Tabular Data Science Final Project - **Outliers Detection**

Michael Teranovsky
Roei Ben Zeev

20 March 2022

1 Introduction

The problem we aimed to solve is Outliers detections. Our solution is to detect those outliers that far by 2 std from mean/median.We had great results in our experiment for data sets that have a high number of numerical features.

2 Problem Description

In data analysis, anomaly detection (also referred to as outlier detection and sometimes as novelty detection) is generally understood to be the identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well defined notion of normal behaviour.Such examples may arouse suspicions of being generated by a different mechanism, or appear inconsistent with the remainder of that set of data.

Outlier detection is often an important step in data pre-processing to provide the learning algorithm a proper dataset to learn on. This is also known as Data cleansing. After detecting anomalous samples classifiers remove them, however, at times corrupted data can still provide useful samples for learning. A common method for finding appropriate samples to use is identifying Noisy data.

3 Solution overview

Sometimes a dataset can contain extreme values that are outside the range of what is expected and unlike the other data. These are called outliers and often machine learning modeling and model skill in general can be improved by understanding and even removing these outlier values.
Our goal is to identify and remove the outliers from your machine learning

dataset.
We would like to automate the data pipeline and make a pre-process mechanism to handle the outliers and notify the user that those samples can affect the model result and distribution.
We will detect the outliers with two different methods and just remove the outliers from the data set.
the methods are:

- All object the far 2 or more times of Standard deviation then the mean.
- All object the far 2 or more times of Standard deviation then the median.

4 Experimental plan

We will try different methods to identify the outliers and check each method to find which method gives us the best results.
We will determine our results based on comparison between the model result with the outliers samples and the results after cleaning the outliers with the different methods.

We will split our data into a training set and test set. on the same train set, we will run the different methods to remove outliers, and run the product on the test set. We do the same with the full training set with the outliers and determine which technique returns the best results.
We will compare the result of the experiment to sklearn library outliers handling function (e.g IsolationForest, EllipticEnvelope, LocalOutlierFactor) and determine if our approach may give us a better result.

5 Experimental evaluation

As mentioned in the paragraph above, we compared the results by running the same model on the data set after we remove the outliers by out 2 methods, and running the model on different sklearn outliers detection methods. we notice that for data set with high amount of numerical features we achieved an outstanding results and better results with out method as we can see in comparison we made on the house pricing dataset that have 21 very numerical features.

	Method	Model's Performances
1	STD Median	0.978315
0	STD Mean	0.977289
2	EllipticEnvelopen	0.912819
5	None	0.911485
3	LocalOutlierFactor	0.909727
4	IsolationFactor	0.908610

As we can see above the model with our 2 methods achieved significantly better result then other sklearn outlier detection methods and even without any outlier detection method.

Also if we will look at the images we added in the project folder, we can see the different in the feature distribution. we can notice clearly the removing of the outliers and by that getting more accurate data with less "noisy" data that can affect out data set. we can see that most of data from edged has deleted and we get most of the data around the "central" of our data set.

Now lets see what happen in data set with a small amount of numerical features. As we can see in the car prices data set. our methods give us a better result then not using outlier removals but other sklearn methods give us a better results.

	Method	Model's Performances
4	IsolationFactor	0.965940
3	LocalOutlierFactor	0.826687
2	EllipticEnvelopen	0.817243
1	STD Median	0.767888
0	STD Mean	0.766859
5	None	0.743349

6 Related work

According to this article we can see an explanation about Global outliers. A data point is considered a global outlier if its value is far outside the entirety of the data set in which it is found:
"A global outlier is a measured sample point that has a very high or a very low value relative to all the values in a dataset. For example, if 99 out of 100 points have values between 300 and 400, but the 100th point has a value of 750, the 100th point may be a global outlier."

According to this, the outliers we detect in our 2 methods are considered as global outliers. we measure out outliers to be far by 2 std from mean/media of the feature.
The article explain the idea of global outliers with visualization example. We improve the technique to automated mechanism to detect the outliers and remove them.

7 Conclusion

For colclusion, we notice that in case that we have high amount of numerical features the method improve the model results and also as we see in the histograms at the added images, we can notice the removal of the outliers and improved in the feature data distribution and even gave us better results then Sklearn outliers detection methods. In data sets with a low amount of numerical feature we notice the our method didn't affect the model results much then not using it but some other sklearn outlier detection methods gave us better results. So according to the data above we can conclude that out methods gave us a very good results and did not harm the results adversely.

We had a great time implementing this project idea, we learned a lot out the affection of pre processing the data set and how it can improve our result. also we saw how it can improve our data distribution.
Thanks for the semester Amit and we wishing you all the best.