ANLY 530: Absenteeism

Aditya Malik, Abisheik Mani, Gaurav Gade, Pradeep Paladugula

Harrisburg University of Science and Technology

Abstract

Competitiveness, market share, professional development and personal support are some of the key factors linked to a promising new market. As organizations evolve in a changing market, the goals and the employee expectations change. This results in various consequences as a by-product of the value creation process. One of them is employee absenteeism which could negatively impact the organization's goals and in turn, the bottom-line results. The aim of this project is to investigate and apply some machine learning algorithms to predict absenteeism at work. Our results indicate that the Decision tree model outperformed other conventional models with 81% accuracy and higher precision and recall scores. Reason for absence, transportation expense and social smoking were found to be major drivers of absenteeism in workplace. The proposed model will help guide organizations in designing policies intended to reduce absenteeism and turnover.

Keywords: Machine Learning, Prediction algorithm, Classification, Employee Absenteeism

ANLY 530: Absenteeism

Unplanned absences due to health-related issues or sickness can have a significant impact on worker productivity and cause disruptions to normal operations impacting profitability. Studies have indicated a link between health and productivity which somehow eludes the radar of many organizations as they make big investments in human capital and personal development. Center for Disease Control and Prevention (CDC) reports the productivity loss related to health problems cost U.S. employers $1,685 per employee per year, or, $225.8 billion annually. These indirect costs affect all employers, even those who avoid medical costs by not funding health insurance. Analytical techniques using Machine learning models can be useful in extracting the relationship between employee data pertaining to absenteeism due to sickness and other factors such as social, physiological, educational and lifestyle choices, understanding the patterns and predicting absenteeism at work. In this project, we use the dataset from a courier company that includes employee records from a three-year period starting July 2007 to July 2010. The project will compare different machine learning algorithms such as Decision Tree Algorithm, Random forest etc. and choose a best performing model that can predict the number of hours of absence.

Related Work

In the highly cited paper (Porter & Steers, 1973) the authors categorize the factors causing absenteeism into organizational, immediate work-related, job-related, and personal. Several studies have explored these four factors in different ways. The impact of sickness absenteeism is more pronounced in industrial firms, high capital-intensive companies, and small- and medium enterprises (Elena & Francois, 2018). The authors report an increase of 1 percentage point in the rate of sickness absenteeism entails a productivity loss of 0.24% based on by applying a

modified version of the Ackerberg et al's (2015) control function method, which explicitly removes firm fixed effects. Machine learning approaches are highly suited to the analysis of absenteeism data as a popular research tool reports Tewari et al. In (Gonen and Izack, 2020), ordinal CART model with information-gain measure is used on the absenteeism dataset. The authors report CART model performs better than Naïve Bayes model by 20% (AUC = 0.65 and AUC = 0.54, respectively).

Data Description

The dataset used in the project consists of records of absenteeism from work during a three-year period starting July 2007 to July 2010 from a courier company. The train and test data split is 90% / 10% with 666 training data and 74 test data observations respectively. There are a total of 20 descriptive features and a target feature. The feature details are provided in the Tables section. Absenteeism time expressed in hours is the target feature for our analysis. The features 'Hit target' and 'Weight' consists of 1 and 2 missing values respectively. The missing values are handled by omitting them as they are very few compared to entire dataset and the impact is expected to be insignificant. Also, non-predictive features such 'ID', 'Workload average day', 'Hit target' are excluded as they do not add much value. The features 'Height' and 'Weight' are excluded as the 'Body Mass Index' feature is calculated from these two features. Since the target feature is a continuous variable, it is categorized into 3 groups (Group 1: No absentee hours, Group 1: up to 6 absentee hours, Group 3: More than 6 absentee hours).

