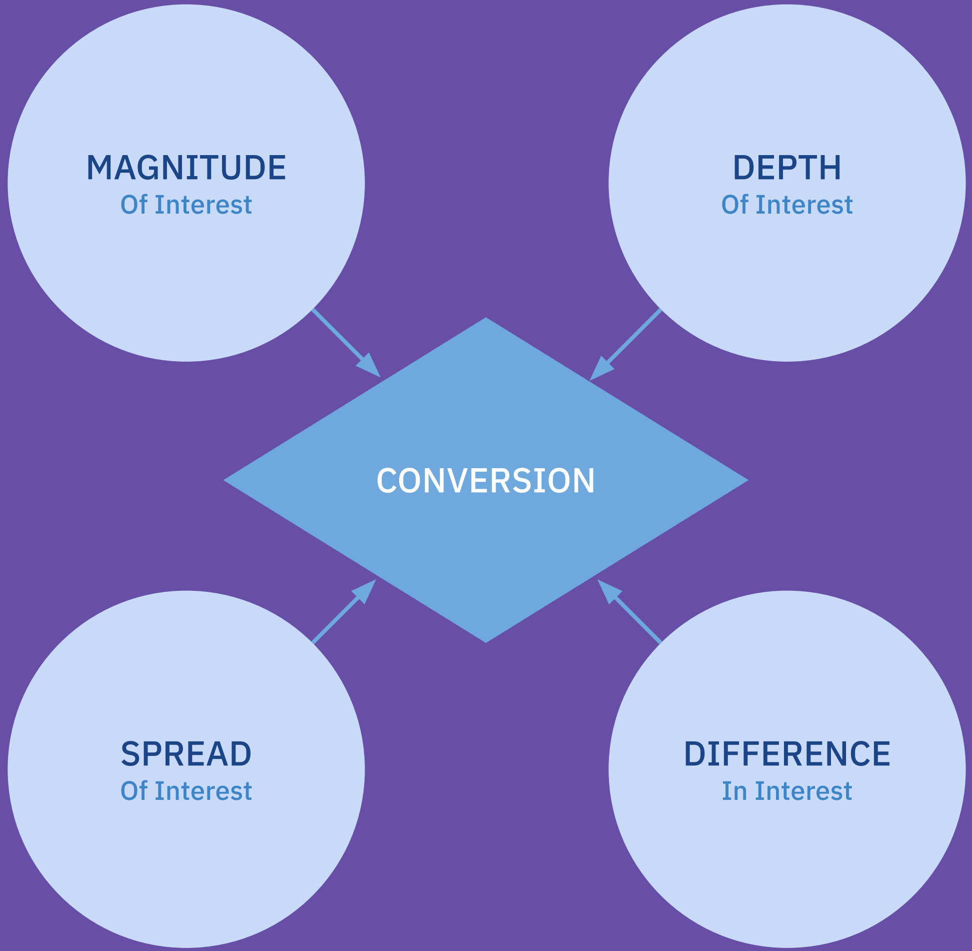
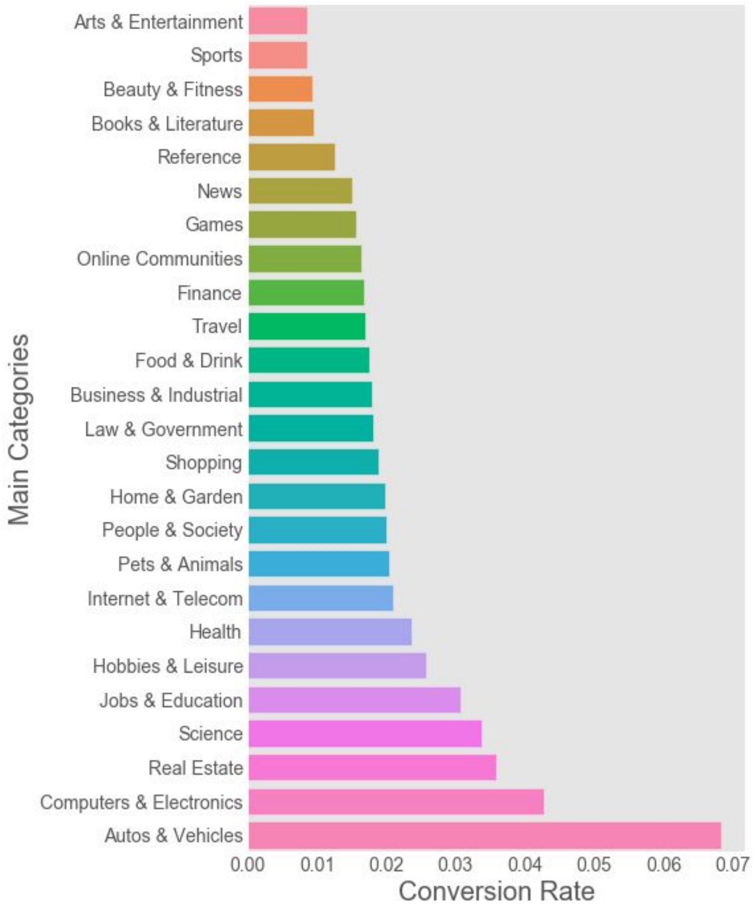
Modeling Conversion Using Hierarchical Interest Metrics

