Paper 0372-2017

Outline Outliers: Adding a Business Sense

Alex Glushkovsky, BMO Financial Group

ABSTRACT

Outliers, such as unusual, violated, unexpected or rare events, have been intensively in focus by researchers and practitioners providing their impacts on estimated statistics and developed models. Today, some business disciplines are focusing primarily on outliers such as defaults of credit, operational risks, quality nonconformities, fraud, or even the results of marketing initiatives in highly competitive environments with low response rates of a couple percent or even less. This paper discusses the importance of detecting, isolating, and categorizing business outliers to discover their root causes and to monitor them dynamically. Addressing not only extreme values or multivariable densities detecting outliers, but also addressing distributions, patterns, clusters, combinations of items, and sequences of events will allow for opportunities to be established for business improvement. SAS® Enterprise Miner™ can be used to perform such detections. Thus, creating special business segments or running specialized outlier oriented data mining processes, such as decision trees, allows for isolation of business important outliers, which normally will be masked in traditional statistical techniques. This process combined with "What-If" scenario generation prepares businesses for future possible surges even when having no current specific type of outliers. Furthermore, analyzing some specific outliers may play a role in assessing business stability to corresponding stress tests.

INTRODUCTION

Outliers are all around us. They exist not only in data or analytics, but they are present in our everyday life. For example, pedestrians crossing the street on a red light. Actually, that may be a wrong example for outliers as sometimes they represent the majority. Anyway, some of our activities are based on outliers – just think about arts or sports. Of course, business is not an exception. Some business disciplines are built entirely around outliers, such as the lottery or casinos, and focus primarily on outliers. For example, defaults of credit, operational risk events, fraud, reliability issues, or even the results of marketing initiatives in highly competitive environments with low response rates of only a couple of percent. Outliers, such as unusual, violated, unexpected or rare events, have been extensively investigated by researchers and practitioners providing their impacts on estimated statistics and developed models.

Today, many publications can be found discussing outlier detection. Thus, (Ben-Gal, 2005) provides a compelling overview concerning univariate and multivariate methods, parametric and non-parametric approaches of outliers detection, robust measures, single-step and sequential procedures, Statistical Process Control (SPC), both classical and based on ARIMA models, as well as data mining, such as distance-based, Replicator Neural Networks (RNN) and clustering. His publication includes a broad list of references concerning outlier detection. Three major methods detecting outliers are discussed in (Mamdouh, 2010): univariate, regression models, and clustering.

Research on sample size and decision criteria to detect outliers is discussed in (Cousineau and Chartier, 2010). Also, the paper addresses non-linear transformation approaches, recursive and non-recursive methods, multiple and nonlinear regressions, as well as, treatments of outliers, such as replacement by mean or by other possible values.

The topic has been presented in previous SAS® Forums. Thus, practical methods of outlier detection are presented in (Polfliet, 2016).

Control of rare events in health care by applying Statistical Process Control (SPC) on a number of normal events between two consequent accidental events or on time periods between two consequent accidental events is discussed in (Kaminski, 1992; Glushkovsky, 1994; Ransdell, 2016).

Commonly, there are conflicting requirements to have sufficient sample sizes to build robust statistical analysis or models and extremely limited information concerning outliers. Even though, the modern “big data” trend of storing a lot of data means that sample sizes of outliers are not too small anymore, but dealing with outliers is still very challenging, starting with their identification, and requires business judgement combined with analytical approaches. Practically, SAS® Enterprise Guide® and SAS® Enterprise Miner™ can be used to perform outlier detections.

The paper discusses the importance to not ignore or automatically filter or replace outliers, but to detect, to isolate, and to classify business sound outliers, to discover their predictive factors (i.e., root causes), and to monitor them dynamically.

The article is not a comprehensive overview or introduction of some methodologies to detect outliers; rather it is a view on possible analytical approaches toward outliers and their roles in business improvement.

CLASSIFICATION OF OUTLIERS

There are a number of definitions of outliers (Weisberg, 1985; Barnet, 1994). Traditionally we can describe outliers by the following not mutually exclusive terms:

- Rare
- Unusual
- Uncommon
- Unexpected
- Violated
- Error

Outliers can be identified by their extreme values, by their low frequencies, or by rare events. Thus, a rare outlier can be revealed to a low frequency or to a low probability density, unusual or uncommon ones can be associated to a shortfall or a large distance value, an unexpected outlier can be related to a large residual of the model or to a significant error of the estimation, a violated outlier can be judged against the defined specifications, and an error can be caused by a typo or inconsistencies, mostly logical.

For example, the tail of the distribution usually has a low probability density (infrequent) and the observations there possess extreme values (unusual). This is quite a common dependence.

Different, not mutually exclusive, and “fuzzy” definitions mean difficulties identifying outliers using different analytical techniques, shortfalls understanding their triggers and causes, and lack of monitoring. Nevertheless, some generalized view can be presented covering different definitions of outliers.

Focusing on business sense, two mutually excluding categories of outliers can be defined:

- Inherent
- Error

An inherent outlier means that it possesses a distinctly different property from the rest of the elements (Hawkins, 1980) but can be present due to the nature of the business. Therefore, it is “organic” for the business. In contrast, any error can be recognized as an outlier and it is not acceptable and ideally should be fixed and prevented from occurring again in the future. Inherent outliers may have a positive or negative impact on a business. In the article, we will focus on outliers that create opportunities for a business improvement leaving outliers with neutral impacts on a business out of scope.

In business, one of the major goals is to implement improvement changes, i.e., boosting positive variances while blocking and preventing negative ones (Shewhart and Deming, 2012). The fundamental

idea of Statistical Process Control (SPC) is a distinction between assignable and common variances (Juran and Godfrey, 1999; Duncan, 1986; Montgomery, 2005, Woodall, 1986). Identified distinction between common and assignable variances allows for initiation of improvement programs. Looking at outliers from such a prism allows us to view them as assignable variances. It means that out-of-control signals dynamically identify outliers with respect to historical observations. Furthermore, as an example, it means that extreme values are not necessarily outliers if they are not assignable variances.

