MATH42715: Machine Learning
Assignment No. 2

Exploring and Predicting Absenteeism using the Absenteeism Dataset and Machine Learning Models

Executive Summary

Employee absenteeism has a tremendous impact on the workplace as well as the employee. It takes place due to numerous reasons, and the degree of absenteeism i.e. the number of hours away from work, has significant ramifications for the productivity of teams within the workplace. This report examines the Absenteeism at work dataset to evaluate the degree to which machine learning approaches can provide insight into how different factors influence absenteeism. This data set was collected over a period of three years (2017-2010) from records of absenteeism at a courier company in Brazil and includes data on when the absence took place (Season, Month, Day), the reasons for absence, the no. of hours of absenteeism, as well as few personal employee attributes.

In this report, several models (specifically, an Unpruned and Pruned Regression Tree, Bagging and RandomForest Model, a Neural Network, and a Generalised Additive Model) were deployed to, on one hand, gain a better understanding of the relationships between variables and how it influences the number of hours of absenteeism, and on the other, to consider different machine learning models and their ability to accurately predict the number of hours of absenteeism. This may provide managers (as well as employees) a better understanding of the incidence of absenteeism, whether work needs to be re-assigned, or whether internal work protocols need to be adjusted.

The unpruned regression tree was constructed using 8 variables while the pruned regression tree included only 3 nodes. The bagging and RandomForest model, on the other hand, was fitted with 4 variables and 1500 trees (as this proved to improve predictive performance) while the neural network was a simple model that adjusted its weights and biases during the training process by utilising 50 neurons through which the data was fed. The final GAM model was constructed with smoothing splines and only 9 variables. The Generalised Additive Model scored the best predictive performance, followed closely by the Bagging and RandomForest model. The Unpruned and Pruned regression trees had lower predictive performance while the neural network had the lowest.

Each of the modelling techniques adopted provided insight into the dataset as well. For example, there is a tendency for a higher hours of absenteeism when employees are afflicted with illnesses (see Reasons 1-21), and significantly lower number of absentee hours for non-illness related reasons (including medical consultations). This signals the need for clear protocols for both employers and employees of how to record an absence on the grounds of illness, how line managers are to address them, and how long-term illnesses are to be managed and supported to curb significant operational costs. High absenteeism in this regard may also be due to limited workplace support (for example, disability support), and should be investigated further. Similarly, the GAM showed that the number of absent hours increase after a particular age (between 45-50 years), while Thursday appears to be the day with the lowest number of absentee hours. Similarly, as the no. of children increases, absenteeism appears to increase, indicating that parental responsibilities may impact how employees approach their work. These findings may warrant further analysis and briefing of line managers, and embedding of work-life balance approaches to reduce absenteeism.

1. Introduction

1.1. Area of Inquiry: Absenteeism in the Workplace

Employee absenteeism can have a significant impact on the workplace, as well as the employee's productivity. On one hand, a high incidence of absenteeism can significantly affect not only the efficiency of the workplace but its continued function as employees are responsible for specific tasks. Similarly, absenteeism can affect an employee's ability to approach and complete their assigned tasks effectively, and further impact their career progress.

In this light, it is important to understand what leads to employee absenteeism and its incidence. Such an approach may provide those that manage employees better information to assess how long an employee maybe absent for based on a series of factors, accommodate their absences in light of pending work targets, and if the no. of hours of absenteeism is too high, re-allocate their work to other divisions. This can also assist an employee as it provides them with some support from the workplace when it comes to mitigating the impact of absenteeism on their work.