Exploratory Data Analysis[1]

In order to have a thorough understanding of the absenteeism dataset, we wanted to explore the data from multiple angles to determine trends to help us when building the prediction models. We looked at all the variables and compared them against the target response variable to determine insights. Broadly, we split the analysis into 2 major overarching categories:

1. **Category I:** This category mostly focused on factors not personally related to the employee. This category includes variables such as reason for absence, transportation expense, day of the week and month of absence. A few clear observations here were a direction correlation between transportation expense and rate of absenteeism, as well as reason for absence and the rate of absenteeism. We did see that employees who have received disciplinary failures are significantly less likely to be absent, though this observation is tempered by the fact that the count of employees who haven't received a disciplinary failure significantly outnumber the ones who have. Looking at the calendar attributes, Mondays and the Summer months show the highest average absenteeism trends. Digging deeper into the Seasons, we find some inaccurate data trends, where a single month is attributed to multiple seasons. This caused us to not use the Seasons variable when building the prediction models. Figure 1 shows the trends which we observed in this category.
2. **Category II:** This category mostly focused on factors personally related to the employee. This category includes variables such as the number of children and pets for each employee, and the employee's age and level of education. Here, we can clearly see a correlation between the employee's number of pets and children compared to their rate of absenteeism, with absenteeism increasing with an increase in the count of pets and

children. Other observations here showed that an employee's age, BMI, and level of education do not affect their absenteeism rate. However, employees who smoke and/or drink have a higher chance of being absent compared to those who don't. Figure 2 shows the trends which we observed in this category.

Technical Approach

Testing and Evaluation

The target feature (absenteeism hours) in the dataset was grouped into three categories and labeled as group 0, group 1 and group 2 respectively to apply supervised learning models. For the supervised models, the performance of each model was evaluated using accuracy and error rate. To compare performance of different supervised learning models, a number of metrics such as precision, recall, specificity and F1 – score was used in addition to accuracy and error rate (impurities). The best performing model was selected based on the scores obtained on the majority of above metrics. In the second part of the analysis, the target feature was treated as a continuous variable and Multiple Linear Regression (MLR) was applied to predict the absenteeism hours based on the selected independent variables. The MLR model was tested using adjusted R square value to determine the percentage of variance explained by the model. The first iteration of the model coupled with the accuracy scores obtained from confusion matrix is used as a baseline. If the results show an accuracy less than 50%, a second iteration is performed by changing model parameters such as number of trees in the case of Random forest or the number of nodes in Decision Tree models. The percentage reduction in error rate is used to evaluate performance improvements of subsequent iterations.

Model Performance

Decision Tree model applied on the absenteeism data set yielded an accuracy of 81% with reason for absence, seasons and age being the key variables. Random forest, an ensemble model was applied on the same dataset and quite surprisingly, it resulted in relatively lower accuracy (70%) than DT but marginally higher than Linear SVM (69%). Pruning the number of trees to the range of 30 -50 did not improve the model performance. SVM using RBF kernel provided the same results as Naïve Bayes model with an accuracy of 70%. Overall, Decision Tree performed better on all the metrics compared to the other models with 86% precision and 87% recall. Random forest performed better than SVM and Naïve Bayes in precision and specificity. For absenteeism prediction, MLR resulted in an adjusted R square value of 0.1281 indicating that 12.8% variance in absentee hours is explained by the model. The standardized residuals plot indicated linearity for most part with some deviation caused by outliers. The significant factors predicted by the model were reason for absence, day of the week and disciplinary failure. The low R square value indicates that this model is not suitable for prediction and may require significant improvement. Based on the results, our recommendation is Decision Tree model for predicting absenteeism in workplace.

Next Steps

The performance of the models could be improved by implementing cross validation techniques and perhaps, using feature selection or principal component analysis (PCA) for dimensionality reduction. Other learning models such as ID3 or C4.5 could be applied to get better prediction accuracy. Screening outliers using factors such as Mahalanobis cutoff, Interquartile range (IQR) or z-scores could help improve performance of models susceptible to

outliers in the dataset. Another recommendation is to test the models on other absenteeism datasets to better determine model reliability.