The link between SPC and outliers is not a new paradigm and has already been discussed (Ben-Gal, 2005).

A list of possible outliers in business may include traditional objects such as variable values, including missing ones, events, records, clusters, and samples as well as non-traditional objects, such as:

- Distributions
- Patterns
- Combinations of items (baskets)
- Sequences of events
- Models

Addressing not just extreme values or multivariable densities detecting outliers, but objects such as those listed above that are conditioned to be assignable variances, will open more opportunities for business improvements.

Moreover, assignable variances may even include the following examples:

- Optimization points of the objective functions. For example, pricing that maximizes profit
- Extraordinary features of products, services, or processes
- Non-dominated strategies that form Nash equilibrium
- Efficient frontier, which is a set of outliers

The critical element detecting inherent assignable outliers is the clear business sense accompanied by a significance to deviate from core elements possessing common variances.

In general, adding business sense to outliers and focusing on improvements always requires two relationships to be discovered: (1) between outliers and their root causes, and (2) between outliers and business sound target variables, i.e., Key Performance Indicators (KPIs). The latest should be defined based on business objectives. Essentially, business sound outliers either impact target variables directly or influence drivers of the target variable. Moreover, assignable outliers usually have leveraged effects on business results. Considering the sample sizes of the outliers, that chain of relationships, in most cases, is a very challenging task to be discovered.

IMPROVEMENT CYCLE DRIVEN BY OUTLIERS

The improvement cycle based on outliers includes five major elements (see Fig. 1) and it is quite similar to the classical Deming - Shewhart's continual improvement PDSA cycle (Deming, 2010; Shewhart and Deming, 2012).

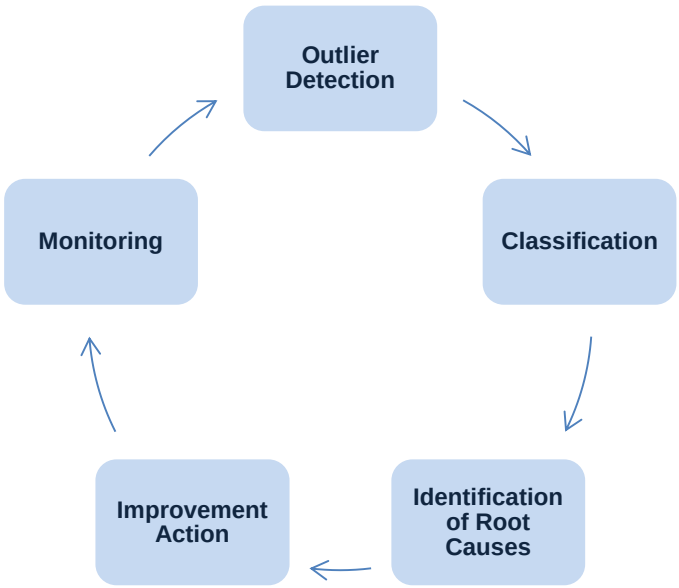

Figure 1. Outlier Based Business Improvement Cycle

Logically, the first two steps of the cycle are detection of assignable outliers and the classification of them as inherent ones or errors. The detection step can be done deterministically using business knowledge for obvious cases or analytically while still applying sound business judgement. Assignable outliers can then be analyzed further to understand their underlying causes which may lead to improvements. This can be done by applying root-cause analysis supported by data mining or machine learning techniques focusing on possible drivers.

Generally, root causes can be triggered by changes in macroeconomic or market environments, competitor initiatives, regulations, or inner business matters.

By analogy to Capability Maturity Model (CMM$^{(SM)}$) (Paulk et al, 1993), levels of business interaction with outliers can be defined as:

- Ignorance (pay no attention, no detections, no isolations)
- Detection (ad hock removal or replacement)
- Control (implementation of special tools such as SPC)
- Management (establishment of improvement actions)
- Optimization (systematic and integrative improvement processes toward a business goal in a constrained environment)

Common mechanisms dealing with outliers today are simple observations during reporting or their removal during data preparations for modeling (level I and II), and data quality management routines mostly during data inputs and loadings (levels II - IV). Higher levels of business interaction with outliers require an enterprise wide involvement.

The most mature level V of dealing with outliers perhaps can be observed in . . . nature (see Figure 2). The climbing plant actually applies the continues improvement strategy by spreading outliers in different directions, checking the outcomes, and then selecting the most beneficial one for future growth.

Figure 2. Utilization of "Outliers" Is a Natural Growing Mechanism of the Climbing Plant.

OUTLIERS AND ANALYTICAL APPROACHES

Common analytical objectives are to estimate statistics or parameters, to discover relationships, to score, to predict, to detect common patterns, to discriminate, or to classify. All those listed above can be affected by outliers. From another side, analytical approaches can be used to detect outliers and their root-causes.

Here is a list of some analytical approaches, which may be helpful in detecting outliers:

- Associations
- Clustering
- SPC
- Neural networks
- Regressions
- Decision trees

The last three models may be used for root-cause identifications as well. However, the neural networks approach is not transparent since it consists of a lot of elementary models and, therefore, the obtained results cannot be drilled down by variables or segments getting more insights on business issues.

OUTLIERS AND MODELS

There are three-way relationships between models and outliers as presented in Figure 3.

Modeling is a well-known and an effective mechanism to detect outliers (Weisberg, 1985; Mamdouh , 2010). On the another hand, model quality itself can be impacted by outliers. Research on the influence of outliers on the quality of predictive classifying models, including decision trees, and comparison to other models, such as kNN, Naïve Bayes, and logistic regression, is described in (Kalisch et al, 2016).

