

# **LEVERAGE FEATURES FROM CLICKSTREAM**

**PREDICT IF THE USER WILL PURCHASE / DROP OUT**

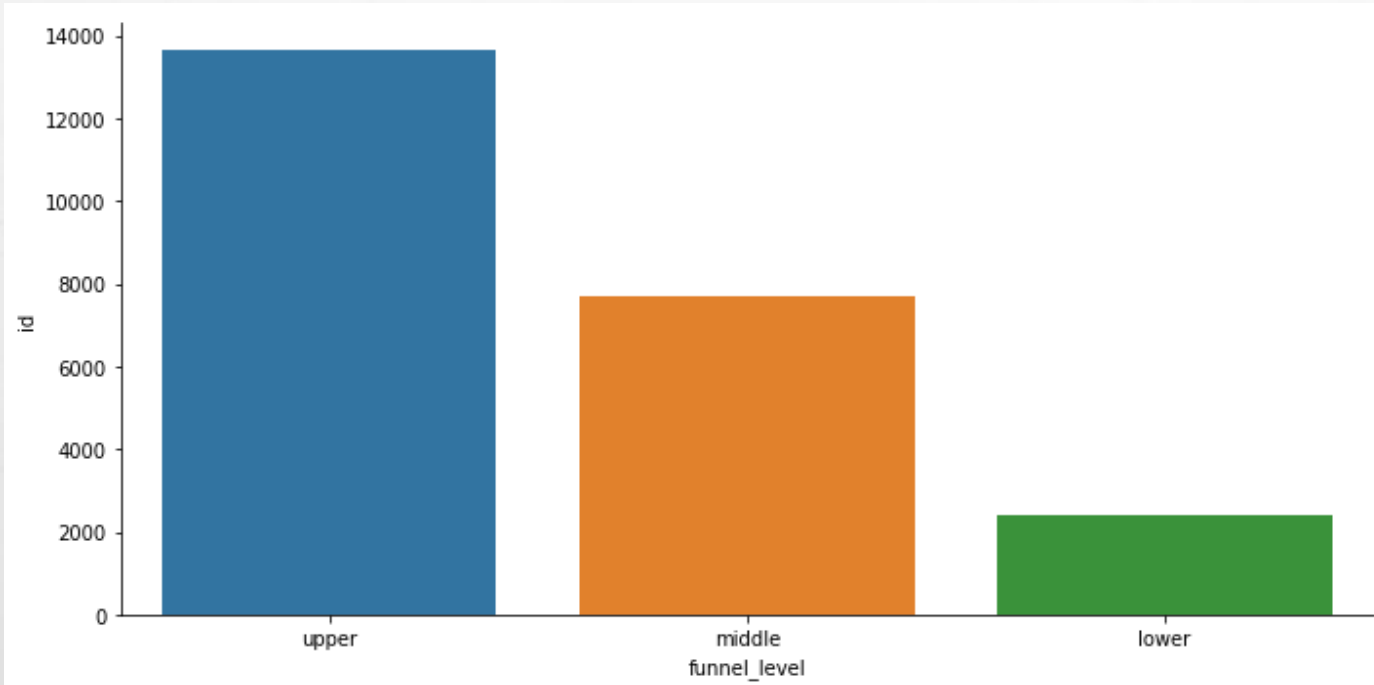


**MANKAYARKARASI C**

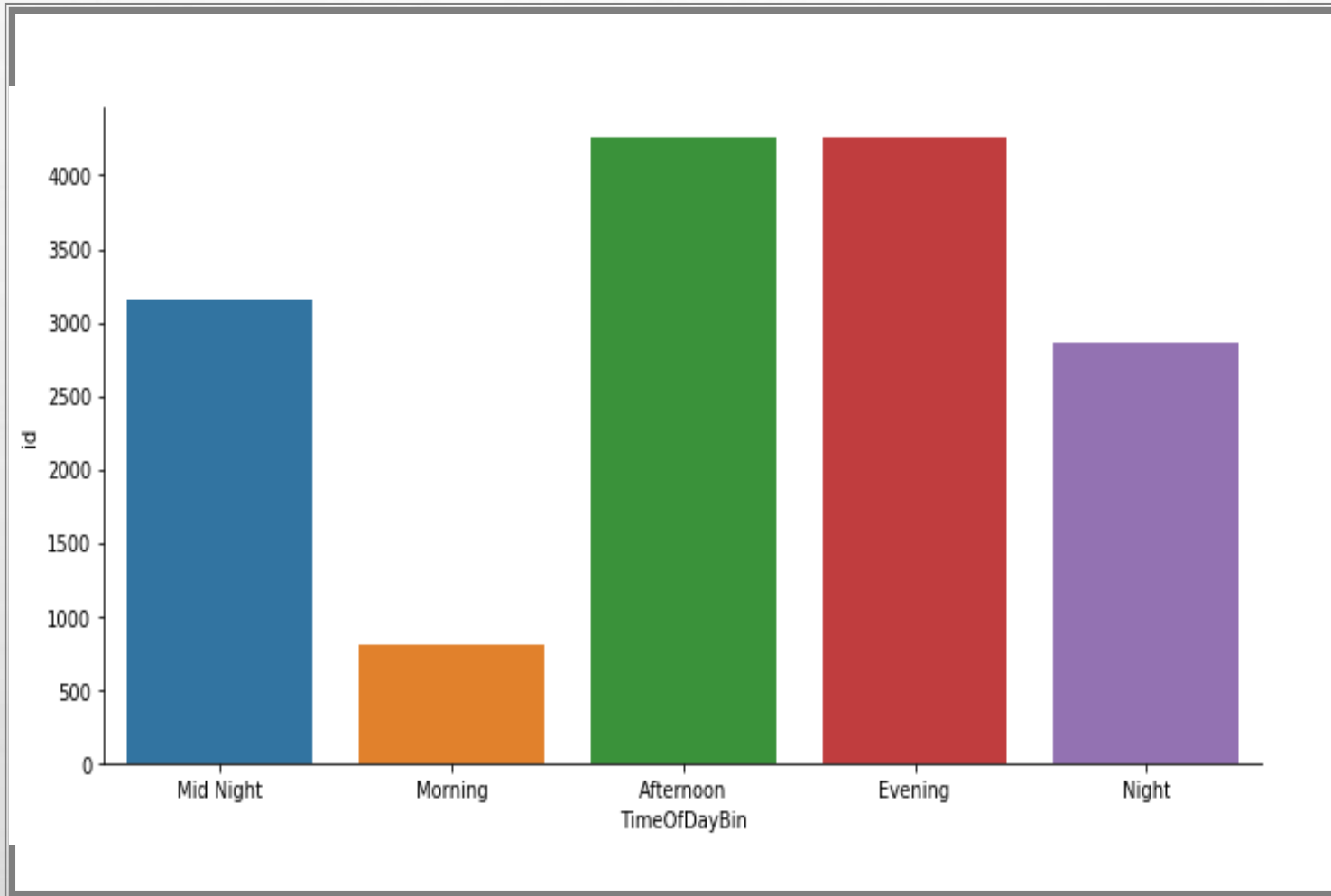
# PROBLEM OVERVIEW

- **GIVEN A SEQUENCE OF CLICK EVENTS PERFORMED BY USER DURING A TYPICAL SESSION, THE GOAL IS TO PREDICT WHETHER THE USER IS LIKELY TO MAKE A PURCHASE OR DROP OUT OF THE PATH.**
- **EACH RECORD/LINE IN THE FILE IS AN ACTION OR A CLICK DONE BY A USER IN THAT SESSION, EACH SESSION IS A UNIQUE USER. THEY WILL HAVE MULTIPLE RECORDS IN A SESSION BASED ON THE ACTIONS AND CLICKS THEY ARE DOING TO COMPLETE AN ORDER OR ADD TO A CHECKOUT OR BROWSING AND EXITING WITHOUT CHECKOUT OR ORDER.**
