August 23, 2018

DETECTION AND VISUALIZATION OF ANOMALOUS PRINTERS USING MACHINE LEARNING

HP INC

Detection and Visualization of Anomalous Printers using Machine Learning

The method described below will predict printer models showing anomalous behavior by going through the data and find out exceptional behaving models quickly and more accurately. It will generate interactive plots to observe different patterns and distributions followed across individual countries and worldwide.

CDAX database contains case notes comprising of general information written by the Customer Support Service (CSS) agents for every call received. These notes contain information about printer model number, the location of the printer, resolution code (such as issue resolved, issue unresolved, order placed for parts of the printer, order placed for replacement of the printer, etc.) It was observed that most printer models follow a general trend of distribution over resolution code and other parameters. But there were printer models for which 65% of the time the call resulted in the replacement of the printer. For another printer model, 95% of the time the call resulted in an intervention (field executives visit the customer's location to fix the bug) that didn't belong in the high-volume call generator models nor in the high-end models. The exceptional behavior can only be identified by analyzing all the possible values of the parameters and manually doing it is not a feasible task. In the current scenario, data analysts will have to go through the dataset (from CDAX database) and plot these features (resolution codes) against each other to find the printer models showing exceptional behavior. The data is huge and features can range from 10 to 50 making it more difficult to handle and time-consuming. Plotting multiple features together is even more difficult to execute leading to low prediction accuracy.

The solution aims to identify printer models which don't follow the general distribution trend over resolution code and are not ever part of high volume call generator models. It also aims at providing an intuitive way of depicting the anomalies to observe patterns and distributions across printer models. Fig 1, below shows the various components of the algorithm.

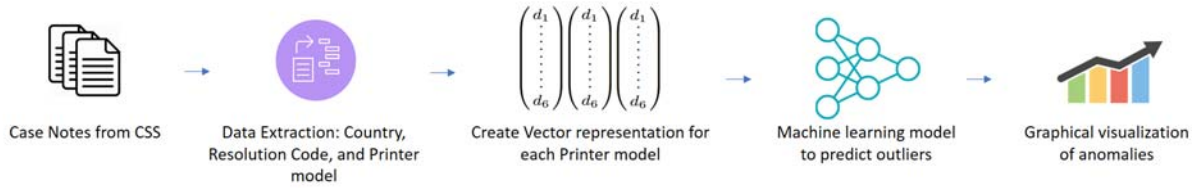

Figure 1.

Information is extracted about the country, printer model, and resolution for each log. The data contains a total of 105 different countries, 1700 distinct printer models and a total of 27 different types of resolution codes, which is a large number, making it difficult to analyze. To reduce the number of distinct resolution codes, it is clustered into 6 main categories i.e. Resolved, Unresolved, Printer Ordered, Parts Ordered, refer to 3rd party and Intervention (on-site visit).

To find anomalies based on printer models, each printer model was represented as a multi-dimensional vector of percentage distribution of resolution code. For example, printer model x received 1000 calls in total and the distribution is:

531 - Resolved, 153 - Unresolved, 92 - Printer Ordered, 114 - Parts Ordered,

52- Refer to a 3rd party and 58 – Intervention

Therefore, printer model x would be represented as: - [53.1%, 15.3%, 9.2%, 11.4%, 5.2%, 5.8%]

The vectors created in the previous module are passed through a machine learning algorithm to create a model which will predict the anomalies in the data. Isolation Forest algorithm is used to plot these vectors in free space. The algorithm selects each vector at a time and tries to isolate it from the other vectors by plotting random planes perpendicular or parallel to the axis. Isolation is done in an iterative manner for each vector. Vectors which are close to each other are difficult to isolate than the ones which are farther. An anomaly score is evaluated for each vector based on the difficulty level to isolate it from the other vectors. Printer models with low anomaly score are treated as anomalies in the data. This method is independent of size, complexity, and number of features resulting in high accuracy and low time consumption. The solution can be implemented on country-specific data as well.

The anomalies are depicted in an intuitive and interactive way by generating a plot between a total number of case logs for different printer models' vs their anomaly scores to visualize different patterns followed across different printer models.

On the X-axis, a total number of case logs for each printer model is plotted and their anomaly score on Y-axis. Each point on the graph represents a printer model highlighting the cause for their exceptional behavior. It helps to identify the distribution pattern and the relation between different anomalies i.e. how close or far they are relative to each other.

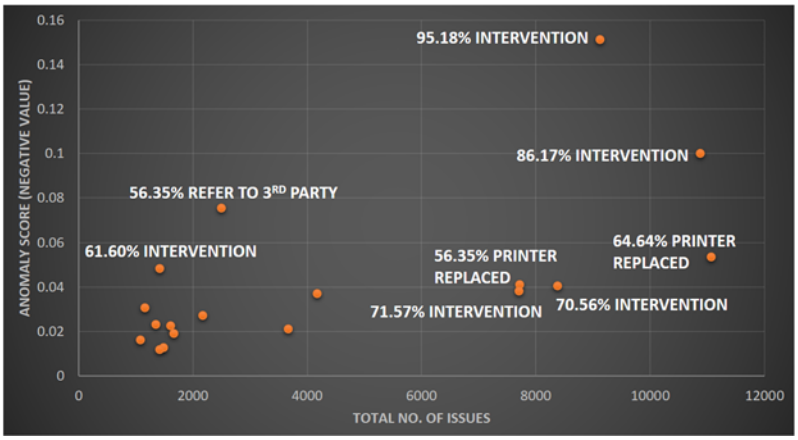

Figure 2.

References: - Isolation forest API –

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.htm

Disclosed by Shameed Sait M A, Himanshu Tiwari, Rajendram Ambrose and Rao Madhusoodhana, HP Inc.