UDACITY

< Return to Classroom

Identify Customer Segments

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Congratulations for passing this project. You have very good hand in python. Thoroughly enjoyed going through your code. So keep it up ! For your future reference, you could use following links,

https://www.analyticsvidhya.com/blog/2016/07/practical-guide-data-preprocessing-python-scikit-learn/

https://onlinecourses.science.psu.edu/stat505/lesson/11

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/

https://www.kaggle.com/cgump3rt/investigate-missing-values

https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e

Preprocessing

All missing values have been re-encoded in a consistent way as NaNs.

Brilliant work in re encoding of missing values 👍

Columns with a large amount of missing values have been removed from the analysis. Patterns in missing values have been identified between other columns.

Nice work in identification of outlier columns based on patterns in data 👍

Mixed-type features have been explored, resulting in re-engineered features.

Brilliant work in re engineering mixed type variables 👍

The data has been split into two parts based on how much data is missing from each row. The subsets have been compared to see if they are qualitatively different from one another.

Brilliant work in setting the threshold value cut off based on patterns in data 👍

Categorical features have been explored and handled based on if they are binary or multi-level.

Excellent work in encoding of categorical variables including non numerical variables 👍

Dataset includes all original features with appropriate data types and re-engineered features. Features that are not formatted for further analysis have been excluded.

Brilliant work in Including only those columns that are required for further analysis 👍

A function applying pre-processing operations has been created, so that the same steps can be applied to the general and customer demographics alike.

Nice work in putting everything together in a function 👍

Feature Transformation

Feature scaling has been properly applied to the demographics data. Imputation has been performed to remove remaining missing values.

Nice work here feature transformation 👍 Alternatively you could also use 'RobustScaler' https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

Principal component analysis has been applied to the data to create transformed features. A variability analysis has been performed to justify a decision on the number of features to retain.

Fantastic work in implementation of PCA and choosing the number of components based on the variance explained 👍

Weights on at least three principal components are used to make inferences on correlations between original features of the data. General meanings are ascribed to principal components where applicable.

Excellent work in explaining about the weights of principal components 👍

Clustering

Multiple cluster counts have been tested on the general demographics data, and the average point-centroid distances have been reported. A decision on the number of clusters to use is made and justified.

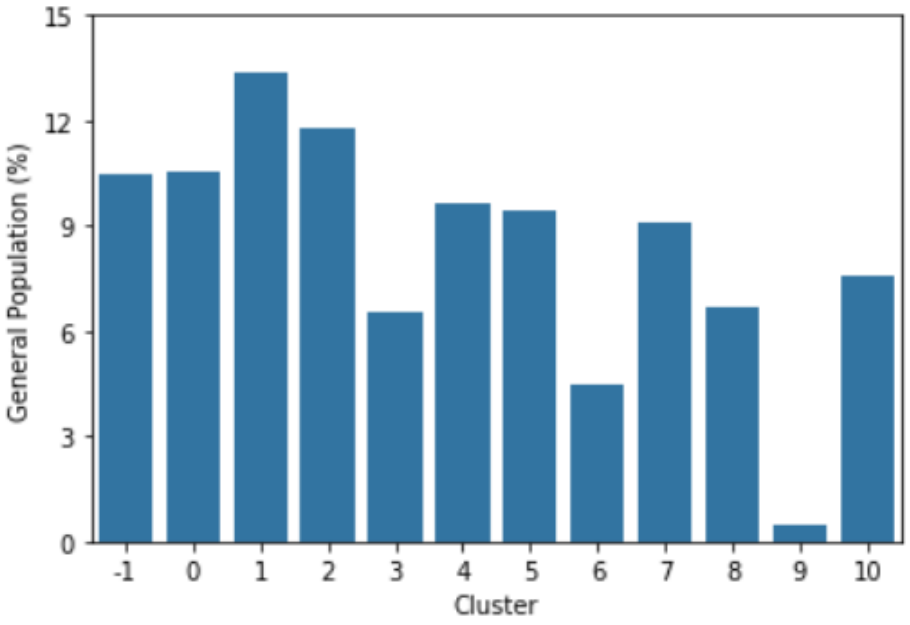
Brilliant work in implementation of kmeans 👍 Alternatively you could also use 'minibatchkmeans' for better performance 👍
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html

