Report on

Employee Absenteeism

using

Predictive Regression Models

Matthew Nnadozie

May 2022

Contents

Chapter 1. Introduction

Workplace absenteeism refers to employees' habitual or deliberate nonattendance to work, high rate of absenteeism has a huge impact on company strategies, from reduced productivity to increased indirect costs such as continued salary payments to indirect costs like payment for ad hoc and overtime wages. Absenteeism has been one of the daunting challenges faced continually by the human resources department in every organization (Cucchiella, et al., 2014). According to the data from the U.S Bureau of Labour and Statistics (United States Dept. of Labor, 2022), about 3.2 percent of the workforce was away from work on any particular day. Due to the adverse impact of absenteeism on productivity, it is therefore very important not only to identify the causative factors but to understand how they contribute to absenteeism in the workplace. The common already identified factors include bullying and harassment, poor staff morale, power reward and recognition systems, mental health issues, mobility challenges, workplace stress, burnout, etc. A recent study assessed the effect of smoking by current and former smokers on absenteeism (Halpern, et al., 2001), with the former recording higher absenteeism rates. A substantial relationship was also found between alcohol use and absenteeism (McFarlin & Fals-Stewart, 2002).

A clear understanding of the causes of absenteeism will help organizations develop better policies and support systems to help curb them and which will consequently impact the business bottom-line positively

The purpose of this report is to review the factors responsible for employee absenteeism with a view to predicting absenteeism rate using employee data collated over 3 months. Understanding the reasons for absenteeism would ultimately enable the business to reduce the losses associated with it. This project focuses on the use of the linear regression method to implement the predictive models.

1.1 Research Problem Statement

Workplace absenteeism has a significant cost implication for any organization, from an increase in cost and workload to a reduction in productivity and profit margin. if the factors underlying workplace absenteeism could be determined and future occurrences predicted then attempts can be made to minimize its effect on the organization.

1.2 Business Objectives

1. To determine the underlying factors causing absenteeism in the workplace

2. To predict absenteeism and prevent its occurrence.

.

1.3 Data

The dataset used is the absenteeism CSV dataset, it presents information on employees and their associated hours of absence from work over 3 months. The data frame consists of 740 observations classified under 16 variables: The variables are listed below.

i. Transportation expense
ii. Distance from Residence to Work measured in Kilometres
iii. Service time measured in years worked
iv. Age
v. Workload in hours
vi. Hit target
vii. Disciplinary failure (1 = yes, 0 = no)
viii. Education (1 = school, 2 = undergraduate degree, 3 = postgraduate degree, 4 = doctorate)
ix. Children
x. Drinker (1 = yes, 0 = no)
xi. Smoker (1 = yes, 0 = no)
xii. Pet
xiii. Weight in Kilograms(Kg)
xiv. Height
xv. Body mass index
xvi. Absenteeism in hours (Target variable)

1.4 Data Processing

This involves the manipulation of collected data to produce information that is meaningful and useful. It consists of three main processes.

1. Data cleaning which is the removal of errors from the dataset
2. Data preparation involves the removal of data that does not fit into the dataset and replacing them with useable data
3. Data transformation which is the transformation of data from one format to another.

The absenteeism dataset consists of 740 observations categorized under 16 variables, of which are integers except the variable workload which is a floating number. The variables Disciplinary failure, Education, Drinker, Smoker are categorical features and as such will be processed as factors to give better insight into the data.

```
data.frame':   740 obs. of  16 variables
$ Transportation.expense          : int
$ Distance.from.Residence.to.Work: int
$ Service.time                    : int
$ Age                             : int
$ Work.load.Average.day           : num
$ Hit.target                      : int
$ Disciplinary.failure            : int
$ Education                       : int
$ Children                        : int
$ Drinker                         : int
$ Smoker                          : int
$ Pet                             : int
$ Weight                          : int
$ Height                          : int
$ Body.mass.index                 : int
$ Absenteeism.time.in.hours       : int
```

Chapter 2. Exploratory Data Analysis

Exploratory Data Analysis(EDA) provides better insight into the structure of a dataset, it reveals the distribution of the variables in the dataset, the relationship between the variables and helps in identifying incorrect data points such as outliers so that they can be easily removed as the presence of outliers can impact the accuracy of a model. An investigation of the entire dataset was performed to determine the key characteristics of the dataset, the relationships between the variables, and how the variables relate to the target variable.

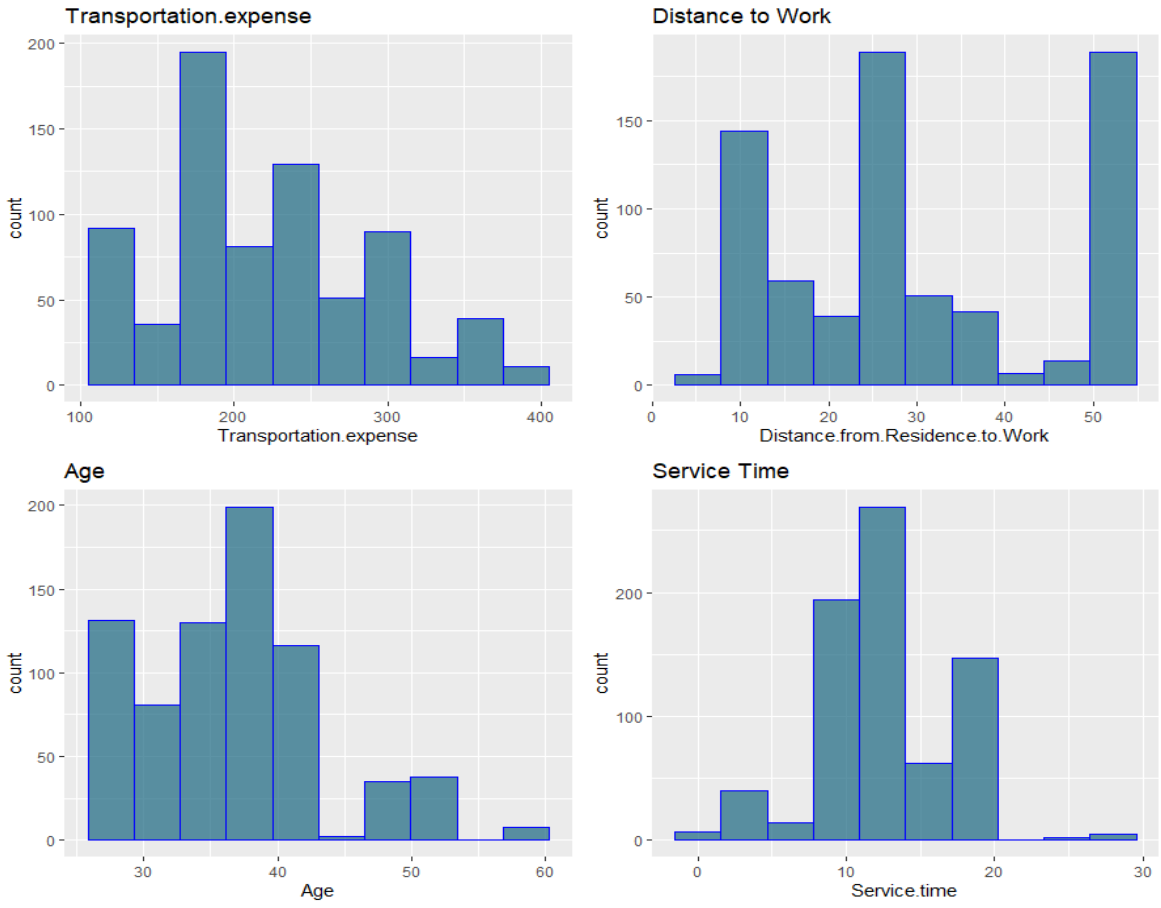
2.1 Missing Value Analysis

The dataset provided was reviewed for missing values. A value is defined as missing for a variable if it is useful but unavailable for the specific analysis (Graham, 2009). Missing data problems are widespread in research and can affect the result of any analysis (Graham, 2009). The review for missing values shows that there are no missing values in the absenteeism dataset. This is shown in the graph below.

Fig 1. Missing value analysis plot

2.2 Univariate Graphical Exploratory Data Analysis

Graphical Exploratory Data Analysis was performed to determine the underlying characteristics and distribution of the features in the dataset using histograms. Histogram plots are commonly used to show the frequency distribution. To start with, individual features that are numerical were analyzed and their distributions are visualized below.



