UDACITY

Back to Machine Learning Engineer Nanodegree

Creating Customer Segments

REVIEW

HISTORY

Requires Changes

1 SPECIFICATION REQUIRES CHANGES

Hi there, it's Cláudio! Thanks for sending all the required files for the review process and for all code executing without any problem.

Congratulations for your project submission and for the quality presented in this challenge. You really did a great job.

However there is a item from the rubric that needed to be reviewed. I don't think you will have any problem to fix that.

I hope you had enjoyed doing this project and put in practice important concepts from machine learning. I will leave my contact below in case you have any doubt about this review as well to stay connected.

That's all. Enjoy machine learning and keep it up the great work. I will look forward for your next project submission.

Finally, I wanted to share a interesting tool from google that helps machine learning engineers to understand the data really fast and then make a decision on what type of algorithm it will fit better that data, called: Facets - Visualizations (https://pair-code.github.io/facets/). Definitely check it out.

The power of machine learning comes from its ability to learn patterns from large amounts of data. Understanding your data is critical to building a powerful machine learning system.

Facets contains two robust visualizations to aid in understanding and analyzing machine learning datasets. Get a sense of the shape of each feature of your dataset using Facets Overview, or explore individual observations using Facets Dive.

Thank you.
Cláudio

email: cgimenest@uol.com.br
Linkedin: https://www.linkedin.com/in/claudiogimenestoledo/

Data Exploration

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

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

Good job:

> Grocery and Detergents_Paper have some positive score (approximately 0.60). Milk has almost no score (approximately 0.10). The other three are not predictable (negative scores).

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

Great job.

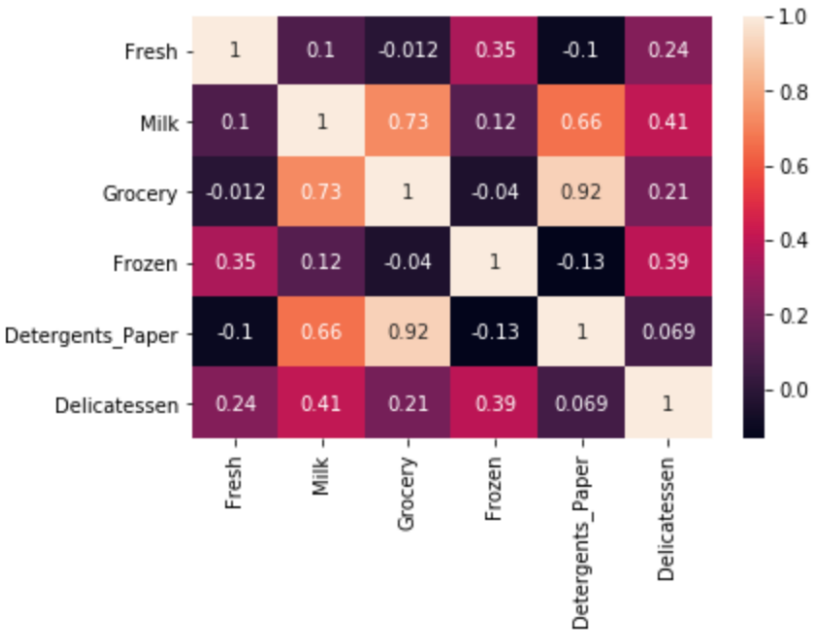
> Grocery <-> Detergents_Paper are correlated. Milk <-> Detergents_Paper, and Milk <-> Grocery are also correlated but not to the same degree.

Suggestion:

- I would recommend you to implement the heat map with values which makes this job easier, take a look into an example below:

```
In [6]: #Adding graphs for correlation to help out throught the process
        import seaborn as sns
        sns.heatmap(data.corr(), annot=True)
```

```
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1a732790>
```



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.

Almost there

Almost there!

Actually there are only 5 data points which are double-counted. Try to identify from the tables you have provided where a indice is also counted at another table.

For example:
65, 66, etc...

Bonus:

- Here is an great article discussing about the strategy of drop or not outliers, definitely check it out: http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

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.

Great explanation.

Bonus:

- Here I will leave some articles about A/B testing in AI and how their are complementing each other:
  https://hackernoon.com/ai-as-complement-to-a-b-test-design-e8f4b5e28d92
  https://www.dynamicyield.com/ab-testing/
  https://www.mediapost.com/publications/article/305837/ab-testing-vs-ai-conversions.html

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.

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

RETURN TO PATH

Student FAQ