# Exercising the Numbers

EECS 349, SQ 2016
Caroline Grace Alexander (cga815), Samir Joshi (sjy034), Yannick Mamudo (ymt657), Matt Niemer (min168)

***Task***

Northwestern University's largest gym is Henry Crown Sports Pavilion, further referenced to as HCSP. At times, HCSP can be very crowded, and certain types of equipment, for example treadmills or squat racks, may not be available based on the number of people at the gym. Our goal was to write a ML algorithm that would estimate the size of the crowd at the gym at any given time in the future. This would allow patrons to make more informed decisions about whether to go to the gym or not, based on their specific needs.

***Data***

Every time a student enters HCSP, they must swipe their WildCard. We began the construction of our data set by obtaining a list of past WildCard swipes from students as they enter HCSP. Teisha Berry, a software specialist working with Northwestern Recreation was able to send a set of spreadsheets detailing the date and time of every recorded entry swipe (one for every person entering the building) from April 2015 to April 2016. The names were omitted and we created a python file to parse the spreadsheet into a .csv file with each row representing a half hour time slot between April 2015 and April 2016. We then assigned the number of patrons at HCSP during each half hour time slot, assuming that patrons stayed at the gym for about thirty minutes.

Once the attendance data was compiled, we added weather attributes. Upon recommendation from Professor Downey, we pulled historical weather data in the Chicago area from the website *http://mesonet.agron.iastate.edu/request/download.phtml?network=IL_ASOS* and with a python script added weather data for each half hour time slot. We also added the Northwestern school quarter and relative progress through the quarter to each time slot.

Our final data set has 14222 entries and is composed of 11 attributes:

- Day of Week - Numeric
- Time - Numeric
- Attendance - Numeric/Nominal
- Quarter - Nominal
- Place in Quarter - Numeric
- W1 (Temperature F) - Numeric
- W2 (Relative Humidity) - Numeric
- W3 (Wind Speed) - Numeric
- W4 (Precipitation) - Numeric
- W5 (Cloud Cover Level 1) - Nominal
- W6 (Cloud Cover Level 2) - Nominal
- W7 (Cloud Cover Level 3) - Nominal

Attendance was our classifier label. In order to try a wide variety of methods and algorithms, we initially created two different .csv files: one in which the attendance was left as a numeric attribute and one in which attendance was sorted into four different bins to make it a nominal attribute. After inspecting our data, we binned our attendance numbers as follows:

- 0 – 50: Not Crowded

- 50 – 100: Moderately Crowded
- 100 – 150: Pretty Crowded
- > 150: Jam Packed

We believe that these bins give a good estimate of the different sizes of crowds at HCSP from our own experience using the facilities. Binning in this way helped us evaluate the performance difference between trying to predict a nominal grouped range of attendances versus trying to predict the actual numeric attendance.

In addition to the numeric/nominal data sets, we created another distinctive split within our data. We found a large amount of entries with 0 under the attendance label, simply because HCSP was not open during those half hour time slots. Hence, we decided to create filtered data sets without those entries to see whether this would positively affect performance.

With these differences in mind, our training/testing data was compiled into four different .csv files with the following differentiators:

- Nominal Attendance, including zero attendances
- Nominal Attendance, without zero attendances
- Numeric Attendance, including zero attendances
- Numeric Attendance, without zero attendances

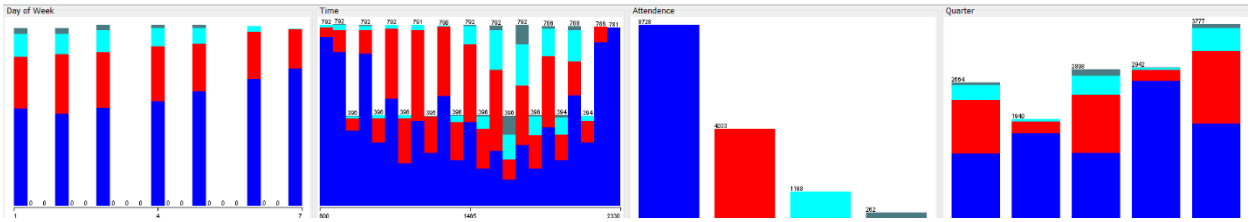
***Methods***

We tested most of the algorithms we learnt throughout the course using the Weka package. The list of ML methods includes Decision Trees, Nearest Neighbor, Linear Regression, Multilayer Perceptron and Logistic Regression.

We used 10-fold validation to build models using the aforementioned algorithms. At any point during the 10-fold cross validation, 90% of the data will be used for training and 10% will be used as the validation set.

***Results***

Using the Weka explorer, we analyzed relationships and trends between many of the different attributes and attendance. Figure 1 shows each attribute graphed against the attendance attribute. Immediately we can see spot some trends in the data. Attendance is heavier earlier in the week and in the afternoon. HCSP also experiences more traffic when the temperature (W1) and relative humidity (W2) are higher. Temperature and relative humidity rise and fall together, so it makes sense that their graphs have a similar shape.



