

**PREDICTING ABSENTEEISM OF EMPLOYEES AT WORKPLACE USING**

**LOGISTIC REGRESSION & GRADIENT BOOSTING ALGORITHM**

By

**SAHIL NAGPAL (0670042)  
PRAMINA PATIL (0663967)**

**Applied Modeling & Quantitative Methods**

**Trent University**

**Fall 2020**

# TABLE OF CONTENTS

[TABLE OF CONTENTS 2](#_Toc58324214)

[ABSTRACT 4](#_Toc58324215)

[INTRODUCTION OF THE PROJECT AND SUBJECT 5](#_Toc58324216)

[Motivation of the Research 5](#_Toc58324217)

[Problem Statement 5](#_Toc58324218)

[Research Questions 5](#_Toc58324219)

[Research Scope 6](#_Toc58324220)

[DATASET AND EXPERIMENTAL PROCEDURE 7](#_Toc58324221)

[Description of the Dataset 7](#_Toc58324222)

[Data Preprocessing Methodology 9](#_Toc58324223)

[Exploratory Data Analysis 9](#_Toc58324224)

[Normal Distribution 12](#_Toc58324225)

[Outlier Detection 13](#_Toc58324226)

[Clustering of Data 14](#_Toc58324227)

[Feature Selection 15](#_Toc58324228)

[Feature Scaling 16](#_Toc58324229)

[Standardize the data 17](#_Toc58324230)

[MODELING AND CLASSIFICATION 18](#_Toc58324231)

[Logistic Regression 18](#_Toc58324232)

[Train the Model 18](#_Toc58324233)

[Gradient Boost Regressor 19](#_Toc58324234)

[Training the Model 19](#_Toc58324235)

[RESULTS AND OUTCOMES 20](#_Toc58324236)

[Logistic Regression 20](#_Toc58324237)

[Accuracy of the fitted Model 20](#_Toc58324238)

[Finding the beta intercept and Coefficients 20](#_Toc58324239)

[Backward Elimination 22](#_Toc58324240)

[Gradient Boosting 22](#_Toc58324241)

[Accuracy of the fitted Model 22](#_Toc58324242)

[SIGNIFICANT FINDINGS OF THE STUDY 25](#_Toc58324243)

[Findings using the classifiers 27](#_Toc58324244)

[MODEL DEPLOYMENT 28](#_Toc58324245)

[CONCLUSION AND PROGNOSTICATION 30](#_Toc58324246)

[REFERENCES 33](#_Toc58324247)

# [ABSTRACT](#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 illnesses to be detected by excesses of occurrences. The primary aim of the early absenteeism analysis was to predict the characteristics and groups of employee absence and the reason of absence that make them implement superior 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 considered to minimize higher absenteeism at work. We now use two prominent machine learning algorithms in this project, namely Logistic Regression and Gradient Boosting, to predict employee absenteeism time and to discover the insights that prompt employees to perform greater absenteeism at work. Through this we have also find the probability of an employee of being absent at the workplace based on evocative parameters. Therefore to find out the best classifier that delivers the maximum prediction accuracy, compare the various machine learning algorithms. We used an existing dataset of a courier organization in Brazil to determine the amount of time of employees' absenteeism. Similarly, we have also crafted a threshold assessment of the extreme and normal absenteeism grounded on the hours absent at workplace. Using the machine learning algorithm, we revealed that the best facts of exhibiting higher absenteeism at work are the justification for absence class as diseases corroborated by the International Code of Diseases (ICD) and the transportation cost from home to work.

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 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 an 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 EXPERIMENTAL PROCEDURE

## Description of the Dataset

The dataset we used in this study was obtained from the UCI Machine Learning Repository, which was created from July 2007 to July 2010 at a courier company in Brazil with records of employee absenteeism at work. Firstly, 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 | Patients follow up |
| 23 | Medical consultation |
| 24 | Blood donation |
| 25 | Laboratory examination |
| 26 | Unjustified absence |
| 27 | Physiotherapy |
| 28 | Dental consultation |

## Data Preprocessing Methodology

We need to clean up the data and translate it into the correct format before feeding the data into the model. The output of the machine learning model is completely dependent on the data fed in the algorithm. The Machine Learning model does not generate the proper outcome on raw data.

### Exploratory Data Analysis

Primarily, we checked the affiliation between the columns and make an appropriate visualization which better depicts the regular outlines, patterns, and figures present in the data.

