Understanding and Improving Conversion Strategies at Granify

This report is organized into the following sections. First, an exploratory analysis is performed on the feature to identify typical usage patterns that emerge. The performance of various conversion strategies is measured for each of the usage patterns identified earlier. Finally, in conclusion, a summary of all findings and future research scope is provided.

Exploring the Feature Space

The distribution of the features shared in the data set is presented in fig. 1. The minmax normalization technique is chosen to scale the features given. The motivation for choosing minmax comes from the fact the features are bounded and do not follow any specific distribution.

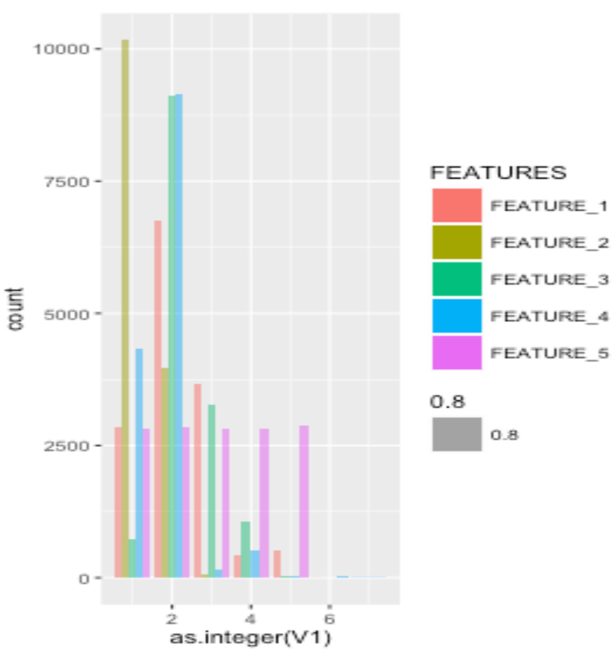


Fig.1 Distribution of Features

Table 1 gives the correlation matrix of the scaled features. From the table it is clear that Features 1,5 are uncorrelated with any of the other features. Features 2,3,4 are correlated amongst themselves.

	F1	F2	F3	F4	F5
F1	1.00	-0.04	0.00	0.04	-0.01
F2	-0.04	1.00	-0.15	0.09	-0.01
F3	0.00	-0.15	1.00	-0.13	-0.01
F4	0.04	0.09	-0.13	1.00	0.01
F5	-0.01	-0.01	-0.01	0.01	1.00

Table 1: Correlation Matrix of Scaled Features

Features 1 and 5 contribute to ~67% of the total variance of the feature set. To get a complete overview of a particular session usage knowledge of feature 1 and 5 is quintessential. Granular information from features 2,3,4 is not required and an aggregated segmented view of the same should suffice. A simple kmeans clustering on features 2,3,4 for different values of k was performed and the within sum squared (WSS) distance was measured in each case. The knee point of the WSS graph was found to be at k=4. With this choice of k, we get the following 4 clusters given in table 2.

	Size	Annotation
C1	9764	Low_F2 + MID_F3 + LOW_F4
C2	3289	HIGH_F2 + MID_F3 + MID_F4
C3	644	HIGH_F2 + HIGH_F3 + LOW_F4
C4	505	Low_F2 + HIGH_F3 + HIGH_F4

Table 2: Cluster results. (Centers shared in the appendix)

In the next section we describe how each of the conversion strategies perform for various values of Feature 5, Feature 1 and cluster membership.

Conversion Strategy Performance

Delta in the conversion percentage is chosen as the performance measurement criteria. A chi square independence test shows that the responses for CG and TG have different distributions. The impact is given in table 3.

Strategy	Coverage	CG Conv. (%)	TG Conv. (%)	Delta (%)
AD 1	22.41%	9.23	10.63	1.43
AD 2	15.30%	4.72	5.15	0.43
AD 3	62.27%	0.99	0.98	-0.01
Overall	100.0%	3.41	3.65	0.24

Table 3: Strategy Performance

The AD1 is the most successful campaign while AD3 is least successful. Fig 2, 3, 4 describe the performance of TG (**Red**) vs CG (**Blue**) for AD 1 for various values of features 1,5 and cluster memberships.

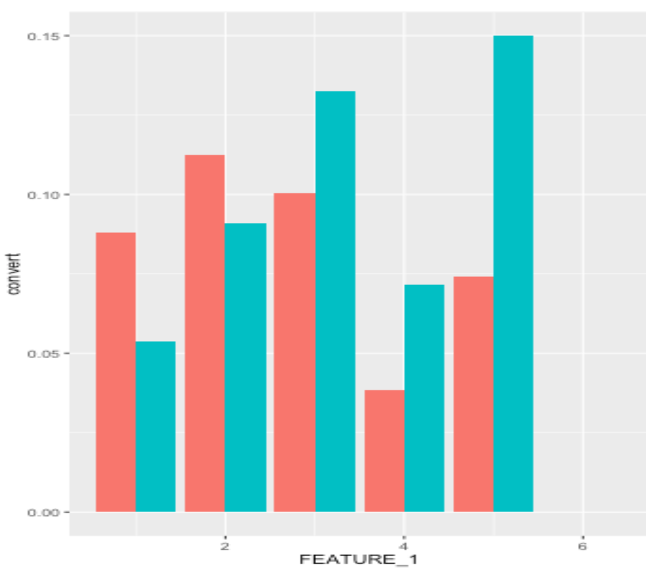


Fig 3: AD 1 conversion for different feature 1 values

From the figures it is clear that

- AD_1 performs poorly for higher values of feature 1 (>= 3)

