

**PREDICTING ABSENTEEISM OF EMPLOYEES AT WORKPLACE USING**

**LOGISTIC REGRESSION ALGORITHM**

By

SAHIL NAGPAL (0670042)  
PRAMINA PATIL(0663967)

A Report submitted in the partial fulfilment of the requirement for the degree of Master’s of Science in Big Data Analytics

Department of Computer Science

Trent University

Fall-2020

**ABSTRACT**

Workplace absenteeism plays a key role in demonstrating a company's efficient and profitable ability. Thus the understanding of employee absenteeism becomes the principle for a company across its different aspects. Since the right assessment of the profile of workers allows such morbidities to be detected by excesses of occurrences. The primary aim of the early absenteeism analysis was to predict the characteristics and groups of employee diseases that make them perform higher absenteeism at work. However it is still to be explored to forecast the absenteeism period of workers using machine learning classifiers and thus to find out the evidence that should be taken into account to minimize higher absenteeism at work. In this project, we have applied two prominent machine learning algorithms namely Logistic Regression and Gradient Boosting to predict absenteeism time of employees and to find out the insights that cause employees to perform higher absenteeism at work. Through this we have also find the probability of an employee of being absent at the work place based on evocative parameters. Meanwhile comparing the different machine learning algorithms to find out the best classifier which produces the highest prediction accuracy. We have used an existing dataset of a courier company in Brazil in order to predict the absenteeism time of employees. The dataset contains 21 categories of the reason for absence which are attested by the International Classification of Disease (ICD) and 7 other categories without the ICD that have proved to be effective in detecting the absenteeism at work. Similarly we have also crafted a threshold assessment of the extreme and normal absenteeism grounded on the hours absent at work place. Based on the seven evaluation metrics such as True Positive, True Negative, False Positive, False Negative, Sensitivity, Specificity, and Accuracy we have evaluated the model performance in predicting absenteeism at work. Using the machine learning model we discovered that the reason for absence class as diseases that are attested by International Code of Diseases (ICD), and the transportation expense from home to work are the topmost facts of performing higher absenteeism at workplace.

**Keywords**: Machine Learning, Absenteeism, Classification;

**INTRODUCTION OF THE PROJECT AND SUBJECT**

* **Motivation of the Research** - We all had the opportunity to work at several national and international organizations where we had different experiences in terms of efficiency, working climate, and place. All of those experiences filled us with a quest to learn the basic facts of workplace absenteeism. In addition, identifying the causes and trends of employee absenteeism becomes important in several dimensions for a company as the proper determination of the profile of workers enables the detection of excesses of such morbidity occurrences. And if it could be properly done, it would make it easier for businesses to increase efficiency and prestige.
* **Problem Statement -** The early research restricts the research by creating a neuro fuzzy network using only the Artificial Neural Network (ANN) algorithm to figure out the reason for absence, specifically the diseases that cause greater absenteeism. However, there are still plenty of unexplored alternatives, especially the use of more algorithms to not only predict absenteeism, but also the overall factors that should be carefully considered. For a inimitability we have calculated the probability of a person being not present along with the comparison of the algorithms.
* **Research Questions**

1. Estimate the precise probability of the absenteeism established on the convinced possible reasons extant in the data.
2. How well the machine learning algorithms accomplishes in the prediction of the absenteeism at workplace and which one achieves well ?

* **Research Scope -** The focus of this study is restricted to the workers of organizations in Brazil. The data on absenteeism used in this analysis is based on a courier company in Brazil. Since only a few pieces of research concentrated exclusively on forecasting Employee absenteeism has taken place over the past few years, there are still plenty of opportunities to experiment on different country's organizational data with different machine learning algorithms in line with enhancing the prediction performance.

**DATASET AND EXPERIMEMTAL PROCEDURE**

**Description of the Dataset**

The dataset we used in this study was obtained from the UCI Machine Learning Repository (Martiniano, 2018), which was created from July 2007 to July 2010 at a courier company in Brazil with records of employee absenteeism at work. First of all the Universidade Nove de Julho Postgraduate Program in Informatics and Information Management used the dataset for their academic study. The data is consisting of the 700 Rows with 12 Attributes namely ID, Reason for Absence, Date, Transportation Expense, Distance to Work, Age, Daily Work Load Average, Body Mass Index, Education, Children, Pets, Absenteeism Time in Hours. The International Code of Diseases (ICD) attests to the absence of diseases in 21 groups. The world standard diagnostic method for epidemiology, health care and clinical diagnosis is classified by International Classification of Diseases (ICD). The ICD of 21 Categories are listed below:

