VIACOM®

Social Media Targeting

Instructor:

Dr. Nathaniel C Lin

Submitted by:
(Group 4)
Ana Paniagua
Fatimat Lukan
Sumit kashyap
Shengyuan He
Qing Yu

Table of contents

Part 1- Introduction (Business question, explain the questions we are trying to answer)

Part 2 - Data Understanding & Data Preparation (data cleaning, explaining fields, and how we join the datasets,etc)

Part 3 - EDA

Part 4 - Modeling (CPM Imputations, Clustering)

Part 5 Deployment (how to apply the model, recommendations)

Part 6 References

Introduction and Business question

Viacom being the 9th largest media company in term of revenue, gets its big chunk of profit through advertisement.

Viacom operates about 170 networks, reaching 700 million subscribers in approximately 160 countries.

Through our understanding of Viacom business requirement, we will answer the following business question:

1. What is the predicted CPM for different facebook pages.
2. What pages are performing good in different demographics.
3. Where are the gaps in the market.
4. Is it worth pursuing the offer?
5. What value we can provide to the partner advertisers.

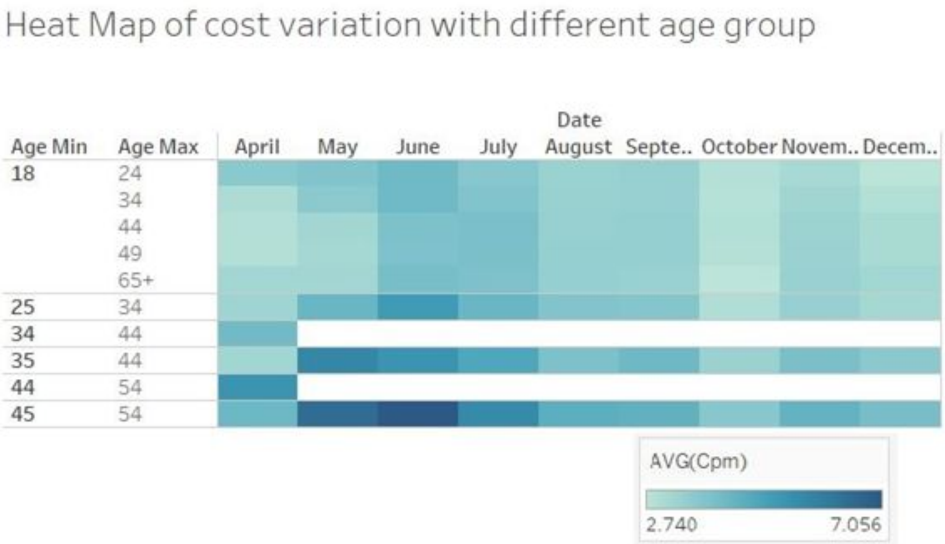
Data Understanding

- Viacom provided 14 files in total. Each of the 13 files represented all the metrics for a particular month Jan 2018 to Jan 2018, and the CPM dataset.
- Each dataset had the following attributes:Hid (page id), date (date of collected data in that month), name (metric types), value (value recorded for the metrics).
- The datasets was uploaded to R studio for cleaning and exploration
- All months dataset was joined using rbind function in R studio to make further analysis.
- The CPM dataset was used to make modeling predictions regarding the cost per page impression
- Predictions were done based on the join of all year and cpm dataset

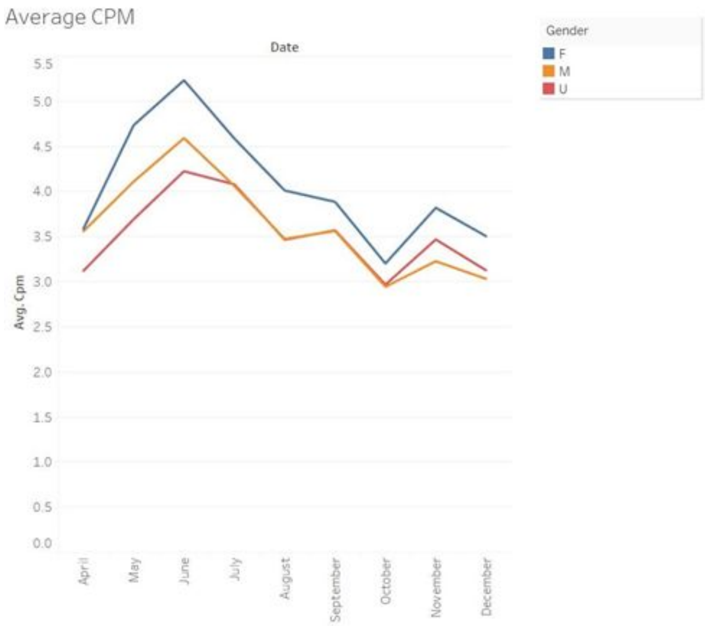
EDA

CPM Data:

Heat map describing how the CPM is changing with date and age group.



