Proceedings of Machine Learning Research 77:95–101, 2017 KDD 2017: Workshop on Anomaly Detection in Finance

An Automated System for Data Attribute Anomaly Detection

David Love DAVID.K.LOVE@AEXP.COM
American Express

Nalin Aggarwal NALIN.AGGARWAL@AEXP.COM
American Express

Alexander Statnikov ALEXANDER.STATNIKOV@AEXP.COM
American Express

Chao Yuan CHAO.YUAN@AEXP.COM
American Express

Abstract

We introduce DataQC, an automated system for data attribute anomaly detection for the purpose of improving data quality. Large organizations can have non-standardized or inconsistent data quality checking practices being followed across different departments. The key motivation behind the development of such a system is to 1) achieve a standard for anomaly detection 2) facilitate quick identification of obvious anomalies 3) reduce human judgment in data anomaly detection 4) facilitate prompt corrective action by data scientists. Most of the methods and techniques used during the development of this system are well known and have been widely used by finance professionals who deal with data. Our contribution is to provide a system that improves overall efficiency, interpretability, and objectivity for detecting data attribute anomalies.

Keywords: Anomaly detection, Data Quality Control

1. Introduction

Financial services companies collect, use and analyze data in a number of ways in the quest to make smarter, faster decisions that are beneficial to their customers and profitable for the company. The recent expansion of information technologies and e-commerce have led to the exponential growth of customer data that can be measured over different granularities (time periods, transactions, customer segments). As a result of this data influx, there is an increased likelihood of a data attribute taking unexpected or erroneous values, causing poor data quality, and thus impacting decision support systems and computer models that use this data.

Errors in data can come from many places. One obvious cause is simply human error, which may result in junk values within a continuous attribute, e.g., a FICO score of 8000. Other errors come from problems in automated data processing systems. For example, a broken data linkage between a data receiver and sender could result in an unusually large number of missing values in one or more data attributes. Irrespective of the reason, missing data can have huge influence on the models or data products developed by data scientist and mishandling often results in faulty interpretations made by the business.

© 2017 D. Love, N. Aggarwal, A. Statnikov & C. Yuan.

Data anomalies can occur even without errors. When collected over a long period of time, distributions can be observed to drift as the underlying behaviors change. For example, a typical financial institution may see the proportion of its population with low credit bureau scores change along with changing economic conditions. This population change could have more widespread impact among attributes defining the behavior of population like income, credit available to them etc. Fundamental data shifts like these can adversely affect the performance of key models and tools deployed by the organization.

Finally, anomalous data attributes could be due to the presence of genuine outliers in the data. Such cases, despite being a reality of the data, have to be treated properly before being processed through a typical modeling framework. Although many machine learning techniques are robust to the presence of outliers, their presence can still affect key steps like feature engineering and feature selection.

In this paper, we present an automated system for detecting anomalous data attributes called DataQC. Using DataQC, data scientists can easily run a comprehensive set of data quality tests against a previously verified reference dataset to quickly highlight how the data has changed from a prior analysis. These results enable the data scientist to focus their investigative efforts on parts of the data most likely to need special attention and ultimately deliver updated models with speed and reliability.

This paper is organized as follows: Section 2 describes the basic statistical methods used in DataQC; Section 3 describes how the methods are combined to generate an overall score on the quality of the data; Section 4 describes some visualization techniques used in DataQC; and Section 5 discusses the results of the use of DataQC.

2. Methods

To detect anomalous data attributes, DataQC leverages a panel of descriptive statistics that are suitable for big data. The analytical procedures in the system are run against a benchmark dataset that is provided by the data scientist. This "gold standard" dataset could be data used in previous experiments that has been checked manually for its correctness (absence of anomalous attributes). DataQC identifies anomalous attributes which have significant differences between the benchmark data and the tested data in terms of key descriptive statistics noted in this section. In the end, the system outputs a score between 0 and 100 where the score of 100 signifies data with zero anomalies while the score of less than 60 points to significant anomalies and deviations between the benchmark and tested data.

2.1. Statistics for Anomalous Attributes

Several of the tests described below in this section use Cohen's d and h statistics to represent the severity of any distributional changes between the benchmark and test datasets. Cohen's d and h measure the effect size (in units of the standard deviation) of changes in mean and proportion, respectively (Carson, 2012; Coe, 2002; Cohen, 1977). These statistics are preferable to standard hypothesis testing methods like the Student's t-test because, in the case of big data, extremely small changes in the distribution will reach statistical significance (Cohen, 1990; Sharpe, 2004). For the purpose of data quality control, we are interested only in changes that are large enough to have practical significance, which can be achieved with the effect size statistics.

Table 1 summarizes the usual recommended cut offs for small, medium, and large effect sizes (Ellis, 2009; Carson, 2012; Ferguson, 2009).

