Using Random Forest and a Autocorrelations for Approach Predicting Absenteeism from Work

Jillian Burke, Sebastian Brown, Reian Festejo, Shane Sharareh, & Wallace Ward

Introduction

In this study we use a dataset with the records of employees from a Brazilian courier company that have missed work. Our task is to determine which of the 19 predictive variables included in the dataset have the largest association with the rate of absenteeism at work, as well as to create a model that will predict how many hours an employee will miss based on the input variables. These steps are implemented using a random forest approach.

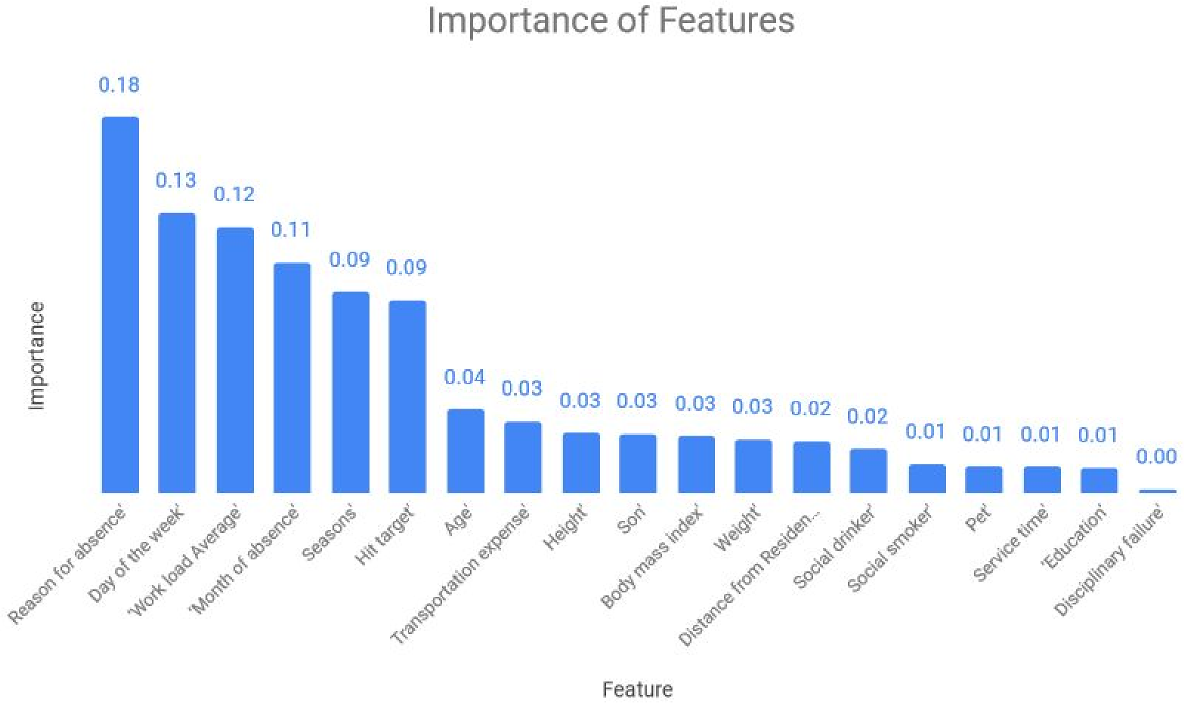
To examine if there are any patterns in the times at which workers are most likely to be absent, on a particular day or the week or month of the year for example, we examined the data using an autocorrelation function. This method is indicated based on the ranking of the variables from the random forest, which show that both day of the week and month of the year are highly associated with the amount of work missed, as well as by preliminary plots of the data that show the presence of cycles in the number of hours missed.

Feature Selection

The SciKit Learn package for Python 3.6 was used to run a random forest analysis of the entire data set and then rank the importance of the features. There were 5 features that contributed 1% or less to the total importance: “social smoker,” “pet,” “education,” “service time,” and “disciplinary failure.” Together these 5 features contribute only 5.2% of the total importance to the model and they were not included in the random forest classifier.

The most informative features in the set was the reason for absence, which is a categorical variable with 28 levels including values like injury, pregnancy, diseases of the respiratory system, etc. It is noteworthy that both the day of the week as well at the month of the year when the absence began are both highly informative variables, the 2nd and 4th most informative variables respectively.