Graph shows higher CPM for females

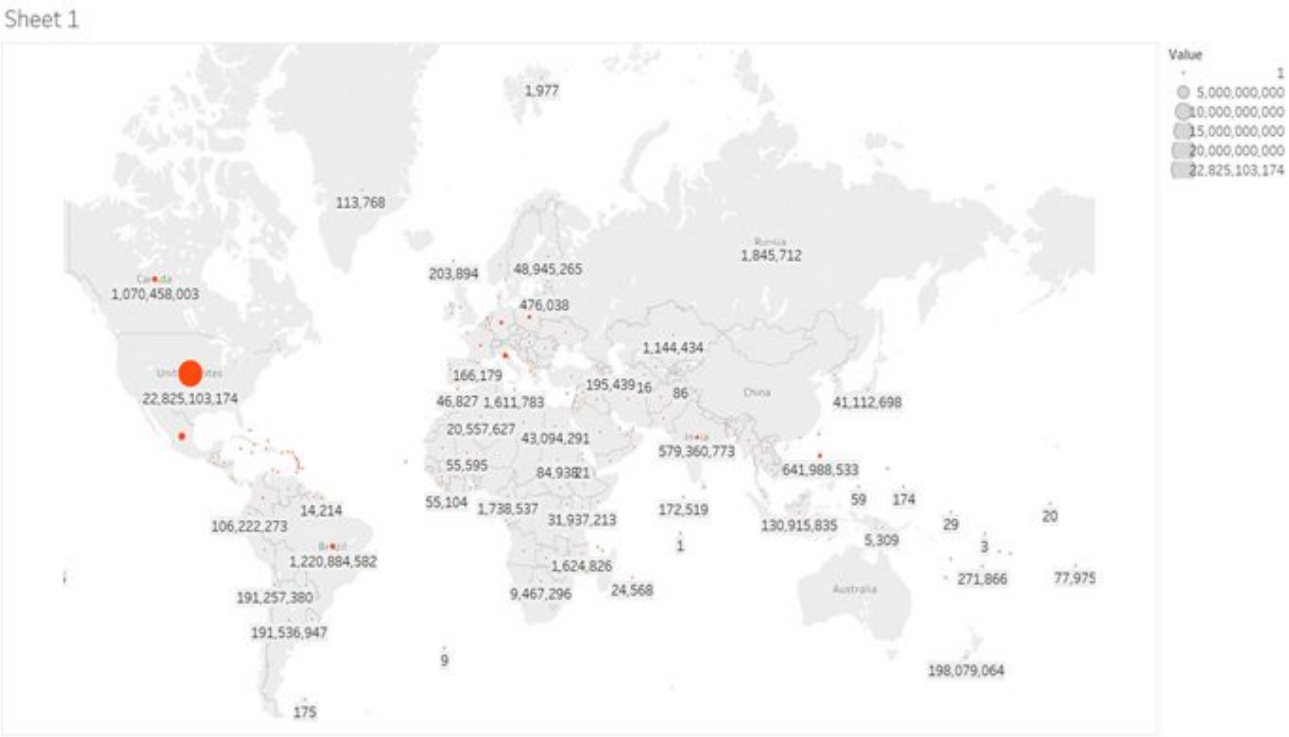


EDA

Page level Data:

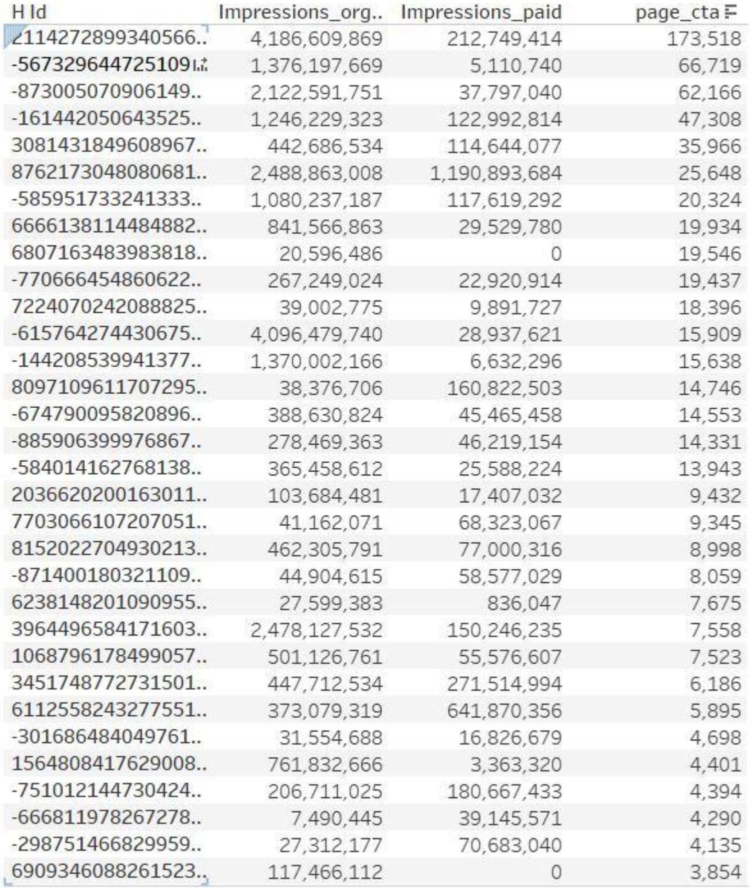
Facebook popularity

Around the globe.



