**April 7th/2022**

**Agenda:**

* Go over Progress Report 2
* Explain the three primary learning objectives
  + Phase 3 in the Progress Report 2

**Gunpreet:** Use other Regression metrics besides R-square, explain why you chose R-square

* Did you think of other metrics in the method of evaluation? **Justify your choice**
* **Suggestion:** Have an appendix comparing two different error model evaluations and explain your choice

**1st Primary Learning Objective:**

**Kimberly:** Graphs have nothing to do with lower funnel and creatives

* **Saul**: Only using upper funnel creatives

**Gunpreet:** Publisher 4 is more sensitive so no matter what the performance isn’t really affected by those factors. [Finding is interesting and different from what CM is used to]

Gunpreet & Kim: The Optimal mix does not appear to be answered (Optimal mix of elements)

* Position the analysis a little bit different given the time available (talk about optimal mix of number of creatives/optimal mix message of the market e.t.c)
* Optimal Mix of Creatives (be cautious)

**2nd Primary Learning Objective:**

**Saul:** Removed 2021, Middle funnel have higher CTR than publisher 4 even though they’re almost the same

**Gunpreet:** Will have to think about it

***(Gunpreet 1&2 are similar in terms of how they were done)***

**3rd Primary Learning Objective:**

**Saul:** Similar trend, suggests increasing spending for Publisher 4 or recreating

**Saul:** Lower funnel buy Flow Entry, doesn’t seem impactful, Publisher 3 have a little higher by-flow conversation rate than Publisher 4

**Gunpreet:** Skip by-flow and focus on primary learning objectives and put the secondary ones on the appendix

**Impact of creatives in market on performance**

**Saul:** Did the same thing but aggravated weekly instead, found the CTR for upper, middle, and by-flow/lower funnel

**Gunpreet:** How do you define in-market?

* **Saul:** Launch market with sum 1 for creatives after launch

**Gunpreet:** Look at average instead of the sum for the plot graph

**Exploratory Data Analysis**

**Kimberly:** Switch it up and make it averages

**Modeling**

**Gunpreet & Kim**: What factors are the most important in the prediction of the KPI. (Creative before launch is what CM knows but they don’t have other information). Provide insights that can be changed on the creative (say you need a creative for mid funnel and publisher 2 what performs better). More about relying on a creative rather than the performance data (which is produced as a result of the creative). When picking features it is not about the data.

* Have a model that can predict performance
* ***Don’t lose sight of the Model***

**Presentation suggestion**

**Consider:** Balance technicality (e.g., Heteroskedasticity or Durbin-Watson, keep it in the back of the appendix to explain it to those that need it). Also talk about the business aspect of it.

**Next Meeting:**

* **(1-2pm) MST Monday/Tuesday**

**Project Objectives**

* **Priority #1 - Building a model that predicts**
* **Priority #2 - Answering the primary learning objectives** 
  + **1 - Do feature selection**
  + **2 - Approach it differently but if you don’t have time it is okay**
  + **3 - Impact of message on the market**
* **Priority #3 - Having it in Tableau or something else (have a wireframe of what the dashboard will look like in Tableau)**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**March 31/2022**

**Feyre:** Current team update & Overview of meeting:

* Profile report didn’t work (just try adding things to the algorithms and seeing the accuracy as we go)
* One person keeps adding the new variables to the algorithm while two others working on the feature
* See if it is possible to combine the UCIs?
* **Yutong** -> Do some feature engineering
* **Feyre** -> Work on EDA
* **Saul** -> Finished the EDA (including 2 primary learning objectives)
  + Uploaded the features
* Break apart the Funnel specific
* **Everyone** -> Use SkLearn and try to get the highest prediction power
* Got some baseline for prediction, runed it, and focused on improving prediction

***Upcoming tasks and updated deadline:***

* **Progress Report #2 [Create a template for after March. 17th]** - April 7th @8am
  + Share your update
  + Send a completed version to CM by Monday **April 4th**
* **Code** - April 14th
  + 2 people start working on this
* **GUI** - April 14th
  + 1 person start working on this (Muhammad &/ Yutong)
* **Report [create a template]** - April 14th
  + 1 person create a template (Feyre)
* **(PPT)/Presentation** - April 14th @9:15am MST
  + 1 person create a template (Feyre)

ProfileReport didn't work, will have to do more manually

Division of tasks, and plan going forward

* Progress Report
* GUI
* ML Algorithm/Improving the performance

**Gunpreet:** making sure that we have limited/filtered dataset

**SAUL:** Filtering significantly reduced our observations, now only about 2500 obs**.**

* Got rid of outliers, using a maximum threshold
* Tried other techniques, but it removed too many observations
* Grouped by funnel and publisher, for creatives in each week
* ClickThru rate seems to be strange,

**Gunpreet Re ClickThru:**

* Did you look at how many creatives per funnel?
* Value count on funnel(see what split is)
* 60% of creatives are upfunnel
* Maybe having more messages in upper funnel is not good, or overloading of messaging causes lower CTR
* Lets be more focused, funnel specific KPI’s

