Unsupervised Learning: Large Scale Data Analysis

Ilan Prais, Yonatan Ehrenreich, Itay Gradenwits

March 4, 2022

Abstract

In the following research we used many unsupervised learning techniques to find the hidden relations in large scale data-sets. In order to do it, we compared different clustering methods with the desire of finding the optimal number of clusters in the data-sets. We also tried to fit the clustering to an external classification, in order to expose hidden relations in the data. In addition, we reduced the dimension of the data in order to plot the data and the association between the clusters and the external variables. Finally, we found the anomalies in the data and tried to evaluate the association between the anomalies and the external variables. For the repository with the complete code and results, please click here.

1 Introduction

In this research we will be analyzing two large scale data-sets. The first data-set is "US Census 1990", which contains about 2.5 million records of the US 1990 Census. The second data-set is "Online Shoppers Purchasing Intention" data-set, which contains information about the circumstances of purchasing for about 12000 people, and whether they finally purchased. In the research we will explore different clustering methods on these 2 data-sets, compare them, and visualize the best results in order to find the hidden factors of the data-sets and infer as much as we can about the data.

2 Methods

2.1 Data Preprocessing

First, we removed from the data the columns related to the external variables. These variables weren't part of the clustering, as we wanted to learn how much information the clustering can give us about these variables. For the census data-set we observed that the data is categorical, so we transformed the data into one hot encoding, so we can treat the data as continuous and calculate euclidean distances between points. As for the online shoppers data-set, we observed the data is mixed; both categorical and numerical. So, we transformed the categorical features to one hot encoding as we did before, and normalized the one hot encoding together with the numerical data to force them both to be on the same scale, resulting in one data-set which can be treated as continuous data and calculate euclidean distances between points.

2.2 Clustering

We ran 5 different clustering methods on our data; K-Means, Gaussian Mixture, Spectral Clustering, Agglomerative (Hierarchical) Clustering, and DBSCAN.

2.3 Evaluation Metrics

To evaluate how "good" a clustering method is, we calculated the silhouette score for a group of possible parameters of the model. We chose the parameters for the method as the parameters which gave the highest silhouette score. Specifically for K-means, we used elbow method as another evaluation metric. For finding the parameters of DBSCAN, we used k-distance elbow instead of silhouette because silhouette between 2 DBSCAN methods with different parameters is problematic - more outliers will always give a higher silhouette score, which makes the results less reliable.
To find out how much a clustering method is associated with an external variable, we calculated the mutual information between the clusters and the real labels, and found the clustering methods which maximizes the mutual information.

2.4 Statistical tests

Our desire was to find the best clustering method for each data-set. We used statistical tests to prove the statistical significance of our results. We used Anova test to calculate the statistical significance that different algorithms give different results. Then, to calculate the statistical significance of one algorithm being better than a second algorithm (by a specific evaluation method), we used a paired t-test "battle". In every statistical test, we used 0.01 as the recommended significance level.

2.5 Anomaly Detection

We used the DBSCAN and One Class SVM methods for detecting anomalies in the data. DBSCAN is a density based algorithm, so it helped us find the points which are outliers. SVM separates a class of observations from another, and using it as one class will cluster the "normal" points hence will find the outliers.

2.6 Dimension Reduction

For the Census data-set we primarily used t-SNE for reducing the dimension of the data. We reduced the dimension to 2 in order to be able to visualize the clustering. However, we used Multiple Correspondence Analysis for visualizing the anomaly detection. For the Online Shoppers data-set we used Factor Analysis of Mixed Data for reducing the dimension of the data, since this data-set is mixed numerical and categorical.

3 Results

3.1 Parameter Choice

As explained in *methods* section, we used silhouette score to evaluate the parameters for all the clustering methods. Since the Census data-set contains about 2.5 million records, it could not be clustered directly, so we had to calculate the silhouette score by taking 30 random samples of the data and for each sample calculate the silhouette score after applying the clustering methods on it, and finally average the silhouette scores of all of the samples. From the central limit theorem, we got that the average of the silhouette scores of the sample converge to the expected value, which is the expected value of the silhouette score of any random sample of the data, which is exactly what we wanted to achieve.

The results for the silhouette score of each algorithm are shown in Figure 1.