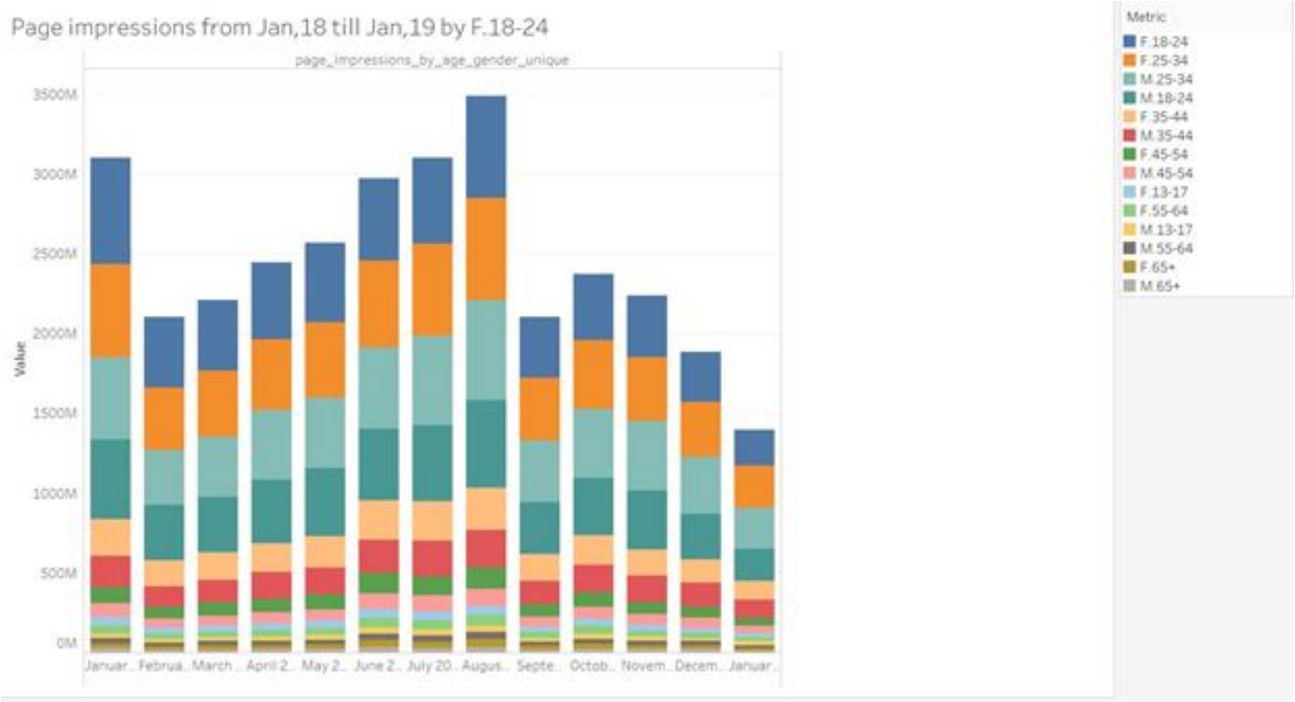
Map based on Longitude (generated) and Latitude (generated). Size shows sum of Value. Details are shown for Metric.

