UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

5 SPECIFICATIONS REQUIRE CHANGES

Hello Student,
You almost get it done, keep it UP!
There are only few minor mistakes you need to amend, I believe you already understand most of the concept in this project.

Data Exploration

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

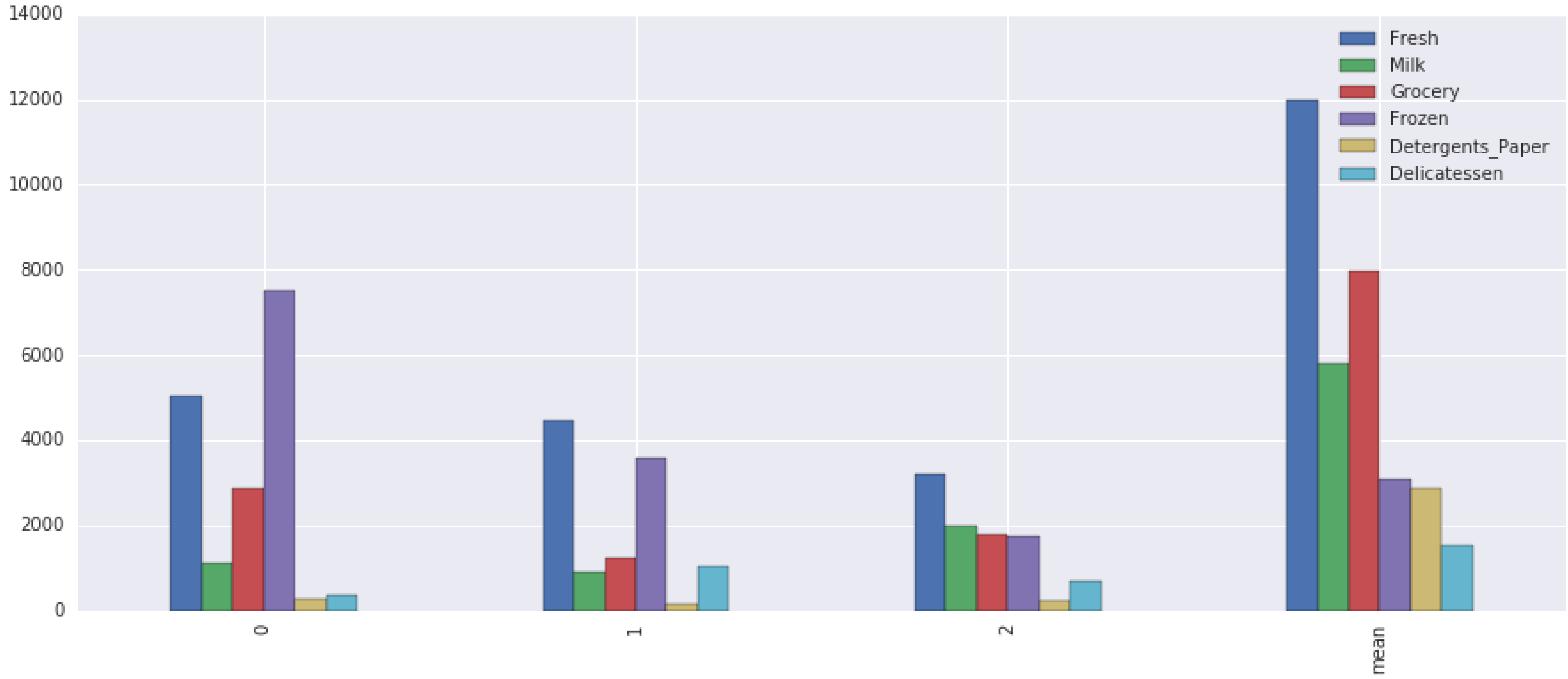
Well done commenting on the different types of establishments the three customers could represent.

Pro Tips:

COMPARING TO DATASET AVERAGE

You could quickly draw a bar plot to visualise the amount of each product purchased for each sample, together with the dataset mean.

```
# Import Seaborn, a very powerful library for Data Visualisation
import seaborn as sns
samples_bar = samples.append(data.describe().loc['mean'])
samples_bar.index = indices + ['mean']
_ = samples_bar.plot(kind='bar', figsize=(14,6))
```



This will make comparing the three different sample points with each other much easier.

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

Great analysis to identify that the amount of `Detergents_Paper` purchased is not necessary to identify specific customers!