1.2. Dataset: "Absenteeism at work"

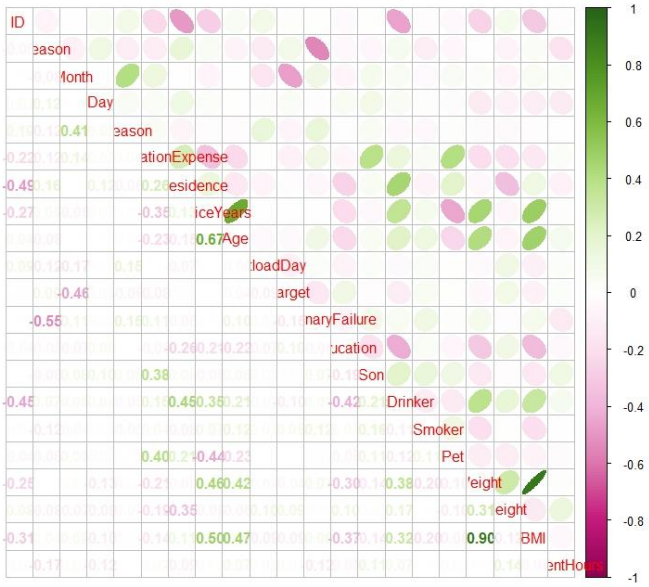
In order to investigate the incidence of absenteeism, this exercise utilises the "Absenteeism at work" dataset available on the UCI Machine Learning Repository. This dataset was created by Andrea Martiniano, Ricardo Pinto Ferreira, and Renato Jose Sassi by utilising records of absenteeism at work at a courier company in Brazil. The data was collected between July 2017 and July 2010. The dataset has 740 observations and 21 variables which are both continuous and categorical as follows; (01) The ID column identifies each employee individually while the (02) Reason for absence is a categorical variable with levels 0-28. Level 0 is where there has been no absence. The levels 1 to 21 relate to absences attested by the International Code of Diseases (ICD). These absences have been stratified into 21 categories as follows: (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 behavioural 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. The remaining categories 22 to 28 relate to absences that do not relate to the ICD: (22) patient follow-up, (23) medical consultation, (24) blood donation, (25) laboratory examination, (26) unjustified absence, (27) physiotherapy, (28) dental consultation. The other variables include (03) month of absence, (04) the day of the week (with 2-6 levels corresponding to Monday-Friday), (05) Seasons (with 4 levels corresponding to Summer-Spring), (06) Transport expense, (07) Distance from residence to work (in kilometres), (08) Service time, (09) Age, (10) Workload Average per day, (11) Hit Target, (12) Disciplinary failure (1 if yes, 0 if no), (13) Education (1=high school, 2=graduate, 3=postgraduate, 4=masters and/or doctorate), (14) Son (no. of children), (15) Social drinker (1=yes, 0=no), (16) Social smoker (1=yes, 0=no), (17) Pet (no. of pets), (18) Weight, (19) Height, (20) Body mass index, and (21) Absenteeism time in hours.

1.3. Brief Exploratory Data Analysis and Data Pre-processing

The dataset has a variety of predictors ranging from time (Season, Month, Day) to personal attributes (Age, Height, Weight, BMI) and health choices (Social Drinker, Social Smoker). It also provides data on their workplace behaviours (Hit Target, Disciplinary Failure) as well as practical considerations related to travel and related expenses. Most significantly, this dataset provides detailed information regarding the reasons for absence from illness-related to non-illness related reasons. Accordingly, the absenteeism in hours in the workplace variable which records the no. of

hours an employees is absent, can be selected as the target variable. The question that is explored is two-fold: what factors contribute to the varying degrees of absenteeism in the workplace, and to what degree can the number of hours an employee would be absent depending on the other predictors be predicted? In order to explore the relationships between the target (AbsentHours) and predictor variables, the following models were deployed: (1) Unpruned Regression Tree, (2) Pruned Regression Tree, (3) Bagging and Random Forest Model, (4) Neural Network, and (5) Generalised Additive Model. However, a word of caution is pertinent at this stage. Absenteeism at work is a complex issue as employee behaviour cannot be strictly understood in a linear sense: the complexity of human behaviour makes predicting the degree to which an employee would be absent a complex proposition. Therefore, the insights gained into its causes and effects need to be combined with a behavioural approach for the best impact.

Figure 1.
Correlation Panel: Correlation Coefficients of **Absenteeism** Variables



In order to provide an overview of how these variables relate to each other, the correlations of each of the variables were plotted. In figure 1, the correlation coefficients, and the degree to which the variables are (or are not) correlated to each other can be seen. In this plot, it is clear that Weight and Body mass index are highly correlated (0.9). Unsurprisingly Age and Service Years are also somewhat correlated (0.67) given that as an employee gets older, they are able to continue in service for longer if they choose to remain with the courier company. There are both positive (e.g. Service years and Weight) and negative relationships (e.g. Season and DisciplinaryFailure) between variables

The original dataset contains 740 observations and 21 variables. For convenience of analysis, the dataset titled "Absenteeism_at_work" was renamed "Absenteeism" and the variables (with spaces) renamed to "Reason", "Month", "Day", "DistResidence", "ServiceYears", "Season", "TransportationExpense", "WorkloadDay", "Target", "DisciplinaryFailure", "SocDrinker", "SocSmoker", "BMI", and "AbsentHours" respectively. Thereafter, the dataset was checked and it was confirmed that there were no missing values in the dataset. In order to facilitate the implementation of the models and data analysis, the following pre-processing steps were also taken.

- The ID column identifies the employee whose information is recorded. There are 36 employees. As the individual employees' absenteeism is not under review, the information in this column is irrelevant. Therefore, the "ID" column was removed.
- The correlation plot showed that the variables "BMI" and "Weight" are highly (positively) correlated (0.9). This is unsurprising as the Body Mass Index (BMI) is calculated using the height and weight of a person. It is well known that one of the assumptions of linear regression is the absence of two independent variables being highly correlated to the other (multicollinearity). In this instance, a Generalised Additive Model (GAM), two decision tree models, and a neural work is implemented. Decision trees are not affected by a relationship between independent variables, and neural networks are arguably less sensitive to multicollinearity. However, a relationship between variables may be problematic when modelling a GAM. Indeed, it has been argued that the presence of multicollinearity may result in nonlinear relationships between those variables in the GAM; a phenomenon termed "concurvity" (Amodio et al, 2014).[1] This, in turn, can result in the estimated coefficients in the GAM to be unstable. As a comparative assessment of the predictive

[1] Amodio, S., Aria, M., & D'Ambrosio, A. (2014). On concurvity in nonlinear and nonparametric regression models. *Statistica, 74*(1), 85–98.

performance of each model is considered, it is necessary to make this comparison as fair as possible. Therefore, the variable "BMI" was removed (and "Weight" retained). This decision was also influenced by the fact that BMI is calculated using both height and weight. The loss of data caused by its removal is minimised as both the variables "Height" and "Weight" remain in the dataset. Though not affected by multicollinearity, its removal may also assist in simplifying the decision trees and improving interpretability.