By
Robyn Kwok, Samuel Appiah-Kubi, Bassim Eledath



Exploratory Data Analysis

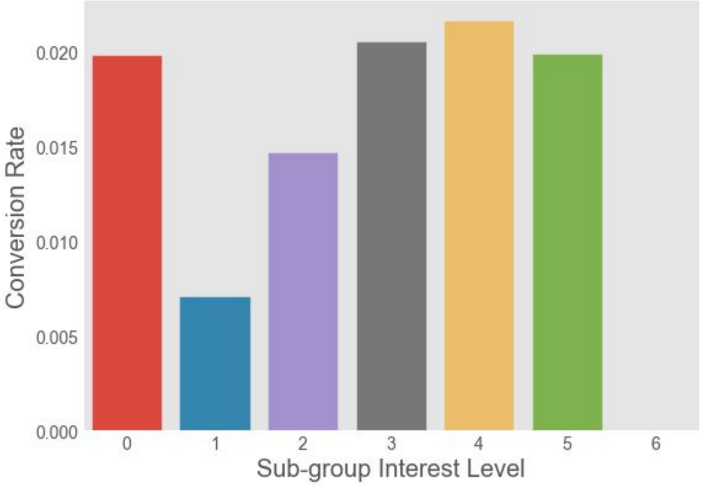


Engagement score: weak indicator of conversion

- No correlation between categories with high engagement scores and categories with the most conversion

Depth of main interest is significant

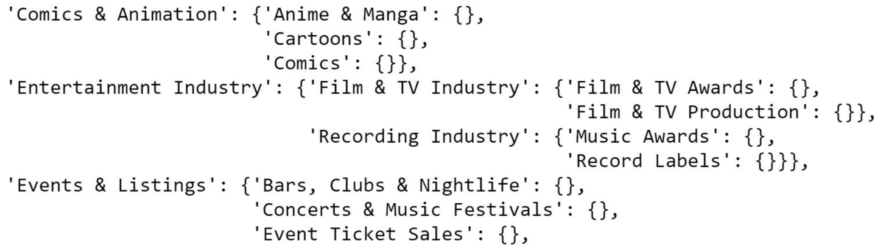
- Users with very general (depth = 0) or very niche (2 < depth < 6) main interest had the highest conversion rates.