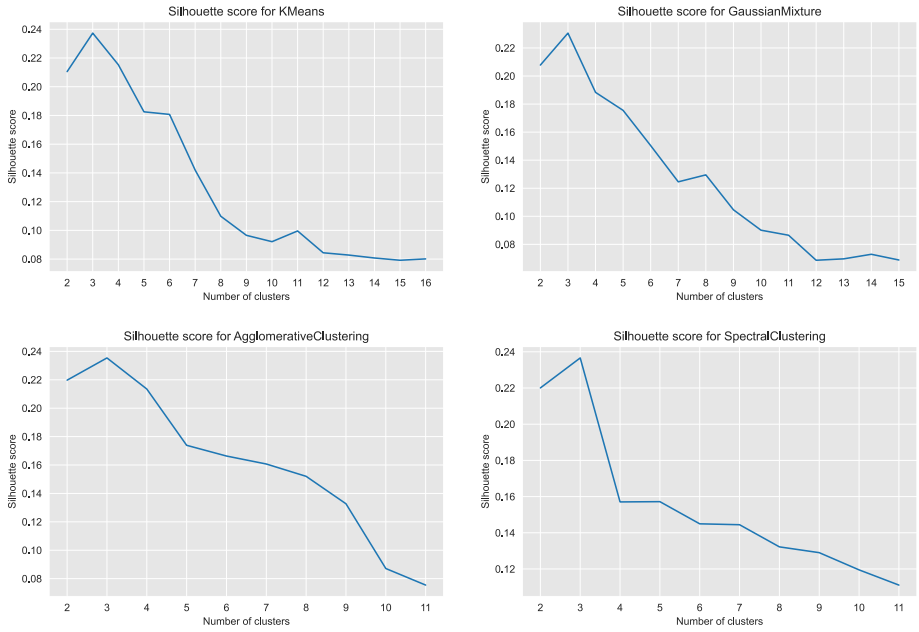

Figure 1: Silhouette score of all models

From the silhouette score graphs, we inferred that the optimal number of clusters in the Census data-set is 3 for Gaussian Mixture, Spectral Clustering and Agglomerative Clustering.

On top of that, we used elbow method on the K-Means cost function in order to get another aspect on the optimal number of clusters for K-Means.

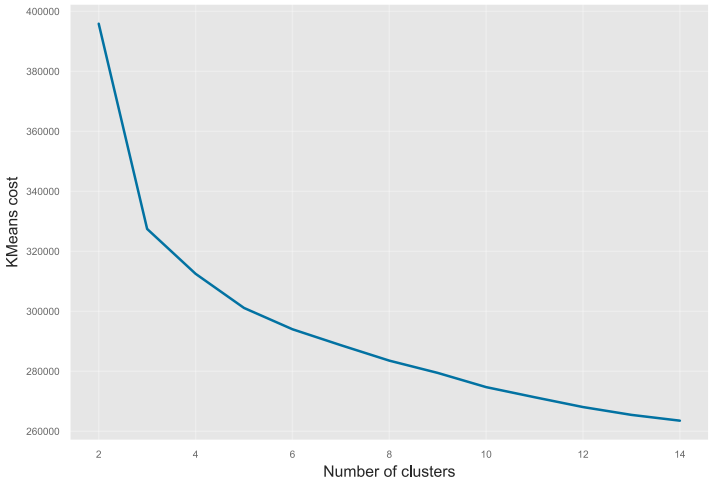

Figure 2: Cost graph for K-Means

Note that the elbow in the graph in Figure 2 indicates that the number of clusters should be around 5. For the parameter we chose the average between the silhouette and the elbow,

hence 4 clusters.
Finally, the average k-distance graph for DBSCAN shown in Figure 3.

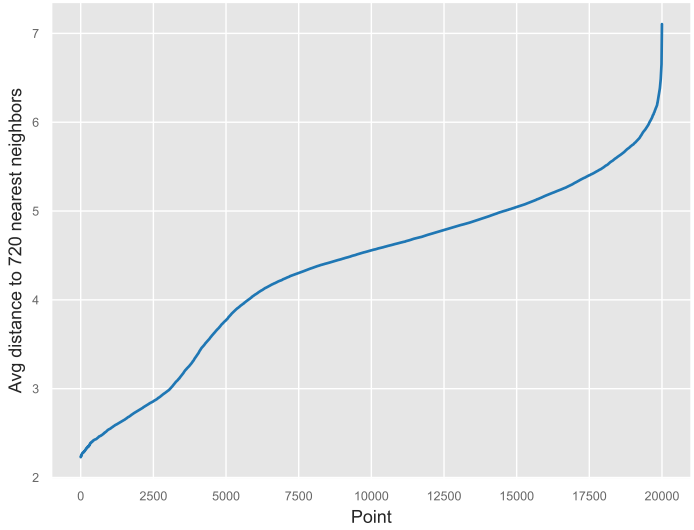

Figure 3: Average distance graph to the neighbors of the points

We saw that the elbow is at 5.5, so we chose $\epsilon = 5.5$ which gave us 1 cluster. We chose 720 for *minsamples* as the recommendation is $minsamples = 2 * dim$.

