Absenteesim Prediction

Nihar Garlapati
Harrisburg University of Science and Technology
Harrisburg,USA
NGarlapat@my.harrisburgu.edu

Divyani Kanawat
Harrisburg University of Science and Technology
Harrisburg,USA
DKanawat@my.harrisburgu.edu

Akhila Saineni
Harrisburg University of Science and Technology
Harrisburg,USA
ASaineni@my.harrisburgu.edu

***Abstract*—Absenteeism is defined as the practice of regularly staying away from work or school without a good reason. In this paper, we will attempt to predict the behavior of absenteeism at a particular organization by using various attributes associated with the employees such as reason for prior absenteeism, Month, Day & Season of absenteeism, family-based information such as number of children, pets and other characteristics of the individual such as social smoker or drinker and attributes such as Weight, Height of the individual.**

I. Introduction

In the new era of competitive organizations, the importance of employees has been constantly increasing. Employees are valuable assets of an organization and most times the key to success. The employers have understood that a content and a motivated employee brings in a lot more profitability and contribution to the organization. Having said that it is important for organizations to take care of the wellbeing of the employees for them to attract individuals to work in their respective organizations.

In this following report, we will be reviewing a dataset by a courier company that includes the absenteeism records of their employees and the reason for their absence. The reason for their absence has been certified with the ICD codes. The goal of the following project is to predict the number of absent hours based on the different attributes associated with the employee. Records of Absenteeism from work during the period of July/2007 to July/2010

II. Exploratory Data Analysis

A. Independent Variables

1. Individual identification (ID), 2. Reason for absence (ICD), 3. Month of absence, 4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6)), 5. Seasons (summer (1), autumn (2), winter (3), spring (4)), 6. Transportation expense, 7. Distance from Residence to Work (kilometers), 8. Service time, 9. Age, 10. Workload Average/day, 11. Hit target, 12. Disciplinary failure (yes=1; no=0), 13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4)), 14. Son (number of children), 15. Social drinker (yes=1; no=0), 16. Social smoker (yes=1; no=0), 17. Pet (number of pet), 18. Weight, 19. Height, 20. Body mass index

B. Dependent Variables

In this exercise, we will be predicting the number of absent hours. However, we will not be predicting the continuous variable of how many hours would an individual be absent, but instead we will bucket the absenteeism into 3 buckets.

1. Low (1)
2. Medium (2)
3. High (3)

When the number of absent hours is less than 6, We would classify such occurrences as low, between 6 to 12 as medium and anything above 12 will be classified as high

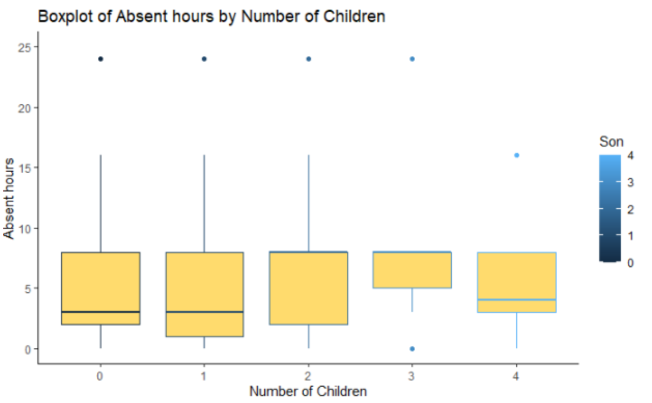
C. Correlation of the variables

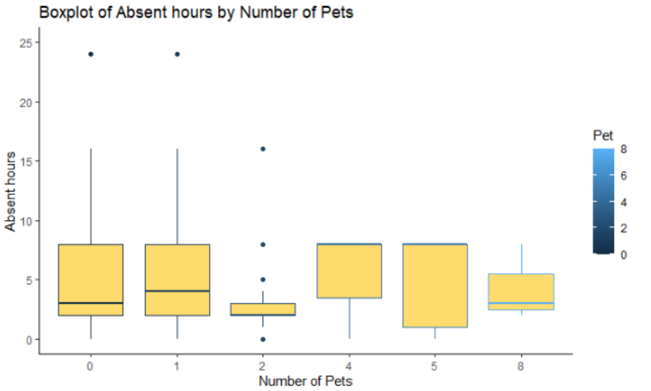
The correlation matrix enabled us to identify that Weight and BMI of the individuals are highly correlated. There were no other coefficients that are either very high. In order to avoid multicollinearity, we will be excluding the Weight of the individual from the modeling

D. Key observations of the dataset

The reasons most used by employees to be absent are reason 13,19,23, and 28. These reasons are Medical consultation, Dental consultation, diseases of the musculoskeletal system and connective tissue, and causes of morbidity and mortality.