Conclusion

The team looked at the 'Absenteeism at Work' dataset to determine trends which could help predict absenteeism and build prediction models to predict the average absenteeism hours per employee. The exploratory data analysis culminated in us determining direct correlations between absenteeism and reason of absence, transportation expense, smoking/drinking, and number of pets/children. Based on insights gained during the exploratory data analysis, we built 6 prediction models to predict employee absenteeism. 5 models used categorial data to predict absenteeism, of which the Decision tree model performed with the highest accuracy followed by the Radial Basis SVM model. The team also used the Multiple Linear Regression algorithm to build a prediction model to predict absenteeism hours as a continuous variable. Subsequently, the team wants to use the currently built models against other similar datasets to test its effectiveness and use other more powerful prediction algorithms to try and more accurately predict an employees absenteeism.

References

Grinza, E., & Rycx, F. (2018). The Impact of Sickness Absenteeism on Productivity : New Evidence from Belgian Matched Panel Data. *IZA Discussion Paper Series*, *IZA DP No.*(51), 28. (Grinza & Rycx, 2018)

Singer, G., & Cohen, I. (2020). An objective-based entropy approach for interpretable decision tree models in support of human resource management: The case of absenteeism at work. *Entropy*, *22*(8). https://doi.org/10.3390/E22080821
(Singer & Cohen, 2020)

Porter, L.W., & Steers, R.M. (1973). Organizational, work, and personal factors in employee turnover and absenteeism. *Psychol. Bull.*, *80*, 151. (Porter & Steers, 1973)

Ackerberg, D. A., Caves, K., Frazer, G., 2015. Identification Properties of Recent Production Function Estimators. *Econometrica 83 (6)*, 2411-2451. (Ackerberg et al., 2015)

Tables

Table 1

Description of dataset

Name	Type	Overview of Information
Absenteeism time in hours	Response	Count of absenteeism hours
Individual identification (ID)	Explanatory	Unique identifier for the employee
Reason for absence (ICD)	Explanatory	21 categories for absences attested by the International Code of Disease
Month of absence	Explanatory	Month when employee was absent
Day of the week	Explanatory	Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6)
Seasons	Explanatory	summer (1), autumn (2), winter (3), spring (4)
Transportation expense	Explanatory	Amount spent on transportation by the employee
Distance from Residence to Work	Explanatory	kilometers
Service Time	Explanatory	
Age	Explanatory	Age of the employee
Workload Average/day	Explanatory	Average workload for the employee
Hit Target	Explanatory	

Name	Type	Overview of Information
Disciplinary failure	Explanatory	yes=1; no=0
Education	Explanatory	high school (1), graduate (2), postgraduate (3), master and doctor (4)
Son	Explanatory	Number of children
Social Drinker	Explanatory	yes=1; no=0
Social Smoker	Explanatory	yes=1; no=0
Pet	Explanatory	If the employee has a pet or not
Weight	Explanatory	Weight of the employee
Height	Explanatory	Height of the employee
Body Mass Index	Explanatory	BMI of the employee

Table 2

Comparison of Model Performance

Metric	Decision Tree	Random Forest	Linear SVM	RBF SVM	Naïve Bayes
Accuracy	81%	70%	69%	70%	70%
Error Rate	19%	30%	31%	30%	30%
Precision Rate	86%	78%	61%	61%	64%
Recall Score	87%	65%	77%	79%	77%
Specificity	89%	81%	83%	87%	82%
F1 - score	87%	71%	68%	69%	70%