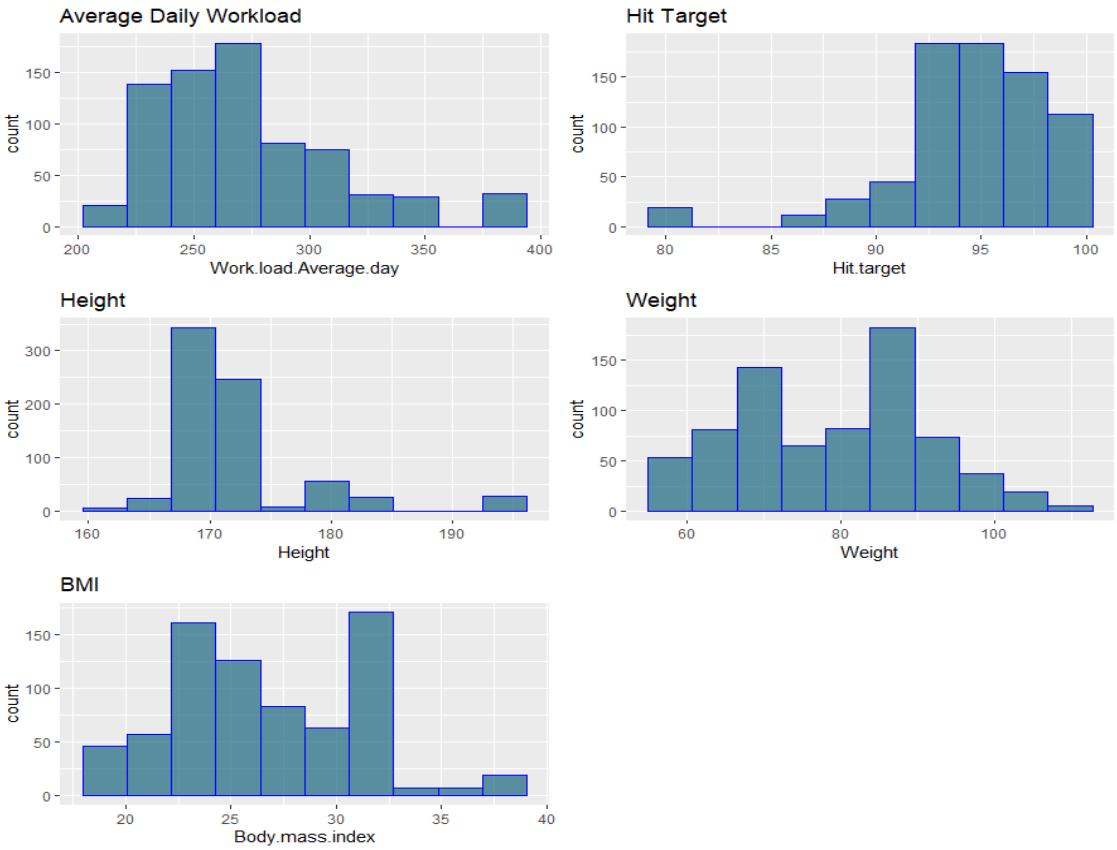


Fig 2. Histogram plots showing the distribution of the numerical features in the dataset.

The images above showed that the features in the dataset are mostly not symmetrically distributed as they are either skewed to the left or right, bimodal or multimodal distribution. Reviewing individual plots,

i. Transportation Expense

The graph of Transportation expense shows a multimodal distribution, an in-depth analysis reveals that about 45% of participants in this category spend between £ (100-200) on transportation over the period of the investigation.

ii. Distance to Work

The graph of this feature also shows a multimodal distribution. To make better meaning of this feature, a review of its relationship with transportation expense is carried out, it is expected that people who live further away from the business location would pay higher transportation fares.

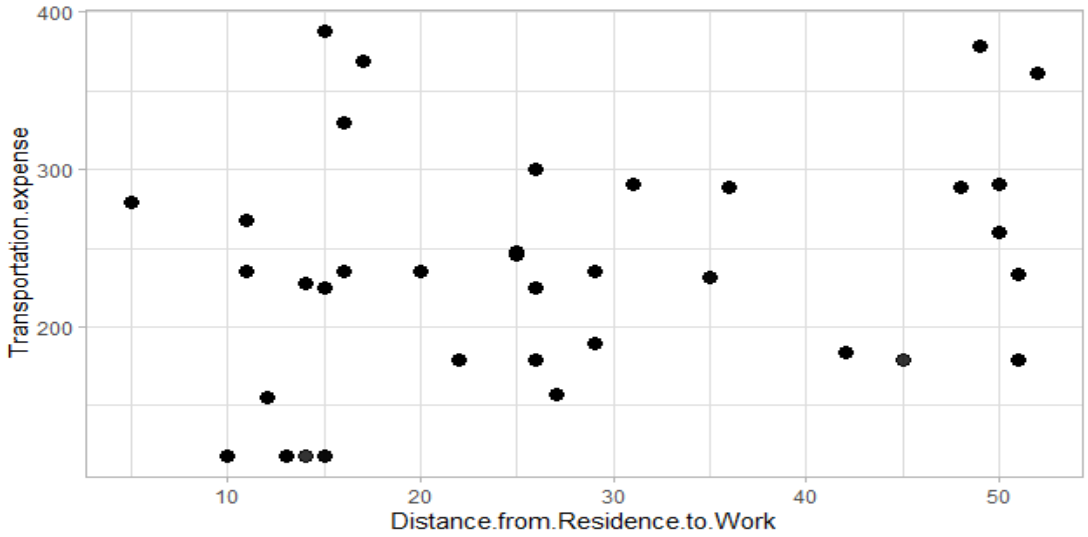


Fig 3. Scatterplot showing the relationship between transportation expense and distance to work

However, the scatterplot above does not show a linear correlation between the two features as expected, the only explanation for this behavior is if a different mode of transportation was used over similar distances leading to a sharp variation in transportation expense. This observation will be addressed during feature selection as only relevant predictor variables would be selected.

iii. Age

The graph shows a right-skewed distribution, it also revealed that most of the employees in the study were below 45years, there is also the presence of outliers in the age feature. This is further investigated and removed in the data pre-processing step.

iv. Other numerical features

The other numerical features in the dataset are either right or left-skewed as shown in the plots. They are further investigated for extreme values in the pre-processing step. The categorical features were also reviewed using bar plots to investigate and to determine the relationship that exists between the features and the target variable (see appendix 1)

2.3 Multivariate Graphical Exploratory Data Analysis

To determine the correlation between the independent and dependent features and the covariance that exists between independent features we deployed the correlation matrix.

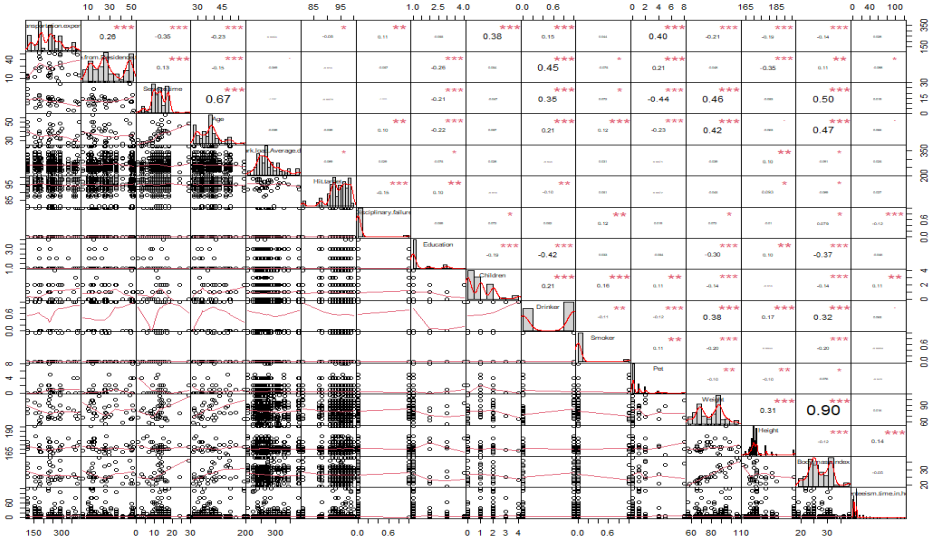


Fig 4. Correlation chart showing relationships and coefficient of variables in the dataset

The correlation chart revealed the presence of two independent variables (Weight and Body Mass Index) that are highly related with a correlation coefficient of 0.90. This is a case of multicollinearity.

Multicollinearity refers to a situation where two or more independent variables in a dataset are highly related with a linear absolute correlation coefficient of above 0.80. Multicollinearity affects the accuracy and reliability of any machine learning model due to the following reasons (Kassambara, n.d.). To deal with this challenge, weight - one of the variables with a high correlation coefficient was dropped from the dataset. The other variables do not strong negative or positive correlation.

Chapter 3. Methodology

The implementation of this project followed the sequence of phases as outlined in the Cross-industry standard process for data mining (CRISP-DM). Firstly, the underlying project objective and requirement from the business perspective was understood, followed by a detailed review of the data to gain insight into its intrinsic characteristic. Univariate and multivariate exploratory data analyses were performed to understand the data. The second phase involved data pre-processing activities to clean, format, and select only data features that are relevant to building the predictive model. This was followed by selecting the appropriate modeling technique and model building. In the final phase, the results obtained were evaluated using commonly accepted regression evaluation metrics, production of project reports, and review of the project.

3.1 Data Pre-processing

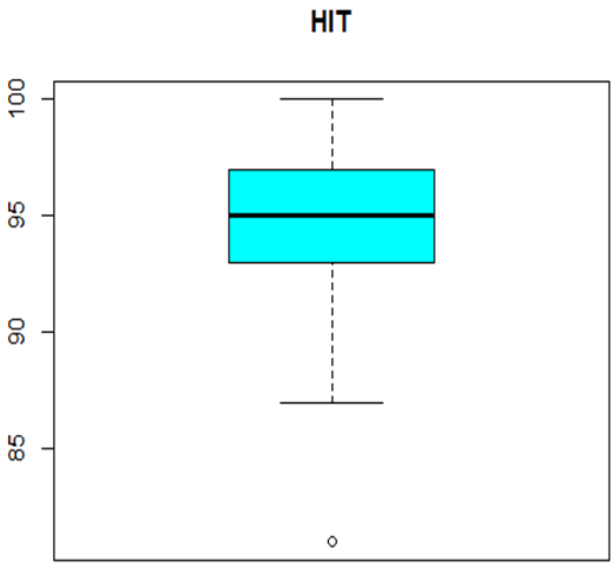
Having investigated the distribution of each of the variables in the dataset to determine the nature of the distributions. The graphs from the EDA show that the variables are not normally distributed but are either skewed to the left or right and in some cases have multi-modal distributions. Skewness most times are caused by the presence of outliers in the observation. Outliers are observations that have unusually high or low values.