- The summary() output shows that the lowest value in "Month" is 0. Closer examination shows that there are 3 rows where the month is zero. The remaining levels in the column are from 1 – 12 indicating months January to December. It is likely that the 3 rows where month is zero is inaccurately recorded, and therefore, the relevant rows were removed.
- The dataset contained character and numeric data. In order to ensure uniformity and facilitate data analysis, all data were converted to numeric data.

After data pre-processing and cleaning was implemented, the dataset contains 19 variables and 737 observations.

2. Modelling and Analysis

2.1. Fitting a Regression Tree

In this dataset, the target variables is the no. of absentee hours. Therefore, the output is to predict the no. of hours absent (as opposed to classifying the absenteeism as being high or low, for example), and thus, a regression tree is required because the target variable is continuous. A regression tree has the advantage of being highly interpretable, and it also provides a clear overview of which variables were used and at what points they were split in order to predict the output. Because of these reasons, a regression was selected and fitted to the data.

2.1.1. Fitting an Unpruned Regression Tree

The dataset is fit to a regression tree using the tree() function (and "AbsentHours" is excluded). The regression tree used only 8 variables (Reason for absence, age, height, whether the employee is a social drinker, distance to work from the residence, season, workload, and target) to construct the tree. It has 11 leaves (terminal nodes) i.e. the final predictions it makes in terms of the variables. The sum of squared errors for the tree (i.e. residual mean deviance) is 10.4.

The training residual mean deviance (which shows the degree of error remaining after the regression tree is constructed i.e. the model's goodness of fit) is 85.96. However, this does not provide much insight into the performance of the regression tree model on unseen data. Therefore, to compute the test error, the data was divided on a 2/3rd split (490 observations for the training set; 247 observations for the test set). Thereafter, a tree was fitted using the training data, and its performance evaluated by using the test data. The test MSE is much higher than the training error at 218.2868. Therefore, the single regression tree performs better on the training set (due to likely overfitting the data), and demonstrates poorer test set performance. This is possible due to the tree being too complex.

2.1.2. Fitting a Pruned Regression Tree

A pruned tree may demonstrate a lower test error rate and better predictive performance. This is because pruning removes variables (branches) that may not contribute significantly to prediction accuracy. Therefore, next a pruned regression tree was fitted to the data. In order to prune the tree i.e. to determine the complexity of the tree, the cv.tree function was used as it runs a k-fold cross-validation to find the deviance as a function of the no. of folds of the cross-validation. One of the benefits of using k-fold cross-validation is that it facilitates the use of the data to its full potential by using it to both train and test the data. Both plots in figure 2 show that the lowest cv-error is when there are 3 nodes.

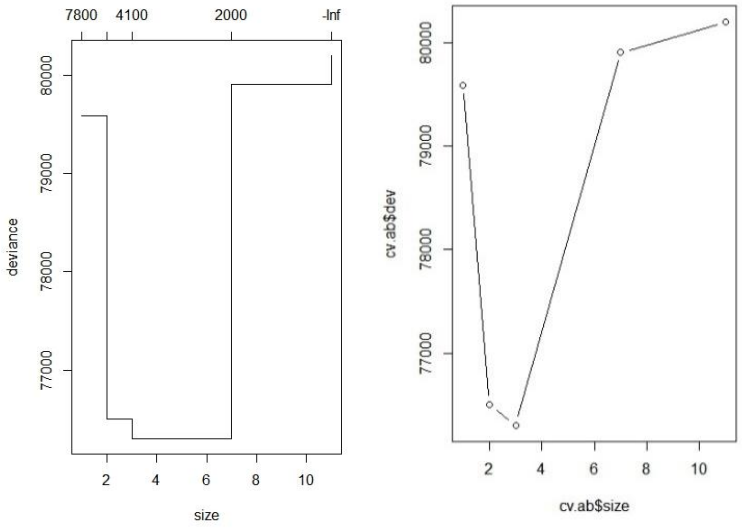


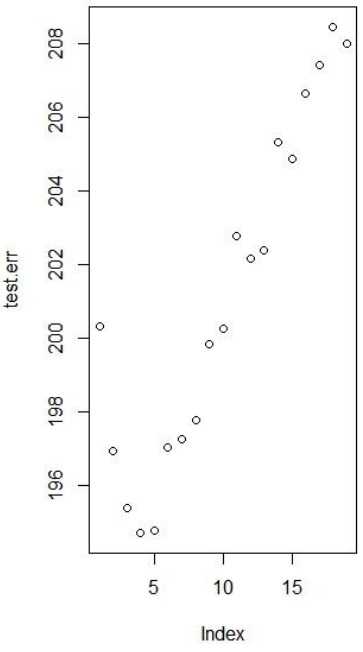
Figure 2.
CV-error when Pruning

Accordingly, the tuning parameter of 3 was selected by favouring more insight as the optimum pruned tree. The pruned tree has a test MSE of 201.3994, and is thus better at predicting the no. of absent hours than the unpruned tree. However, while this pruned tree performs better than the unpruned regression tree, its prediction is a significant deviation from the actual value, and in terms of predictive performance, a rather poor model.

