NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

THE EFFECTIVENESS OF MACHINE LEARNING-BASED ANOMALY DETECTION ALGORITHMS APPLIED TO DEFENSE CONTRACT FINANCIAL DATA

by

Keith D. Edmonds

December 2020

Thesis Advisor: Robert A. Koyak
Second Reader: Colby J. Smithmeyer

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** December 2020	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
THE EFFECTIVENESS OF MACHINE LEARNING-BASED ANOMALY DETECTION ALGORITHMS APPLIED TO DEFENSE CONTRACT FINANCIAL DATA

5. FUNDING NUMBERS

6. AUTHOR(S) Keith D. Edmonds

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

In fiscal year 2020, the U.S. Army spent nearly $77 billion on contracts. Auditors employ various techniques, including anomaly detection, to select contracts that merit scrutiny. But in a resource-constrained environment, auditors can review only a limited number of contracts. Using data obtained from USAspending.gov, we consider how anomaly detection combined with dimensionality reduction can be used to recommend contracts for investigation.

We analyze over 20,000 fixed-price Army contracts between fiscal years 2017 to 2020, using more than one hundred combinations of dimensionality reduction and anomaly detection techniques, and formations of artificial anomalies. A consistent finding is that dimensionality reduction using principal components or autoencoders is not demonstrably beneficial. This finding may be due to the discrete nature of the USAspending.gov data and may not apply to other data sets. The best performance is obtained using isolation forests for anomaly detection without dimensionality reduction.

14. SUBJECT TERMS	**15. NUMBER OF PAGES**
anomaly detection, financial data, machine learning, benchmarking, benchmark, defense contract, contracts, Army, isolation forest, IF, auditors, USAspending.gov, dimensionality reduction	85
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

THE EFFECTIVENESS OF MACHINE LEARNING-BASED ANOMALY DETECTION ALGORITHMS APPLIED TO DEFENSE CONTRACT FINANCIAL DATA

Keith D. Edmonds
Major, United States Army
BS, Columbus State University, 2003
MBA, Troy University, 2004

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 2020

Approved by: Robert A. Koyak
Advisor

Colby J. Smithmeyer
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

In fiscal year 2020, the U.S. Army spent nearly $77 billion on contracts. Auditors employ various techniques, including anomaly detection, to select contracts that merit scrutiny. But in a resource-constrained environment, auditors can review only a limited number of contracts. Using data obtained from USAspending.gov, we consider how anomaly detection combined with dimensionality reduction can be used to recommend contracts for investigation.

We analyze over 20,000 fixed-price Army contracts between fiscal years 2017 to 2020, using more than one hundred combinations of dimensionality reduction and anomaly detection techniques, and formations of artificial anomalies. A consistent finding is that dimensionality reduction using principal components or autoencoders is not demonstrably beneficial. This finding may be due to the discrete nature of the USAspending.gov data and may not apply to other data sets. The best performance is obtained using isolation forests for anomaly detection without dimensionality reduction.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

COVID-19	Coronavirus disease 2019
CSV	comma-separated values
DA	Department of the Army
DATA Act	Digital Accountability and Transparency Act
DBSCAN	density-based spatial clustering of applications with noise
DCAA	Defense Contract Audit Agency
DCMA	Defense Contract Management Agency
DoD	Department of Defense
DoN	Department of the Navy
FAO	federal action obligation
FPC	flexible procedures for clustering
FY	fiscal year
GMM	Gaussian mixture model
IF	isolation forest
kNN	k-nearest neighbor
LOF	local outlier factor
NAICS	North American Industrial Classification System
PCA	principal component analysis
SSE	sum of squares of errors
UCI	University of California at Irvine
VAE	variational autoencoder

