

**Week 3&4 Predictive Modeling**

ALY6080 Integrated Experiential Learning

Fall 2019

Submitted on: Nov 06th , 2019

By:

ANUJ GOLESAR

**Instructor: Prof. Nathaniel C Lin**

**INDEX**

* INTRODUCTION
* ANALYSIS
* Data pre-processing
* Joining Datasets
* Part 1 : Joining Datasets
* Part 2: Calculation of Impactful parameters
* Part 3 : Analyzing the results
* Part 4: Specify how the model results may impact Viacoms business objectives
* Conclusion

**INTRODUCTION**

Viacom monetizes its social presence by creating custom branded content for advertisers that get published on Viacom’s social channels. In this assignment we will be exploring the page level data and CPM ( cost per 100 ) data for determining the changes in pricing model offered by Facebook to Viacom. Based on analysis on different datasets we have tried to determine the impact of individual variables on CPM estimates.

**ANALYSIS:**

For dealing with huge dataset we decided to break down it and work with the ones which hold potential to impact in CPM estimation in future . For this assignment we are using following dataset:

1. Cpm Estimates
2. Page level impression Organic data
3. Page Level impression Unique Gender

**Part 1 : Joining Datasets.**

For merging the whole dataset we used Jupyter notebook for importing data and manipulating it on large scale.

We combined all the files with .csv extension in one file using following commands

**ext='csv'**

**files=[i for i in glob.glob('\*.{}'.format(ext))]**

And stored the whole combined file to the local storage

**# combine all in one**

**finalCSV= pd.concat([pd.read\_csv(f) for f in files])**

Since we have several type of how the page impressions were created , we need only few of them . Thus we break down the whole page level which consist the records from both the years 2018 and 2019 into the separate .csv files which contains only specific data such as only page level impression organic data and so on .

We achieved this with the help of following commands

**for c in finalCSV.name.unique():**

**finalCSV[finalCSV.name==c].to\_csv(str(c)+'.csv' , index = False )**

Further we imported the CPM estimates data on which we will be building our prediction models with different combinations of the datasets that we have acquired till now.

Cpm data tells us about the The value of the cpm (cost per 1000 impressions) in USD .

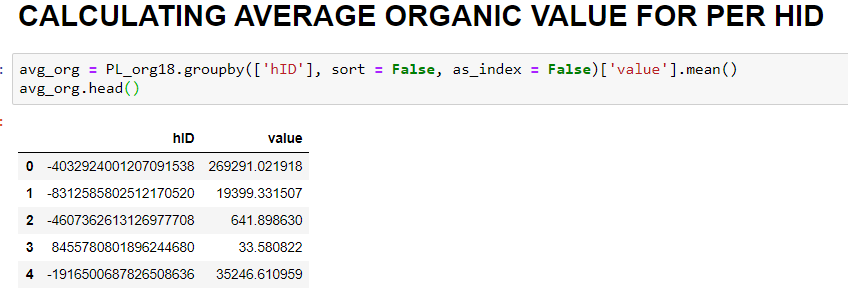
**Part 2: Calculation of Impactful parameters**

In our analysis we further determined the average organic promotion value for per hID in page impression organic data . which means it tells us about the average value per hID ,where hID is the identification number of the web page from where the traffic has come from.

For that purpose we imported the page level organic data in earlier , on executing following commands we determined the average value for organic traffic for each hID

To test the model accurately we split the dataset into training data and testing data for each of the three model ,

**Xa\_train, Xa\_test, ya\_train, ya\_test = train\_test\_split(Xa, ya, train\_size = 0.7, random\_state = 222)**



**Part 3 : Analyzing the results**

Our main moto of the analysis to determine the possible changes that could have been made by Facebook in their pricing model that is being offered to the company Viacom for the duration of two years ( 2018-2019 ) . For that purpose we will build our prediction model based on data that we have gathered by two different methods of data partitioning which are as follows