2.1.3. Bagging and Random Forest Model

In order to improve the predictive performance, another model i.e. Bagging and Random Forest, was deployed. First bagging was performed (bagging is a special instance of random forest when m = p). First the model was implemented with all 19 variables (mtry=19) and with 10 trees (ntree=10). 10 trees were randomly selected to first test the model. The performance of this bagged model was then tested on the test set. The test MSE associated with the bagged regression tree is 204.1492, and thus performs worse than the pruned regression tree.

Figure 3.
No. of Variables and Test Error



In the R documentation, it is noted that the number of trees to grow ("ntree") "should not be set to too small a number to ensure that every input row gets predicted at least a few times." This indicates that the ntree value is dependent on the number of rows in the Absenteeism dataset. Therefore, a random forest was performed with the no. of trees increased to 1500 (to provide each row to get predicted at least a couple of times) but this increased the test MSE to 209.5456. While keeping the no. of trees at 1500, the no. of variables to be evaluated at each split was reduced to 10. This reduced the test MSE to 198.8056 which is below the test error of the pruned tree. This indicated that the no. of variables affects the predictive performance of the model.

Therefore, in order to assess which number of variables would have the best (lowest) test error, a loop was used. Through this, it was identified that the optimum number of variables is 4 with a test error of 194.5598. Accordingly, it is seen that random forests yielded an improvement over bagging in this instance.

2.2.2. Neural Networks

Neural networks are a type of deep learning technique which is capable of learning and predicting complex relationships between variables. In very simple terms, a neural network model consists of different layers: input, hidden, and output. In the input layer there are "neurons" which take the input data, then transforms the data through weights and bias terms, and then applies an activation function. This function is important as it enables the model to explore complex patterns that are non-linear. During the training process, the model learns how to adjust the weights and biases of the neurons in accordance with the input data so that it can reduce the difference between the actual output and the predicted value. Given the utility of neural networks in managing complex data, it is a sound model to implement in terms of the Absenteeism data.

Standardisation of Data | First, the data needs to be standardised. To do this, the scale() function and the model.matrix() is used. The latter converts all factors into dummy variables while the scale() function standardises the matrix so that the mean of each column is zero, and the variance is one.

Determine Model Structure of the Neural Network | Next, the model structure that describes the neural network is designed using the library(keras). First, an object "nnmodel" is created, and details regarding the successive layers in the model are added in sequence using the keras_model_sequential() function. The first layer i.e. "layer_dense" has 50 hidden units and the activation function used is Rectified Linear Unit (ReLU). This activation function is favoured as it helps to avoid vanishing gradients due to activation. Further, the input is also indicated to be the number of columns in the Absenteeism dataset. The next layer is the "layer_dropout" which randomly sets to zero a fraction (i.e. 0.4 in this case) of the inputs to the layer during training. Finally, the model has a final layer i.e. the "layer_dense". Because this is a regression problem, the final layer has a single unit with no activation function. This means it is used to produce a single output value.

Compile the Neural Network | Thereafter, the neural network was compiled. To compile the model, the mean squared error (mse) is specified as the loss function (this is commonly used in regression), while RMSprop (which optimises gradient descent) is specified as the optimizer. RMSprop adjusts the learning rate based on the magnitude of the gradient. The mean absolute error is specified as the evaluation metric. The MAE measures how accurately the model is capable of prediction.

Fit the Neural Network | In order to fit the model, the training data and parameters were provided. The two parameters are epochs and batch_size. A small batch size of 32 was selected with due consideration to the size of the dataset. This means that at each step, 32 training observations are selected at random to compute the gradient. As the training set has n = 490, an epoch is 490/32 = 15.3 Stochastic Gradient Descent (SGD). "Epochs" specify the number of times the model is to iterate over the training set.

