UDACITY

‹ Return to "Data Scientist Nanodegree" in the classroom

Identify Customer Segments

REVIEW

HISTORY

Meets Specifications

Hello Udacity Student,
This submission was a commendable one. The work was exceptional! You did an amazing job and should be very proud of yourself. After reviewing this submission, I am impressed and satisfied with the effort and understanding put in to make this project a success. All the requirements have been met successfully 💯 %. Congratulations on making it through this project.👏 Based on the skills demonstrated in this work, I encourage you to exploit your ability to keep learning new things, you will be amazed how great a problem solver you are. The Udacity team wishes you success in the upcoming projects . All efforts are appreciated, please keep the learning flame burning. Have a nice wonderful day!🏆

> Nothing specifically, I have a lot of commented out code because I was jumping back and forth a lot calling back to the procedures to clean the data (for the function) and to relate the dataframes/arrays that were relevant for the work. So, if it seems like there's a lot of commented code, that's why!

Very well justified.👍

Suggestion

You could remove all the commented code prior to a submission when you are comfortable with your work. This makes is it more presentable, professional and easier to navigate.🙏

- Why should we comment code?
- Top 10 tips of writing cleaner code

Preprocessing

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

Awesome!
The implementation successfully investigated the patterns in the amount of data missing in each column and identifies `6` columns as outlier columns having more than `20%` proportion of missing values. Moreover, these columns were also dropped. Good work.👏

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

Exceptional effort!
The submission made use of a control statement to re-encode missing value codes to `NAN` using missing value codes given in `feat_info`'s last column.✔

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

Good work!
Two mixed-type features, `PRAEGENDE_JUGENDJAHRE` and `CAMEO_INTL_2015`, have been successfully engineered into two new features each. This is good work.👍

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

Brilliant job!
The handling and exploring categorical features is exceptional. The implementation paid special attention to one binary feature, `OST_WEST_KZ`, which has been correctly re-encoded numerically replacing `W` with `1` and `0` with `0`. Nicely done!

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.

Nice job!
Data points have been successfully split into two sets with a good threshold of `35` as could be observed from the provided visualization. Also, the subsets have been compared to see if they are quantitatively different from one another.
However, an ideal value would be a value between `9–32`. Given from the plot, `25` will be a great threshold.

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

Awesome work with this function!
The function correctly incorporates all the above code segments of converting missing value codes to `NaN`, removing outlier columns, dividing dataset into 2 parts and re-engineering new features. This is good job.
The threshold value here could also be adjusted.

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

Nice here! The submission did drop features like `PRAEGENDE_JUGENDJAHRE` and `CAMEO_INTL_2015` which are no longer relevant in their original formulations before moving on. This is good work with the cleaning.

Feature Transformation

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

Good job applying feature scaling using standardScaler class to the demographic data and also removing all remaining missing values.

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.

Great job using weights on at least three principal components to make inferences on correlations between original features of the data.
The submission provided a detailed discussion on both the highly negative and positive values for the three principal components.💪

Pro Tips

Visit these links for more insights.

- Interpret the key results for Principal Components Analysis
- How to interpret/analysis principal component analysis (PCA)

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 work applying PCA to the data that resulted to transformed features. A justification by the turning point is also provided on the decision of the number of features to retain.

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.

Good work with the decision on the choice of cluster. This is reflected from the elbow in the curve of average distance from points to their assigned cluster centers. You got this on point!

Pro Tips

- Finding the optimal number of clusters for K-Means through Elbow method
- kmeans elbow method

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

Positive
Nice one with the plot! Great work with the observation in this section, indicating the kinds of people that are part of a cluster that is overrepresented and underrepresented in the customer data compared to the general population.

External resources

- Find Your Best Customers with Customer Segmentation in Python
- A Quick Look at Market Segmentation
- How to use Customer Segmentation To Learn
- Segmentation: The Basics For Understanding Your Customers
- Customer Segmentation - SlideShare

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.

Awesome! All the steps of cleaning, feature transformation, dimensionality reduction, and clustering models are applied properly to the customer demographics data without creating new sklearn objects.

DOWNLOAD PROJECT

RETURN TO PATH

Rate this review