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

Hey,

You have done well in updating the project from last time but one small issue still remains. I have tried to elaborate on the issue and if it's still unclear to you, you can always clarify it on the forums. We look forward to your next submission, keep at it!

Data Exploration

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

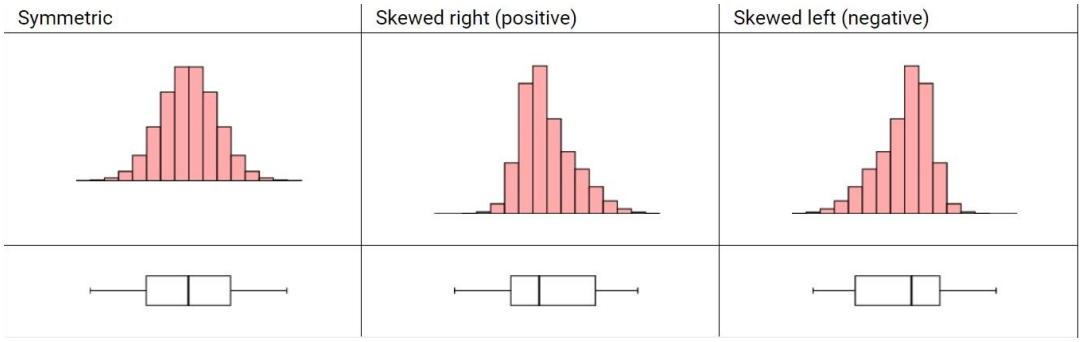
Good job commenting on the establishment that could be represented by each sample point by looking at the dataset statistics.

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

Great work getting the R^2 score here and commenting on the feature's relevance.

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

Nice work identifying the correlated features and commenting on the distribution of the data. Is the data left skewed or right skewed?



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.

The double counted outliers have been correctly identified. Another reason to remove these outliers could be their impact on clustering algorithms. Clustering algorithms are sensitive to outliers and algorithms like KMeans give a lot of weight to outliers (as it tries to optimize the sum of squares).

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.

> However, these two also indicate that there might be corelated features in the dataset that we haven't noticed beforehand and might therefore helpful to segmentate customers.

I'm not sure what you mean here. The second and third principal components are not correlated. We require you to comment that fresh and delicatessen features have a negative correlation with each other in the third dimension as they have opposing signs. It is true that both of these features contribute equally to the dimension as you should be looking at their absolute values, but since they have opposite signs, it means that one increases as the other decreases.

You also need to interpret the fourth dimension.

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.

Some key differences between the models:

Speed/Scalability

- K-Means faster and more scalable
- GMM slower due to using information about the data distribution — e.g., probabilities of points belonging to clusters.

Cluster assignment

- K-Means hard assignment of points to cluster (assumes symmetrical spherical shapes)
- GMM soft assignment gives more information such as probabilities (assumes elliptical shape)

You can read more on the differences between the two models here: https://www.quora.com/What-is-the-difference-between-K-means-and-the-mixture-model-of-Gaussian

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.

As the last reviewer described, having balanced clusters can sometimes take precedence over silhouette scores. In this case, two clusters does give a balanced distribution as you can see in the plot below. Hence 2 should be the optimal choice.

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.

Nice job determining the establishments by looking at the dataset statistics.

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

Good work comparing your intuition with the results from the clusterer.

Conclusion

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

You should update your answer here as it indicates that you selected 6 customer segments whereas only 2 were selected. You could also add more detail by describing how the control and variation groups would be decided for each segment and how the results from multiple tests would be combined.

You can read more on A/B testing from the following links:
https://en.wikipedia.org/wiki/A/B_testing
https://www.quora.com/When-should-A-B-testing-not-be-trusted-to-make-decisions/answer/Edwin-Chen-1
http://multithreaded.stitchfix.com/blog/2015/05/26/significant-sample/
http://techblog.netflix.com/2016/04/its-all-about-testing-netflix.html
https://vwo.com/ab-testing/

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

You have correctly noted that the created customer segments can be used to turn this into a classification problem.

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.

Even though there is some overlap in the central region, the overall alignment is pretty good!

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