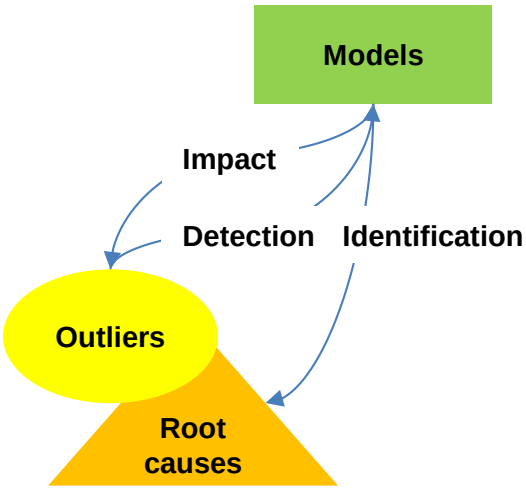

Figure 3. Relationships between Outliers and Models

The impact of outliers on a model can arise during training as well as during the usage of the trained model. Both effects should be in focus. The latest can be handled by setting a monitoring process of the implemented model quality and its inputs. It will allow for detection of changes due to an outlier's presence.

Outliers can exist in target variables as well as in potentially predictive inputs, or both.

Outliers in the target variable should always be specially treated since they represent direct business cases. Of course, the presence of outliers in the binary target variable cannot be detected directly. Outliers of interval, nominal, or ordinal target variables can be handled by applying simple segmentation to isolate them based on business judgement or univariate statistical techniques. After that, this segment (or segments) can be part of further model training processes. Performance as well as the sample size of this segment should be monitored periodically.

Outliers in input variables impact model training as well as model implementation. It should be noted that different types of models have different sensitivity to this issue when regressions are extremely volatile. The popular suggestion is filtration or replacement of outliers to mitigate these impacts.

Filter node in SAS® Enterprise Miner™ can be used to perform outlier removal of class or interval variables using univariate parametric and non-parametric methods. This approach can be modified not to filter but to flag outliers in the dataset instead.

If outliers are filtered from the training data set, then it will improve the robustness of the model, such as regression. However, if similar outliers to the excluded outliers occur again during the implementation stage, then the model becomes not applicable in that case.

If outliers have been replaced by some arbitrary values during modeling, such as means or more sophisticated multivariable estimates, then it will require the same artificial process to be applied for a new data set so that results will not reflect the nature of the business.

The simple yet effective approach when dealing with outliers in input variables is not to remove or replace them, but to apply robust types of models, which are less sensitive to outliers, and/or to isolate them by creating special segments.

It should be noted that a model can be an outlier object by itself. Comparing different models by the "Model Comparison" node in SAS® Enterprise Miner™, it may be observed (see Figure 4) that one model significantly over performs other models or has a much higher lift on a part of the lift chart, which represents an important area from the business perspective (usually an initial one).

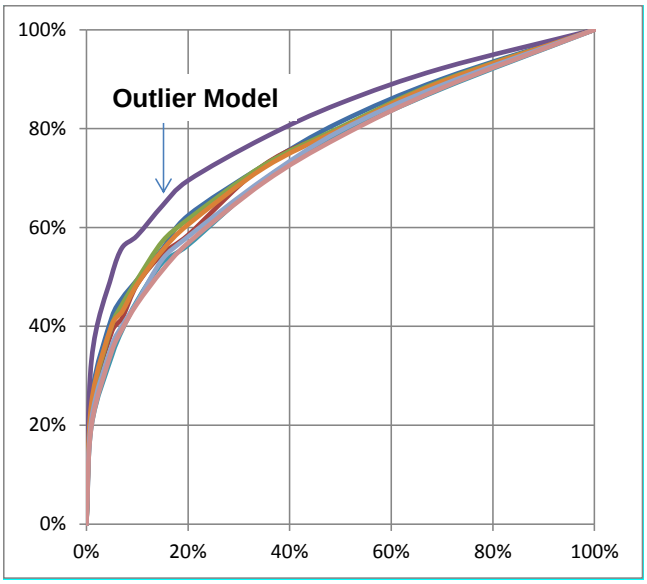

Figure 4. Lift Charts Comparing Different Models of the Same Problem.

Of course, this observation should be assured against an overfitting effect. If the latter is not the case, then it is important to understand the origin of such over performance: “Is it due to data sampling potentially reflecting overfitting?”, or “Is it due to the existence of outliers in input variables?”, or “Is it due to the survived variables and their transformation or grouping?”, or “Is it due to the applied methodology?”, or “Is its due to specified model parameters?”. Understanding the origin of the model’s over-performance may even lead to further improvement to model performance and robustness.

IDENTIFICATION OF CAUSES

After outlier detection, the next immediate question is “What is the cause for it”? In some cases, it may be an obvious answer considering the nature of the outlier. However, in many other cases, the modeling exercise can be helpful against target identifying outliers. It can be done using transparent models such as decision trees or logistic regression for binary target variables. Of course, these exercises will be successful only dealing with assignable outliers and having data on relevant predictive variables.

The key issue here is to prepare potentially predictive variables on the same level as the outliers. For example, if outliers are on an elementary record level, then potentially predictive variables should be compiled on this level as well. However, if it is a segment level outlier, then input variables should be organized on a relevant segment level (preferred) or be on a record level populating associated flags of the outlier segment across all records.

Practically, outlier root-cause modeling can be performed in SAS® Enterprise Guide® using the Rapid Predictive Modeler (RPM) feature with Intermediate method or in SAS® Enterprise Miner™ using Decision Tree, Regression or Credit Scoring (Interactive Grouping and Scorecard) nodes. Selecting different models, interactions, non-linear and monotonic effects should be considered. All mentioned models provide good transparency reviewing results.

Having success identifying root causes of outliers will open up the opportunity for guided business improvements.

OUTLIERS AND SAMPLE SIZE

The larger the sample size, the more outliers may be inside the sample. Small sample sizes may have zero presence of outliers. Similarly, the longer the observation period, the more unusual outliers can be seen. This effect can challenge the analysis of marketing campaigns with an unbalanced random split

between treatment and control groups especially having low or moderate effects (Glushkovsky and Fabian, 2017).

When detecting outliers, the sample size matters. Adjustments of criteria detecting outliers according to the sample sizes is presented in (Cousineau and Chartier, 2010).

Today we observe a “big data” boom. “Big data” means larger sample sizes for outliers and a more versatile mix of different types of outliers. It means that “big data” brings even more importance to detect, control, and manage outliers toward business improvements.

