Predicting Workplace Absenteeism in Brazil: Insights from Employee Behavior and Productivity Data

Shalini Singh and Upashana Suresh Kumar

2024-12-13

1 Introduction

1.1 Project Description and Domain

This dataset is relevant to the field of Data Analytics, specifically in the Human Resources field. It has been used in academic research at the Universidade Nove de Julho - Postgraduate Program in Informatics and Knowledge Management in Brazil. For this project, we are analyzing the dataset to understand and predict patterns of workplace absenteeism, focusing on employee behavior, productivity, and factors that affect absenteeism. Our primary question of interest is: What are the key factors that influence workplace absenteeism? To delve deeper, we examine a sub-question: How do health conditions and demographic factors contribute to absenteeism?

1.2 Overview of Analysis

The data was analyzed through linear regression and KNN (K-nearest neighbors), both using cross-validation. The focus of this analysis is the outcome variable in the dataset: absenteeism time in hours.

For the linear regression model, cross-validation was used, and the model was manually refined by selecting the three most significant predictors. These predictors were determined using the varImp function, which ranks predictors by their importance in the model. In the case of linear regression, varImp evaluates importance based on absolute value of the t-statistic of each variable.

Similarly, the KNN model was chosen through cross-validation and the use of the varImp function. For KNN, varImp assesses importance based on each predictor's contribution to model accuracy and its role in distinguishing neighbors when predicting the outcome. The optimal K-value for the KNN model with the significant predictors was found to be $K = 15$, balancing model complexity and predictive performance.

By applying the varImp function to both linear regression and KNN, we ensured that the selection of predictors and model parameters was statistically driven, optimizing the models for predicting absenteeism time in hours.

1.3 Dataset Description

The dataset, sourced from a courier company in Brazil, comprises 20 variables and 740 observations collected between July 2007 and July 2010. It includes both numeric and categorical data types. Demographic variables such as age, sex, and BMI are complemented by behavioral variables like drinking, smoking, along with contextual variables such as seasons, month of absence, and transportation expenses.

The outcome variable, absenteeism time, is numeric and measured in hours, with values ranging from 0 to 120. The summary statistics are displayed below:

Table 1: Summary Statistics for Absenteeism Time

Stats	Values
Min.	0.00
1st Qu.	2.00
Median	3.00
Mean	6.294
3rd Qu.	8.00
Max.	120.00

Notably, the dataset contained no missing values or obvious outliers, ensuring our data was well-suited for analysis. The dataset contained a disciplinary failure column which we did not include in our analysis. This is due to observed collinearity between the disciplinary failure and reason for absence predictors. Every instance of 1 for disciplinary failure correlates to a 0 value for reason for absence. We also omitted the ID column in our analysis.

The categorical variables in the dataset include season (levels: 1 = Winter, 2 = Spring, 3 = Summer, 4 = Fall), education (levels: 1 = High School, 2 = Graduate, 3 = Postgraduate, 4 = master and doctor), social drinker (levels: 0 = No, 1 = Yes), social smoker (levels: 0 = No, 1 = Yes), day of the week (levels: 2 = Monday, 3 = Tuesday, 4 = Wednesday, 5 = Thursday, 6 = Friday), month of absence (1 = January, 2 = February, 3 = March, 4 = April, 5 = May, 6 = June, 7 = July, 8 = August, 9 = September, 10 = October, 11 = November, 12 = December), ID (levels: 1 to 36), reason for absence (levels: 0 to 28) where each level correlates to a medical issue, which can be found in table 2 below.

Table 2: Reason For Absence

Levels	Medical Reasons
0	None
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 behavioral 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
22 to 28	Unknown

The remaining variables are numeric, including age (ranging from 27 to 58), weight (ranging from 56 to 108

kg), height (ranging from 163 to 196 cm), transportation expense (ranging from 118 to 388), distance from residence to work (ranging from 5 to 52), service time (ranging from 1 to 29), work load average day (ranging from 205.9 to 378.9), hit target (ranging from 81 to 100), son (ranging from 0 to 4), pet (ranging from 0 to 8), and body mass index (ranging from 19 to 38).

