Reuben Karchem

**DrivenData: Blood Donor Data Set**

1. **What is your project question?** Given an individual’s past blood donation history, did s/he make a donation in March 2007? This question is being asked by the competition website, Driven Data. Specifically, they are looking for the probability that an individual made the March 2007 donation. The competition’s evaluation metric is a log loss function (the current leaders have scores under .4).
2. **Raw data – Where did it come from? Why did you choose it?** The data is currently hosted by the UC Irvine Machine Learning Repository, but the donor data comes from a mobile blood donation center in Hsinchu City, Taiwan. The data set was donated to UC Irvine in October 2008, and includes donor data dating back to 2000. The data set includes 4 features, and 1 outcome feature. The features include ‘Months Since Last Donation,’ ‘Months Since First Donation,’ ‘Total Blood Donated in c.c,’ and ‘Total Number of Donations.’ The features all have the integer data type. The outcome feature –March 2007 donation- is binary (0 for no, and 1 for yes).

I chose this data set primarily because I wanted a simple data set with intuitive features that I could evaluate using the different models we’ve learned in this course. I want to walk through each of the model types and adjust the parameters and features, examine the relationship between features with different visualizations, and not get stuck cleaning a data set for weeks. It makes for a bit of a dry examination, but I didn’t want to spend too much time getting lost in the meaning of my data or dedicate the majority of my efforts to converting the data into a workable dataframe.

1. **Data exploration - how did you explore the data? What visualizations did you use? What descriptive statistics did you look at? What did they tell you?**

I first looked at the descriptive statistics for each feature to get a rough idea of their values and distribution. I then did a series of box plots for each feature, grouped on their feature outcome (donor or non-donor). I was hoping to see whether March 07 donors and non-donors were different from one another based on these four features. I found that there was some differentiation for the feature ‘Months since last donation’ and ‘Number of donations.’ I also noticed that there appear to be a couple of outliers for each feature, but have not yet determined what to do about them (exclude them? scale the feature?)

I also visualized the features in scatter plots against one another to see if there were any basic relationships between the features. That is when I discovered that the features ‘Total Blood Donated in c.c,’ and ‘Total Number of Donations’ communicated essentially the same information – the scatter plot formed a perfect line since each donation is 250 c.c. It appears that there are no instances of donors giving a double donation or a donor stopping mid donation. I have not yet determined what I should do about the features being seemingly redundant. I may try running identical models excluding one feature at a time to see which performs better.

1. **Data preprocessing – How did you clean your data? Were there missing values? Did you create new features?**

I intentionally chose a data set that did not require cleaning. I explored trying to create new features, but was unable to tie the timeframes provided (months since first donation, months since last donation) to specific dates. I was hoping I could incorporate time series features about current events in Taiwan or economic data (Taiwan does pay donors), but was unable to do so. I’ll have to rely on features that have been transformed from the features provided in the data set.

Looking at the distribution of the data points (scatter plots and box plots), I thought I’d try transforming the data. Alex pointed out that at least one of the data set’s features was on a vastly different scale than the others and suggested scaling or normalizing the features. I used the preprocessing module from Scikit-Learn to re-scale my features. I used the standard scaler and the minmax methods. The scale of the visualizations improved, but of course the relative distance between the points did not change. The range for the features was greatly reduced (Months since first donation decreased from ~98 months to 4 standard deviations.

I created new dataframes with the features all on the minmax scale, and re-scaled to standard deviations. I also intend to explore some other transformations, but these were the first I found in SciKit Learn’s documentation. I’ll try additional feature transformations and see if any improve my models.

1. **Model creation - Which models did you use why? What were your results? How does this compare to the null model? Which parameters did you change (e.g. k for knn)? Did this improve your model?**

So far I have only tested my model on the training data set. I haven’t calculated the null model, but I believe my outcome feature is about 75% non-donor to 25% donor, so the null model is already pretty high. I have looked at logistic regression, and KNN models on the original dataframe, the dataframe with standardized features, and the dataframe with MinMax scaled features. The logistic regression and KNN models had marginal improvement over the null model. Conceptually, I understood that re-scaling my data would not change the relative differences of the data, and as a result the models would not see marked improvements.

1. **What is your final model or final insight? What features did you use?**

I definitely want to try using the decision tree models to see if that could result in improved model accuracy. After adding in another model type, I’ll start adjusting model parameters to see which (logreg, knn, and decision tree) is ‘best.’ I am currently using the accuracy metric, but I might also look at using a different evaluation metric. If the business case is to correctly identify the individuals who will donate in March 2007, and the blood donation center wants to concentrate on calling those individuals, then I’ll look at finding a model that has high specificity. I would also like to try models not using all of the feature columns (especially volume donated and number of donations). There are additional model parameters that I still need to read more about to understand what changes to make to understand whether they would be useful to change.