- **THE FUNNEL NAME WILL GIVE INFO OF HOW THE CUSTOMERS TRAVERSE THROUGH THE WEBSITE , CUSTOMER START WITH UPPER FUNNEL WHERE THEY DO LEARN AND DO PRODUCT VERIFICATION AND THEN THEY GO TO MIDDLE WHERE THEY SELECT PRODUCTS AND THEN FINISH UP IN LOWER WITH PAYMENT AND OTHER ADDRESS INFO.**

# EXPLORATORY DATA ANALYSIS



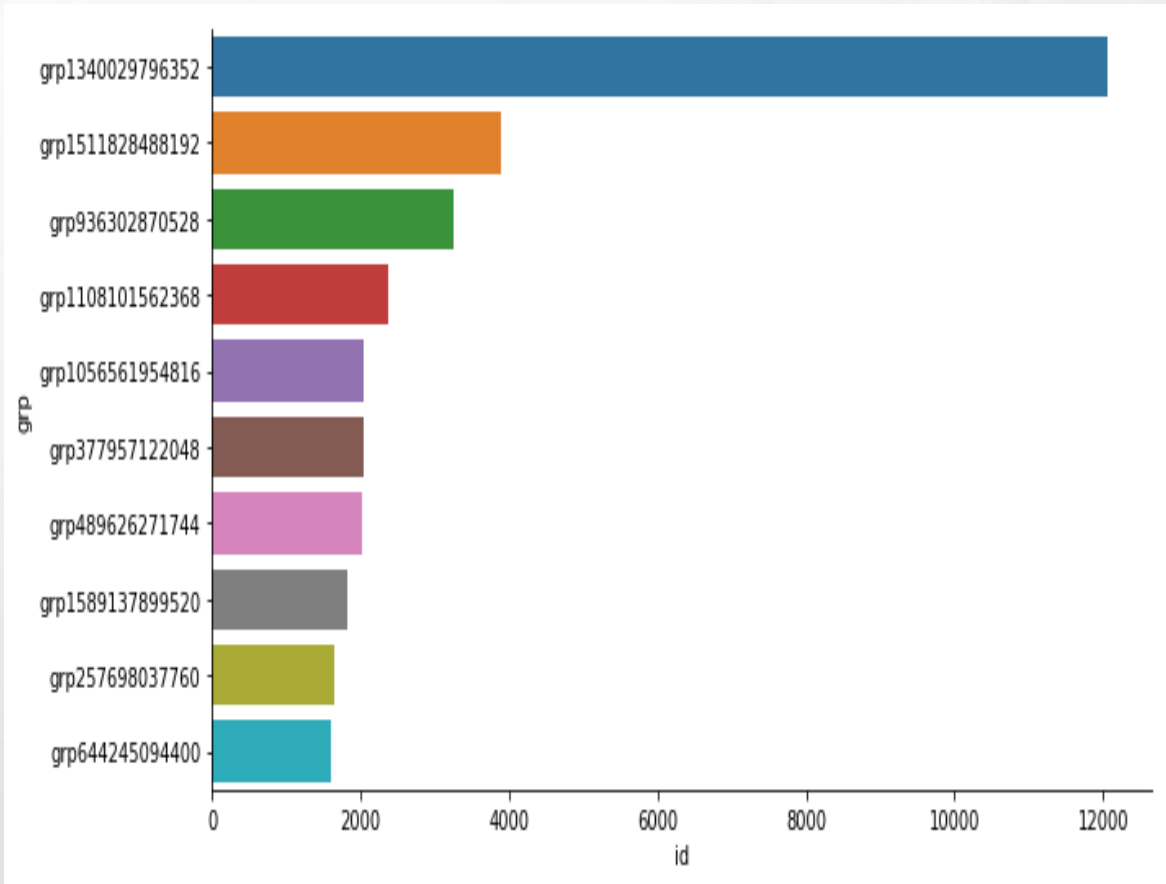
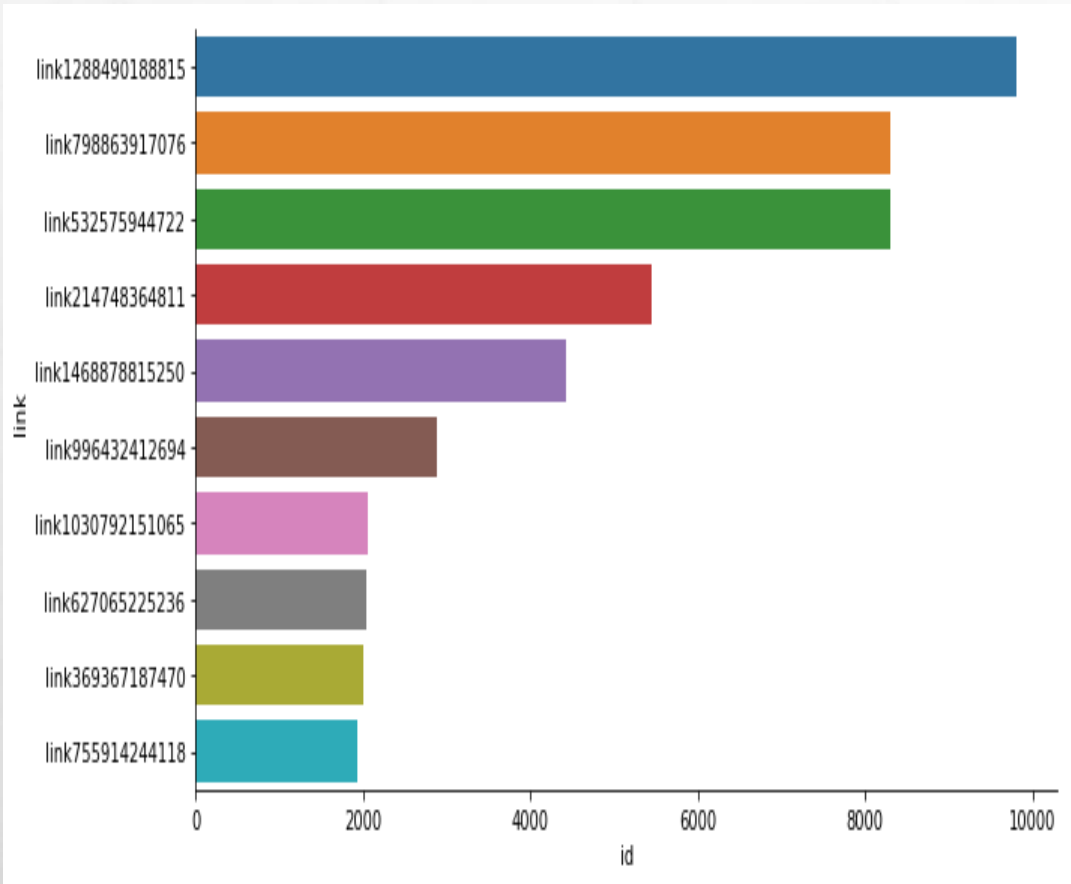
- **UPPER FUNNEL LEVEL IS 61%. SO MORE USERS ARE LEARNING ABOUT THE PRODUCT AND PRODCUT VERIFICATION.**
- **MEDIUM FUNNEL LEVEL IS 29%. THESE USERS SELECT PRODUCT AND ADD TO CART**
- **LOWER FUNNEL LEVEL IS 10%. SO VERY LESS USERS MAKE PAYMENT AND PROVIDE ADDRESS INFO.**