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Each year, the U.S. federal government spends hundreds of billions of dollars on goods and services, most of which is acquired through contracts. It is important to monitor those contracts to ensure that taxpayers' money is spent efficiently and properly. Due to the millions of contracts that are managed, it is impossible for the federal government to scrutinize each individual contract in detail. Data from the USAspending.gov (2020) website, which is managed by the U.S. Department of the Treasury, provides information on federal contract awards to the U.S. public. Although the information that it provides is valuable, it does not provide the level of detail needed to conduct thorough reviews of contracts. In order to obtain more detailed information, it would be necessary to examine the records of specific contracts, which would be prohibitively expensive and time consuming. In this thesis, we examine the extent to which the USAspending.gov data can be used to give an initial indication of contracts that warrant additional scrutiny. This would allow investigative resources to be applied in an efficient manner.

The purpose of our research is to identify algorithms that are effective for generating a recommender's list of contracts for further review. This study poses two questions: 1) How do dimensionality reduction methods compare in their performance? 2) How do anomaly detection algorithms compare in their performance?

Data to support our research was obtained from the USAspending.gov website, limited to fixed price U.S. Army contracts for the purchase of computer and electronic products, between fiscal years 2017 to 2020. A total of 22,491 contracts meet these conditions. Selection of features of these contracts that we use for our analysis results in 86 measured variables. The data is then split into a learning set and test set. The learning set, based on fiscal years 2017 through March 2020, contains 21,193 observations. The test set, based on April through September 2020, contains 1,310 observations. Additionally, we form 12 synthetic, anomalous observations that we append to the test data set. The objective of using dimensionality reduction is to remove noise from the data and extract essential information with a smaller number of variables. We apply three-dimensionality reduction methods to the learning data set: no dimensionality reduction, implying 86

dimensions; principal component analysis (PCA), implying 20 dimensions; and autoencoder, implying 20 dimensions. The next step for our research involves the use of anomaly detection algorithms.

The objective of the anomaly detection algorithms is to obtain a subset of records that are the best candidates for further inspection due to their unusual characteristics. We compare four anomaly detection algorithms: isolation forests, k-means clustering, density-based spatial clustering of applications with noise (DBSCAN), and k-nearest neighbors.

To answer the research questions and identify algorithms that are effective for generating a recommender's list of contracts for further review, we compare 120 distinct settings of dimensionality reduction and anomaly detection algorithms applied to the test set data with various types of seeded anomalies. We observe no benefit from the use of PCA or an autoencoder relative to no dimensionality reduction, holding all other factors fixed. This surprising result may be explained by the fact that the USAspending.gov data is highly categorical, with only one continuous variable. We do not imply that our results would be applicable to data set that contains many continuous variables. Of the two methods that reduce dimensionality, the autoencoder is found to be more effective than PCA. Additionally, isolation forests, with no dimensionality reduction, outperforms the other techniques across a wide range of circumstances. With no dimensionality reduction and across a range of settings, on average isolation forest found 10.8 out of 12 seeded anomalies. Dimensionality reduction, however, adversely affects the performance of isolation forest on the USAspending.gov data to a greater extent than the other anomaly detection methods.

Based on these results, the most efficient combination of algorithms to use with the USAspending.gov data set is isolation forest without reducing the dimensionality. In another data set, such as one having many continuous variables, the results may be different.

References

USAspending.gov (2020) Award data archive. Accessed October 30, 2020, https://www.usaspending.gov/download_center/award_data_archive.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

First and foremost, I would like to thank my wife for enduring the pain of this study. You carried a heavy load keeping everything afloat as I stayed glued to my computer. As a young Soldier, I did not truly understand the significance of your contribution. As a seasoned Soldier, I realized the "house" is still standing because of you. Thank you and I love you! To my kids, I appreciate your understanding during this time. It is my hope that you see this as an example to follow. To my parents, especially my mom, who would always rush to get off the phone so that I had maximum time with my studies, thank you for your support over all these years. Ms. Voli, thank you for keeping the "beast" off my back at times.