Upon initial analysis, we identified reason for absence as a key variable of interest in relation to absenteeism time in hours. Below are the corresponding plots and descriptive statistics for this variable, as well as a plot illustrating the distribution of absenteeism time by employee ID.

The plots reveal important patterns in absenteeism, highlighting key areas of interest. For instance, the Reason for Absence plot indicates that absenteeism is significantly higher among employees with certain health issues. The second plot identifies specific employees who took more days off, providing valuable insights into individual absenteeism trends. (see Figure 1 and Figure 2)

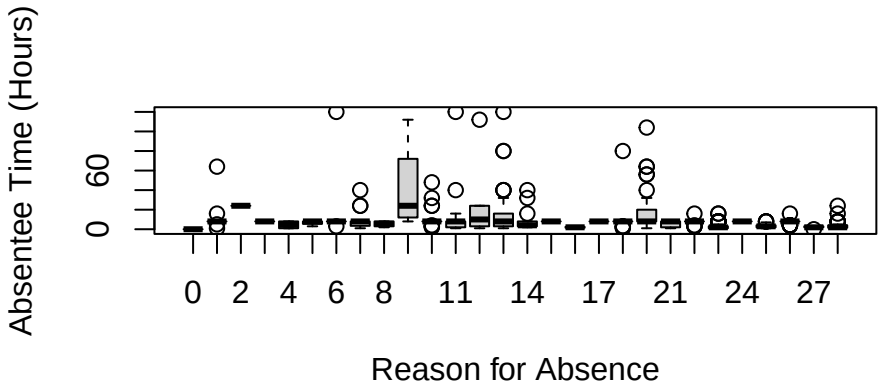

Figure 1: Absenteeism Time Categorized by Reason for Absence

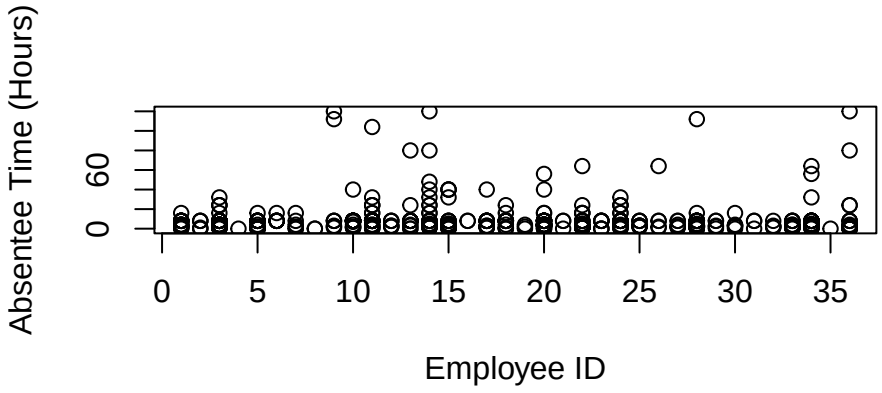

Figure 2: Absenteeism Time Categorized by Employee ID

2 Model 1: Fitting Linear Regression with Cross Validation

We fitted a Linear Regression model, incorporating all variables as predictors to predict absenteeism time in hours. We applied 10-fold cross-validation.

2.1 Linear Regression Model Fit With All Predictors

Table 3: Relative Importance of Predictors for Linear Regression Model.