Top Pages to target based on there cta

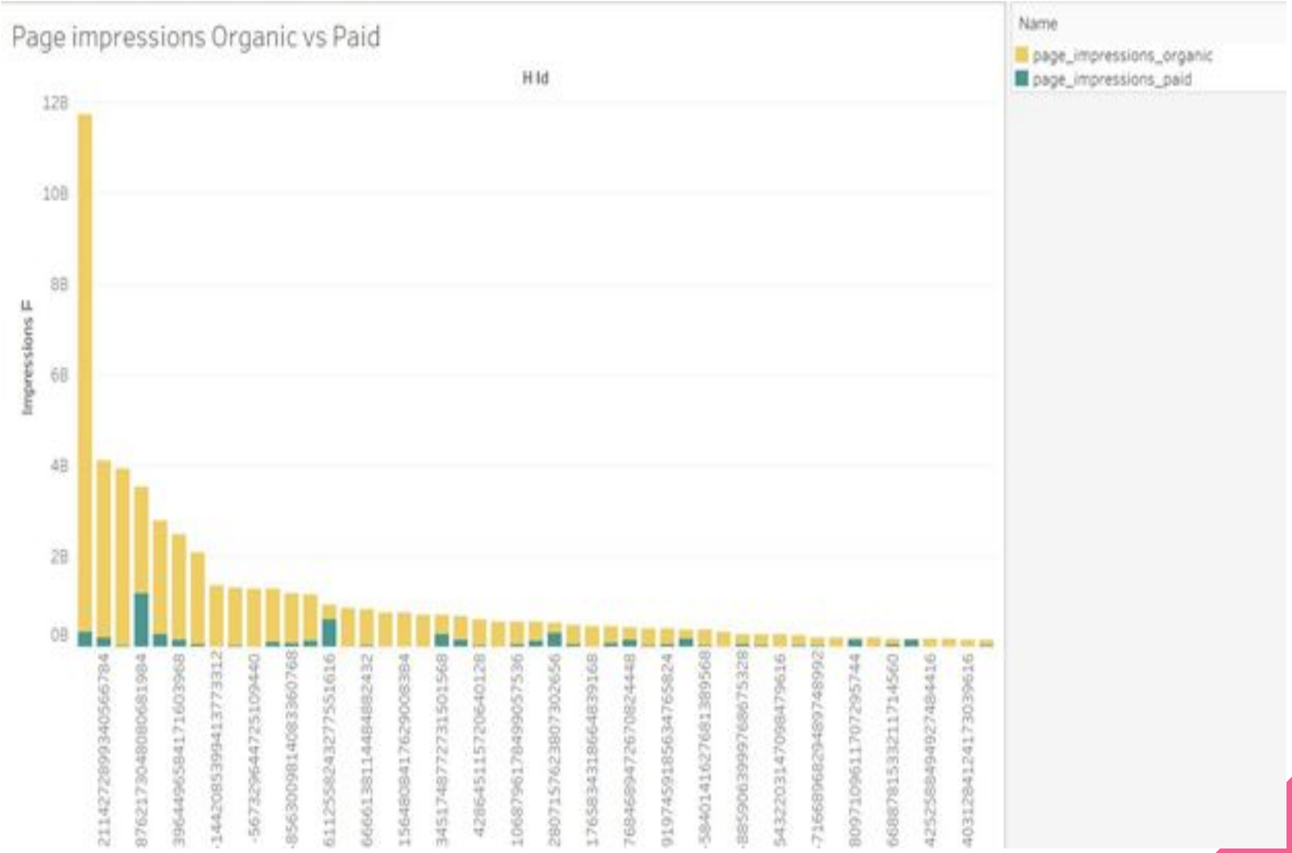


H Id	Impressions_org..	Impressions_paid	page_cta
2114272899340566..	4,186,609,869	212,749,414	173,518
-567329644725109	1,376,197,669	5,110,740	66,719
-873005070906149..	2,122,591,751	37,797,040	62,166
-161442050643525..	1,246,229,323	122,992,814	47,308
3081431849608967..	442,686,534	114,644,077	35,966
8762173048080681..	2,488,863,008	1,190,893,684	25,648
-585951733241333..	1,080,237,187	117,619,292	20,324
6666138114484882..	841,566,863	29,529,780	19,934
6807163483983818..	20,596,486	0	19,546
-770666454860622..	267,249,024	22,920,914	19,437
7224070242088825..	39,002,775	9,891,727	18,396
-615764274430675..	4,096,479,740	28,937,621	15,909
-144208539941377..	1,370,002,166	6,632,296	15,638
8097109611707295..	38,376,706	160,822,503	14,746
-674790095820896..	388,630,824	45,465,458	14,553
-885906399976867..	278,469,363	46,219,154	14,331
-584014162768138..	365,458,612	25,588,224	13,943
2036620200163011..	103,684,481	17,407,032	9,432
7703066107207051..	41,162,071	68,323,067	9,345
8152022704930213..	462,305,791	77,000,316	8,998
-871400180321109..	44,904,615	58,577,029	8,059
6238148201090955..	27,599,383	836,047	7,675
3964496584171603..	2,478,127,532	150,246,235	7,558
1068796178499057..	501,126,761	55,576,607	7,523
3451748772731501..	447,712,534	271,514,994	6,186
6112558243277551..	373,079,319	641,870,356	5,895
-301686484049761..	31,554,688	16,826,679	4,698
1564808417629008..	761,832,666	3,363,320	4,401
-751012144730424..	206,711,025	180,667,433	4,394
-666811978267278..	7,490,445	39,145,571	4,290
-298751466829959..	27,312,177	70,683,040	4,135
6909346088261523..	117,466,112	0	3,854

Total contribution of different age and gender groups over the year



Total views in combined organic and paid on each page

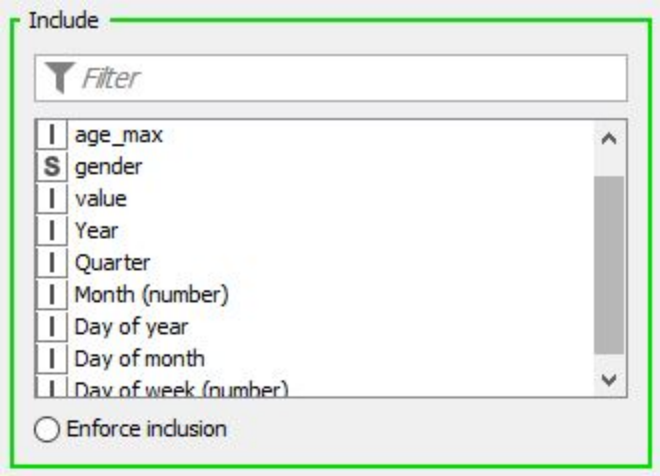


CPM imputation

Data Preparation

- The first step we took was to clean the cpm and page level datasets in R.
- The page level data set were merged together to form a table called year datasets
- The female and male column were concatenated into one column and named gender, f represents female, male represents male while U represents both.
- We changed the data type of the age-min and age-max column from character to numeric
- Imported the data into KNIME
- Read the cpm data by using the file reader and diced the date column to extract year,month, weeks and days to be able to delve more into the date.