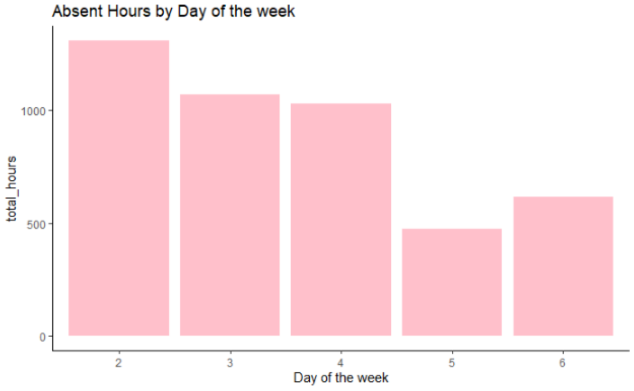
With more children and pets, the number of absent hours seem to be increasing

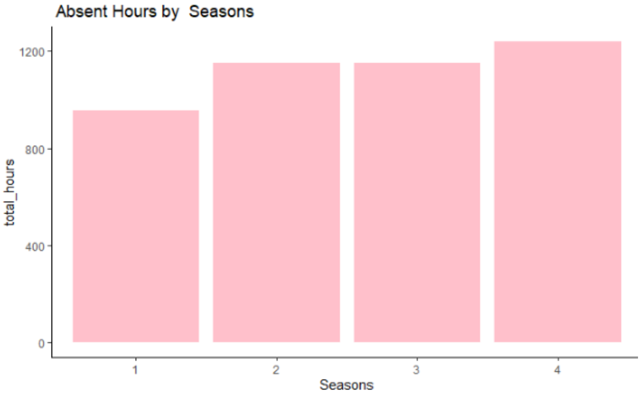


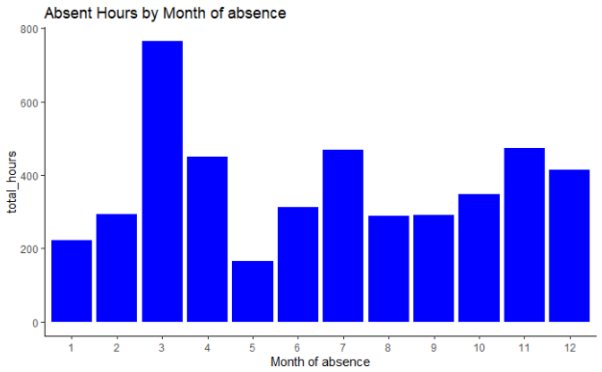


We have seen that the number of absent hours have variation based on different seasons, Day of Week also the month of the absence.

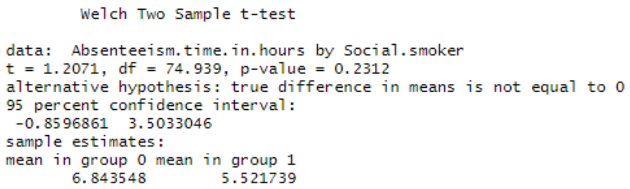
As you can see below, The Winter season has the most number of absent hours, This is inline with the CDC reporting which is basically a season for flu, Mondays seem to be having the highest number of absent hours and March is the month with the most number of absent hours.



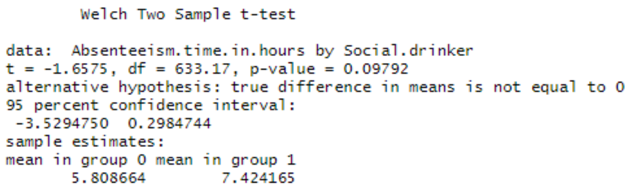




By running a two-sample t test, We were able to identify that there is no significant difference in the number absent hours based on weather an individual is either a social smoker or drinker.



```
        Welch Two Sample t-test

data:  Absenteeism.time.in.hours by Social.smoker
t = 1.2071, df = 74.939, p-value = 0.2312
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.8596861  3.5033046
sample estimates:
mean in group 0 mean in group 1 
       6.843548        5.521739 
```



```
        Welch Two Sample t-test

data:  Absenteeism.time.in.hours by Social.drinker
t = -1.6575, df = 633.17, p-value = 0.09792
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.5294750  0.2984744
sample estimates:
mean in group 0 mean in group 1 
       5.808664        7.424165 
```

E. Data Preperation

The dataset consisted of <10 missing values from the age, weight and hit target variables. All of the variables were continuous, and the number of missing values is very low, Therefore we have performed a mean substitution to replace the missing values with the mean of the variable.

