**Capstone Final Report**

Mixed Media Modeling & Customer Acquisition

Jethro Perez

**I. Introduction**

In the world of electronic commerce (e-commerce) CPG companies must balance year over year growth in revenue while maintaining a net positive ratio of lifetime value over cost per acquisition. The ratio is known as LTV/CAC and is one if not the most important key performance indicators (KPI) for any company whose business model lies on customers consuming a company’s goods on either a one time or recurring basis.

Maximizing this KPI can be divided into two problems, one centered on classifying customers based on average basket size (also known as average order value) and targeting customers who purchase the most (i.e. maximizing LTV). The other which is the focus of my project is maximizing customer acquisition by lowering the average cost per acquisition (CPA).

**II. Problem**

In the realm of marketing a mixed media model (MMM) is used by savvy data intensive organizations to minimize CPAs. A MMM assess the impact of each of your marketing partners in driving new customer acquisition.

By building a MMM marketers can see which channels drive the highest volume of new customers and conversely which don’t. Using this information executives can alter their budgets which normally rely on third party data from advertisers who have different reporting standards and measuring techniques.

Building a model that answers these questions can range from simple to complex depending on the state of your business as well as your marketing plan (one could say those two are correlated with each other). For example, building a model for a local restaurant who advertises in newspapers and local ads is rather simple compared with a Fortune 500 company that sells a product in multiple countries and spends billions of dollars on advertising budgets across multiple different partners.

In the case of my company we operate across the United States and spend millions of dollars acquiring customers for multiple different products. We have a baseline set of advertisers which we spend most our yearly budget on while reserving some for testing new partners who may unlock new markets that have yet to be unlocked.

**III. Data Wrangling**

The data that was gathered for this model came from our third-party advertisers as well as our internal data model which I build and maintain.

Acquiring the former was an exercise in generating reports from the respective advertisers reporting platforms or email correspondence in the absence of the former. To have substantial sample size *n* given the number of potential predictors *p* I decided on aggregating data on a daily level going back from 2016 through 2017 for four different products. In total I gathered 578 observations and 31 different variables.

The data here has been anonymized and scaled for privacy purposes. Figures stated in my model will be different from this project but the overall process (modeling, results and deck) will be the same when recreating it for company members. I decided to also limit this project model only product 1 which is our largest and most important. The findings of this can easily be reproduced to the other three products.

When gathering advertiser data I intentionally omitted customer counts which each partner collects. The reason for this is due to a phenomenon known as over attribution. This occurs because advertisers have different methodologies when measuring the impact of their platform. For example, a social media company like Twitter can track users who click on advertisements and purchase on a website. That mechanism does not work for a company that sells radio advertisements who may rely on the use of specific promotion codes to measure customer acquisition.

To make matters more complex customers use multiple devices (personal computer, work computer, mobile phone, etc.) and don’t always click on ads but use search engines to engage with the brand. Having multiple touch points results in advertisers stealing attribution from each other due to customers naturally having multiple different touch points prior to converting. Over attribution results in overstated customer counts compared to actual figures, therefore I needed to gather that data using our internal data model.

Once I had data from each partner I anonymized and joined all the data into a CSV document. The format was intended to have each day as an observation and each column as a different advertisers spend on a product that particular day and the number of actual customers acquired on that date.

**IV. Exploration & Analysis**

I began exploring my dataset using some of concepts I adopted after learning to use *dplyr* and *tidyr*. This is where I ran into one of the biggest problems in my entire model. To use data in *tidy* format each value must have a key to pair to it. In the case of advertisers and the amount spent that requirement is met, however we don’t know how many customers were acquired from that specific advertiser on a given day we only know that in aggregate. To get a 1 for 1 matching for customers by advertiser we need to solve for over attribution. This is done by either building or adopting a multi touch attribution model which is outside the scope of this project.

The inability to have data in this format limited my ability to facet advertisers as a function of amount spent and customers in a scatterplot. But as we all know real world problems will not have complete data that is clean and ready to model.

I overcame this hurdle and began to explore this data through the lens of a time series. My logic here was a belief that the cost of customer acquisition (CPA) is a function of both the mix in media spend as well as some seasonal components. An easy example of this sales for an electronic retail company which sees huge fluctuations in customer sales as the holiday season approaches. My findings suggested that there was no such evidence in any seasonal components but this can be a false assumption, only having 1 ½ years’ worth of data is hard to confirm this theory.

I followed this by looking at some basic scatter plots which I did log transformations to make spotting trends easier. There are no categorical variables in this data set so using boxplots or histograms a bit moot. I followed by computing some basic summary statistics including the mean, variance and max/min to see how my data was distributed. I inferred from this that my data was best as is and didn’t need any log transformation or removal of outliers.

