Prediction of Absenteeism at Work

Author: Shravya Guda

Group members: Pavel Mesa Neimane, Rosanna Gogliotti, Catherine Aslinger, Sarah Baylor

University of Notre Dame

Department of Applied Computational and Mathematical Sciences

October 31, 3018

TABLE OF CONTENTS

Introduction

The dataset describes absenteeism in a work environment based on 21 variables of employment parameters. The dataset is provided under public domain by the UCI Machine Learning Repository, and was donated to the repository on 4/5/2018. The data itself describes individuals in the time period between July 2007 to July 2010 from a courier company in Brazil, according to the UCI Machine Learning Repository.

Pre-Processing

The data has 21 columns and 740 observations. To Preprocessing portion of the investigation, first the data was examined for missingness. It was found that there was no missing data within the dataset. All data within the dataset is either numeric or categorical.

Next, the categorical data was transformed into ordered factors to handle the categorical class more effectively, and the numeric data was isolated to a smaller data frame. This was done in preparation for our Exploratory Data Analysis. Additionally, column names were renamed from the original file into shorter and more manageable names for the purposes of this analysis.

Then, the ranges of individual variables were examined. It was found that the Month variable had three entries with a value of 0, which did not make sense in the context of the data, because months range from 1 to 12. Because there were only three entries out of 740 observations with the zero value in the Month column, they were removed from the data set for the analysis. After removing these values, the numerical variables within the dataset were scaled to prevent unnecessary skewing of analysis results.

The final step in pre-processing the data was the discretization of the Absenteeism variable. We first visualized the variable, seen in Figure 1. This investigation was geared more towards determining the causes of extended absenteeism, so the data was discretized into two categories; the first being observations of absenteeism eight hours or less and the second being observations of absenteeism greater than eight hours. Upon examining the discretized Absenteeism variable, it was found that the variable was very imbalanced, with significantly more observations of absenteeism incidences under eight hours than there were over eight hours. As such, the observations representing eight hours of absenteeism or less are the majority class, and those over eight hours are the minority class, as seen in Figure 2.

Exploratory Data Analysis

The first step taken in Exploratory Data Analysis was the creation of exploratory plots. All variables were plotted in a scatter plot against the discretized absent time variable to see any initial trends, as seen in Figure 3. Next, the variables were plotted against each other to observe any potential correlations between variables, and example of which is seen in Figure 4. Finally, the variables were plotted individually as a histogram against absent time to notice any trends in the individual variable in a histogram, examples of which can be seen in Figure 5.

Upon inspection of the graphs it was discovered that some variables would likely not be useful in predicting absent time, as there was little to no trend with absenteeism. An example of this is disciplinary failure. As seen in Figure 6, there are few disciplinary failures in the majority class of absent time under eight hours, and no disciplinary failures in the minority class of absent time over eight hours. That being said, all variables were included in the data set during initial

modeling stages, and later removed from the data set for model optimization if they were indeed not good predictors of absenteeism.

Initial Models

Five model algorithms were selected to create initial models for the data. These models were:

a) K Nearest Neighbor (KNN): This model classifies the test observation based on the class of close neighboring observations.
b) General Linear Modeling (GLM): This model performs logistic regression to create a model.
c) Decision tree: This model creates a series of partitions at nodes to classify the data.
d) Random Forest: This model is based in Decision Trees but aggregates many Decision Trees together to avoid dominance and overfitting of any particular variable that can occur with Decision Trees.
e) Support Vector Machine (SVM): This model separates the data into classes using linear planes.

The objective of each initial model was to accurately predict the minority class- absent time greater than eight hours. The models were compared on four metrics, as seen in Figure 7: Sensitivity, F-measure, Geometric Mean, and Model Error.

$$Sensitivity = \frac{\#\ of\ true\ positives}{Total\ number\ of\ observations}$$

Equation 1. Sensitivity is defined as the ratio of true positives to total observations.

As seen in Equation 1, Sensitivity is a measurement of the model's ability to accurately determine whether or not an observation is in the minority class, otherwise known as the rate of true positives.

$$F1 = \frac{2 * precision * sensitivity}{precision + sensitivity}$$

Equation 2. F-measure, also known as F1 score, is defined as the harmonic mean of precision and sensitivity.

As seen in Equation 2, the F-measure is the harmonic mean of precision and sensitivity, ranges from zero to one, with one being a high performing value and zero being a low performing value. The F-measure also aids in determining the quality of prediction accuracy in a model.

$$g = \sqrt{precision * sensitivity}$$

Equation 3. Geometric mean, known as g, is the geometric mean of precision and sensitivity.

The Geometric mean is the geometric average of precision and sensitivity, as seen in Equation 3. The purpose of the geometric mean is to measure how well the model predicts both the majority and minority class- in this case, short and extended absences, respectively.

$$Error = \frac{Number\ of\ False\ Positives + Number\ of\ False\ Negatives}{Total\ number\ of\ observations\ in\ test\ dataset}$$

Equation 4. The error rate is the number of misclassified test observations divided by the total number of test observation.