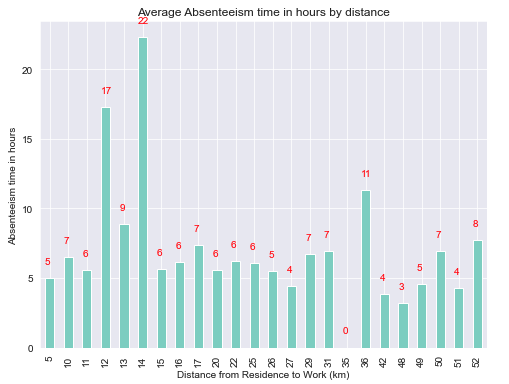


Fig 1: Relation between Absenteeism & Distance

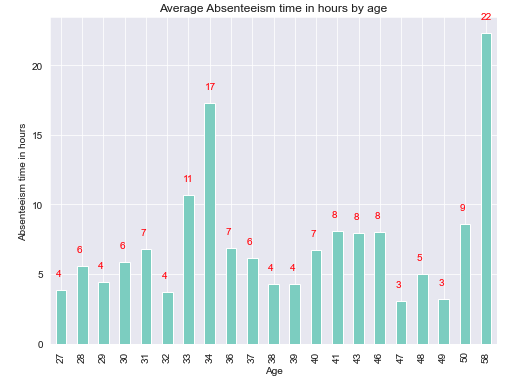


Fig 2: Relation between Absenteeism and Age

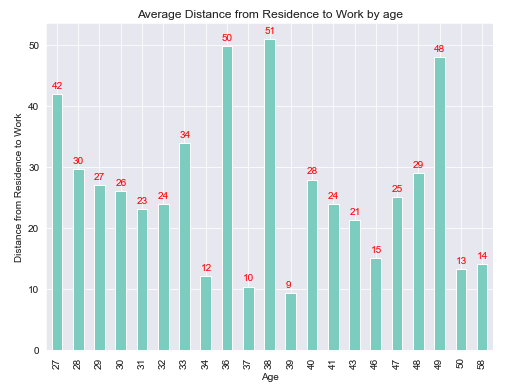


Fig 3: Relation between Distance to work and Age

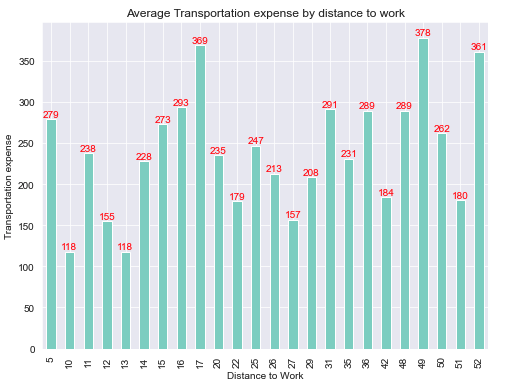


Fig 4: Relation between Distance to work and Transportation Expense

### Normal Distribution

Secondly, 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.

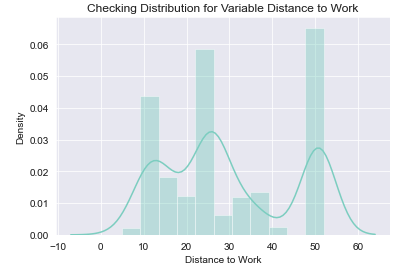
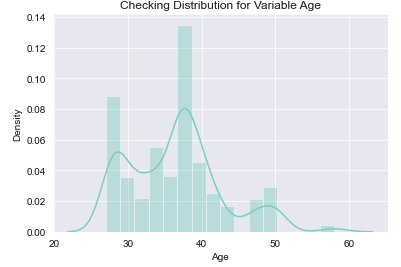
 

Fig 5: Distribution for Distance to Work Fig 6: Distribution for Age

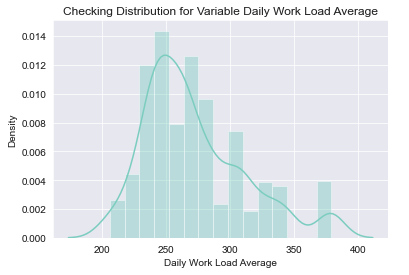
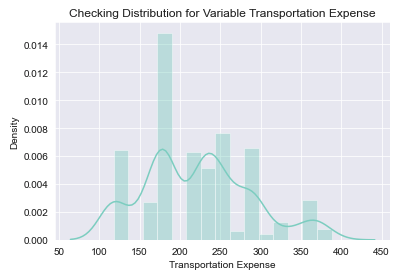
 

