the type of data and purpose of the model.

In this project we have done the following preprocessing tasks along the way, wherever necessary.

When doing Exploratory analysis, it is not useful to perform it on continuous variables. So we first convert them into categories by using bins. We select a range of values of each continuous variables and give that range a name. Each range is a bin.

And while doing statistical analysis, we get to know the significance of some features and decide whether to keep them or drop them. This also comes under preprocessing.

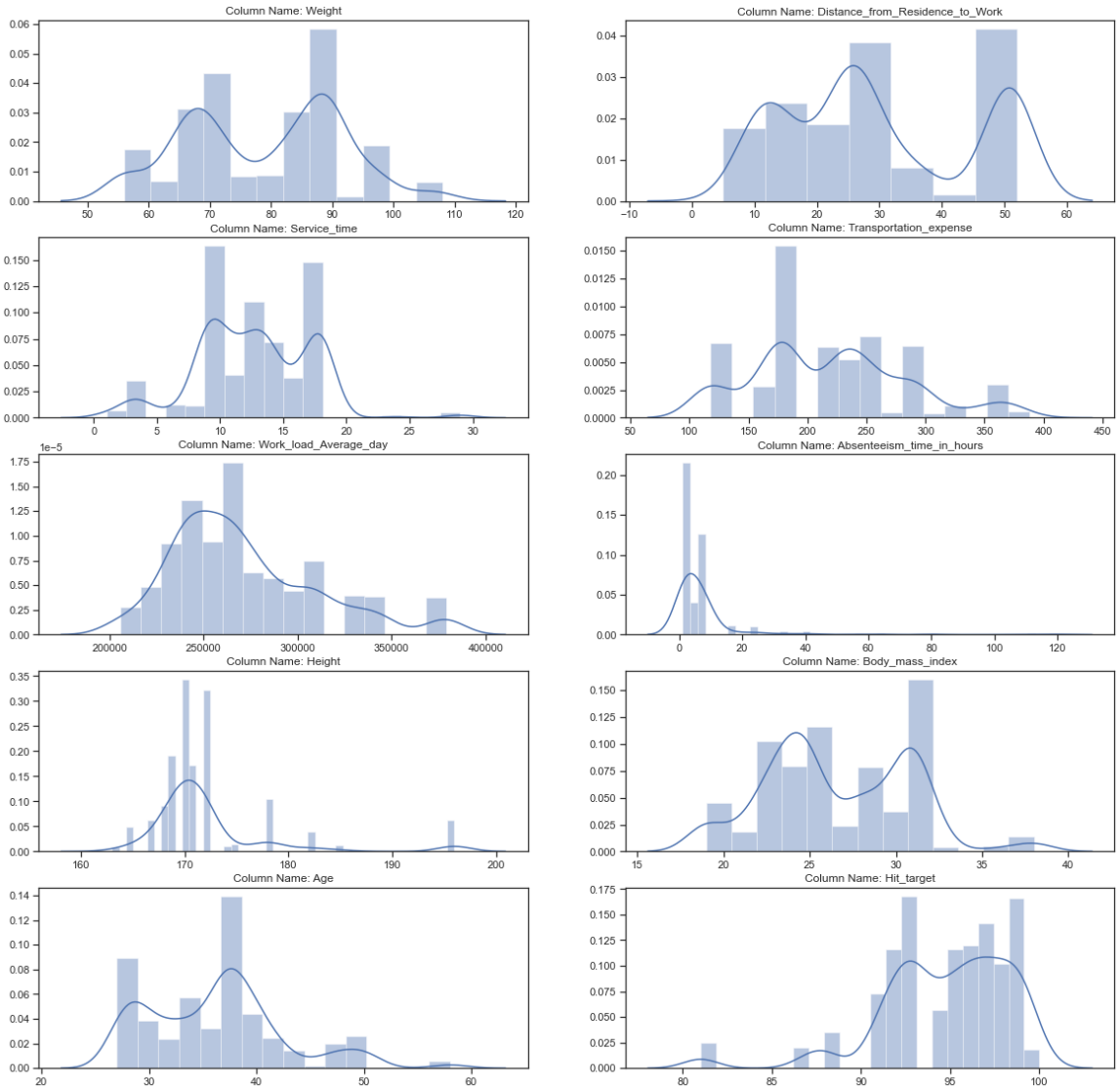

Figure 4.3

We have also observed that our data follows skewed distribution which means the data distribution is not symmetric. Such data generally requires normalization. Also we use normalization to get the range of all features to be equal. This sometimes ensures more accuracy for the model.

4.4. Statistical Analysis

Chi-squared test is used to do statistical analysis to test null hypothesis. Null hypothesis is true when two variables seem to have a relation but in reality they just happen to be so due to chance or randomness. So in the chi-square test, we prepare a contingency table which gives us the p-values. If this value between two variables is 0, they are perfectly related. If it is 1, they are completely unrelated. Generally we get 0 when between the same variables and the significance range is taken as 0 to 0.05.

This is done only on categorical variables.

We found that Disciplinary Failure has 1 chi-square value. So this can be removed.

4.5. Correlation Analysis

We do correlation analysis to see if any continuous variables are correlated and if any are correlated, we can drop a few variables. Here we used a heatmap from Seaborn library.

Dropping some of the correlated variables comes under preprocessing. Having correlated variables only creates noise. So we only need one variable and as this variable is correlated with some other variables, we are covering all of them while feeding the data into the model.

