**Project 1, Task 3: Blackwell Electronics**

Lessons Learned

Anders Dowd - Spring 2022 Data Analytics Certification Program

This was my first foray into the field of data analytics, so every element of the project was new to me. Sure, I was familiar with .csv files and the basics of charting and graphing information, but from step one I was learning new things. Cleaning data was a new concept to me, but as with much of what I ended up learning from this course, the concept was logical and immediately made sense. I know we are about to dive into the deep end of preprocessing and exploratory data analysis, but I felt like I could have handled a little bit more extensive work in those areas in the first task. The data required almost no clean up to be useable.

The course was helpful in the way the different modules we used in Python were gradually rolled out. Prior to beginning this course, I had never written a single line of code in Python, so the strategy of unveiling a few features at a time kept it from becoming too overwhelming and helped me build some foundational skills relatively quickly and painlessly. I thoroughly enjoyed familiarizing myself with some of the plotting features of MatPlotLib and Seaborn and figuring out ways to tweak my parameters to get visually appealing graphics.

I also appreciated how right off the bat I was having to learn by doing. I quickly realized that most of the time, when I had a question, I was able to find the answer on my own somewhere on the internet. The resources available are endless and it is abundantly clear to me even at this early stage that a lot of the job is not necessarily knowing all the answers, but having the skillset to find the answers when the question is asked. So far the course has been a strong reinforcer of that basic truth.

Learning the difference between discrete and continuous data, as well as categorical and numerical data, and how to effectively manage each type informed my broader approach to the tasks presented. Of course now these concepts seem obvious to me, but it was one of those fundamentals of data prep that I had to learn even though in hindsight I realize I’ve seen these techniques my whole life. For instance, binning age into groups is a common practice in any survey I’ve ever taken, but prior to this course I didn’t have a term for it in my head. I also learned rather quickly once I started doing it that the more bins you use, the more specialized the groups become. This can be good for categorizing data, but can make predictive modeling more difficult since the categories become more nuanced the more abundant they are.

One technique that proved helpful for me (at least in this early stage) was developing a naming system that allowed me to create categorical and numerical values of the same variables. There were times when I was trying to display the data in some way or another but the Python command required a certain type of data that was different from what I had. I realize that down the road I may figure out a better system, but for now it was helpful for me to essentially copy a variable under a different name and reclassify its type so that I could easily call back to either format whenever I needed.

As I worked through the tasks, I developed an understanding of what types of graphs to use for various situations. For instance, bar graphs, point plots, and pie charts were all great tools for quantifying and comparing multiple clusters of data. Scatterplots, however, were less helpful, especially when all of my variables were discretized. I look forward to working with more continuous data and delving into how scatterplots can be useful.

The section of Course 1 that I got hung up on the longest was when we got to the classification algorithms. Not knowing exactly what kind of results I could realistically hope for, I spent more time than I should have tweaking parameters and rerunning the algorithms only to continue to get very similar results. I don’t think this is necessarily a bad thing, since even unsuccessful attempts can provide valuable learning experience. But going into future projects that require these types of algorithms, I think I will be quicker to zone in on one set of parameters and stick with them as long as they provide me useful results, especially when we get into bigger data sets that will only take that much longer to process. There was a lot of cumulative waiting-around time during this course while I let the computer run these algorithms, particularly the Gradient Booster.

Another pitfall I encountered as a result of having unrealistic expectations was when I thought I could essentially manipulate the data to get a better model, rather than just working with what adjusting the parameters could get me. Now, when I initially ventured down this path, it should be noted that I didn’t realize I was in fact manipulating the data; I thought I was helping the program learn how to interpret the data. When I first realized that binary target variables were the most successful in predictive modelling, I thought if I could break down a continuous variable (age) into multiple binarily binned data points, I could help the computer see a clearer pattern in the independent variables and therefore make better predictions. So I proceeded to create new columns for “over/under” a series of ages (30, 40, 50…all the way up to 80) which returned a 0 for under and a 1 for over. This gave each row a series of 0s and 1s that could help pinpoint the age of a customer. I then ran this new dataframe through the algorithms with a discretized age variable as the target (in this case, broken down into 5 generational groups). Lo and behold, I had a model with around 90 percent accuracy! But upon doing so, I immediately recognized my folly: I was essentially using age to predict age. The new datapoints I had created would not be possible to know without already knowing the target variable. It was a “duh” moment for me, but one that I only reached by doing the wrong thing first.

That’s an important point to reiterate: even doing the wrong thing can provide good information. When learning by doing, sometimes you have to do the wrong thing to learn what not to do in the future. To that end, I would not change anything about the processes of this course. Everything I did, even when it was wrong, helped build the foundations of what little I do know about data analytics thus far, and it will help me more quickly figure out the right course to take in the future.