Outliers and its impact

on the machine learning models

An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.

Let's take an example, we do customer profiling and find out that the average annual income of customers is $0.8 million. But, there are two customers having an annual income of $4 and $4.2 million. These two customers' annual income is much higher than the rest of the population. These two observations will be seen as Outliers.

What are the types of Outliers?

An outlier can be of two types: ***Univariate*** and ***Multivariate***. Above, we have discussed the example of a univariate outlier. These outliers can be found when we look at the distribution of a single variable. Multi-variate outliers are outliers in an n-dimensional space. In order to find them, you have to look at distributions in multi-dimensions.

What causes Outliers?

Whenever we come across outliers, the ideal way to tackle them is to find out the reason for having these outliers. The method to deal with them would then depend on the reason for their occurrence. Causes of outliers can be classified into two broad categories:

1. Artificial (Error) / Non-natural

2. Natural.

Let’s understand various types of outliers in more detail:

- **Data Entry Errors**: - Human errors such as errors caused during data collection, recording, or entry can cause outliers in data. For example, the annual income of a customer is $100,000. Accidentally, the data entry operator puts an additional zero in the figure. Now the income becomes $1,000,000 which is 10 times higher. Evidently, this will be the outlier value when compared with the rest of the population.

- **Measurement Error**: It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty. For example, There are 10 weighing machines. 9 of them are correct, 1 is faulty. Weight measured by people on the faulty machine will be higher / lower than the rest of the people in the group. The weights measured on the faulty machine can lead to outliers.

- **Experimental Error**: Another cause of outliers is experimental error. For example: In a 100m sprint of 7 runners, one runner missed out on concentrating on the ‘Go’ call which caused him to start late. Hence, this caused the runner’s run time to be more than other runners. His total run time can be an outlier.

- **Intentional Outlier**: This is commonly found in self-reported measures that involve sensitive data. For example, Teens would typically under-report the amount of alcohol that they consume. Only a fraction of them would report actual value. Here actual values might look like outliers because the rest of the teens are under-reporting the consumption.

- **Data Processing Error**: Whenever we perform data mining, we extract data from multiple sources. It is possible that some manipulation or extraction errors may lead to outliers in the dataset.

- **Sampling error**: For instance, we have to measure the height of athletes. By mistake, we include a few basketball players in the sample. This inclusion is likely to cause outliers in the dataset.

- **Natural Outlier**: When an outlier is not artificial (due to error), it is a natural outlier. For instance: In my last assignment with one of the renowned insurance companies, I noticed that the performance of the top 50 financial advisors was far higher than the rest of the population. Surprisingly, it was not due to an error. Hence, whenever we perform any data mining activity with advisors, we used to treat this segment separately.

Impacts having an outlier in the dataset:

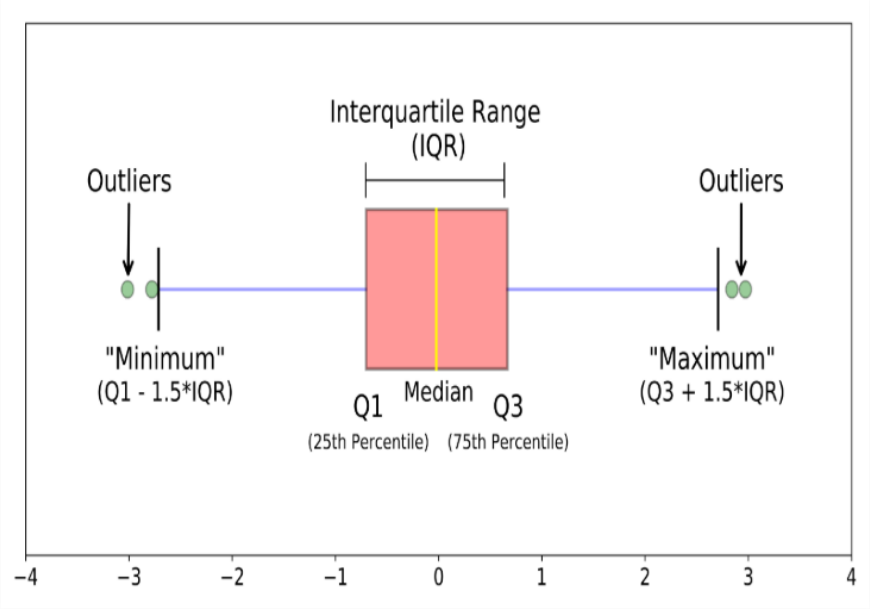
1. It causes various problems during our statistical analysis.

2. It may cause a significant impact on the mean and the standard deviation

Various ways of finding the outlier.

1. Using scatter plots. (we can visualize outlier present in dataset using scatter plot)

2. Box plot.



1. using a z score (given data point is an outlier if the z score is greater than 3).

2. using the IQR interquartile range (if the data point is not lies between [Q1 -1.5*IQR, Q3 + 1.5*IQR]).

ipython notebook for finding the outlier implementation:

https://github.com/sunilkumarcheruku/MachineLearningAlgorithms/blob/master/Identifying_Outlier.ipynb

How to remove Outliers?

Most of the ways to deal with outliers are similar to the methods of missing values like deleting observations, transforming them, binning them, treat them as a separate group, imputing values and other statistical methods. Here, we will discuss the common techniques used to deal with outliers:

Deleting observations: We delete outlier values if it is due to a data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.