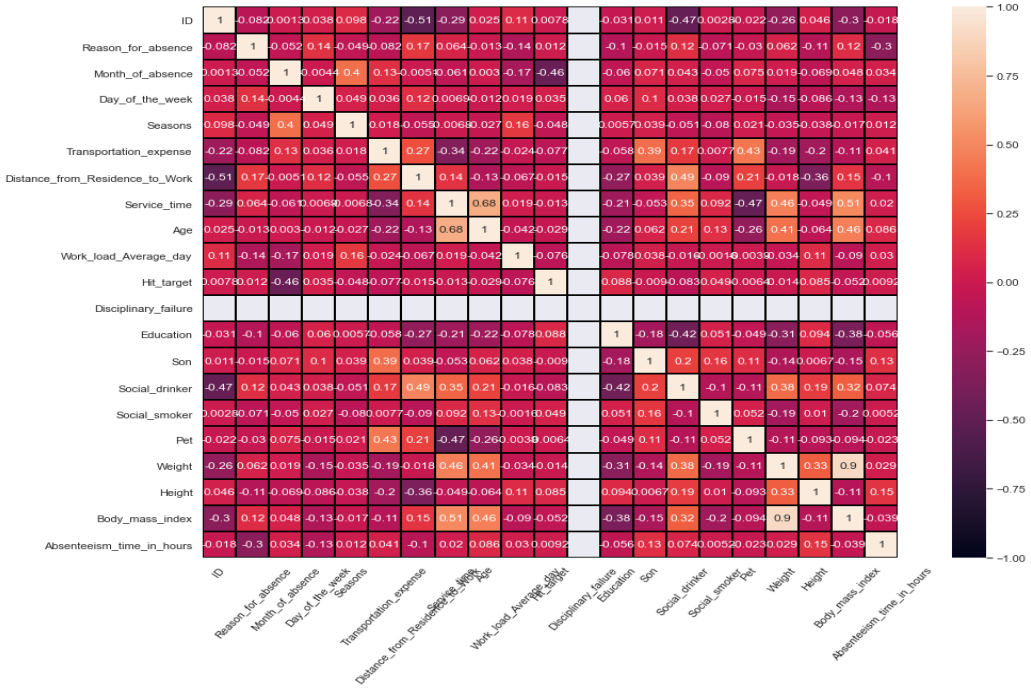

Figure 4.4

4.6. Exploratory Data Analysis

In Exploratory Data analysis, we explore the relationship between independent variables and a dependent variable. Here the dependent variable is also our target which is Absenteeism time in hours. Below is an example for this.

- Reason for Absence vs Absenteeism time in hours

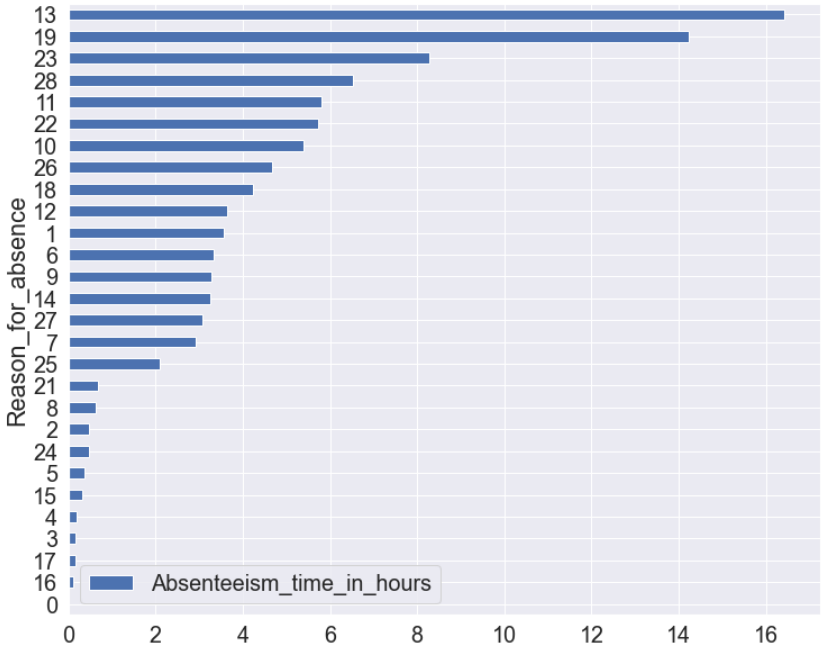

Figure 4.5

4.7. Feature selection

From the tasks done before like statistical analysis using a chi-squared test to check null hypothesis and correlation analysis, we will be selecting a few variables and dropping a few others. Based on p-value and correlation values these are the features that are selected to feed into the model.

The results and difference in accuracies from removing some features is discussed in the results section.

Removal of features based on r2 score

Before going into the model we will be checking the significance of features again based on the r2 score. For that we first apply a basic stats model using OLS regression. We applied OLS regression on the scaled data frame from standard scaler. And then checked the r2 score of predicted values against the test values.

And the r2 score is found to be 0.143. This serves as a base value to remove any features. If removal of any feature changes the r2 score then that feature is significant. If the score is decreased, our accuracy is increased because of that feature and we should not remove that feature. Else if the score is increased because of removal of a feature, then it suggests that this feature is contributing to less accurate results. But that does not mean that we should directly remove this feature. As our dataset is small, some features may lead to less r2 score but they may get important once we have a larger dataset. So the data analyst needs to study the importance of the feature before removing it.

If the feature is not changing this base value much, it suggests that this may not be significant and it can be removed. It doesn't affect the results in any way. If there are many features that are like this, then we may need to deal with them.

After feature selection from statistical analysis and correlation analysis, we have checked their significance based on change of r2 score and found that Age is not contributing much to change in r2 score. But this is not removed because it also did not decrease the accuracy of the model in prediction. It may lead to some noise in larger datasets but in this project we decided to keep it.

Removal of features based on p-value

In OLS regression summary, we get to know the p-values of all the features. P-value is the evidence against null hypothesis. The features with p-value less than 0.05 are considered significant. And all others are removed.

