Zach Olivier
ISYE 6420
Course Project

Project Summary

Goal of the project is to analyze marketing channel performance for a small SaaS startup. Data is an anonymized list of leads (website sign-ups) with corresponding indicators if that lead eventually purchased a product (wins), broken out by the originating lead source (marketing channel and marketing subchannel). Data has been provided by the company and analysis has not been performed before acquiring this data. Original data is provided in the *bayesian_marketing.csv* file.

Based on the data, there are over 20+ unique channels that leads can originate from, including: cost per click, direct sales, display ads, organic search, paid social ads, paid video ads, and referral partner ads. Within each channel are a host of subchannels including: Facebook, google, affiliate backlinks, salesforce, YouTube and others. Using Bayesian analysis, the aim is to determine which mix of channels and subchannels is best to maximize the amount of purchases (wins) and inform Marketing leaders which channels to invest in or divest from.

Bayesian methodology is relevant to this problem as we are dealing with only 6 months of daily marketing wins data. Further, channel investment level and rate are variable throughout the dataset. With new channels just starting, it is hard to estimate performance as channels will low volume may have artificially high conversion rates. For example, a Facebook conversion rate (present throughout dataset and one of the largest sources of leads) may be just 1%, while a new channel could generate only 20 leads but have a conversion rate of 10%. Based on only conversion rates, we may incorrectly divest from Facebook into the new channel.

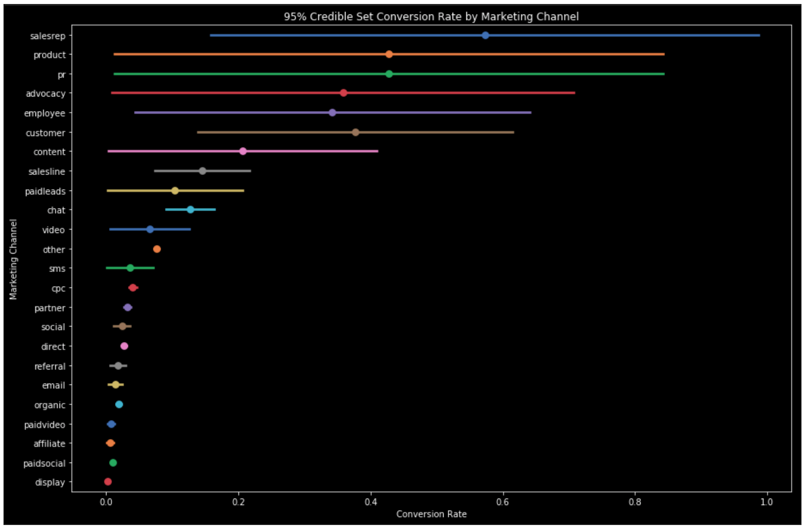
The plan to analyze this data will happen in two stages:

- Estimate Bayesian credible sets for conversion rate for each channel and subchannel
    - Estimating conversion rates will take advantage of the Binomial – Beta conjugate pair, allowing us to sample and build credible sets directly from the Beta posterior distribution
    - Analysis will use a non-informative Jeffrey’s prior to allow the sample size of each lead source to inform the length of the credible sets
    - Ranking lead sources by estimates conversion rate and length of credible set will show which channels we have the most confidence in, and eliminate falsely investing in channels with high conversion point estimate but wide credible set band
    - Details of the methodology can be found in the *analyze_channel* and *analyze_subchannel* functions within the *project_functions.py* script

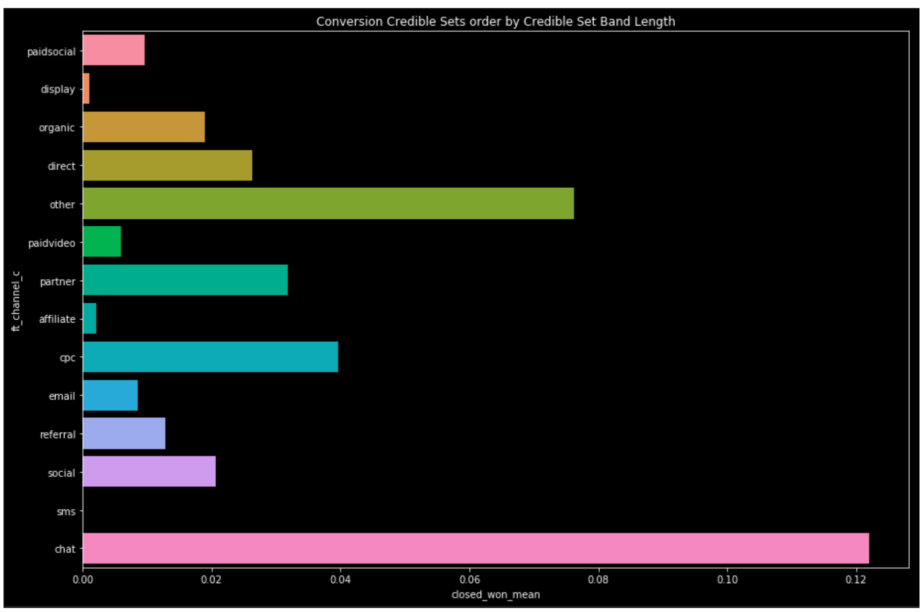
- Model the probability to win for each channel and subchannel
    - Multiple models will be developed to determine which channels and subchannels result in the highest probability of a win
    - Each model will be a Logistic Regression fit on a combined set of channels and subchannel, channel only, and subchannel only
    - Results from these models will help drive investment into specific channels i.e. channels with the largest effect of the probability of conversion
    - All models will be fit within the Bayesian framework provided by pymc3
    - Each model takes approximately 5 minutes to fit – model results are printed after each model fit is completed
    - Details of the methodology can be found in the *channel_subchannel_model*, *channel_model*, *subchannel_model* functions within the *project_functions.py* script