1. Out of sample
2. Time dependent holdout

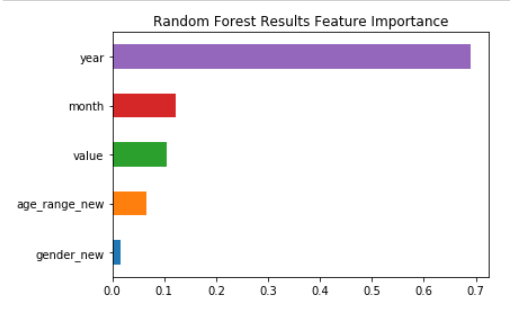
For determining how the pricing model is behaving over the period of January 2018 to August 2019 we will build several models with different combinations of the datasets. We have build random forest regression model for predicting the CPM values over that CPM data along with combination of page level data

**Following results shows us how the**

|  |  |  |
| --- | --- | --- |
|  | Accuracy : |  |
| Data | With Organic Value | Without Organic value |
| Model 1  [ CPM data 2018 and 2019 , Organic Data for 2018 ] | Accuracy : 88 % | Accuracy : 80 % |
| Model 2  [ CPM data for 2018 , Organic data for 2018] | Accuracy : 84 .34% | Accuracy : 81.4 % |
| Model 3  [ CPM data for 2019 . Organic data 2019 ] | Accuracy : 61.5 % | Accuracy : 33.34 % |

From above results table we can observe the accuracy of the models that we have built . We have calculated the accuracy of the random forest model with organic impression value and without it as well.

In model 1 results we trained the model on CPM data of 2018 and 19 and organic impression data of 2018 , however the model produced decent accuracy which to determine what parameters are important in predictions we could take help of following graphs ,



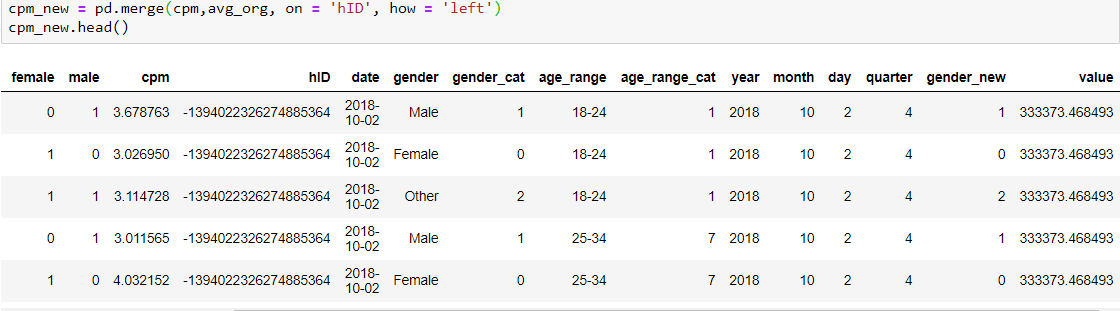
From above graph for CPM and organic data it shows that year is most important in prediction along with month and value of organic traffic impression , which tells us that the pricing model has been uniform for the year 2018 and for start of 2019 .

In second model we trained the model with CPM data of 2018 and organic impression data of 2018 , for this we performed the time dependent hold data splitting on whole CPM data , as follows

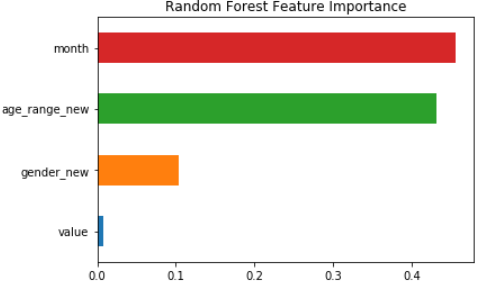
**CPM\_18 = cpm\_df[(cpm\_df ['year'] == 2018)]**

**CPM\_19 = cpm\_df[(cpm \_df['year'] == 2019)]**

Also before training model we performed the merge operation of CPM data and Organic data on hID



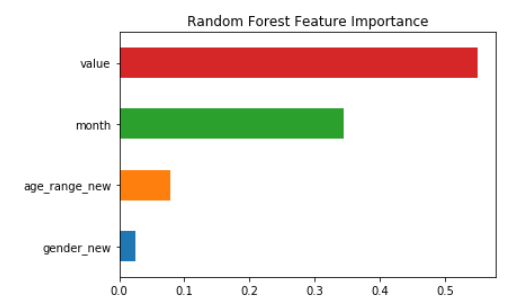
For Model 2 we got the accuracy to be around 83 % and 81 % for with and without value of organic impression , which certainly suggests that the value of the organic impression data for each hID has no effect on the Pricing Model of Facebook for the year 2018 , as we can observe the value of Value variable in previous plot and this plot :



Here the feature importance shows value contributes comparatively less to the predictions .