But in small datasets we need to be careful while removing features based on p-value. As small datasets are not large enough to represent the real world data, the p-values don't exactly match their significance in real life. So along with considering the p-values, human analysis and decision is also required.

Based on p-values, the features that give maximum r2 score are as follows

```
cols_four = ['Reason_for_absence', 'Day_of_the_week', 'Distance_from_Residence_to_Work',  'Age',\
             'Education', 'Son', 'Social_drinker', 'Social_smoker',  'Weight', 'Absenteeism_time_in_hours']
```

4.8. Scaling

We have seen that our data follows skewed distribution. This may sometimes result in less accurate results. Generally we do normalization to the data that follows skewed distribution. To normalize our data we are using minmaxscaler technique. This changes the scale of all values of the data to the range of 0 - 1.

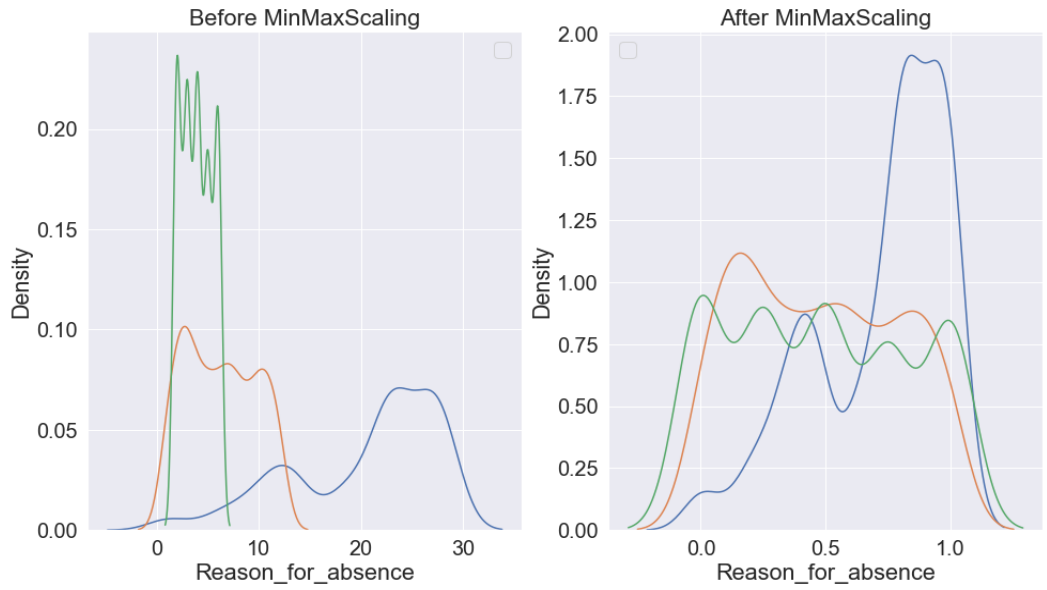

Figure 4.6

We check and compare the accuracy of these two scaling techniques and also without scaling. The results are discussed in the results section.

4.9. Splitting of data

After the model building is done, we need to find the accuracy of the models. But for that if we used the same data we feed into it, we may get higher accuracy that the real one. So we first split the data into train data and test data. We only feed train data into the model and then use the test data to find the accuracy of the model.

75% of the data is split into train data and 25% is used as test data.

4.10. Model Building

We have used three different algorithms to build machine learning model. We then compare the scores of these three models and try to improve the score by parameter tuning. We also try to get a consistent score along the three models.

The algorithms we used are

1. Linear Regression - This is used to find a relationship between independent or predictor variables and dependent or response variables. This algorithm does not intend to find a deterministic relation but finds a statistical relation. Deterministic relation is where you can find the response variable accurately using the predictor variable. Statistical relation deals with probability. How many times they have occurred together and tells us how many times they may occur together in the future. For example, the formula to convert grams into pounds is a deterministic relation. But the relation between height and weight is a statistical relation.
2. Ridge Regression - Ridge regression shrinks the data towards populated areas. This eliminates the bias which is caused by outliers.
3. Lasso Regression - Lasso regression also used shrinkage but towards a center point like mean. This is useful then we have all data populated near the mean or when the outliers are minimal. So let us compare the score of Lasso regression also.