Data Processing

Aggregation

- Interest data is hierarchical in nature:
  - Aggregate to avoid curse of dimensionality
- Construct metrics to:
  - Account for loss of information after aggregation
  - Make inferences on the relationship between user interests and conversion



```
'Comics & Animation': {'Anime & Manga': {},
                       'Cartoons': {},
                       'Comics': {}},
'Entertainment Industry': {'Film & TV Industry': {'Film & TV Awards': {},
                                                  'Film & TV Production': {}},
                           'Recording Industry': {'Music Awards': {},
                                                  'Record Labels': {}}},
'Events & Listings': {'Bars, Clubs & Nightlife': {},
                      'Concerts & Music Festivals': {},
                      'Event Ticket Sales': {},
```

***Figure**: Hierarchical nature of interest data*

Objectives

Main objective: <u>Inference</u>, not prediction

Key Questions Answered:

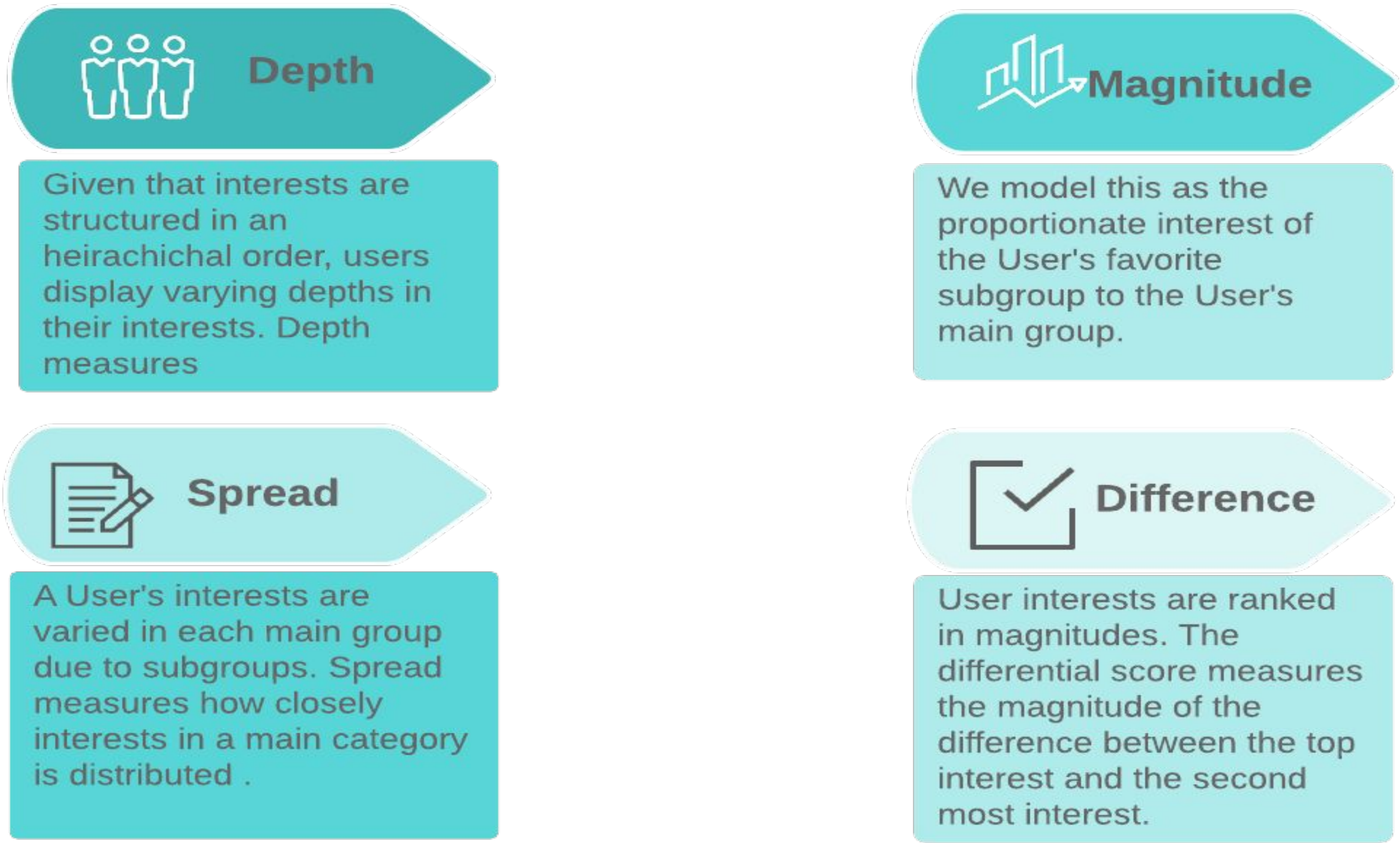
1. **Magnitude of user interest**
   Does a higher interest in favorite topic signify higher conversion rate?