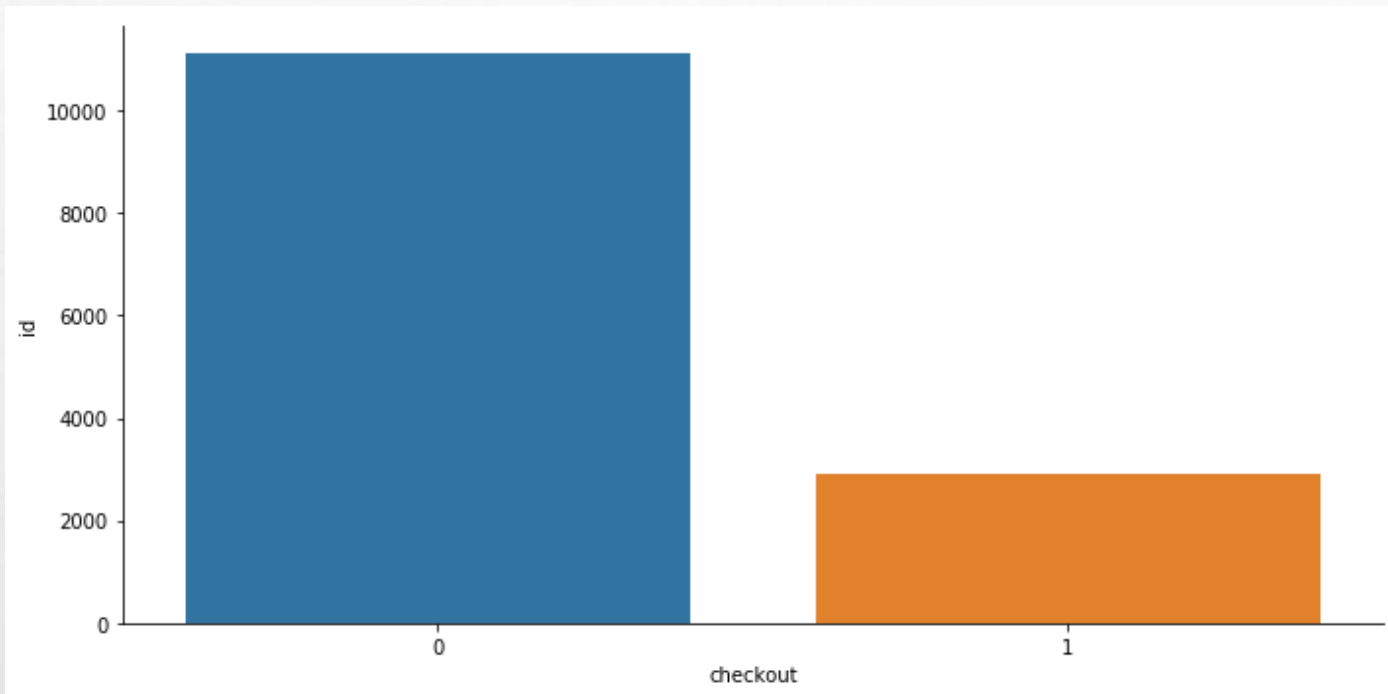


## EXPLORATORY DATA ANALYSIS

- **USERS ARE MORE ACTIVE IN AFTERNOON AND EVENING SESSION COMPARED TO NIGHT/MID NIGHT TIME.**
- **MORNING TIMINGS, THE USERS ARE BUSY AND NOT ACTIVELY SHOPPING FOR ITEMS.**