Figure 1. Features Importance as predicted by a Random Forest Method



Random Forest Classification

Two random forest classifiers were created. For the first approach only the 14 most informative variables were included and observations from 2007-2008 were used to train the classifier, while data from 2009-2010 was used for testing. A new outcome variable was created by binning the number of hours missed into 1 of 3 categories.

1. Low: 1-40 hours = 1 week or less
2. Moderate: 41-80 hours = ~1-2 weeks
3. High: 81-120 hours = more than 2 weeks

This approach appeared to be highly accurate. Using only 14 features and 20 trees, the classifier was able to achieve 97.9% accuracy at predicting level of absenteeism in the testing data, however we have very unequal sample sizes for each level of the outcome variable and in the testing set not a single instance of the “high” level of absenteeism. This make the actual accuracy difficult to determine. Using a random split of the data into training and testing sets in order to get more balanced observations across the levels of the outcome variable, rather than using the first two years to predict the second two years, may help solve this problem.

Table 1. Confusion matrix of Random Forest Prediction of Level Of Absenteeism

Predicted

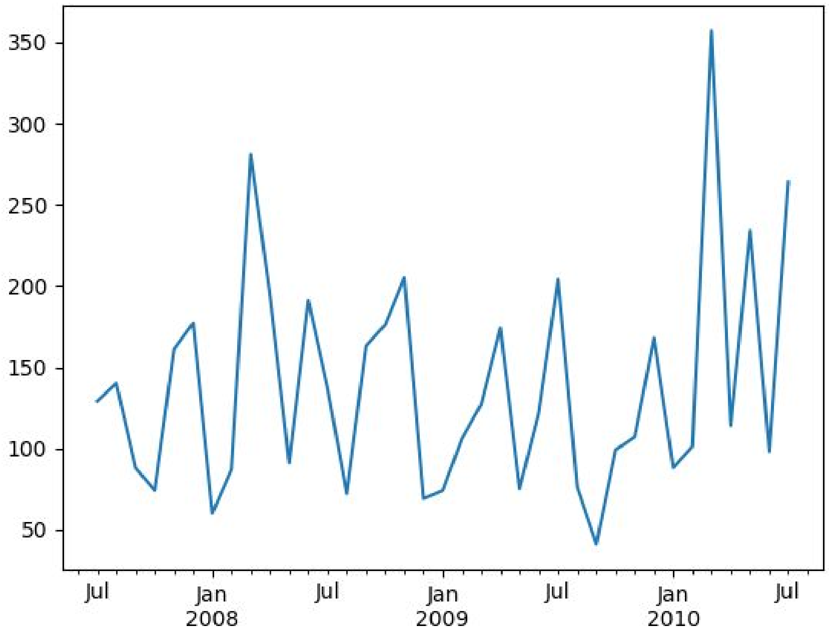
	Low	Med	High	Σ
Low	369	0	0	369
Medium	3	1	0	4
High	5	0	0	5
Σ	378	1	0	378

Overall Accuracy: 0.9788359788359788

Autocorrelations

We began by simply visualizing the data in order to assess if there is evidence of cyclical patterns in the data. To do this we added the number of hours missed in each month and plotted them over time (see Figure 2). From this plot, it appears that number of hours missed occurs in cycles, with a peak in absences following January in 2008 and January 2010.

Figure 2. Total Hours Missed vs. Date (month and year) X-axis: date. Y-axis: Number of hours of work missed.



We next examined the entire data set in chronological order by creating an autocorrelation plot. (see Figure 3). Autocorrelations can be used to show cyclical patterns in a data set. In this method, data is offset from itself at regular intervals known as a lags. At lags where the peaks line up with the valleys there will be a high positive correlation and at lags where the peaks line up with other peaks there will be high negative correlation. Thus, a cyclical data set will have a sine wave pattern with diminishing amplitude of the peaks and valleys of the waves.

Figure 3. Autocorrelation of Absenteeism at Work. The Y-axis: magnitude of correlation. X-axis number of lags. The blue line shows the 95% confidence interval. Correlation values that exceed these bounds are significant at the alpha=0.05 level.