Project Results

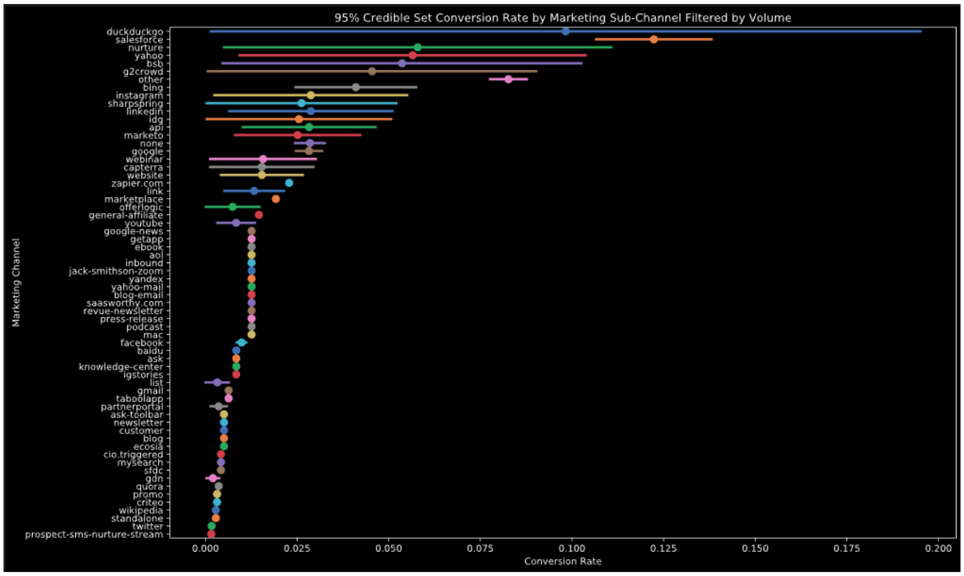
- Usefulness of the Bayesian credible sets is clearly shown here – looking at conversion rates only, we would show channels such as Sales Rep, Product, PR, Advocacy, and Employee as having conversion rates of greater than 30% (global conversion rate is just 3%). Looking at the credible set bands we can see the lack of data is causing a wide interval, sometimes stretching from 5% to 90% conversion for these low volume channels.



- Using the same credible sets, we filter to the channels that have the shortest intervals from lower to upper posterior equi-tailed density. These shed light on the true performing channels that also provide an adequate amount of lead volume. Channels like CPC, Chat, and Partner have relatively high conversion rates with narrow credible set intervals.



- The same methods were then applied to the marketing subchannel data, allowing us to quickly sort through and find the subchannels that have the highest "confidence" of their conversion rate, as well as segment out those without enough data to be sure about their estimate



- In total, three models were fit to the data to estimate channel and sub channel contribution to odds of converting the lead to a win.
  - Model 1: Channel and Sub Channel Model
    - Key Channels / Sub Channels based on model coefficients:
    - Paid Video, Google, Affiliate Back-Links, Salesforce
  - Model 2: Channel Model
    - Key Channels based on model coefficients:
    - Paid Video, Partner
  - Model 3: Sub Channel Model
    - Key Sub-Channels based on model coefficients:
    - Salesforce, Affiliate Back-Links, Google

Channel and Sub Channel Logistic Regression Model Results:

	mean	sd	hpd_2.5	hpd_97.5
ft_channel_c_cpc	0.052919	1.190172	0.037020	0.073543
ft_channel_c_direct	0.028674	1.086572	0.024421	0.033725
ft_channel_c_display	0.038058	1.963509	0.009598	0.138822
ft_channel_c_organic	0.026099	1.217857	0.018301	0.040010
ft_channel_c_paidsocial	0.140627	1.174864	0.100649	0.189136
ft_channel_c_paidvideo	1.019196	2.656550	0.151130	6.700427
ft_channel_c_partner	0.041395	1.091075	0.034815	0.049108
ft_subchannel_c_facebook	0.064214	1.171023	0.047547	0.086737
ft_subchannel_c_gdn	0.059623	2.030173	0.015928	0.255212
ft_subchannel_c_google	0.762706	1.202467	0.532430	1.103138
ft_subchannel_c_link	0.273434	1.416275	0.139704	0.531683
ft_subchannel_c_salesforce	0.499993	1.088801	0.421955	0.587573
ft_subchannel_c_youtube	0.010466	1.341650	0.005844	0.018145

Conclusions

Based on our Bayesian analysis we can easily determine which channels and sub-channels are noise and which channels can add value in win conversion. Most of the channels and subchannels shown have extremely wide credible set intervals, indicating we are unsure of their true conversion based on the data sample. For channels and subchannel with narrow credible set intervals, we can identify Chat, CPC, Partner, Organic leads as having the highest relative conversion rate that we are confident in. Further, based on our Bayesian logistic regression models we can identify Paid Video, Google, Affiliate Back-Links and Salesforce leads has having the highest overall contribution to win conversion odds. Based on this analysis we can recommend investing more in these channels while divesting in channels such as GDN, YouTube, and Facebook.