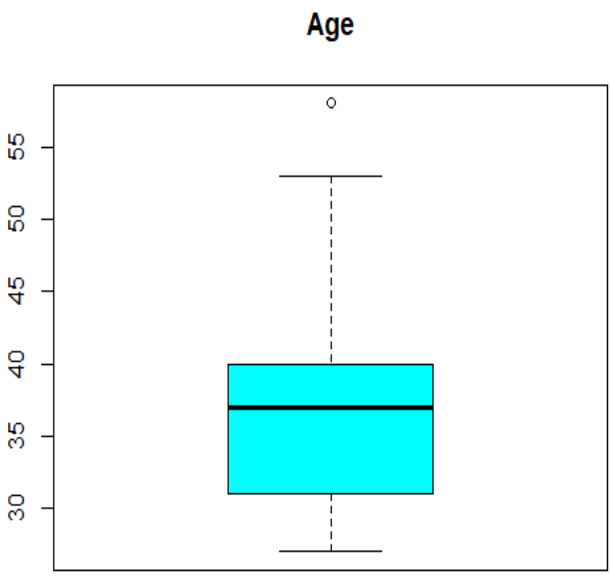
3.2 Outlier Analysis

Boxplots were used to confirm the presence of outliers in the variables. Outliers are unusual or abnormal values in the dataset and are capable of significantly altering any regression model. The presence of outliers in the variables was firstly investigated to determine if they provide meaningful insight into the data before deleting them.

The boxplots below show the presence of outliers in the following features;

i. Transportation expense'
ii. Service time
iii. Age
iv. Workload
v. HIT
vi. Height





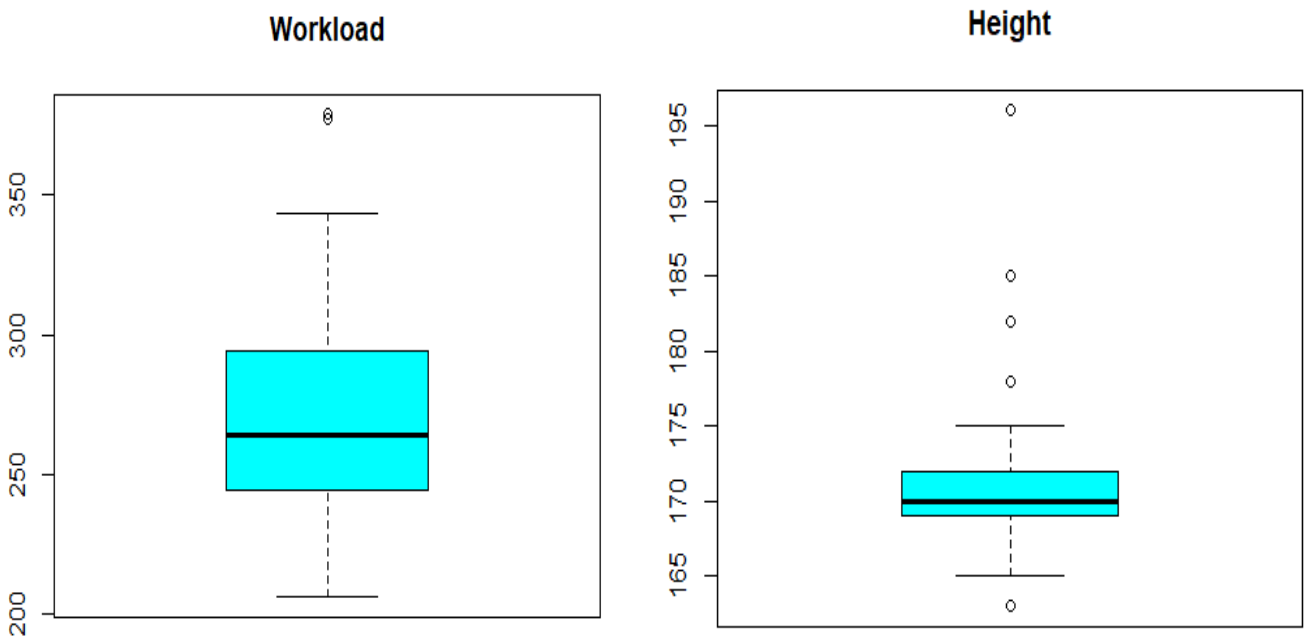
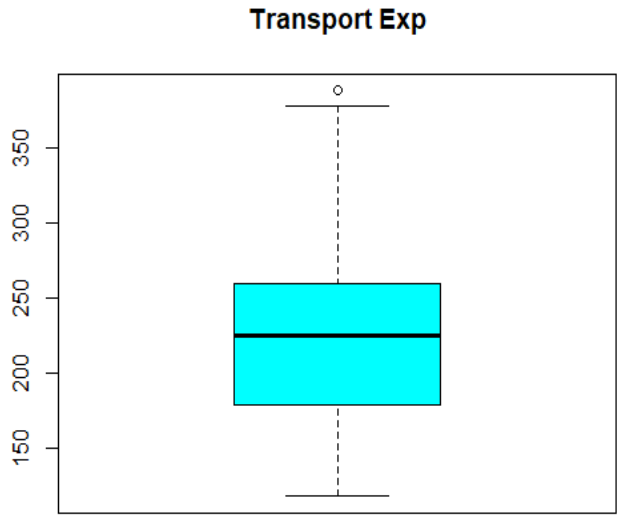


Fig 5. Analysis of outliers using boxplots



3.3 Feature scaling

The features in our dataset have different measurement units and their magnitude varies widely, it is therefore important that the features are scaled down to a fixed range to enable the machine learning algorithm objective function to work properly. For a regression model, the objective function is to optimize the convergence of the gradient descent at some global minima (Naik, 2022) and to help achieve this, individual numerical features in the dataset were normalized to a value between 0 and 1 using a Min-Max scaler.

3.4 Feature Selection

Machine learning models are faced with the challenge of having to deal with a large number of input features, therefore, making it difficult for the models to make accurate predictions. Feature selection refers a method in machine learning for identifying data with relevant attributes for model building by eliminating irrelevant, redundant, or highly correlated data without losing much information. Implementing feature selection helps in better understanding of the data, reducing computational requirements, easier interpretation, and increases generalization by reducing the variance(Muthukrishnan and Rohini, 2016).

Least Absolute Shrinkage and Selection Operator (LASSO, or Lasso) regression shrinks variable coefficients, and variables with zero coefficients are eliminated, it also performs feature selection. Ridge regression work by shrinking variable coefficients to prevent overfitting, both methods work by shrinking the dimension of the data by making the estimated regression coefficient close to zero. Ridge regression reduces the squared sum of coefficients (L2 regularization) while LASSO reduces the absolute sum of the coefficients (L1 regularizations) (Muthukrishnan and Rohini, 2016)

Elastic net mixes the attributes of ridge and lasso regression models. These regression models have inherent feature selection capabilities and will be implemented for this project.

3.5 Model Justification

The aim of implementing these models is to ultimately adopt upon comparison a final model that is parsimonious, easy to interpret, and accurately predicts future data. The lasso model picks a subset of features that improves the interpretability of the multiple linear regression model. Ridge and lasso models and by extension the elastic net, work to eliminate multicollinearity problems which can adversely impact the analysis in general and severely limit the conclusions that could be drawn from the project(Çiftsüren and Akkol, 2018).

3.6 Model Building

The aforementioned regression models capable of predicting the target feature were developed after a comprehensive data pre-processing.