Note, that for the comparison between the algorithms we used the closest ϵ to the elbow which results in at least 2 clusters. We found that ϵ to be 4.5. The reason behind the desire for 2 clusters is that we need at least 2 clusters to run the silhouette evaluation.

Finally, We can summarize the parameter choice into a table.

Method	number of clusters	additional parameters
KMeans	4	none
Gaussian Mixture	3	none
Agglomerative Clustering	3	none
Spectral Clustering	3	none
DBSCAN	1	$minsamples = 720$, $\epsilon = 5.5$
DBSCAN	2	$minsamples = 720$, $\epsilon = 4.5$

Figure 4: best parameters for each method summary - Census data-set

As for the shoppers data-set we used the same methods, and we found the optimal parameters shown in Figure 5.

Method	number of clusters	additional parameters
KMeans	2	none
Gaussian Mixture	3	none
Agglomerative Clustering	2	none
Spectral Clustering	2	none
DBSCAN	1	$minsamples = 150$, $\epsilon = 2.3$

Figure 5: Best parameters for each method summary - Shoppers data-set

3.2 Finding the Best Clustering Method

We wanted to find the method which is statistically significantly better than the other algorithms by the silhouette score. For that, we took 30 random samples and clustered each sample using the five algorithms, and for each clustering calculated its silhouette score. Finally, we have gotten enough results to perform some statistical tests.

We started with the Census data-set. First, we performed Anova test for calculating the statistical significance for the five algorithms having different results. The p-value of Anova test was 9.15e-41 which is below 0.01 (the recommended significance rate), so we could reject the Anova null hypothesis. From that, we inferred that the five algorithms have statistical significant different results.

Next, used paired T-tests for finding the best algorithm. We found the best algorithm by starting with an arbitrary algorithm, and for every algorithm of the five, compare it with the current best algorithm and continue with the one which is statistically significant better. This "battle" method is similar to the method of finding the maximum element in an array. The results are shown in Figure 6.

First Algorithm	Second Algorithm	p-value	T-statistic
K-Means	Gaussian Mixture	9.13e-05	-4.37
DBSCAN	Gaussian Mixture	6.34e-17	-14.34
Spectral Clustering	Gaussian Mixture	0.33	0.98
Agglomerative Clustering	Gaussian Mixture	0.52	0.64

Figure 6: Paired T-tests on silhouette score for Census data-set

As we can infer from the results, Gaussian Mixture is the statistically significant best algorithm by its silhouette score. The Gaussian Mixture clusters on the Census data-set are shown in Figure 7.

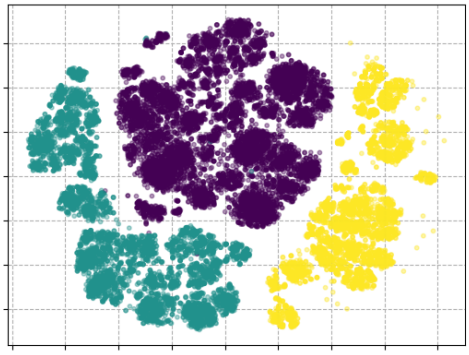

Figure 7: Gaussian Mixture Model Clustering

Thereafter we did the same for the Shoppers data-set. Note that we did not include DBSCAN in this comparison as nearly every ϵ gave us 1 cluster which is not suitable for silhouette and association. The Anova p-value this time was 9.267e-19, which is below 0.01, so we could reject the Anova null hypothesis. Next, we used paired T-tests to find the statistically significant best algorithm, same as we did on the Census data-set.

First Algorithm	Second Algorithm	p-value	T-statistic
K-Means	Gaussian Mixture	8.49e-10	8.09
K-Means	Spectral Clustering	0.28	1.09
Agglomerative Clustering	K-Means	0.0007	-3.67

Figure 8: Paired T-tests on silhouette score results for Shoppers data-set

As we can infer from the results, for the Shoppers data-set, K-Means is the statistically significant best clustering method by silhouette score. The K-Means clusters on the Shoppers data-set are shown in Figure 9.