Fig 7: Distribution for Daily Workload Average Fig 8: Distribution for Transportation Expense

### 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.

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

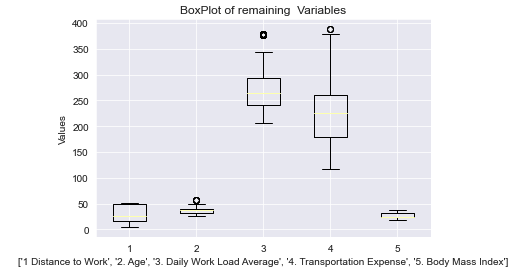


Fig 9: Outlier Illustration

### Clustering of 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.

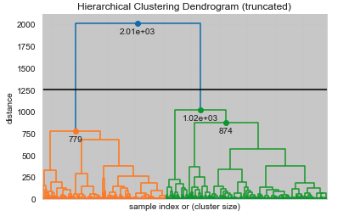


Fig 10: Hierarchical Clustering Dendrogram

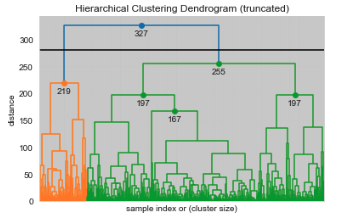


Fig 11: Hierarchical Clustering Dendrogram

### Feature Selection

We need to access the value of each predictor variable in our analysis. 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.

