# Selecting the Appropriate Outlier Treatment for Common Industry Applications

Kunal Tiwari
Krishna Mehta
Nitin Jain
Ramandeep Tiwari
Gaurav Kanda
Inductis Inc.
571 Central Avenue #105
New Providence, NJ

## ABSTRACT[1]

Outlier detection and treatment is a very important part of any modeling exercise. A failure to detect outliers or their incorrect treatment can have serious ramifications on the validity of the inferences drawn from the exercise. There are a large number of techniques available to perform this task, and often selection of the most appropriate technique poses a big challenge to the practitioner. In making this selection, the sophistication of the technique needs to be balanced with the ease of implementation and the impact on model performance. In this paper we focus on the impact some of these techniques have on the more common applications of modeling. We evaluate 5 different outlier detection and treatment techniques namely: Capping/Flooring, Sigma approach, Exponential Smoothing, Mahalanobis distance and the Robust Regression approach. The impact of these techniques is evaluated in a linear and logistic regression framework, the two most common modeling approaches relevant for a lot of industry applications.

## 1. INTRODUCTION

Outlier detection is an important task in data mining activities[2] and involves identifying a set of observations whose values deviate from the expected range. These extreme values can unduly influence the results of the analysis and lead to incorrect conclusions. It is therefore very important to give proper attention to this problem before embarking on any data mining exercise.

Research in this area is voluminous and there are whole arrays of techniques available to solve this problem. A lot of these techniques look at each variable individually, whereas there are others which take a multivariate approach to the outlier problem. A challenge often faced by the practitioner is to select optimally among these techniques.

In this paper, we evaluate the appropriateness of five outlier detection techniques for two of the most popular industry applications, linear and logistic regressions. In short listing the techniques for the study, we paid a lot of attention to the implementability of the techniques in an industry setting.

## 2. OUTLIER APPROACHES

As mentioned before, we selected 5 different outlier treatment approaches for closer evaluation. In this section, we will briefly describe each of the approaches.

### 2.1 Capping/Flooring Approach

A value is identified as outlier if it exceeds the value of the 99th percentile of the variable by some factor, or if it is below the 1st percentile of given values by some factor. The factor is determined after considering the variable distribution and the business case.

The outlier is then capped at a certain value above the P99 value or floored at a factor below the P1 value. The factor for capping/flooring is again obtained by studying the distribution of the variable and also accounting for any special business considerations.

### 2.2 Sigma Approach

With the sigma approach, a value is identified as outlier if it lies outside the mean by + or – “x” times sigma. Where x is an integer and sigma is standard deviation for the variable.

The outlier is then capped or floored at a distance of”y” times sigma from the mean. “y” is equal to or greater than “x” and is determined by the practitioner.

### 2.3 Exponential Smoothing Approach

To detect outliers with this approach, the curve between P95 and P99 is extrapolated beyond P99 by a factor “x”. Any values lying outside this extended curve are identified as outliers. Similarly the curve between the 5th percentile and 1st percentile is extended to Minimum value by some factor. Values lying below this extended curve are outliers.

[1] Please send all correspondence to kmehta@inductis.com

[2] For an extensive discussion on this see [1]

Any value which lies beyond the extended curve is treated by an appropriate function, which maintains the monotonicity of the values but brings them to an acceptable range.

### 2.4 Mahalanobis Distance Approach

Mahalanobis distance is a multivariate approach and is calculated for every observation in the dataset. Then every observation is given a weight as inverse of the Mahalanobis distance. The observations with extreme values get lower weights.. Finally a weighted regression is run on to minimize the effect of outliers.

Mahalanobis distance is different from Euclidian distance in that:

- It is based on correlations between variables by which different patterns can be identified and analyzed
- is scale-invariant, i.e. not dependent on the scale of measurements
- it takes into account the correlations of the data set

Here is how it is derived:
D2 = y'*y = y * i2 = (x – x avg)'*C-1*(x – x avg)
Where D (= +√D2) equals the Mahalanobis-distance.

y = C-1/2(x – x avg); Where x = original (log-transformed) data-vector
C = estimated covariance matrix
y = transformed data-vector
x avg = estimated vector of averages

For more on Mahalanobis technique see [2], [3].

