Detecting Anomalies

Team B2: Aash Gohil, Chunxiaqiu(Tommy) Yang, Phyllis(Fangfei) Cao, Zihan Cui, Zixing Li

Project Overview: Founded by two veterans at Frontier Capital Management, Assette delivers comprehensive software solutions to assist its clients with their financial data management. Assette's current processing procedure refreshes its client databases on a batch basis, however, there remains possibilities that Assette receives data that has not been properly curated. In an attempt to increase Assette's business processing efficiency, our team's mission is to develop a series of statistical validations to detect anomalies in data submitted during the ingestion process.

Methodologies

1. Modified Z-Score

- The standard z-score are limited when the data are not normally distributed or the data/sample size is small, also sensitive to extreme values
- The modified z-score is a standardized score that measures outlier strength or how much a particular score differs from the typical score
  - It is less influenced by outliers when compared to the standard z-score because it relies on the median for calculating the z-score
  - The modified z-score is calculated from the mean absolute deviation (MeanAD) or median absolute deviation (MAD)
  - The values are multiplied by a constant to approximate the standard deviation

2. K-Nearest Neighbor (KNN)

- K-Nearest Neighbours algorithm detects anomalies using the distances of k-nearest neighbors as anomaly scores.
  - If an observation is much far from the other observations then that observation is considered to be an anomaly.
- The key parameter in kNN is n_neighbors, which determines the number of neighbors to use for calculating distances from the point of measurement
- Model outputs: (a) distances between data points and (b) associated index values which can be used for detecting anomalies

Data

2 Datasets - Primary Dataset & Comparison Dataset

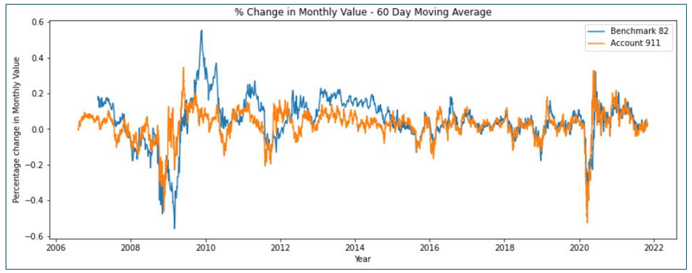
- Implement methods on original Account Performance Factor datasets (502 accounts).
- Also Implement on the created Comparison between Account and Benchmark Dataset (244 benchmarks).

BenchMark History		
Number of benchmark		244
	Num Benchmark	Year
oldest account	1	1969
newest account	2	2019
maximum accounts	98	1994
median accounts	98	1994

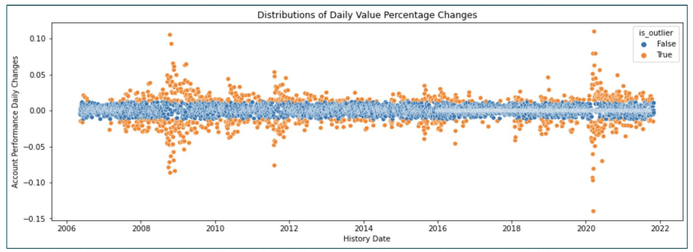
AccountPerformanceFactors		
Number of accounts		502
	Num Account	Year
oldest account	1	1979
newest account	18	2021
maximum accounts	36	2015
median accounts	19	2011

Relationship between BMH and APF	
Num of benchmark	Num of Accounts
1	168
2	293
3	25
4	19
5	2
6	5
7	1

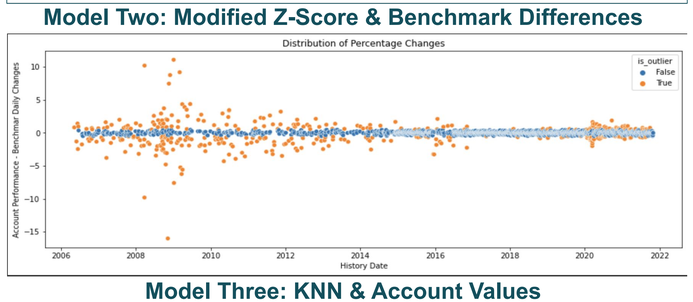
Sample Account and Benchmark



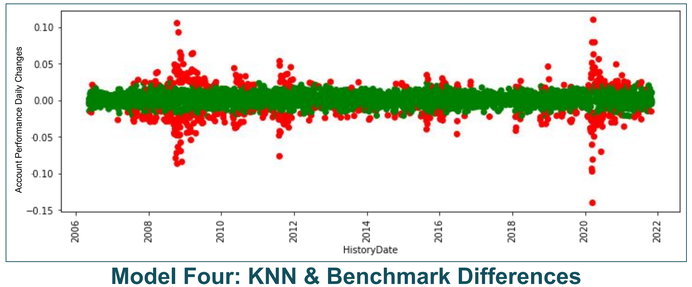
Model One: Modified Z-Score & Account Values



