Absenteeism at Work

Brandon Vandergriff, Stephen Meshotto, Jacob Hill, Gabriel Kiprono
Advisor: William Eberle

Problem

As the economic side of the world continues to grow, so do the demands associated with it. In lieu of these demands, absenteeism at work can lead to interruptions throughout a company or with a company's workflow. The overarching question is, can these be predicted despite being unintentional or habitual?

Methods

- Data normalization - a few points of our data were outliers: less than one percent were either invalid or too far off the spectrum to be useful. While it was decided that these points were valid in the grand scheme of things, they were not useful in the modeling process.

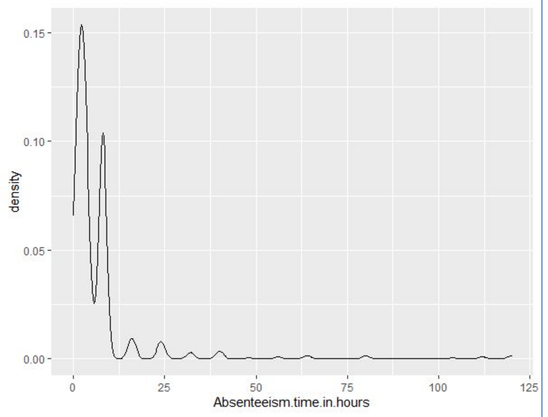

Figure 1: *The density of absent hours in the data set compared to the total amount. Notice how the plot is right skewed.*

- Data modeling - using the data, we were able to create a linear model for predictive analytic measures.
- Machine learning - after splitting the data into training and testing sets, we were able to predict with moderate accuracy if an employee would be gone for less than a day or not.
- Quantile regression - after finding which attributes had a high significance, we used these to build a data model.

Methods

- R and RStudio. The rationalization behind using R revolves around it being the leading programming language for statistical analysis. The libraries we used within R are ggplot2 for visualizations, quantreg for the model, and ROCR for model evaluation.

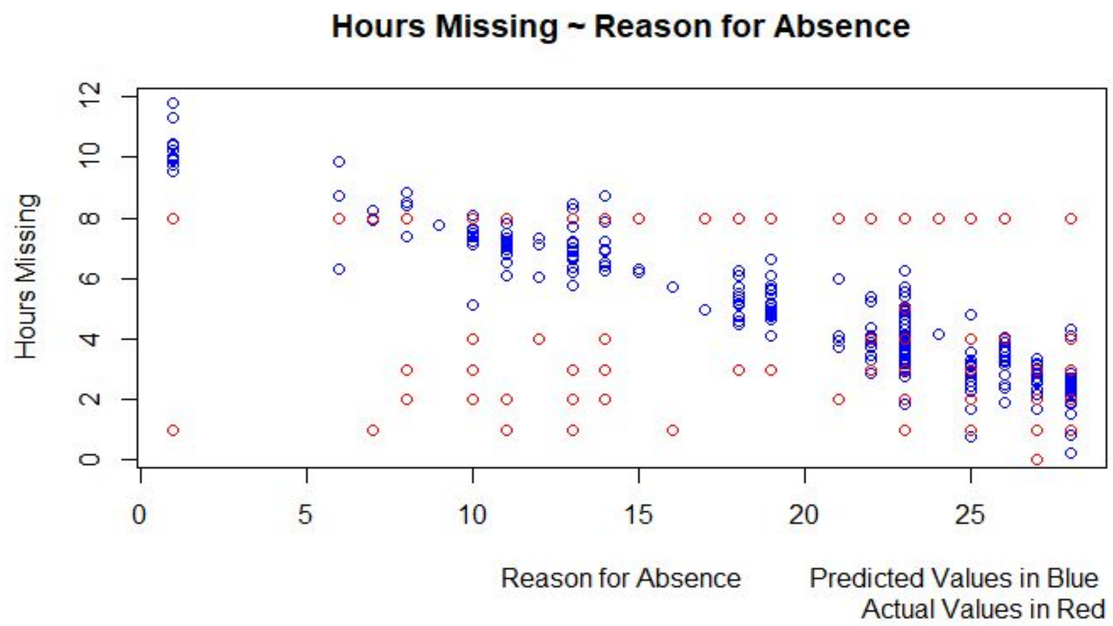

Figure 2: The actual data set values plotted with the linear regression model's predicted values.