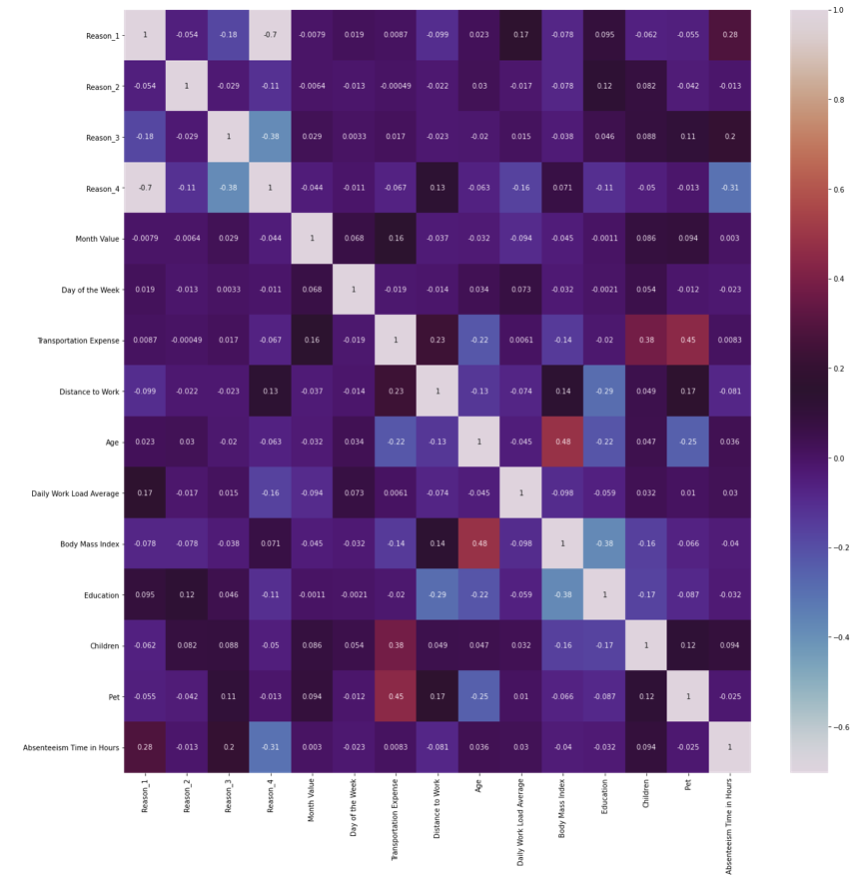


Fig 12: Correlation Matrix among variables

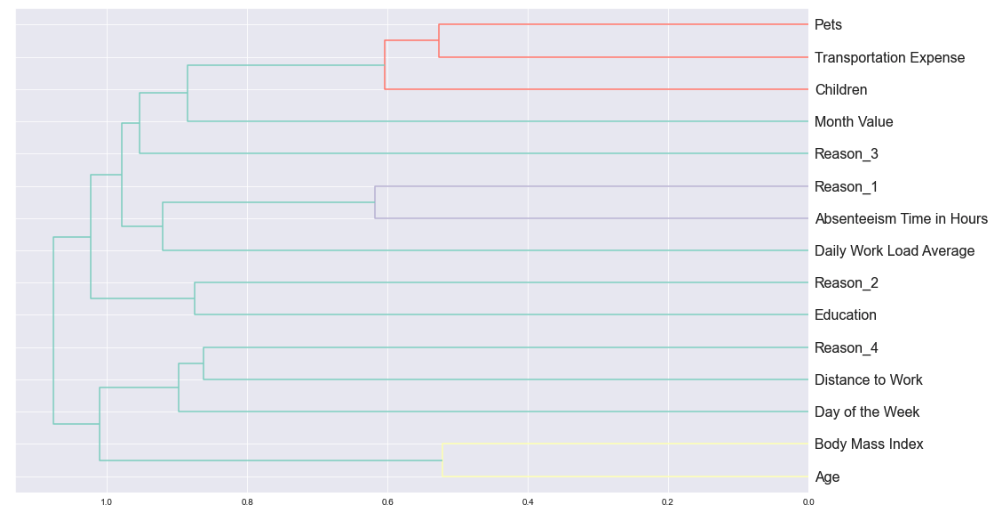


Fig 13: Dendrogram for Absenteeism 1

### 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

As data of different magnitude (scale) can 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 underperforming for many other (but not all) algorithms, standardization is one of the most common pre-processing tools. 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.

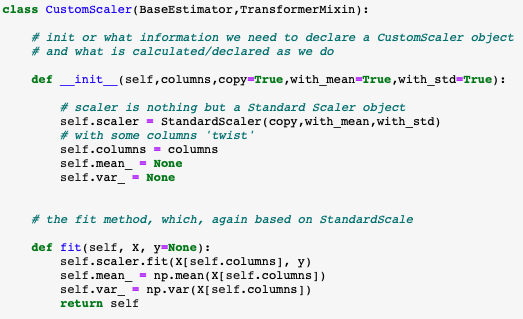


Fig 14: Class CustomScalar

# MODELING AND CLASSIFICATION

In this research, we have used a total of two popular machine learning techniques for the purpose of classification. 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.

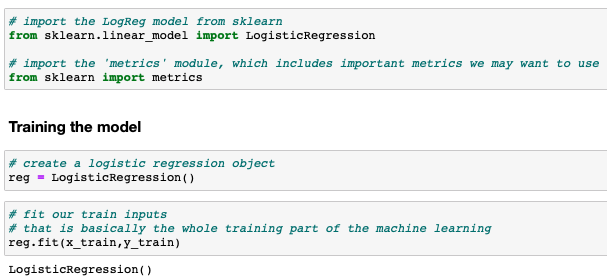


Fig 15: Train Model

## Gradient Boost Regressor

Gradient boosting classifiers are a group of machine learning algorithms merging many weak supervised learning together to generate a strong predictive model. "The idea behind "gradient boosting" is to take a bad learning hypothesis or weak algorithm and make a set of tweaks to it that will improve the power/hypothesis. This type of Boosting Hypothesis is based on the idea of an approximately precise probability of learning (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.