Multiple Linear Regression Model

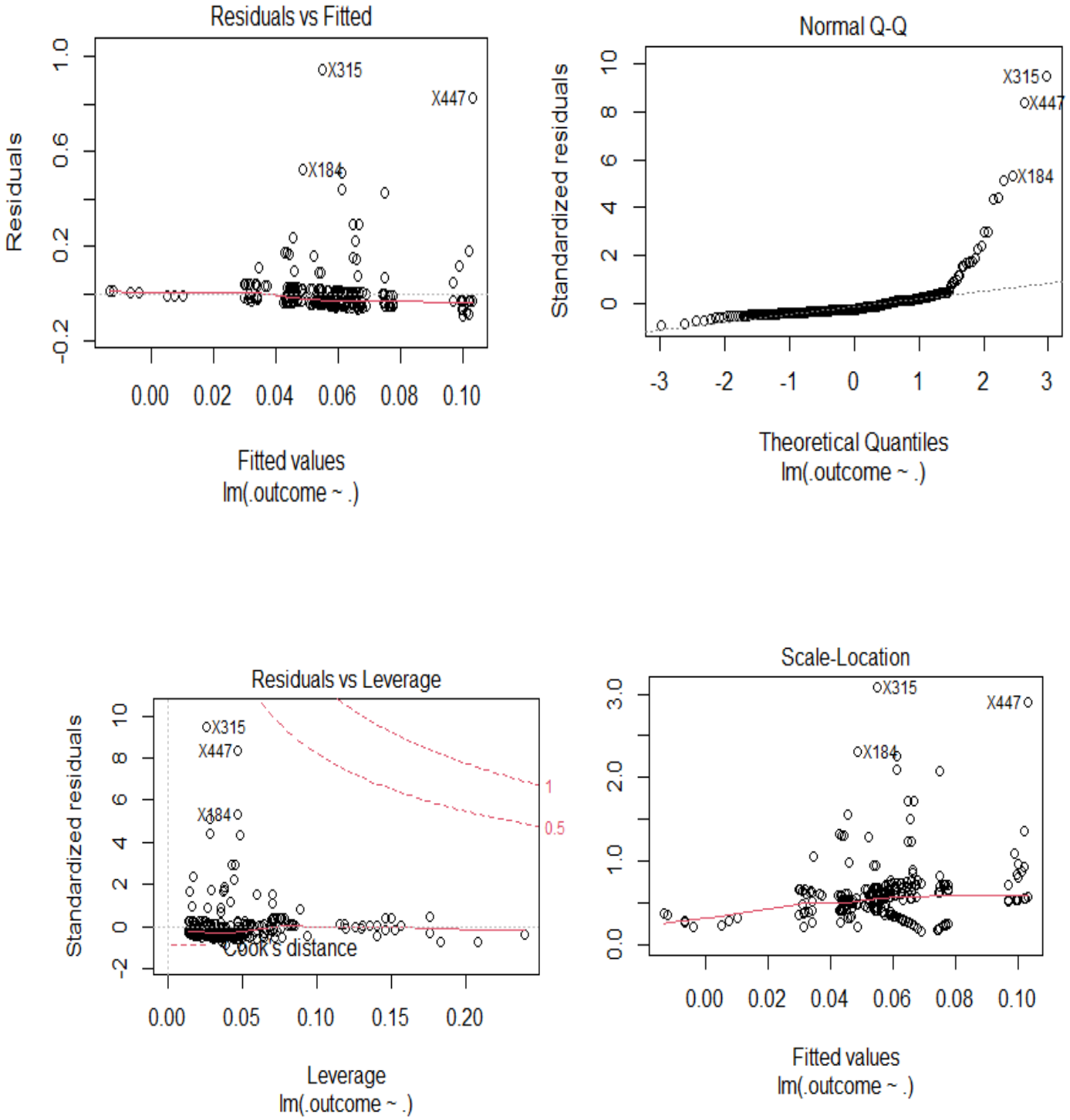
Multiple Linear regression is a form of linear regression that permits the use of two or more independent features, it uses the same theory as linear regression and makes some assumptions on the relationship between the dependent and independent features, and on the distribution of the independent features. These assumptions are;

i. A linear relationship between independent and dependent features
ii. Independent features have normal(Gaussian) distributions and are not correlated.

To train the model in R, the method “lm” (linear model) was chosen and applied to all the independent features in the dataset using a single line of code. 344 samples and 14 predictors were used to train the model, and resampling was done 10-fold and repeated 5 times using K-fold cross-validation. The model showed that the feature “Disciplinary. Failure” was more statistically significant compared to others in reaching this result. A summary of the model evaluation is tabulated below.

MAE	RMSE	R-Squared
0.04781804	0.09148503	0.02789485

Fig 6. Plots Multiple Linear Regression Model Residuals



Ridge Regression Model

The Ridge regression model is also known as L2 regularization, acts as an extended linear regression model by adding a penalty (squared magnitude of the coefficients) to the loss function multiplied by lambda (R & R, 2016).

$$L = \sum(\hat{Y}i - Yi)^2 + \lambda \sum \beta^2$$

This model works by attempting to shrink the coefficients of all the features in the model without eliminating any of the features irrespective of their importance in predicting the target variable.

To build the Ridge model in R, the package “glmnet” was used and applied to all the independent features in the dataset after the variables were converted into matrices.

Alpha for ridge regression is chosen as zero and determines the weight to be used while lambda is a hyper-parameter estimated using the specified cross- validation and determines the strength of the penalty to be applied to the coefficients. Lambda was chosen as a sequence between 0.0001-1 with a length of 5. 344 samples and 14 predictors were used to train the model, and resampling was done 10-fold and repeated 5 times using cross-validation. A summary of the resampling results across tuning parameters for the model evaluation is tabulated below.

lambda	MAE	RMSE	R-Squared
0.000100	0.04789289	0.09199562	0.02777537
0.250075	0.04668458	0.09002399	0.03111771
0.500050	0.04697209	0.08999440	0.03176539
0.750025	0.04708608	0.08999343	0.03203423
1.000000	0.04714706	0.08999523	0.03218070

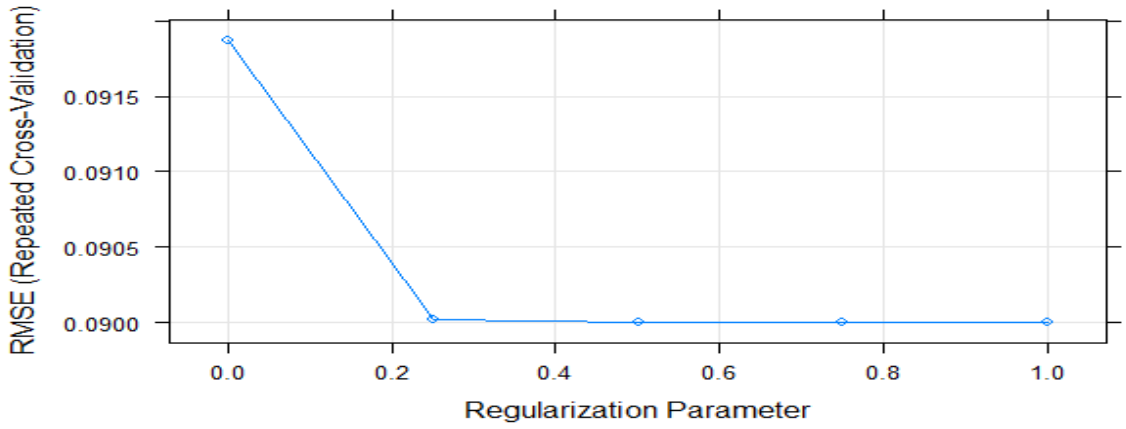


Fig 7. Plot of RMSE Vs Lambda

The Root Mean Square Error (RMSE) was the deciding factor in selecting the optimal model using the least value and alpha kept constant at a value = 0.

The Selected tuning parameters

Fitting alpha = 0, lambda = 0.75 on full training set

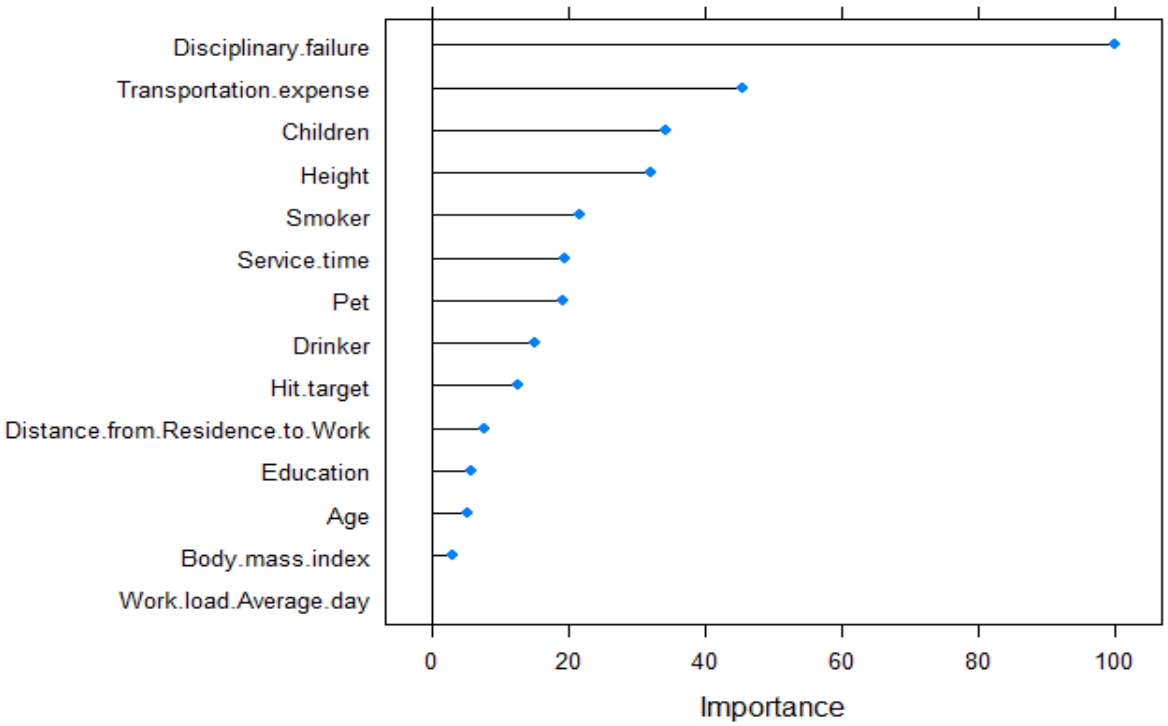


Fig 8. Plot of Variable Importance

The plot of variable importance shows the order of importance for the independent features in predicting the target variable. Disciplinary failure is the most important variable followed by transportation expense.