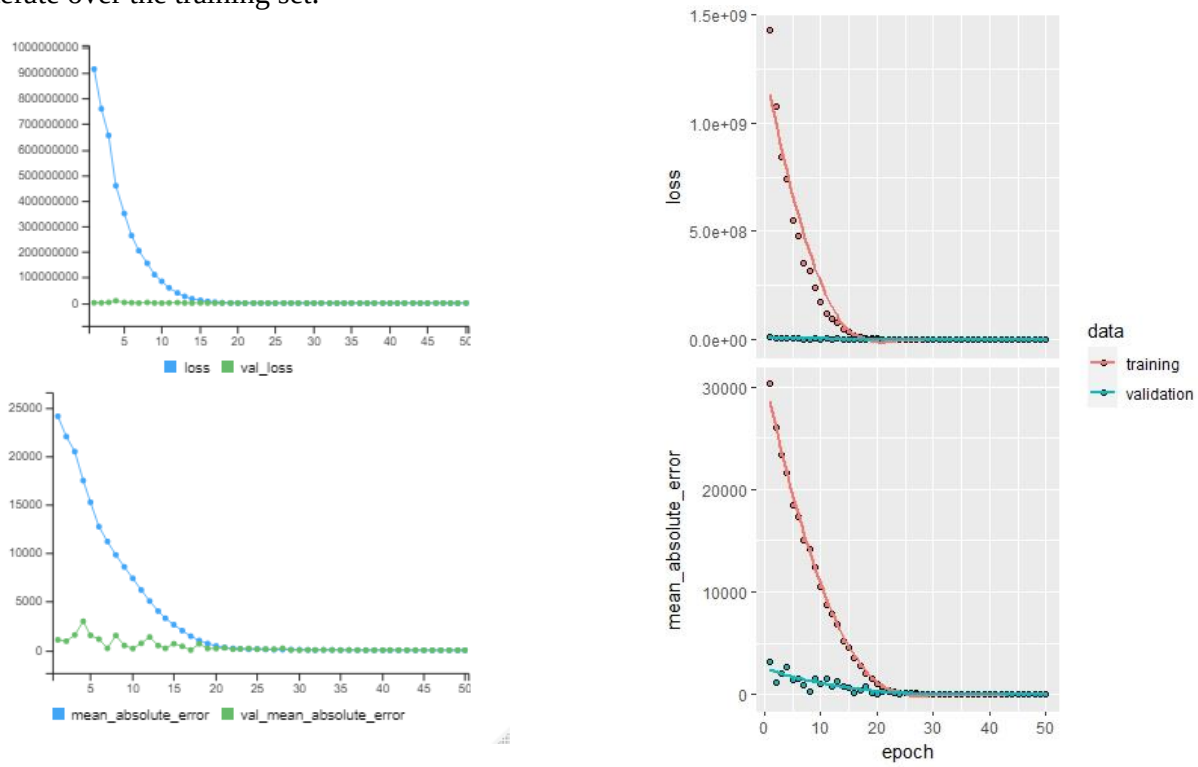


Figure 4. Training Loss and Validation Loss

The neural network was fitted at different values of epoch and the test mean square error computed for the predictions. In Figure 4, for example, the training loss can be seen to stop decreasing and plateau indicating that the weights and biases of the model have been changed and updated to such an extent that the continued iterations are unlikely to improve the model.

2.2.3. Generalised Additive Models

Finally, a GAM was selected as it has the advantage of being extremely flexible and being able to fit different methods (e.g. step functions for categorical variables, and non-linear polynomials for continuous variables). To select the variables to include in a GAM, it is first necessary to select a subset of variables. There are a number of subset selection methods (best subset selection). First, Forward Subset Selection was adopted using the regsubsets() function (it is less susceptible to overfitting). By identifying the minimum Cp value, a subset of 6 variables were identified i.e. "Reason, "Day", "Age", "DisciplinaryFailure", "Son", "SocDrinker". However, Forward Subset Selection assumes that the relationship between the variables are linear. In the present exercise, the exact relationship between the variables are not known. In order to account for both linear and non-linear relationships between variables, best subset selection, ridge regression, lasso regression, and pcr were all implemented to assess which has the highest median value in terms of prediction. In order to do so, first, the data was split into training ($2/3^{rd}$ i.e. 464 observations) and testing ($1/3^{rd}$ i.e. 232 observations) data. Next, the best subset selection model (using minimum BIC) was trained using the regsubsets function, ridge and lasso models using the cv.glmnet function (alpha=0 and alpha=1 respectively with 10 k-folds), and pcr using the pcr function over 50 repetitions. Thereafter, predictions were generated using the predict functions and the predictive performance evaluated.

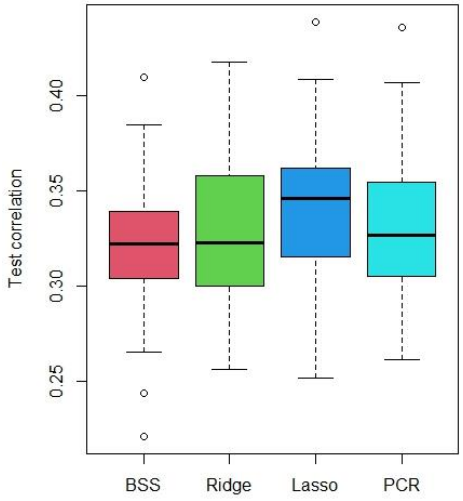


Figure 5. Boxplot of Predictive Performance of Best Subset Selection, Ridge Regression, Lasso Regression and Principal Component Analysis