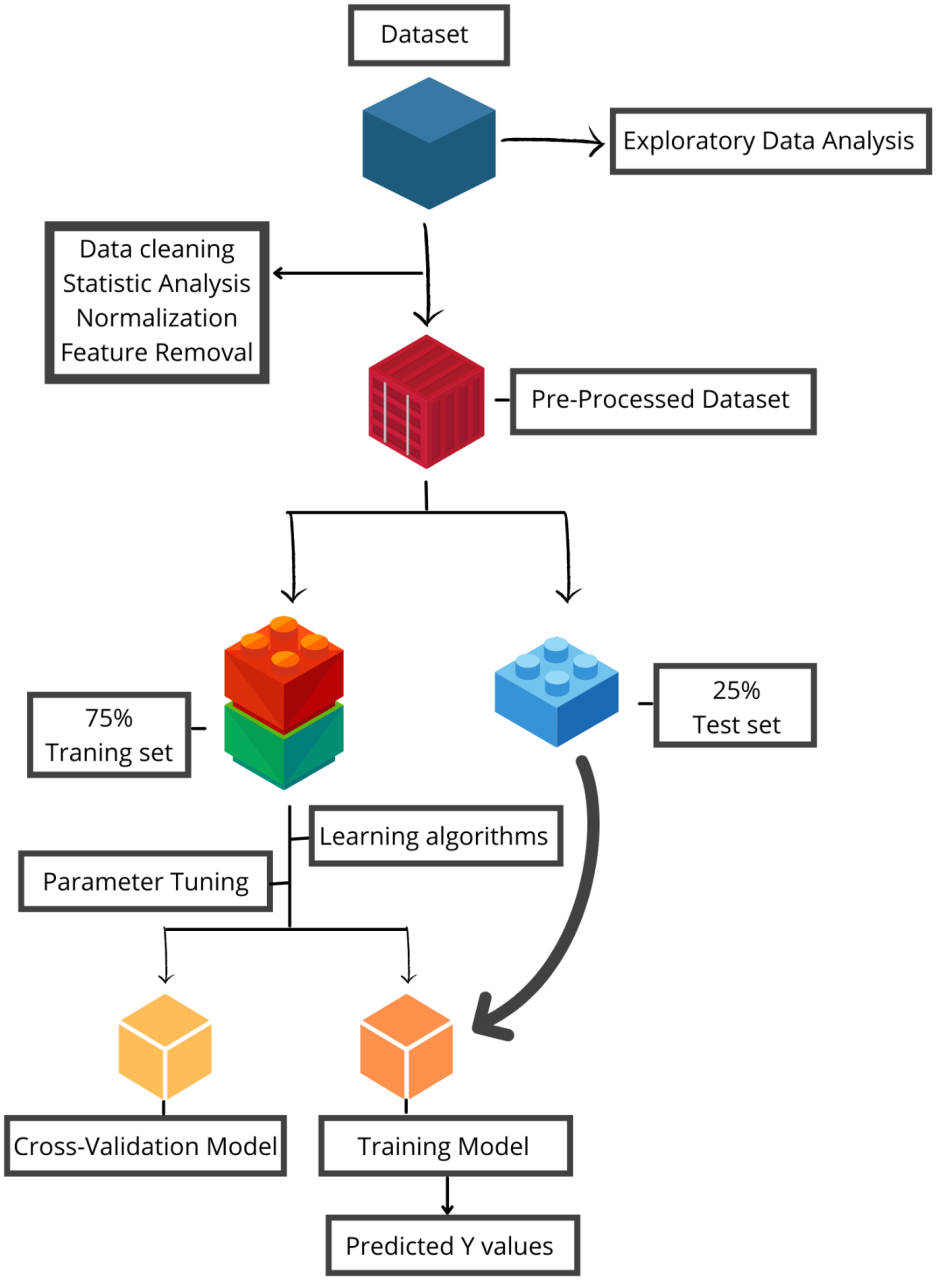
Dataset
Exploratory Data Analysis
Data cleaning
Statistic Analysis
Normalization
Feature Removal
Pre-Processed Dataset
75%
Traning set
25%
Test set
Learning algorithms
Parameter Tuning
Cross-Validation Model
Training Model
Predicted Y values

Figure 4.7

4. 11. Parameter Tuning

After the model is built we do parameter tuning to increase the r2 score or accuracy of the models. We automate this process by using GridSearchCV from sklearn library. It selects the best hyperparameters necessary to build a more accurate model.

4.12. Testing - Cross Validation

We have already split the data into training data and testing data. We can test our model accuracy on this test data. But to have more critical tests and to know the accuracy of real world data, we use cross validation. Cross Validation using k-fold approach. It splits the data into k-number of samples and then uses one sample as test data and all others as training data. It does the same with all k samples and gives us the average of all the accuracies. This also prevents having false higher accuracy due to the leakage of test data into models when normal split is done.