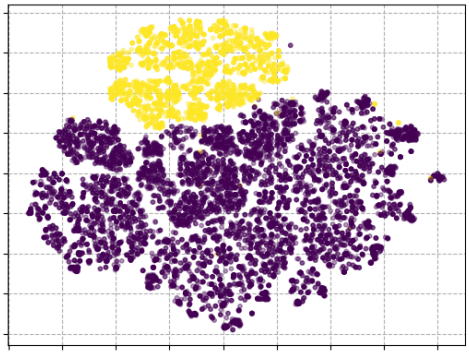

Figure 9: K-Means Clustering

3.3 Associating a clustering method with every external variable

We wanted to find a clustering method which best fits each external variable. for the Census data-set, we attempted to fit the clusters with four external features:
dAge; represents the age of the citizen, 8 categories, *dHispanic*; represents the Hispanic origin of the citizen, 10 categories, *iYearwrk*; represents the latest working year of the citizen, 8 categories, and *iSex*; Represents the sex of the citizen, 2 categories.
For the Online Shoppers data-set, we attempted to fit the clusters with the external feature *Revenue*, which tells whether a purchasing has occurred or not, therefore 2 categories.

We started with the Census data-set. As we did in the previous section, we took 30 random samples, clustered each sample using the five algorithms, and calculated the mutual information between the clustering and the four external features. First, we took the average mutual information of the samples and generated the results shown in Figure 10.

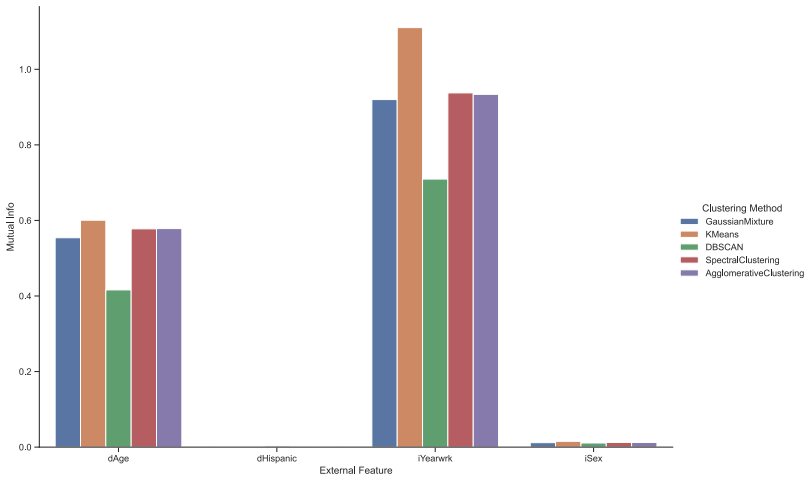

Figure 10: Average mutual information of each external variable

Next, we performed some statistical tests to prove the results from Figure 10 are statistically correct.

First, we explored the *dAge* feature. The Anova p-value for the results of the mutual info between the different clustering methods and the *dAge* feature was 2.2e-11, which meant that the five algorithms have statistically significant different results.
Next, we performed a paired T-test "battle" to find the statistically significant best algorithm by the mutual information of the clustering with the *dAge* feature.

First Algorithm	Second Algorithm	p-value	T-statistic
K-Means	Gaussian Mixture	0.007	2.8
DBSCAN	K-Means	4.57e-06	-5.33
Spectral Clustering	K-Means	7.12e-17	-14.29
Agglomerative Clustering	K-Means	7.44e-15	-12.33

Figure 11: Paired T-tests with dAge on the Census data-set

As we can infer from the results in Figure 11, K-Means is the statistically significant best algorithm by the mutual information of the clustering and the *dAge* feature, confirming the results in Figure 10.

Subsequently, we did the same statistical test for the other external features.
For *iYearwrk*, the Anova p-value was 1.11e-27, which meant that the five algorithms have statistically significant different results. By comparing the algorithms using paired T-tests as before, we inferred that K-Means is the best algorithm by the mutual information of the clustering and the *iYearwrk* feature, matching Figure 10.

As for *iSex*, the Anova p-value was 0.0001, which meant that again the five algorithms have statistically significant different results. Here, we inferred that K-Means is the best algorithm by the mutual information of the clustering and the *iSex* feature, again confirming Figure 10.

