Modeling Conversion Rate Favorability With Various Dimensions of User Topic Interest

Abstract

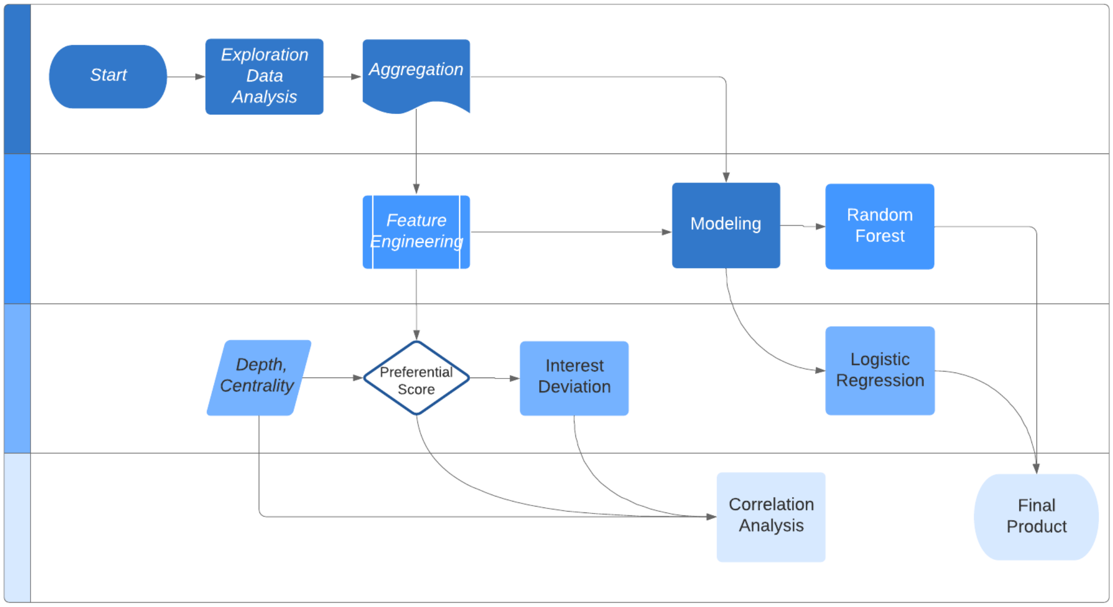
Conversion Rate Optimization is a technically challenging field due to the breadth of measurable and immeasurable influences driving user behavior. For example, a user going through an online blog may be influenced to click an ad for reasons that are overt (such as time spent on web page) or hidden (such as age of user, that is not logged). This paper aims to measure the influences pertaining to user preferences (by topic of interest). We explore various ways of measuring the ground truth interest (inestimable in theory) of the user via data manipulation and model estimation. In sum, the dimensions we address and questions we answer in this paper are as follows:

1. Magnitude of user interest: Does a higher interest in favorite topic signify a 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 his/her biggest interest, how much bigger in magnitude is it from his/her second biggest interest, and does this influence conversion rate.