2. **Depth of user interest**
   Do users with more niche interests have higher conversion rates?

3. **Spread of user interest**
   Are users who have multiple interests more likely to convert?

4. **Preferential difference in interest**
   How strongly in favor is the user for their main interest?

Feature Engineering - Constructing Metrics



Depth

Given that interests are structured in an heirachichal order, users display varying depths in their interests. Depth measures

Magnitude

We model this as the proportionate interest of the User's favorite subgroup to the User's main group.

Spread

A User's interests are varied in each main group due to subgroups. Spread measures how closely interests in a main category is distributed .

Difference

User interests are ranked in magnitudes. The differential score measures the magnitude of the difference between the top interest and the second most interest.

Results and Recommendations

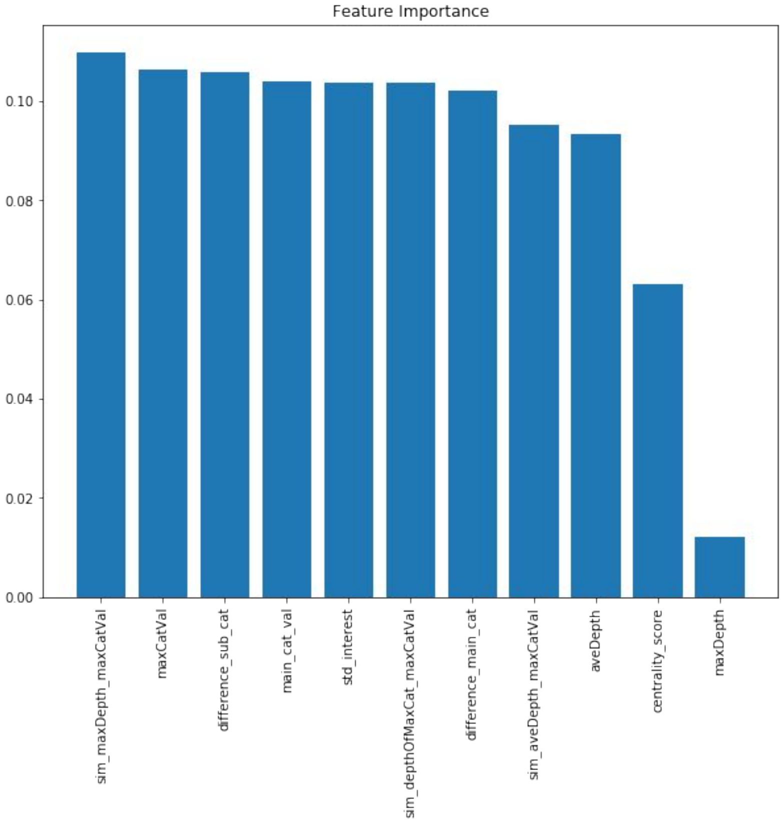
Results

- **Model fit**: Both our logistic and random forest classifiers fit the data poorly (AUC ~ 0.5)
- **ANOVA** (p-val = 0.05):
  - Significant relationships:
    - **Negative**: Average depth, max proportionate interest (sub-group)
    - **Positive**: Differential interest, max proportionate interest (main group)
  - Insignificant relationships:
    - Centrality, max depth, deviation of interest
- **Correlation analysis**: Inconclusive, given proportionate interest does not signify anything about frequency.