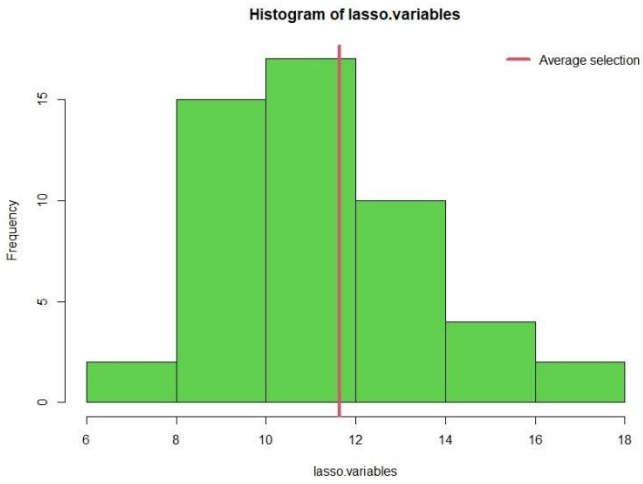


Figure 6. Histogram of No. of Variables Selected by Lasso Regression

The highest median value in terms of prediction i.e. the best predictive performance was demonstrated by lasso regression. The histogram above shows that lasso has the best prediction (with 11 variables). Thereafter, a lasso regression was fitted with the glmnet() function, and the variables where the coefficients have not been coerced to zero were identified. The variables with the highest coefficients were "Reason", "Day", "DistResidence", "Age", "Education", "SocDrinker", "DisciplinaryFailure", "Height". These were selected to fit a General Additive Model. These variables include both continuous and categorical variables. A GAM is capable of fitting both continuous and categorical variables. However, to facilitate better interpretation, the categorical variables i.e. were "Reason", "Day", "Education", "SocDrinker", "DisciplinaryFailure" were converted to factors.

When these variables are plotted, there are no obvious relationships between the variables. Given that this data relates to demographic, behavioural, and personal preferences, clear and definite relationships are unsurprisingly difficult to discern from a pairs plot. Therefore, the relationships between the no.of absentee hours and the continuous variables i.e. "DistResidence", "Age" and "Height" were explored by systematically fitting each variable to a linear model, and then

polynomials from 2-10 degrees in relation to "AbsentHours". Thereafter, a ANOVA test was conducted.

In examining the relationship between "AbsentHours" and "DistResidence, the hypothesis that the decrease in RSS is not significant is checked and rejected if the p-value is smaller than a significance level of 0.05. Accordingly, a 4-degree polynomial was selected (0.1625). In terms of the relationship between "AbsentHours" and "Age, the results of the ANOVA test suggest that a quadratic model is most suitable (p-value of the quadratic model = 0.002465). The third continuous variable is "Height". A similar procedure is followed, and the results of the ANOVA test suggest are unclear as the linear model appears to be the most suitable (as the p-values of all the polynomials are greater than 0.05). However, when "Height" is plotted against "AbsentHours", linearity is not apparent. While a linear model is adopted for the purposes of fitting the GAM model, this is a variable that requires greater examination given more time.