# GRAPHS FOR TOP 10 GROUPS AND LINKS





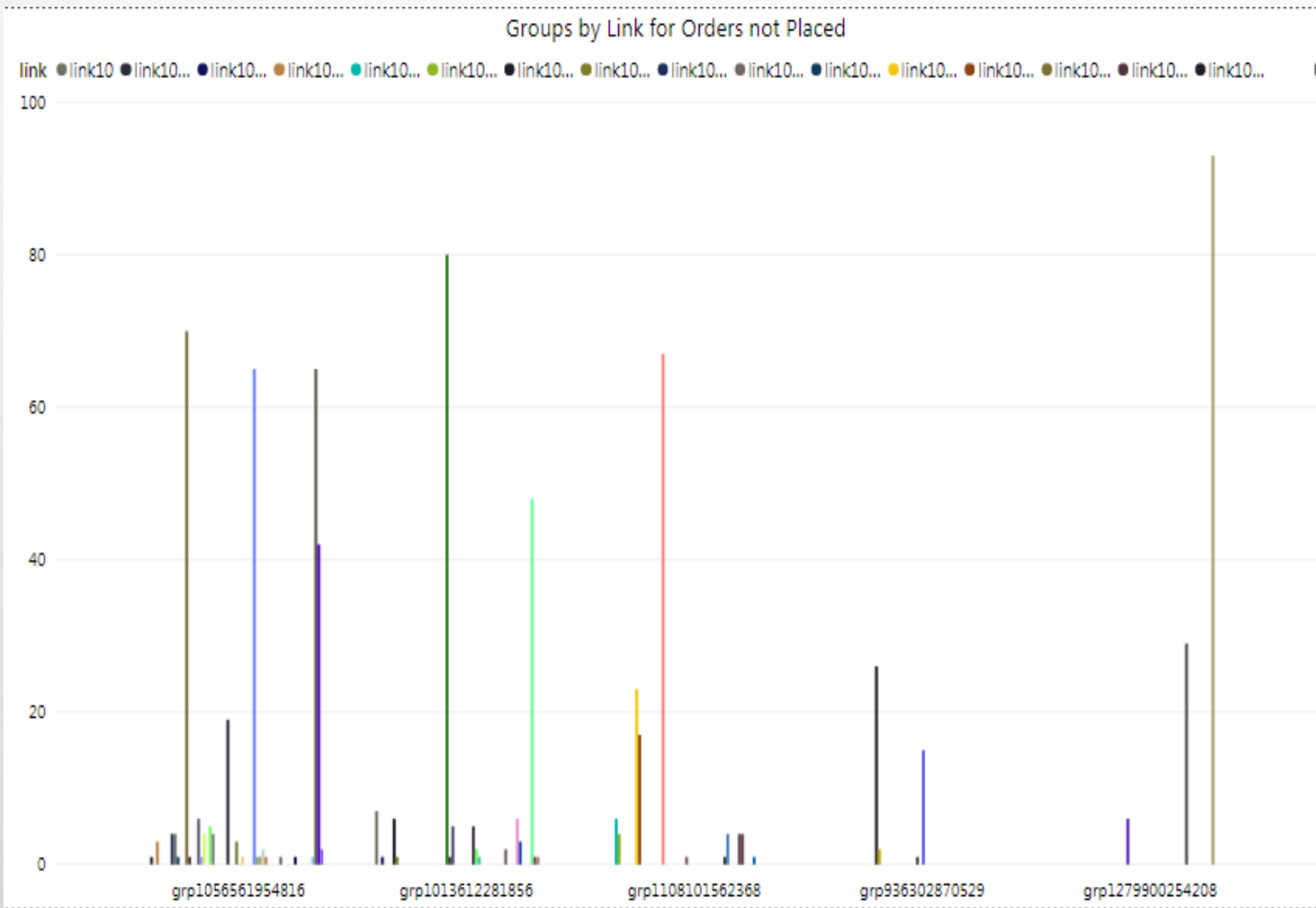
**79.38K**  
Adds not Clicked

**120.04K**  
Adds Clicked

## EXPLORATORY DATA ANALYSIS

- **MOST OF THE USERS ADD ITEMS TO CART BUT DO NOT PROCEED TO PURCHASE ITEMS.**
- **60% OF THE USERS ARE INTERESTED IN ADVERTISEMENT AND 40% ARE NOT.**