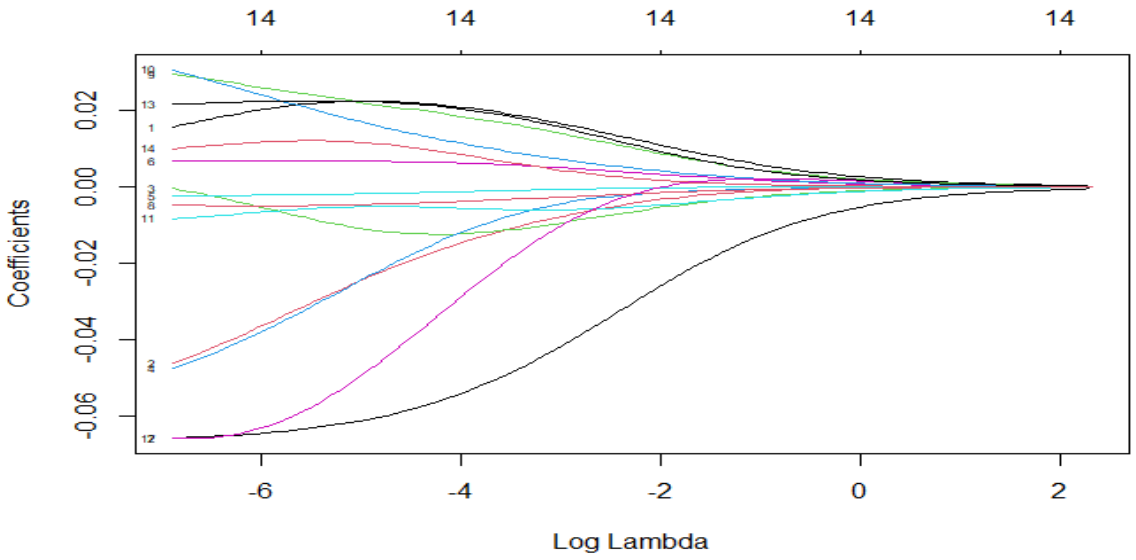


Fig 9. Plot of Coefficients Vs Log Lambda

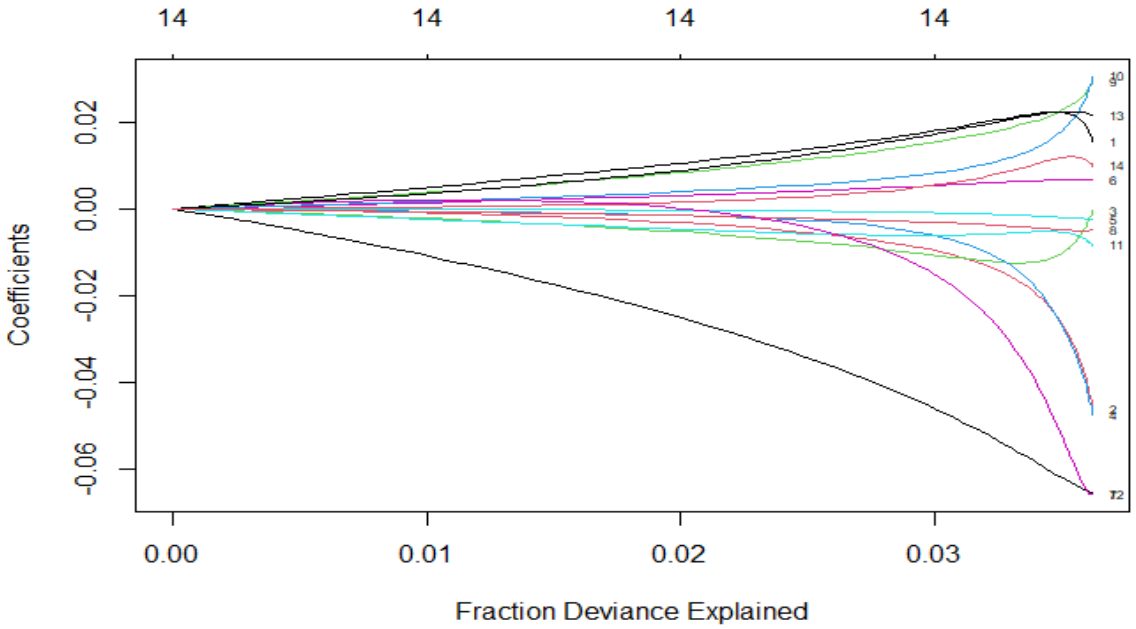


Fig 10. Plot of Coefficients Vs Fraction Deviance Explained

Lasso Regression Model

The Least Absolute Shrinkage and Selection Operator (Lasso) modifies the loss function in the linear regression model by assigning a penalty to the absolute values of the model coefficients.

$$L = \sum(\hat{Y}i - Yi)^2 + \lambda \sum |\beta|$$

The penalty reduces the value of some coefficients to zero and subsequently eliminates variables with zero coefficients in a process known as feature shrinkage. One of the advantages of using Lasso regression is that it does both shrinkage and feature selection (R & R, 2016).

To build the Lasso model in R, alpha was set to 1, and lambda was chosen as a sequence of values between 0.0001 -1 with length set to 5. The method “glmnet” was used and applied to all the independent features in the dataset.

344 samples and 14 predictors were used to train the model, and resampling was done 10-fold and repeated 5 times using cross-validation. A summary of the resampling results across tuning parameters for the model evaluation is tabulated below

lambda	MAE	RMSE	Rsquared
0.000100	0.04766109	0.09370217	0.02754032
0.250075	0.04731141	0.09189828	NaN
0.500050	0.04731141	0.09189828	NaN
0.750025	0.04731141	0.09189828	NaN
1.000000	0.04731141	0.09189828	NaN

The Root Mean Square Error (RMSE) was the deciding factor in selecting the optimal model using the least value and alpha kept constant at a value = 1.