Finally, for *dHispanic* The Anova p-value was 3.86e-26, which again meant the five algorithms have statistically significant different results. This time we concluded that DBSCAN is the best algorithm, confirming Figure 10. We summarized the results into Figure 12.

Feature	Most Associated Algorithm	Mutual Information
dAge	K-Means	0.6
iYearwrk	K-Means	1.1
iSex	K-Means	0.015
dHispanic	DBSCAN	0.003

Figure 12: Best clustering method for each external feature

From the results in Figure 12 we deduced that there is an especially high connection between *dAge* and *iYearwrk* to the K-Means clustering.

Finally, we reduced the dimension of the data and visualized the clustering together with the external feature classification. Each color represents a class of the external feature, and the lines go from the point to the center of the cluster produced by the clustering method.

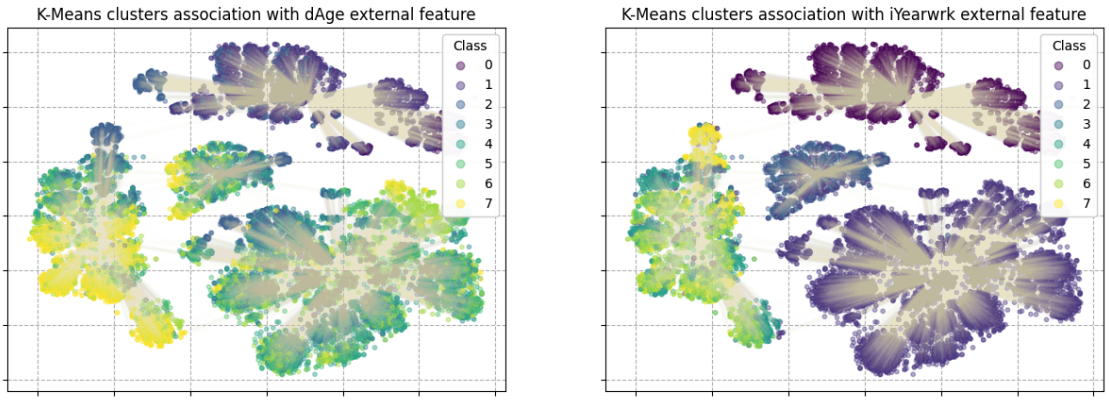

Figure 13: K-Means clusters association with external features

In Figure 13 we can clearly see the strong association between the K-Means clusters and the *dAge*, *iYearwrk* classifications.

Next, we performed similar tests for the Shoppers data-set. In this data-set we had 1 external feature, *Revenue*. So, we preformed the statistical test between all the methods, for the mutual information between the clustering given by the method and the *Revenue* feature. The paired T-tests results between the different clustering algorithms are shown in Figure 14.

First Algorithm	Second Algorithm	p-value	T-statistic
K-Means	Gaussian Mixture	0.002	-3.27
Spectral Clustering	Gaussian Mixture	0.0017	-3.36
Agglomerative Clustering	Gaussian Mixture	0.002	-3.31

Figure 14: Paired T-tests with dAge for Census data-set

We inferred from Figure 14 that Gaussian Mixture is the statistically significant best algorithm by the mutual information of the clustering with the *Revenue* feature. However, the mutual info is very low at 0.004, So we concluded that there is no meaningful association between the clustering and the feature.

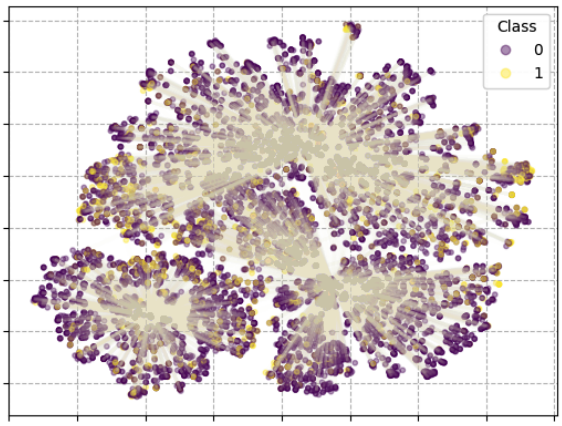

Figure 15: Gaussian Mixture association with with Revenue external feature

From Figure 15 we ascertain that there is little to no relation between the clusters and the classes of the *Revenue* feature, which makes sense as the mutual information is very low.