As seen in Equation 4, the final model metric was the model error, calculated from the Confusion Matrix. The model error is the proportion of incorrectly classified observations to the total number of test dataset observations.

Box plots, as seen in Figure 7, were created to compare these four key model metrics between the five initial models tested. While the Support Vector Machine model had the lowest model error rate, the KNN model performed best in Sensitivity, F-measure, and Geometric Mean, with values of 0.23, 0.24, and 0.47 respectively. Because the classification of the minor class is crucial, especially given the data set is so unbalanced, Sensitivity was valued as the most important key metric. Therefore, even though SVM presented the lowest error rate, KNN was regarded as the better overall model.

It is worth noting that GLM also performed about as well as KNN did in the comparison of the initial models. However, KNN was still selected as superior as it displayed the principle of Occam's Razor more effectively- it is a simpler model overall than GLM and given that they perform similarly, the simpler model was considered as superior.

Model Optimization

The goal for the optimized KNN model was to increase the Sensitivity measurement by addressing the imbalance in the data and also selecting only relevant predictors for the model. To address the imbalance in the data, a Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset. The SMOTE technique underrepresents the majority class and creates synthetic samples of the minority class, effectively modifying the training dataset to be 50% majority class and 50% minority class.

To select variables that are good predictors of absent time, a random forest algorithm was applied to predict the importance of each variable using the gini index. The data set was then ordered based on the importance of each variable. The five most important variables indicated by this method were Seasons, Reason for Absence, Service time, Month, and Work load.

To optimize the K value and to select the ideal combination of predictor variables, a Monte Carlo Cross Validation was performed on the training data set. 500 trials were run with 2/3 samples from the training set on all possible permutations of the variables and odd K values, for a total of 95,000 tests. The Sensitivity was compared for the top 5 models, as seen in Figure 8. These were re-tested with new training samples to verify that the models were not overfit to the initial training sample. After confirming the fit was good, the top model was selected. The top performing model uses a K-value of 19 and 14 out of the initial 19 predictors, dropping Children, Pets, Social Smoker, Social Drinker, and Disciplinary Failure from the final model.

Comparison of Models

Figure 9 compares the Sensitivity of the optimized model to that of the initial KNN model. The initial KNN model had a median Sensitivity of 0.23, but the optimized model has a median Sensitivity of 0.81- a significant improvement in this measurement. This improvement indicates that the optimized model can more accurately predict the minority class of extended absences than the initial model can. Figure 11 displays a comparison of the confusion matrix for each model. The number of true positives in the optimized model are a little over double that of the initial KNN model, but the number of false positives are also greater. Given the significant

improvement of the model, it was determined the benefit of accurately predicting the true positives outweighs the instances of the false positives in this optimized model.

Conclusion and Next Steps

The optimized KNN model with 14 variables and a K value of 19 performed significantly better than the initial KNN model that was created. The optimized KNN model was verified with Monte Carlo cross validation and correctly predicted 27 true positives, compared to the 13 that the initial KNN model predicted. The optimized model does have a greater frequency of false positives, so further research could be conducted in this case to determine a way to mitigate these occurrences to make the model even stronger.