***Table-1.1 Reason for absence with ICD***

|  |  |
| --- | --- |
| **Code** | **ICD Description** |
| 1 | Certain infectious and parasitic diseases |
| 2 | Neoplasms |
| 3 | Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism. |
| 4 | Endocrine, nutritional and metabolic diseases |
| 5 | Mental and behavioral disorders |
| 6 | Diseases of the nervous system |
| 7 | Diseases of the eye and adnexa |
| 8 | Diseases of the ear and mastoid process |
| 9 | Diseases of the circulatory system |
| 10 | Diseases of the respiratory system |
| 11 | Diseases of the digestive system |
| 12 | Diseases of the skin and subcutaneous tissue |
| 13 | Diseases of the musculoskeletal system and connective tissue |
| 14 | Diseases of the genitourinary system |
| 15 | Pregnancy, childbirth and the puerperium |
| 16 | Certain conditions originating in the perinatal period |
| 17 | Congenital malformations, deformations and chromosomal abnormalities |
| 18 | Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified |
| 19 | Injury, poisoning and certain other consequences of external causes |
| 20 | External causes of morbidity and mortality |
| 21 | Factors influencing health status and contact with health services. |

***Table-1.1 Reason for absence without ICD***

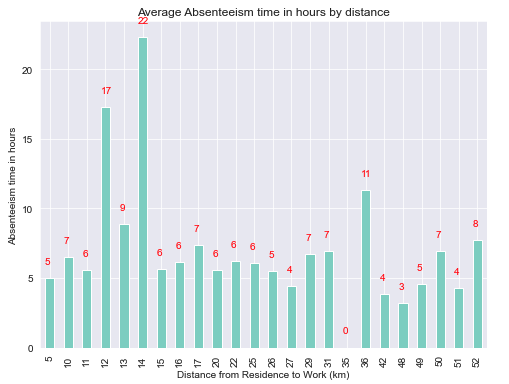
|  |  |
| --- | --- |
| **Code** | **Description** |
| 22 | Patient follow up |
| 23 | Medical consultation |
| 24 | Blood donation |
| 25 | Laboratory examination |
| 26 | Unjustified absence |
| 27 | Physiotherapy |
| 28 | Dental consultation |

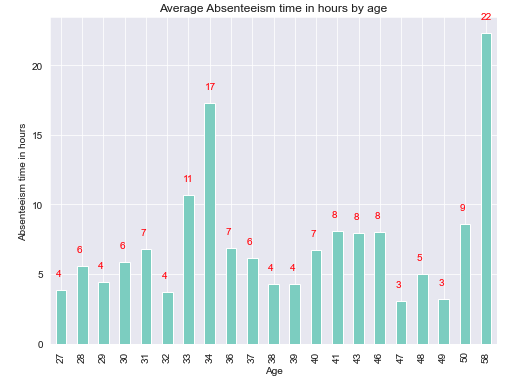
**Methodology**

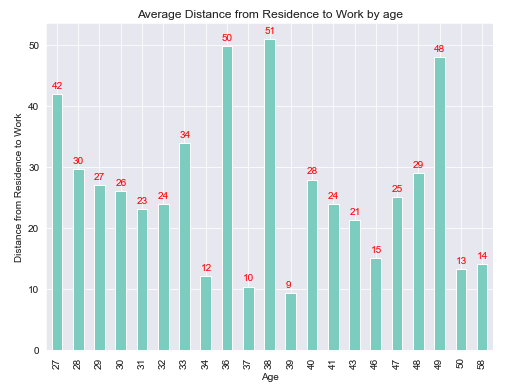
We need to clean up the data and translate it into the correct format before feeding the data into the model.

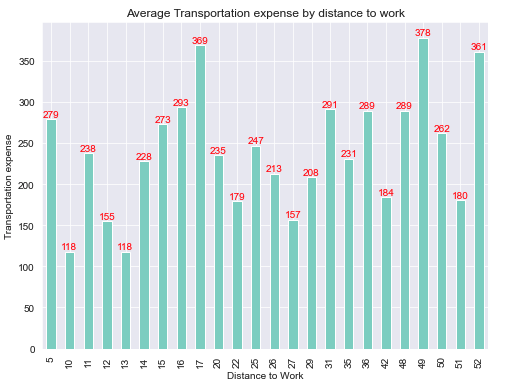
**Data Preprocessing**- The output of the machine learning model is completely depend on the data feed in the algorithm. The Machine Learning model does not generate the proper outcome on raw data.

* **Exploratory Data Analysis** – Firstly , we will always see the affiliation between the columns and make an appropriate visualizations which better depicts the regular outlines ,patterns and figures present in the data.

