UDACITY

Return to "Data Scientist Nanodegree" in the classroom

Identify Customer Segments

REVIEW

HISTORY

Meets Specifications

Hi Udacity Learner,
Congratulations on completing the project!.
This was a really challenging project and you have shown lots of expertise. Sure you have learned a lot and we encourage you to
keep up with this hard work. Have a nice day and good luck forward.
Receiving your feedback is always a pleasure.

```
So PLEASE SPARE a few moments of your Precious Time to comment (in the comment section) after rating this review.
```

> What issues did you face in the project?
> How long you took to complete this project?
> Any suggestions or ideas you may have on the project.

I'll look forward to reading from you. Thanks a lot!

Preprocessing

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

Good job encoding the missing values.
Missing value codes given in feat_info's last column have been used to convert all codes to NaNs. ✓
Taken care of the 'X' and 'XX' . ✓

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

Six outlier columns in the dataset is correctly identified and removed. ✓
Correctly identified the columns that have the same or similar counts of missing values. ✓

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.

Data has been split into two parts based on the number of missing values in each row. ✓
Compared the distribution of at least 5 columns of split-off data. ✓

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

Categorical features have been properly processed. ✓
The binary feature, OST_WEST_KZ, is re-encoded numerically before it can be used for further analysis. ✓

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

Two mixed-type features, PRAEGENDE_JUGENDJAHRE and CAMEO_INTL_2015, is correctly engineered into two new features each. ✓

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

Good job cleaning the dataset to include only relevant columns. 👍
Have dropped features which are no longer relevant in their original formulations before moving on. ✓
Note: Always remember to exclude the original columns from the final for any new columns that you have engineered. Otherwise, their values will interfere with the analysis later on the project.

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

All the pre-processing operations like main feature selection, encoding, and re-engineering steps were clubbed

into one single function for future use. Good job!!

Feature Transformation

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

All the missing values were imputed by the mean value and feature scaling is performed. Good job.

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.

Good plot and applying PCA to the data .
The principal component cutoff is justified by drop off in variance. ✓
Tips: More insights on how to choose the number of components in PCA

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.

Discussed the strongest positive and strongest negative features of the three principal components. ✓

A simple note for your future reference and understanding of PCA:

Each PC has one dimension, and the mid-point has value 0. The sign (positive or negative) tells you the direction that a given variable in that PC is going on a single dimension vector.
For example, if you have 5 variables, the first PC has an eigenvalue of 0.8, and the loadings of each variable in this PC are -0.8, -0.5, 0, 0.2, and 0.5, you can conclude that:
1) Variable 3 doesn't play any role in explaining the variation on PC1 (Var3 has value = 0)
2) Var4 has a small role, whereas the others have sizable roles in explaining the variation due to that PC.
3) Var1 will have a greater impact than Var2 and Var5.
4) There is a perfect contrast between Var2 and Var5
5) Finally, the PC scores derived from this PC (linear function of this PC and the observed values for those variables) will show that individuals with negative PC scores will tend to have greater values of Var1 and Var2, and lower values for the remaining Vars, whereas individuals with PC scores greater than 0 will tend to have greater values of Var4 and Var5, and lower of the remaining.

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.

Cluster counts on the data are tested and average point-centroid distances are correctly reported. ✓
Trick: A quick (and rough) method is to take the square root of the number of data points divided by two, and set that as the number of clusters. The elbow method and kernel method work more precisely, but the number of clusters can also depend on your problem.

For example, if I was trying to separate a population into 3 shirt sizes, I would be optimizing for the best location of the 3 clusters, rather than optimizing for the number of clusters that best segment the data.
$k=\sqrt{3}$

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

Customer demographics data have been handled in a fashion consistent with the general demographics 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.

Compared the proportion of data in each cluster for the customer data to the proportion of data in each cluster for the general population. ✓

Identified the target customers. ✓

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review