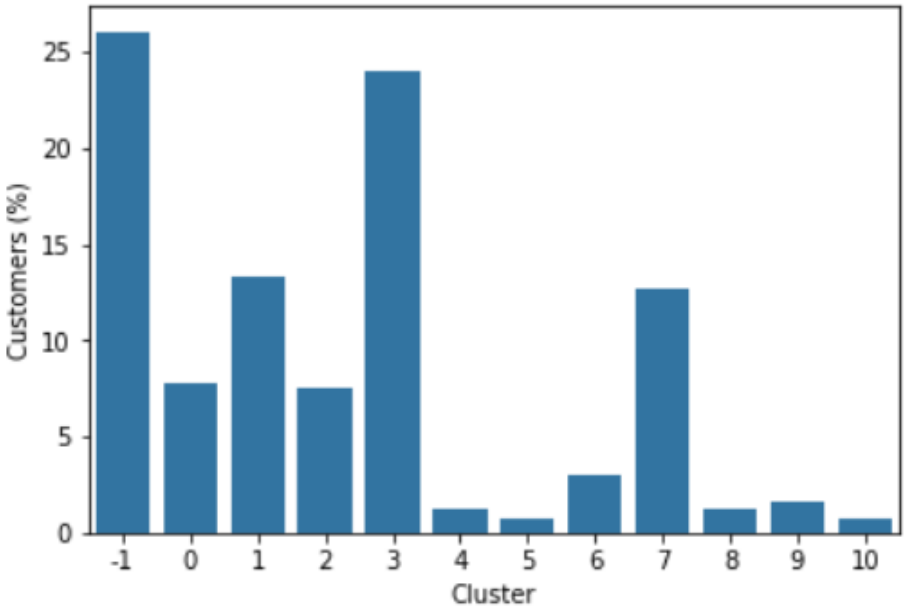
Cleaning, feature transformation, dimensionality reduction, and clustering models are applied properly to the customer demographics data.

Brilliant work in applying the same steps to customer data 👍

A comparison is made between the general population and customers to identify segments of the population that are central to the sales company's base as well as those that are not.

Very good analysis 👍 Alternatively you could also use cluster centers for analysis as shown in below hypothetical example,





Based on above two plots we can say that,

The mail-order sales company is most popular with people in clusters -1, 3, and 9, while it is least popular with people in clusters 4, 5, and 10. One interesting thing about cluster 9 is that it is a fairly small cluster in the general population.

Let us assume the cluster centers for cluster 3 (for 13 principal components) as below,

```
array([-3.00694898, 0.49115263, 2.8052062 , -0.75417665, 2.23232499, 0.17817959, -0.0561161
5, -0.12680474, 0.01261986, 0.18779835, -0.18657745, 0.07406087, -0.20793672])
```

And cluster centers for 5 as below,

```
array([ 4.35826558, -1.3257423 , -2.48667625, -0.35543514, 0.29251846, 0.10548082, -0.027900
59, -0.27247945, -0.14171943, 0.44250003, 0.14115323, 0.04784254, -0.08440762])
```

Then we can write the discussion section as below, by referencing to data dictionary and the weights of principal components (Refer Step 2.3: Interpret Principal Components) as below,

Cluster 3 has its largest coefficients on the first, third, and fifth components. The moderate positive weight on the second component suggests a slight lean towards older people. The negative weight on the first component corresponds with people in urban or more dense neighborhoods and/or lower financial affluence. The positive weight on the third component corresponds with males and male-oriented financial and social profiles. The fifth component is most strongly associated with the mainstream vs. avantgarde affiliation, via the MOVEMENT_TYPE and GREEN_AVANTGARDE features. The positive coefficient indicates tendencies towards the avantgarde rather than the mainstream.

Cluster 5 is in almost the entirely opposite direction as cluster 3. The strongest weights are on the first and third components and they carry the opposite signs. This corresponds with rural and less dense neigborhoods, more wealthy households, and female-oriented profiles. While there isn't a large coefficient on the fifth component, there is a moderate negative coefficient on the second component, again in the opposite sign of cluster 3. This corresponds with a lean towards younger people and their associated profiles.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this project