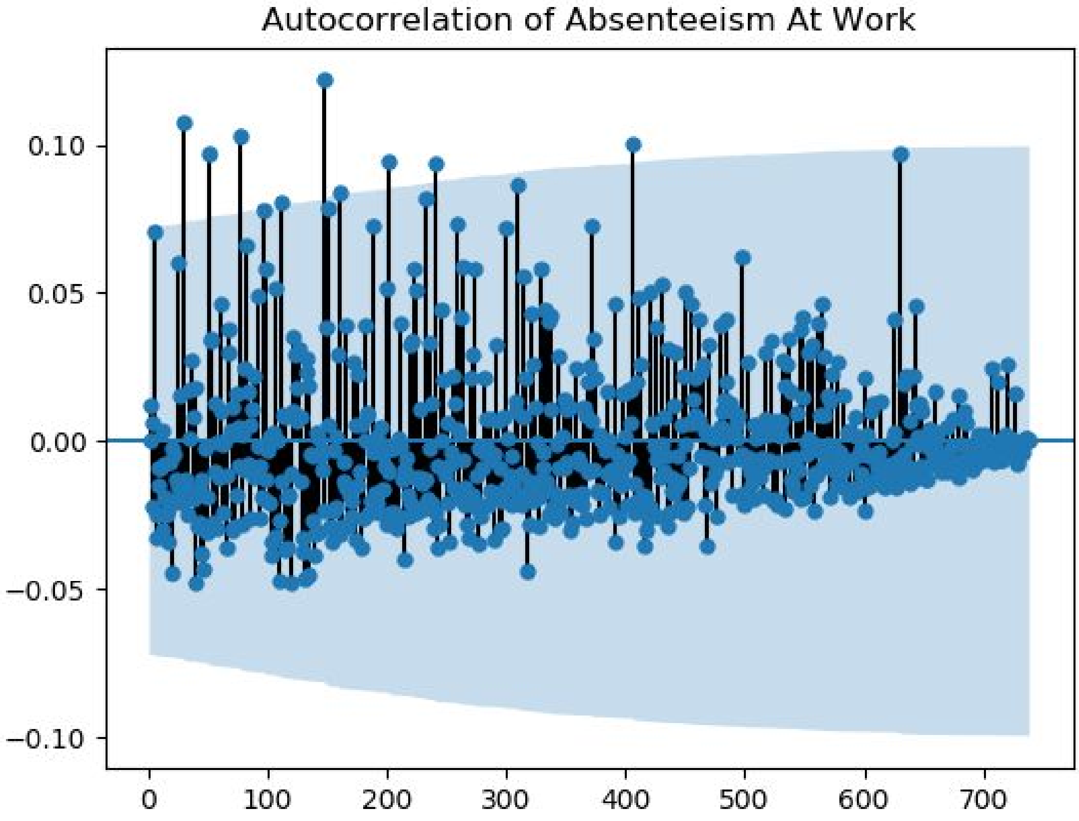
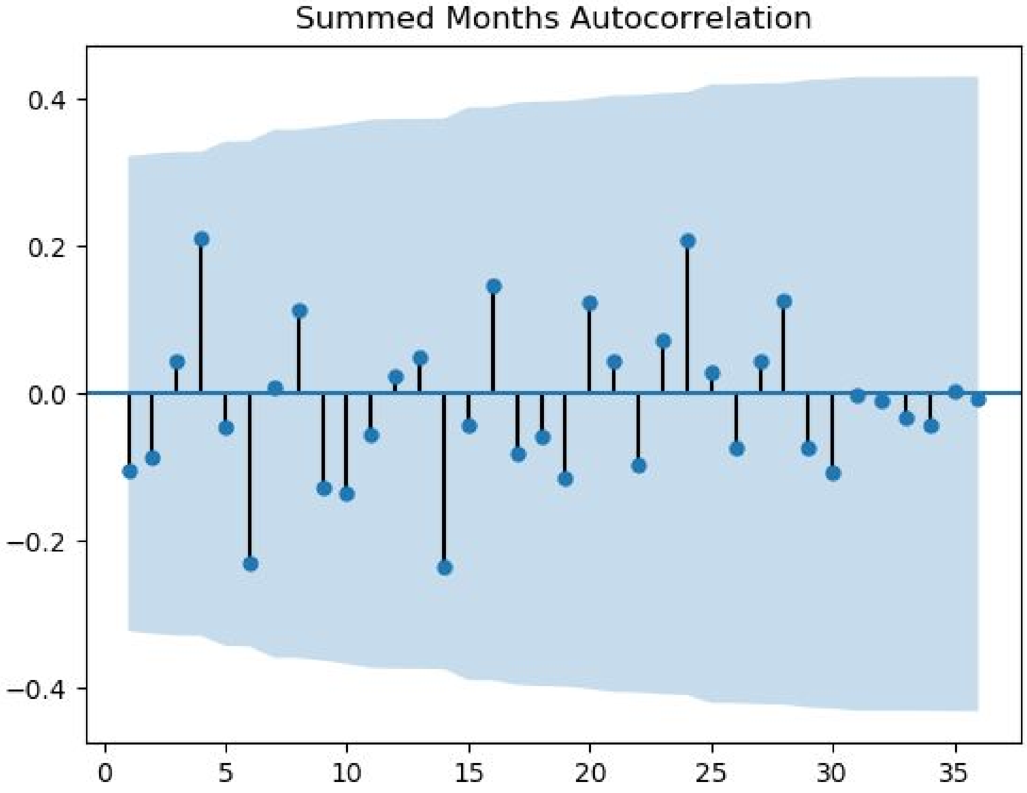