Transforming and binning values: Transforming variables can also eliminate outliers. The natural log of value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows dealing with outliers well due to the binning of the variable. We can also use the process of assigning weights to different observations.

Imputing: Like imputing missing values, we can also impute outliers. We can use mean, median, mode imputation methods. Before imputing values, we should analyze if it is a natural outlier or artificial. If it is artificial, we can go with imputing values. We can also use the statistical model to predict values of outlier observation and after that, we can impute it with predicted values.

Treat separately: If there are a significant number of outliers, we should treat them separately in the statistical model. One of the approaches is to treat both groups as two different groups and build individual models for both groups and then combine the output.

Impact of Outlier on KNN algorithm:

It depends on your implementation of KNN but it can have an impact on your error. If you're using KNN where K=1 then you're telling your model to only find the training example that is closest to the point you're searching for and return its class. If you use K>1 you're telling it that you want to find the closest K training examples and then do a majority vote with those examples. Using K>1 will smooth out your decision boundaries and, assuming there isn't a clump of outliers, negate any impact that outliers will have on your predictions.

As long as K>1 and there aren't a cluster of outliers in your data then you have nothing to worry about since KNN's majority vote will negate the effects of outliers.

Impact of Outlier on Naive Bayes:

There are different flavors of Naive Bayes, so the answer depends a bit on the use case.

One potential issue with outliers is that unseen observations can lead to 0 probabilities. For example, Bernoulli Naive Bayes applied to word features will always produce 0 probabilities when it encounters a word that wasn't seen in the training data. Outliers in this sense can be a problem. However, all these and similar issues of Naive Bayes have well-known solutions (like Laplace smoothing, i.e. adding an artificial count for every word) and are routinely implemented.

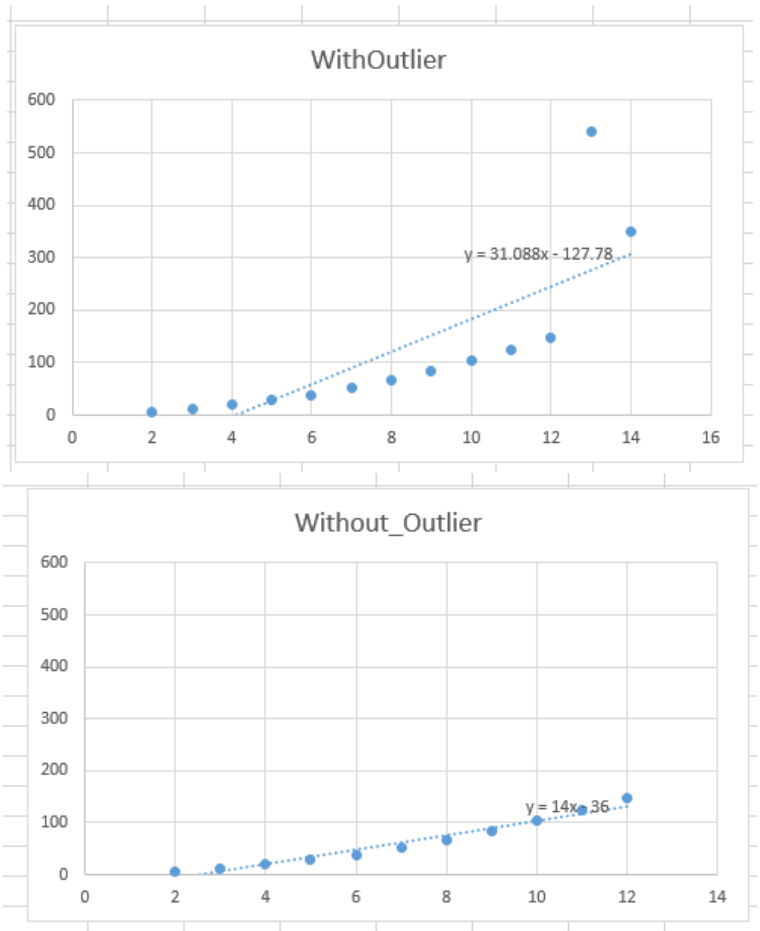
In Gaussian Naive Bayes, outliers will affect the shape of the Gaussian distribution and have the usual effects on the mean, etc.

So depending on your use case, it still makes sense to remove outliers.

Impact of Outlier on Linear Regression:

Outliers can have a dramatic impact on linear regression. It can change the model equation completely i.e. bad prediction or estimation. Look at the scatter plot and linear equation with or without the outlier.

Look at both snapshots, equation parameters changing a lot.



Impact of Outlier on Logistic Regression:

Logistic Regression performance not getting affected by Outlier, due to it can handle outlier using the sigmoid function.

Impact of Outlier on SVM:

Outliers have the capability to make your model poor. The margin will shrink and the decision boundary will be sub-optimal resulting in poor classification.

In the presence of outliers, you need to use a more general version of the Support Vector Machine that is with soft margins.

Impact of Outlier on Decision Trees:

Due to outliers, the depth of the decision tree increases and the model will get over fitted.

Ensemble models don't have a problem with Outliers due to sampling and aggregation.

we can neglect outliers using min_samples_split and min_samples_leaf