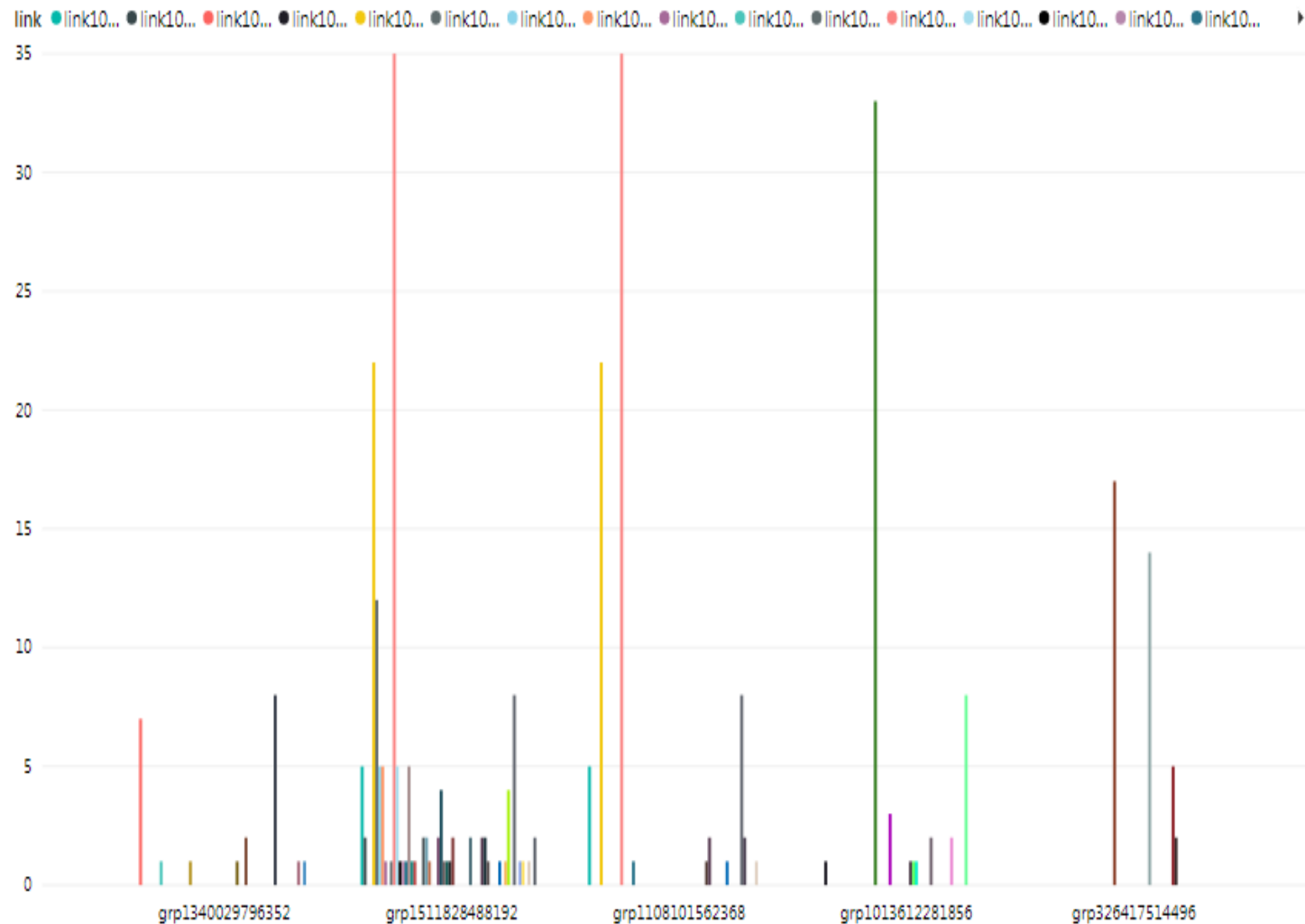


# EXPLORATORY DATA ANALYSIS

FOR USRS WHO DO NOT CONTINE PURCHASE (TOP 5 GROUPS ARE CONSIDERED WITH LINK COUNT):

THE USERS CLICK ON VARIOUS LINKS. THIS MEANS THEY ARE STILL NOT CONFIDENT ABOUT THE PURCHASE.