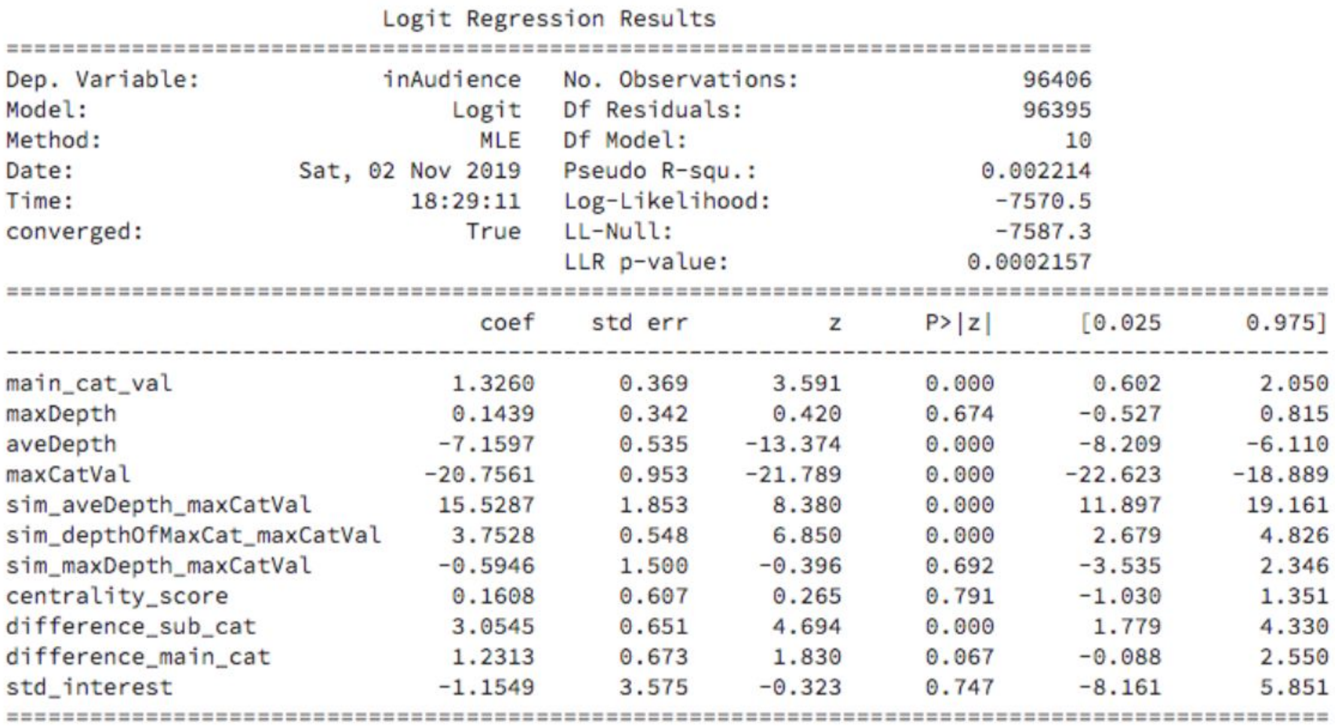
Recommendations

- **Use frequency metrics**: In conjunction with interest metrics, frequency metrics will enable us to separate active vs inactive users, as this is a potential confounding factor. This will improve inference.
- **Complex models**: As the nature of data is hierarchical (high cardinality), for prediction, we recommend deep learning models. To correct for class imbalance we recommend algorithms such as SMOTE.
- **Other metrics to consider:** Besides frequency, other metrics such as device used, duration spent etc. can work well with user interests in modeling conversion.