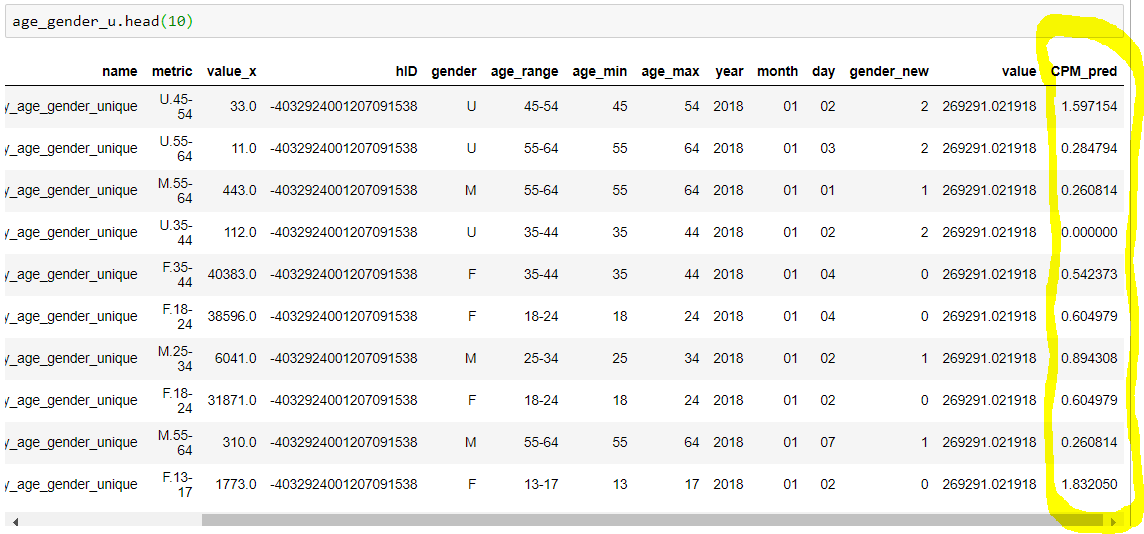
Now for third model we have used 2019 data for both CPM and Organic data to build the model ,

For which we got the accuracy of 31.43 % without value of Organic impression data . Which shows the considerable drop in the accuracy of the model in the year 2019 , where as after training model with organic impression value we got the accuracy almost double to that of the without organic value which is 61.5 % . From this observation we can judge that there has been significant change in the pricing model of the Facebook in the year 2019 . The organic value which was not impactful in year 2018 has been impactful in year 2019 moreover the contribution of demographic data in CPM prediction also seems to be less as compared to previous year.



The above plot of feature importance for Model 3 suggest that as compared to year 2018 , features such as age range and gender has low impact on CPM predictions .

In last part of analysis , we predicted the CPM values for page level age\_gender\_unique data as the testing data for the Model 3 , we got following results

