Thomas P McGuire

IST707 Intro to Data Analysis

Winter 2021

Predicting Absenteeism at Work

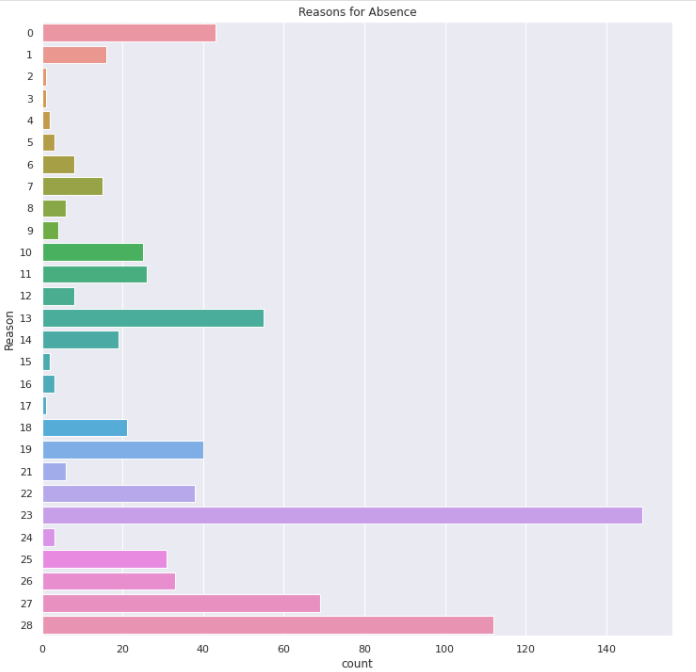
**INTRODUCTION**:

The goal of this report is to show how to tackle any data-mining problem with ease and efficiency. I will start with some exploratory data analysis after introducing the dataset. I will discuss how I preprocessed the data and how I determined which models I used in my attempts to classify the data. I will then discuss how the analysis went and all of my conclusions based on the data.

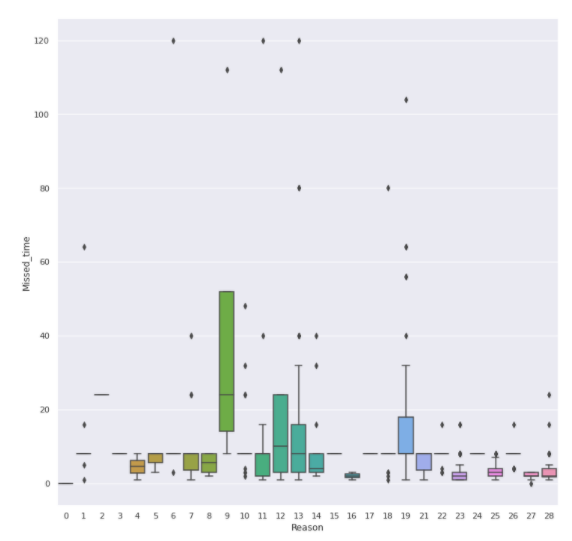
Absenteeism at work is a critical problem that every business faces. Employees becoming sick affect every possible business from fast food joints to the top of corporate America. During the past year of the pandemic, many companies were forced to adapt to stay at home orders. This has alleviated some of the causes that force employees to miss work. Unplanned employee absence can cause numerous problems in any given work place. Projects can be postponed or rushed through, feelings of resentment between employees for taking on additional work, mistakes of all kinds can be the result of unplanned any employee missing work. If these unplanned absences could be better understood, predicted and classified then perhaps we could classify why an employee might call out sick based on some of their traits and characteristics. Is it possible to predict why an employee might call off from work if we know simple facts like how many children they have or if they smoke and drink? By the end of this report, I hope to have an answer.