Groups by Links for Orders placed

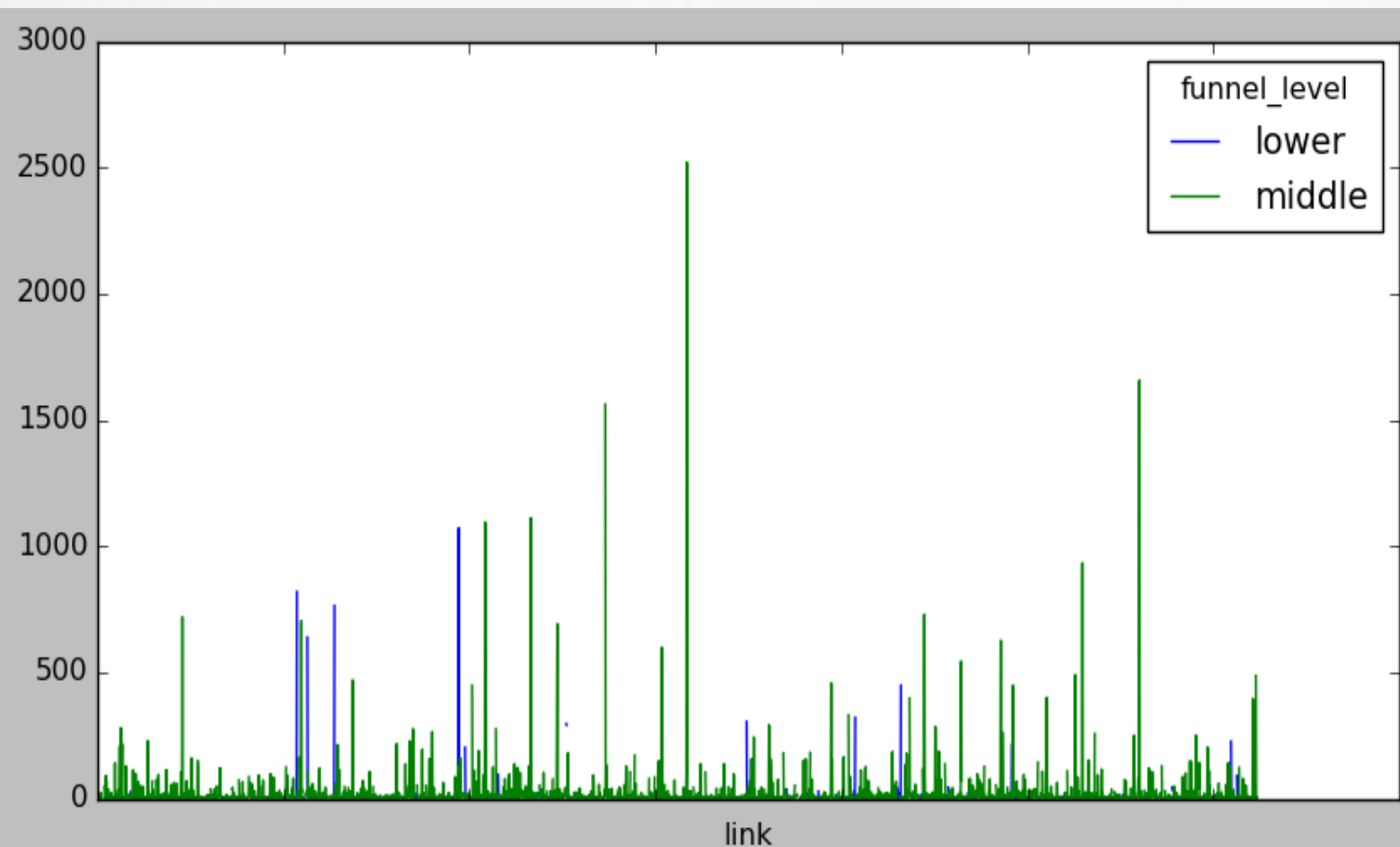


# EXPLORATORY DATA ANALYSIS

- **FOR USRS WHO COMPLETE PURCHASE (TOP 5 GROUPS ARE CONSIDERED WITH LINK COUNT):**

**THE USERS CLICK ON SPECIFIC LINKS – MAYBE THE INTERESTED ITEM AND THE PURCHASE IS COMPLETE.**





# EXPLORATORY DATA ANALYSIS

- **FROM THE GRAPH WE INFER THAT LOWER FUNNEL LEVEL USERS CLICK LINKS IS NOMINAL. THIS MEANS THE USERS ARE CONFIDENT ABOUT PURCHASING THE ITEMS ONLY.**
- **MIDDLE FUNNRL LEVEL USERS HAVE COMPARATIVELY MORE LINK CLICKS SO THEY SELECT THE PRODUCT BUT STILL NOT CONFIDENT ON PURCHASE.**

# DATA MODELLING

- **HYPOTHESIS:**

- **USERS ARE INTERESTED TO LEARN AND VERIFY THE PRODUCT BEFORE PURCHASING THE PRODUCT**
- **MORE USERS ARE INTERESTED IN ADVERTISEMENTS**
- **USERS ARE ACTIVE DURING AFTERNOON, EVENING TIME COMPARED TO MORNING TIME**
- **USERS WHO COMPLETE PURCHASE CLICK ON SPECIFIC LINKS MOST OF THE TIME(MAYBE PRODUCT DETAILS INTERESTED IN).**
- **USERS WHO DO NOT PURCHASE /DROP OUT CLICK ON RANDOM LINKS. THIS MEANS THEY ARE STILL EXPLORING AND NOT CONFIDENT OF PURCHASE TO BE MADE.**
- **MOST OF THE USERS ADD ITEMS TO CART BUT DO NOT PROCEED TO PURCHASE.**

# FEATURE ENGINEERING

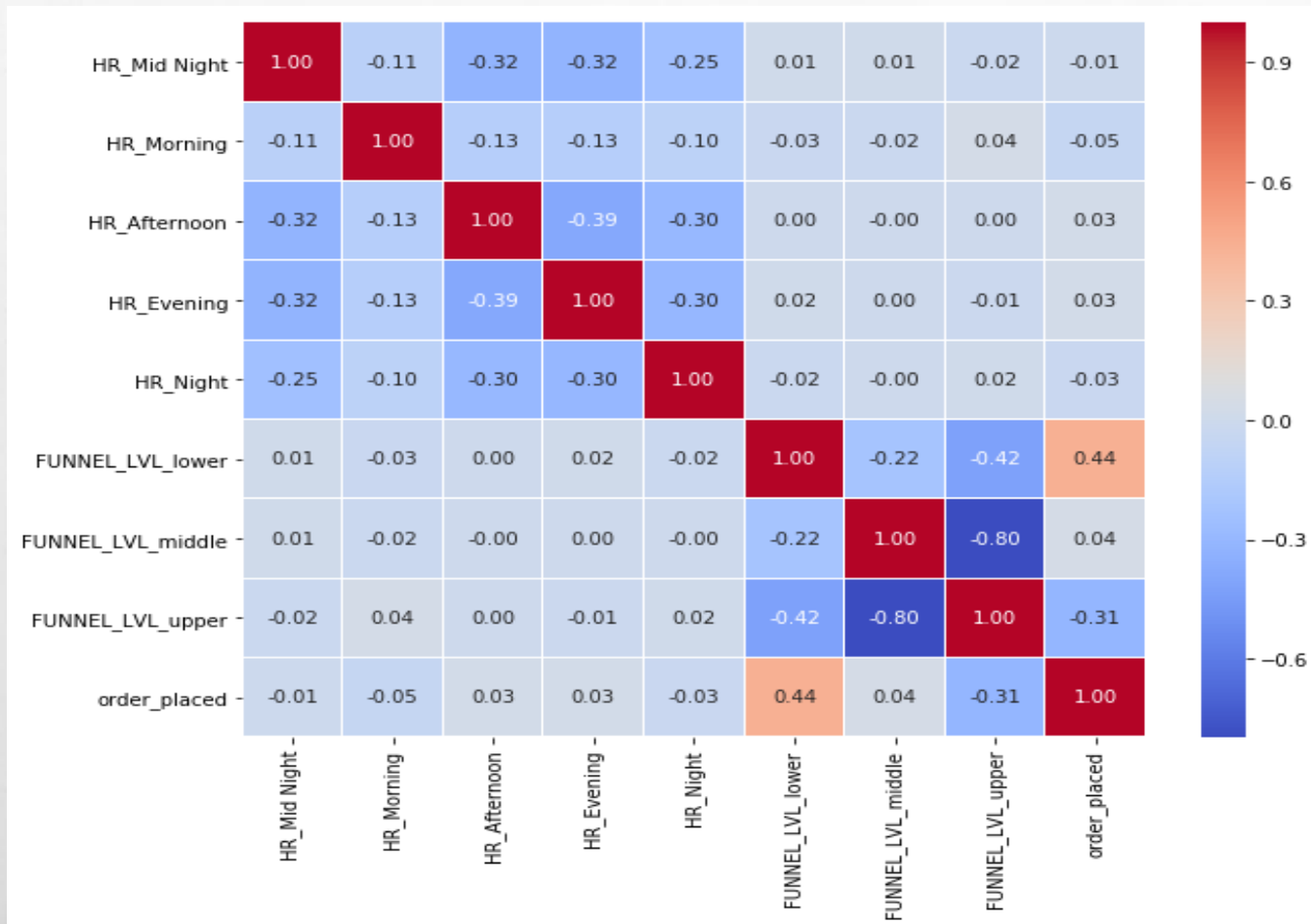
- **WITH THE DATA GIVEN, CONSIDERING THE TIMESTAMP ALL THE PURCHASE DATA IS GIVEN FOR THE SAME DATE – 30<sup>TH</sup> APRIL 2018. SO HOURS IS CONSIDERED AND ADDITIONAL FEATURES MORNING, AFTERNOON, EVENING, NIGHT, MIDNIGHT IS ADDED.**
- **FUNNEL TYPE PLAYS A MAJOR ROLE SO HAS USED THIS CATEGORICAL DATA AS FEATURES**
- **SESSION ID, GRP ID, AD ARE DROPPED AS THIS DOESN'T PLAY SIGNIFICANT ROLE AS FEATURES.**
- **THE GIVEN DATA HAS BEEN PREPROCESSED (SCALED) AND PCA IS USED FOR DIMENSIONALITY REDUCTION**