Date dice

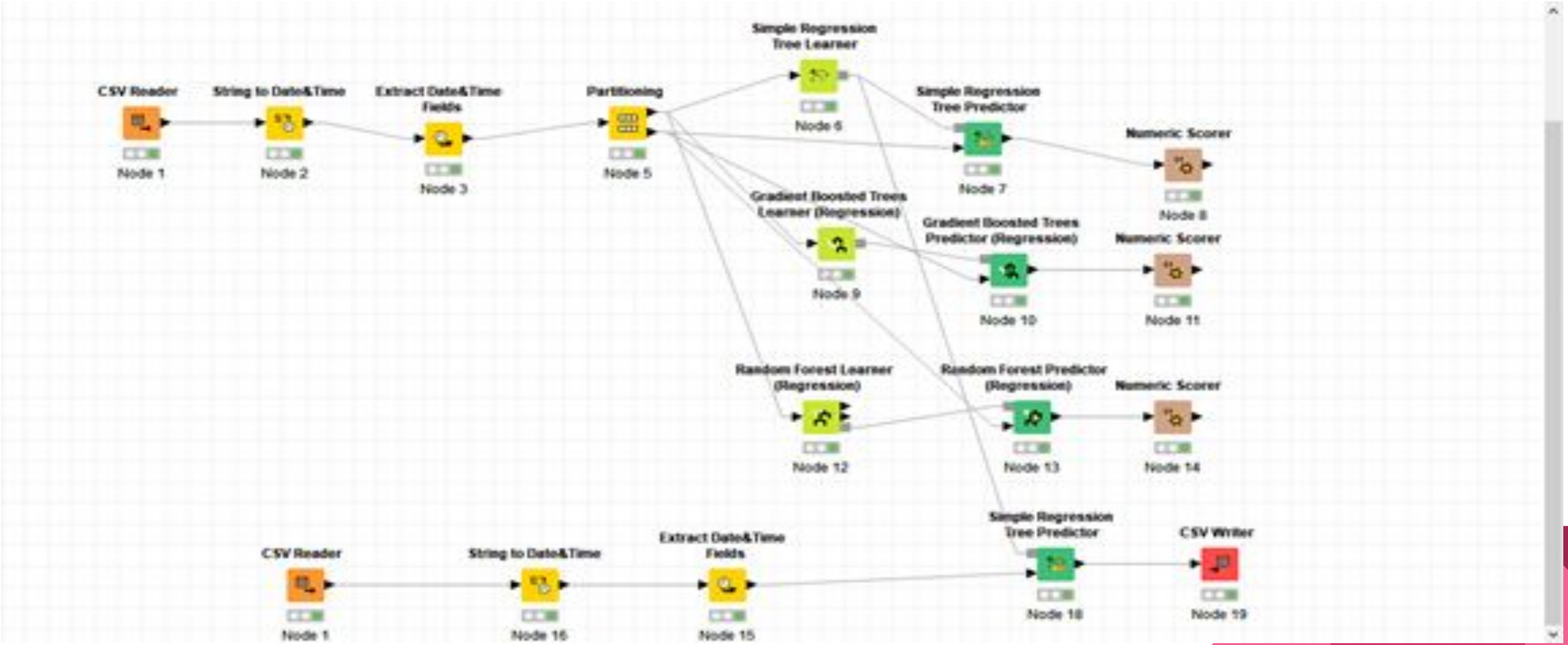


Include

I age_max
S gender
I value
I Year
I Quarter
I Month (number)
I Day of year
I Day of month
I Day of week (number)

○ Enforce inclusion

Diagram of the nodes used in KNIME



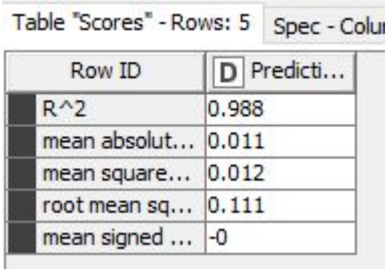
CPM imputation in KNIME

- We partitioned the data to train the cpm at 75% and test the remaining 25%
- Random forest and decision tree were used to analyze the cpm to find it’s accuracy.
- The random forest gave a positive result of 0.979, while the decision tree gave a positive correlation of 0.988 which means there is high correlation between CPM, age, gender, HId and the year
- Since the 2 models are good, we decided to use the decision tree to predict the cpm in the year datasets

CPM imputation

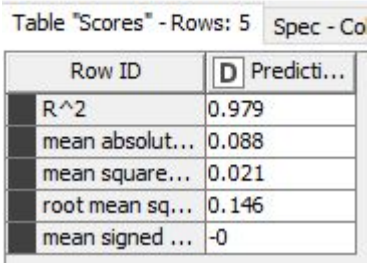
Evaluation

Regression tree



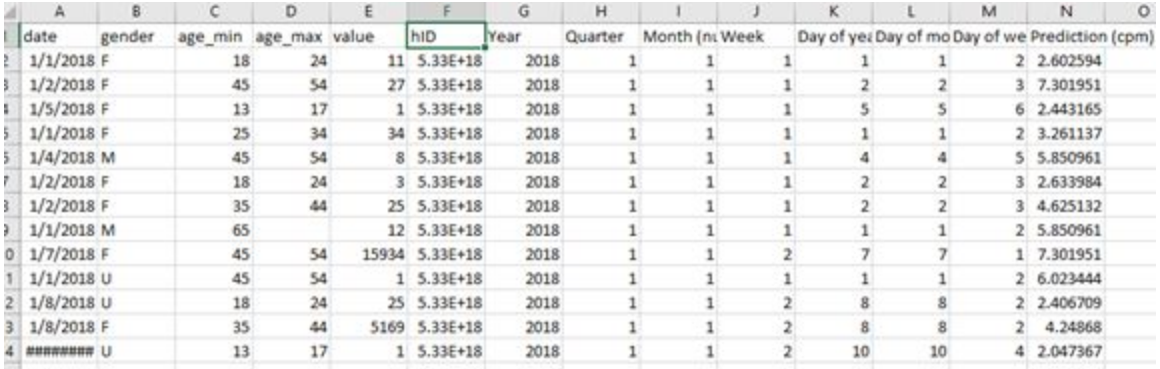
Row ID	D Predicti...
R^2	0.988
mean absolut...	0.011
mean square...	0.012
root mean sq...	0.111
mean signed ...	-0