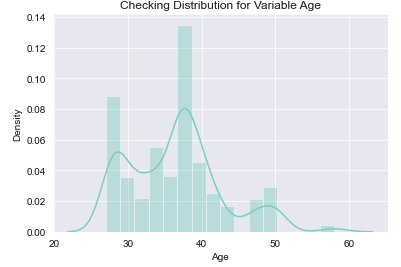
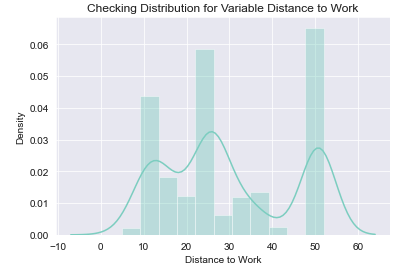


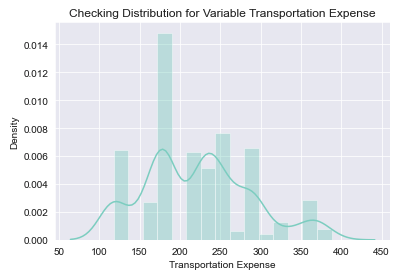
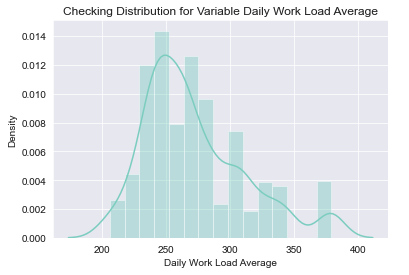






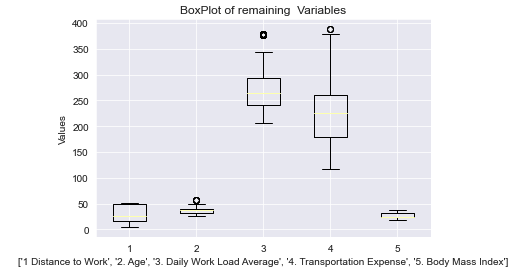
* **Normal Distribution** - Second , we should always check the normal distribution of each column which additionally helps to extract the useful information for the outcome and speaks about the data.



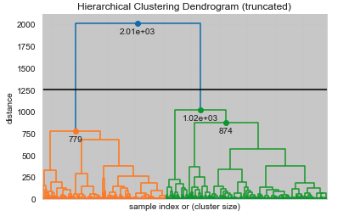


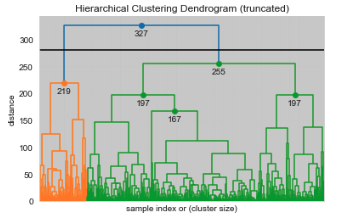
* **Outlier Detection** – From such probability distributions, we can clearly note that most of the variables are skewed. The presence of outliers and extreme values in the data can most likely be explained by Skew in these distributions. The existence of outliers is one of the other pre-processing steps, apart from testing for normality. We use a classic method of eliminating outliers in this. Using boxplots, we illustrate the outliers.

Deal with outliers : Outliers are nothing but a problem which deviates the accuracy of the model from its precision, in order to handle we first find out the outlier values and replace them with their closest median values.