### 2.5 Robust-Reg Approach

Robust-Reg is also a multivariate approach. It runs regression on the dataset and then calculates the residuals. It then computes the median of residual after which it takes difference of median and actual residual and calculates MAD (mean absolute deviation) as:

Median/0.6745 (where 0.6745 is value of sigma)

Now it calculates absolute value as residual/mad and gives weight with respect to absolute value. This process is run iteratively and gives weight to each observation. Outliers are given 0 weight and are deleted from final regression. For more on robust regressions, see [4], [5].

## 3. THE EVALUATION PROCESS

Our focus of application is in the direct marketing and consumer finance areas. In these areas as well as in a lot of other industry and research settings, the problems analyzed involve either a binary or a continuous dependent variable. Therefore, we have evaluated the outlier treatment approaches for the linear and logistic regression problems. Due to the difference in the nature of these problems, we have used separate but appropriate performance metrics for each of them. Four different datasets were used to derive the results. The results presented are average across the datasets. Wherever appropriate, results were calculated for validation datasets.

For logistic regressions the performance metrics used were Lift, percentage concordance and c-value. For the linear regressions, Lift, R-square, percentage error bands and coefficient of variation for the process were used for the evaluations.

The metrics used to evaluate the performance of the various techniques are fairly standard. A brief description of these metrics follows.

Lift at x is defined as the percentage of total responders captured (for logistic regression) or the percentage of total volume captured (for linear regression) in the top x percentile of the model output. In our experimental study, we chose lift at the 10th percentile as it is a very common cutoff used in different marketing exercises. The results reported broadly hold for lifts at the 20th and 30th deciles as well

Percentage concordance provides the percentage of total population which has come into concordance according to the model. A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value. For more on concordance see [6].

C-value is equivalent to the well known measure ROC. It ranges from 0.5 to 1, where 0.5 corresponds to the model randomly predicting the response, and a 1 corresponds to the model perfectly discriminating the response. For more on C-value see [6].

R-square is the proportion of variability in a data set that is accounted for by a statistical model. The R-square value is an indicator of how well the model fits the data. Higher is its value better will be the method used to build the model. It is one of the most common metrics used to evaluate a very broad class of models.

Coefficient of variation is a measure of dispersion of a probability distribution. Lower its value better will be the method used to build the model. For more on coefficient of variation see [7].

Percentage error bands are formed by breaking the percentage of error in prediction into bands. The percentages of population present in different error bands determine the efficiency of the model. It is desirable to have more of the observations fall into the lower error bands.

## 4. RESULTS

In this section we present the findings of our empirical study. We first discuss the results for the logistic regression model, followed by the linear regression results.

### 4.1 Logistic Regression[3]

**Lift Curve:**

The sigma approach gave the highest lift at the top deciles, followed by the capping and flooring approach. Figure 1 illustrates the results.

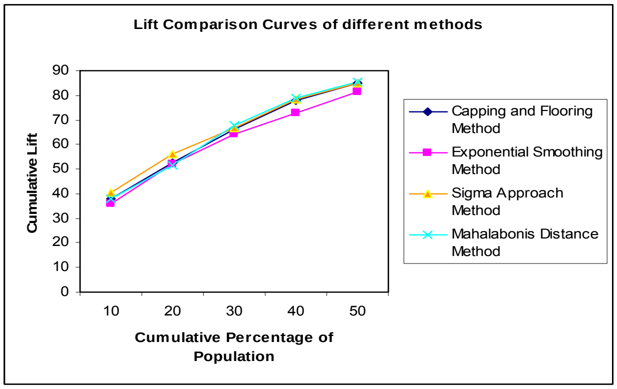

**Figure 1: Lift comparison curves for various methods using logistic regression**

**Concordance:**

Sigma approach provided the best concordance. The Mahalanobis distance measure again gave the worst results.