Figure 3 show that while there are a few few statistically significant points of positive correlation, in the raw data, in general the data is not significantly correlated. We also do not see any major cyclical patterns (sine waves) in the autocorrelation plot.

To any correlations in the number of absences by month over the course of the 4 year study, we first summed the hours of absences by month as in Figure 2 and created an autocorrelation plot for this (see Figure 4). The plot clearly depicts the cyclical pattern that was seen in Figure 2, however there are no significant correlations present (alpha = 0.05).

Figure 4. Autocorrelation of Absenteeism at Work per Month The correlation of the data set with the absentee hours of each month summed together with itself. X-axis: number of lags. Y-axis: correlation. The zero lag point has been removed as it always has a correlation value of 1 since there is no lag in the second data set. The shaded area represents 95% confidence interval.



Month Order and Day Order Autocorrelations

To further analyze the data using autocorrelation functions, the data was first arranged into two distinct patterns. The first pattern, called “month order” groups the data by the month and then sorts these groups by year and then finally by the day of the week.

Example of “month order”:

Jan. 2007 Mondays,
Jan. 2008 Mondays,
…
Jan. 2007 Tuesdays,
Jan. 2008 Tuesdays
…
Feb. 2007 Mondays,

Feb. 2008 Mondays,
etc.

The second pattern, “day order”, groups the data by day of the week first, and then sorts these groups by the month and then finally by the year.

Example of “day order”:

Mondays in Jan. 2007,
Mondays in Feb. 2007,
…
Mondays in Dec. 2010,
…
Tuesdays in Jan. 2007,
Tuesdays in Feb. 2007
…
Tuesdays in Dec. 2010,
Wednesdays in Jan.2007,
etc.

For both data sets, the number of hours absent on each day of the week, for each unique month-year combination were summed. For example, all hours of absences that occurred on the second day of the week, Tuesdays, within the first month of the year during the first year would be combined. We refer to this as a “generic day” because a generic Tuesday does not represent a specific day within a particular month and year, but rather all Tuesdays within that month and year. This reduced the number of data points and make examining the cyclical patterns within the data set easier. These “generic days” were then numbered in order so that each day had a unique identifier that reflects the chronological order of the data that we call a “generic day number”.

Month Order Autocorrelations

Figure 5. Month Order: Total Hours Missed vs Generic Day Number
Absenteeism in hours missed (Y-axis) for each “generic day” within the data when ordered by month. Generic day numbers (X-axis) correspond with earlier months and higher generic day numbers correspond with later months.

