REVIEW

Meets Specifications

The key takeaway is that a feature which cannot be predicted by other features should not be removed from the dataset for the sake of reducing the dimensionality of our dataset, since its information content, however useful that might eventually prove to be, is not contained in the rest of the features.

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 by other parameters. The higher R^2 is a less relevant attribute since it doesn't bring any new information to the analysis.

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 work here!

✓

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

✓

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

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.

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.

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.

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.

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.

Conclusion

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

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

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.