Predictor	Overall
Reason.for.absence9	100.0000000
Reason.for.absence19	99.3689948
Reason.for.absence13	94.9799155
Reason.for.absence12	73.6069497
Reason.for.absence11	59.2736992
Reason.for.absence6	58.8169350
Reason.for.absence10	52.9133933
Reason.for.absence22	44.2550779
Reason.for.absence7	42.8994341
Reason.for.absence18	42.5444870
Reason.for.absence14	39.9079563
Reason.for.absence1	39.6493641
EducationPostgraduate	37.9632955
ID	37.5194831
Age	35.2174705
Reason.for.absence26	34.7195216
Day.of.the.weekThursday	33.2725976
Reason.for.absence27	30.9658247
EducationGraduate	30.2620088
Reason.for.absence2	29.7670840
Son	27.9253938
Reason.for.absence23	26.9088123
Reason.for.absence8	23.2275720
Reason.for.absence21	22.1637186
Reason.for.absence25	21.7129980
Day.of.the.weekFriday	21.6513556
Reason.for.absence28	20.6172798
SeasonsSummer	18.2393911
Social.drinkerYes	17.6946000
Reason.for.absence24	16.8822952
Distance.from.Residence.to.Work	16.5064481
Reason.for.absence5	15.4586812
Work.load.Average.day	15.4557462
Pet	15.4382416
SeasonsFall	14.9135165
SeasonsSpring	13.2915000
Reason.for.absence17	11.8215283
Reason.for.absence4	11.7863477
Day.of.the.weekWednesday	11.7361097
Social.smokerYes	11.6352442
`EducationMaster/Doctor`	11.6023002

Predictor	Overall
Transportation.expense	10.6584084
Reason.for.absence3	10.0794891
Reason.for.absence15	8.5006729
Month.of.absenceMay	8.4586647
Body.mass.index	8.1155469
Reason.for.absence16	7.9470131
Month.of.absenceFebruary	6.7672341
Month.of.absenceJanuary	6.6195996
Hit.target	5.9233907
Month.of.absenceJuly	5.8856005
Day.of.the.weekTuesday	5.2390879
Height	5.0168174
Weight	3.9109138
Month.of.absenceMarch	3.6783758
Month.of.absenceApril	3.4845443
Month.of.absenceOctober	2.9224535
Service.time	2.9179157
Month.of.absenceNovember	2.6081495
Month.of.absenceJune	1.0326588
Month.of.absenceDecember	0.4835373
Month.of.absenceAugust	0.1571383
Month.of.absenceSeptember	0.0000000

The three most significant predictors are reason for absence, education, and age. These are the variables we included in our next model. We omitted the ID predictor in our analysis.

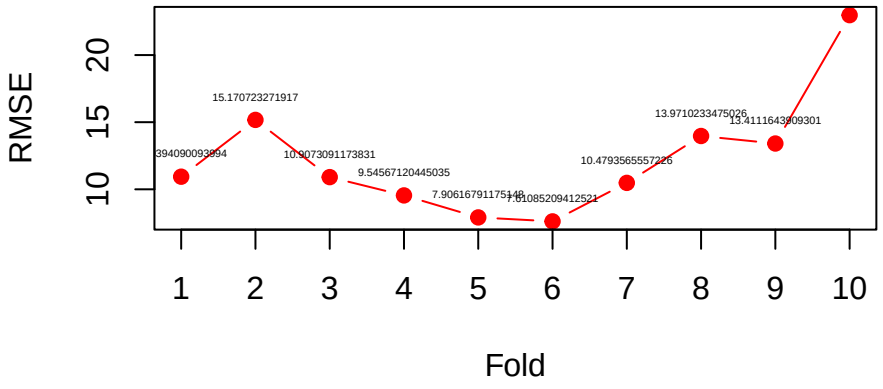

Figure 3: RMSE of Linear Regression Model Using All Predictors Through Cross Validation

The manual model has the lowest RMSE at fold 6 at 7.610852, making it the optimal fold. The overall performance of the folds is poor, aside from folds 5 and 6.

2.2 Linear Regression Model Fit With Significant Predictors:

We fitted a Linear Regression model, incorporating reason for absence, education, and age as predictors to predict absenteeism time in hours. We applied 10-fold cross-validation.

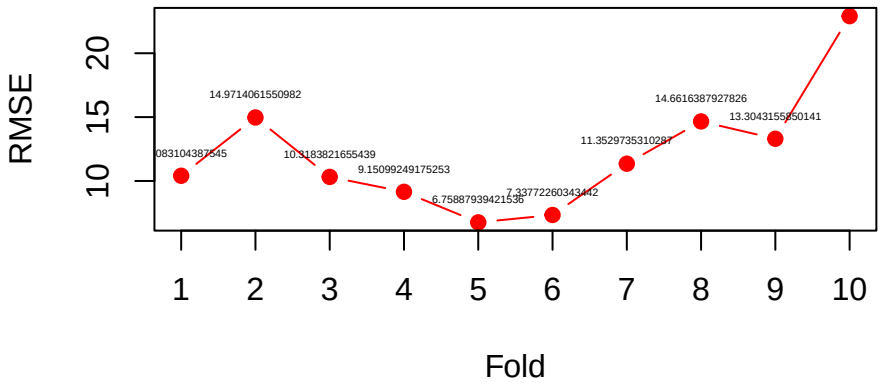

Figure 4: RMSE Linear Regression Model Using Significant Predictors Through Cross Validation