The Selected tuning parameters

Fitting alpha = 1, lambda = 1 on full training set

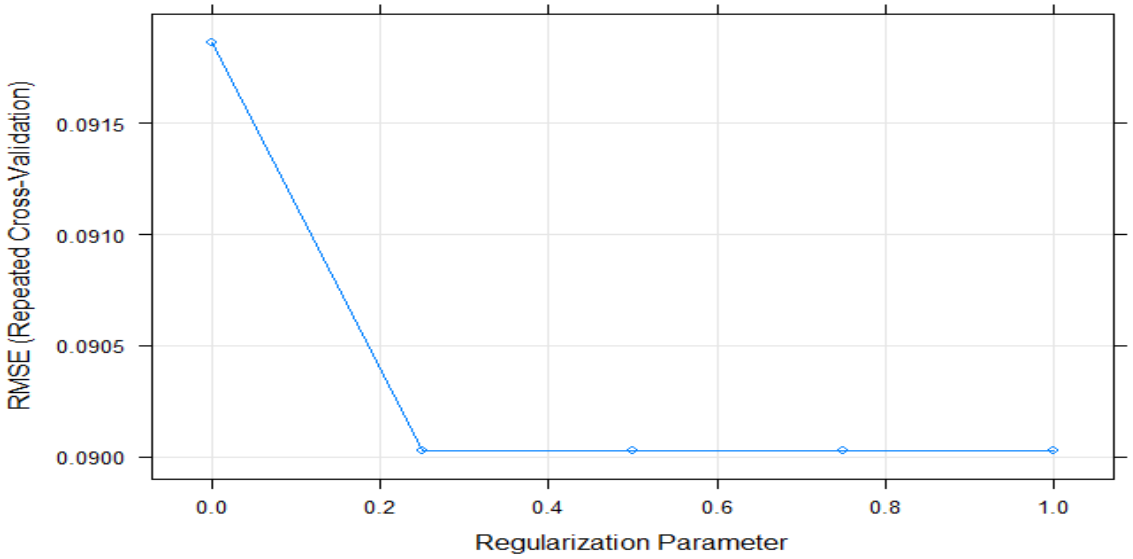


Fig. 11 Plot of RMSE Vs Lambda

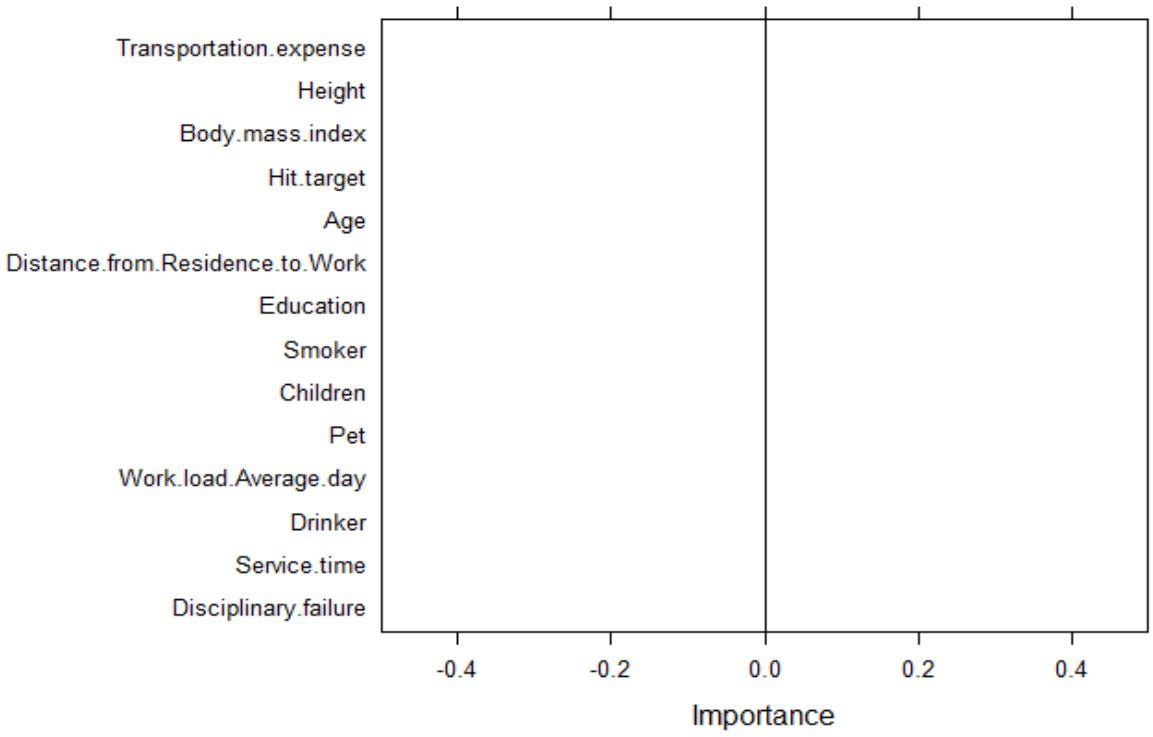


Fig 12 Plot of Variable Importance

From the plot it can be seen that no feature was chosen as having better importance than the others in predicting the target variable, all the features lie at the zero importance mark.

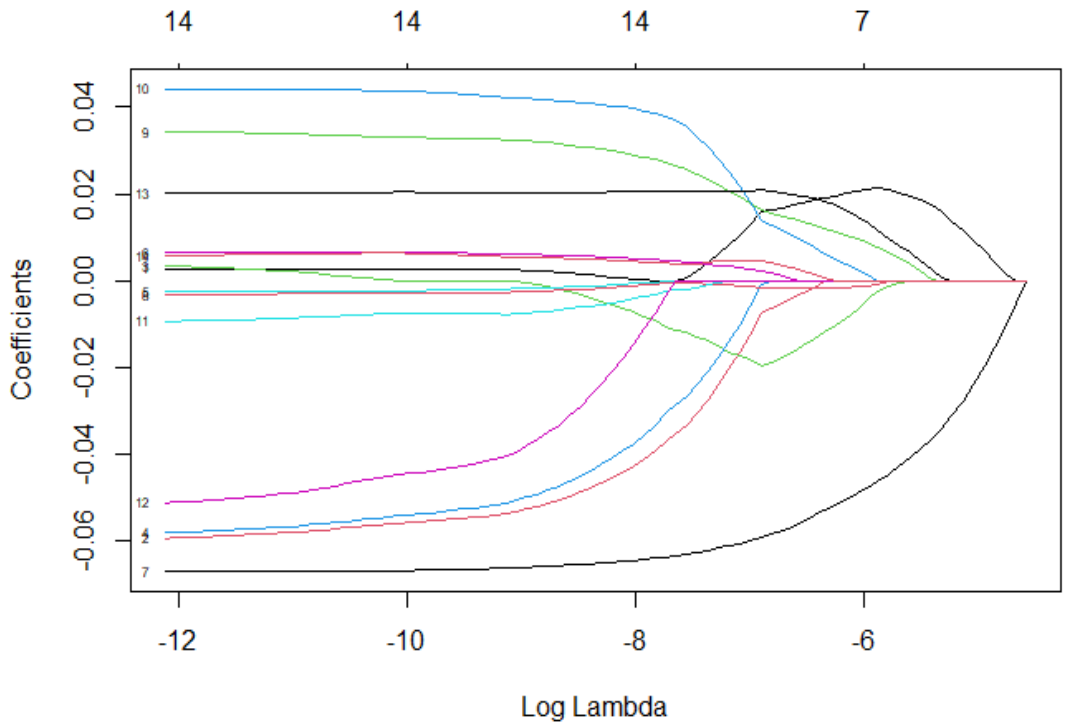


Fig 13. Plot of Coefficients Vs Log Lambda

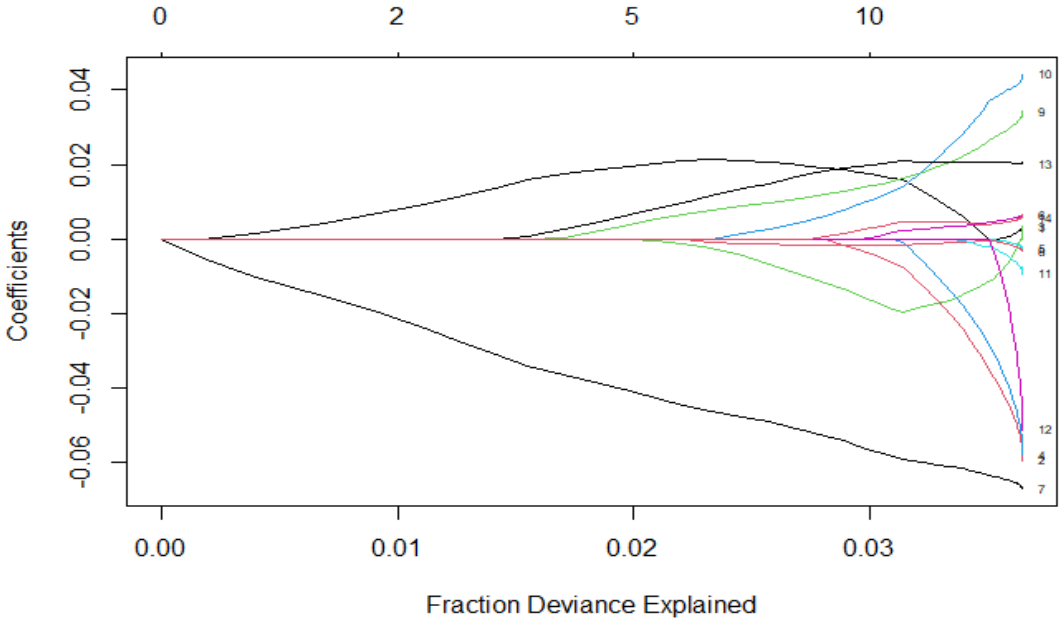


Fig 14. Plot of Coefficients Vs Fraction Deviance Explained

Elastic Net Regression Model

The Elastic Net regressions model combines the attributes of lasso (L1) and ridge (L2) regression. The L1 penalty reduces the weights of all the coefficients while allowing some to be reduced to zero. The L2 penalty on the other hand reduces the weight of the coefficients but does not eliminate any of the features from the model (R & R, 2016).

$$L = \sum(\hat{Y}i - Yi)^2 + \lambda \sum \beta^2 + \lambda \sum |\beta|$$

The model was trained in R using the "glmnet" package. A sequence of values between 0-1 is chosen for alpha in order to get an optimal alpha value for the model and lambda was set to .0001 to 0.2

344 samples and 14 predictors were used to train the model, resampling was done 10-fold and repeated 5 times using cross-validation. The best model selected was determined by the RMSE with the least value. Alpha = 0.111, lambda = 0.0501 were the final values adopted for the model.

