**XN Project: Sponsor Project**

Hricha Adhikarj, Mounika Jakkampudi, Qiao Ma, Bowen Sun, Bingzheng Yan, Sunil Raj Thota

College of Professional Studies, Northeastern University

**XN Project: Sponsor Deliverable**

To understand which audiences should be used “to optimize total impressions, total reach, total video views, total engagements, engagement rate, high video retention, and low cost” (Viacom CBS) for Viacom CBS, our team created models and visualizations that address the business questions from the Business Sponsor.

**Executive Summary**

ViacomCBS delivers premium content to audiences across traditional and emerging platforms worldwide. Through television, streaming and digital content, studio production, publishing, live events, merchandise and more, we connect with billions of people. (ViacomCBS, n.d.)

**Goal and Objectives**

The aims and objectives often provide analysis and evaluation of visual charts that readily offer insights and enable a strategy for the following business questions:

* Total Power (kW) and Energy (kWh) Delivered to Installation
* 2.Total Power (kW) and Energy (kWh) Consumed
* 3.Power (kW) and Energy (kWh) from Solar Panels
* 4.Power (kW) and Energy (kWh) from Genset

**Project Scope**

Analysis using the data provided by the sponsor, Viacom CBS, using tools at our disposal, including, Tableau, R and Python. Data and tools acquisition outside of the aforementioned items are outside of the scope of this project, including refreshed data.

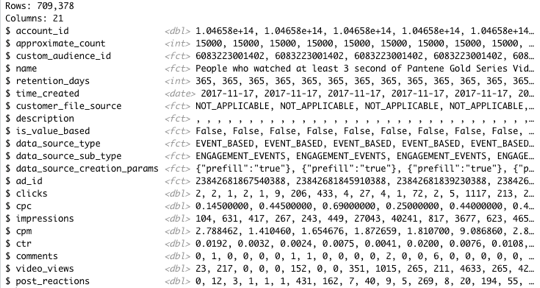
**Models, Tools, and Techniques**

The main programming languages used for this project are R and Python. The main data visualization program used is Tableau. Models and techniques used for analysis include linear regression, random forest, and XGBoost. However, a major pitfall for running models was hardware, so random forest and XGBoost were unable to complete their training cycles due to low memory allocation. This can be addressed by either accessing a frame sponsored by a company for analytics, such as Google Collaboratory, or some other platform.

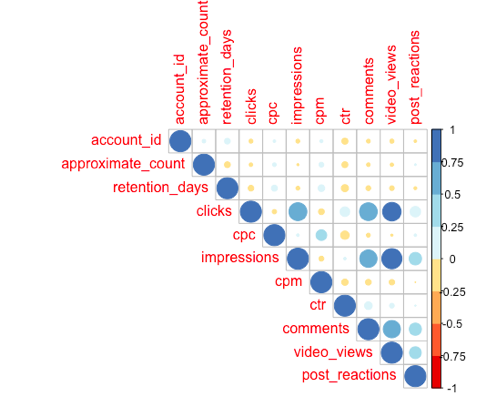
**Data Preparation & Exploratory Analysis**

The Studer Innotec dataset has multiple folders with various log files, Xtender, and Messages information in CSV format.

21 attributes and 885931 observations. We could immediately see that the dataset had some null values, roughly 20% which needed to be addressed. Since the null values were roughly 20% of the total number of observations, we decided to drop them. The next step was to ensure all the variables were in the right format. The date variable was converted to a proper date-time format and the all the string category variables were converted to factors and the data was prepared for analysis. A glimpse of the dataset after data cleaning:



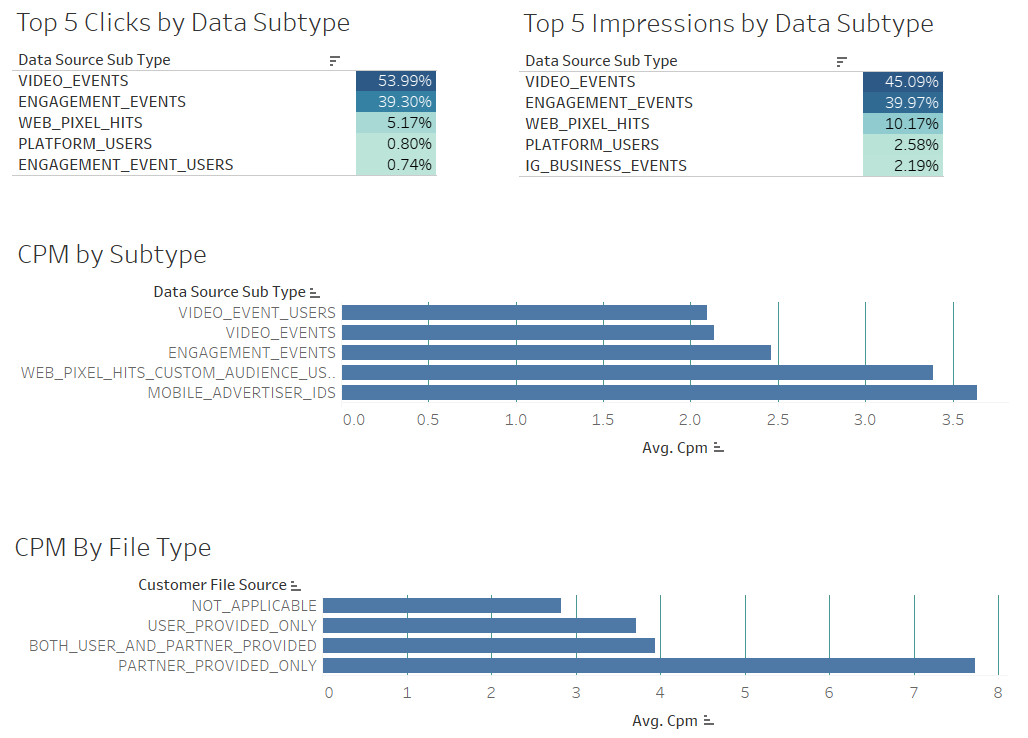
A correlation matrix is a table with coefficients for a given set of numeric variables to determine if a relationship exists between them. The coefficient tells what kind of a relationship exists between the variables and helps determine the strength of the relationship (Facer, 2020). If the correlation coefficient is closer to +1 then the variables are positively correlated meaning if one increases, then the other increases and vice versa. If the coefficient is closer to –1 then the variables are negatively correlated meaning if one increases the other will decrease and vice versa. The image below is a correlation matrix of all the numeric variables in the dataset The dark blue color indicates that the variables are positively correlated, and the red color indicates that the variables are negatively correlated.



Straight away we can see that there is a strong positive correlation between impressions and clicks. The clicks and impressions are also positively correlated to video views, comments and post reactions. We also notice that cpm, ctr and cpc are not really correlated to anything very strongly. This needs to be considered when developing a predictive model.

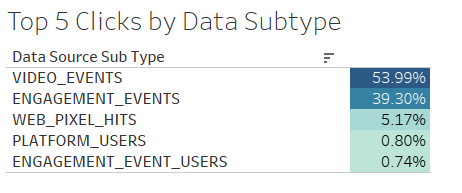
The data was prepared further to include only unique factors of custom audience id. So, our final dataset has only 3109 observations with the average values being calculated for cpm, cpc &ctr. This greatly reduces the number of observations and helps remove redundant data. This dataset was used exclusively for predictive modelling only.

**Tableau Dashboard Walkthrough**



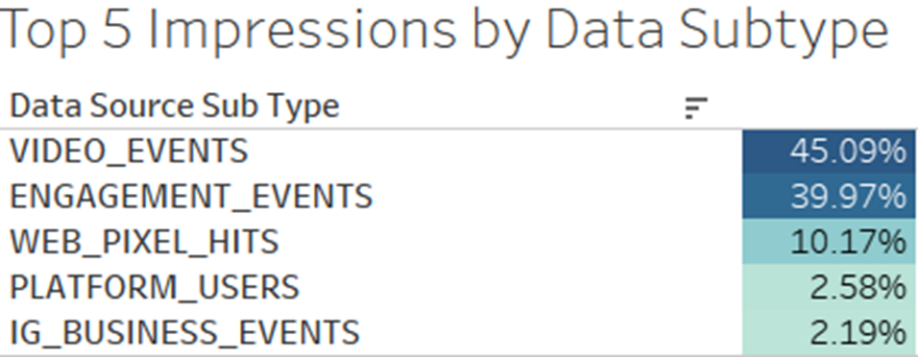
***Top 5 Clicks by Data Subtype***

This chart shows the percent of total clicks broken down by Data Source Sub Type. It is filtered on the top largest percent to highlight the top subtypes. The larger percentages are represented by darker colors dictated by the greater amounts of clicks in total.