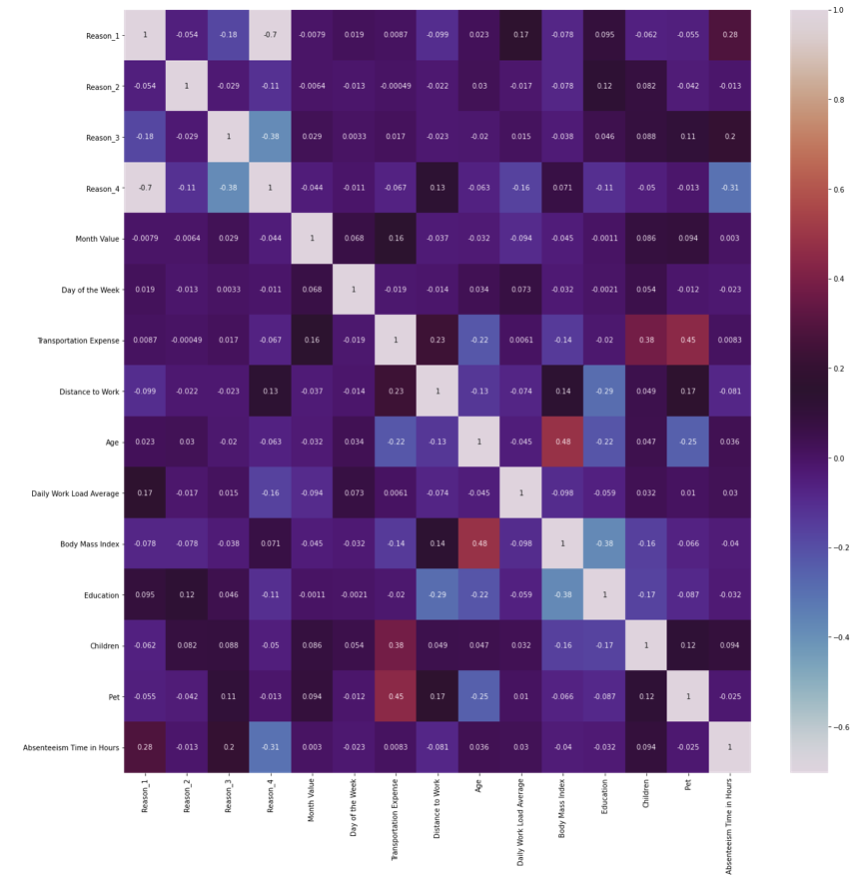


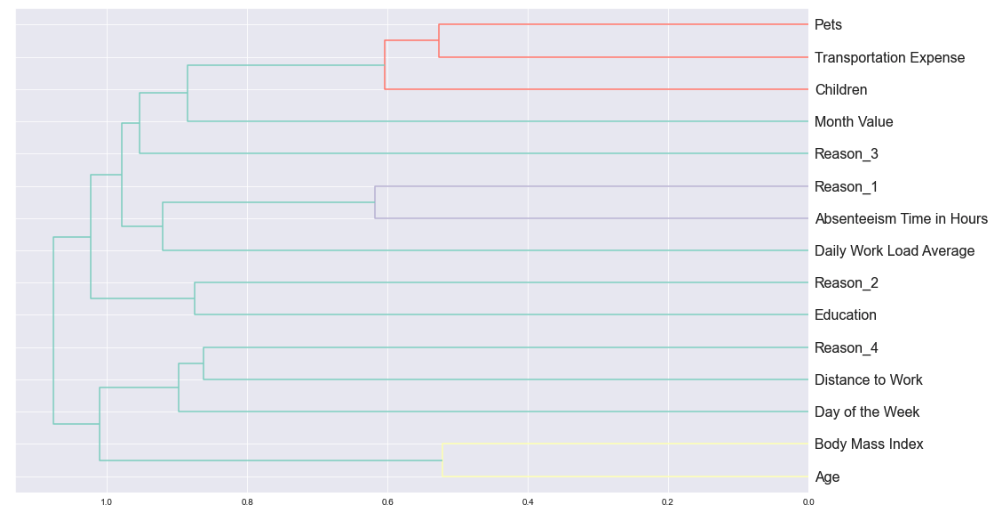
* **Clustering of the Data** - The algorithm begins with all data points allocated to their own cluster. Two of the closest clusters are then combined into the same cluster. In the end, when there is just a single cluster remaining, this algorithm terminates.





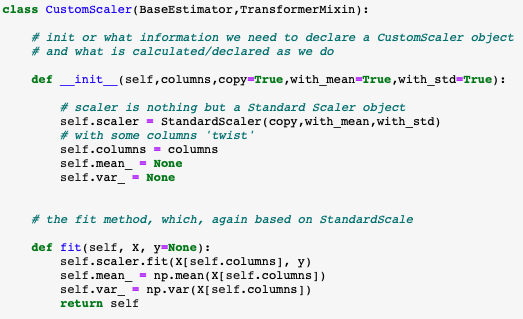
* **Feature Selection** - We need to access the value of each predictor variable in our analysis before conducting some kind of modelling. There is a probability that several variables are not at all important to the problem of class prediction in our research. The selection of subsets of corresponding columns for model construction is known as the Feature Selection. To minimize overload, we use the technique of feature selection to extract meaningful data characteristics.





* **Feature Scaling** - A technique used to standardize the range of independent variables or data features is function scaling. It is also known as data normalization in data pre-processing and is normally done during the pre-processing stage of the data.
  + **Standardize the data** - Standardization is one of the most popular pre-processing tools as data of different magnitude (scale) may be skewed to high values, we want all inputs to be of similar magnitude, this is a particularity of machine learning in general, with unscaled data performing badly for most (but not all) algorithms.
    - In our case, we need to scale['Reason 1 ',' Reason 2 ',' Reason 3 ',' Reason 4 ',' Education ', pet and' children '] because these are the columns that contain categorical data, but in numerical form, so we need to transform them. We need to choose the variable that needs to be transformed or scale.
    - **You can ask why we are actually performing the analysis manually and column wise ?**

Since evaluating data features wisely is always good, it gives us confidence in our model and we can easily interpret our model analysis.

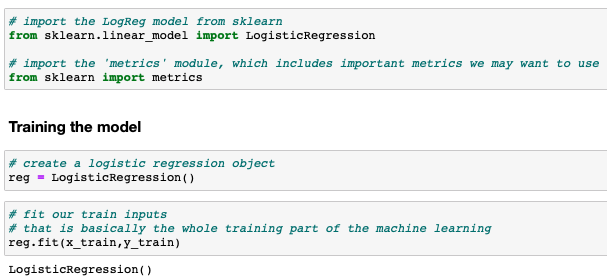


**Modeling and Classification**

In this research, we have used a total of two popular machine learning techniques for the purpose of classification. In order to lay out options and analyze the potential effects of selecting those options, we have realized that regression based classifiers can provide a highly effective framework.