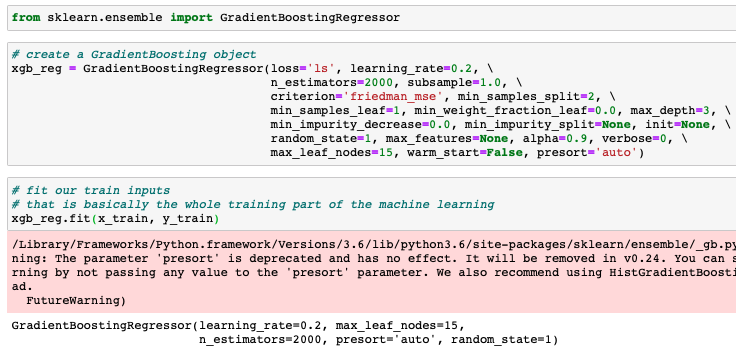


Fig 16: Gradient Boosting

# 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 must look at other parameters.

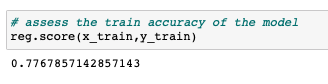
****

Fig 17: Accuracy of the Model

### Finding the beta intercept and Coefficients

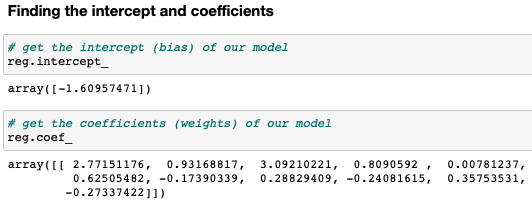


Fig 18: Finding intercept and coefficient

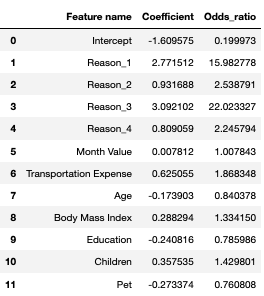


Fig 19: Standard coefficients

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

**A feature is not particularly important to predict the outcome if the standard coefficient is around 0 or the odds ratio is approximately 1.**

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

### Backward Elimination

As our data has unusable features which eventually deviates the accuracy of the model, we have introduced the concept of the Backward Elimination to counter this. 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 model’s sensitivity and the specificity comes into the picture which can be calculated by the True Positive, True Negative, False Positive and False Negative.

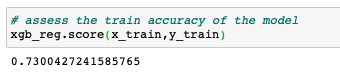
****

Fig 20: Accuracy of fitted model

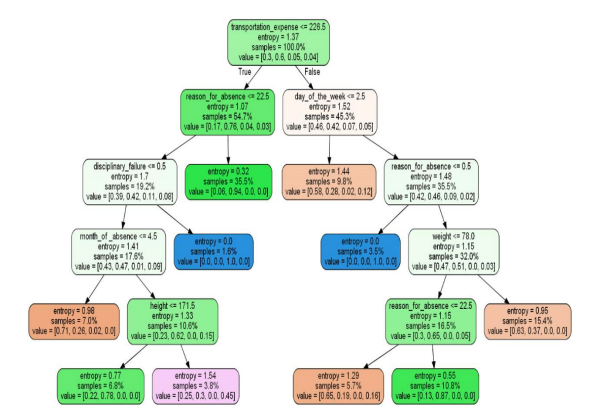


Fig 21: Splitting the data

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 we 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**