OUTLIERS AND DISTRIBUTIONS

Dealing with samples at different time periods or segments of data based on specified conditions, it is possible to observe uncommon distributions as shown in the example (see Figure 5,a). In this case, the values of all samples are within the same range (-3;+3), but the distribution of the variable is different for one sample. This means that one distribution in that example is an outlier. In fact, some outliers of distributions can be easily detected by statistics such as mean, medium, or standard deviation. In the example presented in Figure 5, a, the first two statistics will work well for that matter. However, in some cases, such as those presented in Figure 5, b, more sophisticated criteria should be applied. In this example, the outlier distribution cannot be detected by just considering the means, mediums, or standard deviations. It requires more sophisticated tests such as a two-sample Kolmogorov-Smirnov one to detect overall cumulative distribution differences or the application of a simple comparison of percentiles to detect cumulative distribution differences of specific regions (https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test). Possible detection techniques here can be pairwise comparisons or more quick detections against average distribution (Glushkovsky and Billard, 1998).

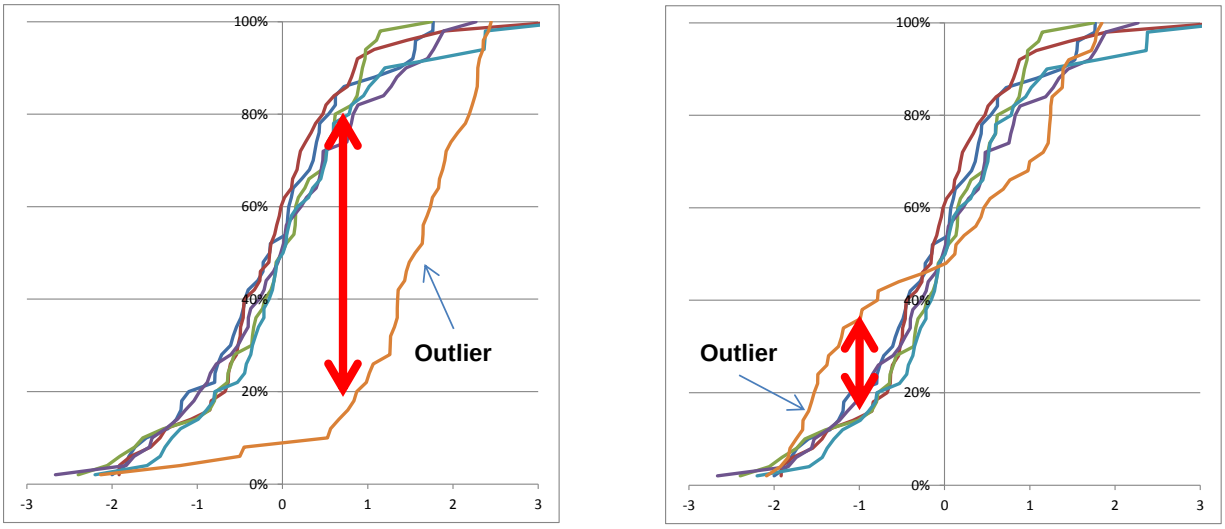

Figure 5. Empirical Cumulative Distributions with Example of Outliers:

a) Significant Shift in Mean and Medium;

b) Bimodal Distribution

Detecting the outlier of distribution of the latest type (bi- or multi-modal), means that the underlying process has assignable cause(s) generating high concentrations around two regions. Identification of the cause of that abnormal behaviour may lead to business improvement and consistency.

OUTLIERS AND PATTERNS

Outliers of patterns can be recognized by observing multi-dimensional objects, such as images, heat maps, or eventually just underlying data organized as multi-dimensional arrays, which are grouped by another dimension(s). Usually, these objects are two-dimensional and grouped by periods and/or other dimension(s) of interest. Detection of outliers in contingency tables is presented in (Rapallo, 2011).

Detecting outliers in patterns utilizing SPC approach has been discussed in (Glushkovsky and Billard, 1998). To find a pattern outlier in this case, a simple SPC technique for a Pearson correlation coefficient can be applied. Each pattern is analyzed against an average pattern or a master one. It reduces calculations dramatically from $O(N^2)$ to $O(N)$. Example in Figure 6 shows how this approach detects an outlier pattern. Obviously, the pattern number 4 is an outlier with correlation coefficient of -0.92 to an average pattern compared to about 0.97 of the others.

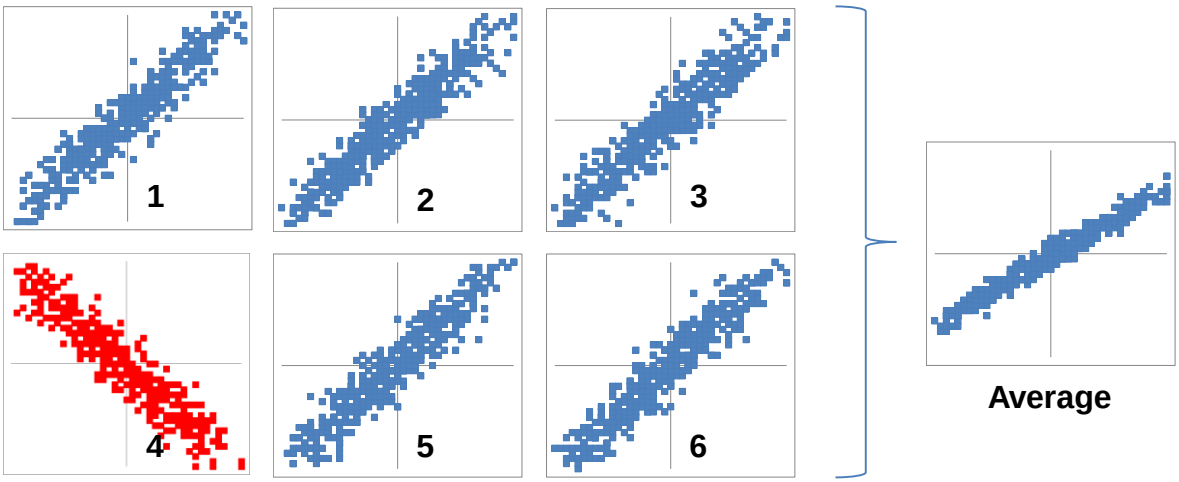

Figure 6. Example of Sequential Patterns Control.