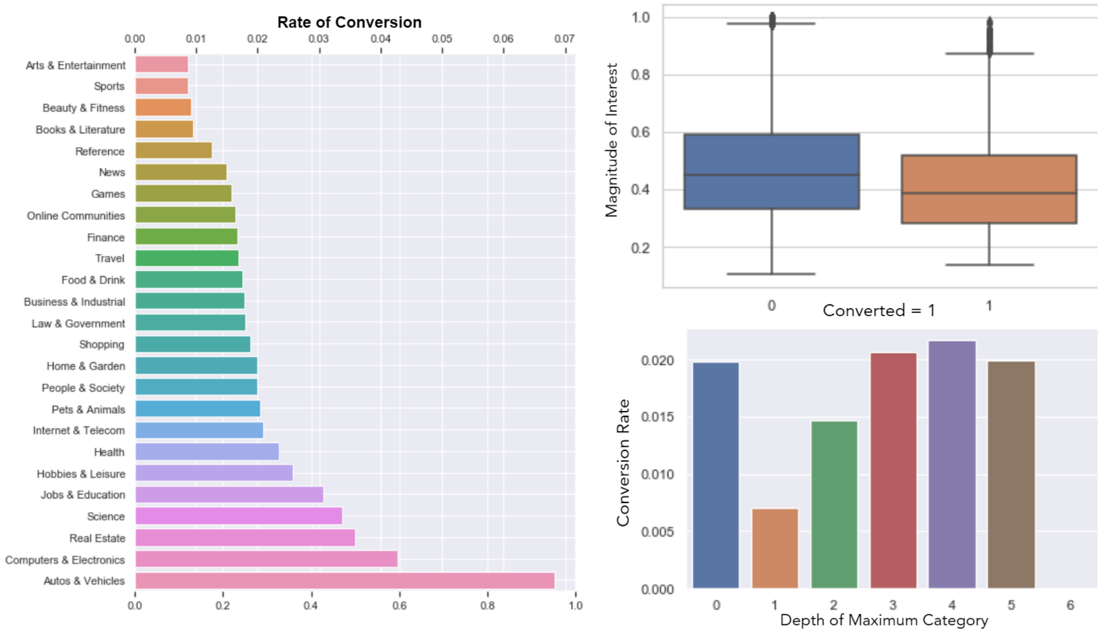
1. Introduction

Overview

At a high level our analysis pipeline is as follows:



Exploratory Data Analysis (EDA)



1. **Left:** We first mapped the rate of conversion per main category, the highest level of long term interests. The reason we discounted short term interests is because only ~20,000 rows contained non-empty short term features. Moreover, the conversion rate indicates if a user has converted in the past with no indication of time. Hence it is hard to reason the impact of short-term interests on conversion.
2. **Right-upper:** Engagement scores for users' main interest groups are significantly lower for successful conversations. When running logistic regression, we found a negative coefficient (-1.0601) between the conversion rate and average engagement score. We began to suspect that higher engagement scores on a user's main interest group could suggest that a user's overall engagement is lower. Noting also that engagement scores to 1.0, high engagement scores on one category does not imply high frequency, and may indicate lower frequency since a user may be engaging with fewer categories overall.
3. **Right-lower:** users with very general (depth=0) or very niche (depth 3-5) most-engaged categories had the highest conversion rates.

2. Data

Data Aggregation

When we did our EDA, we realized the interests of users had high cardinality i.e. there are more than 1400 interest groups. For example, the Computers and Finance interest groups had 6 sub-groups each within them, which meant our model fitting process would suffer unless some dimensionality reduction took place. We thus decided to aggregate by main interest groups.

Feature Engineering

Our feature engineering process was driven by the main questions we proposed at the start of our analysis. We first mathematically formulate them and proceed with explaining them in more detail.

Mathematical Formulation

1. Depth, $d(v)$:

$$d(v) = \begin{cases} 1 & \text{topic, v} \in \text{V, has no sub-label} \\ \sum_{v \in V} d(v_{sub}) & \text{topic, v} \in \text{V, has sub-label} \end{cases}$$

2. Centrality, c(i):

$$\frac{\text{subgroups covered in } maxproportionategroup_{individual}}{\text{total subgroups in } maxproportionate_{subgroup}}$$

3. Magnitude of Interest, m(i), the maximum proportion of interest

$$m(i) = max(ProportionalInterest_{individual})$$

4. Interest Deviation, $\omega(i)$

$$\text{Given that s} = \sqrt{\frac{\sum_{v \in V}(v_1 - \overline{v})^2}{N-1}} \omega(i) = s \forall \text{m of User, u}$$

5. Differential Score, s(i)

$$s(i) = max_1(i) - max_2(i)$$

Description

1. **Depth:** different users have varying depths in their interests. For example, some users have niche interests that are deeply nested under more popular interest groups - fewer people are interested in the 'Gossip' sub-interest than the parent 'News' interest. We add the depth of the user's biggest interest (sub group level) as a feature.
2. **Centrality:** some users expose themselves to multiple sub-groups in their main interest group. For example, a user whose main interest in the parent-group computers may be interested in multiple sub-groups such as software, hardware etc. which may influence conversion.
3. **Magnitude:** A user's magnitude of interest in his/her main interest may drive conversion and we model this as the user's proportionate interest in his/her favorite sub-group and main-group (separately)
4. **Deviation:** We model the deviation of users' interests as the standard deviation in proportionate interests. We hypothesize a user with more varied interests (and therefore less proportionate standard deviation) is less likely to convert as opposed to a more polar user.
5. **Differential score:** An extension of magnitude (a marginal estimate rather) which measures the difference in proportionate interest between the highest and second highest interest of the user. This score is calculated both at the aggregate and sub-group level for each user.