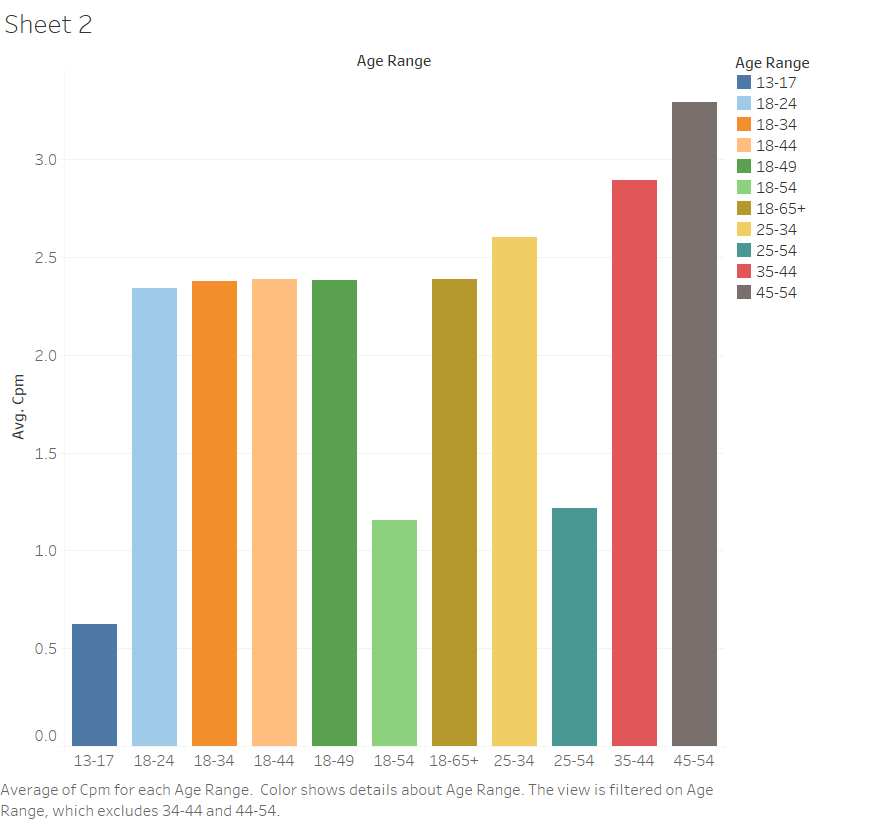


In predicted CPM column we can see that the values have gone low in year 2019 .

**Part 4: Specify how the model results may impact Viacom’s business objectives**

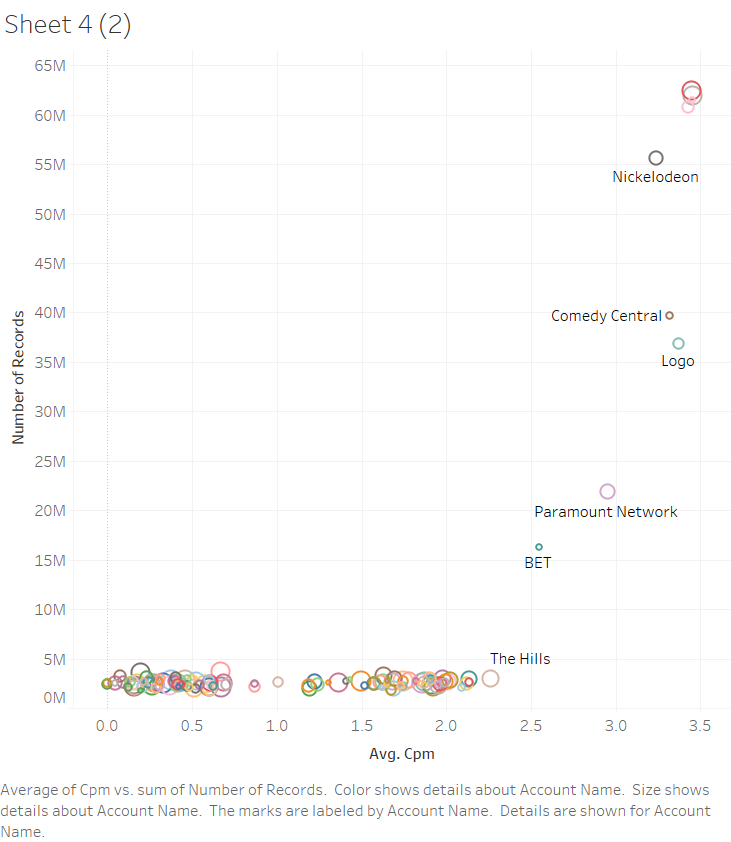
In our analysis we built 3 models to know how the Pricing model is behaving and which are the variables such page level impression values , demographic data and how they are contributing to the prediction of the CPM costs. The model has certainly suggested that Viacom needs to rethink about which account names are largest contributors to their CPM estimates , which user group is more responsive to the adds . What are the reason that the traffic is not landing on the target page.   
On further investigation on current data using rest of the page level data above questions can be answered.

