Predicting Absenteeism in the Workplace

By: Myra Rust

Executive Summary

This report provides analysis and evaluation of machine learning techniques used to predict employee absentee hours. Methods of analysis include leveraging historical data, machine learning, and predictive modeling to forecast how many hours an employee will be absent based on demographic information. Multiple models were evaluated, and performance was measured on regression performance statistics, including R-squared. The best performing model was able to account for 4.3% of target variance. This result is not optimal and indicates that demographic information, at least that used in the project, is not sufficient to make accurate forecasts. However, analysis did identify a potential alternate forecasting model that could be used to make more accurate forecasts.

Introduction/Background

Workplace absenteeism negatively impacts businesses by reducing overall productivity, negatively impacting employee morale, and cutting down on business profits. “The Centers for Disease Control and Prevention (CDC) reports that productivity losses linked to absenteeism cost employers $225.8 billion annually in the United States, or $1,685 per employee.” (Stinson, C., 2015). This project will use data collected from a business on 36 employees and their absenteeism over a period of three years to conduct research using data science methodologies and techniques to identify possible causes and patterns in absenteeism.

The objective of this project is to answer the following research questions:

Can absenteeism be forecasted using employee demographic information?

What are the main factors of absenteeism?

Are there patterns in absence rates by season, month, or day of the week?

Providing employers with information on when higher absenteeism will occur and reasons surrounding the absences will provide them with information that can be used to assist in absence management and ultimately lead to better productivity, a more positive employee experience, and higher profit margins.

Preliminary Analysis

Data Source: The dataset contains 740 observations with 21 features detailing reasons for employee absences and personal/demographic information of the employee. This dataset was recorded by

a courier company in Brazil between July 2007 to July 2010.

Link to dataset: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

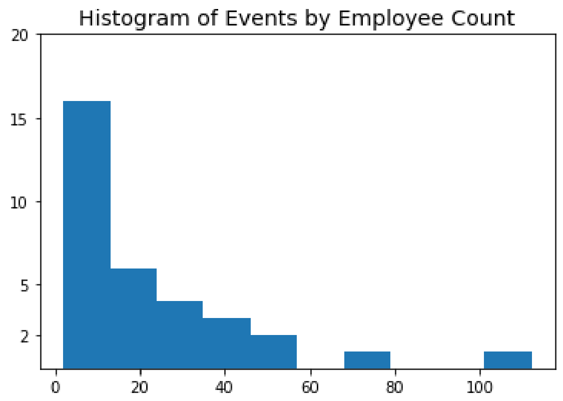
Data Cleaning: The dataset consisted mostly of numbers and required very little cleaning. 44 Observations with zero hours absent (target variable) were removed, column names were changed to be more user friendly, weight column was converted from kilograms to pounds, and height column was converted from centimeters to inches.

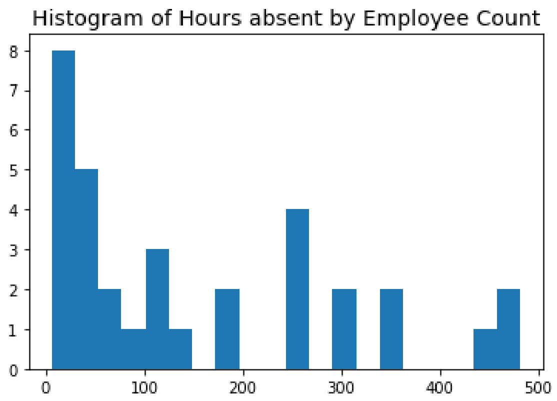
Features List:

1. Individual identification (id)
2. Reason for absence (reason) – Includes 21 categories linked to International Code of Diseases (ICD) medical reasons and 7 categories for other more-routine like medical events.
3. Month of absence (month)
4. Day of the week (weekday) - (2) Monday, (3) Tuesday, (4) Wednesday, (5) Thursday, (6) Friday
5. Seasons (season)
6. Transportation expense (trans_expense)
7. Distance from Residence to Work in Kilometers (distance)
8. Service time (service_time)
9. Age (age)
10. Daily Work load Average (workload)
11. Hit target (hit_target)
12. Disciplinary failure (discipline_fail) - (0) No, (1) Yes
13. Education (education) - (1) High school, (2) Graduate, (3) Postgraduate, (4) Master and doctor
14. Number of Children (children)
15. Social drinker (drinker) - (0) No, (1) Yes
16. Social smoker (smoker) - (0) No, (1) Yes
17. Number of Pets (pet)
18. Weight in pounds (weight)
19. Height in inches (height)
20. Body mass index (bmi)
21. Absenteeism time in hours (hrs_absent)

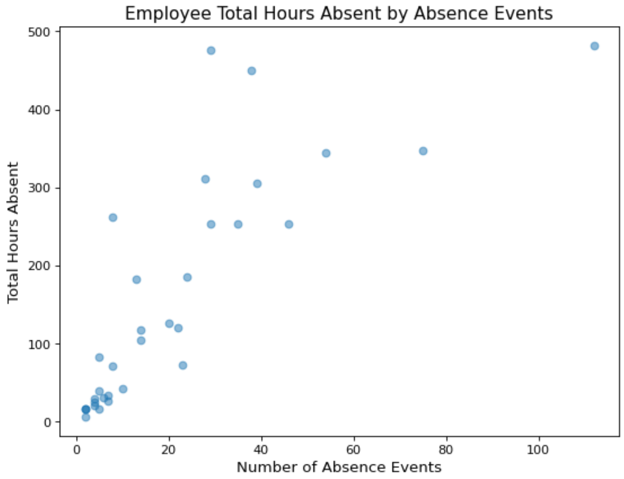
Preliminary Observations:

Workplace absences occur for many different reasons but there are individuals that incur more absences than others. This dataset tracked absences from 36 employees over three years. The following histograms depict employee counts for number of absence events and number of hours absent.

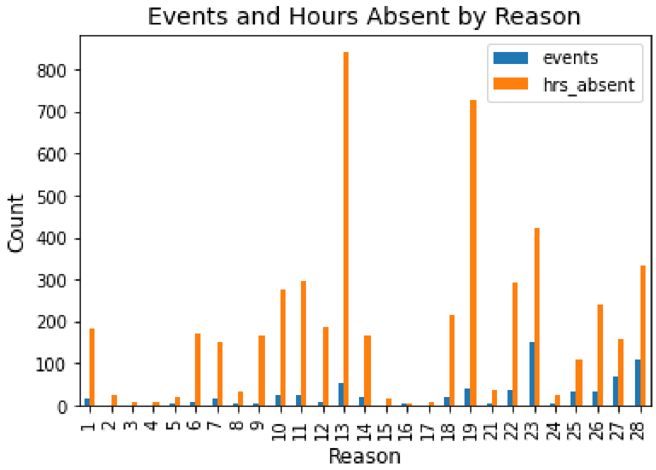

Histogram of Events by Employee Count


Histogram of Hours absent by Employee Count

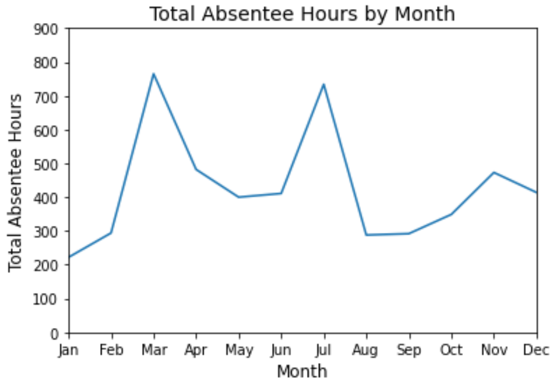
As you can see most employees had between 0-10 absence events, but there were some individuals that had more, including one employee that had 113 absence events in the three year period. As far as number of hours absent, we see the same trend. Most employees fell into the range of either 0-25 hours absent or 25-50 hours absent, but a few employees had over 400 absentee hours in three years. This coordinates with what is displayed in the below scatterplot. As the number of absence events increases, so does the number of hours absent, with the majority of employees falling in the lower left range and a small number of employees incurring the majority of absence events and hours absent.