**DATASET DESCRIPTION:**

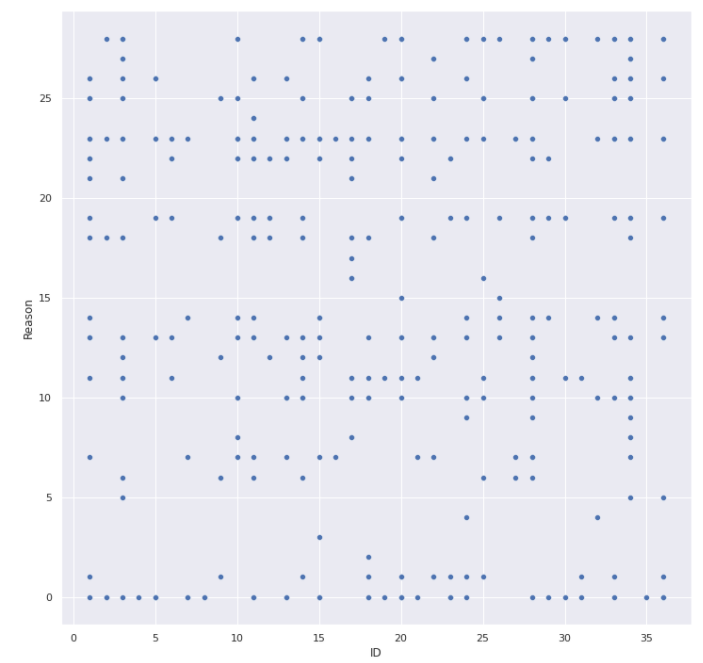
This dataset was taken from the UCI machine learning repository1 and was donated by a courier company in Brazil. The data comes from between July 2007 and July 2010. There were 28 different possible reasons for absence that had to be reported to International Code of Diseases, ranging from diseases of the digestive system to blood donations and dental consultations, along with a category for no reason. There were seven reasons that did not require a patient follow up. Those include laboratory examination, physiotherapy and medical consultation. There are 740 employee absences during this time period. Within the dataset, it offers several details about its employees, such as height, weight, if they have children or pets and if they smoke or drink. The list of variables also included details about their level of education, daily workload and the extent of their absence in hours.



As we can see from this histogram there is a large amount of variability between reasons for being sick. Only one reason had no data points associated with it, reason 20, which was External causes of morbidity and mortality. Among the rest of the data, there are 14 reasons that had less than 10 data points across all 740 rows, totaling 45 data points. This might cause enough problems within the data that they might need to be removed. I ran a handful of different plots and could not find any obvious correlation between any employee and why they might call out from work. There is some variety in how much time an employee missed for a given reason of absence, which makes sense as similar illnesses could affect people differently.



Based on employee ID, there are 36 employees working at this facility. One of the employee numbers could have been reused, as there was a sudden and drastic change in their age. Another employee has over 120 hours missing from work due to illness. My goal is to classify each absence by the reason of absence and if within the scope of this report, attempt to classify how many hours a person could be out of work for a given illness. While this is a relative small dataset of only 740 data points, I believe it could still leading to some very interesting conclusions.

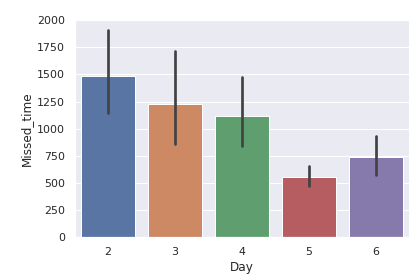


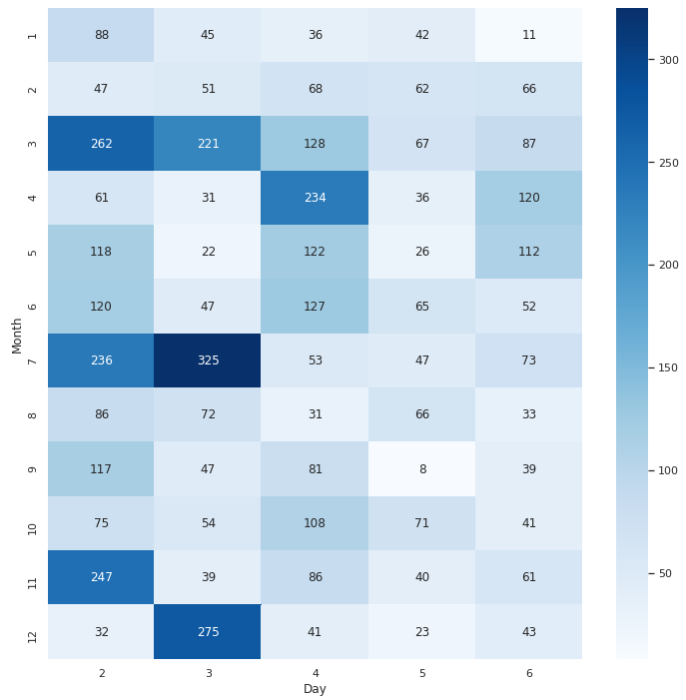
**DATA PREPROCESS:**

Because of where this dataset came from, there was no much need for preprocessing except to get the data arranged to my liking. Some of the columns names were changed to make the dataset less unwieldy. For the sake of cluster analysis, I scaled all columns with the exception of Reason for absence and the employee ID, which were then removed before the analysis. For the classification part, I separated the data into train and test sets as I only had one dataset to work with. The training set was 70% of the original data and the test set was 30%. Performing a quick calculation, I found that the no information rate was 20% so if a model performs worse than this, it is not a good model for this dataset.