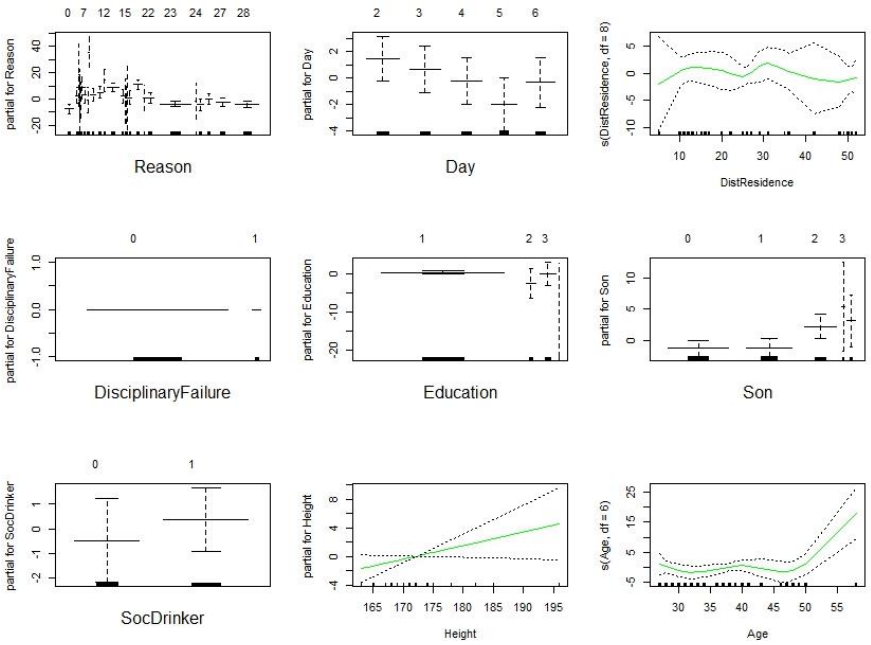


Figure 7. GAM fitted with 9 Variables

Thereafter, a GAM with splines, specifically, b-splines, natural splines, and smoothing splines, for each of the continuous variables. B-splines have d+K while a natural cubic spline K+1 degrees of freedom. While the ANOVA for Parametric effects for the GAM with smoothing splines shows that Reason is highly statistically significant, it also shows "Son", "SocDrinker", "Education" and "DistResidence" are also (to a lesser degree) significant. To assess which of these GAM Models are capable of predicting with the highest accuracy, it was trained on a 2/3rd data split and tested on the remaining 1/3rd. The test MSE for the GAM with the b-splines were the highest at 188.2952 while the GAM with the smoothing splines recorded the lowest at 182.8407. Therefore, the best performing GAM Model is the model with the smoothing splines fitted to the continuous variables.

3. Discussion and Further Work

3.1. Discussion of Results

Having fitted several models to the dataset, the results of the models are examined here. In order to make the comparison of models fair, the performance of each model was evaluated on the same training and test data split, and the test mean squared error was computed used to compare predictive performance.

First, an unpruned regression tree was fitted. This tree only used 8 of the variables to construct the tree. According to this unpruned regression tree, the reason for absence is the most important factor in determining the no. of absentee hours. It is immediately clear that if the reason is one classified above 20 i.e. if an employee is absent for a non-ICD disease related reason like medical or dental consultation, physiotherapy, or blood donation (i.e. reasons from 22 onwards), the absent hours are low (around 3 hours and 40 minutes (3.664)). This no. of hours is also predicted for reason 20 (external causes of morbidity and mortality and 21 factors influencing health status and contact with health services which are classified as ICD disease related reasons.

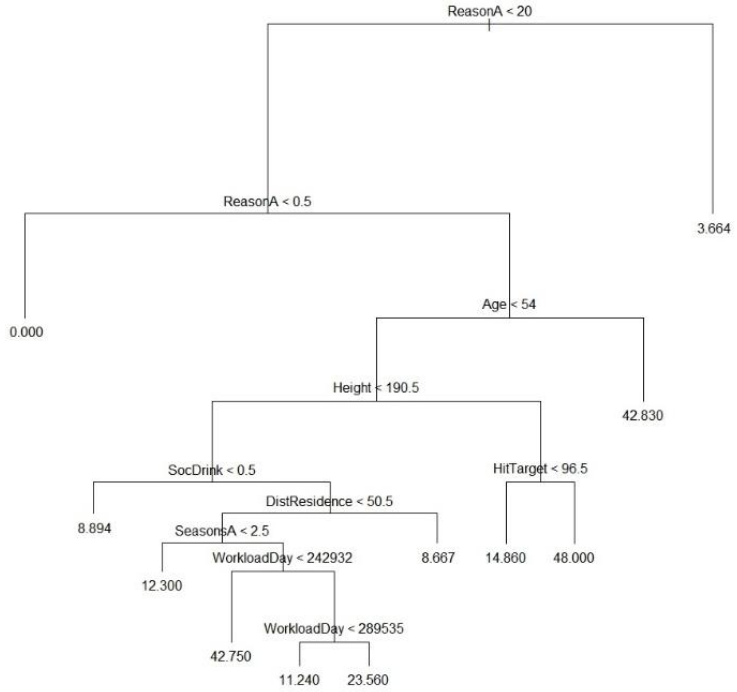


Figure 8. (Unpruned) Regression Tree

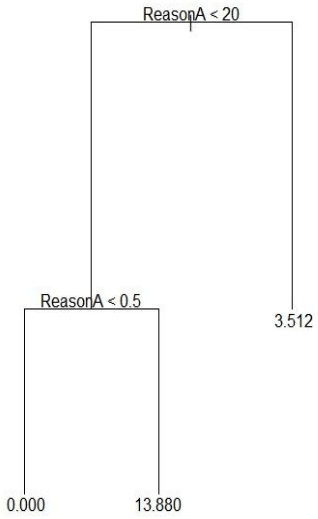


Figure 9. Pruned Regression Tree

On the other hand, if the reason is a ICD disease related reason apart from 20 and 21 above, the no. of absent hours will range between 8 hours and 40 minutes (8.667) and 48 hours. Of these, those above 54 years are likely to be absent around 42 hours and 50 minutes (42.830), while the no. of absent hours of those below the age of 54 years depends on their height, whether they are social drinkers, hit target, distance to residence, the season, and the workload. However, the **test MSE of the unpruned regression tree is 218.2868**.

Next, the tree was pruned to assess whether this would increase predictive performance. After selecting the tuning parameter of 3 for reasons specified in 2.1.2. above, it was noted that reason for absence remains the most important predictor. The **pruned tree has a test MSE of 201.3994**, and is thus better at predicting the no. of absent hours than the unpruned tree. As the square root of this is around 14.19, this indicates that regression tree leads to test predictions that are (on average) within approximately 14 hours and 11 minutes of the true median no. of absent hours. While this pruned tree performs better than the unpruned regression tree, its prediction is a significant deviation from the actual value, and in terms of predictive performance, a poor model.