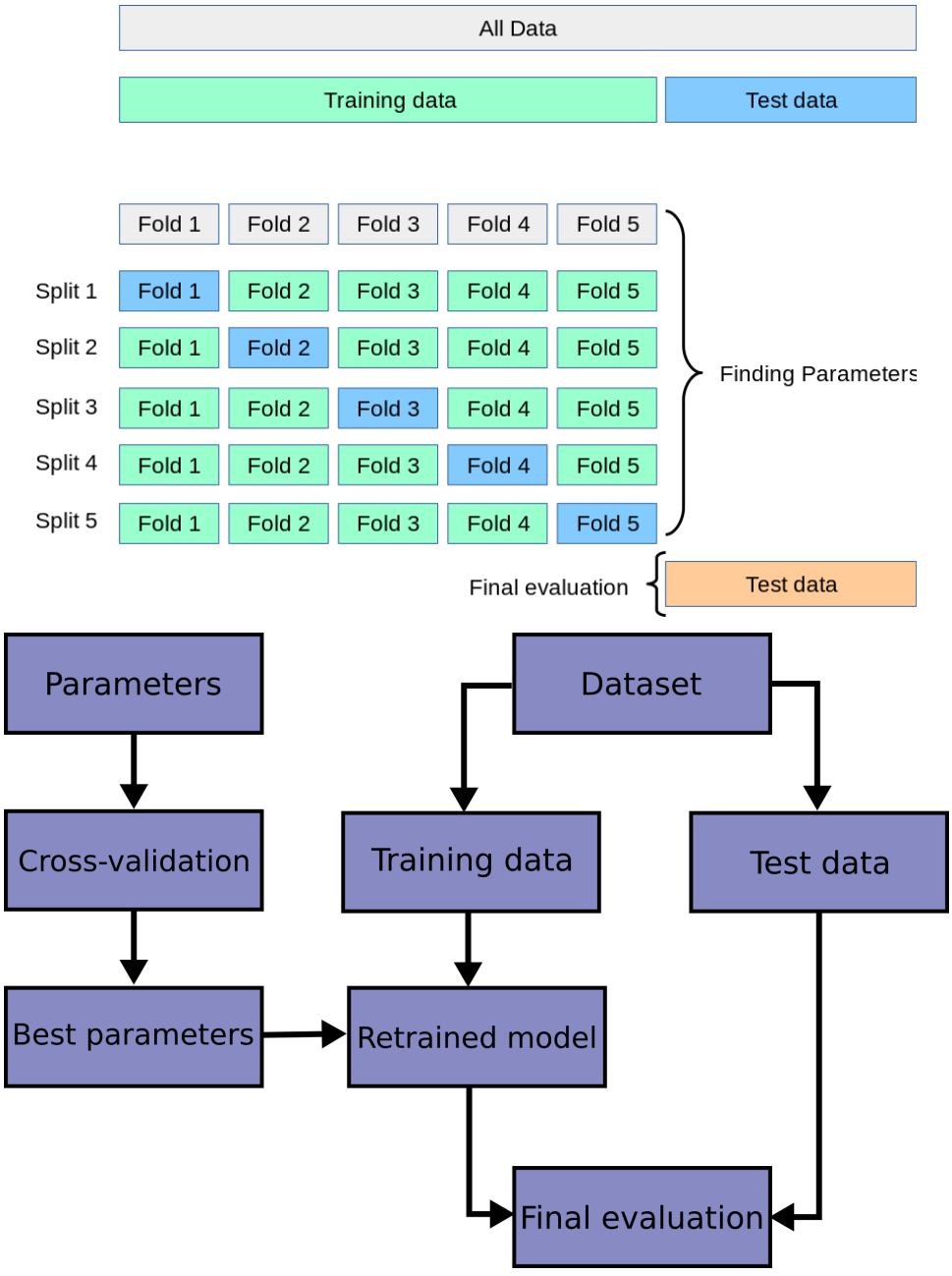

Figure 4.8

(Images from scikit-learn.org)

5. Results

5.1. Exploratory Data Analysis

Let us first see the results from Exploratory Data Analysis

In figure 4.5, we can see that most Reason number 13 is contributing to most of the absences. 13 is the code for Diseases of the musculoskeletal system and connective tissue.

Measures to be taken - It is possible that these are caused by bad working posture of the employees. This may be because of no-break work or bad work facilities. So companies should conduct a study on the workplace environment with this insight and get to know if it is indeed related to the company environment, it should take preventive measures like giving more frequent breaks, or improving the design of seating places, the ergonomics of workspaces etc. This can not only stop absenteeism rates but may also increase productivity of employees.

EDA of Day of the week vs Absenteeism

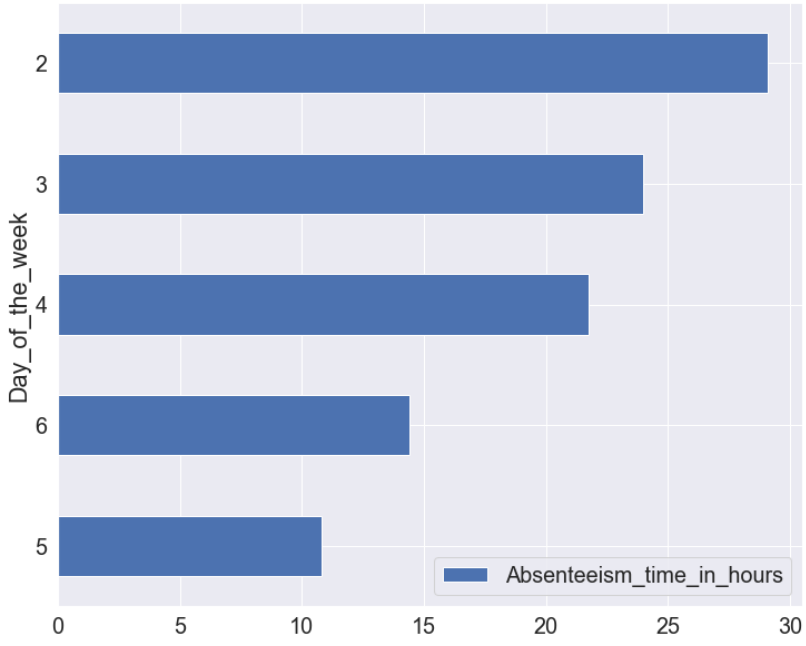

Figure 5.1

We can observe that many of the employees are absent on Mondays. The measure that we can take from this insight is to decrease the activity of the company. If employees are absent on this day, then it is better to not schedule any major work on this day. This can reduce the cost of lost productivity.