Random forest



Row ID	D Predicti...
R^2	0.979
mean absolut...	0.088
mean square...	0.021
root mean sq...	0.146
mean signed ...	-0

CPM of All hIDs



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
date	gender	age_min	age_max	value	hID	Year	Quarter	Month (n	Week	Day of ye	Day of mo	Day of we	Prediction (cpm)	
1/1/2018	F	18	24	11	5.33E+18	2018	1	1	1	1	1	2	2.602594	
1/2/2018	F	45	54	27	5.33E+18	2018	1	1	1	2	2	3	7.301951	
1/5/2018	F	13	17	1	5.33E+18	2018	1	1	1	5	5	6	2.443165	
1/1/2018	F	25	34	34	5.33E+18	2018	1	1	1	1	1	2	3.261137	
1/4/2018	M	45	54	8	5.33E+18	2018	1	1	1	4	4	5	5.850961	
1/2/2018	F	18	24	3	5.33E+18	2018	1	1	1	2	2	3	2.633984	
1/2/2018	F	35	44	25	5.33E+18	2018	1	1	1	2	2	3	4.625132	
1/1/2018	M	65		12	5.33E+18	2018	1	1	1	1	1	2	5.850961	
1/7/2018	F	45	54	15934	5.33E+18	2018	1	1	2	7	7	1	7.301951	
1/1/2018	U	45	54	1	5.33E+18	2018	1	1	1	1	1	2	6.023444	
1/8/2018	U	18	24	25	5.33E+18	2018	1	1	2	8	8	2	2.406709	
1/8/2018	F	35	44	5169	5.33E+18	2018	1	1	2	8	8	2	4.24868	
#########	U	13	17	1	5.33E+18	2018	1	1	2	10	10	4	2.047367	

Findings

- There are 530 pages
- The highest amount of cpm across all hIDs and pages is $9.09
- The average amount of cpm is $3.56
- The highest number of page impressions are female between the age of 18-24
- The lowest number of page impressions are the teenage page
- The age range between 34-54 has the highest amount of cpm and they are female

Limitations

- Despite model showing a good correlation, we still can not get the actual cpm values for some of the H.ID as other variables cannot be incorporated into it.
- Training Data is too small to build a model with high accuracy.

Recommendations

- Since the highest cpm is $9.09, it means that all pages where customers ad are posted will generate profit to Viacom as they will charge their customers $25 . They will generate profit of $15.94
- The page with the highest cpm is the page of female adult of working age, which means, they are capable of paying the amount.
- Viacom should always target the page of age group 34-54 that are female
- If ads are posted on the age group 28-24, the page will generate more impressions as the age group tends to use facebook more than older adults
- The project worth execution as it will generate profit for viacom

Model to evaluate the conversion funnels

For example:

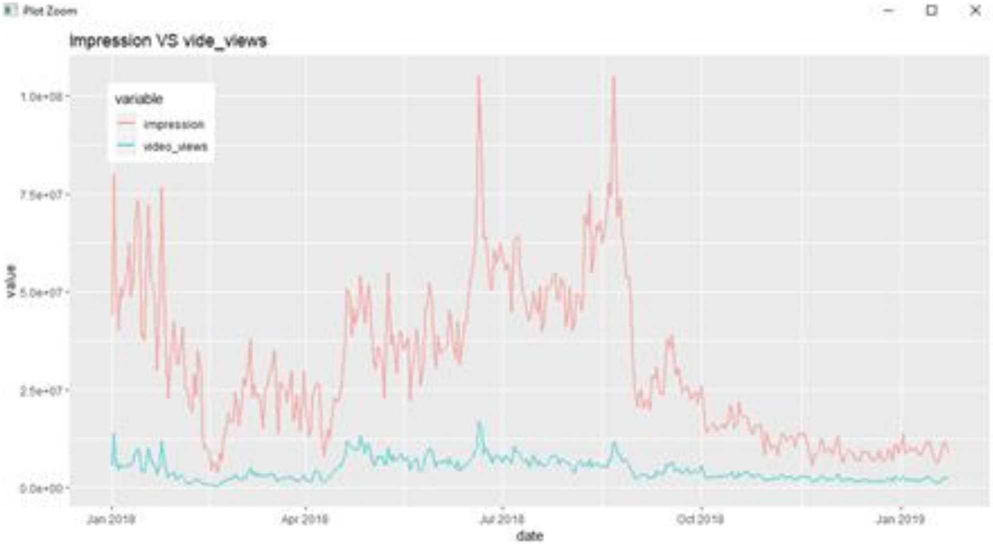
Some hID have high impression values and high video_views values.

Some hID have low impression values but high video_views values.

….More examples?