**ANALYSIS EXPERIMENTS:**

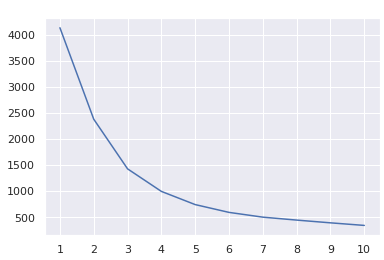
One of the first tests I ran while inspecting the data before running through the machine learning models was to see if there was any one day in particular that had the most absences. I was quite surprised to see that it was fairly evenly distributed over the 5 working days though Monday had the greatest amount of hours missed. Thursday had the least amount of time missed by employees by almost 200 hours. Tuesdays in July actually had the highest amount of missed time across the 3 years.



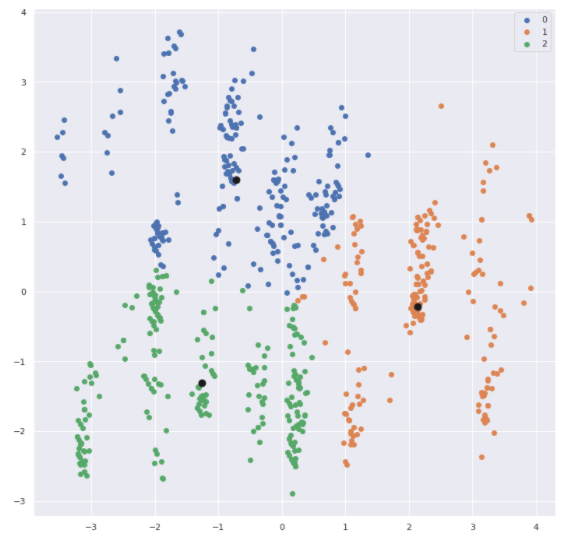


Clustering will be used to see if there is any correlation in this dataset between the reason for absence and several of the other factors, such as age, children, season, if the employee is a smoker or not and level of education. This dataset will then be run through several machine learning models to see if we can predict a reason for absence based on the rest of the factors. The models I have choose to use are Naïve Bayes, K-Nearest Neighbors, Support Vector Machine, Random Forest and Gradient Boosting.

**CLUSTERING RESULTS:**

****

I set up my Kmeans model and decided to use 3 clusters, as this was what the calculations suggested. I used PCA to convert the clusters into showable space in 2 dimensions. The confusion matrix for the accuracy of this model is hosted on Github2 along with the original dataset and the hosted code that I used for this report. Even without the employee id, I was happy with how the clusters formed. That is not to say that they were accurate but there is distinction between them all. On the confusion matrix, it does not seem to have clear patterns between the 3 clusters in terms of reason for calling out sick.

****

In the confusion matrix, there was a relatively even split of the original 740 points across the 3 clusters with no clusters dominating any of the reasons for absence. This is unsurprising as many of the employees called out sick for many different reasons, so each of the smaller clusters within a cluster could be each time the employee called out sick.

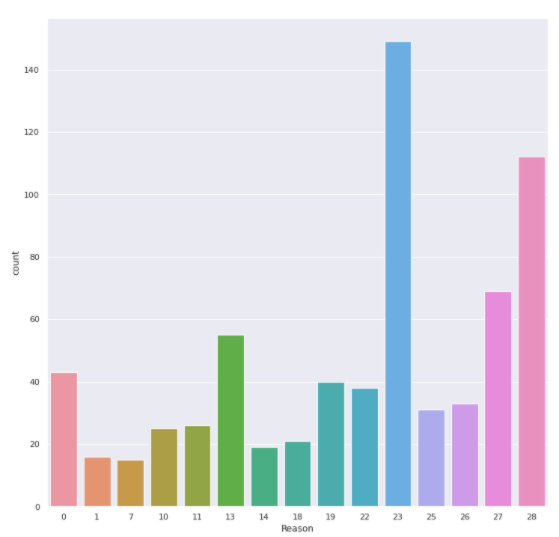
**CLASSIFICATION RESULTS:**

I ran 5 separate models on the data after gently cleaning the data to ensure no missing data points. K-nearest neighbors, Support vector machine, Naïve Bayes, gradient boosting and random forest were the models I chose to run. I set up them using their original conditions. I chose these models as these were the ones covered in class and they are some of the most common ones used in the field of data science.