EDA of Social Drinker vs Absenteeism

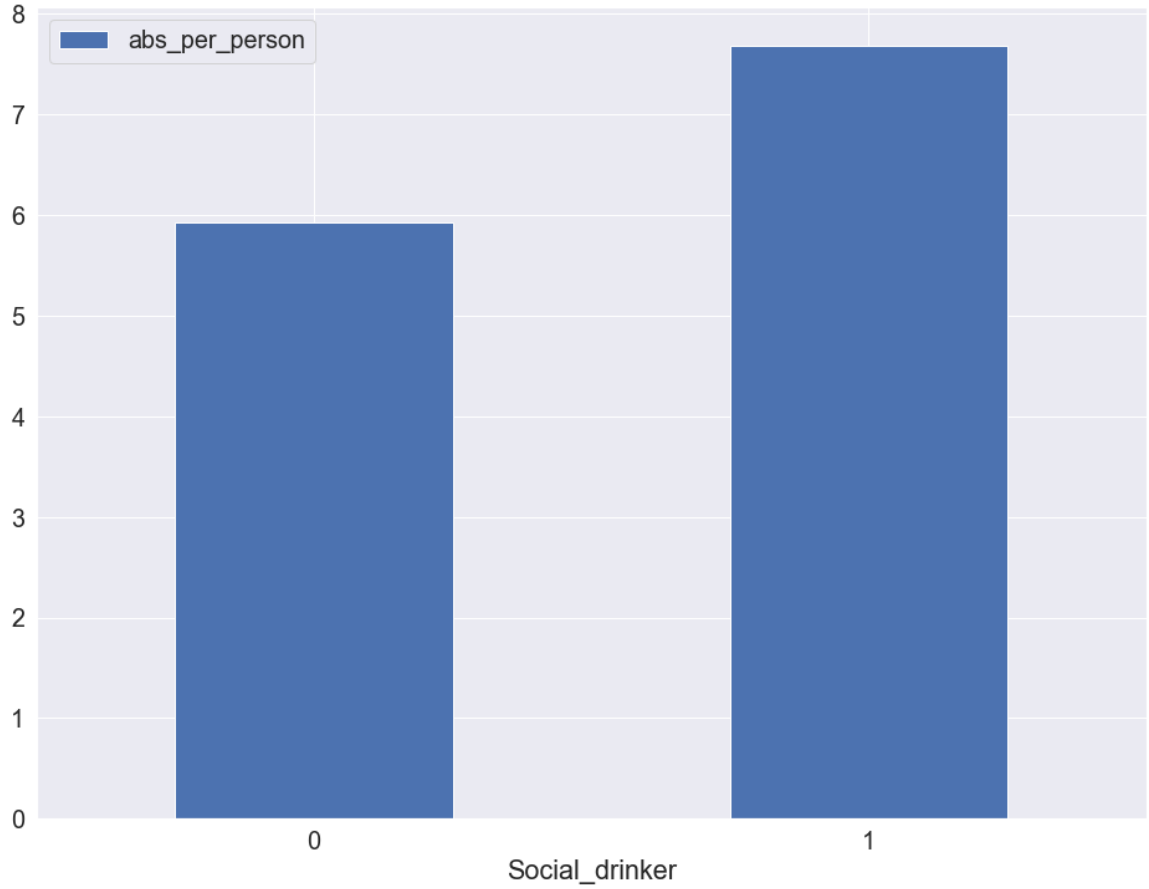

Figure 5.2

We can observe that employees who are social drinkers tend more towards absenteeism. But the difference is not significant enough to draw any concrete conclusion. So this should be taken care by domain experts to take any measures if necessary.

EDA of Son vs Absenteeism

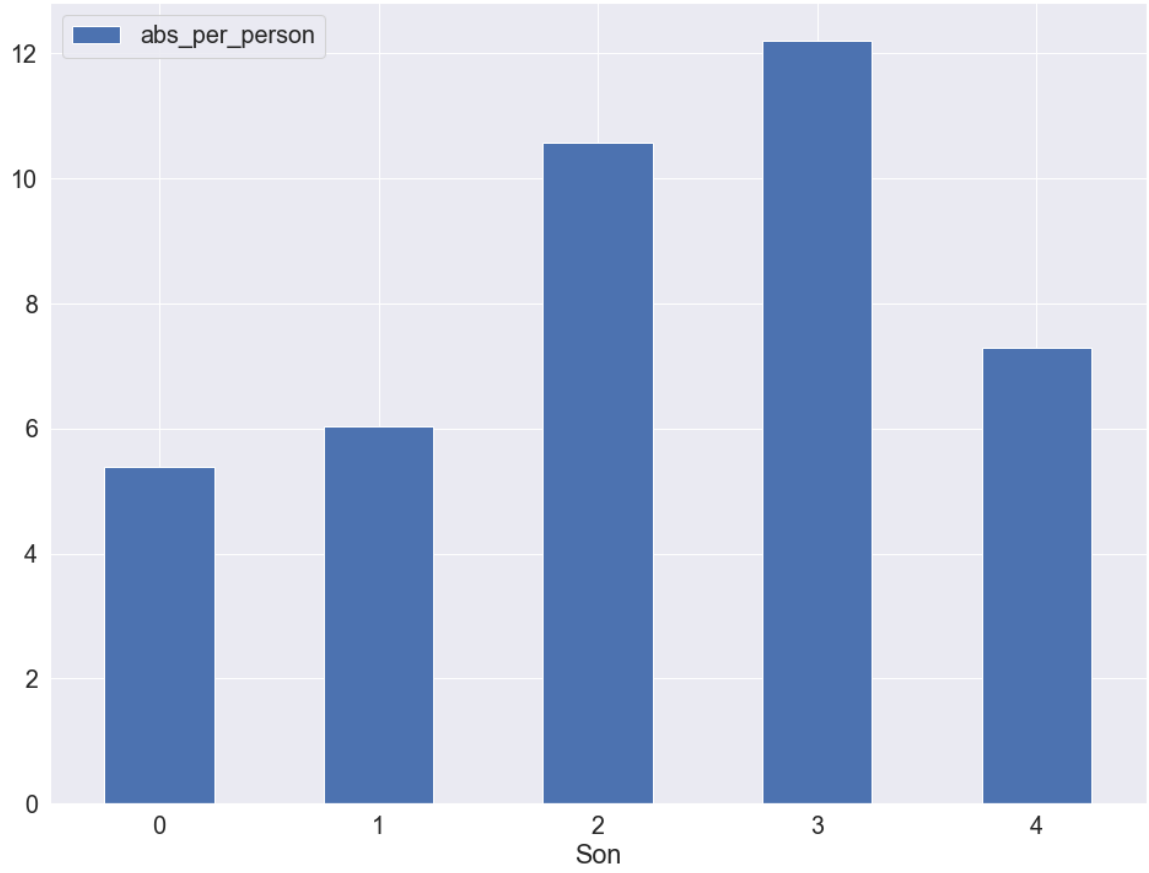

Figure 5.3

Here 'Son" means the number of children. So we can see the steady increase of absenteeism with the increase in children the employee has. This may be because of attending to their needs like attending their schools for some reason or to take them to hospitals when needed. These needs clearly increase with the number of children. The company may not be able to do anything to stop these absences but it can think of employee satisfaction by providing them with allowances or health policies related to the children.