The manual model has the lowest RMSE at fold 5 at 6.758879, making it the optimal fold. The overall performance of the folds is poor, aside from folds 5 and 6.

The predictions from this model result in an RMSE of 6.190467.

2.3 Evaluating the Linear Regression Model

After predicting the fitted model on the reserved test data set, the RMSE is 6.190467. Reason for absence 9 (Diseases of the circulatory system), 2 (Neoplasms), and 12 (Diseases of the skin and subcutaneous tissue) are the most impactful predictors within this model, indicating that reason for absence is extremely significant to the outcome variable. Health conditions have a far-reaching impact on quality of life, so a correlation to absentee time makes sense.

The original model fit with all predictors has and RMSE of 7.610852 at the optimal fold, compared to our manually fitted model, which has an RMSE of 6.758879. This indicates that our fitted model is better suited to the data, signifying that the predictors we chose are significant.

3 Model 2: KNN with Cross Validation

3.1 KNN Model Fit With All Predictors:

We fitted a KNN model, incorporating all variables as predictors to predict absenteeism time in hours. We applied 10-fold cross-validation, repeated 3 times.

Table 4: Relative Importance of Predictors for KNN Model.

Predictor	Overall
Reason.for.absence	100.0000000
Height	59.8077061
Day.of.the.week	47.7984278
Distance.from.Residence.to.Work	39.9071057
Son	32.1373380
Age	24.0401530
Body.mass.index	7.1614848
Social.drinker	5.0873193
Hit.target	4.6514201
Education	4.4914220
Month.of.absence	4.4905614
Pet	2.8685880
Service.time	1.7868468
Weight	0.9653383
Transportation.expense	0.3889794
Seasons	0.3275883
Work.load.Average.day	0.1005658
ID	0.0595895
Social.smoker	0.0000000

The top three predictors by variable importance are: reason for absence, height, and day of the week. These are the variables included in our manual model.

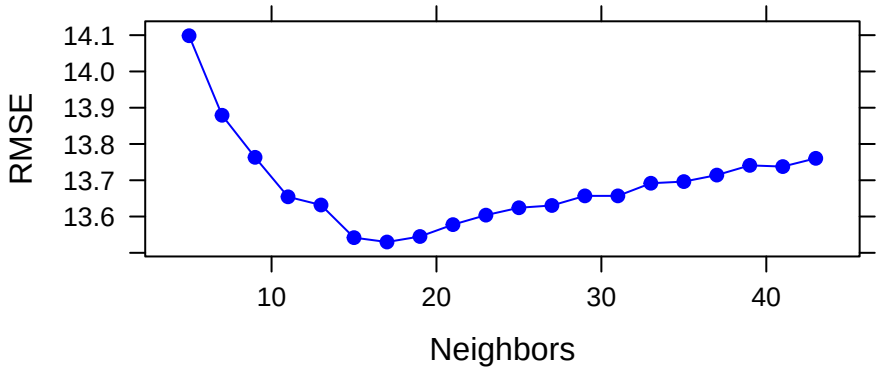

Figure 5: RMSE of KNN Model With Significant Predictors (Through Cross Validation)

The optimal value of K chosen by the model is K = 17. The RMSE for the optimal value of K is 13.52974.

3.2 KNN Model Fit With Significant Predictors:

We fitted a KNN model, incorporating reason for absence, height, and day of the week as predictors to predict absenteeism time in hours. We applied 10-fold cross-validation, repeated 3 times.