Appendix A: Figures

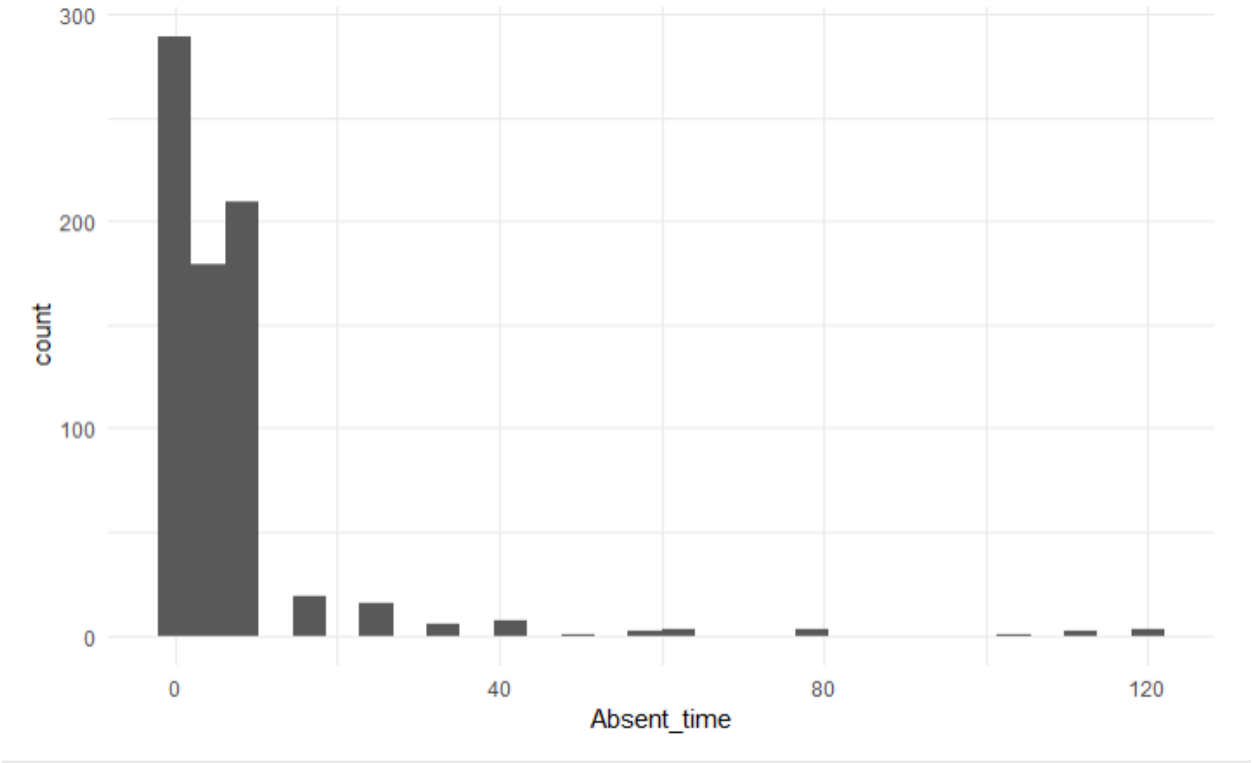


Figure 1. Histogram of Absent Time prior to discretization.

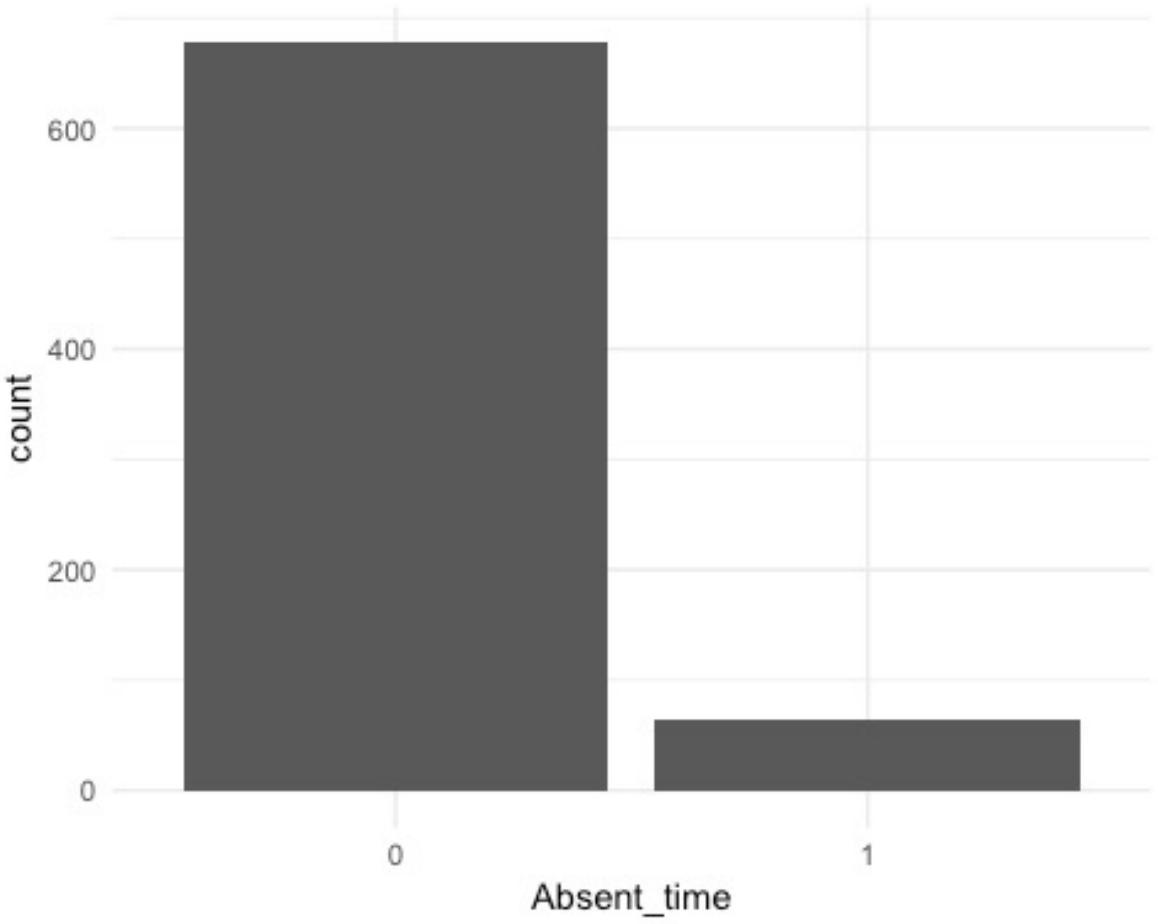


Figure 2. Histogram of discretized Absent Time.