To support this we can have some insights from following plots ,

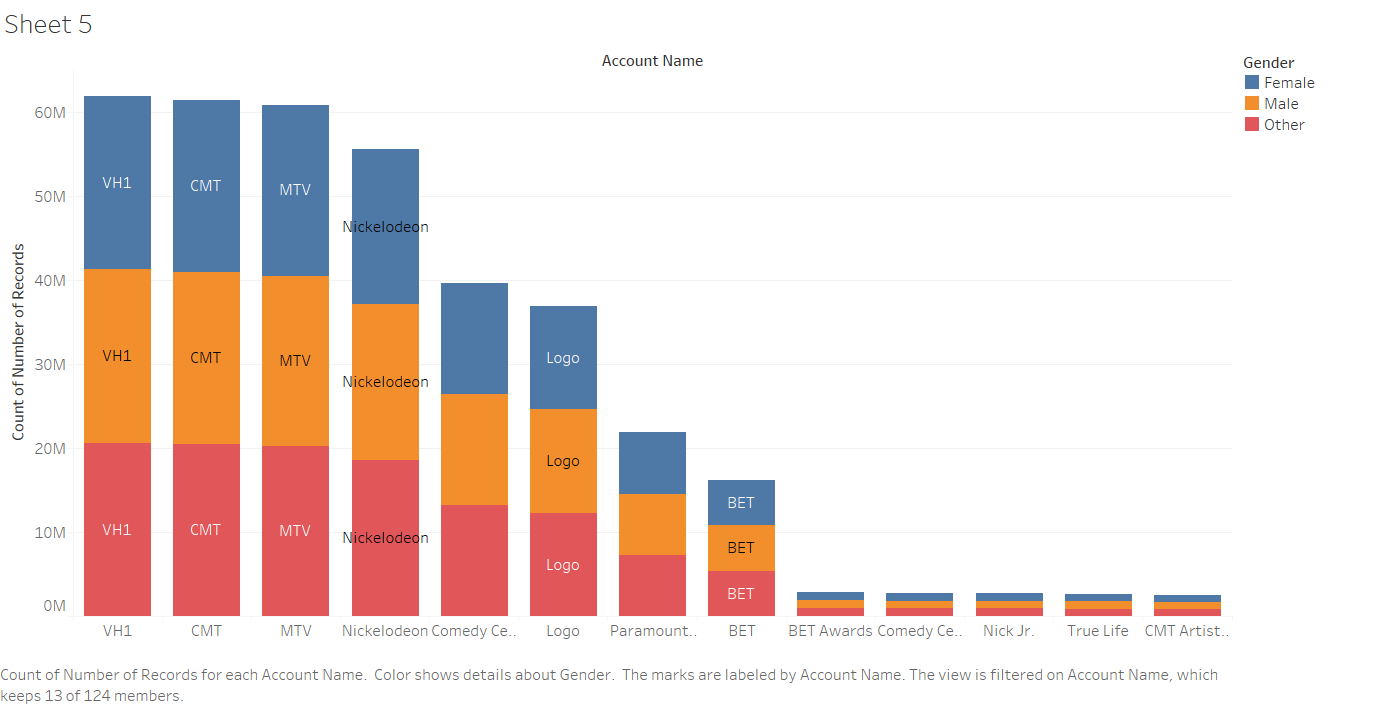


Here , we can see that viewers in the range 45-54 are contributing to the average CPM and should be targeted.

In next scatter plot , we considered the account name data to see how it is contributing to the average CPM value over the span of two years which shows that Nickelodeon , VH1 ,MTV and comedy central has been significant in contribution and thus more ads should be routed on this accounts.



With help of CPM account name data we found some facts that can help Viacom to change their business objectives , Which says that there are more number female , males and others who are creating impression on the accounts named as VH1 , CMT , MTV and Nickeledeon . Which suggests that Viacom must focus more these channels for maximizing their profit by have their social presence on some of the top viewed accounts.



**Conclusion :**

From the above analysis we could determine that Facebook has made several changes in their Pricing model for Viacom as in year 2018 demographic data has significant contribution to the CPM predictions , Whereas in 2019 it has gone down. Due to unavailability of continuous data for year 2019 we further need to investigate more and determine the significant parameters that could have been considered by Facebook in their revised Pricing Model .