Exploratory Data Analysis Part I

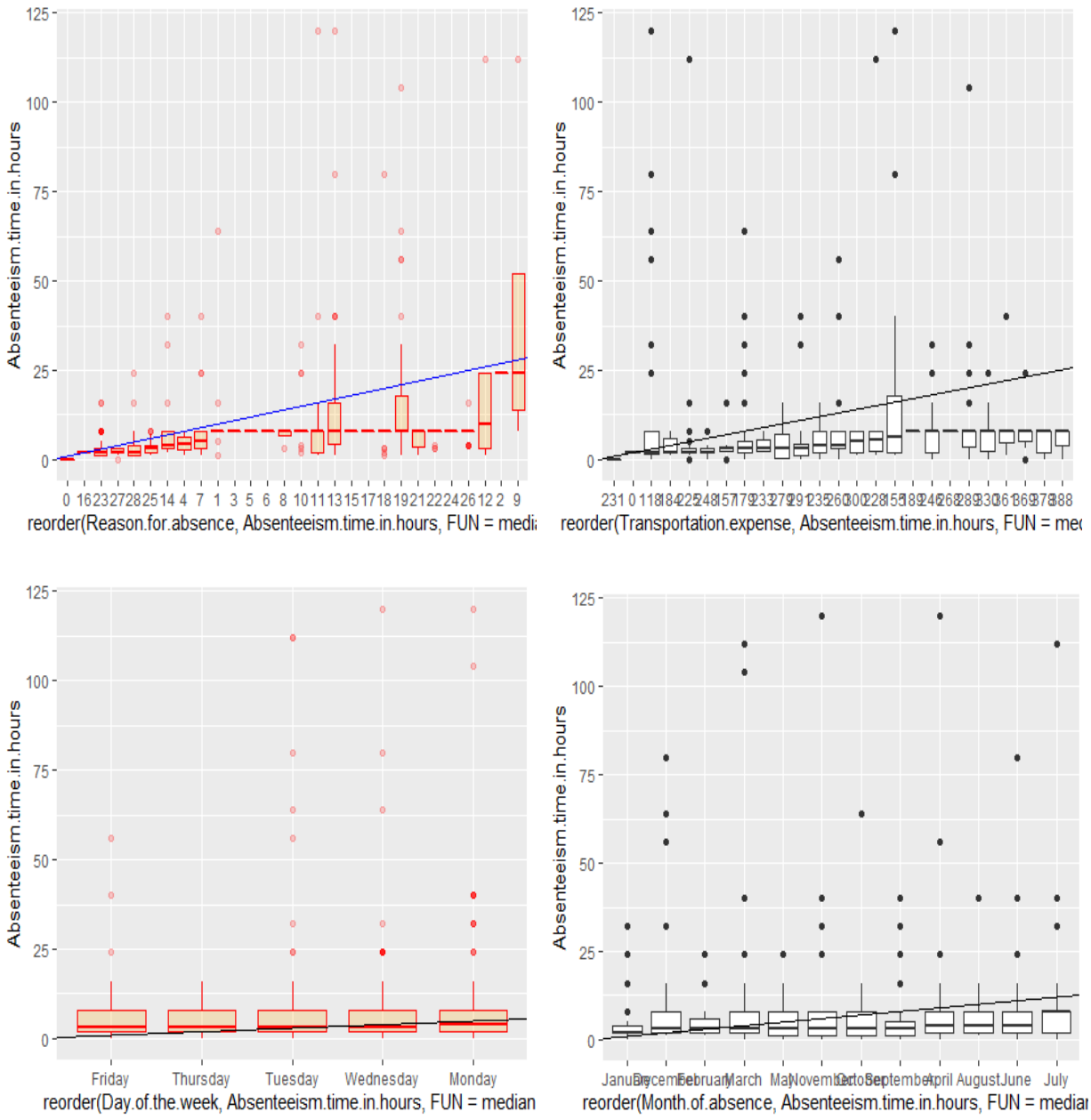


Figure 1. The distribution of absenteeism data vs. reason for absence, transportation expense, day of the week and month of absence