Lastly I created a correlation matrix using the *pairs* and *corr\_matrix* function to get a better idea of how each variable varied with each other. The linear relationships I saw on each of the predictors variables with the response ignited the idea for the initial algorithm that I would use to apply machine learning to solve the question at hand.

**V. Model Selection**

Given the structure of my data set I believe multiple linear regression is the best approach and starting point. I began by building a multiple linear regression with my baseline model being one that simply uses all the predictors to measure the response.

I randomly split my data into a training and test set using a 70-30 mix. When using the baseline model I simply regressed all predictor variables on the response to see what the simplest model will yield. The summary of the model yielded a paltry adjusted r-squared value of 0.533 and only 7 out of 9 predictors being statistically significant. The model had an intercept of 204 meaning that if we spent $0 on advertising we would on average acquire 204 customers. The poor r-squared value lead me to the path of refining my model.

I started removing predictors with large p-values out of the model one by one and noticed that my r-squared value wasn’t moving. This indicated that I need to take a closer look at my data to look for any patterns. I noticed a bit of multicollinearity between a few predictors which violates one of the requirements necessary for OSL to work. Limiting this required adopting another method for variable selection as well as interaction terms.

I introduced a few interaction terms to see if this would improve the performance of my model and to my delight they did! Adding 3 interaction terms improved my adjusted r-squared up to 0.6403 or a 20% improvement over the baseline. This makes sense intuitively because advertisers and the users who engage in them don’t live in a vacuum. Most people who convert due to advertising are exposed to it a few times to keep it on their mind and may be influenced into buying something because they see ads everywhere they look.

Given the large boost I needed to devise a way to search every possible model that can be built using each distinct combination of predictors and interaction terms. I used stepwise selection using the *regsubsets* command. I created the best 3 models for each combination starting at 1 predictor 28 predictors. I then looked for the most optimal model using BIC, Mallows Cp and adjusted R squared to find the optimal number of predictors to use.

All three statistics showed that 8 is the optimal number of predictors which yielded the following model.

*prod1.customers = A.prod1 + L.prod1 + A.prod1:D.prod1 + A.prod1:I.prod1 + A.prod1:L.all.1 + I.prod1:K.prod1 + I.prod1:L.prod1 + L.all.1:L.prod1*

This says that the optimal mix in media occurs when you spend on advertiser A and L independently and spend on pairs of A and D together, A and I together, etc.. As stated in the Introduction to Statistical Learning book there are a few advertisers that have synergy or work with each other. The result of this work boosted the adjusted r-squared to 0.684 or a 28% improvement from the baseline model.

The presence of multicollinearity made me believe that the last model may still be flawed so I decided to use a ridge regression which performs well with data like this. Ridge regression uses a penalty factor lambda to adjust each of predictors via a shrinking the beta values based on this penalty. I ran a sequence of over 2,000 models at various lambda values. I found the optimal value and created model. This model unfortunately performed poorly with an adjusted r-squared of 0.537. This makes sense with a basic ridge regression because it simply shrinks the beta values to a very small value but doesn’t remove them.

After generating these models, I gauged their performance via k fold cross validation for the ridge to simply using the test set those using OSL. This led to the conclusion of using the discovered using the subset selection process.

**VI. Summary**

As I finished this project I found that there is more complexity to the problem at hand then what was first assumed. I am happy with the model I constructed but I believe with some improvements in the data I can reach one which will perform better.

This leads me to my first recommendation which is to use the presence of the synergies found in my model to better understand which advertisers interact and how many different touch points are there. My model only assumes pairwise connections but there can be interactions with 3 or 4 distinct combinations working together.

If we build a multi-touch attribution model or purchase a service we can understand how many touch points it takes a user to convert. Having this data will give us a deeper understanding of these relationships and their magnitude. Understanding this will assist in variable selection which in turn may produce a better model.

While this is being worked on I recommend testing this model by adjusting spending allocations and spending considerably less on those that don’t appear in this model. Doing so diverts money that would have been wasted on channels that don’t produce any new customers into ones that have the best cost efficiencies. If this works you can further extend the use of the model to predict the volume of new customers over a horizon of time

My last recommendation would be to further enhance the depth of data that is collected. One factor that was not discussed but impacted the results is the presence of customers who purchased without the influence of advertising. This model assumed that 100% of the variation of customers acquired was due to advertising. That is not the case in most organizations which can use brand recognition to influence purchase behavior. The simplest example here is coca cola which is something that almost everyone remembers but isn’t necessarily because you saw them on Facebook. There are a few current estimates of what percentage of customers are from the Non-Paid channel as it is called. If we collected this and removed it from the equation we would have better model performance.

Overall I am confident that the model presented can be used in practice and iterated on with new data as well as different modeling techniques as we learn more about the impact of advertisers on customer’s decision to purchase. All of this of course is easier with more data!