I ran Naïve Bayes before the rest of the models, which gave an accuracy of 28.8%. After Naïve Bayes, I scaled the dataset in to ensure that all variables were taken into account evenly. I was quite shocked to see that none of the models could rank higher than 50%. The highest was RandomForest around 46 to 49%. KNN and SVM had model accuracies at 36% and 41%, respectively. Gradient boosting got about 42.8%.

I also attempted to run a deep learning model to see if that could perhaps pick up on some potential “hidden” variables that lay within the data. I was fairly disappointed by the result when the model was done. Not only was it far worse than the each of my more simple machine learning models, across 10 epochs, it failed to classify with an accuracy of greater than 3%. This could be to the size of the dataset rather than the dataset not having good patterns or my own failure in not setting up the deep learning model appropriately.

Out of curiosity for myself, and unsure if it was in scope of this project, I changed my target variable from Reason of absence to the total time in hours each person was out because perhaps there were not enough patterns in the data to correctly classify the reason for absence based on an employee’s details. I hoped that it could use the addition of the absence reason to predict how long an employee would be out of work. As the data was already used to run the models previously, the only preprocessing necessary was to change the target column. I also dropped all the rows that had reasons for absence that account for less than 10 across the 3 year period. About 45 rows were dropped, leaving me with around 695 rows and 15 Reasons for absence. The new value for no information gain was 26.5% with 8 hours of missed time being the most common.



When I switched to my new target and reran my models, I was surprised again. Random Forest remained about the same at 48%. KNN and SVM improved in their accuracy while Naïve Bayes and Gradient Boosting decreased in their accuracy. Perhaps it would be easier to predict how much sick time an employee would take if their ailment and employee details were known. Even my deep-learning model improved to around 15%. None of my models could break 50% accuracy and 2 of them failed the standard of being a better classifier than new no information gain.

**CONCLUSION**:

From my experiments with clustering and with classification using various machine-learning models, I believe I can safely say that this dataset is difficult for machines to perform with effectiveness. Or to take more accountability, I could not tweak the dataset to perform well in this text mining case.

If I had been more inspired and not switched datasets in the middle of this project, I would have broken this down more to see if I could accurately classify who smokes, who has children or pets. The dataset is complicated by the fact that across 740 rows in the dataset, there are only 36 employees working at this courier business. 3 out of the 36 employees did not take any time off during the 3 years or if they were sick, they did not miss any time at work. This means that across the dataset, there is a large repetition of data. Thus if classifying the data by reason it can lead to the models being confused why 2 otherwise identical rows of data would have two different classifications. There are some rows that had the same employee call out for different reasons but had different hours of absence. Thus, the reason for the second half of my experiment, trying to classify each absence by hours of absenteeism, which actually gave some of the rows some difference.

Another potential pitfall that could be causing the model problems is that this is attempting to track human behavior. Machines notoriously have difficulty tracking the human behavior. Perhaps some of these employees played hooky and gave a bogus reason. There is no reason to not trust these people, but even just a small fraction of employee calling out sick for illegitimate reasons would sway a dataset of this size. A quick search informed me that medical certification of illness is required and employees must present signed doctor’s medical certificates to their employers3. Perhaps this is why there were so many of the 7 reasons that did not require a patient follow up. I would be very interested in seeing where this company was in the country or seeing how a similar sized company in another country would have compared. There would be a lot of different comparisons, especially if the company in another country was roughly the same size and performed the same tasks. While it could lead to some less savory reasons if handled by poor management, being able to predict how long any illness could keep an employee from working could be beneficial to the other employees in knowing how long their coworker would be unable to work.

This is not to say that employees should not take time off from work. Burnout by itself can cause halts in productivity in addition to resentment towards management for not balancing workloads. Employees’ mental health is important. Scheduled absences should be encouraged for employees to reduce burnout and increase productivity. There has to be understanding by management to keep enough employee staffed so that the burnout is minimized, the resentment towards the other employees and management is small and employees do not feel compelled to come in to work feeling unwell, especially as we get closer to the end of this pandemic.

While I may not have gotten the results I wanted from the clustering and the machine learning models, I am fairly happy with how I attacked this problem. Modeling and accurately predicting human behavior is not going to ever be easy. There is some great business value in being able to predict how long someone will be out for say a doctor’s appointment or dental visit.

**Citations**:

1 https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

2 https://github.com/TomPMcGuire37/absenteeism

3 https://www.asinta.com/news/benchmark-glance-brazilian-employee-benefits-january-2016/#:~:text=Brazilian%20Sick%20Leave%20Requirements%3A%20Employees,by%20employers%20at%20full%20pay.&text=After%2015%20days%2C%20social%20security,employees%20are%20on%20unpaid%20leave.