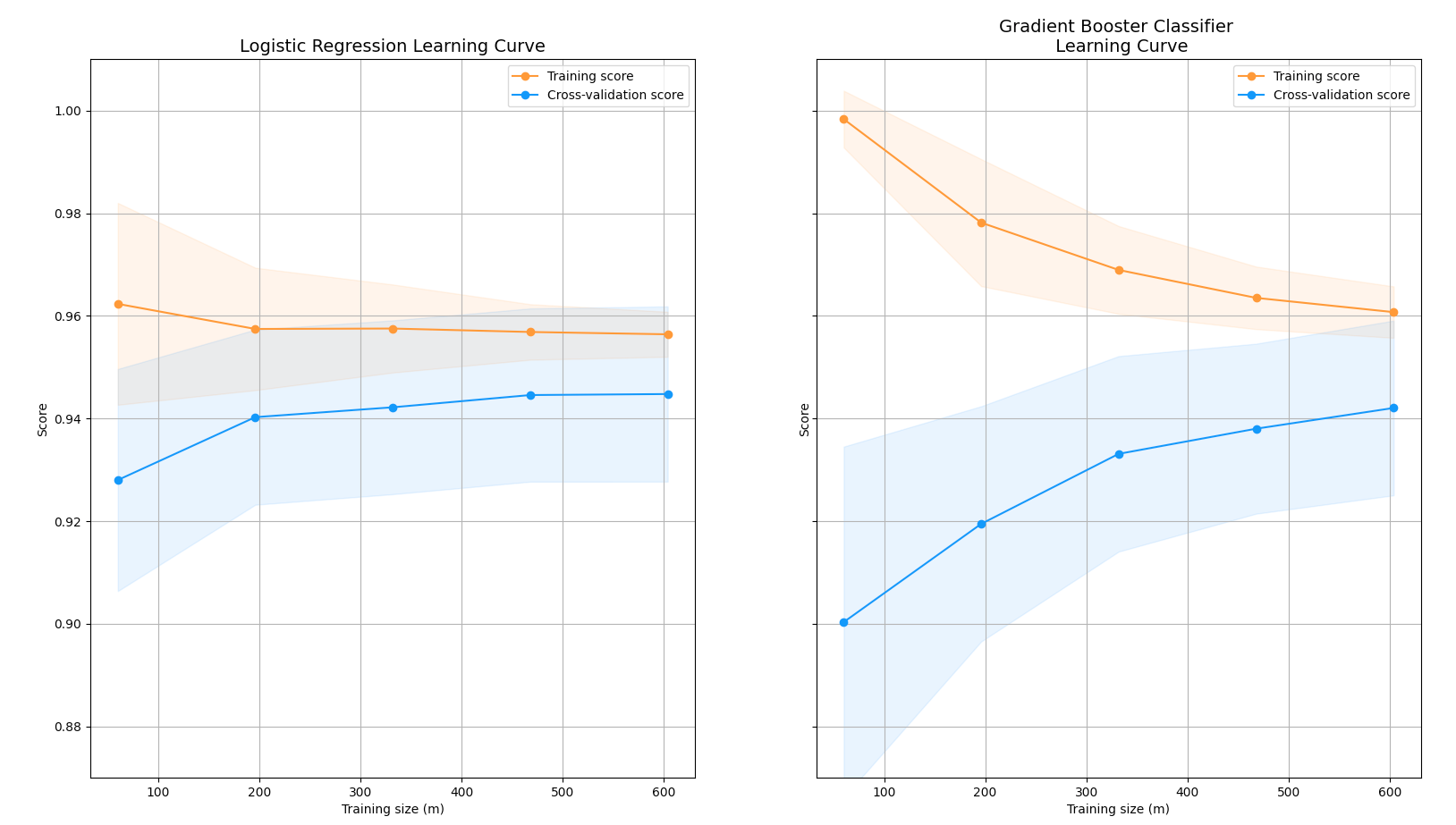
****

Fig 22: Logistic Regression Learning Curve Fig 23: Gradient Booster Classifier Learning Curve

# SIGNIFICANT FINDINGS OF THE STUDY

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 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 returns 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 normalize these ideals.

The following characteristics confirm the correctness of these probabilities:

1. The outcome of probability is always in between 0 and 1.
2. The sum of these probabilities amounts to 1.

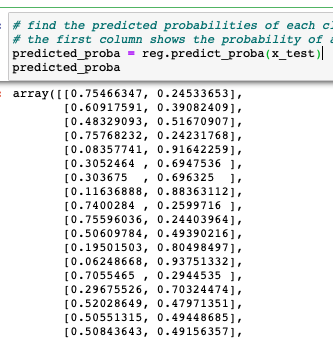


Fig 24: predict\_proba method

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 merged the two new columns – ‘Probability’ and ‘Prediction’ in the predicted data.

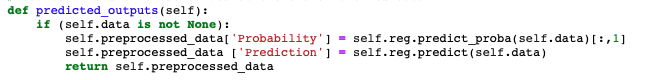


Fig 25: predicted\_outouts function

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

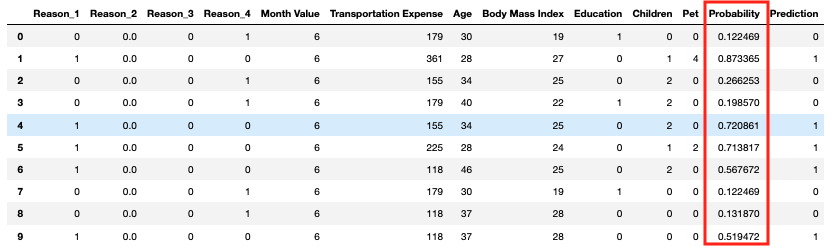


Fig 26: Calculated Probability

## Findings using the classifiers

For the process of classification, we 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. We split the data with a test ratio of 0.2 and 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 to determine how to deploy a model. We save the model using the Python Pickle Module.

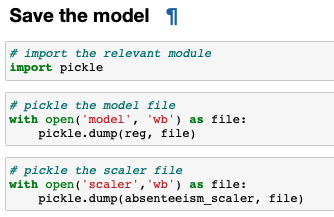


Fig 27: Importing pickle module

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
2. With this assignment, we will share a zip file which consist of:

* absenteeism\_module.py (This python file consists 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



Fig 28: Importing absenteeism\_module

1. Create a variable of your choice and import the model and scale under the absenteeism model method



Fig 29: Importing model and scale

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



Fig 30: Loading and cleaning the data

1. Now the fourth and the final step, use your variable which you have used for storing the model and scale values

(We have a dedicated method for generating the probability output and store your results in a .csv file for visualization and other purpose)



Fig 31: Storing results to CSV

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. Your 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.