Table 1: Cohen's d and Cohen's h cut offs

Cohen's d / Cohen's h	**Effect Size**
0.2	Small
0.5	Medium
0.8	Large

2.2. Missing Rates

Problems in data pipeline linkages can result in missing values appearing in the final output dataset. To test for this possibility empirically, DataQC tests for significantly large differences in rate of missing values for each attribute in the dataset. Following the recommendations from Section 2.1, we adapt use of the Cohen's h test statistic, defined as

$$h = 2 \left| \arcsin \sqrt{p_T} - \arcsin \sqrt{p_B} \right|, \tag{1}$$

where p_B and p_T are the missing rates in the benchmark and test datasets, respectively (Cohen, 1977).

Cutoffs for the Cohen's h are chosen based on the guidelines listed in Table 1. Any attribute with an h value falling above the cutoff is marked as anomalous and in need of further investigation.

2.3. Means

DataQC checks for changes in distribution using three complementary methods. The first and simplest method is to check for changes in the mean value of each attribute. DataQC focuses on the effect size of difference in the means by calculating Cohen's d statistic,

$$d = \frac{\bar{x}_T - \bar{x}_B}{\sigma_B},$$

where $\bar{x}_B$ and $\bar{x}_T$ are the sample means of the benchmark and test data, respectively, and σ_B is the standard deviation of the benchmark data (Cohen, 1977).

Cutoffs are again selected sing the guidelines in Table 1. Any attribute with a d value above the cutoff is marked as anomalous.

Since the mean is sensitive to extreme values in the data, DataQC calculates the mean (and variance) by flooring and capping the observations at the 1st and 99th percentiles.

2.4. Single Attribute Distribution

Testing for the difference in the means cannot find all changes in distribution of an attribute. To determine whether other, more subtle changes in the distribution have occurred, DataQC constructs and compares the histograms of the benchmark and test datasets. Bins for the histograms are selected by a linear spacing between the 1^{st} and 99^{th} percentiles of the benchmark data, with two additional bins to count values outside this interval. The relative frequency of each bin is determined, and differences in these frequencies are assessed using Cohen's h statistic defined in equation (1). An attribute if marked as anomalous if at least one bin falls above the cutoff value.

2.5. Multiple Attribute Distribution

The histogram method from Section 2.4 may fail to detect changes in the interaction between multiple data attributes. Simply extending the histogram-based analysis of the previous section suffers from the curse of dimensionality in two ways: the number of possible combinations of attributes grows exponentially with the number of attributes, and the density of observations decreases as more dimensions are included. To counteract these problems, DataQC builds a Random Forest classifier (Breiman, 1996; Dietterich, 2000) to distinguish between the benchmark and test datasets. The Random Forest method was chosen because it works well across a variety of problems and requires the very little classifier-specific data preprocessing, e.g., random forest can be used without data scaling and with only sentinel imputation values. An anomaly score is generated for the test dataset overall (not the individual attributes) using the Gini coefficient of the classifier evaluated on a holdout set combined with the attribute importance rankings from the classifier.

2.6. Extreme Values

The final aspect to compare between benchmark and test data is the existence of extreme values. DataQC concentrates on examining the largest value in each dataset, and calculates the standardized value for both benchmark and test data,

$$E = \frac{\max(x) - \bar{x}_{B,99}}{\sigma_{B,99}}, \tag{2}$$

where $\bar{x}_{B,99}$ is the mean of benchmark data above the 99^{th} percentile, and $\sigma_{B,99}$ is the standard deviation of the same subset. Equation (2) is calculated for both benchmark and test datasets, notated E_B and E_T, respectively.

DataQC examines differences in extreme values using two rubrics: (a) difference in normalized extreme values $(e_T - e_B)$ and (b) ratio of extreme values $\frac{e_T}{e_B}$. Attributes are marked as anomalous if both statistics are sufficiently large.

3. Overall Anomaly Scores

DataQC provides an overall anomaly score for each of the above tests. The Random Forest classifier method described in Section 2.5 generates a score from the Gini coefficient on a holdout set and the attribute importance scores generated by the classifier. All other methods create anomaly scores from the percentage of attributes not marked as anomalous from each test. This system was tested across many different data sets. Table 2 lists categories for the severity of data anomaly were selected based on user feedback.

Table 2: Score ranges for low, medium, and high levels of anomalous attributes

Score	Attribute Anomaly Level
100 – 80	Low
80 – 60	Medium
< 60	High

In practice, all the data sets showing high attribute anomaly level must be investigated by the analyst.

3.1. Levels of Analysis

It is often desirable to reproduce the analysis from this section through different levels of the data. There are two levels that are most common: repeating the analysis broken down by time frame and by segment.

DataQC offers the automatic capacity to repeat all tests (except Section 2.5) for each time frame and for each segment. Scores are generated individually for each analysis that is performed.