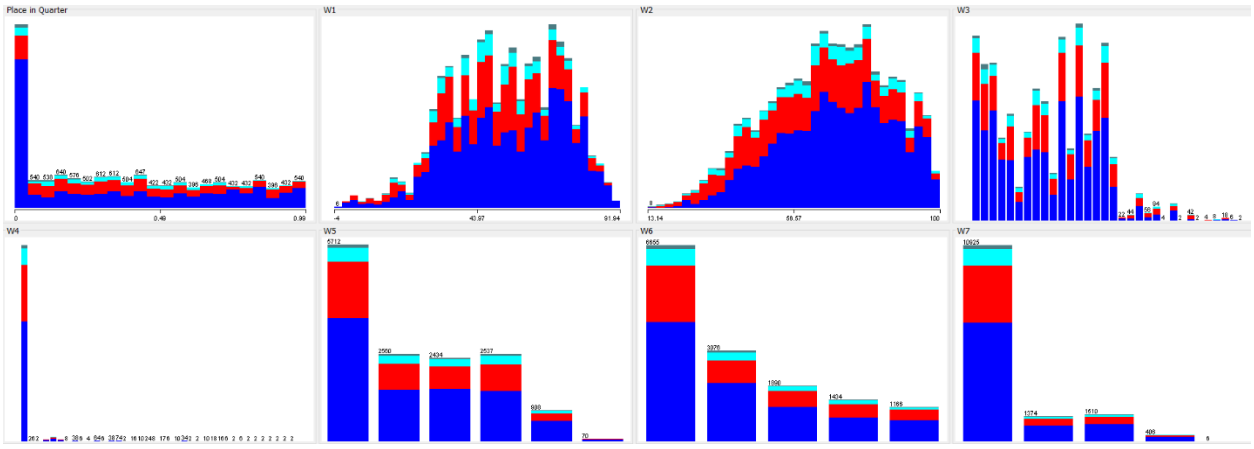



*Figure 1: Attributes vs. Attendance (Nominal)*

After inspecting the data, we decided to measure performance by using the accuracy of the model on 10-fold validation. Figures 2 and 3 illustrate the accuracy of our different algorithms' performance in Weka.

We can see from our graphs that removing examples where the attendance was zero generally decreased the 10-fold cross validation accuracy for our different algorithms. The algorithm with the greatest performance was the Nearest Neighbor variant, KStar. Reporting an accuracy of 83.215% in 10-fold cross validation, it classifies our data quite well.