Results

Data Description

Our data set lists the hours that employees missed by a Brazilian courier company from July 2007 to July 2010. The data includes categories about employee health like body mass index, social factors like whether the employee drinks or smokes, and other physical characteristics such as distance to work from residence and total service time. In addition to these characteristics, information about how long a particular employee missed and the reasoning behind the absence is included. All of these were essential in building our quantile regression model.

Data Preparation and Analysis

To prepare our data for modeling, we first cleaned the data. Our model had a few data points at with reason 0, which doesn't exist. After cleaning, we fed prediction reason for absence, age, day of the week the absence occurred, the employee's weight, height, and BMI.

Regression Model

Due to the dataset's tendencies, we chose to use Quantreg's Quantile Regression Model. This type of model is less affected by outliers, but tends to have less relevant coefficient numbers when the model is created. After comparing a standard linear model to a quantile model, the quantile regression seemed to fit the best, as seen in Figure 2.

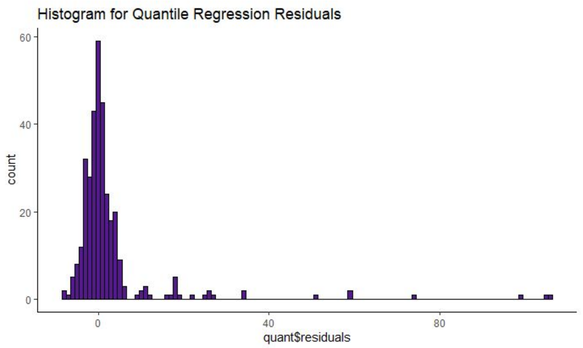

Figure 3: *The validity of the quantile model can be viewed through the density of values around zero. The further the number from zero, the worse the regression is.*

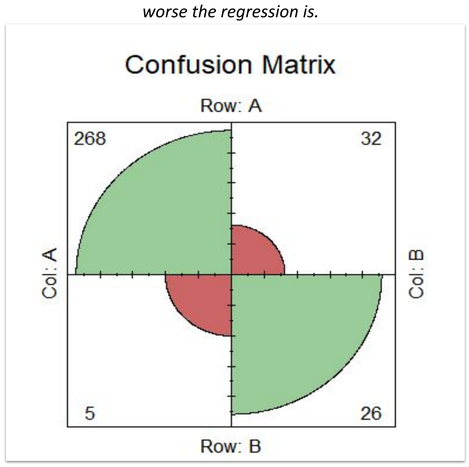

Figure 4: *The confusion matrix from a single run of training and testing.*

Conclusion and Lessons Learned

Lessons Learned

From our model, the average absences in a year is between 5-7 hours per individual excluding the outliers(the rarest reasons) the most significant categories are commuting distance, service time, family issues the tendency of smoking and drinking. Smoking and has been associated with the respiratory diseases, doctors' follow up, medical consultation and lab consultation which are the outliers observed in the data just like the body weight and height. there is not enough insights that can be drawn from this dataset as it only applies to Brazil, a third world country in South America. There could have been more factors to consider like weather.

Interesting Observations

Taller people were more likely to miss less, but when they did they missed for longer periods of time.

Using the linear model, there was almost no correlation between the social tendencies and time missed, such as if they were a social drinker or smoker, or whether or not they had a pet.

Conclusion

The average accuracy of this model is 85%. We were able to predict with high confidence whether or not someone would be missing for less than 8 hours or not. The average sensitivity for this model was around 95%, and the average specificity was around 10%.

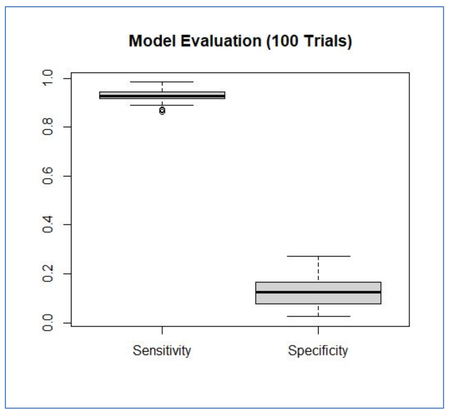

Figure 5: *The sensitivity and specificity after running 100 different test trials on this data set.*

Acknowledgements/References

Martiniano A., Ferreira R., Sassi R. (2010). *Absenteeism at work* [Data Set]. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work#