3. Modeling

With our data finally processed, we fit two binary classification models - a logistic regression model and a random forest classifier. For both the logistic regression and random forest classifier, the performance is poor with an AUC of 0.5. The coefficient estimates and feature importance graphs for the logistic model and random forest classifier are provided in the Appendix section.

4. Conclusion, Recommendations and Learnings

Based on our analysis, we make the following observations and recommendations:

1. There isn't enough signal in the proportionate interest data: Our models show a poor AUC, which is indicative of a poor model fit. Although our ANOVA shows some variables to be statistically significant, their effect is marginal on conversion rate. We also performed a two sample t-test on both the converted and non-converted group (not included in report) and the differences were insignificant (t-test was done based on propensity scores of both samples).
2. **ANOVA results:** Directionally, based on ANOVA, we see between the statistically significant features (<0.05), the one with the largest magnitude coefficient is maxCatVal, which has a negative coefficient of -20.4. This is the engagement score of the user's most-engaged interest. We see that a lower engagement score correlates to a higher conversion rate. This corroborates our initial hypothesis that high engagement scores on the single most-engaged interest indicates low overall engagement.
3. **Inclusion of interest frequency:** We believe the data on interest should better reflect the recency and frequency of interest rather than just magnitude and depth. High engagement could manifest in low magnitude scores for a single interest since the user could have engaged with a large array of interests, hence dividing their attention to individual subcategories. Frequency, we believe, would be a better signal of conversion, and this cannot be derived from proportionate interest.
4. **Future steps:** We may attempt to implement future models using a more balanced class. Algorithms like Synthetic Minority Over-sampling Technique (SMOTE) can be used to balance the data classes, however this will not compensate for the poor signal in the predictor variables. We would also like to include other dimensional data, such as geo-location of the user, to better predict conversion.

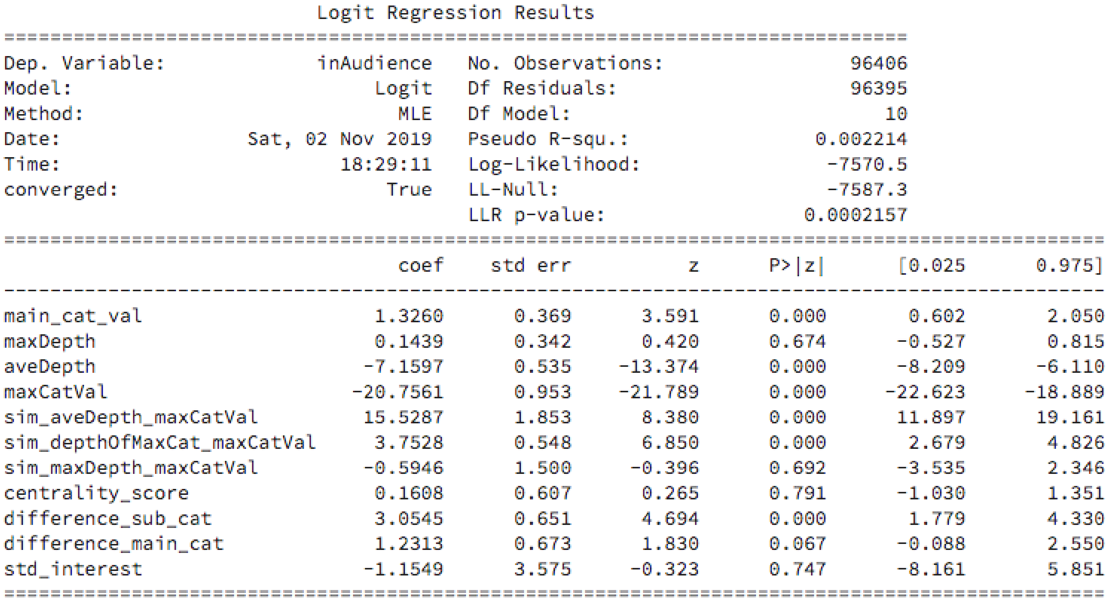
5. Appendix

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

Logistic Model 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
===============================================================================================
```