Exploratory Data Analysis Part II

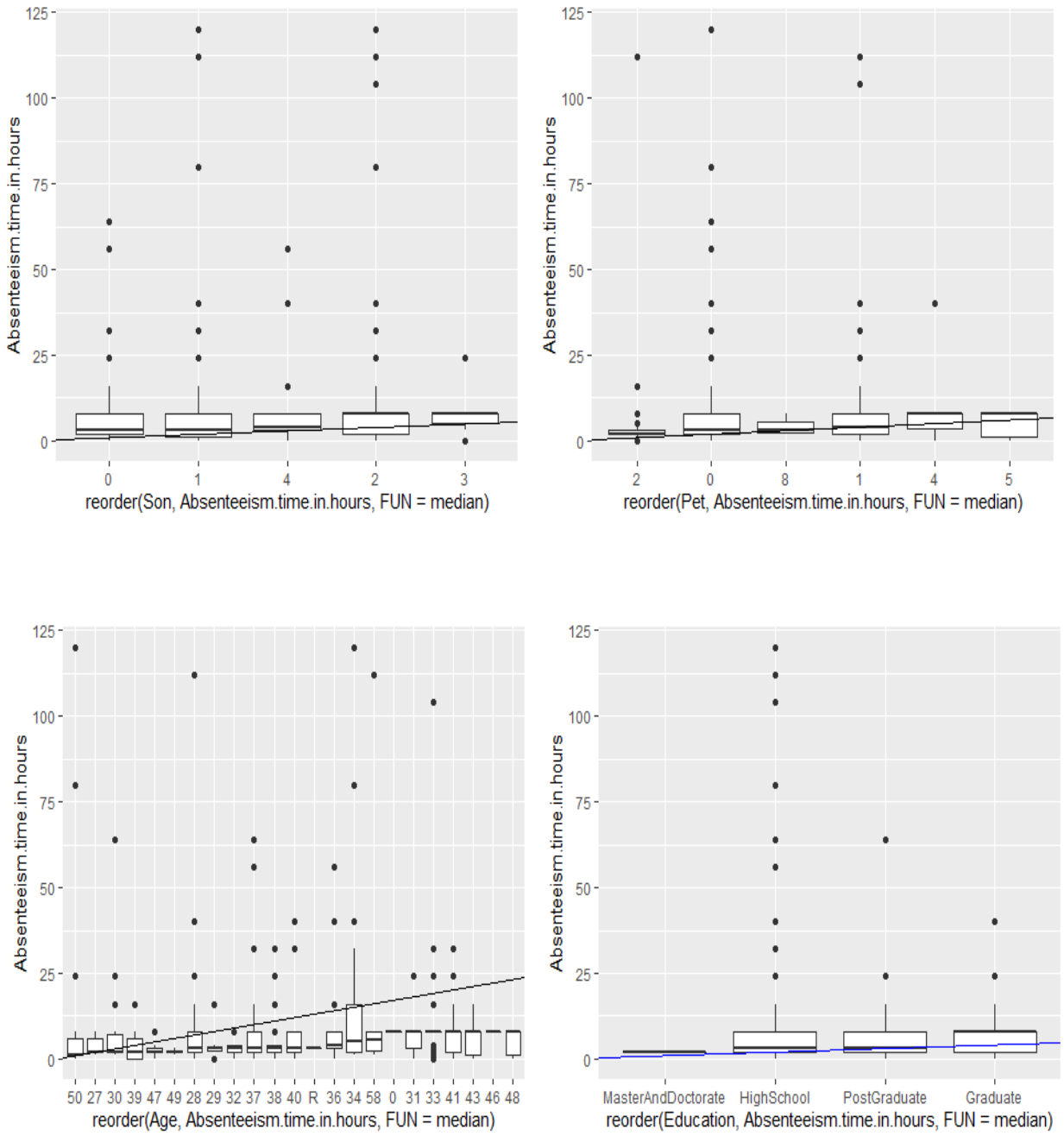
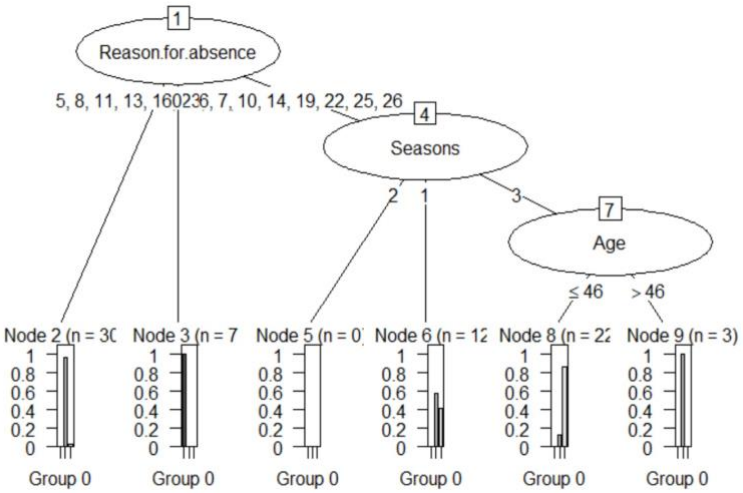