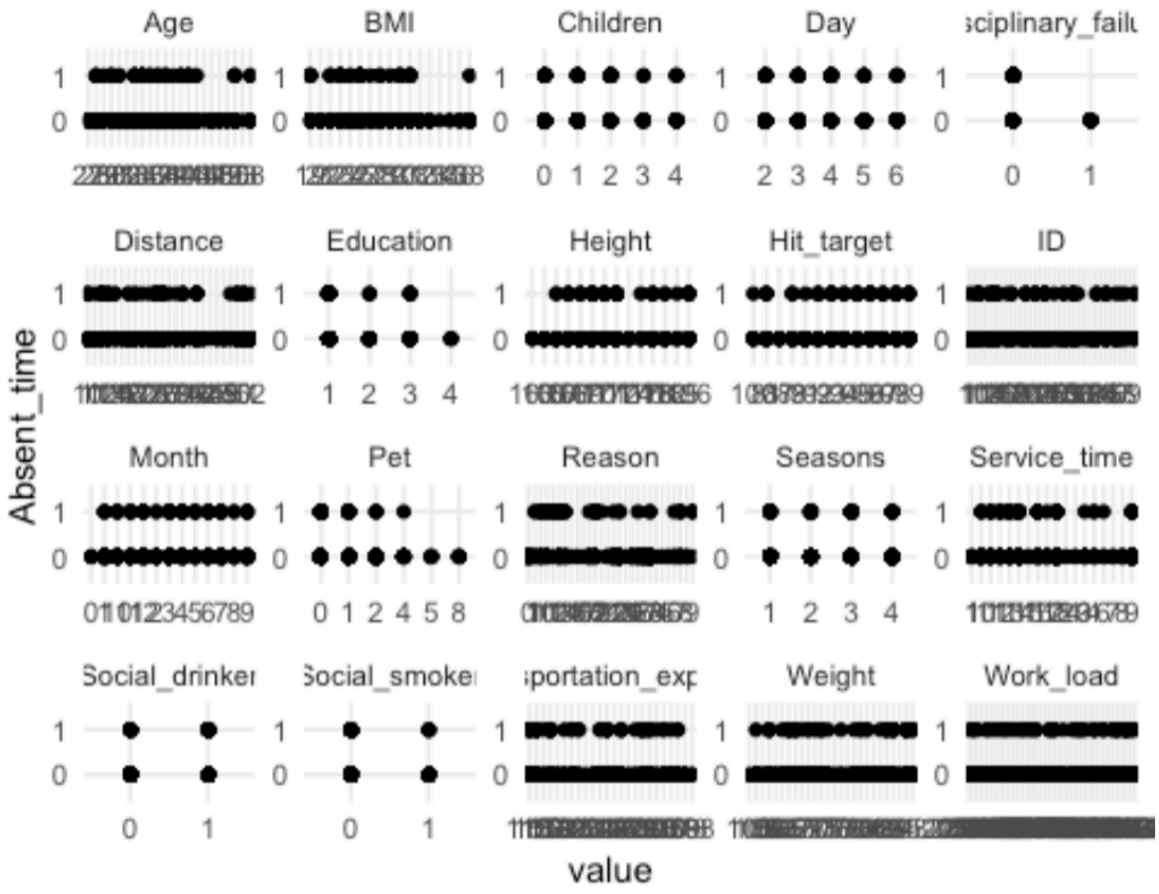


Figure 3. Scatterplot of Absent Time versus all other variables in the data set.