To my thesis advisor, Dr. Robert Koyak—I don't know what to say. I know you're not much for flowery words, so thank you. Thank you for accepting me and becoming my thesis advisor. Thank you for being patient with me as I learned and attempted to apply this new skill set. Thank you for the probably no fewer than 50 hours of one-on-one Zoom meetings. Thank you for the 0045 hours email replies. Thank you! I know, this is starting to sound like a "chant." So, simply stated, without your steadfastness in ensuring I completed the work, I would not be celebrating my departure from NPS with degree in-hand. You practically carried me across the finish line. You've taught me so much with this study, and I'm very appreciative of it.

Professor Whitaker, thank you! The assistance you provided with the AWS service, although it may appear small, helped out immensely in this study. Your kindness helped me maintain my sanity while running resource intensive codes.

To my fellow "greensuiters" at TRAC Monterey—LTC Wade, thank you for bringing me into the fold and allowing me to work this problem set. Kurt, thank you for getting me up and running. Last, but definitely not least, Colby. Colby, as my second reader, in that final play when I needed a quick assist, you did not hesitate to lend a helping hand. Thank you!

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

> Any fool can tell a crisis when it arrives. The real service to the state is to detect it in embryo.
>
> —Isaac Asimov (as quoted in Luebke 2020)

Asimov's observation is especially important when we attempt to detect anomalies in large datasets. As Luebke observed:

> Although Asimov wrote the quote above about a crisis in the futuristic Galactic Empire, it can also be applied to data. The quote eloquently summarizes why catching abnormalities in the earliest, "embryotic" state can save you from crises such as downtime, by helping you find the root cause of an issue faster, and reducing mean time to repair. (2020)

Each year, the U.S. federal government spends hundreds of billions of dollars on goods and services, much of which is through contracts. It is important to monitor those contracts to ensure that taxpayers' money is spent efficiently and properly. Due to the millions of contracts that are managed, it is impossible for the federal government to scrutinize each individual contract in detail. Data from the USAspending.gov (2020) website, which is managed by the U.S. Department of the Treasury, provides information on federal contract awards to the U.S. public. Although the information that it provides is valuable, it does not provide the level of detail needed to conduct thorough reviews of contracts. In order to obtain more detailed information, it would be necessary to examine the records of specific contracts, which would be prohibitively expensive and time consuming. In this thesis, we examine the extent to which the USAspending.gov data can be used to give an initial indication of contracts that warrant additional scrutiny. This would allow investigative resources to be applied in an efficient manner.

To address this topic, we focus on contracts made by the Department of the Army (DA) during fiscal years 2017 through 2020 that involve purchases of computing and electronic equipment. Approximately, 25,000 contracts are considered with a combined value of over $14 billion. A challenge in meeting this objective is that the USAspending.gov data provides 276 attributes to describe each contract. We examine the

extent to which statistical techniques for reducing dimensionality can help to meet this objective.

Although it is valuable to identify specific contracts that are well isolated from others in their attributes, another useful application is to provide a list of contracts that most warrant investigation when investigative resources are limited. In the latter case, it may be that no contracts are isolated from others. Our approach can be used to manage investigative resources without implying that any specific contracts are anomalous. Our approach belongs to the category of unsupervised learning techniques: we do not have information on contracts that would label them as problematic or anomalous.

A. THESIS PURPOSE

The purpose of this study is to identify algorithms that are effective for generating a recommenders list of contracts for further review. With the numbers of algorithms available that can be used to identify and score anomalies, there is little knowledge on how well they work with data from USAspending.gov. Benchmarking these tools will help identify those algorithms that are especially promising in generating a recommender's list for further investigation.

B. RESEARCH QUESTIONS

This study will examine behaviors related to contracting, both on the part of those receiving contracts and on the part of contracting activities in the DA, by analyzing various anomaly detection algorithms in order to create a benchmark that will seek to answer the following questions:

1. How do dimensionality reduction methods compare in their performance?
2. How do anomaly detection algorithms compare in their performance?