# FEATURE ENGINEERING (TOP VALUES)

COLUMN	COUNT	UNIQUE	TOP	FREQ
• ID	199410	14014	SESSION51539607562	235
• AD	120035	10388	AD369367187493	229
• LINK	199410	5248	LINK798863917076	15485
• GRP	199410	108	GRP1340029796352	93084
• FUNNEL LEVEL	199410	3	UPPER	120772

- AD HAS NULL VALUES, THIS MEANS USERS ARE NOT VIEWING ADVERTISEMENTS. HENCE IT IS PART OF DATA AND NO MISSING VALUE DATA CONCEPT IS APPLIED ON THIS COLUMN.

# CORRELATION HEAT MAP



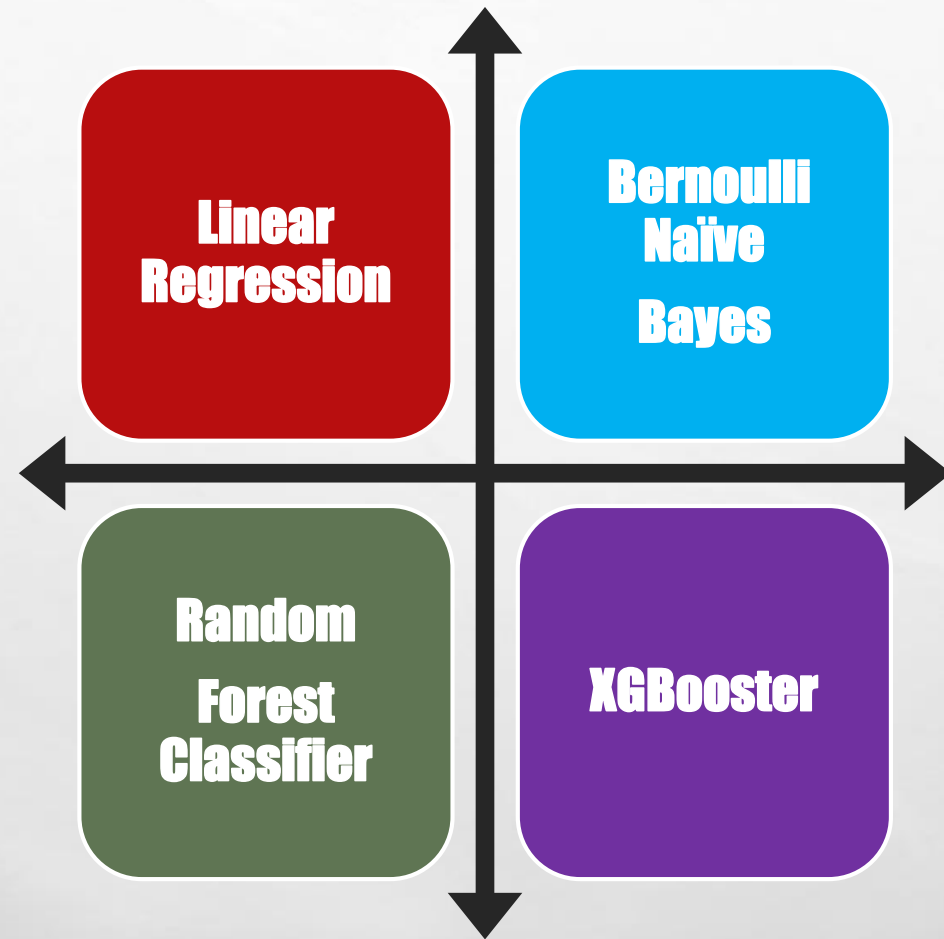
- **WE INFER THAT FUNNEL\_LVL\_MIDDLE AND FUNNEL\_LVL\_UPPER ARE HIGHLY NEGATIVELY CORRELATED, HOWEVER WE DECIDE TO RETAIN THESE COLUMNS AS THEY ARE SIGNIFICANT IN FEATURE ENGINEERING.**

# FEATURE ENGINEERING

- **WITH THE DATA GIVEN, CONSIDERING THE TIMESTAMP ALL THE PURCHASE DATA IS GIVEN FOR THE SAME DATE – 30<sup>TH</sup> APRIL 2018. SO HOURS IS CONSIDERED AND ADDITIONAL FEATURES MORNING, AFTERNOON, EVENING, NIGHT, MIDNIGHT IS ADDED.**
- **FUNNEL TYPE PLAYS A MAJOR ROLE SO HAS USED THIS CATEGORICAL DATA AS FEATURES**
- **SESSION ID, GRP ID, AD ARE DROPPED AS THIS DOESN'T PLAY SIGNIFICANT ROLE AS FEATURES.**
- **FROM CORRELATION HEATMAP THE FUNNEL TYPE DATA “MIDDLE” IS RETAINED AS IT IS USED AS ONE OF THE SIGNIFICANT FEATURE.**