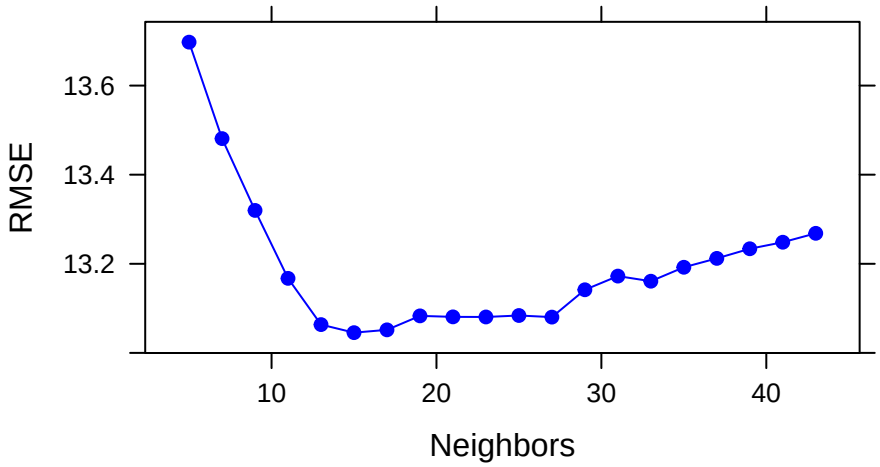

Figure 6: RMSE of KNN Model With Significant Predictors (Through Cross Validation)

The optimal value of K chosen by the model is K = 15, with an RMSE of 13.04544.

The predictions from this model result in an RMSE of 6.756024.

3.3 Evaluating the KNN Model

After predicting the fitted model on the reserved test data set, the RMSE is 6.756024. Originally, the fitted model had an RMSE of 13.04544 at the optimal k-value, compared to 13.52974 from the model built with all predictors. This indicates that the second model is fit better to the data, through eliminating less significant predictors.

4 Comparing Models

Table 5: Predictive RMSEs of Both Models.

Linear Regression	KNN
6.190467	6.756024

The predictive abilities of the linear regression model are better than the KNN model, making it a more suitable choice.

5 Conclusion

The models applied in this analysis—linear regression and K-Nearest Neighbors (KNN), both using cross-validation—provided valuable insights into the factors influencing absenteeism at work.

The linear regression model identified significant relationships between absenteeism and predictors such as "Reason for Absence," "Education," and "Age," with a Root Mean Squared Error (RMSE) value of 6.758879 at the optimal fold, indicating a moderate fit. Specifically, the model revealed that Reason for Absence 9—diseases of the circulatory system—had the strongest positive association with absenteeism, underscoring its critical impact. Employees with circulatory system-related conditions are significantly more likely to take leaves, which may be due to the chronic or severe nature of these illnesses that require medical attention or recovery.

The K-Nearest Neighbors (KNN) also identified significant relationships between absenteeism and predictors such as "Reason for Absence," "Height," and "Day of the Week," with a Root Mean Squared Error (RMSE) value of 13.04544 at the optimal k-value, indicating a relatively weaker fit compared to the linear regression model. The model specifically identified a unexpected predictor: height. This can reflect underlying correlations with health or demographic factors that are not directly captured in the dataset. This result showcases the KNN model's strength in identifying less obvious relationships within the data. However, the higher RMSE indicates that its predictions are less dependable compared to the linear regression model.

The predictive capabilities of these models are fairly similar, at 6.190467 for the linear regression model and 6.756024 for the KNN model, but the linear regression model is more accurate in predicting the outcome variable on a held-out test set.

Overall, the analysis demonstrates that health-related factors, particularly circulatory system diseases, are critical drivers of absenteeism. Absenteeism varies across different employees based on health-related factors, such as chronic conditions like circulatory system diseases, which have a significant positive impact. Other variables, like education, age, and even unexpected factors like height, also influence absenteeism, though their effects can vary in strength depending on the model used for prediction. While the linear regression model provides insights into these relationships, the KNN model highlights additional factors like height and contextual variables, offering a new perspective.

6 References

[1] DATA SET: Absenteeism at Work - UC Irvine Machine Learning Repository:(https://archive.ics.uci.edu/dataset/445/absenteeism+at+work)