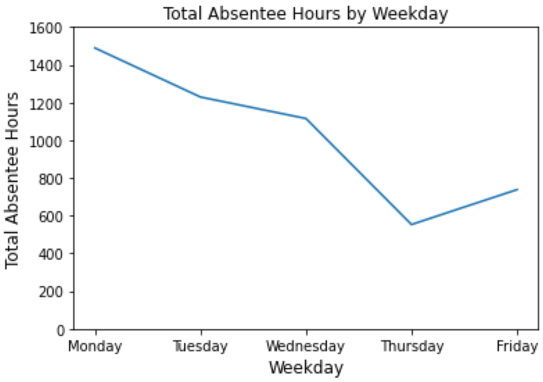

Employee Total Hours Absent by Absence Events

When viewing the medical reasons for the reported absence events, there are a few that seem to account for large numbers of events and the majority of employee hours absent. As shown in the bar plot below. Reasons 23 - medical consultation, 27 - physiotherapy, and 28 – dental consultation have the greatest number of events and account for ~47% of absence events. These events are all considered routine medical appointments. Reasons 13 - Diseases of the musculoskeletal system and connective tissue, 19 - Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified, and 23 – patient follow-up account for ~39% of all hours absent from work. A full list of medical reasons can be found in *Appendix A: List of Reasons*.



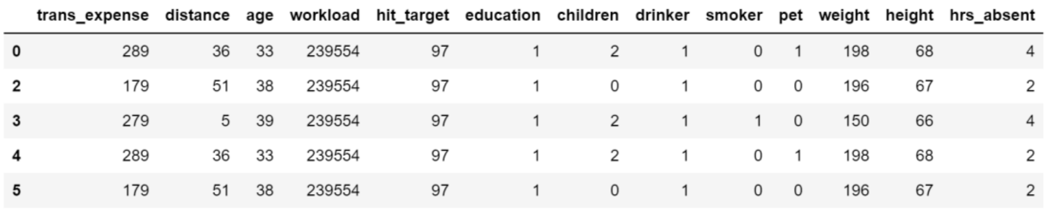
The season appears to have little connection to absentee hours, with the number of total hours absent fluctuating only ~300hrs, between seasons. However, when looking at monthly and day of the week absentee hours, we can see obvious trends. March and July have significant spikes in absentee hours and most absentee hours occur on Mondays, decreasing thereafter until Thursday with a slight rise again on Fridays.





Only a few features appeared to be correlated (age to service_time and weight to bmi). Due to this correlation and based on correlation to target service_time and bmi were removed from the dataset. Also, worthwhile to note is that most features had very little correlation with the target feature (hrs_absent) as well. The correlation heatmap and correlation to target feature visualizations can be found in *Appendix B: Correlation Visualizations*.

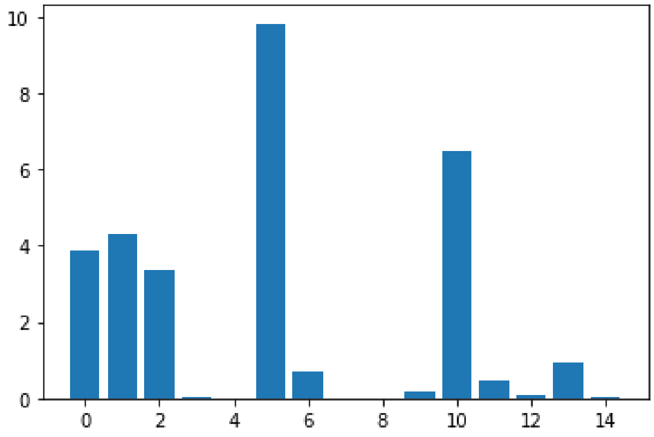
<u>Data Transformation</u>: To continue with our use case scenario features that would not be known at the time of forecasting were removed from the dataset to prevent bias and data leakage. These features included: reason, season, month, and weekday. The dataset at this point, can be viewed below. Once all data cleaning was complete categorical variables were transformed using one-hot encoding, the data was scaled using the MinMaxScaler function. Data was then divided into a 70/30 train-test split with 487 observations in the training set and 209 observations in the test set.