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 here identifying the features which are correlated, and also confirming your suspicions about the features from the previous question! However, there is one point you need to address here:

Required:

- Please comment on the distribution of the features (you instead gave comments on the relationship between `Detergents_Paper` and `Grocery`). You need to comment on the individual feature distributions.

Data Preprocessing

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

Nice work implementing feature scaling :)

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

Nice job catching the outlier 66, 65, 128, 154, and commenting on whether it should be removed or not. Please see my feedback for this section, including what is required:

Required:

```
    # TODO: Calculate Q1 (25th percentile of the data) for the given feature

    Q1 = np.percentile(log_data['Delicatessen'],25)

    # TODO: Calculate Q3 (75th percentile of the data) for the given feature
    Q3 = np.percentile(log_data['Delicatessen'],75)
```

In this part, the question required you to get the percentile for all the given features but not the `Delicatessen` only. As a result the outlier list is not correct, however, `65, 66, 128, 154` appeared more than once coincidentally.

- Please amend.

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.

Fantastic work in identifying the amount of variance explained by the first 2 and the first 4 dimensions.
Also, you did a great job in interpreting the first four dimensions as a representation of customer spending.
Please see some suggestions and comments below for more information on this section:

Suggestions and Comments:

These two links helped me a lot in understanding Principal Component Analysis and what it does, hope it helps you too.
https://onlinecourses.science.psu.edu/stat505/node/54
http://setosa.io/ev/principal-component-analysis/

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

Nice job here!

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.

Great work elaborating on GMMs and K-means, and giving a solid reason as to why you chose [K-means] algorithm for this particular problem.
You might want to provide some citations and reference for your work to make it more credible.
Below are some of my comments, feedback, and suggested reading:

Gaussian Mixture Models

- You did well on giving the advantages of Gaussian Mixture Model.

SUGGESTED READING:

- If you feel as going deeper with regards to Gaussian Mixture Models, check out the following links:
  - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/mixture.html
  - http://www.nickgillian.com/wiki/pmwiki.php/GRT/GMMClassifier
  - http://scikit-learn.org/stable/modules/mixture.html#gmm-classifier

K-Means Clustering

- Your description on the advantages of K-means is very explicit

SUGGESTED READING:

- Check out these links for even more thorough explanations of K-Means Clustering:
  - http://playwidtech.blogspot.hk/2013/02/k-means-clustering-advantages-and.html
  - https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
  - http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm
  - http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Reason for choice of Algorithm:

- Well done here again! Please see the links below for some more information on how to compare the two algorithms:

SUGGESTED READING:

- 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.

Great job identifying the optimal cluster score as 2, and identifying its associated silhouette score

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 attempt here, and just by reading your proposed establishments, they all make sense! Please see my comments and feedback below:

Required:

Kindly note that this question actually requires you to provide the set of establishments each center of the clusters might represent based on statistical description of the dataset. It is a bit similar to the very first question of the project, where you are asked to discuss what type of customer your picked sample should represent.

- Please provide justifications for your claims based on the **statistical description** from the dataset.
- Here is an example:

> The right hand side cluster (Segment 1) appears to be represented by customers who spend above average on Milk, Grocery, DetergentsPaper and below average in Fresh & Frozen -- these establishments could be grocers and retailers.

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

Nice try here, and you did well predicting clusters for each sample point. Please see my feedback below:

Required:

- The question actually has two main parts:
    - 1) It asks to state which customer segments from **Question 8** best represents the sample points?
    - 2) It asks to justify whether the predictions are consistent with these.

However, you only answered the first part of the question, and didn't answer the second.

- Please compare your predicted results to the original or initial interpretation of the representations of the sample points. Do the cluster representation and your sample representation agree with each other? Do the features of the sample points correlate with the feature representations of the assigned cluster? Please elaborate on this more explicitly.

Conclusion

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

Good attempt in answering this section. Please see some feedback:

Required:

You mentioned " The wholesale distributor should take samples from both clusters and collect feedbacks", however, you did not state how would the test conduct, which one would be the test group and which would be the test group.

- Please note that the question requires you to identify how to use the **structured data** to identify which customers will be affected. So it requires you to use the cluster structure to help or assist you in conducting an A/B test

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

Nice job stating that the cluster labels can be used as an **target label** to a supervised learning algorithm!

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.

Well done comparing the clusters from your algorithm to the customer 'Channel' data!

RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review

Student FAQ