3.4 Associating an external variable with every clustering method

For each clustering method, we wanted to find the the external feature that fits best to the clustering, meaning, the feature which is statistically significant best by the mutual information with the clustering. In this section we researched only the Census data-set, as the Shoppers data-set has just 1 external feature therefore there is no comparison to carry out.

As before, we took 30 random samples, clustered each sample using the five algorithms, and calculated the mutual information between the clustering and the four external fields. We

took the average mutual information of the samples and found that for all methods *iYearwrk* has the best association, as shown in Figure 16.

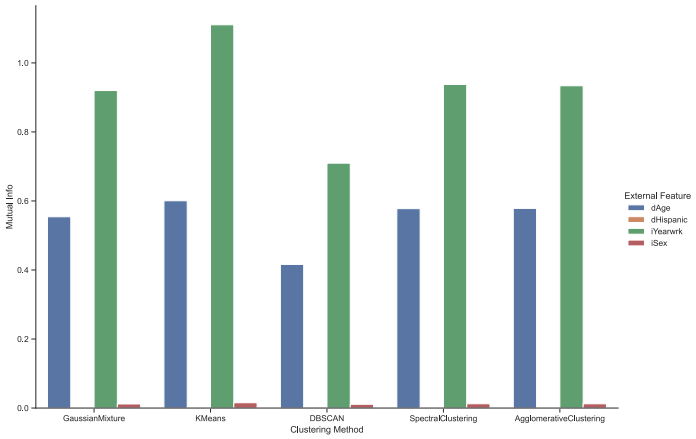

Figure 16: Average mutual information of each clustering method

Next, we performed some statistical tests to prove the statistical significance of the results from Figure 16. In the statistical tests, we could reject the Anova null hypothesis for all tests, and inferred that *iYearwrk* is the statistically significant best feature by the mutual information with the clustering, means, the feature which best fits with the clustering.

3.5 Anomaly Detection

In the final part of the research we wished to find anomalies on the data and any association between them and external variables. For the Census data-set, we used the ϵ value deduced from Figure 3 in section 3.1 to filter the outliers.

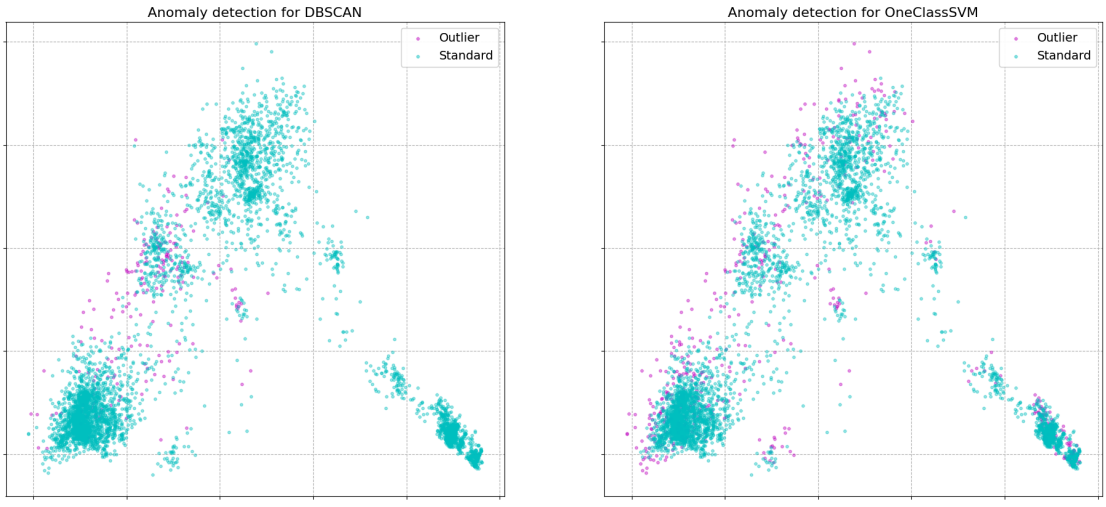

Figure 17: Anomaly Detection for Census data-set using DBSCAN and OneClassSVM

Next, we wanted to associate the anomalies to the external variables. For each method we

calculated the mutual information between the outliers and the external features, and using a paired T-test we found that OneClassSVM had a higher association.