The same concept can be applied by using other criteria depending on the type of cells (or pixels). Thus, for binary cells, the Jaccard index can be considered.

Detecting a pattern outlier that triggered an out-of-control signal, indicates that further analysis should be done focusing on business implications of that pattern and a specific point on the grouping dimension(s).

OUTLIERS AND CLUSTERING

Detection of outliers using clustering is one of the popular choices providing multi variable inputs (Ben-Gal, 2005, Mamdouh, 2010). In this case, clusters with small sizes, as small as one observation, are recognized as outliers. However, this method may separate clusters that have no influence on the target variable. From a business perspective, these clusters of outliers are still in question: Is the estimated average values of the target variable of the outlier cluster different from the rest of the observations or to the nearest clusters? It should be noted that business aspect can be covered directly by applying the supervised decision tree models, which include target variable.

Figure 7 presents an example of adding business sound estimations of KPIs, such as average profitability, to clusters. Thus, considering clusters with small sample sizes, two outliers can be detected. Outlier I has high profitability and especially significantly higher profitability compared to the neighboring clusters. In contrast, outlier II has similar profitability to the entire population and to the neighboring cluster. It means that outlier I may possess an important business case and it is subject for further analysis.

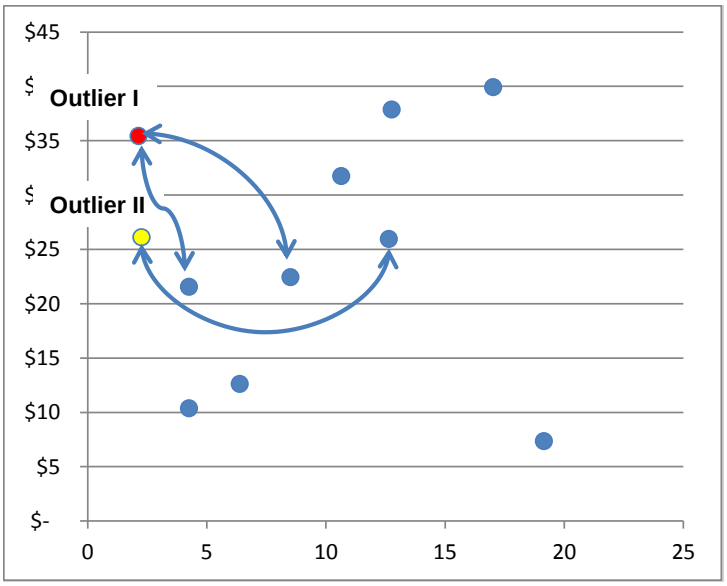

Figure 7. Scatterplot of Clusters Profitability Versus their Sample Sizes (%), where connectors link some neighbouring clusters:

- **Outlier I cluster has a small sample size and significant difference of the target variable to the neighboring clusters;**
- **Outlier II cluster has a small sample size but similar value of the target variable to the neighboring cluster.**

OUTLIERS AND ASSOCIATION RULES

Detection of outliers in transactional data using association rules has been discussed in (Dash and Lie, 2010).

Outliers of events (or transactions) can be associated with the occurrence of a single type of event or in combination with some types of events or with their sequence:

1. Rare type of event occurs (rare item in the basket)
2. Rare association of events per customer or session (combinations of items in the basket)
3. Rare sequence of events per customer or session

In order to detect type (1) outliers, a simple condition can be applied for the proportion to be less than a certain small threshold level, for example, 1%. Observing the distribution of categories of events will support a reasonable setup of the threshold level.

Both the association (2) and sequential (3) type of outliers can be detected by running Association node in SAS® Enterprise Miner™. For that intent, support levels should be set to very small levels.

All detected types of outliers and, therefore, generated rules for types (2) and (3) should be accompanied by estimated target values. For example, let us consider that there is the following outcome concerning association rules and their estimated revenues as presented in Figure 7. In this example however, the dots represent association rules and the X-axis actually presents support for each rule. Of course, a very small support of the rule and high (or low) revenue characterizes a potential business sound outlier. However, in most cases, it is not so obvious and requires a lift analysis of the target variable compared to the neighboring rules. Depending on the business case, neighboring rules can be found by matching left or right hand sides or even just by elements of the rule.

OUTLIERS AND REGRESSION MODELS

Regression models exposed to all three relationships as shown in Figure 3.

Impact of outliers on regression models has been discussed extensively in the past providing their high sensitivity which may produce wrong models in the case of outliers (Stevens, 1983).

On the other hand, trained regression models can be used to detect outliers by observing residuals with significant deviations (Mamdouh, 2010).

Logistic regression models can be helpful to identify root causes. For example, the scorecard model, which is part of the SAS® Enterprise Miner™ Credit Scoring package, can be used for more than just training a model for a binary target variable. This model includes a grouping stage with the transformation of variables to Weight of Evidence (WOE) and the following logistic regression. Example of the results is presented in Table 1 and it includes the list of survived variables, their groupings, and calculated statistics.

		Group	Scorecard Points	Weight of Evidence	Event Rate TARGET_BIN	Percentage of Population	Coefficient
D4	0	1	0	-0.54	21.252	30.64	-0.09
	1, _MISSING_, _UNKNOWN_	2	2	0.28	11.425	69.36	-0.09
D16	1	1	3	-0.49	19.565	28.19	-0.14
	0, _MISSING_, _UNKNOWN_	2	0	0.2	3.384	71.81	-0.14
V17	V17< 2	1	-2	0.31	5.823	8.11	-0.42
	2<= V17	2	-2	0.25	12.418	7.17	-0.42
	MISSING	3	2	-0.05	15.673	84.72	-0.42
V12	V12< 2	1	0	0.26	4.474	9.72	-0.18
	2<= V12< 3	2	-1	0.4	10.948	5.55	-0.18
	3<= V12< 4	3	-2	0.53	12.437	3.67	-0.18
	4<=V12< 10	4	-1	0.42	15.417	10.32	-0.18
	10<= V12, _MISSING_	5	2	-0.14	26.604	70.73	-0.18

Table 1. Fragment of the Scorecard Results

Using this table, it is possible to detect outliers based on a low percentage of population and evaluate their impact on the target event rate (see Figure 8).

