**Lessons Learned**

1. What Worked Well:

>Preprocessing- Age Groups

Dividing the age attribute into 4 bins worked well. There are various ways to create these bins. One of the ways we found to be effective was to just divide the age attribute into 4 equal age groups, meaning an equal amount of years in each group. Another effective way was to divide it into 4 groups with each group representing a certain time frame in life. For example, the first group was 18-22. This age group represents the “potential college” age where people are still going through college or just starting to earn money and therefore do not have a lot of expendable income.

>Preprocessing- Using Excel

Another aspect of the preprocessing process that worked well was using excel to look at and organize the data. Using excel just makes it easier to look at the data and more convenient to organize it than using Weka.

1. What Didn’t Work Well:

>Cross-selling Recommendations

Some of the cross-selling items that made logical sense for the core product had a lower correlation with that core item. We are not sure as to why they didn’t have a good correlation with the item. Maybe there wasn’t enough data to relate them.

>Preprocessing- Using Excel

Using Excel with different weight ratings often produced the same results. Meaning the suggestions for cross-selling often showed the same list of new products no matter which attributes we changed the ‘Chi-Square’ values to. One of the team mates spent a consider amount of time, tweaking and changing these weights just to see if they could make the results change significantly.

>Preprocessing- Age Group Bins

Using more than six bins for the age groups stretched the results into something that really didn’t seem to make sense. No one on our team used the same break down as well. We had to correlate the results to come up a common set of facts.

>Using Classifier algorithms

Sometimes, the algorithms that we used didn’t get the optimal results. In some cases, the recommended algorithm was just forgone and manual data analysis through excel was used. Depending on how much data is present and which classifier is used, the results can vary quite a bit.

1. How would you change aspects of the processes that didn't work well?

>Teaching Imperfections of Data Sets

Help students gain an understanding of the imperfections of data sets earlier in the course/theoretical materials. They should know about missing data and why the size of a data set can make a big difference.

>Preprocessing

Make sure you understand how you are preprocessing the data, and trying to preprocess it in different ways or in more detail.

>More Data

Having more data would help predict certain attributes better. For example, we were trying to find out what attribute would predict the amount spent the best, and we weren’t able to get an accurate prediction with the data that was present. The prediction would have been a lot more accurate if there was more data to pull from.

1. What advice would you give a new data analytics specialist undertaking a similar project?

>Know the Data

You must forget what you know about what the data may mean. Take the time to look at the data and really learn about it. It takes several reiterations of sorting and sifting to come up with information where a valid conclusion can be drawn. Be flexible. Not everyone is going to embrace the technology immediately, especially if they are in the trenches and have previously evaluated the data in a manual fashion. So it is best to really ‘mine’ the data, and have a good understanding of the results seen. This can be done by looking at the graphs, changing the values for classifying the data, and being vigilant about looking for trends. You don’t need to be an expert in the field that the data is from (i.e. a Sale person), but you must dig into the results and become familiar with them. Also, some times changing values in regards to how the data is mined, doesn’t really change the results as well.

>Don’t Use Technical Jargon in Presentations

Using technical jargon can confuse the people who you are presenting to. Especially if they are not data scientists and have no knowledge of the field, which is likely going to be the case in most cases. Getting stuck having to explain certain techniques and technical aspects during a presentation can get long and time consuming. CEO’s and/or board members will not like this.

>Utilizing Excel

Try to utilize excel as much as possible. It is a great tool for looking at the data, organizing it, making charts and graphs, etc.

>Learn the classifiers.

Make sure you know about the classifiers, how each one works and in which situations to use which ones. For example, using M5P for numeric data vs. using J48 for nominal data.