Based on the variability in the workload average/ day and service time, we will be scaling the data from 0 to 1 for these 2 variables

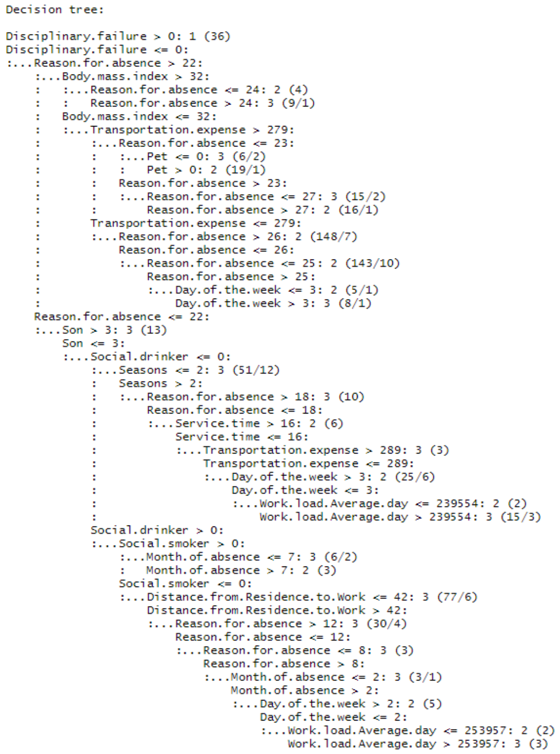
III. Modeling

In order for to achieve the best results, We have implemented several classification models and have picked the best model based on the accuracy, Sensitivity, Recall and F1 scores.

A. Tree Based Classification

The first model that has been implemented was the tree based classification model, The model yielded an accuracy of

