Matthew Reed

Lessons Learned Report

Working for Blackwell Electronics has been a great opportunity for me to learn about the ways businesses think. I have assumed that the field of data analytics would continue to expand as the amount of data increased, combined with the decreasing cost of technology. eCommerce is undoubtedly a huge aspect of sales; the way in which businesses approach handling data and driving marketing from that is eye opening.

Watching different employees debate purchasing data has led me to understand that there is no one right answer for every situation. The questions asked by Blackwell are straightforward but require contextual understanding to answer them the right way.

* Do customers in different regions spend more per transaction? Which regions spend the most/least? (Central spent most, West spent least)
* Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?
* Can we predict the amount a customer will spend per transaction based on other data we have collected about that customer?
* Finally, we need to investigate Martin’s hypothesis: Is there any correlation between age of a customer and if the transaction was made online or in the store? Do any other factors predict if a customer will buy online or in our stores?

Pre-processing the data itself has been a logical effort, though understanding some of the algorithms has been difficult. So far I’ve only had to understand whether I am handling numeric or nominal values but I assume this will change as I become more experienced and handle larger, more complex problems. Discretizing the transaction data allowed me to organize the ages into 4 bins. I believe doing this made the results clearer to Blackwell. Working with J48 and M5P taught me the difference between regression and classification. Reading their outputs was a lot easier to do after I understood these concepts. Regression involves estimating or predicting a response. Classification is identifying group membership. It took me a lot of time to understanding correlation coefficient for the M5P.

Creating the cross-recommendations platform for Blackwell has been very informative. Using 157,818 instances of transaction data and 27 attributes (products), I preprocessed the data using the numeric to nominal filter. I also had to replace some missing values in the data. To do this, I had to use the ReplaceMissingValues filter. Since the classes aren’t defined in the data I set the “ignoreClass” to “True”. To calculate correlation, I used the chi-square test. This test of independence is a useful statistical tool that helps in identifying if two variables are related to each other. Chi-square works with nominal and categorical data, while other algorithms only work with one type. Correlation does not imply causation is an important concept in understanding the chi-square output. Using correlation as a basis for testing relational hypothesis is considered a logical fallacy. A high rank does not necessarily mean that one product correlates strongly with another. For example, there may be a large gap between the merit values of the number 1 and number 2 ranked products, which indicates that number 2 product might not be very strongly correlated to the original product, even though it is number 2. I realized that merit scores do not always work while building the cross-recommendation excel spreadsheet. For example, the SquareTrade 3-Year Computer Accidental Protection Warranty’s product 3 is Microsoft Office Home & Student 2010. However, I recommended product 9, the Alienware AAR4-100000BK because the warranty is for computers. Another example was the Logitech Desktop MK120 as product 3 for several printers but I recommended the ink instead.

Data visualization will be my biggest challenge in the upcoming courses. Infographics seem to be key in keeping a clear narrative easy to understand. Watching a variety of YouTube videos on the subject opened my eyes to the variety of ways to express large volumes of data. O’Reilly and TedxED have been the two main channels so far. Prefuse, Protovis, Vega, and Data-Driven Documents are tools I’ve only recently learned about, and hope to learn to use in the future. I wish that infographic tools were more design friendly. A lot of infographics look the same.

I have a few recommendations for Blackwell Electronics. Blackwell can create a “Geek Squad” that can teach the older customers how to use the laptop as well as other products.

Come up with marketing strategies in the west region to attract more customers. Offer more incentives to purchase in-store like price matching, BOGO. Explore the concept of stores within a store. Create awareness of online shopping in the west region. Set up more brick and mortar stores in the east region.

I am very excited to start using R and understanding the more complicated algorithms. A friend of mine is building a localization project using python and SVMs, so I want to learn about SVMs as well. My plan is to keep watching videos and resources on learning these algorithms so I have a better contextual understanding of which to use. I wonder how much more I can charge for creating the visualizations compared to just processing the data and drawing conclusions. I also wonder if I can create software that will have both the processing and visualization options in one. I hope I can use my experiences here to move into the world of music data, e.g. Spotify and Pandora, to understand listening habits better.