Feature	Mutual Information
dAge	0.0025
iYearwrk	0.0031
iSex	0.0003
dHispanic	0.0067

Figure 18: Mutual information between outliers and external features for OneClassSVM on the Census data-set

We can notice from Figure 18 that the mutual information with all the features is negligible, therefore we deduced there is no distinctive association between anomalies and features.

Finally, we ran the same anomaly detection methods on the Shoppers data-set. The results are shown in Figure 19.

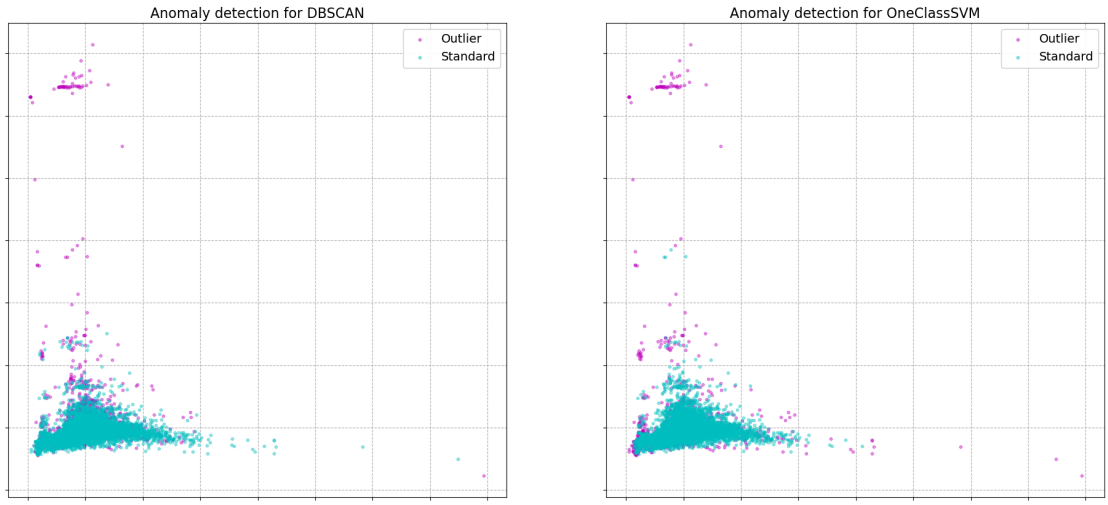

Figure 19: Anomaly Detection for Shoppers data-set using DBSCAN and OneClassSVM

And again, for each method we calculated the mutual info between the outliers and the external feature, and using a paired T-test we found that OneClassSVM had a higher association. However, as before, we notice that the mutual information with the external feature is negligible, therefore there is no distinctive association between anomalies and the feature for the Shoppers data-set as well.

4 Discussion

After examination of the results, we can make some conclusions and smart insight about both data-sets, their structure, and their hidden relations. For the Census data-set we can safely say that there is a strong association between the clusters and the *iYearwrk*, *dAge* features. This conclusion is very interesting as it proves the possibility of Unsupervised Learning techniques being useful and informative.

As for the Shoppers data-set we concluded that there no association between the clusters and the *Revenue* external variable. This could be the result of some of the limitations stated below, or because the methods we used simply aren't suitable for this data. Therefore, for the Shoppers data-set it might be a good idea to use deep data models in order to understand complex relations in the data.

From these results we can understand that every data-set is different. Some data-sets can be easily clustered by certain methods, some by other methods, and some cannot be clustered at all. Both data-sets suffer from limitations and weakness which make clustering the data a harder task. An example of a limitation is both of the data-sets having a large amount of anomalies. In addition, the census data-set is very large - clustering all the rows in one go can take a very long time, thus we had to take samples of the data and verify the results using statistical tests.

In closing, we think a lot had been learned and can be learnt from this research, and generally in this field. Unsupervised Learning is stronger then it seems, and this is just the beginning.

References

[Ala20] Mahbubul Alam. Support vector machine (svm) for anomaly detection. 2020.

[LvdM08] Geoffrey Hinton Laurens van der Maaten. Visualizing data using t-sne. 2008.

[Mah21] Md Sohel Mahmood. Factor analysis of mixed data. 2021.

[Mul20] Tara Mullin. Dbscan parameter estimation using python. 2020.

[San17] Sergio Santoyo. A brief overview of outlier detection techniques. 2017.

[Ver21] Yugesh Verma. How to use support vector machines for one-class classification? 2021.