Figure 2. The distribution of absenteeism data vs. employees who have children, employees who have pets, employee's age and level of education

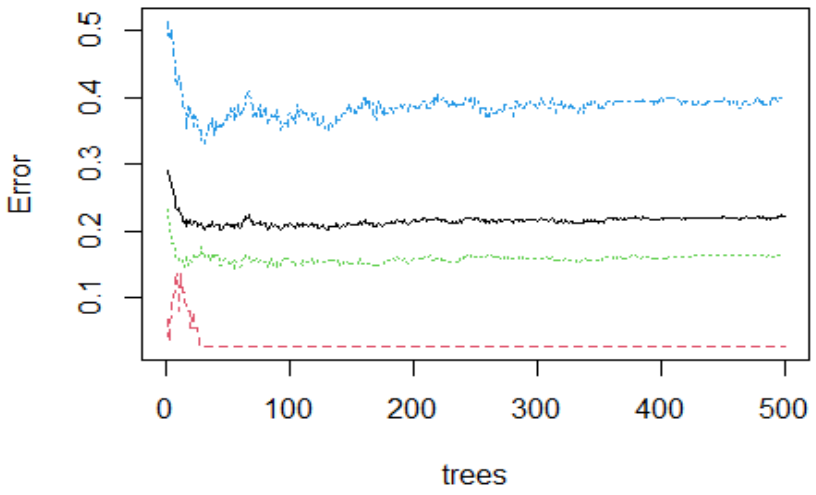
Decision Tree Model



Actual Group	Predicted Group: Group 0	Group 1	Group 2	Total
Group 0	7	0	0	7
	0.095	0.000	0.000	
Group 1	0	31	11	42
	0.000	0.419	0.149	
Group 2	0	3	22	25
	0.000	0.041	0.297	
Total	7	34	33	74