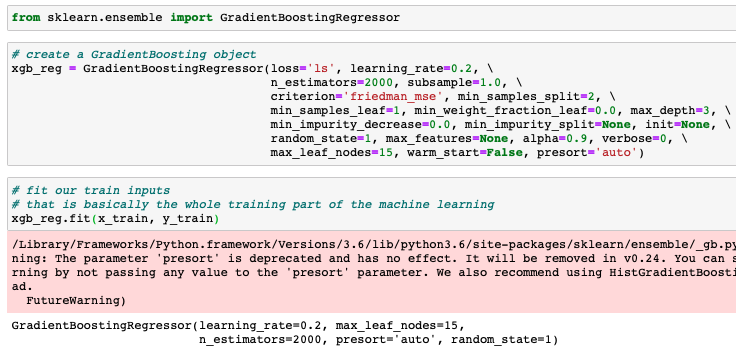
* **Logistic Regression** - A simple classification technique is logistic regression. It belongs to the linear classifier group and is very similar to linear and polynomial regression. Logistic regression is simple and reasonably uncomplicated, and interpreting the results is easy for you. Although it is basically a binary classification process, it can also be extended to multi-class issues.

**Train the Model** – We will first segment the data into the train and the test. On the train dataset, we will establish our model.



* **Gradient Boost Regressor -** A group of machine learning algorithms that combine several weak learning models together to produce a strong predictive model are gradient boosting classifiers. The concept behind "gradient boosting" is to take a poor hypothesis or weak algorithm of learning and make a series of tweaks to it that will increase the hypothesis/power. learner's This form of Boosting Hypothesis is based on the concept of roughly accurate learning probability (PAC).

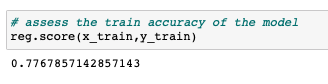
**Training the Model** - We will first segment the data into the train and the test. On the train dataset, we will establish our model.



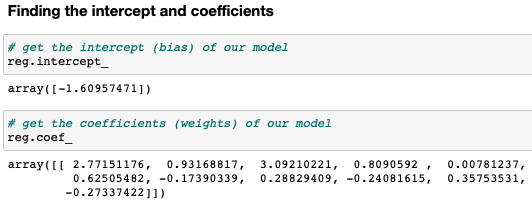
**RESULTS AND OUTCOMES**

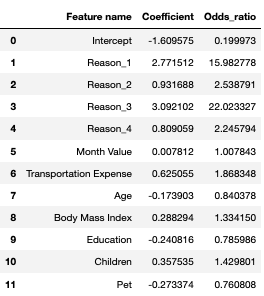
**Logistic Regression**

* **Accuracy of the fitted Model -** The most intuitive measure of performance is accuracy and it is simply a proportion of correctly predicted observations to total observations. One will think that our model is better if we have high precision. Yeah, precision is a great indicator, but only if you have symmetrical datasets where false positive and false negative values are approximately the same. Therefore to determine the consistency of your model, you have to look at other parameters.

****

* **Finding the intercept and Coefficients** – Check the beta value and the coefficients of our model.





Standard Coefficients : The coefficient of regression where all variables have been standardized.

**A feature is not particularly important to predict the outcome if its :**

1. If its standard coefficient is around 0
2. If its odds ratio is around 1

A weight of 0 implies that no matter what actually is the feature value , we will always multiply it by the value 0.  
The odds ratio means that – For a unit change in the standardized feature, the odds increase by a multiple equal to the odds ratio.

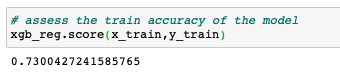
**Backward Elimination** – Since our data has unusable features which eventually deviates the accuracy of the model, so in order to handle this particular problem we have introduce the concept of the Backward Elimination.

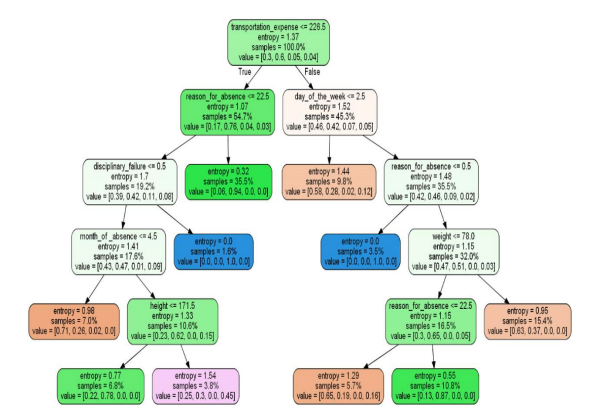
The idea is that we can simplify our model by removing all the features which has close to 0 contribution to the model

**If the coefficient value is less than p < 0.05 then we will take those coefficients for further processing.**

**Gradient Boosting**

* **Accuracy of the fitted Model –** In this case the models sensitivity and the specificity comes into the picture which can be calculated by the True Positive , True Negative , False Positive and False Negative.

