**Capstone Project Milestone Report**

**Introduction**

At the beginning of the course I was asked to present a problem that would utilize the skills I would learn in this course. I began to brainstorm a few ideas that I believe could be solved by someone with my experience and current skill set. I also wanted to tackle a real-world problem that utilized data from my current position as well as solve a pressing question in my organization.

To come to my decision, I took some time to understand the current state of the company and some key performance indicators (KPI) that had pressing issues. This landed me on a problem that every ecommerce company has, customer attribution. My domain experience in the field of marketing biased my decision but the problem at hand can easily be in the top five in terms of importance to the company’s long term growth.

My initial draft proposed that I use third party marketing data that is anonymized for privacy reasons to build a mixed media model (MMM). In marketing advertisers spend money with various vendors for the purposes of either brand recognition, new customer acquisition, or return customer reactivation. Our company is primarily concerned with the second of the three and a challenge we face is understanding how to maximize customer acquisition with a fixed budget.

I would employ some statistical algorithms that I was familiar with (e.g. multi-linear regression and k-means clustering) as well as others that I would learn to build a model that would both minimize cost per acquisition (CPA) and predict future customer growth given a set budget.

**Data Wrangling/Exploration**

I was a bit fortunate with the gathering of the data as it only required that I reach out to all of our vendors via email. The downside of this is that it took some time on their end but once I had it I could begin my project.

I began by consolidating all the sources of data which were strictly in csv format into a single document. I then queried our database to obtain new customer figures for that same date range and joined the two data sets via calendar date.

I then attempted to structure them in a tidy format with each row being an observation (date) and each column being a variable (in this case advertisers spend by product). I put it in this format because of my understanding that each key must have a value to be tidy. The problem that I encountered quickly was that while each advertiser spends on a certain date I do not have a pairing for customer counts by advertiser. My customer counts are aggregated on the daily level but not the advertiser level which is the problem that I am trying to solve!

After my data was cleaned and organized I imported into and data frame to begin the exploration process. I began my exploratory data analysis by using both the ggplot2 and dplyr packages in R. I made some simple box plots, histograms and scatter plots to show the structure of my data. I followed that up by running some simple summary statistics including the mean, standard deviation and IQR by vendor to gleam any helpful insights. I noticed that there were a handful of big advertisers followed by a second tier who only comprise a tiny percentage of the whole advertising budget.

Lastly, I examined the data as a time series object to see if the it took the shape of a time series with a trend, seasonal component and cycle. Doing so gave me confidence to say that this series does not reflect a time series due to the quick nature or of marketing team. Their job requires that they pay attention to CPAs so their budgets change as they get new information.

After viewing the data in several different ways, I noticed that there are patterns that would indicate a linear relationship between advertiser spend and customer acquisition. This led me to believe that employing a linear model would be a good starting point as well as a baseline model.

I also noticed that pockets of spend and customers were divided in a few clusters. That indicated that I could employ a random forest and use the results to test against my baseline model.

Looking back at my initial proposal nothing much has changed in terms of the algorithms that I would choose. The gathering, cleaning and analyzing were all new skills I gained during this course and have helped in validating my initial idea.