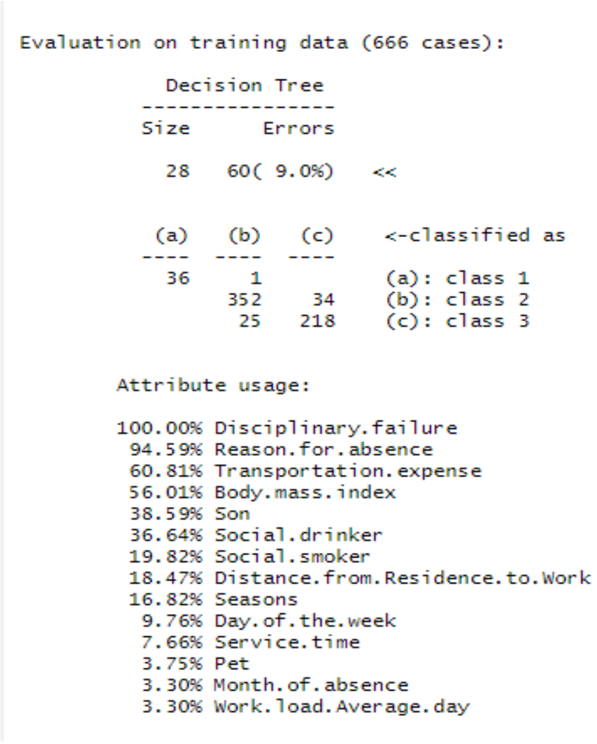
74% and the decision tree of the model is mentioned below.



```
Decision tree:

Disciplinary.failure > 0: 1 (36)
Disciplinary.failure <= 0:
:...Reason.for.absence > 22:
    :...Body.mass.index > 32:
    :   :...Reason.for.absence <= 24: 2 (4)
    :   :   Reason.for.absence > 24: 3 (9/1)
    :   Body.mass.index <= 32:
    :   :...Transportation.expense > 279:
    :       :...Reason.for.absence <= 23:
    :       :   :...Pet <= 0: 3 (6/2)
    :       :   :   Pet > 0: 2 (19/1)
    :       :   Reason.for.absence > 23:
    :       :   :...Reason.for.absence <= 27: 3 (15/2)
    :       :       Reason.for.absence > 27: 2 (16/1)
    :       Transportation.expense <= 279:
    :       :...Reason.for.absence > 26: 2 (148/7)
    :           Reason.for.absence <= 26:
    :           :...Reason.for.absence <= 25: 2 (143/10)
    :               Reason.for.absence > 25:
    :               :...Day.of.the.week <= 3: 2 (5/1)
    :                   Day.of.the.week > 3: 3 (8/1)
    Reason.for.absence <= 22:
    :...Son > 3: 3 (13)
        Son <= 3:
        :...Social.drinker <= 0:
            :...Seasons <= 2: 3 (51/12)
            :   Seasons > 2:
            :   :...Reason.for.absence > 18: 3 (10)
            :       Reason.for.absence <= 18:
            :       :...Service.time > 16: 2 (6)
            :           Service.time <= 16:
            :           :...Transportation.expense > 289: 3 (3)
            :               Transportation.expense <= 289:
            :               :...Day.of.the.week > 3: 2 (25/6)
            :                   Day.of.the.week <= 3:
            :                   :...Work.load.Average.day <= 239554: 2 (2)
            :                       Work.load.Average.day > 239554: 3 (15/3)
            Social.drinker > 0:
            :...Social.smoker > 0:
                :...Month.of.absence <= 7: 3 (6/2)
                :   Month.of.absence > 7: 2 (3)
                Social.smoker <= 0:
                :...Distance.from.Residence.to.Work <= 42: 3 (77/6)
                    Distance.from.Residence.to.Work > 42:
                    :...Reason.for.absence > 12: 3 (30/4)
                        Reason.for.absence <= 12:
                        :...Reason.for.absence <= 8: 3 (3)
                            Reason.for.absence > 8:
                            :...Month.of.absence <= 2: 3 (3/1)
                                Month.of.absence > 2:
                                :...Day.of.the.week > 2: 2 (5)
                                    Day.of.the.week <= 2:
                                    :...Work.load.Average.day <= 253957: 2 (2)
                                        Work.load.Average.day > 253957: 3 (3)
```

The model is considering the first branch as disciplinary failure, followed by reason for absence, BMI etc.

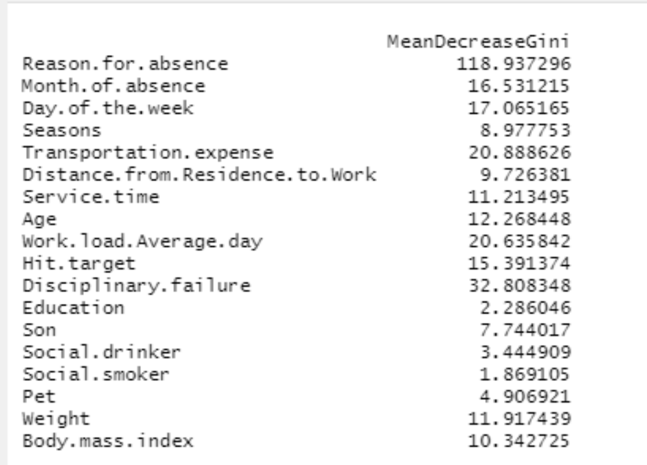
While building the following model, We have excluded weight to avoid multicollineearity. The usage of the attribute is mentioned below



```
Evaluation on training data (666 cases):

            Decision Tree
          ----------------
          Size      Errors

            28   60( 9.0%)   <<


           (a)   (b)   (c)    <-classified as
          ----  ----  ----
            36     1          (a): class 1
                 352    34    (b): class 2
                  25   218    (c): class 3


    Attribute usage:

    100.00% Disciplinary.failure
     94.59% Reason.for.absence
     60.81% Transportation.expense
     56.01% Body.mass.index
     38.59% Son
     36.64% Social.drinker
     19.82% Social.smoker
     18.47% Distance.from.Residence.to.Work
     16.82% Seasons
      9.76% Day.of.the.week
      7.66% Service.time
      3.75% Pet
      3.30% Month.of.absence
      3.30% Work.load.Average.day
```

B. *Random Forest*

The random forest model has yielded an accuracy of 72.9% and the importance of the variables in this model is mentioned below.