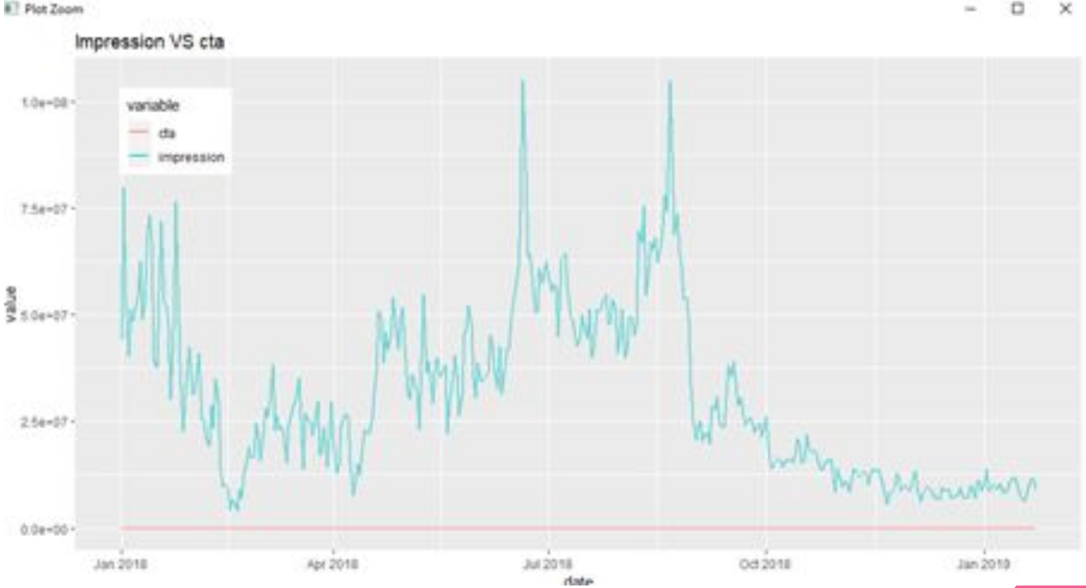
Model to evaluate the conversion funnels

hID ‘5645365176675739648’



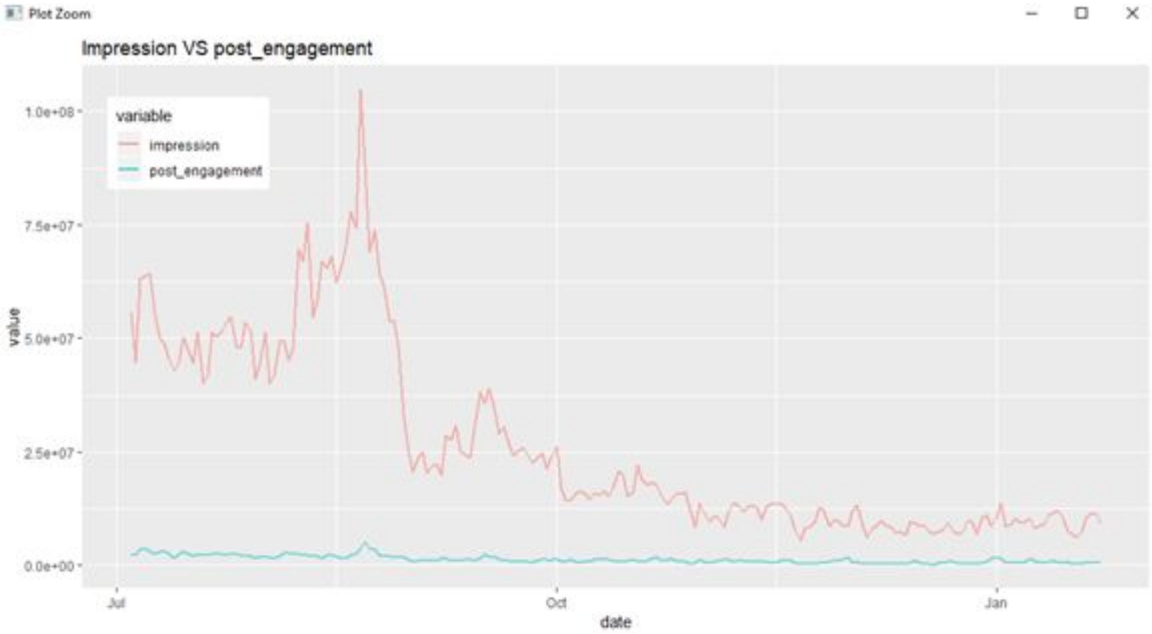
Model to evaluate the conversion funnels

hID ‘5645365176675739648’



Model to evaluate the conversion funnels

hID ‘5645365176675739648’

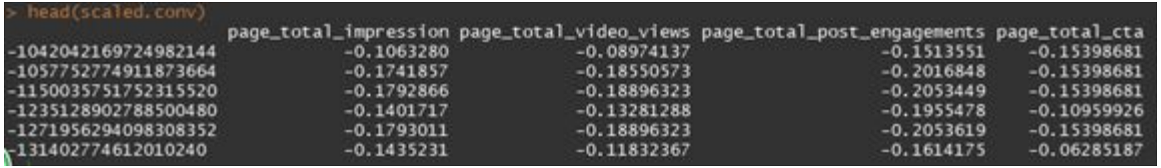


Model to evaluate the conversion funnels

Could we use clustering algorithm to group the hIDs that have the similar feature, like high impression with high_video views?