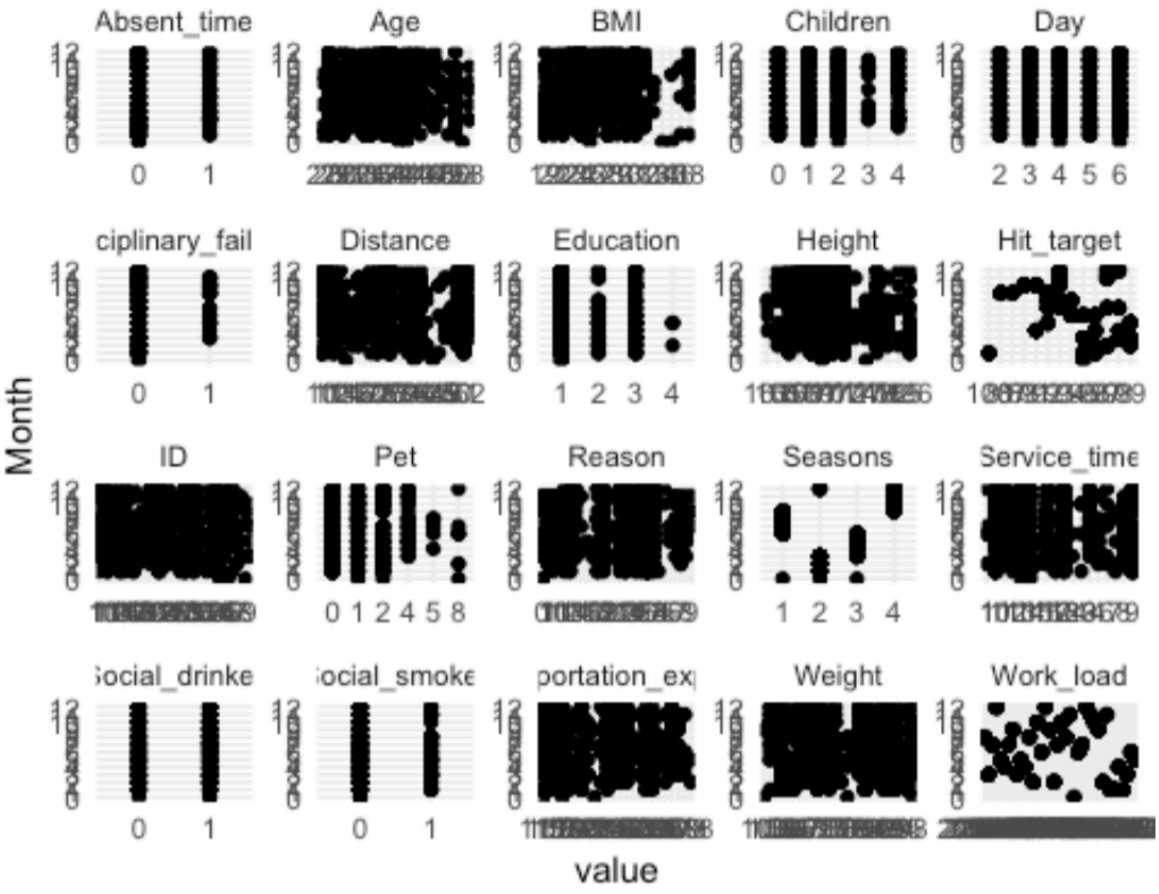


Figure 4. Month plotted against all other variables. Each variable was tested as the response variable to note any correlations.

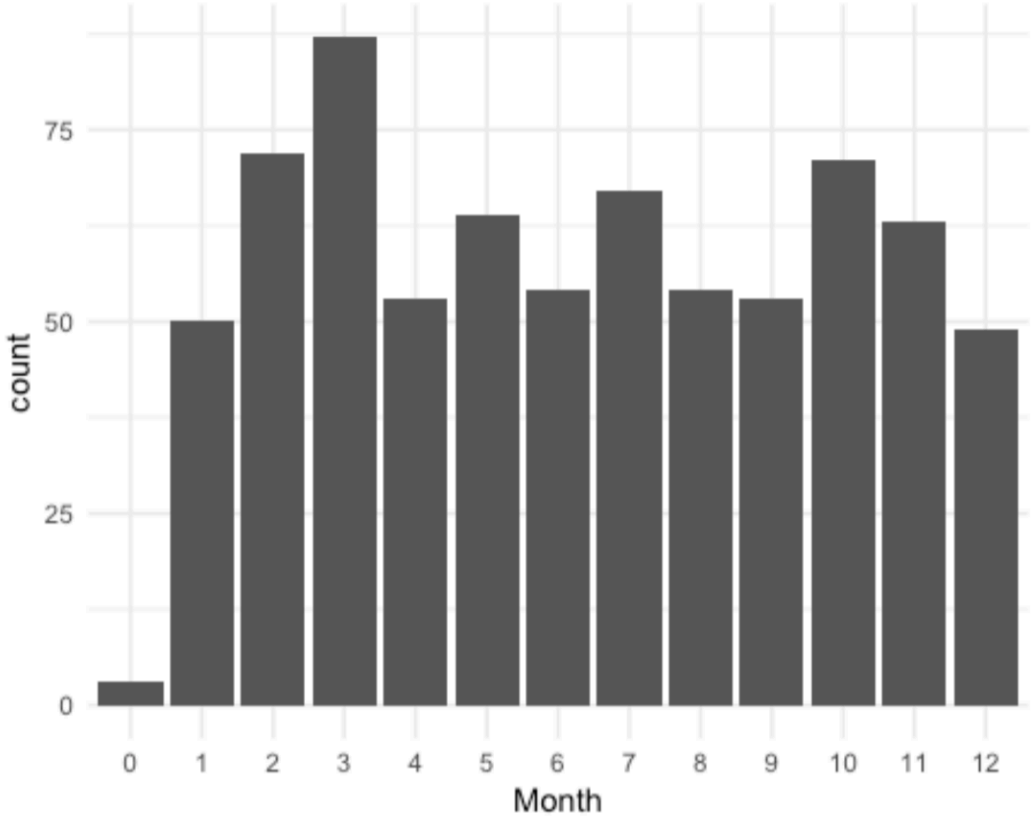


Figure 5. Histogram of Month observation counts. The peak in this graph is seen at Month=3, which can be interpreted as March.