Appendix - Random Forest Feature Importances



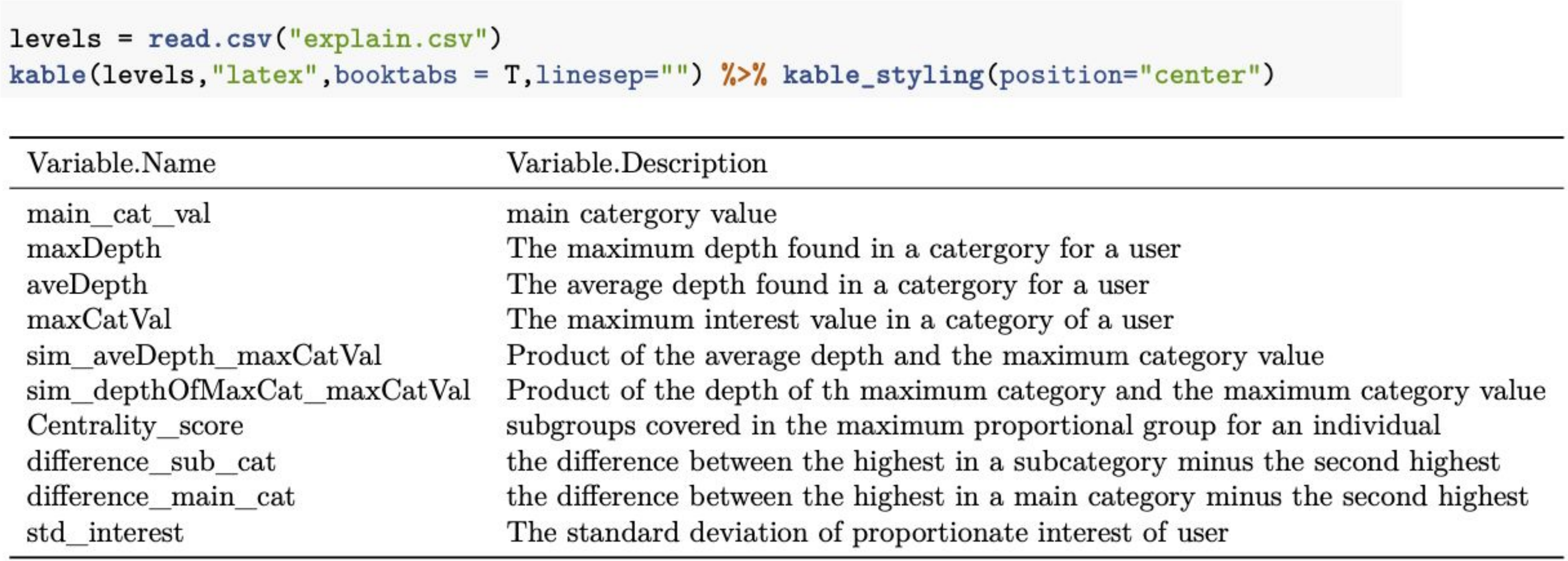
Appendix - Logistic Regression Results



```
                           Logit Regression Results
==============================================================================
Dep. Variable:             inAudience   No. Observations:                96406
Model:                          Logit   Df Residuals:                    96395
Method:                           MLE   Df Model:                           10
Date:                Sat, 02 Nov 2019   Pseudo R-squ.:                0.002214
Time:                        18:29:11   Log-Likelihood:                -7570.5
converged:                       True   LL-Null:                       -7587.3
                                        LLR p-value:                 0.0002157
===============================================================================================
                                  coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------

main_cat_val                    1.3260      0.369      3.591      0.000       0.602       2.050
maxDepth                        0.1439      0.342      0.420      0.674      -0.527       0.815
aveDepth                       -7.1597      0.535    -13.374      0.000      -8.209      -6.110
maxCatVal                     -20.7561      0.953    -21.789      0.000     -22.623     -18.889
sim_aveDepth_maxCatVal         15.5287      1.853      8.380      0.000      11.897      19.161
sim_depthOfMaxCat_maxCatVal     3.7528      0.548      6.850      0.000       2.679       4.826
sim_maxDepth_maxCatVal         -0.5946      1.500     -0.396      0.692      -3.535       2.346
centrality_score                0.1608      0.607      0.265      0.791      -1.030       1.351
difference_sub_cat              3.0545      0.651      4.694      0.000       1.779       4.330
difference_main_cat             1.2313      0.673      1.830      0.067      -0.088       2.550
std_interest                   -1.1549      3.575     -0.323      0.747      -8.161       5.851
===============================================================================================
```