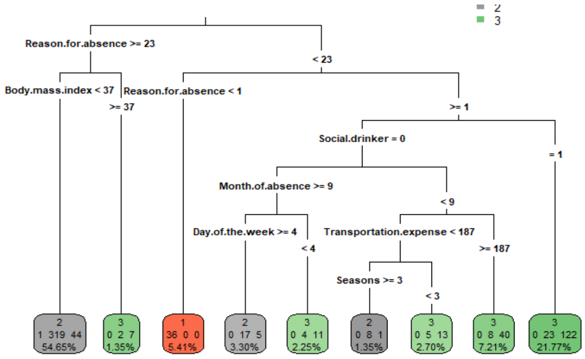
```
                                MeanDecreaseGini
Reason.for.absence                    118.937296
Month.of.absence                       16.531215
Day.of.the.week                        17.065165
Seasons                                 8.977753
Transportation.expense                 20.888626
Distance.from.Residence.to.Work         9.726381
Service.time                           11.213495
Age                                    12.268448
Work.load.Average.day                  20.635842
Hit.target                             15.391374
Disciplinary.failure                   32.808348
Education                               2.286046
Son                                     7.744017
Social.drinker                          3.444909
Social.smoker                           1.869105
Pet                                     4.906921
Weight                                 11.917439
Body.mass.index                        10.342725
```

Based on the GINI score, we understand that the ICD coded reason for absence is the most important feature followed by disciplinary failure & transportation expense.

c. *Regression Tree*

The classification tree with regression has yielded the highest accuracy of 75.6%. The model starts with classifying based on the reason for absence, which seems to be the most important feature even in the random forest model, followed by BMI and Social drinker.

The tree is provided below



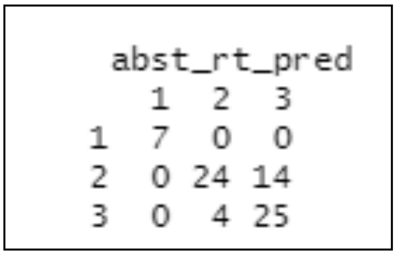
d. *Other Models*

There were other models which were also used such as Naïve Bayes (69%), SVM with Vanilla dot (67.4%), Poly Dot (67%) and RBF dot (66%) & K nearest neighbor (47%). Based on the number of variables used and the complexity of the dataset the above-mentioned models have not yielded high accuracy and therefore will not be discussed further in the following paper.

IV. Conclusion

Based on the above-mentioned findings, we would conclude that the regression tree algorithm has provided the best results for the following dataset in order to predict the absenteeism bucket.
Accuracy 75.6%
Precision (PPV) for Group 2: 63% Group 3:86%
Recall (TPR) for Group 2: 85% Group 3:65%
F1 Score: 0.73

Confusion Matrix:



```
   abst_rt_pred
     1  2  3
  1  7  0  0
  2  0 24 14
  3  0  4 25
```

We have also noticed that there is a very high probability that the model will differentiate individuals with less or greater than 6 absence hours more accurately, since most false positives and false negatives are in the second (Medium) and third (High) buckets.

V. Takeaway

Based on the GINI score of the random forest model & also the usage of the attribute in the regression tree model, we have noticed that reason for absence remains to have at most importance, therefore it is important for organizations to monitor the health and well being of the individuals to create a happier and productive work environment.

Transportation Expense seems to be of high importance; however, this aspect depends on many other aspects such as the city of the workplace along with the socio-economic status of the employee.
Disciplinary Failure also has high importance, both employees and employers need to work together to address such issues.
There is a clear pattern identified where the number of absent hours is varied between different seasons, day and month. Employers can pay attention to such patterns and plan accordingly.
BMI seems to also impact the absenteeism dataset and it is very clear that healthier individuals tend to be less absent.
Children and Pets are definite responsibilities, can result in more absent hours.
Social drinker and smoker are not very significant but seem to impact the number of absent hours.

VI. References

[1] G. Ashish "Employees Are Valuable Assets Of An Organization And The Key To Success" February, 2015

[2] V.Kushal "Tree-Based Methods: Regression Trees" March 2019

[3] T. Neelam "Understanding the Gini Index and Information Gain in Decision Trees" March 2020