**Kim:**

* Too many messages may erode engagement
* Not in the market at the same time, been normalized to weeks
* Think about the overall audience, Upper funnel is supposed to be broad, and gets narrower as we move down funnel
* Upper funnel can be a lot about brand awareness rather than targeting
* Focus should be on funnel… Upfunnel should be specific for that KPI
* Still can be correlation, but more of an insight… not predicting

**Kim/Gunpreet:**

When publishers have few ceratives, its not that convincing, so can just say that we didnt have enough observations

**Gunpreet:**

Can you show the way of calculating the number of creatives and…

**Saul:** I counted the values and summed them for each feature in the market per week.. Grouping by funnel and publisher

**Gunpreet:**

Recommend using absolute values for weeks, not relative values

Give better distribution for funnel specific, and give insights - like for certain publishers/in certain funnels KPI’s do X during “christmas time” for example

**Back to Saul:**

For second Learning Objective - Impact of number of creatives in market on performance

* Will have to redo without the normalized weeks
* Grouped Proportions by publisher (social Only)
* Outliers dropped dramatically when filtered out to only include social

**Gunpreet:**

* Look at changes week on week (growth rates/ decline rates)

**Gunpreet/Kim Feedback:**

* How far along are we into the modeling? Ans: Just started, but trying to figure out which models to use
* Who is leading the modeling? Ans: More of a group/collaborative effort
* Are we still on track? Ans: Yes
* What is the plan with the GUI?: Ans:

Which models are good?

Pycaret? Random forest

For baseline, use linear regression

Use a proper library

Prophet Library from facebook - time series forecasting

Regardless of what library we use, make sure we go through the checks. I.e. deseasonalize the data…stationarity checks, detrending

**-------------------------------------------------------------------------------------------------------------------------------**

**March 17th/2022**

**Saul:** Update on Progress Report 1 (shared as a PDF on Slack) and show EDA

* Will have to rely on manual exploration
* We have some extreme outliers that we still need to clean
* Linear regression -> will create bias and influence our model

**Gunpreet:** Did more of pre-processing than cleaning

**Saul:** We have our own EDA but we are thinking of using pandas.profiling b/c the data is too

large to compile

**Gunpreet:** Use in 1000s row b/c pandas.profiling is slow, and limit yourself to 1-KPI

(e.g., Click through rate) [Recommendation: use sub data instead of the full data]

**Kim:** Break it down into 1-funnel (e.g., isolate up-funnel click through rate and KPI) and see if that works and gives you anything (to see how it can expand the scope of the data that we are looking at.

**Saul:** Once we use EDA we will move to model selection and profiling

* Gave a run down of the EDA (many of which had different values)
  + There wasn’t any extreme outliers and they were visually represented
  + Channel proportion >80% rows were displayed while the rest were social
  + There were more UF while MF & LF were equal
  + Publisher Proportions - there was at the most 5 rows (some will not be used as variables b/c the variation is not much across the majority of the dummy variables)

**Gunpreet:** Publisher Proportions (rather than eliminating the small Publishers with small values) group them up and label them as “Publisher X or another publisher” (helps remove noise and helps you determine how many rows you actually have to draw) - Let the model decide which features/variables are important.

**Kim:** Do not display publisher and social together, break them apart and then find correlation to group them. (Will check if the 20-80 Channel Proportions is not normal) [You are counting each line as a unique line but display has a bunch of difference for ad sizes even though it has the same cids, making display look like it has much more assets]

**Saul:** Team will experiment to see if we can play around with the channels. Remaining graphs are difficult to describe.

**Kim:** Summarize key things that stand out for you and summarize them, you don’t need to explain all the graphs in detail.

**Gunpreet**: What is the determining factor between each plots

**Saul:** Number of observations (for e.g., in the we see that size 3 has more discount observations that are higher and tails off [has either no discount or has high discount on top compared to the size 15]). Most graphs don’t really have the answers we’re looking for.

**Gunpreet:** If there are single isolated creatives, you might as well just remove them]. *Latent Impression* (we may not be putting any money but we may see such impressions which affects the click through rate or was not paused] suggestion: (Bind the click through rate to see the standard distribution rate across different binds and anything above a certain value can be eliminated/try other edumetric ways).

* Can you isolate the spend creative for a week 7 spent for 507 and 509 and see if the normal spent is normal/anything above and determine whether you want for (e.g., you can remove for anything below $10) (Look into Max put for jupiter)
* Include the primary & secondary learning objective (think about how you can do the EDA and data analysis, and some of the answers you might find from the modeling) and try and answer the primary & secondary learning objectives as much as possible
* Rerun the visuals with the outliers removed, write a brief insights, highlight it and then share it with CM reps

**Saul**: Features - began with RFE (Recursive Feature elimination) to make the estimates more robust & more competent than the results

**Gunpreet**: Think that not all impressions are available before launch - We can’t use the variables we have like total impression b/c we are using the feature variables

**Feyre:** Feedback from Dr. Crost

* Like the set up of the ML algorithm (split it into training, validate, and test) -> said that this was the best way to avoid overfitting and get robust prediction.
* Recommends to consider a nested cross-validation approach - For e.g.,:
  + Split data into 6 blocks
  + Use 5 blocks to train and validate the model (e.g., w/ 5-fold cross-validation) this is the inner loop of the nested algorithm
  + Generate predictions for the 6th, which serves as test data
  + Rotate blocks to use a different one as test data - the outer loop nested algorithms
* He stated that this approach would allow us to use all of our data for training, validation, and testing, and in addition get us out-of-valida-sample predictions for all observations instead of just for a small fraction.