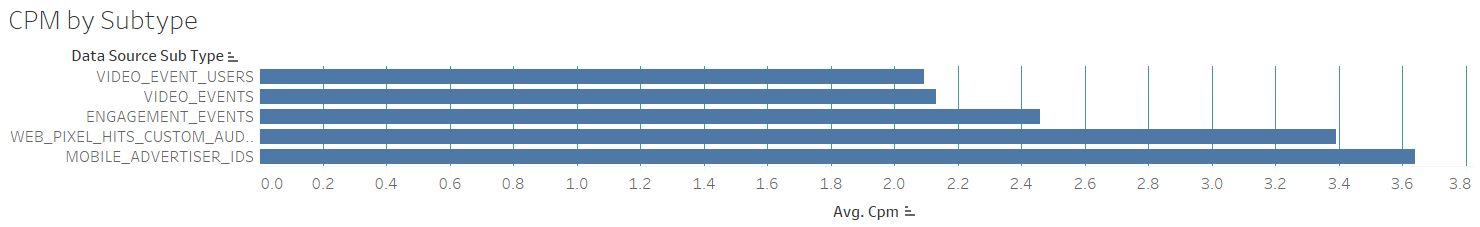


***Top 5 Impressions by Data Subtype***

This chart shows the percent of total impressions broken down by Data Source Sub Type. It is filtered on the top largest percent to highlight the top subtypes. The larger percentages are represented by darker colors dictated by the greater amounts of impressions in total.

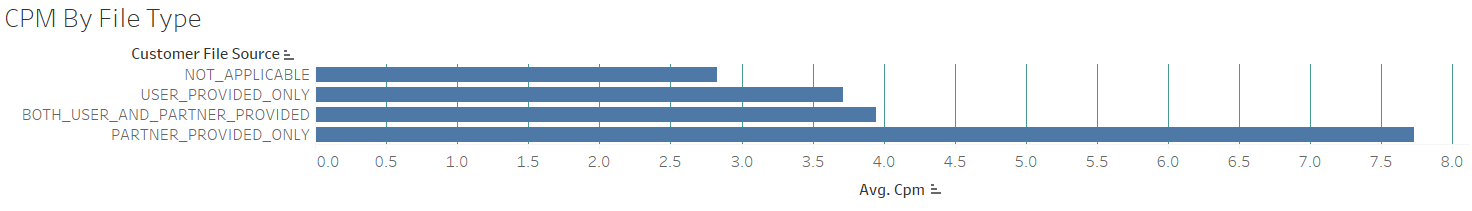


***CPM by Subtype***

This chart shows the average of cpm by Data Source Sub Type. It is sorted in ascending order by the lowest cpm average for the top 5 Data Source Sub Types. This chart has also been filtered to the top five Data Source Sub Types.

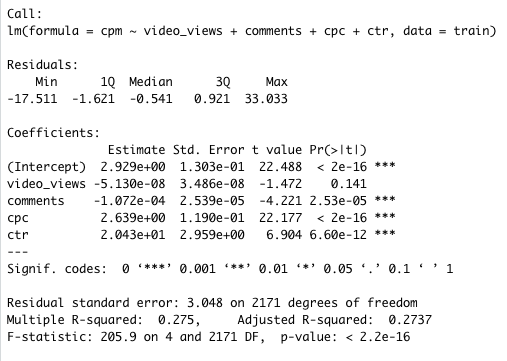
***CPM by Filetype***

This chart shows the average of cpm by Customer File Source. It is sorted in ascending order by the lowest cpm average for the top 5 Customer File Sources. This chart has also been filtered to the top five Customer File Sources.



**Performance of Predictive Modeling**

To answer the business question whether it’s possible to build a model to predict the values of cpm, cpc & ctr. We attempted to predict cpm using the stepwise modelling technique to find the most influential attributes. A reason for doing this was because the findings from the correlation matrix showed that there were no attributes which were really influencing the value of cpm. Using this technique, the attributes that had the most significant impact on cpm that gave the most accurate results were video views, comments, cpc and ctr. These attributes also had very less multi collinearity which means that the predictors were not influencing each other in any way. We constructed a linear regression model using these attributes and the summary of the findings is in the image below.



The model had high root mean square error of 2.9. Usually this is not an issue but given that most of the CPM values are below 10, this is a high deviation. The model is not completely wrong, but we would need to use more advanced analytics techniques to build a more precise and accurate model.

**Conclusion**

Since Viacom CBS is participating in an ever-increasing competitive market, with developed giants, such as Netflix, hulu, Amazon Prime, and Disney+, it is crucial that data analytics are utilized effectively. The dataset provided was from Facebook Analytics and showed insight from various customer types specifically of advertisement interactions. It should be noted that Facebook Analytics is retired, and another method will need to be used to collect data.

There are a few main takeaways from this analysis. Based on the data, in its current state, creating a reliable predictive model is difficult to synthesize. There are low correlation scores between most of the variables important to Viacom CBS (cpm, ctr, cpc). However, impressions and clicks are somewhat correlated to other variables, so it is more feasible to predict one of these two things. The provided dashboard is a sample piece of our company's work that can be integrated into everyday analysis. The charts are interactive with their original format and provide a closer look at any patterns or trends within the dataset. Finally, across the board two Data Source Subtypes have dominated for clicks, cpm, etc. Video events and engagement events seem to be the most interacted with. It would be advised to pursue more of these Data Source Subtypes.