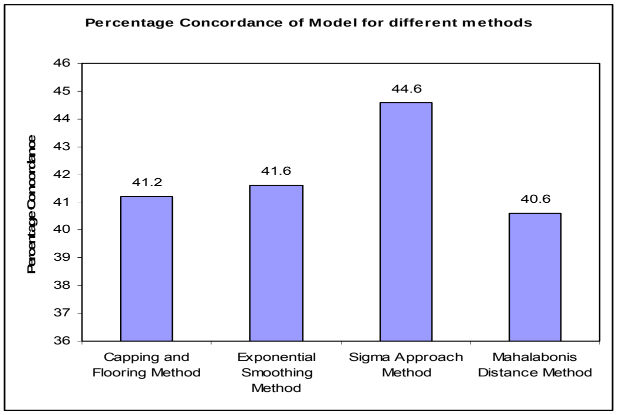

**Figure 2: Percentage Concordance comparison chart for various methods using logistic regression**

**C-value:**

When looking at the c-values, again sigma approach came out to be the best approach.

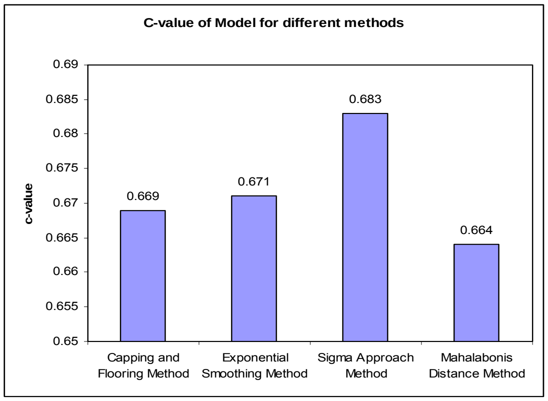

**Figure 3: C-value comparison chart for various methods using logistic regression**

### 4.2 Linear Regression

**Lift Curve:**

When results obtained from different approach were compared for the lift values, Mahalabonis distance approach came out to be the best. (see figure 4).

[3] Robust regression is more appropriate for linear regressions and was not considered for logistic regressions.

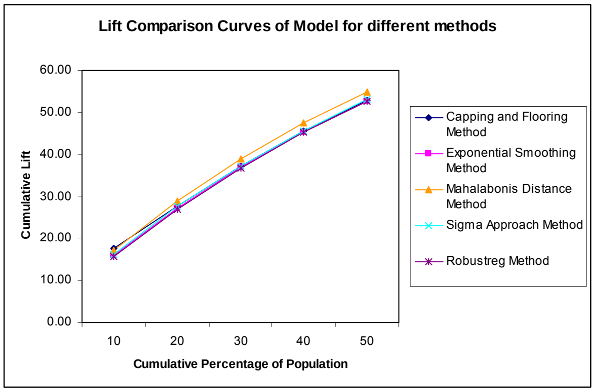

**Figure 4: Lift comparison curves for various methods using linear regression**

**R-square:**

When looking at the R-square values, again Mahalabonis distance approach came out to be slightly better than the other methods (see figure 5).

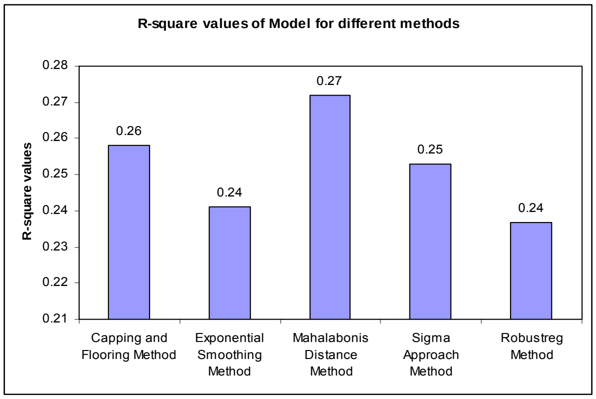

**Figure 5: R-square values comparison chart for various methods using linear regression**

**Coefficient of Variation:**

On the Coefficient of variation metric, Mahalabonis distance approach performed much better than the other methods (see figure 6).