**Feyre:** Find a date (April 12th/14th) for the presentation and determine whether it will be held in person/online?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**March 3rd/2022**

**Saul:** (Update on current status)

* We’ve encoded the data set using DVs
* Cleaned up the Dates w/ the KPI’s
  + Aligned the values by the weeks (so that they start at the same index)
* Next -> Start selecting features for modeling (recursive feature elimination, and user feature importance)

**Gunpreet:** Are you jumping into Modeling/are there an EDA (Exploratory Data Analysis - Analysing the data as received upfront is an important of the deliverable)

* **Wes:** Dr. Crost said just throw whatever you have on the ML Algorithm and figure out the importance

**Gunpreet:** Look at the data itself and decide what is happening, and determine if there are any outliers before jumping into Modeling (Go from future analysis to data analysis). Make sure Data Analysis is a key part of the project for you (do not ignore what the data looks like or means before and after you clean it.

**Gunpreet:** Are you thinking of KPI prediction as classification task/regression task?

* **Saul:** regression b/c we are trying to predict a numerical problem
* **Gunpreet:** [Pick both methods b/c there can be different performance, look at VIF (Variance Inflation(implaition??) features).
  + Compare what the overall gap is selecting one way/the other (they tend to typically be close to each other) you can also try to stack them. [There is no right/wrong method, just justify why you chose to do things the way you did]
* **Gunpreet:** Try regression method/other methods - Even though random forest would give you an answer, sometimes it can give you the wrong answer and will run longer…[Interpretability is a big component of selecting a model] Talk about what the gap is when selecting one model/another.
  + Google how to select the best models for key feature prediction…?

**Gunpreet**: How have you approached Feature Engineering:

* **Wes**: aggregated spending & impressions on the right hand side. Peaks of each one (Max values). Currently talking about grouping & making more interactive terms with the right hand side terms by utilising DVs. Growth of rate refers to change that is taking place on a weekly basis.

**Gunpreet:** Provide the explanatory (exploration part of the data) as part of the deliverable, and should be done either in-parallel/before the modelling (ties to our primary learning objective) [Up to you to decide whether you want to do it before/in-parallel to the modelling step]

* What is the data telling us at the publisher/other level
* Helps to form an informed KPI prediction

**Next steps:**

* **Feyre:** Send out bi-weekly meeting invitation
* **Start working on Progress Report #1 due March 17th**
* **Show Gunpreet & Kim the cleaned data (screenshare)**
* **Get the updated data from Muhammad (do an exploration of the data separately then collaborate what we find)**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Feb. 3rd /2022 - Meeting with CM**

**Current updates:**

* Project Proposal Plan is nearly finished - once we complete final edits we will share it with our project partners along with the project proposal agreement (the agreement that you were shown by Dr. Crost at an earlier time)
  + Includes: Project milestones and deliverables. For now we’ll share a picture until
  + Currently working on a PPT presentation - this will be shared once completed

**Next step:**

* Data - cleaning, analysing e.t.c. - nothing major has been done atm - We might have more questions in the future (and if needed potentially changing the meetings to weekly) However, the current system works.
  + **Gunpreet:** Weekly 30min or bi-weekly to an hour is workable as well
* Next upcoming major task is the 1st progress report which is due on Mar. 10th - We will be providing feedback. (How often would you like to receive updates)
  + ***Gunpreet:*** Give us a couple points of updates on how things are going, how far along we’ve come, if there are any issues, or anything new/modifications being changed - Goal: **To address major points and stay on the point. [Address issues asap]**
  + **Do weekly updates on Tuesday and a few days before the deadline of the project.**

***Open floor:***

* ***Muhammad:*** any formal way to list experience on resume, CM will get back to us next meeting
* ***Feyre* (presenting Saul’s question):** Do we only need to develop three forecasting models based on the funnel specific (upper, mid, and lower)?
* Up to you - can be one funnel specific or three (then you choose whichever one you think is best)
* ***Gunpreet*:** What are your thoughts on the project so far?
* **Gunpreet suggestion:** 
  + Moving forward: Dissect the data and find specific data based on what creatives we need. You might not use the whole data
  + The goal of Data Analysis is to Share the story that you have (Use whatever software you want, you don’t have to be restricted to Tableau). You can use Python, R, Tableau, Excel e.t.c… (Prefer Python b/c we can use it)

**Upcoming task for everyone:**

* **Sign and send the Project Proposal Agreement**