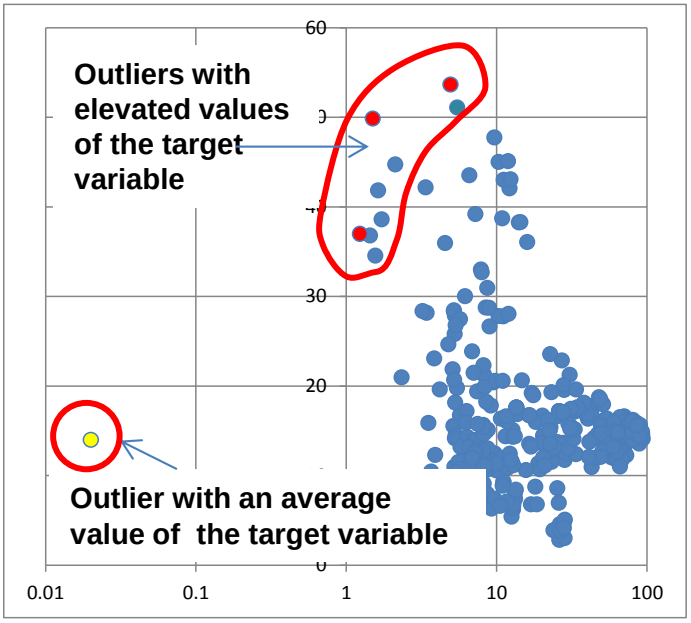

Figure 8. Scatter Plot Representing Event Rate of the Target Variable vs Logarithmically Scaled Percentage of Population

OUTLIERS AND DECISION TREES

The decision trees are quite robust against outliers present in the input variables, simply because they are mixed with others observations within the splits that constrained by minimum node size. Nevertheless, just ignoring them may lead to a loss of business opportunities. As mentioned earlier, outliers of the target variable or both target and input variable(s) definitely have an impact on business.

On the another hand, the decision trees can be used to detect business sound outliers (Breiman *et al*, 1984; John, 1995; Reif *et al*, 2008).

Concerning business implications, it can be done against one-way inputs that significantly impact the target variable. Here are the steps that can be followed identifying univariate special segments:

1. Select inputs for deterministic special segment splits by using business objectives and judgement. Note: it can be any type of variable: nominal, ordinal, binary, or interval numeric.
2. Run univariate decision trees specifying a very small final node size to be allowed. Explore the setup of the maximum number of branches to be, let's say, 5-7, i.e., more than the default of two.
3. Based on some sound statistics, such as purity measures for binary targets, identify candidates for the special segment.
4. Isolate special segments obtained by step 3 only if they have a strong business sense.
5. Run conventional decision tree, gradient boosting, random forest, scorecard, or regressions models on the rest of the population. Step 3 can provide additional guidance for a minimum node size to be allowed in the modeling.
6. Assembly 4 and 5.
7. Defend or reject obtained modeling results.

The identified special segments (step 4) should be accompanied by the following controls of both the sample size and the estimated target value. It will allow business to dynamically react to any changes detected by such controls. It should be noted that even having extremely small sizes, these special segments do not possess risk of overfitting. It is confirmed by the judgement against a strong business sense in step 4. Therefore, training decision tree model to find them (step 2) should not include partition to training, validation, or test samples. The partition is important to train a robust model on the remaining population to prevent overfitting (step 5). The main objective of this approach is to detect business sound special small segments of outliers while preventing overfitting of the whole model which is assured by step 5.

Similarly, multivariate outliers including their interactions can be detected by just running a decision tree model specifying a very small minimum node size. However, business judgement is very difficult or even impossible to perform considering interactions for a tree even of only three levels deep.

Running decision trees detecting outliers, the following aspects should be considered:

- Sample size of the final node
- Estimated values of the target variable of the final node
- Interaction effects of the branches
- Split cut-offs of the nodes

Thus simple proportion of the final node size to the size of the entire modeling sample size can be used to evaluate a support of that segment. These support values can be used then for a preliminary screening of all final nodes to be below certain small limit. Estimated value of the target variable of the final node should be compared to the entire population or to the neighboring node. A significant difference means that this node is a good candidate to be a business sound outlier. Moreover, the latest two aspects

provide some insights on possible root causes of the detected outlier: list of drivers, their interactions, and ranges of values.

COMBINATION OF ANALYTICAL APPROACHES

For more through detection and analysis of outliers and their impact on business, a chain of different analytical approaches can be applied.

In this case, the output of the first stage is included as an input of the next one. For example, first running an unsupervised clustering or association rules analysis with decision tree modeling afterwards, it allows for interactive effects between clusters or rules with other potentially predictive variables and their combined effects on the target variable to be found.

In this case, the decision tree target variable should be a business sound KPI, such as profit, revenue, sales volume, etc. In most cases, outlier clusters or rules can be isolated as special segments and not to be part of the following modelling stage. However, estimation and business judgement of the target variable is still required as well as follow-up monitoring.

OUTLIERS AND TIME SERIES

SPC is a special technique to classify variances as assignable and common ones, which was designed for time series [Duncan, 1986; Woodall , 1986; Juran and Godfrey, 1999; Montgomery, 2005; https://en.wikipedia.org/wiki/Shewhart_individuals_control_chart). Today SPC applications are widely used in modern manufacturing signalling any abnormal changes in the process. In data quality management, it becomes a routine to detect outliers during data loading based on some statistical control techniques. SPC became so critical in many industries that the International Organization for Standardization (ISO) provides guidance for the implementation of a SPC by ISO 11462 standard. Business objectives of that standard are clearly stated as “increasing production efficiency and inherent capability, and reducing interval and cost” (https://www.iso.org/obp/ui/#iso:std:iso:11462:-1:ed-1:v1:en).