Figure 3. Decision tree model view and confusion matrix

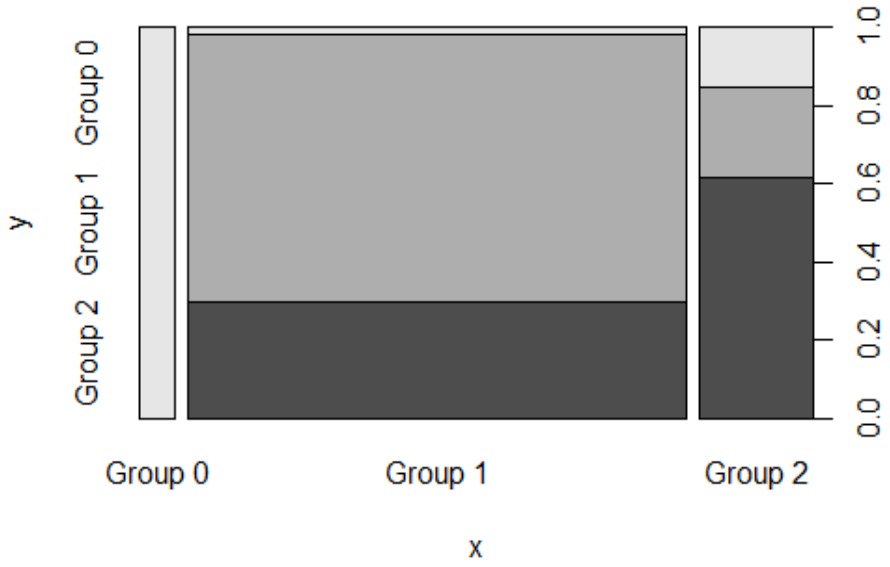
Random Forest Model



Actual Group	Predicted Group Group 0	Group 1	Group 2	Total
Group 0	4	1	2	7
	0.054	0.014	0.027	
Group 1	0	33	9	42
	0.000	0.446	0.122	
Group 2	0	10	15	25
	0.000	0.135	0.203	
Total	4	44	26	74

Figure 4. Random forest model view and confusion matrix

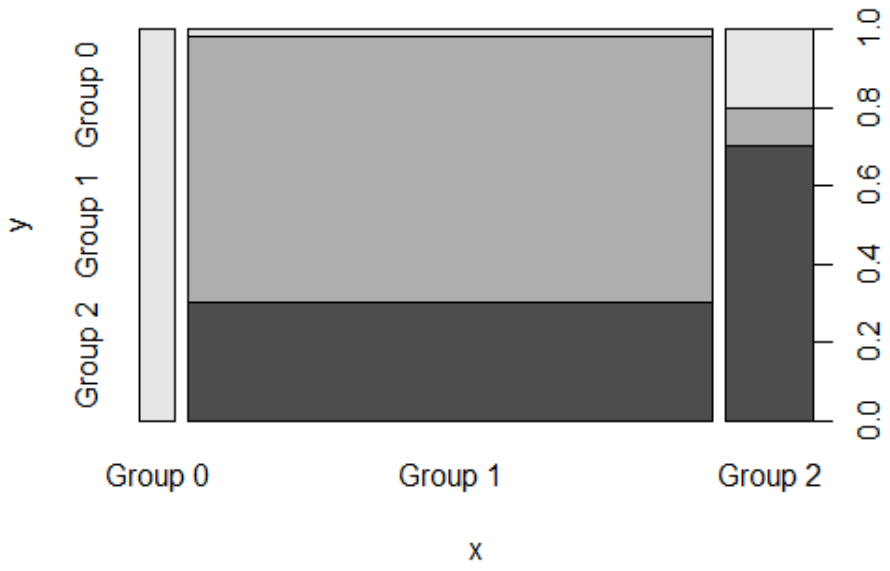
Linear SVM Model



svm_vanilla_pred	Group 0	Group 1	Group 2
Group 0	4	0	0
Group 1	1	39	17
Group 2	2	3	8