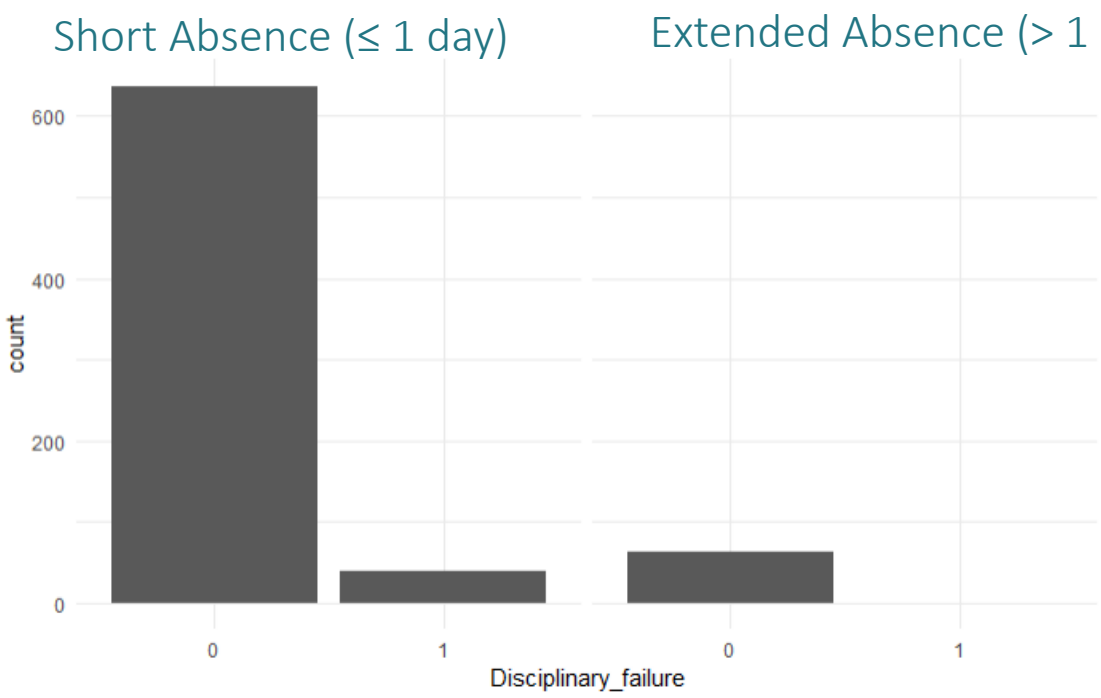
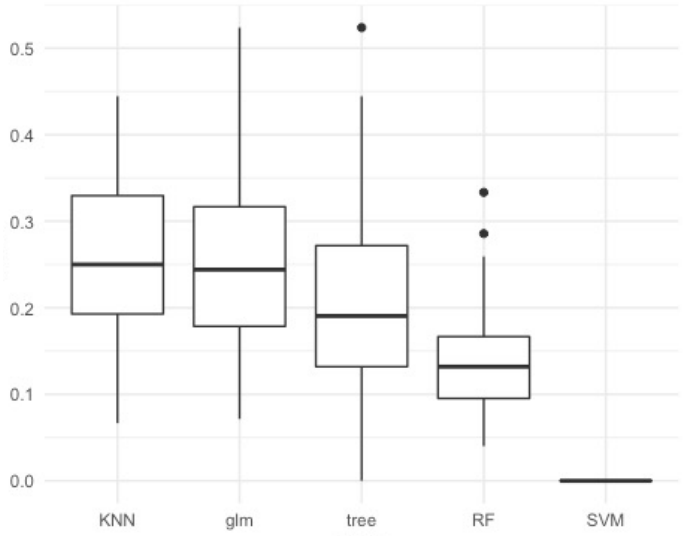
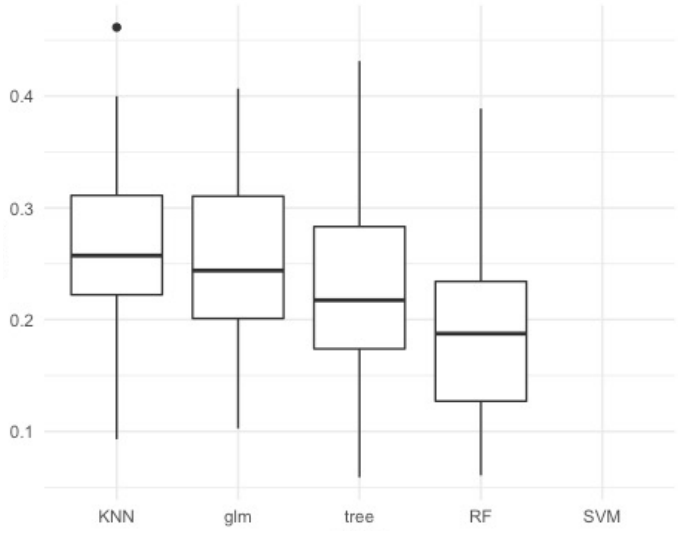


Figure 6. There are few counts of Disciplinary Failure in the majority class and none in the minority class, likely rendering this variable a poor predictor of absenteeism.