C. SCOPE, LIMITATIONS, AND ASSUMPTIONS

Contract data from fiscal year (FY) 2017 to 2020 are used to assess the benchmarking of anomaly detection algorithms. Particularly, we analyze the DA's

obligated contracts for goods and services, focusing on fixed price contracts for products classified as computers and electronic manufacturing.

The data contains extensive information about contracts, including the amount of the contract, who performed the contract, class of materials, and set aside qualifications. Lacking, however, is information on the specific item being purchased, including quantities and unit cost. In FY 2019, for example, General Dynamics Mission Systems was awarded a multi-year contract valued at nearly $112 million for computer and peripheral equipment manufacturing, but quantities are not listed, which prevents us from obtaining costs per unit.

Two primary assumptions are made with regards to the USAspending.gov data set: data accuracy, and consistency of structure over time. We accept the data reported in USAspending.gov, which is provided by contractors and vetted through federal checks (USAspending.gov 2020), to be accurate.

D. THESIS ORGANIZATION

The remainder of this paper is organized as follows. In Chapter II, background and related work, we review the setting surrounding defense contracts and research in anomaly detection. Chapter III, data and methodology, we identify the structure of the USAspending.gov data and provide the method of this study. Chapter IV, results, we compile and present the research results. Lastly, Chapter V, conclusion, we discuss the results and our recommendations.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND AND RELATED WORK

In this chapter, we review research related to anomaly detection that incorporates the use of machine learning algorithms on financial data. This review provides both the setting to which this research applies and insights into current work in the field.

A. USASPENDING.GOV DATA

As of August 31, 2020, the U.S. government had spent $8.7 trillion during FY 2020 (USAspending.gov 2020). Of this amount, the Department of Defense (DoD) was responsible for 11.68% of the total, or just over $1 trillion. Of this amount, approximately $343.3 billion (33.6%) was obligated for defense contracts. The DA was responsible for approximately $76.7 billion (22.3%) of the DoD's contracts budget. For fiscal years (FYs) 2017 to 2019, DoD alone was responsible for obligating nearly $1.1 trillion to defense contracts. Of the $1.1 trillion, DA was responsible for nearly $264.6 billion (24.1%).

In keeping with the requirements of the Digital Accountability and Transparency Act (DATA Act) of 2014, the U.S. government provides spending data on its expenditures for the U.S. public via the USAspending.gov website. In accordance with the DATA Act, the Office of Management and Budget transferred responsibility of USAspending.gov to the Department of the Treasury. As stated on the USAspending.gov webpage,

> USAspending.gov is the official source for spending data for the U.S. Government. Its mission is to show the American public what the federal government spends every year and how it spends the money. You can follow the money from the Congressional appropriations to the federal agencies and down to local communities and businesses. (USAspending.gov 2020)

B. ANOMALY DETECTION

A widely-used approach to managing the expenditures of billions of dollars is through the application of anomaly detection techniques. Anomaly detection, or the process of finding non-conforming patterns, has been applied to DoD problems that include military surveillance, fraud detection, and fault detection (Chandola et al. 2009).

Two DoD agencies are responsible for oversight and auditing for contract: The Defense Contract Management Agency (DCMA) and The Defense Contract Audit Agency (DCAA) (DCAA 2020; DCMA 2020; SBIR 2020). Because auditing is manpower-intensive its use is necessarily limited in a resource-constrained environment. Due to this limitation, the number of investigators needed to review all contracts warranting a closer look is lacking and as such, implementing a selection process based on a recommender list obtained from anomaly-detection metrics would be beneficial. The benchmarking of anomaly detection algorithms, with or without dimensionality reduction, would be helpful to guide the building of a recommender system of this kind.