Figure 5. Linear SVM model view and confusion matrix

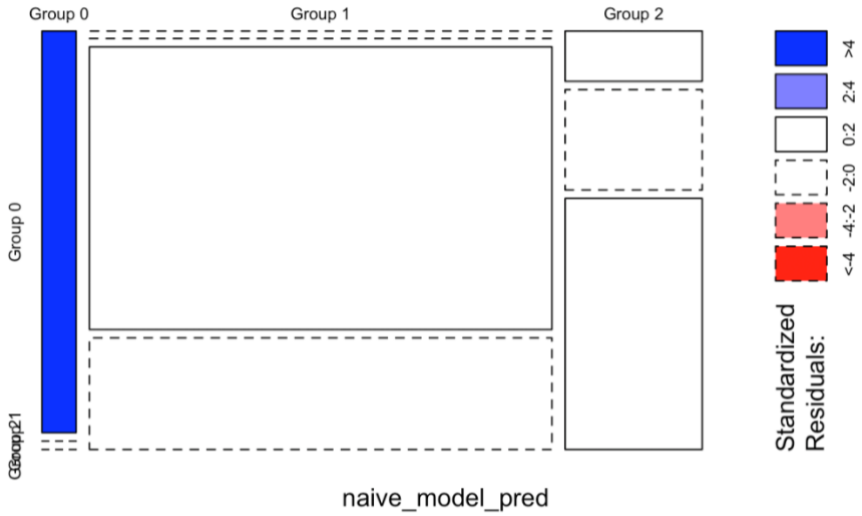
Radial Basis SVM Model



svm_rbf_pred	Group 0	Group 1	Group 2
Group 0	4	0	0
Group 1	1	41	18
Group 2	2	1	7

Figure 6. Radial basis SVM model view and confusion matrix

Naïve Bayes Model



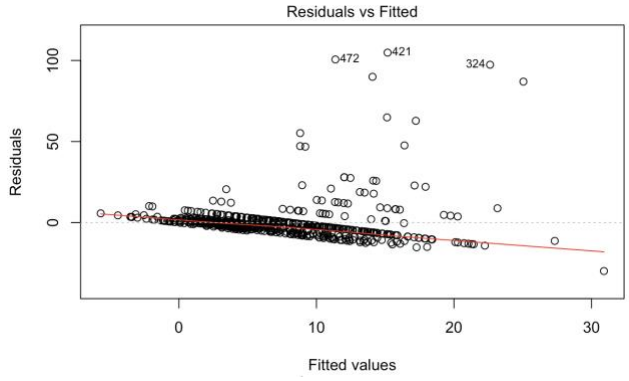
naive_model_pred	Group 0	Group 1	Group 2
Group 0	4	0	0
Group 1	0	35	12
Group 2	3	7	13

Figure 7. Naïve Bayes model view and confusion matrix

Multiple Regression Model

```
Call:
lm(formula = trainData_ec$Absenteeism.time.in.hours ~ ., data = trainData_ec)

Coefficients:
                    (Intercept)               Reason.for.absence                  Month.of.absence
                      14.014316                       -0.538449                       -0.019848
                Day.of.the.week                         Seasons           Transportation.expense
                      -0.767817                        0.420462                       -0.001785
Distance.from.Residence.to.Work                  Service.time                           Age27
                      -0.056781                        0.041334                       -3.716328
                          Age28                           Age29                           Age30
                      -3.470583                       -7.621527                        0.355850
                          Age31                           Age32                           Age33
                      -6.406164                       -6.553237                       -3.085133
                          Age34                           Age36                           Age37
                       1.232106                       -1.087801                       -5.290325
                          Age38                           Age39                           Age40
                      -9.154062                      -10.017114                       -4.040645
                          Age41                           Age43                           Age46
                      -4.588580                      -11.256420                       -7.235828
                          Age47                           Age48                           Age49
                      -9.385046                       -6.400292                      -13.419140
                          Age50                           Age58                            AgeR
                      -6.227916                       11.034681                       -8.259558
          Disciplinary.failure                       Education                             Son
                     -18.899685                       -0.973335                       -0.045155
                Social.drinker                  Social.smoker                             Pet
                       3.634442                        1.794443                       -0.253623
              Body.mass.index
                       0.447214
```



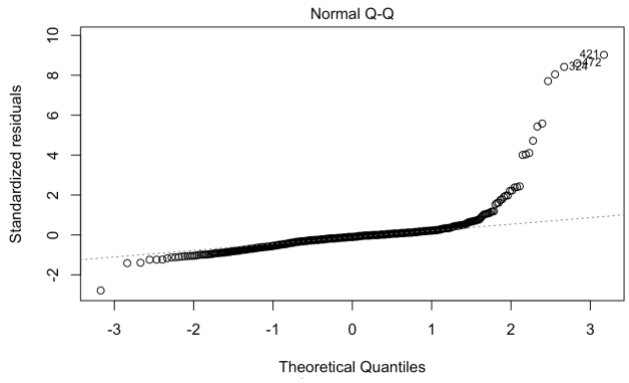


Figure 8. Multiple Linear Regression (MLR) model results