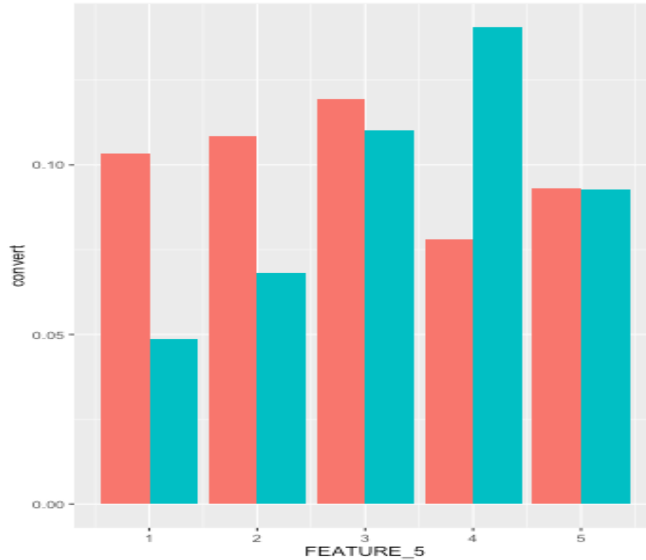


Fig 3: AD 1 conversion for different feature 5 values

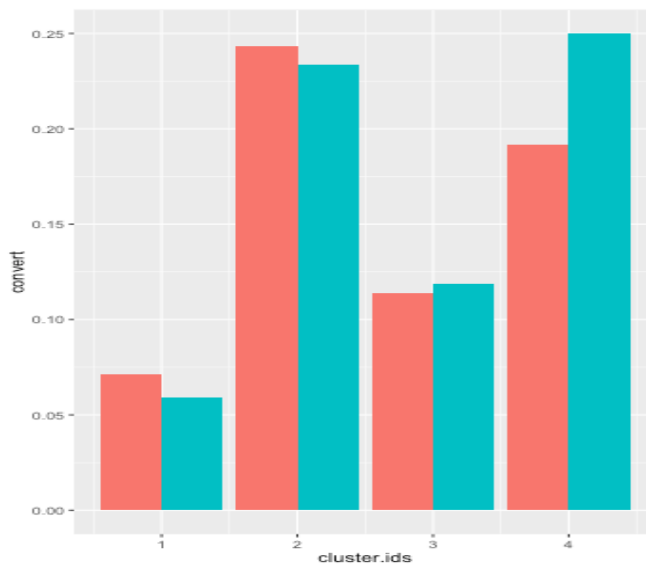


Fig 3: AD 1 conversion for different cluster memberships

- AD_1 performs poorly for higher values of feature 5 (>= 4)
- AD_1 Performs poorly for clusters 3 and cluster 4 where feature 3 seems high

A similar analysis for all strategies gives us the following **key take a ways**:

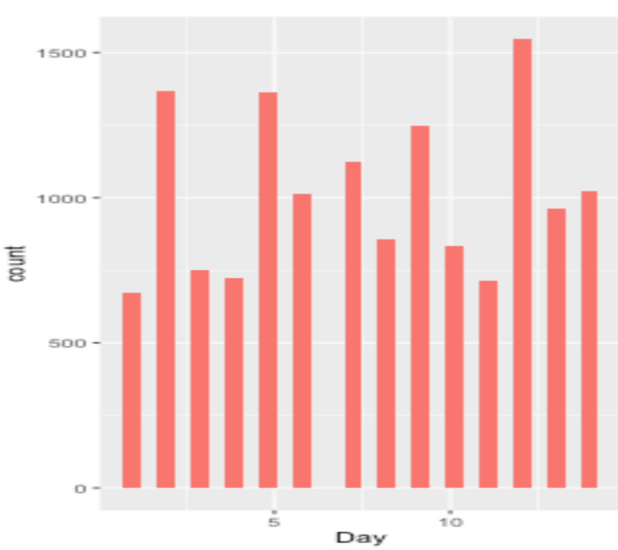
- *All ads perform poorly for feature 1 >=3, AD_2 works well only when feature 1 equals 1*
- *All ads perform poorly for feature 5 >=4, AD_2 works well only when feature 5 equals 2 or 3*
- *All ads perform poorly in cluster 3*
- *AD_1,AD_2 perform poorly when feature 3 is high*
- *AD_3 performs very poorly when feature 4 is low*

Conclusion and Future Scope

Factors that resulted in poor performance of ads were identified. It was shown that in some cases (discussed in the previous section) doing nothing would be more advisable than sending a wrong recommendation. A classifier can be modeled using a "no recommendation" strategy to do the same.

Some days have much higher clicks as compared to others. It might be worthwhile to look at the outliers and understand the cause behind the same.

Fig 5: Day wise trend



It is also observed that the conversion rate is different for different points of time during the day. The TG and CG of our dataset are not representative over the time of day. Some more analysis needs to be done on whether this might affect the overall recommendation strategy performance.