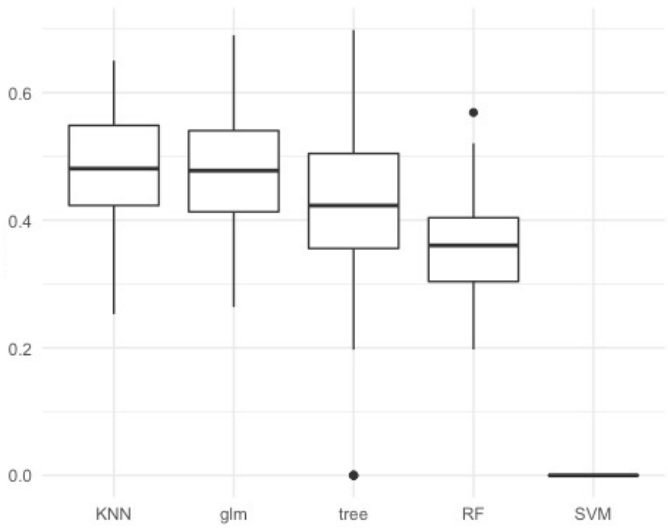
a)


0.5
0.4
0.3
0.2
0.1
0.0
KNN
glm
tree
RF
SVM

b)


0.4
0.3
0.2
0.1
KNN
glm
tree
RF
SVM

c)


0.6
0.4
0.2
0.0
KNN
glm
tree
RF
SVM

d)

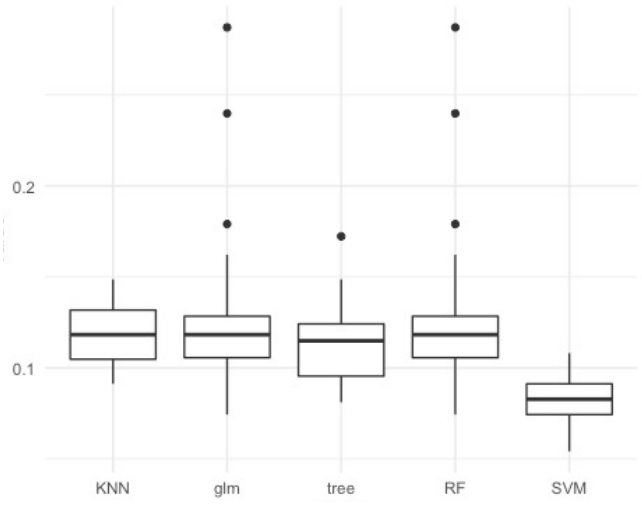


Figure 7. Comparative boxplots of key metrics. a) Sensitivity for KNN is 0.23. b) F-measure for KNN is 0.24. c) Geometric Mean for KNN is 0.47. d) Model error for KNN is 0.12.