Staying competitive in the entertainment industry will only get harder as more consumers shift to online resources for engagements. One way Viacom CBS can stay on top is by investing more time into data analytics. A significant barrier to consider would be the loss of Facebook Analytics that was used to create this dataset. A comparable collection company would need to be considered or possibly an in-house built platform might be called for at this point.

# **References**

Facer, C. (2020, December 3). *How to Create a Correlation Matrix in R*. Retrieved from Displayr: https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/

Jamaledine, R. (2017, March 31). *6 Steps to Successfully Define the Scope of a Project*. Retrieved from potential: https://www.potential.com/articles/project-scope/

Knaflic, C. N. (2015). *Storytelling with Data.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Phillips, J. (2007). *CAPM/PMP Project Management All-in-One Exam Guide.* USA: McGraw-Hill. Retrieved from https://learning.oreilly.com/library/view/capmpmp-project-management/9780071487481/9780071487481\_developing\_the\_preliminary\_project\_scope.html

Viacom CBS. (n.d.). *Analyzing Custom Audiences on Facebook.* Retrieved May 2, 2021, from https://docs.google.com/document/d/1-E5TKshJHltsftlTj9LMEL83b8Znd8Wq/edit

**Appendix: R Code**

library(plyr)

library(ggplot2)

library(tidyverse)

library(gridExtra)

library(corrplot)

library(lubridate)

library(dplyr)

library(RColorBrewer)

library (MASS)

install.packages("pROC")

install.packages("glmnet")

install.packages("Metrics")

library(glmnet)

library (Metrics)

library(pROC)

library(caret)

custom\_aud <- read.csv("~/Desktop/Northeastern/Integrated experiential learning/XN project/custom\_audience\_ads.csv")

glimpse(custom\_aud)

lapply(custom\_aud, function(x){length(which(is.na(x)))})

attach(custom\_aud)

custom\_aud <- custom\_aud %>%drop\_na()

#custom\_aud <- custom\_aud %>% mutate\_at(c("impressions", "cpm", "ctr"), ~(scale(.) %>% as.vector))

custom\_aud[[6]] <- as.character(custom\_aud[[6]])

custom\_aud[[6]] <- as.Date(custom\_aud[[6]], format = "%Y-%m-%d")

custom\_aud[[3]] <- as.factor(custom\_aud[[3]])

custom\_aud[[13]] <- as.factor(custom\_aud[[13]])

glimpse(custom\_aud)

custom\_numericcols=custom\_aud%>%

dplyr::select\_if(is.numeric)

corplot <- cor(custom\_numericcols, use = 'pairwise')

corrplot(corplot, type = 'upper', col = brewer.pal(n = 8, name = "RdYlBu"))

custom\_aud <- custom\_aud %>%

mutate(total\_impressionscost = impressions \* (cpm/1000)) %>%

mutate(total\_cpc = clicks \* cpc)

groupColumns <- c("name", "custom\_audience\_id","approximate\_count","retention\_days", "time\_created", "data\_source\_sub\_type")

dataColumns <- c("clicks", "total\_cpc","impressions", "total\_impressionscost", "video\_views", "comments", "post\_reactions")

custom <- ddply(custom\_aud, groupColumns, function(x) colSums(x[dataColumns]))

custom <- custom %>%

mutate(cpm = (total\_impressionscost/impressions)\*1000) %>%

mutate(cpc = total\_cpc/clicks) %>%

mutate(ctr = clicks/impressions)

custom\_numericcols=custom%>%

dplyr::select\_if(is.numeric)

set.seed(153)

train\_index <- sample(1:nrow(custom\_numericcols), 0.7\*nrow(custom\_numericcols))

train <- custom\_numericcols [train\_index,]

test <- custom\_numericcols [-train\_index,]

model1 <- lm(cpm ~ video\_views + comments

+ cpc + ctr, data = train)

summary(model1)

preds1 <- predict(model1, test)

actuals\_preds1 <- data.frame(cbind(actuals=test$cpm, predicteds=preds1))

head(actuals\_preds1)

lm\_rmse <- rmse(test$cpm, preds1)

lm\_rmse

step\_model <- step( lm(cpm ~ approximate\_count + retention\_days + video\_views + comments

+ cpc + ctr, data = custom\_numericcols), direction = "both")

summary(step\_model)

pred\_step <- predict(step\_model, newdata = test)

step\_rmse <- rmse(train\_y, pred\_step)

step\_rmse