	trans_expense	distance	age	workload	hit_target	education	children	drinker	smoker	pet	weight	height	hrs_absent
0	289	36	33	239554	97	1	2	1	0	1	198	68	4
2	179	51	38	239554	97	1	0	1	0	0	196	67	2
3	279	5	39	239554	97	1	2	1	1	0	150	66	4
4	289	36	33	239554	97	1	2	1	0	1	198	68	2
5	179	51	38	239554	97	1	0	1	0	0	196	67	2

Feature Selection

A total of 15 features remained and correlation, mutual information, and decision tree were evaluated as potential methods for feature selection. Correlation feature selection provided the best results. Using KBest to determine feature importance, 5 features were shown to possess significantly more importance than the others, as seen in the image below. These features were: trans_expense, distance, age, children, and height.



Model Selection and Evaluation

The objective of this project is to forecast absentee hours expected from an employee based on demographic information. Since the value we are trying to forecast is a number, this is a problem of regression. Baseline regression, Linear regression, decision tree regression, and random forest regression models were all run and evaluated on Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R2), which represents the proportion of variance in the target variance that can be explained by the model.

Results

The results are provided in the following table. The linear regression model performed the best of all models, accounting for 4.3% of target variance. However, this is still a low R2 score and the results are less than desired to accurately predict employee absenteeism.

Model	MAE	RMSE	R2
Baseline Regressor			<0.00003
Linear Regression w/all features	0.057	0.122	0.027
Linear Regression w/selected features	0.055	0.121	0.043
Decision Tree Regressor	0.071	0.183	-1.198
Random Forest Regressor	0.060	0.140	-0.276

Conclusion

In conclusion, the obtained demographic information alone is not enough to effectively forecast absence rates by individual employee. Adding additional demographic information into the model could increase forecasting abilities, especially if that information was related to existing medical conditions. However, this type of information is protected by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and it is illegal in the United States for employers to request this information from employees.

Future Work

Explore creating prediction models using season, month, and weekday information as we can see that the analysis of those features had obvious trends and patterns. I believe that it could be used to adequately forecast employee absenteeism. Another option would be to collect additional features information. Examples: Position (junior, management, executive, CEO), Physical demand (desk-based, active movement, manual labor), or does the employer offer PTO and/or sick leave.

Acknowledgements

I would like to thank Professor Catherine Williams and all other professors that are part of Bellevue University's data science program. Without their direction, support, and guidance, this paper would not have been possible. I would also like to thank my daughter for sacrificing her time with me to allow me to pursue this degree for the betterment of myself.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