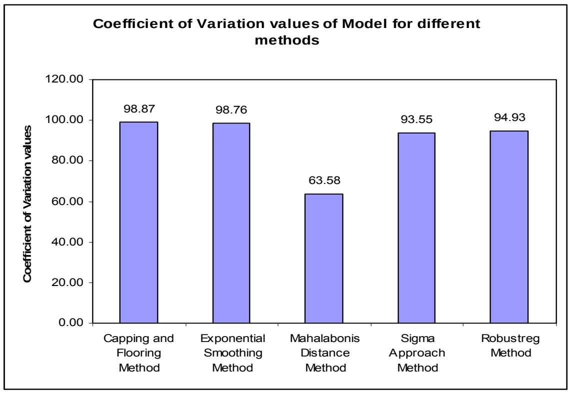

**Figure 6: Coefficient of variation values comparison chart for various methods using linear regression**

**Error Bands:**

Mahalanobis distance method outperforms all other methods on the error band criteria, as it captures a substantially larger amount of population in the lower error bands (see figure 7).

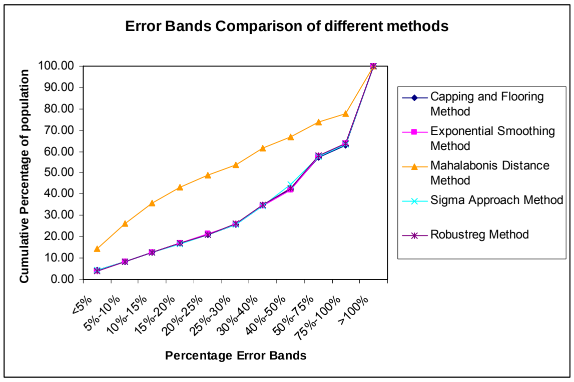

**Figure 7: Population Distribution in error Bands for various methods using linear regression**

Therefore for linear regressions, the Mahalanobis distance method does marginally better than other methods on some criteria but substantially improves model performance on coefficient of variation and error band calculations.

## 5. CONCLUSIONS

.
Our analysis provides evidence that the type of technique used for outlier treatment does influence the model performance.

For linear regression problems, Mahalanobis distance comes out better than any other technique used. However, for logistic regression problems simpler techniques like sigma perform the best and Mahalanobis distance approach performs the worst.

These results have serious implications for industry applications. For example in direct marketing applications, it is fairly common to build customer enrollment and profitability models with the same dataset as a two stage exercise. The first stage involves a logistic model and the second stage a linear regression model. The common here practice is to treat outliers at the beginning of the analysis and then proceed with no additional thought given to the outliers. However our findings suggest that this practice is sub optimal and outliers should be treated differently for the two modeling exercises**.**

## 6. REFERENCES

[1] Pyle D. *Data Preparation for Data Mining*, Morgan Kaufmann Publishers, 1999.

[2] RasmussenJ.L. *Evaluating Outlier Identification Tests: Mahalanobis D Squared and Comrey Dk*, Multivariate Behavioral Research 23:2, 189-202, 1988

[3] http://en.wikipedia.org/wiki/Mahalanobis_distance

[4] Rousseeuw P.J., Leroy A.M. *Robust Regression and Outlier Detection*, Wiley-Interscienc, 1987.

[5] Chernobai A., Rachev S.T. *Applying Robust methods to Operational Risk Modeling*, UCSB Technical Report, 2006.

[6] Allison P. *Logistic Regression Using the SAS System*, SAS Publishing, 1999.

[7] http://en.wikipedia.org/wiki/Coefficient_of_variation

## 7. ACKNOWLEDGEMENTS

[1] SAS is a Registered Trademark of the SAS Institute, Inc. of Cary, North Carolina

[2] Amit Sharma, Inductis, for his guidance on Mahalanobis distance approach

## 8. CONTACT INFORMATION

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

| | |
|---|---|
| Author Name: | Krishna Mehta |
| Company: | Inductis Inc.<br>571 Central Avenue #105<br>New Providence, NJ 07310 |
| Work Phone: | 1-908-743-1105 |
| Fax: | 1-908-508-7811 |
| Email: | kmehta@inductis.com |
| Web: | www.inductis.com |