**Age vs Probability**

We now scrutinize whether the probability of being absent is dependent on the person’s age. For this, 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.

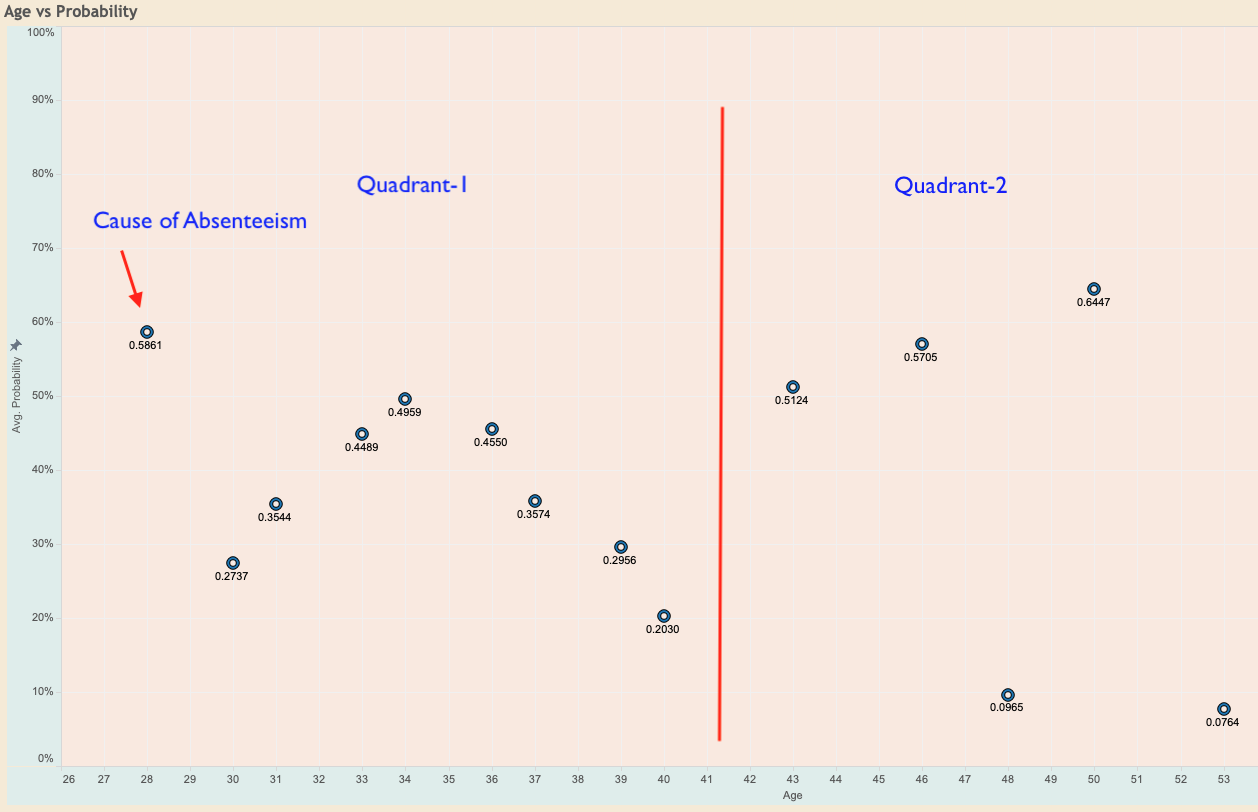


Fig 32: Age Vs Probability

**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 due to this particular reason).



Fig 33: Reason Vs Probability

**Transportation and Children**

The absenteeism can also be instigated by the transportation cost. From the visualization we can see that the transportation charges greater than 200 can cause the absenteeism and it is more common with the employee having either 1 or more children.