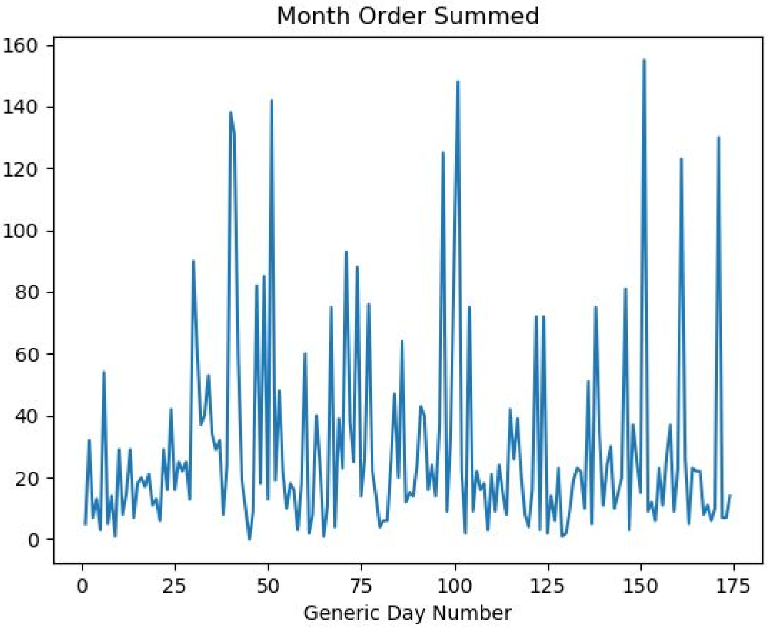
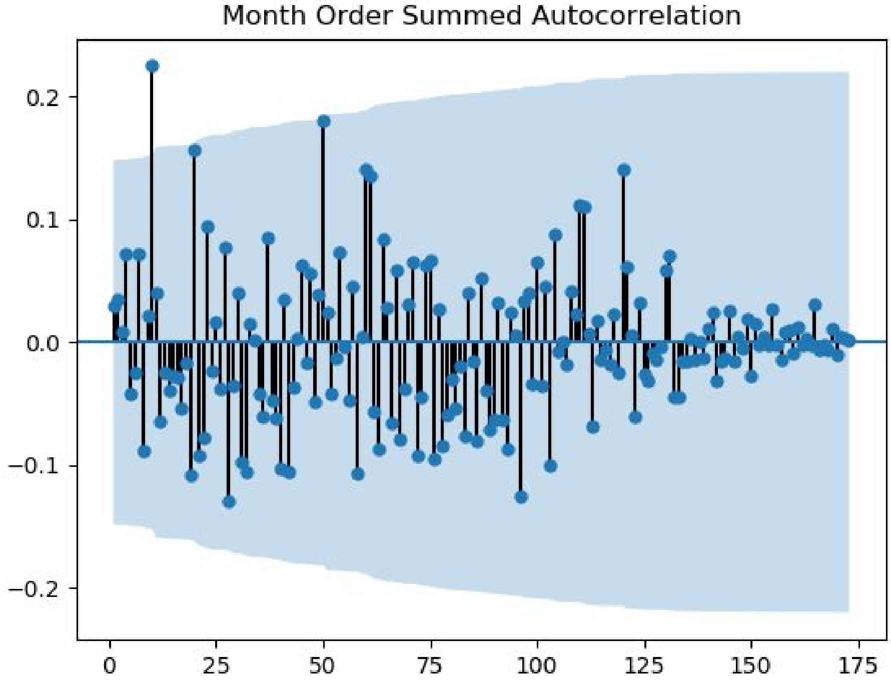


Figure 6. Month Order: Autocorrelation Plot. The correlation between the month ordered data set with a lagged version of itself, known as the autocorrelation. X-axis: number of lags. Y-axis: correlation. The zero lag point has been removed as it always shows a correlation value of 1 since there is no lag in the second data set.



We predicted based on the line graphs of the data (Figure 5) that when ordered by month, absenteeism would be higher during the summer months and lower in the fall, winter and spring regardless of the year. In other words that absenteeism would correlated with summer months. As shown in Figure 6 however, when ordered by month, the data set has a couple of significant correlations at specific lags but most correlations are not significant at the 95% confidence level (indicated by the blue shaded region). Also, the figure does not show any strong cyclical properties as no clear patterns are present visually. This suggests that there is no correlation between absenteeism at any one time of the year.

Day Order Autocorrelations

Figure 7**. Day Order: Total Hours Missed vs Generic Day Number** Figure 7 is similar to Figure 5 but shows hours of absenteeism (Y-axis) when the data is ordered by days as opposed to by months. Lower generic day numbers correspond with earlier days of the week while higher generic day numbers correspond with later days of the week.