The third model deployed was a bagging and random forest model with 1500 trees and 4 variables (see 2.1.3 for rationale). The **RandomForest model has a test MSE of 194.5598**. The experiments carried out with different number of trees and variables, indicated that to some extent, the variables used in the model are important in facilitating greater predictive performance. To examine which variables are considered important in this model, the importance() function was used.

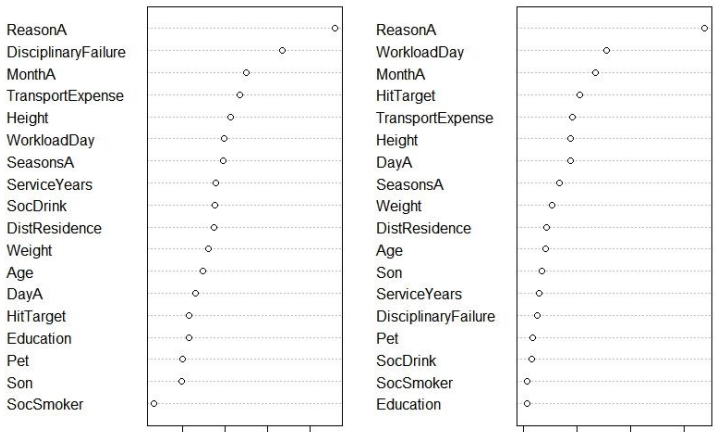


Figure 10. Variables in Decreasing Importance: RandomForest

These results indicate that from all the trees considered in the random forest, the reason for absence is by far the most important variable. In terms of the mean decrease of

accuracy in predictions on the out of bag samples (%IncMSE), the next most important variable is Disciplinary Failure. However, in terms of the total decrease in node impurity (IncNodePurity) which is measured by the training RSS, the next most important variable is the WorkloadDay. While these measures do not agree on an exact 4 variables that are important, it is clear from the Figure 10 that the reason, whether there was disciplinary failure, the workload of the day, and transport expenses are considered quite important in this bagging and Random Forest model.

The fourth model deployed as a single layer neural network. This model was constructed with 50 neurons that takes the input data, and the activation function ReLU which introduces non-linearity to the model and allows the exploration of non-linear complex relationships in the data. During the training process, the 50 neurons takes the input data, examines complex relationships between the data, and adjusts the weights and biases of each neuron to reduce the difference between the actual and predicted output. During this process, 40% of the 50 activations from the layer before are set to zero at random during each iteration. This dropout layer helps to regularise the learning process, prevent overfitting, and make optimistic predictions. The final layer is designed to produce a single output. While the predictions of this model cannot be visually plotted in the same manner a regression tree is, it is noted that when the neural network was fitted at 15 epochs, the test MSE scored 346.6779, and through incremental increase of epochs and continued testing, this **test MSE of the neural network model gradually reduced to 275.4530** at 2000 epochs. In any event, even with 2000 epochs, the neural network with a single layer has a higher test MSE than all the other models tested so far.

The fifth and final model is a Generalised Additive Model (GAM). This model was fitted with both categorical and continuous variables (the latter with smoothing splines). This **GAM model scored the lowest test MSE at 182.8407**. This model also shows how the number of absent hours increases after a specific age (between 45-50) while Thursday appears to be the day with the lowest number of absentee hours. Similarly, as the no.of children increases, absenteeism appears to increase, indicating that parental responsibilities may impact how employees approach their work.

3.2. Further work

The presence of correlated variables in the dataset was dealt with by removing one of the correlated variables. However, with further time, several other techniques can be used to mitigate the impact of multicollinearity. When implementing GAMs, Principal Component Analysis could be implemented (though this could reduce interpretability). For both GAMs and neural networks, impact maybe mitigated using regularization methods (e.g. ridge, lasso) as well.

In terms of the Neural Network Model, a neural network implemented to tackle a problem of Absenteeism at work can benefit from a larger dataset as these models perform better and reduces overfitting. The dataset in this exercise has less than 1000 observations and the predictive performance of the NN model was less than that generated by the other models. With more time and resources, the reason for this should be explored in depth. Further, the predictive performance of the model can be improved through techniques such as regularisation and data augmentation. Training a neural network also takes time and resources. In the present exercise, the neural network was tested on increasing degrees of epochs (from 15 to 2000). This reduced the test error to 275.453. However, with longer time it is possible to assess whether this can be improved. Furthermore, while a neural network may converge (i.e. where the weights and biases are unlikely to be better adjusted) during the training process, overfitting remains a concern. Therefore, additional techniques need to be used to address this such as early stopping. Furthermore, other neural network models can be explored to assess whether, for example, convolutional neural networks, can provide greater insight into the data, enhance generalisability, and predictive accuracy.

In terms of the GAM model, the present model has 9 variables (out of 19) fitted with smoothing splines (for the continuous variables). The model would benefit greatly from experiments to assess whether a different (and smaller) combination of variables can improve the predictive performance of the model.