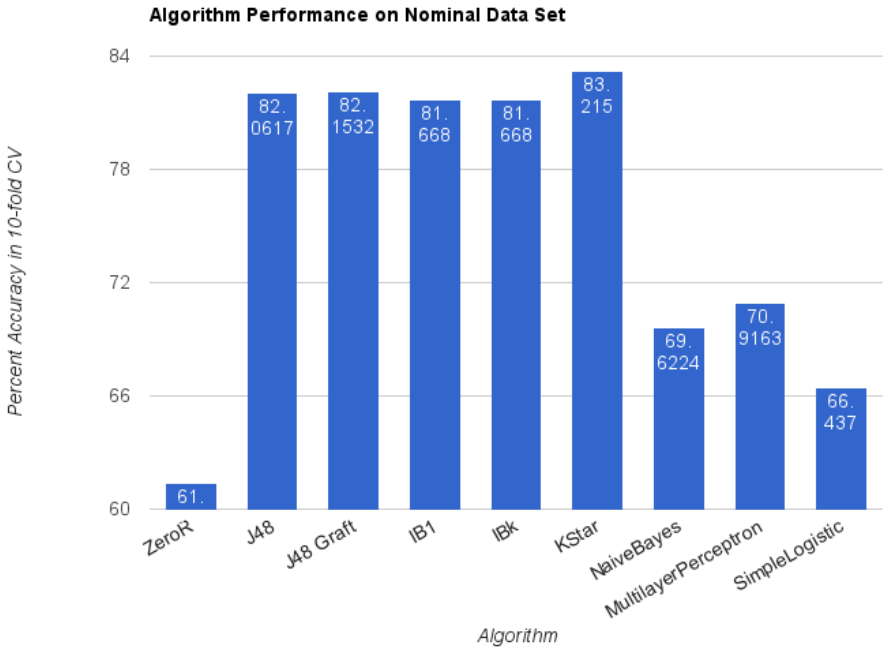



*Figure 2: Algorithm Performance on Nominal Data Set with Zeroes*

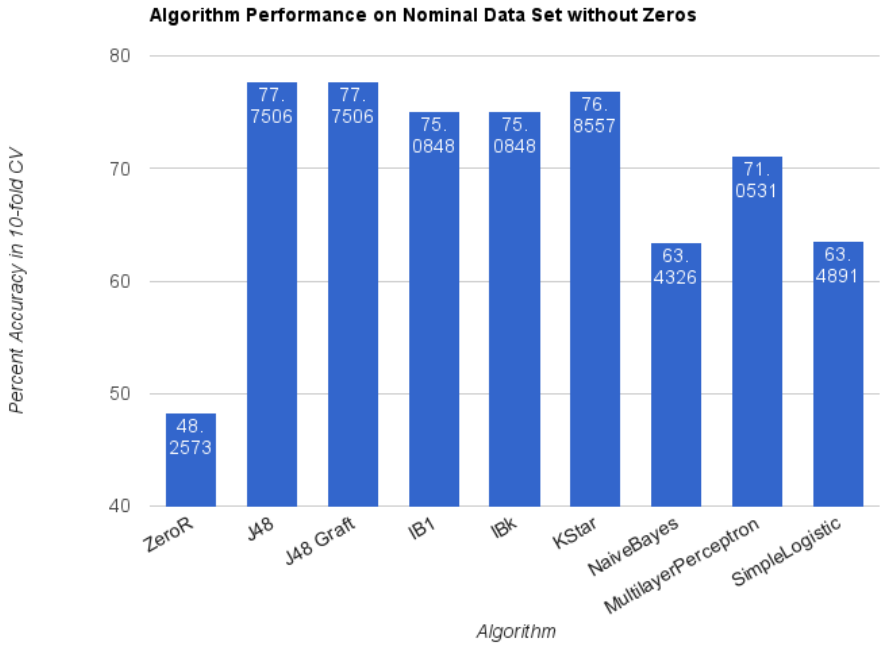



*Figure 3: Algorithm Performance on Nominal Data Set without Zeros*

We only tested one algorithm, Linear Regression on our data sets where attendance was left as a numeric attribute. Figure 4 shows the correlation coefficients for the Linear Regression models trained on our data sets. The model did not perform as well as we had hoped on either data set. A regression model should have a correlation coefficient of at least 0.8 to be considered strong, and ours are significantly lower than that.

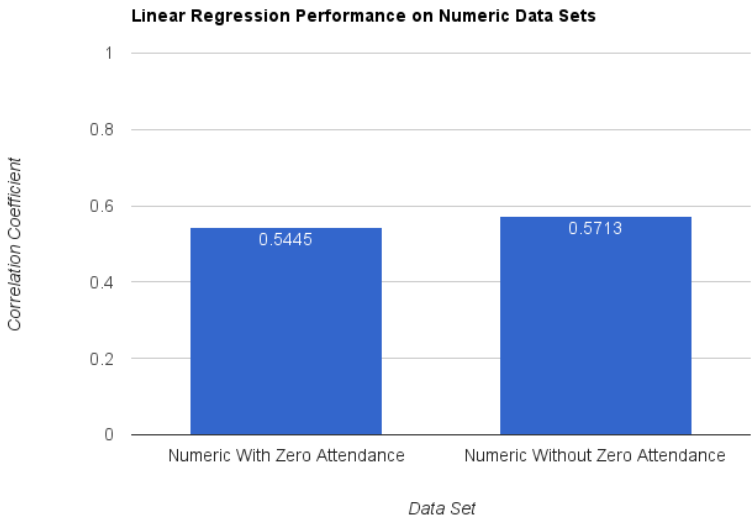



*Figure 4: Linear Regression Performance on Numeric Data Sets*

Analyzing our results, it makes sense that the KStar Nearest Neighbor algorithm gives the best performance on the nominal set. The nature of the data set is well suited so that the strengths of a Nearest Neighbor algorithm are maximized. The number of patrons at HCSP for some half hour time slot is likely to be close to the number of patrons during a half hour time slot with similar attributes, so we can expect Nearest Neighbor to classify quite well, especially when we have binned our attendance numbers into four groups.

Regarding the other algorithms, we can see why they do not perform as well as Nearest Neighbor. Decision Trees (J48 and J48 Graft), the close second to Nearest Neighbor, does very well with discrete attributes. Several of our weather attributes were nominal, discrete values and they likely contribute to the tree's relative success. KStar likely outperforms J48 and J48 Graft because of its entropic distance measure.

Multilayer Perceptron, Logistic Regression, and Linear Regression are similar in that they try to learn a regression or function to classify the class of an example. This approach, while practical in some cases, is not as helpful for our data sets. These three algorithms are not well suited for data sets with several nominal attributes, like ours. Thus with the current set of attributes, it is unlikely that a good function for approximating the attendance at HCSP is realizable.

Naïve Bayes is not well suited for our data sets because of its naïve assumption of conditional independence among the attributes. This assumption is not true (our weather attributes are certainly dependent on each other), so Naïve Bayes is not well suited for this task.

***Future Work***

Future work for this project includes adding and removing attributes to increase the accuracy of the model. Better attribute selection would undoubtedly lead to increased validation accuracy. Given the success rate of the KStar algorithm, it would also be worth looking into implementing the algorithm for a web based application for use by Northwestern students.

***Work Distribution***

The work was distributed evenly. We all discussed, brainstormed, and wrote code for the data parsing/ .csv compilation. Samir Joshi and Yannick Mamudo took more of a lead on the debugging and getting our data in a more appropriate format for Weka. Matt obtained the HCSP and weather data and Caroline Grace Alexander coded the website. Every member participated in the running of the various models.