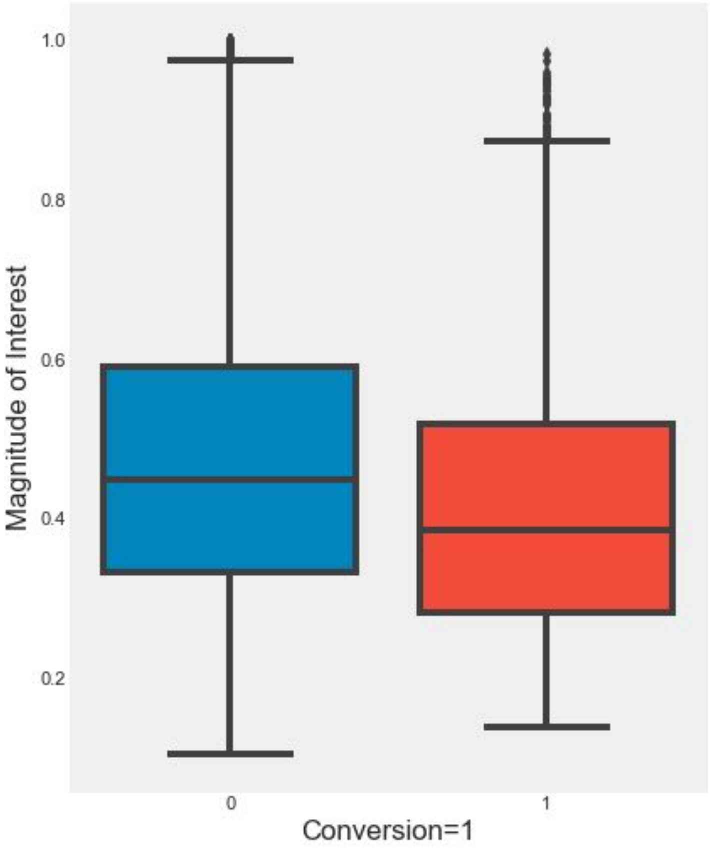
Appendix - Data Dictionary



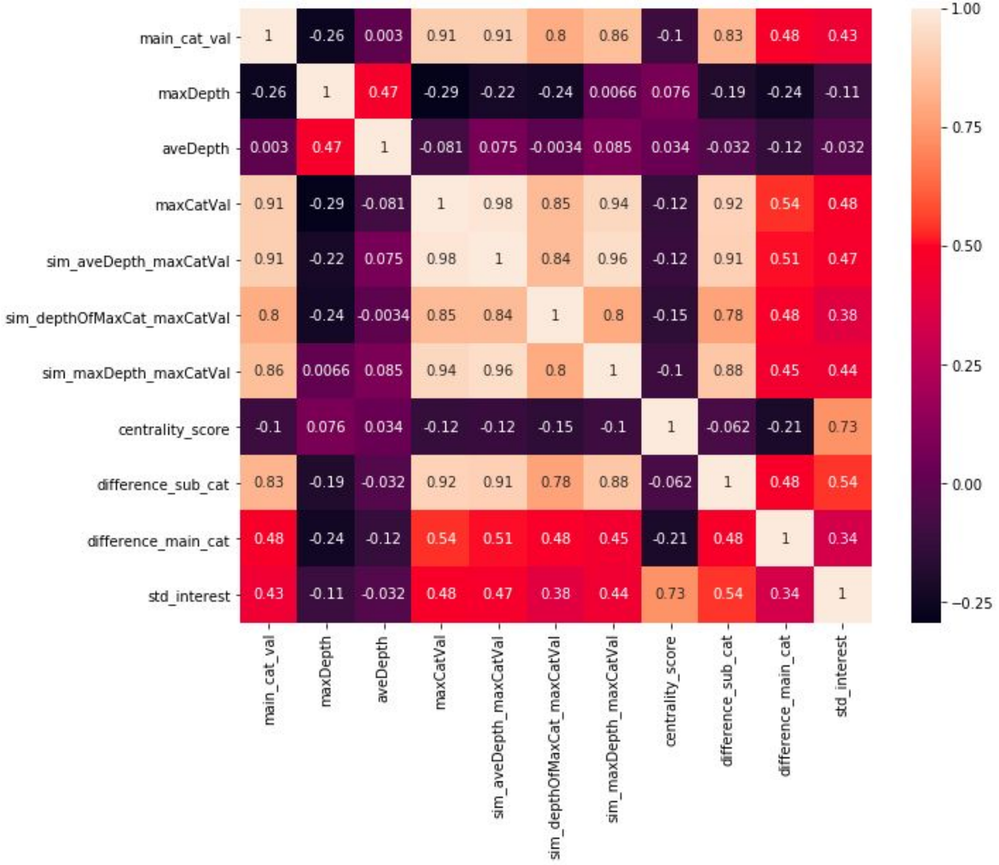
```
levels = read.csv("explain.csv")
kable(levels,"latex",booktabs = T,linesep="") %>% kable_styling(position="center")
```