Model to evaluate the conversion funnels

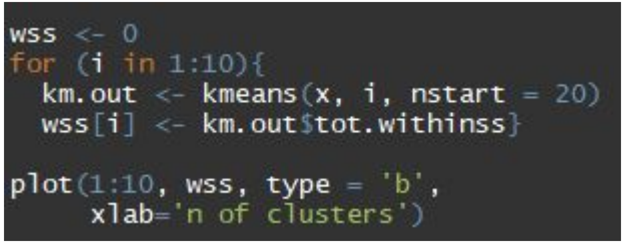
Data Cleaning and Scaling



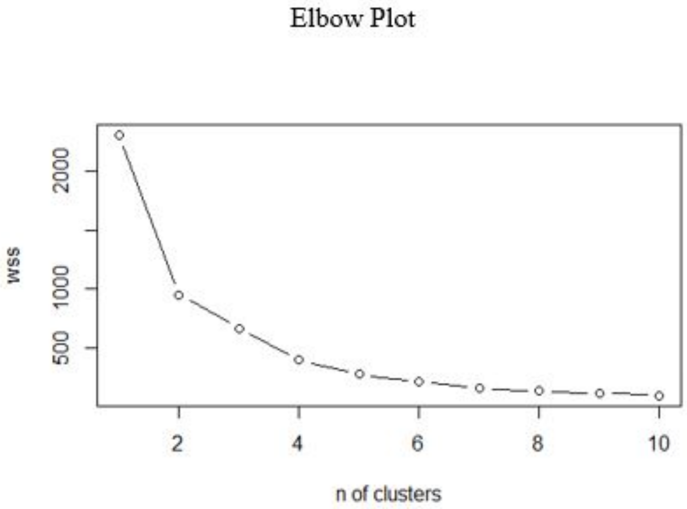
```
> head(scaled.conv)
                     page_total_impression page_total_video_views page_total_post_engagements page_total_cta
-1042042169724982144            -0.1063280            -0.08974137                  -0.1513551    -0.15398681
-1057752774911873664            -0.1741857            -0.18550573                  -0.2016848    -0.15398681
-1150035751752315520            -0.1792866            -0.18896323                  -0.2053449    -0.15398681
-1235128902788500480            -0.1401717            -0.13281288                  -0.1955478    -0.10959926
-1271956294098308352            -0.1793011            -0.18896323                  -0.2053619    -0.15398681
-131402774612010240             -0.1435231            -0.11832367                  -0.1614175    -0.06285187
```

Model to evaluate the conversion funnels

How to choose the number of clusters?



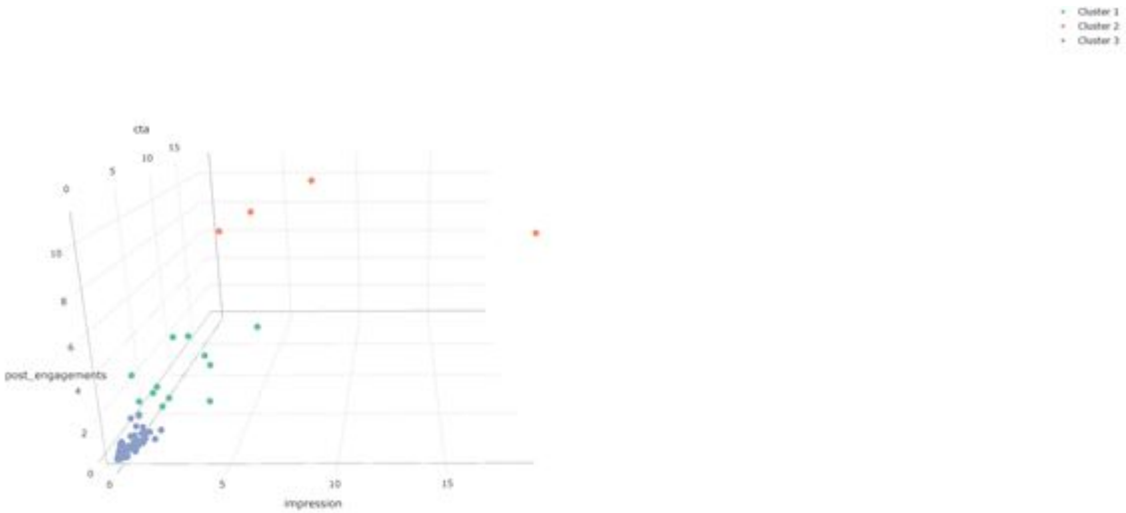
```
wss <- 0
for (i in 1:10){
  km.out <- kmeans(x, i, nstart = 20)
  wss[i] <- km.out$tot.withinss}

plot(1:10, wss, type = 'b',
     xlab='n of clusters')
```



According to Elbow principle, we select 3 clusters.

Model to evaluate the conversion funnels

Clustering:(need to be downloaded) https://drive.google.com/drive/folders/1bry7_h3mQx7cWvlm9ko25l4b7o3qCite?usp=sharing



Recommendation

- One of the advantages of this model is that it can provide the information about the conversion ability for both advertisers and Viacom. Advertiser could choose the best option for themselves. For example, if the advert is a video and the advertiser could choose the cluster that have relatively low impression but with high video_views so that they could reduce their cost. And Viacom could use this information to provide different service for various group. For example, if advertiser's budget is limited, Viacom could use quantity to reach the quality, which is providing more hIDs with the same feature to reach the goal.

References

Viacom - Official Site

https://www.viacom.com/

Week 7&8 Lecture

https://northeastern.blackboard.com/bbcswebdav/pid-20470167-dt-content-rid-55827014_1/xid-55827014_1

Thank You