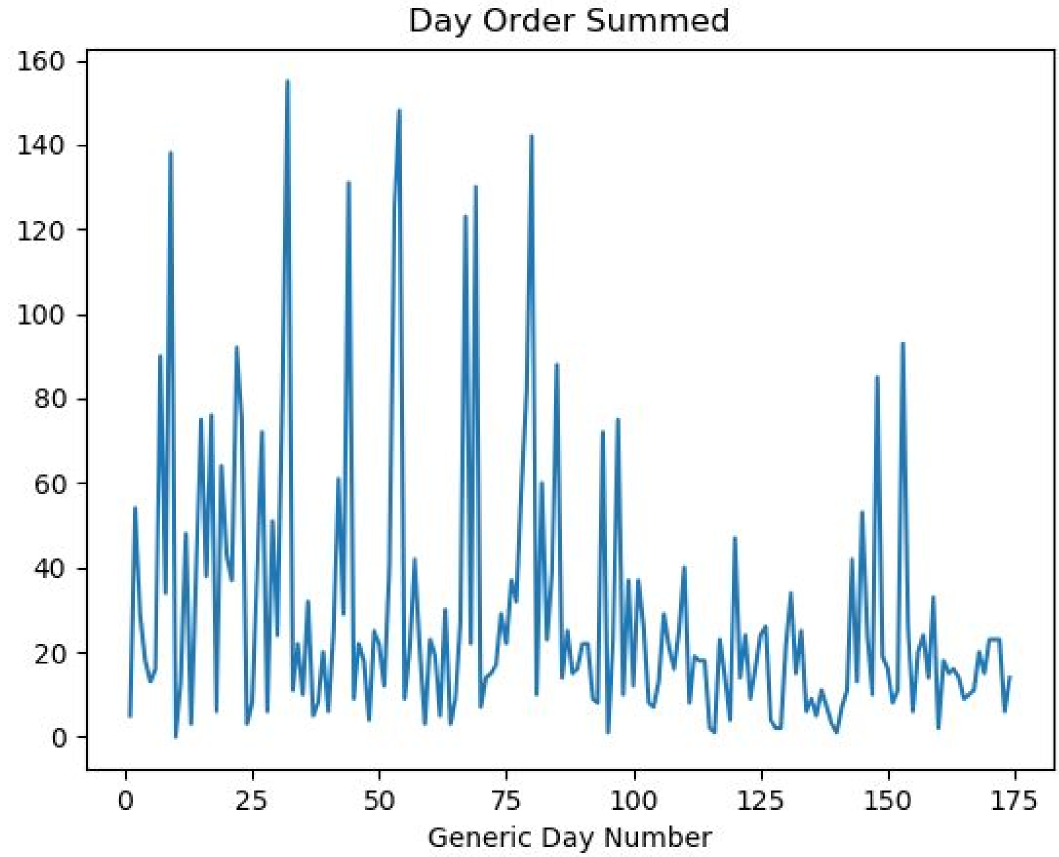
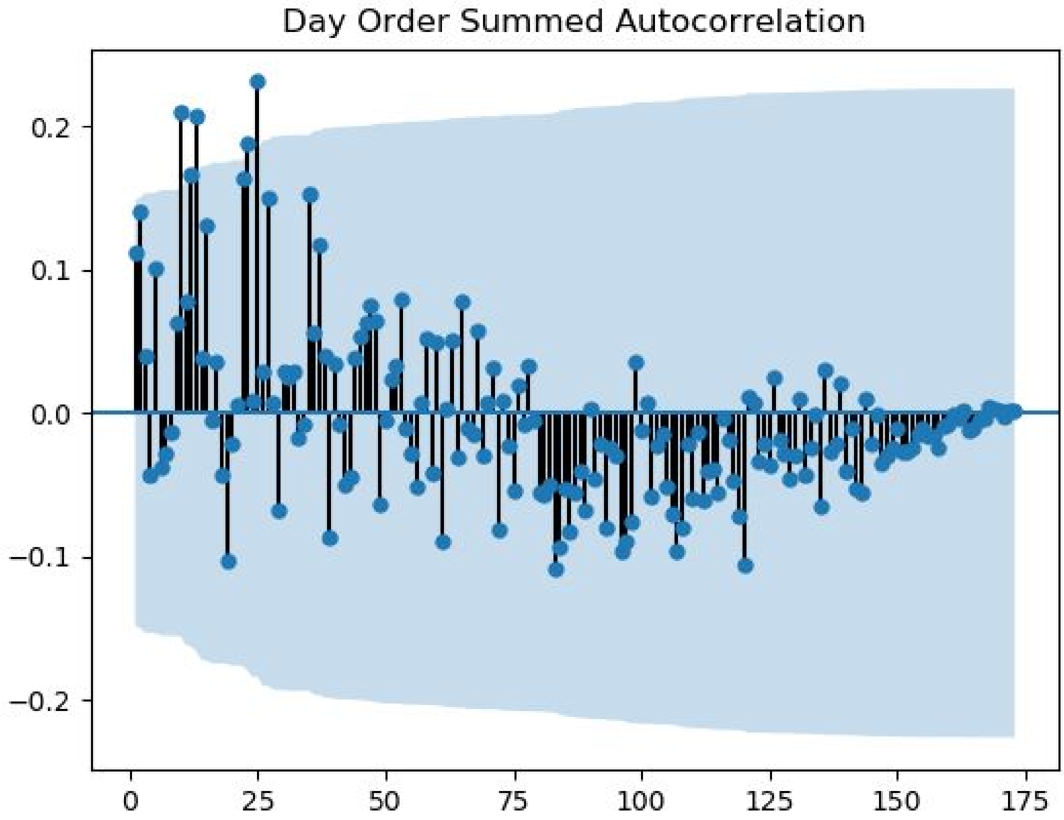


