Customer Segmentation Analysis

William Schill

Customer segmentation is an active part of any complete marketing strategy and is used to identify areas of expansion. In this scenario the author was supplied two small sets of survey data and app usage data. They were merged, cleaned, analyzed and then a simple PCA and KMeans ++ clustering method was used to determine the best segmentation. The best segmentation for this particular analysis was chosen at 4 segments around multiple customer attributes although other segmentations could be more useful.

Data Wrangling

The data sets given in this analysis were relatively small and consisted of both user survey data as well as app usage data. Initial analysis on both of data sets highlighted instances in which user id's had repeat instances. These instances usually differed in country of origin and age bracket. Since there was no discernible way to make these instances unique and multiple variations were taken in an attempt to keep them. It was determined a fair course of action would be to group them together as best as possible using numerical versions of the categorical features. For the survey data this was about 37 instances which was 1.2 % of that data, and for the app usage data this was 35 instances or 1.14% of that data.

The data sets were merged into one using the pandas concatenating method in order to achieve a 1:1 set of unique user ids. Null analysis revealed that with the exception of 5 features, the majority of the features in the data set were < 10% null values. Of these 5, 3 presented the largest problems. The **daily goal** data was handled through the median of the data and both the **primary language motivation follow up** and **other resources** features were expanded using scikit learn's Term Frequency – Inverse Document Frequency Vectorizer (TFIDF). It became apparent through iterations that the daily goal data was far too skewed with a simple NULL fill so it was removed from the data rather than trying to impute that unknown. Additionally, no further action was taken with the Term Frequency portions as they were deemed to be a better side portion of the analysis than directly placed into the segmentation.

After merging of the data, first glance highlighted clear separation of features relative to three separate categories: Demographic Data, App Info and Language relative info. Typical data science methods lead to cleaning of the data and many of the features could be considered categorical or ranked object features than needed clean up and re-mapping of the data in order to be more useful. Additionally, in most cases, the null values were filled in with the mode of the feature. In should be noted that the author is aware of other null value filling methods such as those centered around nearest neighbors, however due to time constraints, the methods were kept simple. In the accompanying notebook, pie charts can be found highlighting the changes to distributions after the null values were replaced. Future efforts should definitely use finer analytical methods.

Most of the features were factored out with the intent on giving them a relative rank, for example, the **subscription** feature had 4 categories which were ranked for analysis. Some features ended up being solely binary.

The features **time spent on survey** and **highest course progress** had negative data in both instances with 91 and 90 points respectively. Since this was nonsensical from the author's perspective, the negative values were replaced with the median values of their respective features and further detail can be found in the corresponding notebook.

A multitude of charts were created in an effort to examine the features as best as possible. All of these instances can be found in the accompanying notebook with some notation regarding each.

Segmentation

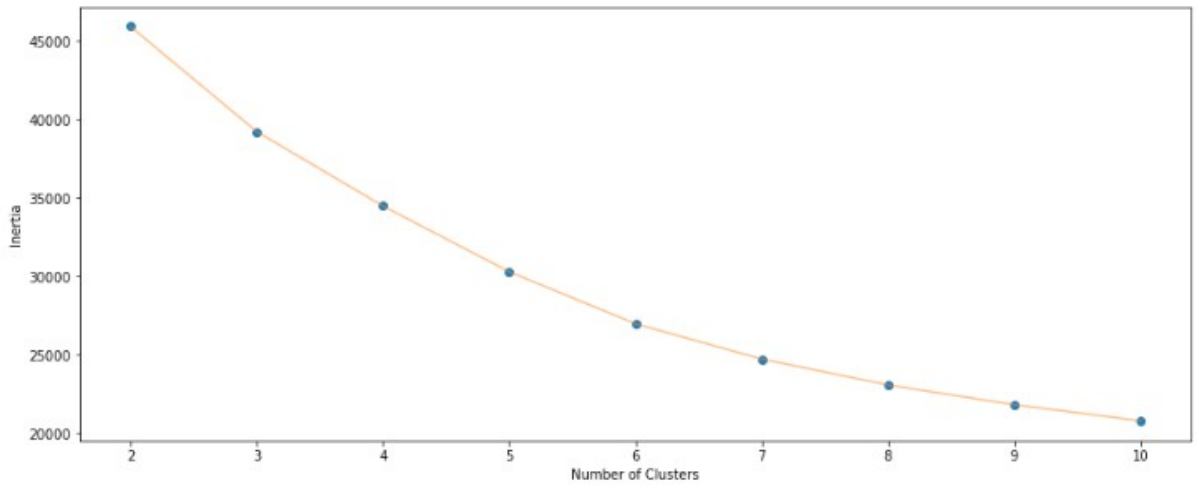
The heart of the problem was to find some relevant segmentation within the data. Now, there are multiple methods available for reducing the dimensions of a data set such as K-Means, PCA, and t-Distributed Stochastic Neighbor Embedding (TSNE). For the sake of brevity and the authors ability to create something both visually meaningful and interpretable, the method used here was K-Means, specifically the K-Means ++ version that focus on randomization specific to avoiding poor clustering in the usual K-Means algorithm.