The applicability of SPC to detect outliers is discussed in (Ben-Gal, 2005) focusing on traditional methods as well as using ARIMA models filtering autocorrelation processes first. Transformations of variables are a very important aspect to setup a proper control. Thus, a simple moving range calculation of the period-to-period changes of a variable allows for control of variance. Traditional SPC techniques assume normally distributed variables, however, using estimations of percentiles, especially bottom 1% and top 99%, it is possible to adapt SPC to be applicable to any empirical distributions (Montgomery, 2005).

The objects for SPC can be any KPIs, critical time series, or patterns along the time. Eventually, it can be applied to any variable, such as interval, ordinal, nominal, or binary, that is stored in a database and which may vary by time or other dimension, for example, by products. Of course, the selected variable for SPC should have strong business reasons for control.

There is a traditional and economical technique to setup SPC based on sampling and not a total control. However, it may possess a risk to skip outliers or be impacted by seasonality effects.

For high quality processes, where occurrences of the events of interest are very rare, SPC can be set by controlling the duration between two consequent occurrences of such events (Kaminski, 1992; Glushkovsky, 1994; Ransdell, 2016). Alternatively, the control can be set counting the number of regular events between two consequent events of interest. For example, SPC on a production line may be set against the number of good products between two consecutive bad ones. Recently, this control became more and more demanded in industries witnessing significant quality improvements to a “Six Sigma” level. It should be noted that such SPC means dynamic control of extremely rare outliers. To some degree out-of-control signals from such SPC can be recognized as “outliers in outliers”, since it provides detection of abnormal changes in outliers. Detection of such changes allows businesses to study effects implementing some improvement programs or getting earlier signals in case of unexpected changes for already high quality processes.

OUTLIERS AND STABILITY

Observing stochastic time series, there are different types of responses that result from the existence of assignable causes:

- Isolated single outlier
- Shift in mean or in variance
- Trend in mean or in variance
- Autocorrelation
- Cycling (seasonality)

All except the first one lead to time series instability. In contrast, having an isolated single outlier followed by common fluctuations means system stability to withstand such spikes. Of course, we should consider only inherent outliers while errors should be isolated, corrected and prevented. This “immunization” effect of a single outlier has been described in (Glushkovsky, 2006), which is reflected by the increased stability index.

Thus, a time series example, which is presented in Figure 9, has a single outlier. After that spike, the process was quickly recovered. The recovery period can be acknowledged by applying run-tests (Juran and Godfrey, 1999). The stability index of that time series equals 2.0, where the higher the index, the more stable the process (Glushkovsky, 2006). However, by removing just that one outlier, the stability index of the time series actually decreases to 1.7.

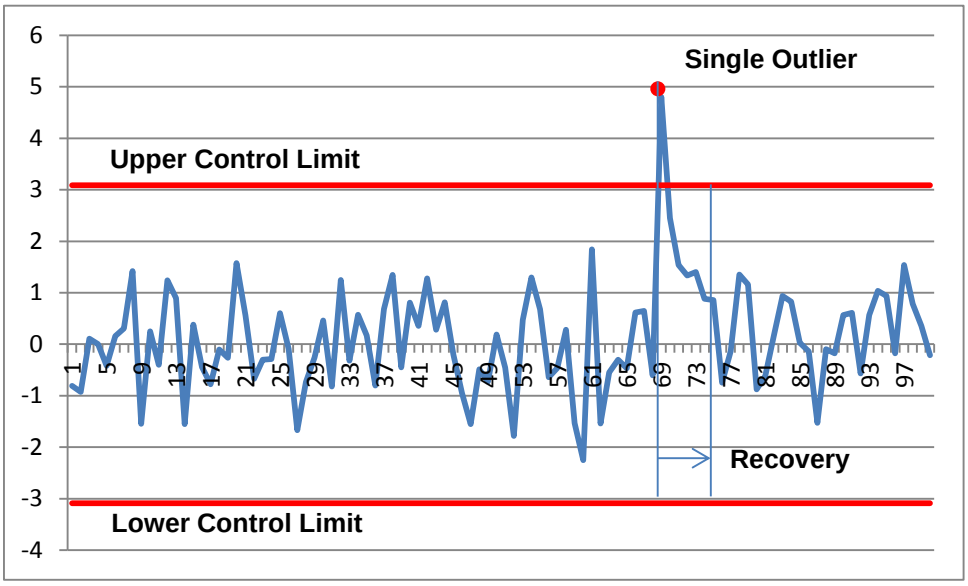

Figure 9. Example of Time Series with a Single Outlier

Analyzing past time series and detecting some specific outliers can be used when assessing business stability to the corresponding stress. Inherent outliers with negative outcomes on business permit for checking stability boundaries, predicting responses, and establishing preventive and corrective actions to such extreme situations. On contrast, outliers with positive impacts can be analyzed in order to retain or even boost their effects. For example, if a single outlier with a positive impact, such as the one shown in Figure 9, has been observed, then the question is: “What should be done to shift the mean of the process up, or even to trigger an upward trend”? To answer that question, data analysis focusing on data prior and during recovery period should be considered.

This process combined with “What-If” scenario generation prepares businesses for future possible surges even when having no current specific types of outliers. The limitation of such an analysis is within the observed ranges of the past significant fluctuations. The extrapolations outside of the observed past ranges cannot be applied due to usually unknown non-linear effects, and, therefore, the confidence of such estimations will be very questionable.

Creating special segments isolating outliers and establishing control of such segments, it extends a view on outliers as actual or potential ones. For example, business may identify a specific type of customer by some criteria. The size of the segment may vary from period-to-period, let’s say depending on a macroeconomic cycle as an example. Then, there may be some periods when there will be no observations in this segment. It means that this segment is defined not just for actual outliers but for potential as well. Focusing on potential outliers, not just on actual ones, opens opportunities for a business to be dynamically prepared for future possible changes. This can be set by generating “What-If” scenarios and assessing possible outcomes. This approach replicates the climbing plants nature strategy (Figure 2).

