



PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!  

Requires Changes

7 SPECIFICATIONS REQUIRE CHANGES

Awesome

You did an excellent work. You just need to fix some details.

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.

Requirement

In the rubric you need to use the statistical description of the dataset. Compare the samples to the mean and percentile of the features to justify your conclusions.

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

Requirement

You predicted very well the R^2 but the conclusion is the opposite. The attributes with lower R^2 are more relevant since they cannot be predicted by other parameters. The higher R^2 is a less relevant attribute since it doesn't bring any new information to the analysis.

[This article](#) mention the high correlation as a factor to remove attributes.

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

Awesome

Your analysis here is great. Can you see how it reinforces the conclusion in the previous question? **Grocery** and **Detergents_Paper** are almost redundant.

It is important to notice that the data is positively skewed and this is the reason the for pre-processing done in the next steps. Check the image below:

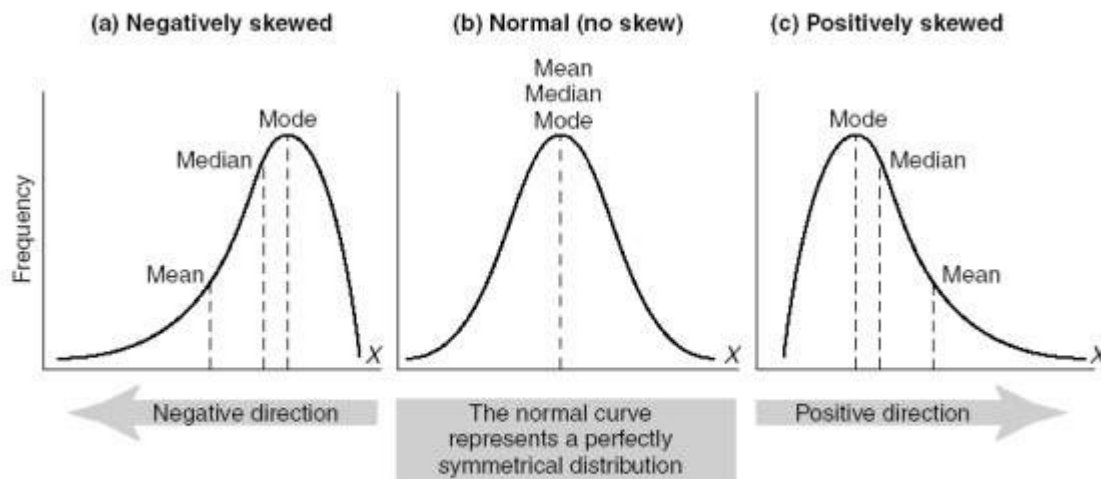


FIGURE 15.6 Examples of normal and skewed distributions

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.

Requirement

You need to review the answer here, you presented several outliers but in the explanation you say there is just one outlier. In the rubric you also need to identify the extreme outliers (data points considered outliers for more than one feature). HINT: There are 5 in total.

Suggestion

[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.

Requirement

The question 5 needs to be answered.

Suggestion

[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.

Requirement

In the rubric you need to use the statistical description of the dataset. Compare the clusters to the mean and percentile of the features to justify your conclusions.

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

Requirement

Here you should compare the Question 1 and Question 8.

Conclusion

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

Requirement

In the rubric you need to answer how an A/B test can be performed on customers. 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](#)



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

[Watch Video](#) (3:01)

RETURN TO PATH

Rate this review

[Student FAQ](#)