The data was condensed to just a few sets of features that were believed to be the most relevant to the segmentation:

```
demo = ['age_factor', 'annual_income_factor', 'emp_status_factor', 'stud_factor']
lang_info = ['usage_factor',
        'primary_language_commitment_factor',
        'lang_prof_factor',   'prim_lang_review_yes',
         'subscriber_factor',]
app_info = [ 'highest_course_progress',
        'highest_crown_count', 'n_active_days', 'n_lessons_started',
        'n_lessons_completed', 'longest_streak',]
```

One might wonder why gender is missing from the demos but in reality, a feature with a generally simple distribution is just as easily examined after the fact. Additionally, from a marketing perspective, what would the purpose be to market towards a specific gender for language learning application? This same approach was taken with other features as well such country of origin.

The initial examination over the standardized data highlighted somewhere between 3-7 segments and in the notebook, we took the time to attempt to name them. PCA was able to get us down to 6 components covering over 70% of the variance and was used to provide a 4 segment grouping for the final work. The silhouette score was also used and that is was lead to using 4 segments over 5,6, or 7.



A simple **groupby** statement was implemented over the data frame segments to get a better understanding of how each segment associated with various features.

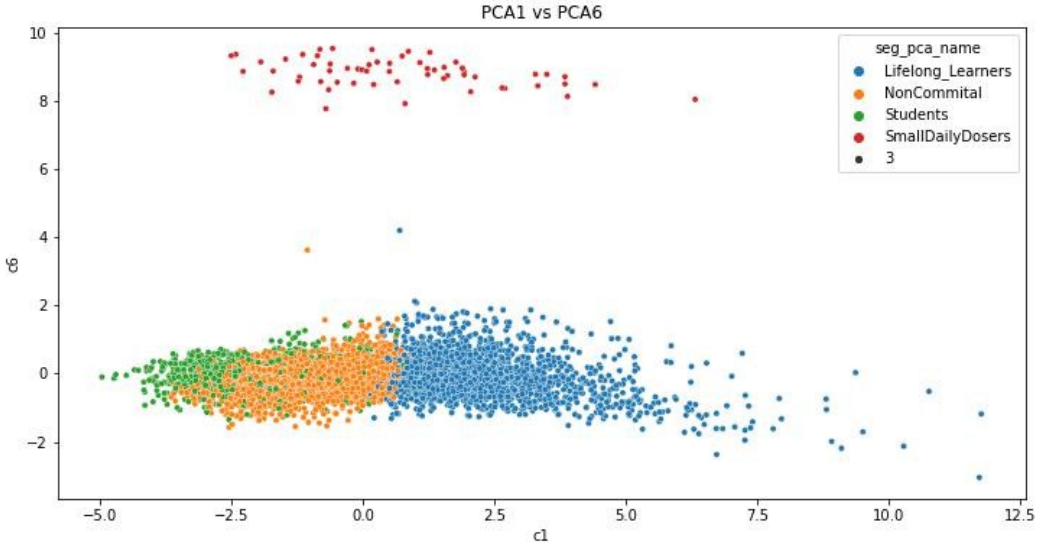
- Segment 0 are older, with the highest income. They are employed and not students and use the app frequently. They are mostly subscribers. - **Life Long Learners**.
- Segment 1 are students. Studying has the more proficient with the language but they are average in course progress and aren't subscribed. - **Students**.
- Segment 2 are older but not the oldest. They are employed and are unlikely to be students however their course progress is similar to that of students. - **NonCommital**.
- Segment 4 are older with decent income and some are students but these tend to be driven by their streak. They are active for long periods of time. Some are subscribed but they might do a little bit each day. - **DailyDosers**

Additional analysis from the PCA segmentation can be found in the notebook as well as against other features such as Gender and Country of Origination.

The import portion of this segmentation is that if the app was to market towards a particular group, it would likely be best to try various campaigns. For example, how does one engage segment 1 (the students) more? They are already in a potential “immersion” environment and a possible method might be a discounted student price to get them to subscribe. Similarly, how do we incorporate those in segment 2 who are also unsubscribed but have more income than students? How do we get them to be more engaged with the app to then move them to Segment 1 eventually?

The interesting portion of this were some of the outliers that were coming out strong in the data. For example, the feature regarding streak length may have benefited from being cleaned of outliers as it ended up creating a segment to itself in a way. Further work could be done to change that impact but it can be seen in the PCA1 vs PCA6 chart below.


PCA1 vs PCA2
c2
c1
seg_pca_name
Lifelong_Learners
NonCommital
Students
SmallDailyDosers
2


PCA1 vs PCA6
c6
c1
seg_pca_name
Lifelong_Learners
NonCommital
Students
SmallDailyDosers
3