****

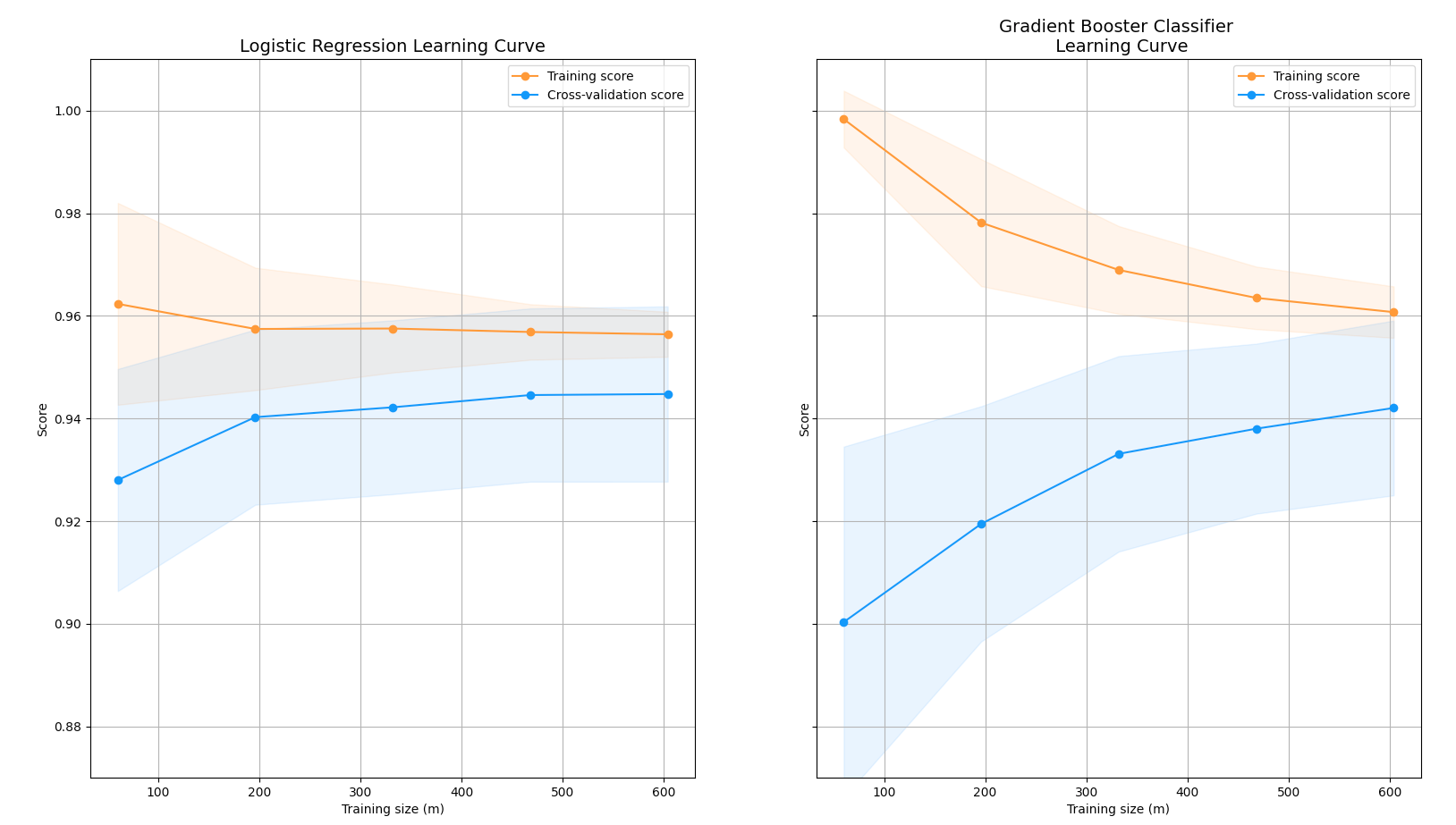


After performing the classification on training test of the split data we have seen the promising results from the classifiers in terms of accuracy and the logistic regression is one the which performs well among all with an accuracy rate of 79 %

For more approval on the scores I have also performed the rotation estimation (cross validation) of the classifying data and the scores are as follows :

The cross validation score of the Logistic Regression : **79.28**

The cross validation score of the Gradient Boost : **74.67**

****

**FINDINGS (WHAT MAKES THIS PROJECT FROM OTHER RESEARCHES)**

**Probability is What we are craving for !!!**

Along with performance of the algorithms (classification techniques) we have also calculated the probability of a person being absent at work. The probability of being absent is what is unique among all other research and operations on this particular data.