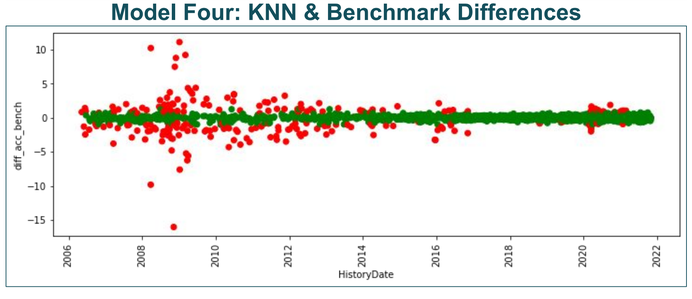
Model Two: Modified Z-Score & Benchmark Differences



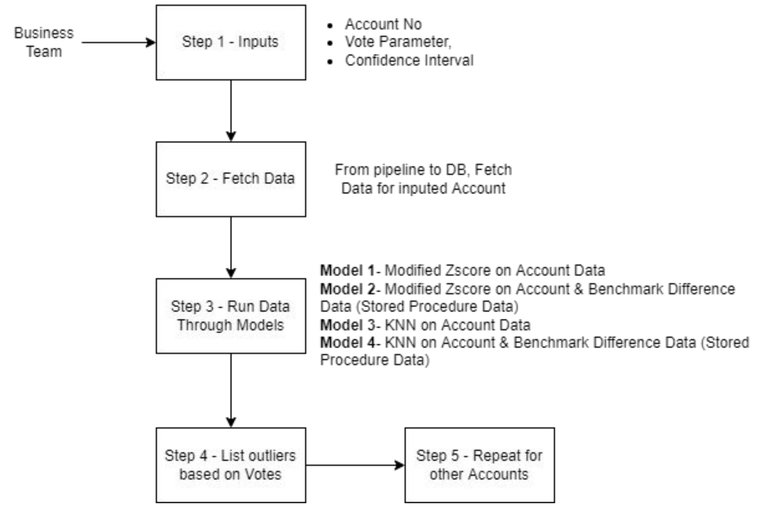
Model Three: KNN & Account Values



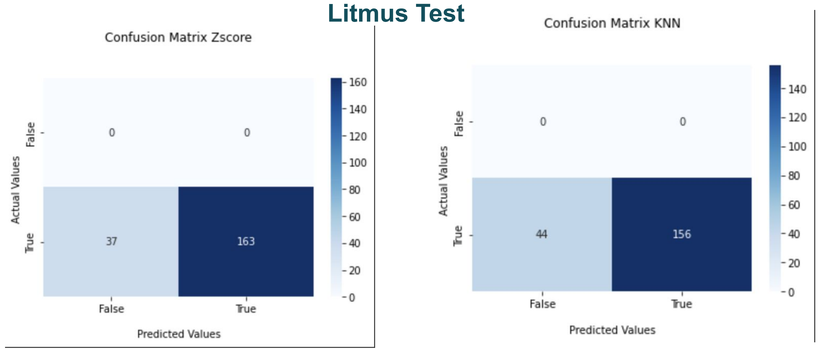
Model Four: KNN & Benchmark Differences



Business Implementation Diagram



Litmus Test



Summary

Litmus test shows that Z-Score is potentially more accurate than KNN method. The validity of both methods can also be tested by observing the proportionally larger amount of outliers during economic events, such as the financial crisis in the late 2000s and the COVID-19 pandemic in the recent years. For the sample account model one through four each produced 824, 363, 460, and 222 outliers. This shows that using benchmark differences could potentially be more conservative than using monthly values alone.

Limitations

One limitation of this project is that there were no available labeled data, so there were no way to know for certain the actual accuracy of the methods. To further improve the analysis, efforts such as experimenting with other models and finding labeled data may help reinforce the reliability of the anomaly detection methods.