Figure 8. Day order: Autocorrelation Plot. The correlations between the data set in day order with a lagged version of itself is shown. The Y-axis depicts the correlation value and the X-axis depicts the number of steps the data is lagged. The blue area is the 95% confidence interval.



Based on Figure 7, we predicted that that absenteeism would be more prevalent at the end and the beginning of the work week rather than in the middle. As in the autocorrelations plot of the data ordered by month (Figure 6), there are few statistically significant correlations when the data is ordered by day. However, in this order, the data set does show the beginnings of a sine wave pattern when examined visually.

As the data has been arranged so that all instances of one day of the week occur next to each other, it approximates a single week. Within the data set the generic day that each day of the week starts at is as follows: Monday begins at 1, Tuesday begins at 36, Wednesday begins at 70, Thursday begins at 105, and Friday begins at 139. There are no Saturdays or Sundays included in the data set. Figure 8 shows is that the data set has a somewhat positive correlation with itself at lower lag values, when the earlier days of the week are aligned with other early days of the week and when the later days of the week are aligned with later days of the week. Around a lag value of 75 the data set begins to have a negative correlation with itself, which would be around when Wednesday begins to be aligned with Monday, when Thursday begins to align with Tuesday, and when Friday begins to align with Wednesday. This suggests that there are discrepancies in the amount of absenteeism at those points. As a follow up to this, the total number of work hours missed for each day are in table 3.

Table 3. Total Number of Work Hours Missed Per Week Day

Monday	1489
Tuesday	1229
Wednesday	1115
Thursday	553
Friday	738

These numbers agree with what Figure 7 shows, as the number of hours missed on Friday and Thursday are much lower than the number of hours missed on Tuesday and Wednesday.

Conclusion

Using random forest analysis the most important feature of the data set was determined to be reason for absence, with day of the week as the 2nd most important feature, and month of the year as the 4th most important. A random forest classifier using only the 14 most important features and 20 trees had a 97.9% accuracy when classifying data points from the last two years as low, medium, or high levels of absenteeism. However, the testing set was not well distributed among the three categories with no instances of the high category. In the future, randomly selecting data points to use as a testing set may give more reliable results.

Following up on the importance of the day of the week and month of the year features, a number of autocorrelations were run with the data set organized into different shapes. The autocorrelation for the data set in the original order but with the months summed showed a cyclical pattern in the data, meaning absenteeism was high at some points in the year and low at others. Similarly, the autocorrelation of the data ordered by day of the week showed that there was a difference in the number of hours missed early on in the week and the number of hours missed later in the week. The autocorrelations for the entire, unmanipulated data set and the data set ordered by month did not show any cyclical patterns. There were very few correlations of significant value throughout all autocorrelations.

Sources

Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.
https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work