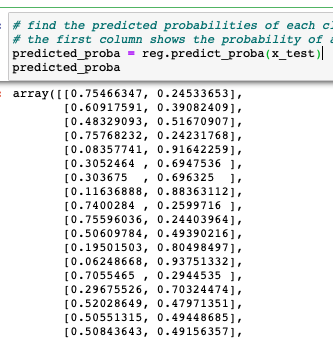
For calculation of the probability, the logistic regression is best model since it comes with the utility (defined function) which helps us to calculate the precise accuracy.

The predict\_prob method of the logistic regression requires the standard scaled test data which is being through the feature scaling and Max Min normalization. It always return the probability estimates. The estimates returned for all classes are ordered by class name. If multi-class is set to "multinomial" for a multi-class problem, the softmax feature is used to find the expected likelihood of each class. Else uses a one-vs-rest method i.e. using the logistic function to measure the likelihood of each class assuming it is positive. Throughout all the groups, and normalise these ideals.

**Are we sure that is a required probability** ?

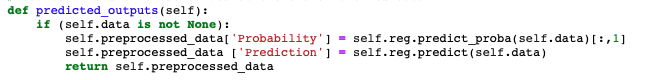
Yes, Since there are following physiognomies which actually corroborates that these are actual and real probabilities :

1. The outcome of probability is always in between 0 and 1.
2. If you count up the probability of being in some given class of an observation, they amount to 1.

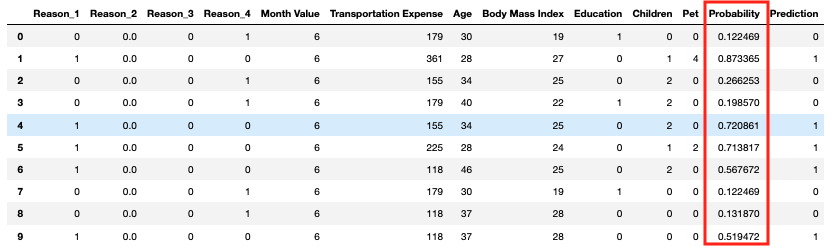


Here, the predict\_proba variables returns the decimal values of the probabilities. To sum up, this returns the array which holds the exact probability of absenteeism based on the test drive data which we have passed to the model.

With the predict\_proba method we have merge the two new columns – ‘Probability’ and ‘Prediction’ in the predicted data.



Now the predicted data comes with the parameters and the calculated probability and the prediction trigger value.



**Findings using the classifiers**

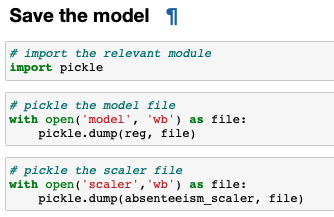
For the classification purpose I have used Logistic Regression and Gradient Boost. For the accuracy measure , we have also created the learning curves (The wider the gap between the training score and the cross validation score , the chances are that model is overfitted). Also if the score of both the learning curve and the training is quite low then we can assume that the model is underfitting the data. For the classification purpose we split the data with a test ratio of 0.2 . The primary goal of our classifiers are to determine the difference between the fraud transactions and non- fraud transactions.

The training score of the classifiers are as follow:

|  |  |
| --- | --- |
| **Classification Technique** | **Accuracy Score** |
| Logistic Regression | 78 % |
| Gradient Boost | 73 % |

**MODEL DEPLOYMENT**

Before the deployment of the model, we first need to consider how end users can interact with the predictions of the model in order to determine how to deploy a model. In order to save the model we are using the Python Pickle Module.



For the sake of the usage, we have created the scale and model file using the pythons pickle module (we have created the **model deployment files** which is being used for creating the API’s , Django and Flask Web Sites).

1. Open any python editor (Integrated Development Environment) or you can also use the Jupyter Notebook. To the customer centric deployment , the results can be generated in just four lines of code.
2. With this assignment, We will share a zip file which consist of:

* absenteeism\_module.py (This python file consist of the methods and functions which will process and executes the input data).
* Absenteeism\_new\_data.csv (This is the input file which a user or a client will bring to use the model).
* Model (This model is the COCOM type executable model which generates the probability of the Absenteeism).
* Scaler (The scaler file act as the input file for any type of method and this is generated using the pickle module).