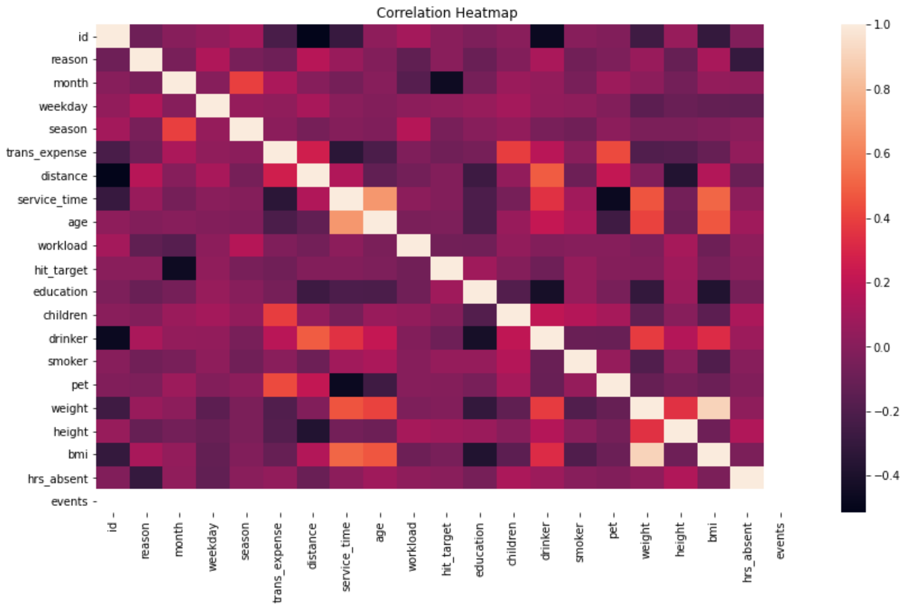
Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp.1-4). IEEE.

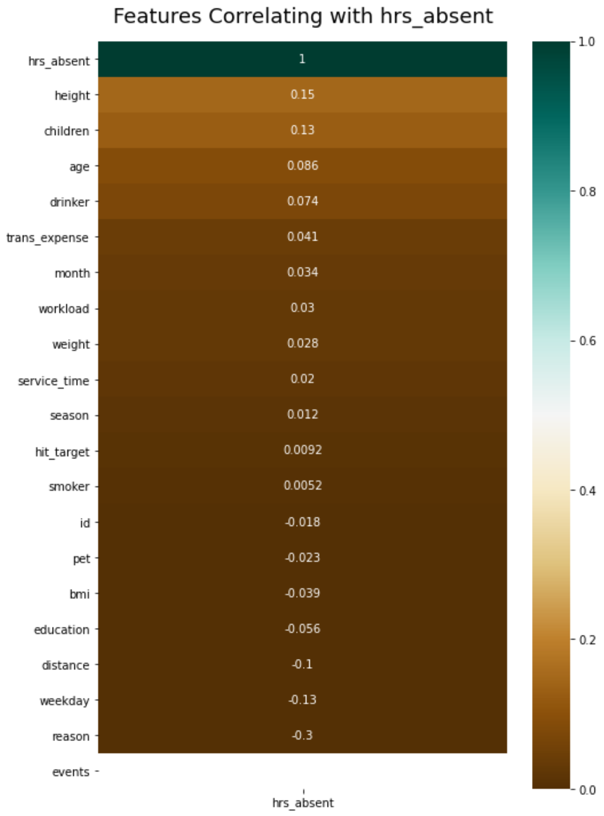
Stinson, C. (2015, January 28). Worker Illness and Injury Costs U.S. Employers $225.8 Billion Annually. CDC Foundation. Retrieved from https://www.cdcfoundation.org/pr/2015/worker-illness-and-injury-costs-us-employers-225-billion-annually.

Appendix A: List of Reasons

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. Patient follow-up
23. Medical consultation
24. Blood donation
25. Laboratory examination
26. Unjustified absence
27. Physiotherapy
28. Dental consultation

Appendix B: Correlation Visualizations


Correlation Heatmap
id
reason
month
weekday
season
trans_expense
distance
service_time
age
workload
hit_target
education
children
drinker
smoker
pet
weight
height
bmi
hrs_absent
events


Features Correlating with hrs_absent
hrs_absent 1
height 0.15
children 0.13
age 0.086
drinker 0.074
trans_expense 0.041
month 0.034
workload 0.03
weight 0.028
service_time 0.02
season 0.012
hit_target 0.0092
smoker 0.0052
id -0.018
pet -0.023
bmi -0.039
education -0.056
distance -0.1
weekday -0.13
reason -0.3
events
hrs_absent