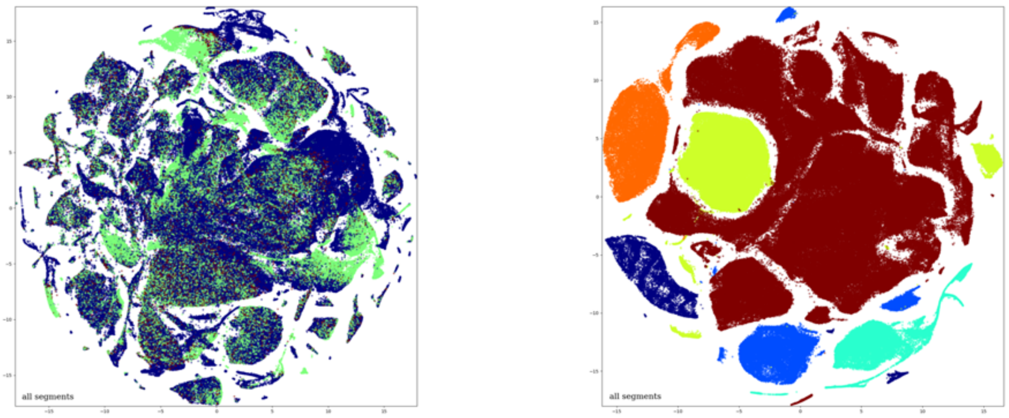
4. Data Quality Visualization

Data visualization is an important part of the analysis process that can reveal interesting aspects of the data that automated tests may miss. DataQC incorporates visualization by using the method of t-Distributed Stochastic Neighbor Embedding (t-SNE) to provide a robust visualization of many-dimensional data (Maaten and Hinton, 2008). t-SNE is a dimensionality reduction technique that preserves small distances between observations at the expense of making larger distances less meaningful. This means that neighboring points in the transformed data are also likely to be neighbors in the original dataset, thus preserving clusters in the original dataset.

DataQC uses t-SNE to provide a tool that can be helpful in answering several questions, outlined in the subsections below. None of the important variables highlighted (benchmark/test distinction, time frame, segment) are included in input for training the t-SNE transformation. Example t-SNE plots can be seen in Figure 1.

4.1. How Similar are the Benchmark and Test Datasets?

The benchmark and test datasets can be highlighted on the t-SNE plots in different colors. This provides a visual method of determining how different the datasets are. If the observations of benchmark and test data are well mixed throughout the plot, then we can conclude that the datasets are similar (e.g., Figure $1(a)$) On the other hand, if there are



(a) Well-mixed observations

(b) HIghly separated observations by segemnt

Figure 1: **t-SNE Plots.** This shows two example cases of t-SNE plots, where the classes are well mixed (1(a)) or well separated (1(b))

areas made up mostly or entirely of one dataset, this can point towards places of interest in the quality control process (e.g., Figure 1(b)).

4.2. Does the Data Change Over Time?

By highlighting the time frame of each observation, DataQC can provide insight on whether the data changes substantially over the set of time frames. Similar to the analysis provided in Section 4.1, well mixed observations would indicate low dependence on time, while distinct regions of solid colors show that time is an important component of this dataset.

4.3. Is Segment-Level Modeling Important?

Often in the practice of modeling one finds the question of the size of model to build: one large model or several smaller models on distinguishable segments of the data? DataQC's visualization techniques can be used to provide insight into this decision. When the segments are highlighted in the t-SNE plot, highly separated segments provide evidence that segment-level models may perform better than a single integrated model. Well mixed segments provide evidence of the opposite.

5. Results & Conclusions

In this paper we have introduced a novel data attribute anomaly detection system called DataQC. This system allows organizations to perform an objective and effective assessment of their data using the same high standard across business units and use cases. As a result, data scientists can have cleaner datasets and the opportunity to build more robust models.

The use of DataQC has reduced data scientists' manual data assessment effort by 90% and has significantly increased the accuracy of detecting anomalous data attributes. Use of this system also ensured consistently high standard for data assessment. To date, DataQC has been leveraged for all key customer financial datasets that constitute the basis of predictive models.

References

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Cristi Carson. The effective use of effect size indices in institutional research. *Citováno dne*, 11:2016, 2012.

Robert Coe. It's the effect size, stupid: What effect size is and why it is important. 2002.

Jacob Cohen. Statistical power analysis for the behavioral sciences (revised ed.), 1977.

Jacob Cohen. Things i have learned (so far). *American psychologist*, 45(12):1304, 1990.

Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40 (2):139–157, 2000.

Paul D. Ellis. Thresholds for interpreting effect sizes, 2009. URL http://www.polyu.edu.hk/mm/effectsizefaqs/thresholds_for_interpreting_effect_sizes2.html.

Christopher J Ferguson. An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5):532, 2009.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Donald Sharpe. Beyond significance testing: Reforming data analysis methods in behavioral research. 2004.