Outliers can be seen as severe factors of an embedded Design of Experiments (DOE) in the business. There can be even more dramatic and proactive steps ahead by introducing outliers with respect to some business drivers of interest. It can be recognized as an extreme DOE. This approach may lead to a very successful impact on business, assuming that risks are constrained.

CONCLUSION

Outliers are not just statistical or data quality issues but business matters. Do not ignore outliers. On the contrary, hunt for them! Use the climbing plants strategy to improve business in the right direction by using outliers as anchors to climb business peaks.

Dealing with outliers should not just be an analyst’s ad hock practice when preparing reports or training models, but systematic approaches beyond standard data quality and integrity assurances.

Systematically implementing special applications to detect outliers, to estimate their performance, to identify their root causes, and to control them will improve business.

DISCLAIMER

The paper represents the views of the author and do not necessarily reflect the views of the BMO Financial Group. All charts and data are simulated for illustrative purposes only and do not reflect the actual business state.

REFERENCES

Barnett, V. and Lewis, T. 1994. *Outliers in Statistical Data*, 3rd edition, John Wiley & Sons, Chichester

Ben-Gal, I., 2005. *Outlier Detection, Data Mining and Knowledge Discovery Handbook*, Springer

Breiman, L., Friedman, J., Stone, C., and Olshen, R. 1984. *Classification and Regression Trees*, Taylor and Francis

Cousineau, D. and Chartier, S. 2010. “Outliers detection and treatment: a review”, *International Journal of Psychological Research*, Vol. 3. No. 1

Dash, M. and Lie, N. 2010. “Outlier detection in transactional data”, *Journal Intelligent Data Analysis*, Vol. 14, Issue 3

Deming, W. 2000. *Out of the Crisis*, , MIT Press

Duncan, A. 1986. *Quality Control and Industrial Statistics*. 5th edition, Richard D. Irwin, Homewood, Ill

Glushkovsky, A. 1994. " 'On-line' G-control chart for attribute data", *Quality and Reliability Engineering International*, Vol. 10, Issue 3

Glushkovsky, A. and Billard T. 1998. "Pattern Control: Correlation Analysis Combined with SPC." *Quality Engineering*, Volume 11, Issue 2

Glushkovsky, A. 2006. "Stability Index of Stochastic Processes: The Statistical Process Control Approach." *Economic Quality Control*, Vol 21, No. 1, 87 – 111

Glushkovsky, A. and Fabian, M., 2017, "Differentiate Effects from the Noise of Promotional Marketing Campaigns", *Proceedings SAS Institute Inc*, SAS Paper 0778

Guidelines for implementation of statistical process control (SPC), 2001, https://www.iso.org/obp/ui/#iso:std:iso:11462:-1:ed-1:v1:en

Hawkins, D. 1980. *Identification of Outliers*, Chapman and Hall

Hawkins, S., He, H., Williams, G., and Baxter, R. 2002. "Outlier detection using replicator neural networks", *In Proceedings of the Fifth International Conference and Data Warehousing*

John, G. 1995. "Robust Decision Trees: Removing Outliers from Databases", *KDD-95 Proceedings*

Juran, J. and Godfrey A. 1999. *Juran's Quality Handbook*. 5th edition. NY, McGraw-Hill

Kalisch, M., Michlak, M., Sikora, M, Wrobel, L., Przystalka, P. 2016. "Influence of Outliers Introduction on Predictive Models Quality", *Proceedings Beyond Databases, Architectures and Structures. Advanced Technologies for Data Mining and Knowledge Discovery: 12th International Conference*, BDAS 2016, Ustroń, Poland, Springer

Kaminsky, F., Benneyan, J., Davis, R., and Burke, R. 1992. "Statistical Control Charts Based on a Geometric Distribution." *Journal of Quality Technology* 24:63–69.

Kolmogorov-Smirnov Test, 2017, https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Mamdouh, R. 2010. *Data Preparation for Data Mining Using SAS*, Morgan Kaufmann

Montgomery, D. 2005. *Introduction to Statistical Quality Control*. 4th edition, Hoboken, New Jersey: John Wiley & Sons, Inc.

Paulk, M., Weber, C., Curtis, B., and Chrissis, M. 1993. "Capability Maturity Model for Software (Version 1.1)". *Technical Report*. Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University.

Polfliet, J. 2016. "Outlier Detection Using the Forward Search in SAS/IML® Studio", *Proceedings SAS Institute Inc*, SAS1760

Preetha, S. and Radha, V. 2012. "Enhanced Outlier Detection Method Using Association Rule Mining Technique", *International Journal of Computer Applications*, Vol. 42, No. 7

Rapallo, F. 2011. "Outliers and Patterns of Outliers in Contingency Tables with Algebraic Statistics", *Scandinavian Journal of Statistics*, 39(4)

Ransdell, B. 2016. "Improving Health Care Quality with the RAREEVENTS Procedure", *Proceedings SAS Institute Inc*, Paper SAS4040

Reif, M., Goldstein, M., Stahl, A., and Breuel T. 2008. "Anomaly Detection by Combining Decision Trees and Parametric Densities", *Pattern Recognition*, ICPR 2008

Shewhart, W. and Deming, W. 2012. *Statistical Method from the Viewpoint of Quality Control*, Courier Corporation

Shewhart Individuals Control Chart. 2016. https://en.wikipedia.org/wiki/Shewhart_individuals_control_chart

Stevens, J. 1983. “Outliers and Influential Data Points in Regression Analysis”, *Quantitative Methods in Psychology*

Weisberg, S. 1985. *Applied Linear Regression*, 2nd Ed., J. Wiley & Sons, Inc., New York

Woodall, W. 1986. “The Design of CUSUM Control Charts”. *Journal of Quality Technology,* 18

ACKNOWLEDGEMENTS

The author would like to thank Matthew Fabian and Lori Bieda of the BMO Financial Group for their support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Alex Glushkovsky
Alex.Glushkovsky@bmo.com