Variable.Name	Variable.Description
main_cat_val	main catergory value
maxDepth	The maximum depth found in a catergory for a user
aveDepth	The average depth found in a catergory for a user
maxCatVal	The maximum interest value in a category of a user
sim_aveDepth_maxCatVal	Product of the average depth and the maximum category value
sim_depthOfMaxCat_maxCatVal	Product of the depth of th maximum category and the maximum category value
Centrality_score	subgroups covered in the maximum proportional group for an individual
difference_sub_cat	the difference between the highest in a subcategory minus the second highest
difference_main_cat	the difference between the highest in a main category minus the second highest
std_interest	The standard deviation of proportionate interest of user

Appendix - Engagement Scores by Conversion Outcome



Appendix - Correlation Analysis



	main_cat_val	maxDepth	aveDepth	maxCatVal	sim_aveDepth_maxCatVal	sim_depthOfMaxCat_maxCatVal	sim_maxDepth_maxCatVal	centrality_score	difference_sub_cat	difference_main_cat	std_interest
main_cat_val	1	-0.26	0.003	0.91	0.91	0.8	0.86	-0.1	0.83	0.48	0.43
maxDepth	-0.26	1	0.47	-0.29	-0.22	-0.24	0.0066	0.076	-0.19	-0.24	-0.11
aveDepth	0.003	0.47	1	-0.081	0.075	-0.0034	0.085	0.034	-0.032	-0.12	-0.032
maxCatVal	0.91	-0.29	-0.081	1	0.98	0.85	0.94	-0.12	0.92	0.54	0.48
sim_aveDepth_maxCatVal	0.91	-0.22	0.075	0.98	1	0.84	0.96	-0.12	0.91	0.51	0.47
sim_depthOfMaxCat_maxCatVal	0.8	-0.24	-0.0034	0.85	0.84	1	0.8	-0.15	0.78	0.48	0.38
sim_maxDepth_maxCatVal	0.86	0.0066	0.085	0.94	0.96	0.8	1	-0.1	0.88	0.45	0.44
centrality_score	-0.1	0.076	0.034	-0.12	-0.12	-0.15	-0.1	1	-0.062	-0.21	0.73
difference_sub_cat	0.83	-0.19	-0.032	0.92	0.91	0.78	0.88	-0.062	1	0.48	0.54
difference_main_cat	0.48	-0.24	-0.12	0.54	0.51	0.48	0.45	-0.21	0.48	1	0.34
std_interest	0.43	-0.11	-0.032	0.48	0.47	0.38	0.44	0.73	0.54	0.34	1