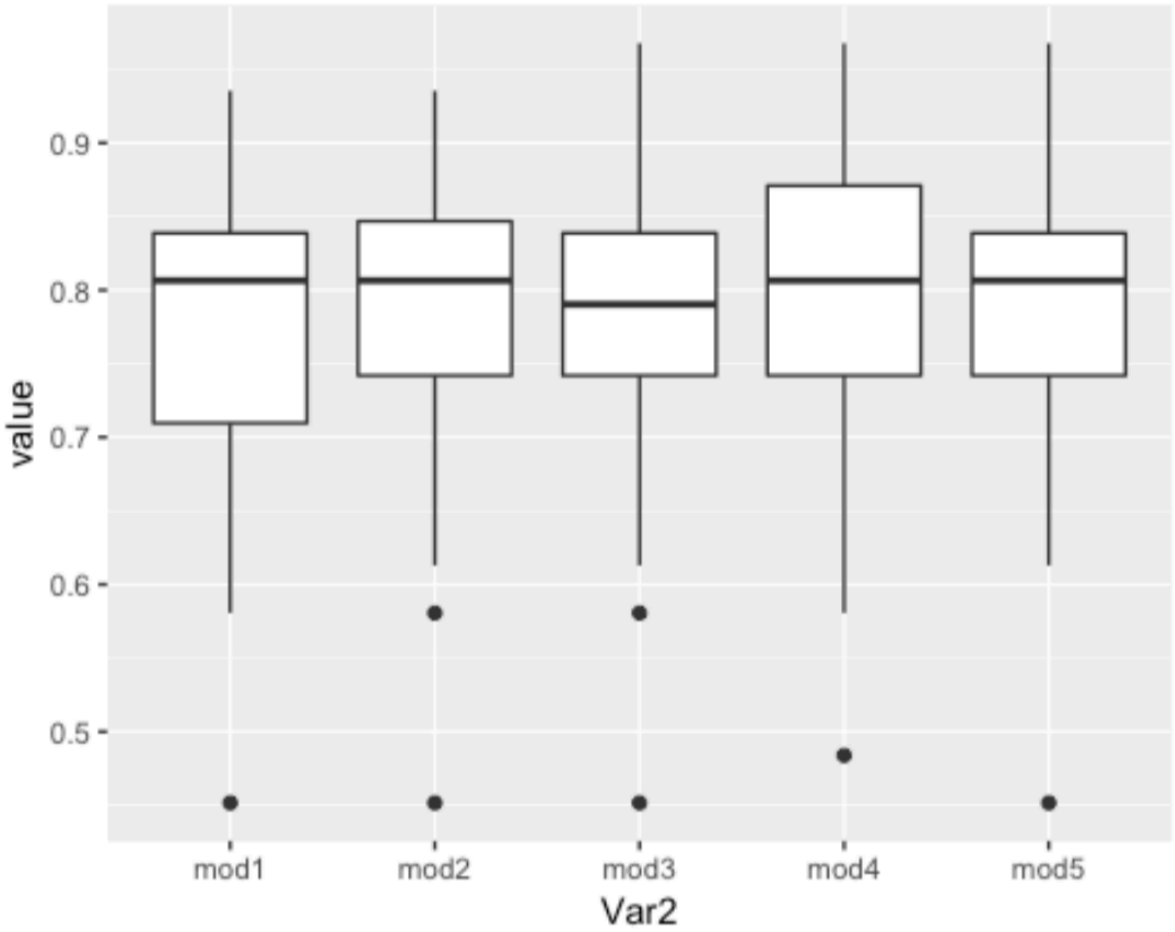


Figure 8. Comparison of sensitivity in the top 5 models after optimization.

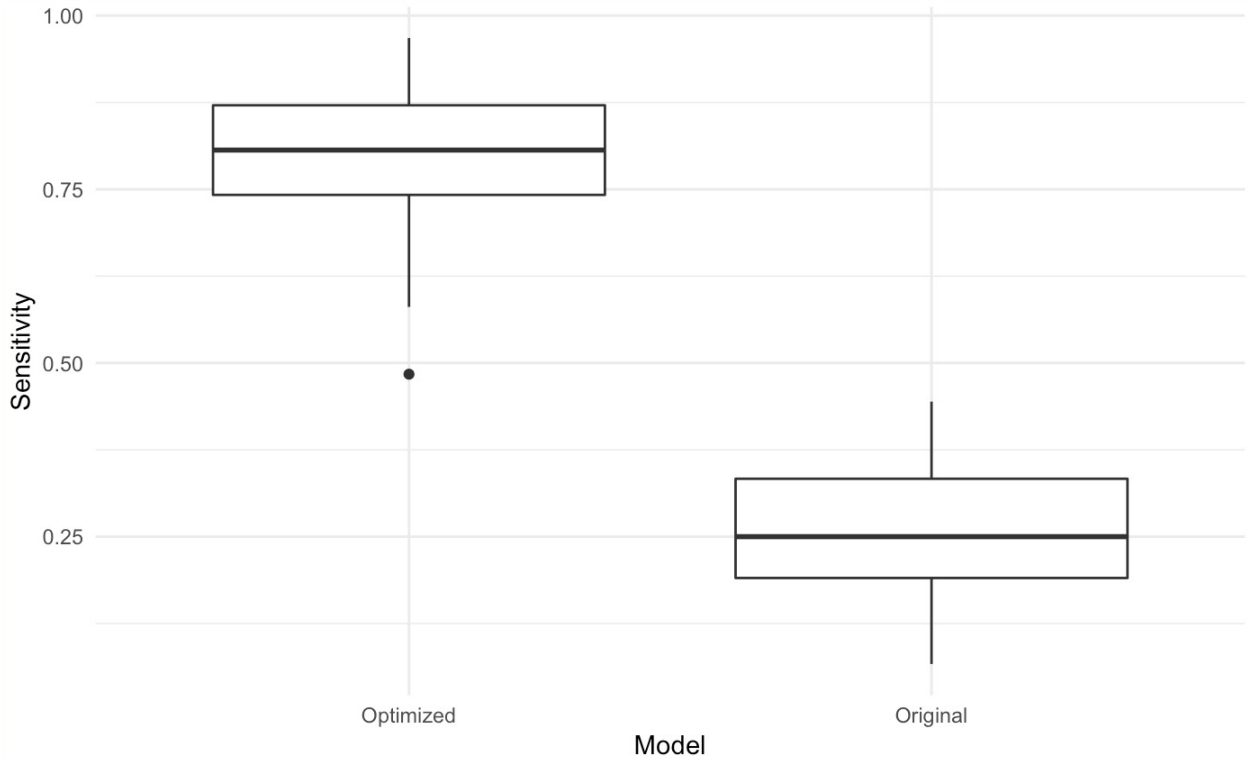


Figure 9: Sensitivity comparison of optimized and initial KNN models. The optimized model performs significantly better than the initial with respect to sensitivity.