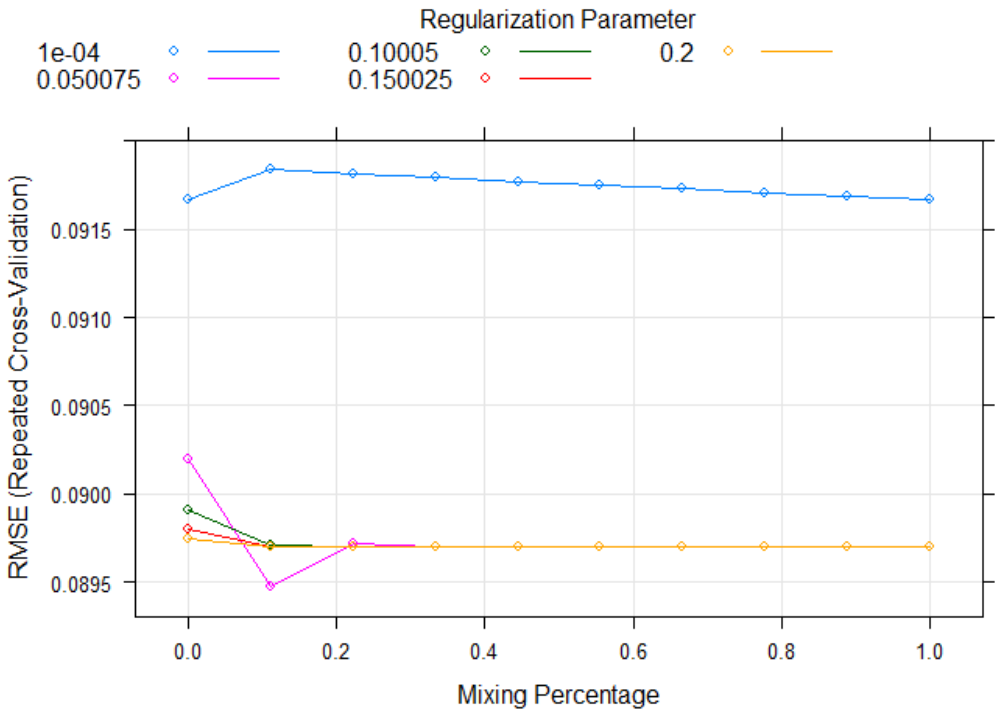
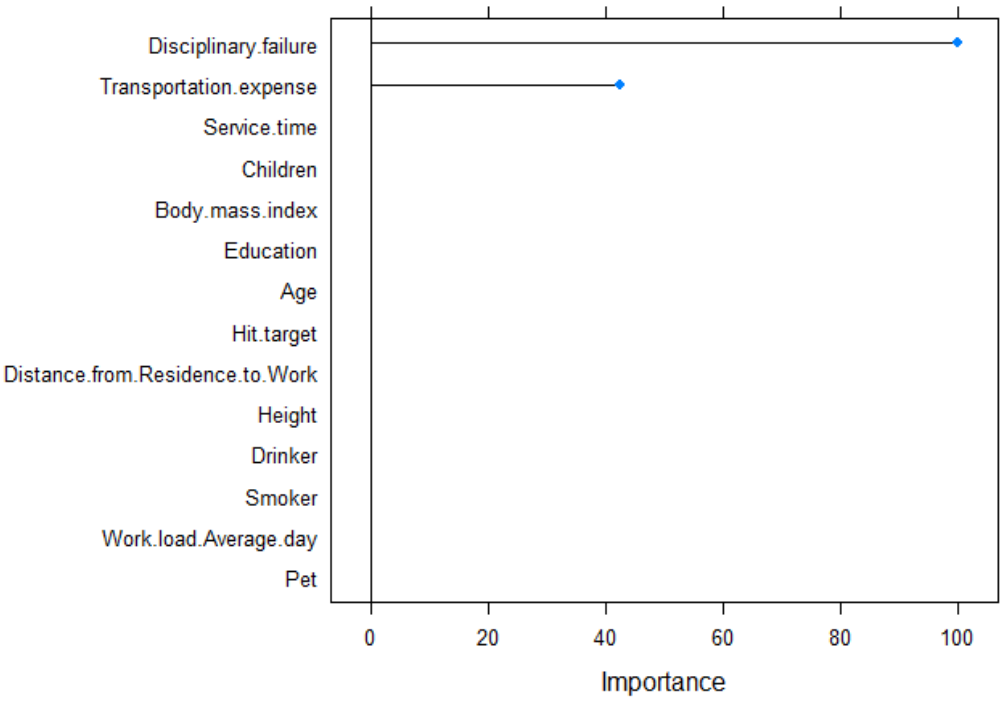


Fig 15. Plot of RMSE Vs Mixing percentages

Fig 16. Plot of Variable Importance



The plot of variable importance shows that the Elastic Net model chose Disciplinary failure and Transportation expense are the important variables while setting other variable coefficients to zero without eliminating them.

Chapter 4. Conclusion

This chapter reviews the performance of the regression models using common regression evaluation metrics such as the Mean Absolute Error(MAE), Root Mean Square Error(RMSE), and R-squared to select the model with the best result.

4.1 Model Evaluation

Regression evaluation metrics such as MAE, RMSE, and R-squared were used in the previous chapter to confirm the suitability of the trained models in terms of performance. The Mean Absolute Error (MSE) is a simple metric that calculates the difference between the actual and predicted values in absolute terms, the aim is to select the model with the smallest value. This metric is also widely used because it is robust to outliers.

The Root Mean Square Error (RMSE) measures the spread of the residuals, the residuals show the distances of the data points from the regression line. RMSE is the standard deviation of the residual, the objective is to select the model with a lower RMSE as a lower value for this metric is an indication of the level of fitness of the data to the model measured in absolute terms.

R-Squared (coefficient of determination) measures the amount of variance in the target variable that can be defined by the independent variables, it is also a measure of “goodness of fit” as it shows how well the data fit the regression model (in percentage terms). The aim is to select a model with a higher value of R2 as it indicates more variability explained by the model.

4.2 Model Selection

The summary of the evaluation metrics of the models is shown in the table below. Model comparison was achieved using the res() function under the “caret” package in R (Rai, 2018).

Model Evaluation Metrics

Models	MAE	RMSE	R-Squared
Linear Model	0.04769953	0.09207278	0.2564521
Ridge	0.04687375	0.08999731	0.2278161
Lasso	0.04728653	0.09002749	NA
Elastic Net	0.04661480	0.08947064	0.1604999

The Elastic Net model performed better on the training dataset as it has a lower value for both MAE and RMSE, however, the R-Squared value for the Elastic Net model is lower compared to the three other models. The R-Squared value does not indicate the correctness of a model and should not be used alone in assessing models' performance but in conjunction with other metrics.

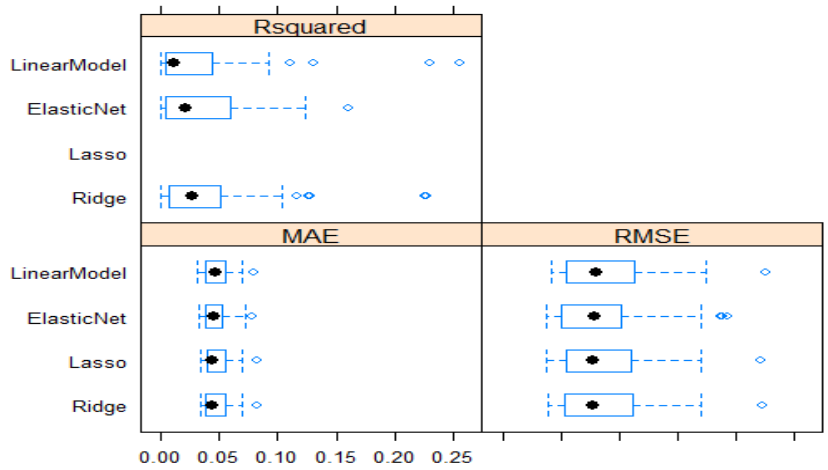


Fig 17. Box &Whisker Plot of R-Squared, MAE and RMSE

The plot above gives a visual summary of the models, it supports the argument that the elastic net model performed better than the other models, for R-Squared and RMSE we can see the presence of more outliers further away from the maximum point, MAE is lower for the elastic net model, the same as RMSE

The best tuning parameter for the elastic net model is (alpha = 0.1111111, lambda= 0.050075) this indicates that the elastic net model is closer to a ridge model than a lasso model. The coefficients of the variable of importance for the model are shown below.

```
15 x 1 sparse Matrix of class "dgCMatrix"
                                          s1
(Intercept)                     0.052697366
Transportation.expense          0.008467831
Distance.from.Residence.to.Work .
Service.time                    .
Age                             .
Work.load.Average.day           .
Hit.target                      .
Disciplinary.failure           -0.019969272
Education                       .
Children                        .
Drinker                         .
Smoker                          .
Pet                             .
Height                          .
Body.mass.index
```

An earlier study found that in terms of prediction accuracy and model selection consistency, the elastic net repeatedly outperforms ridge and lasso regressions (Zou & Hastie, 2005). The elastic net model has also been successfully applied in the following areas, portfolio optimization (Shen, et al., 2014), cancer prognosis (Milanez-Almeida, et al., 2020), and doubly regulated Support Vector Machine which uses a mixture of L1 and L2 penalties (Wang, et al., 2006)

4.3 Model Prediction and Accuracy Estimation

Upon successful model comparison and selection, the final selected model which was earlier used to make a prediction (P1), in-sample on the training data was again deployed to make a prediction(P2), out-of-sample, using the test data.

The predictions P1 and P2 are used to calculate the error metric. The RMSE was selected as the best error metric for estimating the model accuracy. It measures how accurately the model predicts the target variable and indicates the absolute fit of the model to the data. Lower values of RMSE signify a better fit.

P1= Final Model (Selected Model) prediction on training data

RMSE value for P1 = **0.1000334**

P2= Final Model (Selected Model) prediction on test data

RMSE value for P2 = **0.05356039**

The RMSE figures above show that our model did relatively well on both the training and test data, a lower RMSE value for P2 indicates that the model performed better on the test data, an indication of a better prediction accuracy.

4.4 Project Evaluation

Some of the key ways of assessing the success of any project are by determining if, the goal and objectives set out at the beginning of the project were subsequently achieved, meeting the Key Performance Indicator(KPI) at every phase of the project, project deployment and the ability of the project to deliver key actionable insights or optimization suggestions that would enable the business to perform better in the future. This project was initialized by firstly drawing up well-articulated and unambiguous business objectives that are measurable. The following key performance indicators, data collection, data processing and analysis, model evaluation and building, and project report were comprehensively performed and met satisfactorily.

The exploratory analyses of the data provided relevant insights into the characteristics of the data, for example, the relationship between the distance from residence to work and transportation expense was not linear as would have been expected, an indication that employees used a different mode of transportation to work with the assumption that the same shifts (day or night) were worked. The presence of outliers was discovered, some of which could be explained as an error in either data entry, measurement, or experiment. While performing multivariate exploratory data analysis we learned about the presence of independent covariates in the dataset leading to multi-collinearity problem, this was however addressed by the feature selection technique adopted.

Final model selection was done after reviewing the performance of the four regression models using appropriate regression metrics. The final selected model, the elastic net

model revealed that only two predictor variables, transport expense and disciplinary failure are important in predicting the target variable, with these findings, it would be advised that the business implements a disciplinary mechanism in the workplace and also provide transportation incentives such as staff bus, staff car loan or transport allowance as these would reduce the rate of absenteeism and make the business more profitable.

