UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

1 SPECIFICATION REQUIRES CHANGES

Awesome

You did an excellent work. You have just small things to change.

Suggestion

I recommend checking the Google Python Style Guide, there are great tips about how to improve coding, in general: https://google.github.io/styleguide/pyguide.html

Data Exploration

Three separate samples of the data are chosen and their establishment representations are proposed based on the statistical description of the dataset.

Awesome

Great description of each sample.

A prediction score for the removed feature is accurately reported. Justification is made for whether the removed feature is relevant.

Awesome

You predicted very well the R^2 and the conclusion is great. The attributes with lower R^2 are more relevant since they cannot be predicted and the higher R^2, less relevant to use the attribute.
This article mention the high correlation as a factor to remove attributes.

Comment

About your comment of using different `random_state`, the best would be to run the test for 100 different `random_state`s and getting the average.

Student identifies features that are correlated and compares these features to the predicted feature. Student further discusses the data distribution for those features.

Requirement

You still have to answer "How is the data for those features distributed?"
Check the image below, it should help you:

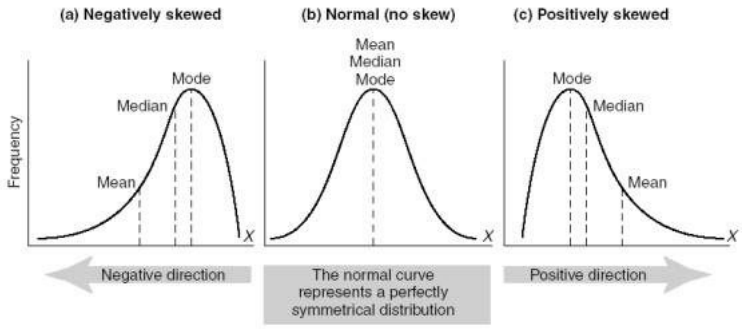

FIGURE 15.6 Examples of normal and skewed distributions

Reference: https://www.quora.com/How-do-outliers-affect-normal-distribution-in-statistics

Data Preprocessing

Feature scaling for both the data and the sample data has been properly implemented in code.

Student identifies extreme outliers and discusses whether the outliers should be removed. Justification is made for any data points removed.

Awesome

Great analysis checking the outliers. The explanation about the removal is great but keep in mind that there is not a definitive answer here.
This article has a interesting discussion about detecting outliers. And this article discusses about dropping or not the outliers.

Feature Transformation

The total variance explained for two and four dimensions of the data from PCA is accurately reported. The first four dimensions are interpreted as a representation of customer spending with justification.

Awesome

You are right. It is important to notice not only the absolute value of each parameter inside the dimension but also the direction (some attributes are in opposite directions).
Only two dimension can represent over 70% of the client data.
This article has a nice visual example of PCA.
This article shows a discussion about what each dimension represents.

PCA has been properly implemented and applied to both the scaled data and scaled sample data for the two-dimensional case in code.

Clustering

The Gaussian Mixture Model and K-Means algorithms have been compared in detail. Student's choice of algorithm is justified based on the characteristics of the algorithm and data.

Awesome

Good description of both algorithms.
There is an article that present this comparison. Or this presentation can help you.
Nice explanation justifying your choice.

Several silhouette scores are accurately reported, and the optimal number of clusters is chosen based on the best reported score. The cluster visualization provided produces the optimal number of clusters based on the clustering algorithm chosen.

Awesome

The groups are well proposed and justified by the data. Well done!

The establishments represented by each customer segment are proposed based on the statistical description of the dataset. The inverse transformation and inverse scaling has been properly implemented and applied to the cluster centers in code.

Sample points are correctly identified by customer segment, and the predicted cluster for each sample point is discussed.

Awesome

Good analysis of the sample points.

Conclusion

Student correctly identifies how an A/B test can be performed on customers after a change in the wholesale distributor's service.

Awesome

Great description about how to deal with the A/B test considering the clusters. This kind of analysis is critical for the business.
It is critical to notice that testing only one cluster, may not give any information about the other cluster. In an A/B test it is important to understand each cluster as a different kind of customer and they should be analysed knowing this difference.
This Netflix article is great about A/B testing and this quora discussion presents the problems you can face in the test.

Student discusses with justification how the clustering data can be used in a supervised learner for new predictions.

Awesome

Great description about how to use this information in the supervised learning. Using this information we end up with more knowledge about our customers.

Comparison is made between customer segments and customer 'Channel' data. Discussion of customer segments being identified by 'Channel' data is provided, including whether this representation is consistent with previous results.

Awesome

Great analysis.

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Student FAQ