EDA of Age vs Absenteeism

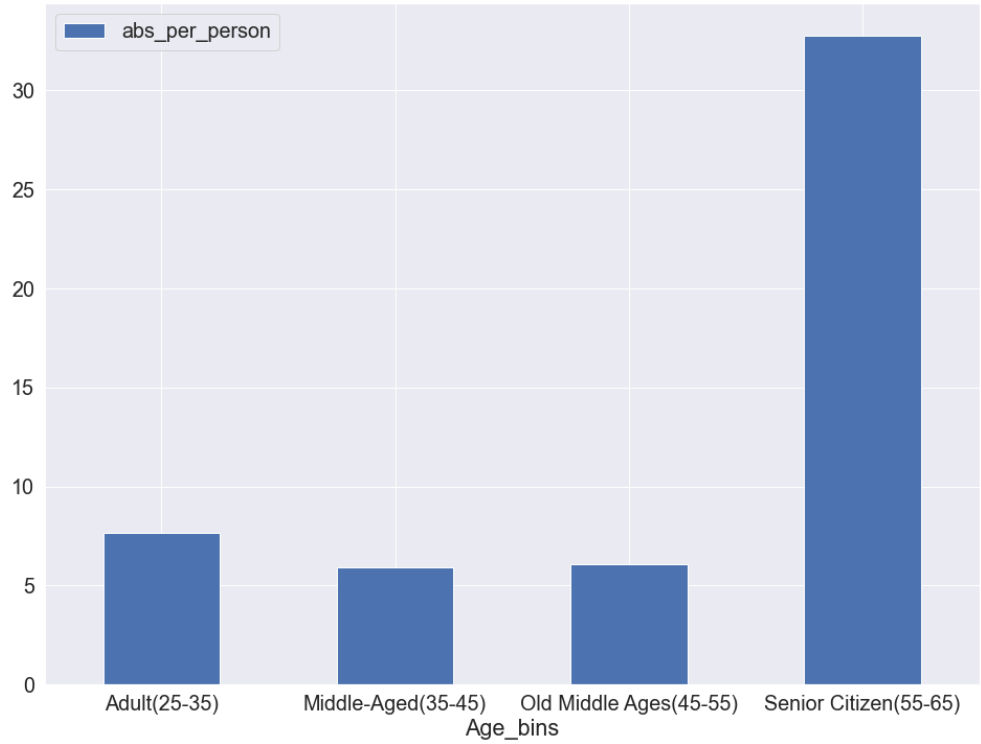

Figure 5.4

The number of Senior Citizens being absent is very high compared to other employees. This can most probably because of more health problems. So as a preventive measure the company may decide to give them more health benefits. But we can also see that there are only 8 senior citizens in our data. So we can also not come to any conclusion based on only 8 observations.

EDA on Hit Target vs Absenteeism

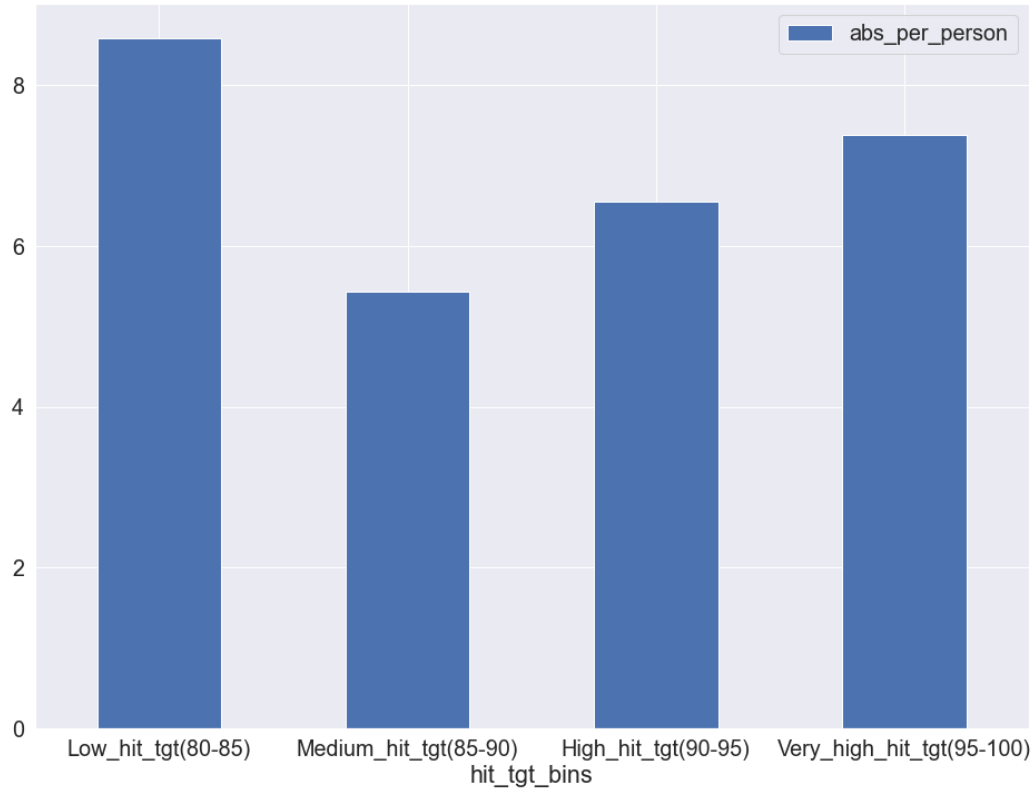

Figure 5.5

This is an interesting observation. People who are given least target and very high target are more tended towards absenteeism. If people are given very high targets, it may lead to stress and negative emotions towards work and this may explain their absenteeism behavior. And if the target is very low, people tend to lose motivation to work and they may procrastinate, So ideal targets would be both challenging as well as practical to achieve.