References

aniruddha, 2020. *Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-

normalization-standardization/
[Accessed 15 March 2022].

Çiftsüren, M. N. & Akkol, S., 2018. Prediction of internal egg quality characteristics and variable selection using regularization methods: ridge, LASSO and elastic net. *Archives Animal Breeding,* 61(3), pp. 279-284.

Cucchiella, F., Gastaldi, M. & Ranieri, L., 2014. Managing absenteeism in the workplace: the case of an Italian multiutility company. *Procedia-Social and Behavioral Sciences,* Volume 150, pp. 1157-1166.

Graham, J. W., 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology,* Volume 60, pp. 549-576.

Halpern, M. T., Shikiar, R., Rentz, A. M. & Khan, Z. M., 2001. Impact of smoking status on workplace absenteeism and productivity. *Tobacco Control,* 10(3), p. 233.

Kassambara, A., n.d. *Correlation matrix : A quick start guide to analyze, format and visualize a correlation matrix using R software.* [Online]
Available at: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software#what-is-correlation-matrix
[Accessed 18 March 2022].

McFarlin, S. K. & Fals-Stewart, W., 2002. Workplace absenteeism and alcohol use: a sequential analysis. *Psychology of Addictive Behaviors,* 16(1), p. 17.

Milanez-Almeida, P., Martins, A. J., Germain, R. N. & Tsang, J. S., 2020. Cancer prognosis with shallow tumor RNA sequencing. *Nature Medicine,* 26(2), pp. 188-192.

Naik, K., 2022. *Live EDA and Feature Engineering playlist.* s.l.:s.n.

Raghunathan, T., 2015. *Missing data analysis in practice.* s.l.:CRC press.

Rai, B., 2018. *Ridge, Lasso & Elastic Net Regression with R | Boston Housing Data Example, Steps & Interpretation.* s.l.:s.n.

R, M. & R, R., 2016. *LASSO: A feature selection technique in predictive modeling for machine learning.* s.l., s.n., pp. 18-20.

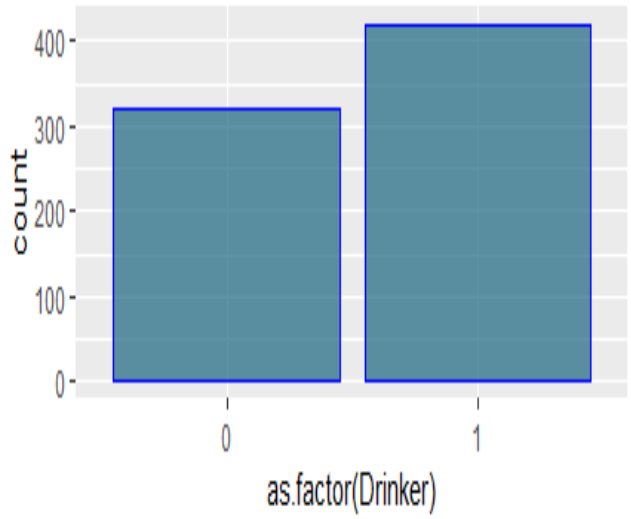
Shen, W., Wang, J. & Ma, S., 2014. *Doubly Regularized Portfolio with Risk Minimization.* s.l., s.n.

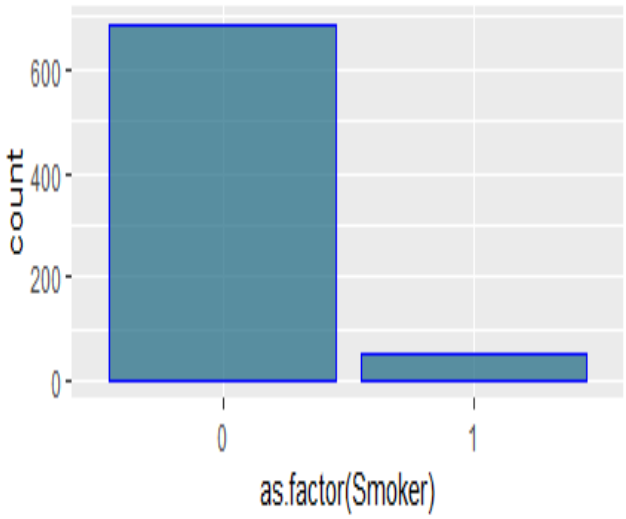
United States Dept. of Labor, 2022. *Labor Force Statistics from the Current Population Survey.* [Online]
Available at: https://www.bls.gov/cps/cpsaat47.htm#cps_eeann_abs_ft_occu_ind.f.1
[Accessed 19 May 2022].

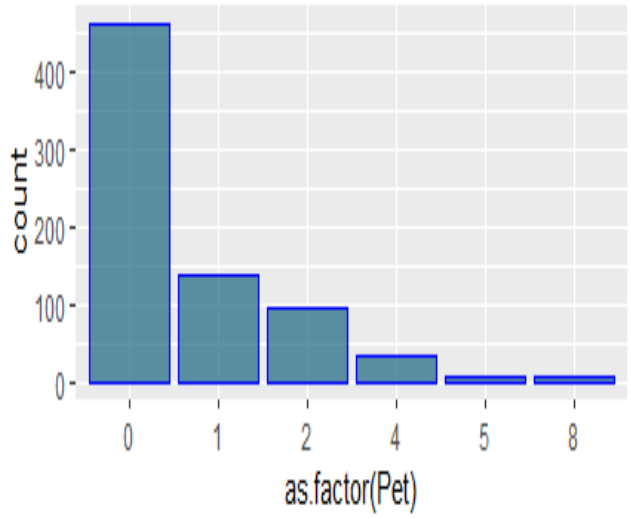
Wang, L., Zhu, . J. & Zou, H., 2006. The doubly regularized support vector machine. *Statistica Sinica,* Issue 1017-0405, pp. 589-615.

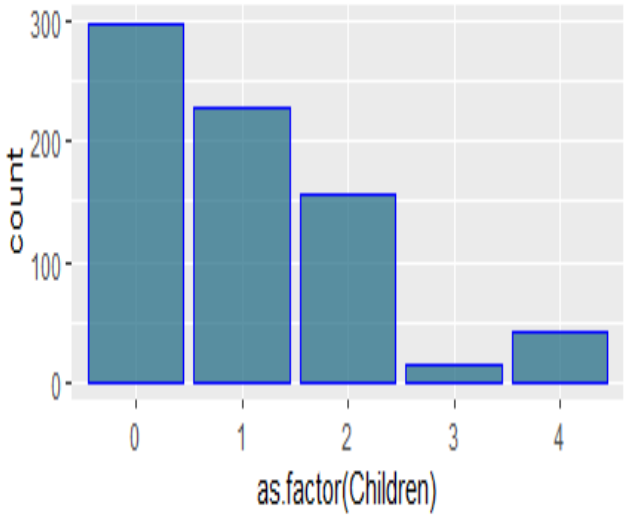
Zou, H. & Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology),* 67(1369-7412), pp. 301-320.

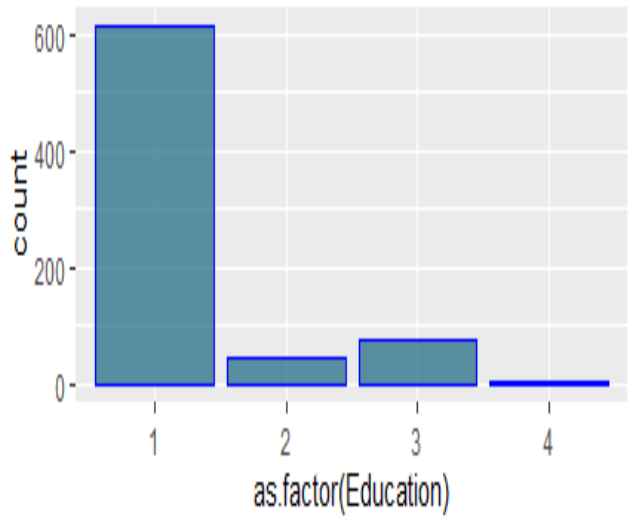
Appendix 1. Bar plots of Categorical variables

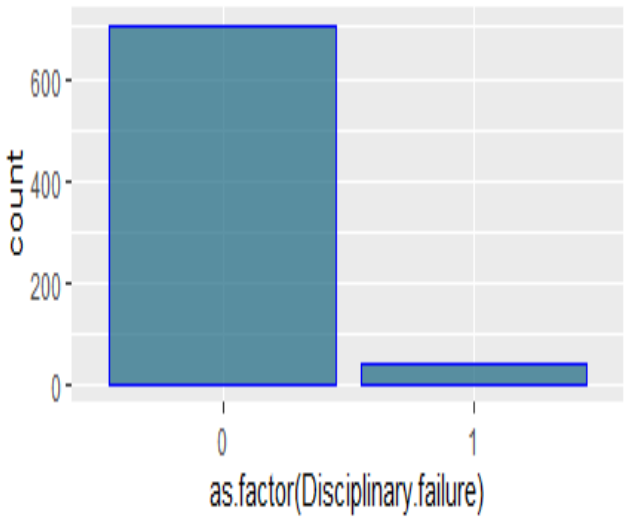












Appendix 2. R codes and Pre-processed data

OpenDocument
Text

R code used in implementing model 1

OpenDocument
Text

Code Exploratory Data Analysis 1

OpenDocument
Spreadsheet

Pre-processed Data 1