Capstone Project Milestone Report

Randy Linder / November 30, 2016

**Project summary**

I will doing an initial step of a much bigger initiative to be more proactive as opposed to reactive in predicting marketing content to put in front of clients. This project will be to hone in on leveraging our digital tracking data to determine appropriate marketing content to recommend to our advisors. If we were to accurately do this, the hopes are that we would could help create: a better-prepared salesforce, more advisor digital engagement (cheaper for CG and potentially more brand loyalty), stickier assets and stronger sales.

The primary question I’d like to answer is “if we customize email campaigns to particular advisors, can we increase digital engagement (as well as better prepare wholesalers for sales visits)?”

For non-digitally-active advisors, we obviously have little to work with in terms of what content pillars have been of interest. However, we believe that if we can identify digitally-active advisors who ‘look like’ non-digitally-active advisors, they are likely to share similar content preferences. This assumption may or may not show to be a decent assumption to make. Similarly, we believe that we can leverage the behaviors of digitally engaged advisors to make modestly engaged advisors more engaged.

This will be an entirely advisor to advisor model (clustering and other distance-based models).

**Hypothesis**

User-user collaborative filtering on the user attributes listed below increases at least one of the digital engagement indicators within 90 days of deployment of the recommender system.

**Assumptions**

1. The level of digital engagement is independent from content preference.
2. Content pillars are exchangeable across all media (mail, email, website material, in-person topics, etc.)
3. The last 90 days of our advisors’ digital behavior is a good representation of their digital behavior in general.

**Deeper dive on data**

I queried our internal data hub to extract two data sets (advisor attributes and web engagement) and then left-joined them.

Advisor attributes: every row is a distinct advisor

* Advisor ID (internal id, first name, last name, CRD id)
* Gender
* Years in the field
* Internal prospecting score (A, B, C, HP)
* Years doing business with CG
* Retail channel (XC, FC)
* Firm name (and internal firm id)
* Biggest asset class in portfolio (and dollar amount, proportion of portfolio)
* Second biggest asset class in portfolio (and dollar amount, proportion of portfolio)
* Biggest share class in portfolio (and dollar amount, proportion of portfolio)
* Biggest account type in portfolio (and dollar amount, proportion of portfolio)
* Total transactions in last 12 months (LTM)
* Total sales in last 12 months (LTM)
* Total redemptions in last 12 months (LTM)
* Total AUM in portfolio

Web engagement: every row is a distinct content pillar

* Advisor ID (internal id)
* Content pillar clicked via web in last 90 days
* Click count

There are various limitations to this data including:

* Asset class coverage, share class coverage and account type coverage have been moved from rows to columns, which limits how advisors look from a product perspective. For example, if an advisor has a CG portfolio composed of 5 asset classes, we will only get insight into the top two. The alternative approach could have been to create a column for every asset class, share class, account type and then populate the advisors assets under management (AUM) for each.
* Mix of categorical and continuous numerical variables which poses a challenge for segmentation, clustering and distance calculations in general. I think this will make our results especially sensitive to the segmenting and clustering algorithms.
* Advisor IDs in our system include historical advisors, as well as clients that are dually registered, which makes it challenging to know with confidence whether a row on the advisor attributes table truly represents a distinct advisor. There are some methods we can use to increase that confidence, but not 100%.

I did a series of data cleaning techniques including: custom-querying from more than 10 different internal tables, spreading asset classes, share classes and account types from rows to columns to preserve advisor as most granular level in advisor attributes table, unioning two data sets with the same column names but content pillars represented, filtering down to only advisors that 1) have an internal prospecting score and 2) are in the XC or FC retail channel and joining the two data sets above at different levels of granularity to create a new table at the most granular level.

**Approach**

* **WoE:** The first step is to understand how correlated each of the attributes above are with digital engagement. For this, I will use a weight of evidence (WoE) methodology to output an information value (IV) for each attribute. We will then take the 10 largest drivers of digital engagement as our primary independent variables from which to build our clusters. We will not use the segmentation outputted by the WoE analysis.
* **Clustering:** We will build clusters on 70% of the digitally engaged population only. We will apply various clustering algorithms including k-means with different distance approaches (Euclidean, Manhattan, other).
* **Content preference profiles:** Once the clusters are built, we can built out the content preference profiles for each cluster. I envision a pie chart of content pillars assigned to every cluster. These will be used as the output of our model.
* **Formulate experiment:** Once comfortable with the clusters and content preference profiles, we will select a subset of unengaged advisors, pass them through to a cluster and output recommended content pillars. The control group will be sent random content pillars while the tested population will be sent model-output content pillars.
* **Other ways to validate model:** Back testing.

1. Validate the model on a subset of digitally engaged advisors who were not part of the training data.
2. Apply model to digitally engaged advisors in different 90-day periods.
3. Study the advisors who made a transition from non-digitally-engaged to digitally-engaged in the observed data. Back test if the materials recommended by the model overlap and/or coincide with those that were engaged with by the advisors.
4. Assuming null hypothesis, calculate the probability that the content preference distribution by segment was random (ie determine if results are statistically significant).