# CLASSIFICATION MODELS

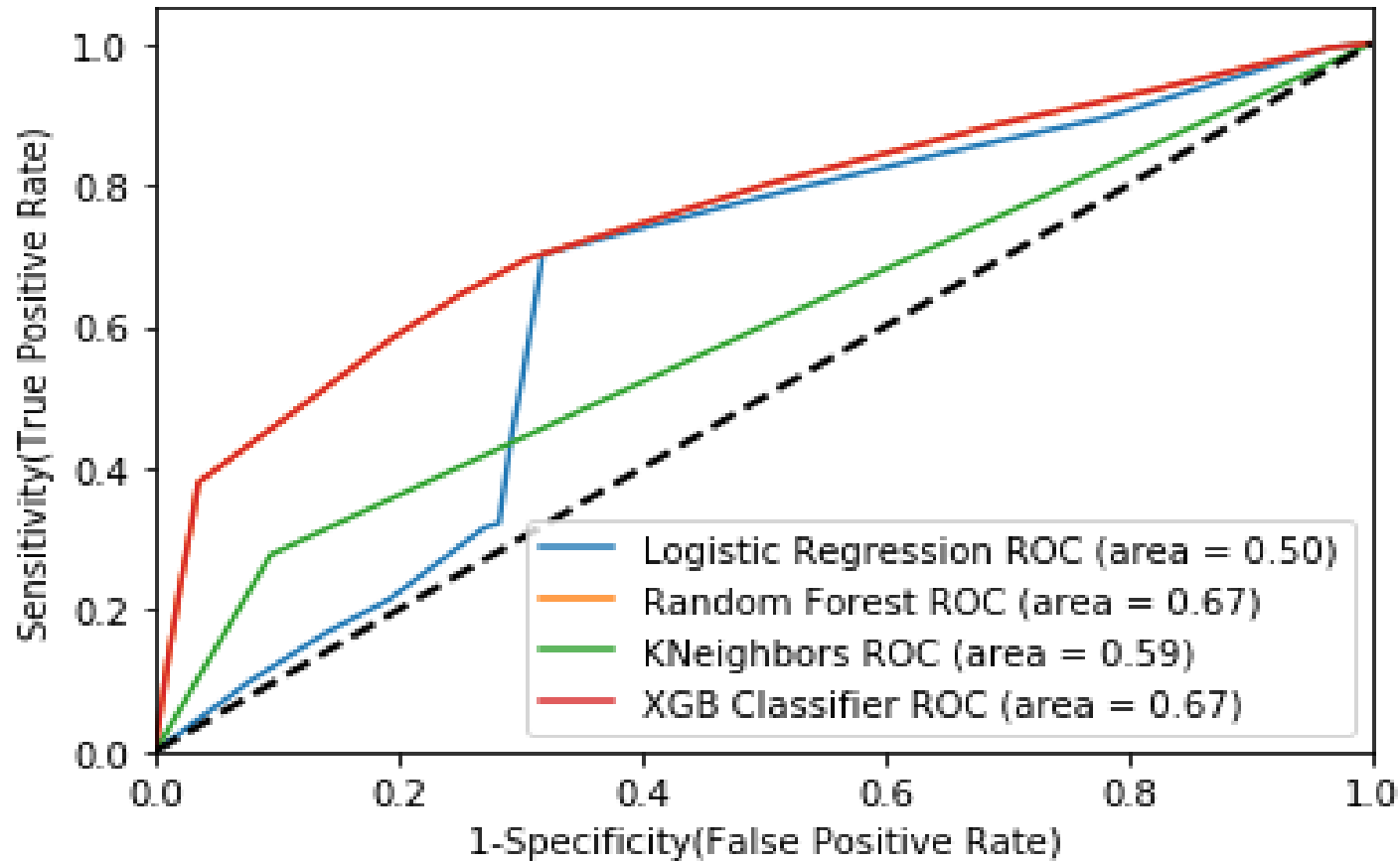


# CLASSIFICATION METRICS

ALGORITHM	LOG LOSS	ACCURACY	F1 SCORE	PRECISSION	RECALL
BERNOULLI NAÏVE BAYES	0.4259	79.9408	0.7102	0.6390	0.7994
RANDOM FOREST Classifier	0.4113	84.7124	0.8274	0.8343	0.8471
XGBOOST Classifier	0.4113	84.7124	0.8274	0.8343	0.8471


# ROC CURVE GRAPH

Receiver Operating Characteristic



- **FROM MULTIPLE ALGORITHM ROC CURVE GRAPH INTERPRETATION, IT IS FOUND THAT XGB CLASSIFIER BEST SUITS THIS PREDICTION.**

# RESULTS

- **COMPARING THE LOG LOSS, ACCURACY VALUES “XGBOOSTER” SUITS BEST FOR THIS SOLUTION.**
- **HENCE USED XGBOOSTER ALGORITHM AND TRAINED THE TEST DATASET .THE OUTPUT IS WRITTEN TO SUBMISSION.CSV FILE**  
  
Microsoft Excel  
ma Separated Val
- **THE HYPOTHESIS PROVED ARE:**
  - **MAJORITY OF USERS ARE INTERESTED IN THE ADVERTISEMENT RATHER THAN PRODUCT SHOPPING.**
  - **LOWER FUNNEL USERS ARE CONFIDENT ON THE PRODUCTS PURCHASE AND HAS MINIMAL CLICKS ON INTERACTION LINKS.**
  - **MIDDLE/LOWER FUNNEL USERS USE VARIOUS INTERACTION LINKS TO VERIFY OTHER PRODUCTS AVAILABLE. PROVIDING PRODUCT COMPARISON WILL HELP USERS CHOOSE THE PRODUCT QUICKLY.**
  - **MAJORITY USERS DROP OUT AFTER LEARNING/VERIFYING THE PRODUCT. SO BETTER PRODUCT DESCRIPTION/REVIEWS WILL HELP THEM TO FINALIZE THE PRODUCT.**