1. Once you open the Jupyter notebook , Import all the module from absenteeism\_module.py



1. Now , create a variable of your choice and import the model and scale under the absenteeism model method.



1. And now , Import the input data (the input data can be raw data with outliers and all other error prone components).



1. Now the fourth and the final step (yes it’s the last step), use your variable which you have used for storing the model and scale values. We have a dedicated method written in for generating the probability output and store your results in a csv for visualization and other purpose.



Note : You can also copy and paste (use) the same code for usage as well. These four lines of code can generate the fruitful results.

Heads Up : You schema must match with the model schema.

**CONCLUSION and PROGNOSTICATION**

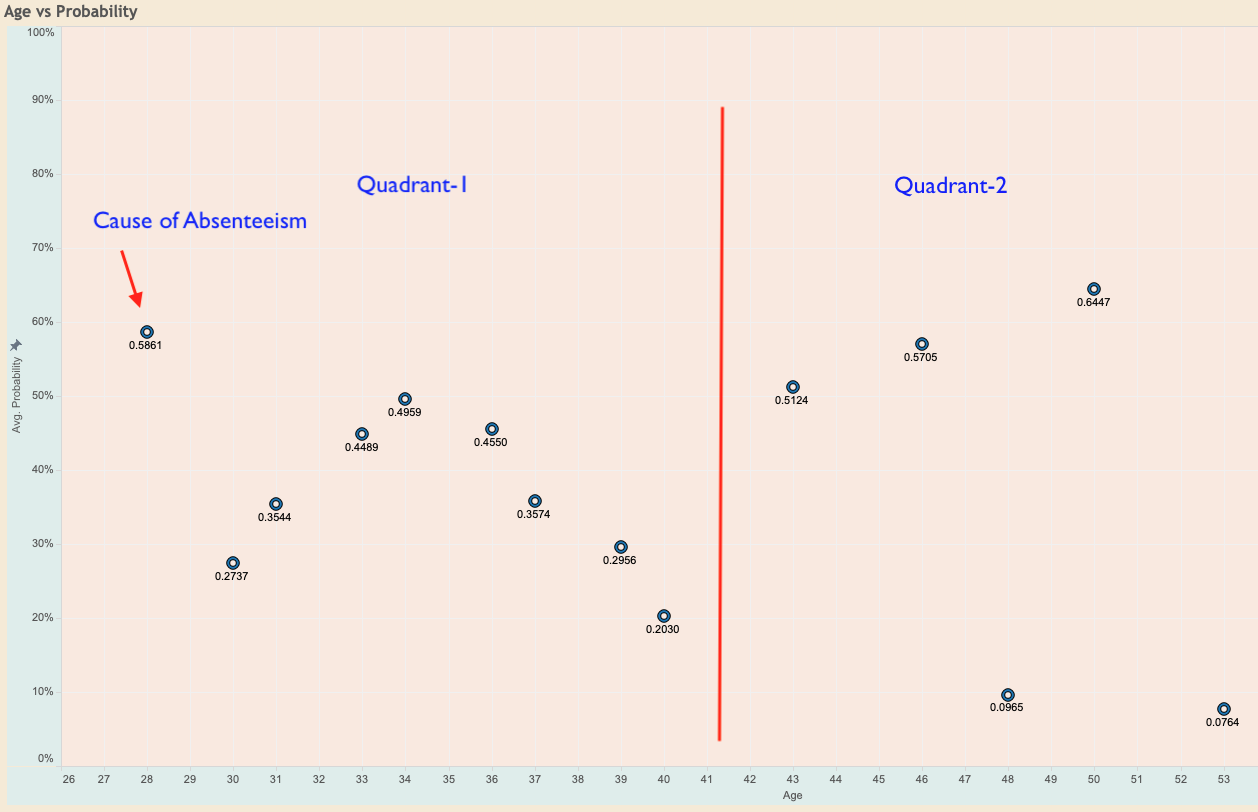
* **Problem Statement -** we have calculated the probability of a person being not present along with the comparison of the algorithms. Estimate the precise probability of the absenteeism established on the convinced possible reasons extant in the data.

For the better visualization and conclusion of the predicted results, we have explicitly used the Tableau for better insights.

Now since we have calculated the probability , we now scrutinize the considerations for

1. Age vs Probability of Absenteeism.

For the age vs probability section, we have taken the average probability and divide the workbook into two quadrants(one with age less than 50 and other with age greater than 50) and came to conclusion that the employees with higher age factors are more likely to absent from the workplace for more than usual time. However, there are some exceptions when an employee with less than age of 50 is absent due to the maintained reasons.



1. Reasons vs Probability

For the Reason vs Probability section, we have taken the probability (not the average since it varies for each reason. Each of the reason has 0 and 1 Flag (0 implies that the person was not absent because of this particular reason and 1 implies that the person was absent due to this particular reason).



1. Transportation and Children