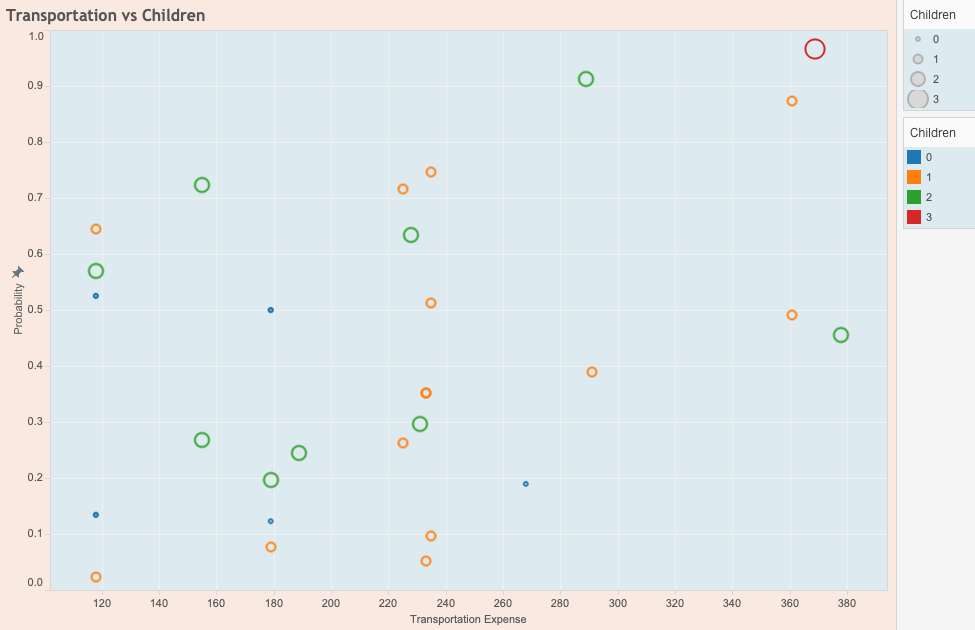


Fig 34: Transportation Vs Children

**Business Logic**

Absenteeism has been costing U.S. companies billions of dollars every year in terms of

productivity, wages, poor quality of goods/services and excess management time.

Absenteeism contributes towards individual’s productivity, performance of the team and companies’ profits.

Using the Absenteeism Analysis, companies should predominantly focus on employees who are marking high absenteeism which is one of the major causes of business loss.

Companies can enhance and focus on several topics like:

* Pay overtime for employees
* Help employees living far away with a rented accommodation option.
* Provide paid leave or a choice to encash leaves
* Arrange health camps on a regular basis to ensure all the employees are fit.

# REFERENCES

1. Datasciencelovers. (2019, December 07). Retrieved December 06, 2020, from <http://www.datasciencelovers.com/tag/absenteeism-prediction/>
2. What is Absenteeism? (n.d.). Retrieved December 06, 2020, from <https://www.betterteam.com/what-is-absenteeism>
3. Madlinger, G. (2018, March 20). The 6-Step Process for Dealing with Employee Absenteeism. Retrieved December 06, 2020, from <https://wheniwork.com/blog/how-to-deal-with-employee-absenteeism>
4. [https://archive.ics.uci.edu/ml/datasets/Absenteeism at work](4.%09https:/archive.ics.uci.edu/ml/datasets/Absenteeism%20at%20work)
5. Author: Erik van Vulpen Erik van Vulpen is the founder of the Academy to Innovate HR (AIHR). He is a globally recognized HR thought leader and teacher in the future of HR. He regularly speaks at conferences about HR training and upskilling. Connect, Author: & Erik van Vulpen is the founder of the Academy to Innovate HR (AIHR). He is a globally recognized HR thought leader and teacher in the future of HR. He regularly speaks at conferences about HR training and upskilling. Connect with Erik on LinkedIn. (2020, June 04). Absenteeism: A 3-Step Analytics Approach. Retrieved December 06, 2020, from <https://www.analyticsinhr.com/blog/absenteeism-3-step-analytics-approach/>
6. H., By, -, Hussain Mujtaba Hussain is a computer science engineer who specializes in the field of Machine Learning. He is a freelance programmer and fancies trekking, Mujtaba, H., Hussain is a computer science engineer who specializes in the field of Machine Learning. He is a freelance programmer and fancies trekking, & Here, P. (2020, June 08). What is Gradient Boosting and How is it different from AdaBoost? Retrieved December 06, 2020, from <https://www.mygreatlearning.com/blog/gradient-boosting/>
7. Nelson, D. (n.d.). Gradient Boosting Classifiers in Python with Scikit-Learn. Retrieved December 09, 2020, from https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/