This can be a very helpful insight while assigning tasks for employees as this has potential to also increase productivity along with decreasing absenteeism.

6. Prediction and Accuracy

By passing the test data into the model, it can predict the hours a particular employee may resort to in the future. But the accuracy is not yet perfect. Let us first see the accuracy or the r2 score with all features considered.

- All features considered - before parameter tuning:

```
R2 Score of Linear Regression before parameter tuning:
0.05834173967520817
R2 Score of Ridge Regression before parameter tuning:
0.059532056095447206
R2 Score of Lasso Regression before parameter tuning:
0.08328698632358489
```

Table 6.1 (a)

- All features considered - after parameter tuning:

|:

metric_params	Linear_Reg	Ridge	Lasso
fit_time	0.005271	0.004293	0.004865
score_time	0.004200	0.003962	0.004382
test_r2	0.054458	0.066896	0.058447
train_r2	0.169229	0.166007	0.092324
test_neg_mean_absolute_error	-6.692392	-6.493442	-5.826414
train_neg_mean_absolute_error	-6.214577	-6.089751	-5.706764

Table 6.1 (b)

- Disciplinary Failure removed from chi-squared test (1st iteration)

Before Parameter Tuning

```
R2 Score of Linear Regression before parameter tuning:
0.058341739675209836
R2 Score of Ridge Regression before parameter tuning:
0.05953205609544743
R2 Score of Lasso Regression before parameter tuning:
0.08328698632358467
```

Table 6.2 (a)

After Parameter Tuning

metric_params	Linear_Reg	Ridge	Lasso
fit_time	0.005723	0.004972	0.004627
score_time	0.003994	0.004254	0.003476
test_r2	0.054458	0.066896	0.058447
train_r2	0.169229	0.166007	0.092324
test_neg_mean_absolute_error	-6.692392	-6.493442	-5.826414
train_neg_mean_absolute_error	-6.214577	-6.089751	-5.706764

Table 6.2 (b)

- Removing the following features (2nd iteration)

- Month of Absence, Seasons, Transportation Expense, Service time, pet, height

Before parameter tuning

```
R2 Score of Linear Regression before parameter tuning:
0.06632854032798619
R2 Score of Ridge Regression before parameter tuning:
0.06735053674376457
R2 Score of Lasso Regression before parameter tuning:
0.08881048546053005
```

Table 6.3 (a)

After Parameter Tuning

metric_params	Linear_Reg	Ridge	Lasso
fit_time	0.005332	0.004678	0.006006
score_time	0.004183	0.004007	0.003932
test_r2	0.080987	0.090572	0.085635
train_r2	0.164703	0.161785	0.154948
test_neg_mean_absolute_error	-6.569173	-6.387078	-6.265604
train_neg_mean_absolute_error	-6.183680	-6.074795	-5.972306

Table 6.3 (b)

- Removing Work load average, hit target and body mass index (3rd iteration)

Before Parameter tuning

```
R2 Score of Linear Regression before parameter tuning:
0.10252542186805658
R2 Score of Ridge Regression before parameter tuning:
0.10339996128368789
R2 Score of Lasso Regression before parameter tuning:
0.12259275051779084
```

Table 6.4 (a)

After Parameter Tuning

metric_params	Linear_Reg	Ridge	Lasso
fit_time	0.002793	0.003197	0.003398
score_time	0.002993	0.002204	0.003185
test_r2	0.080059	0.086477	0.067228
train_r2	0.146238	0.143264	0.082804
test_neg_mean_absolute_error	-6.519162	-6.345896	-5.853249
train_neg_mean_absolute_error	-6.217696	-6.102086	-5.796557

Table 6.4 (b)

In this, "test_r2" is the metric we use to calculate the near-real-world accuracy of the model. We can observe from the above values that removing some features increases the accuracy and removing some decreases the accuracy.

Out of all the observations the features removed till 2nd iteration are the best scores we got and we can conclude that Ridge Regression is the best among the three in accuracy scores. This is because of the nature of our data which contains a lot of outliers. Ridge Regression can handle this by shrinking the data values towards the populated regions instead of towards mean like Lasso regression.

And because of the same reason that a lot of data is outlier data it is appropriate to use Mean Absolute Error while considering the possible error range of the predictions. The values of absenteeism hours in the dataset range from 0 to 120. A mean absolute error of around 6 for this range is also practical and we can use this in real life applications.

7. References

1. M. Raman, N. Kaliappen and C. L. Suan, "A Study on Machine Learning Classifier Models in Analyzing Discipline of Individuals Based on Various Reasons Absenteeism from Work," 2020 International Conference on Decision Aid Sciences and Application (DASA),

Sakheer, Bahrain, 2020, pp. 360-364, doi: 10.1109/DASA51403.2020.9317017. Retrieved from https://www.ieeexplore.ieee.org
2. A. Rista, J. Ajdari and X. Zenuni, "Predicting and Analyzing Absenteeism at Workplace Using Machine Learning Algorithms," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2020, pp. 485-490, doi: 10.23919/MIPRO48935.2020.9245118. Retrieved from https://www.ieeexplore.ieee.org
3. M. Skorikov et al., "Prediction of Absenteeism at Work using Data Mining Techniques," 2020 5th International Conference on Information Technology Research (ICITR), Moratuwa, Sri Lanka, 2020, pp. 1-6, doi: 10.1109/ICITR51448.2020.9310913. Retrieved from https://www.ieeexplore.ieee.org
4. Gayathri, T. (2018). Data mining of Absentee data to increase productivity. International Journal of Engineering and Techniques, 4(3), 478–480. Retrieved from http://www.ijetjournal.org

Copyright Documentation

1. The images used in Figure 4.7 are taken from the website scikit-learn.org.