Continuing, we review several benchmarking studies. Each study is described in three subsections: the problem and approach to answer this problem, data and methodology, and lastly, the author(s) recommendations. The five studies that we discuss are Emmott et al. (2013), Steinbuss et al. (2020), Wang et al. (2019), Bowman (2019), and Lee (2019).

1. Systematic Construction of Anomaly Detection Benchmarks from Real Data

Emmott et al. discuss the lack of a standard methodology for anomaly detection methods, and they address problems that arises from use of specific applications on synthetic data (2013). The authors approach this issue by presenting a methodology for creating synthetic anomalies from real-world data sets.

a. The Study

The authors use data from the University of California at Irvine (UCI) Machine Learning Repository with the following criteria: not time-series, at least 1,000 observations, no more than 200 features, and only numeric data. The authors' methodology includes selecting data sets with specific criteria, defining normal versus anomalous data points, computing point difficulty, and semantic variation and clusteredness. The detection algorithms that the authors consider are one-class support vector machines, support vector data description, local outlier factor (LOF), isolation forest (IF) and split-selection criterion isolation forest, and ensemble gaussian mixture model.

b. Conclusion/Recommendations

The authors recommend that three attributes be controlled in benchmarking algorithms: point difficulty, which measure the distance between an anomalous and a normal observation; relative frequency of identified true anomalies based on differing algorithms; and clusteredness or the separability of clusters.

2. Benchmarking Unsupervised Outlier Detection with Realistic Synthetic Data

Steinbuss et al. discuss the challenges of benchmarking unsupervised outlier detection algorithms with synthetic data (2020). The authors also compare outliers found in existing benchmark data and fully synthetic data. In existing benchmark data, outlier characteristics are various and unknown, while with synthetic data, outliers and regular instances display clear characteristics and therefore allow more meaningful evaluation of detection methods. Their approach is to develop a process for generating of data sets with the idea of reconstructing anomalous observations from existing real-world benchmark data that displays insightful characteristics. Their "classify" algorithm shows how well the authors model fits real-world data, and their "detect" algorithm benchmarks unsupervised outlier detection methods. Competitive outlier detection methods used for benchmarking, as noted by the authors, include IF, k-nearest neighbors (kNN), LOF, and kernel density estimation (2016).

a. The Study

The authors use 19 data sets from the UCI repository suggested by Campos (2016), for their analysis. These data sets contain only numeric attributes, have duplicate values removed, and each attribute has at least ten distinct values. The authors develop a generic process for designing realistic synthetic benchmarks for unsupervised outlier detection. In this process each data set is used to fit a generative model to the regular instances, which is modified to yield a model for outlier generations.

b. Conclusion/Recommendations

The authors' approach reveals promising results, although no single method is optimal for all types of outliers. Much like many domain-specific approaches, better results can be derived from using a tailored approach rather than a generic approach.

3. Progress in Outlier Detection Techniques: A Survey

Wang et al. provides a comprehensive and organized review of the progress of outlier detections as of the time of their study (2019). Specifically, the authors compare the advantages and disadvantages of a number of outlier detection techniques.

a. The Study

The outlier detection methods considered by the authors are applied to different sets of data, including regular and high-dimensional data sets as noted in Koufakou et al. (2010), streaming data sets, network data, uncertain data as used by Aggarwal et al. (2008), and time series data. The authors categorize outlier detection techniques into six main groups: statistical-based, distance-based, density-based, clustering-based, ensemble-based, and learning-based. These groups are then compared based on their performance.

b. Conclusion/Recommendations

Wang et al. conclude that the most appropriate outlier method is predicated on the data set. Based on the USAspending.gov data, the learning-based approach would be the preferred method according to the authors' rubric.

4. Anomaly Detection Using a Variational Autoencoder Neural Network

Bowman (2019) discusses the difficulty in identifying and classifying observations in an unsupervised environment using Department of Navy's (DoN) contract award data. The author addresses this problem by developing a new objective function for training variational autoencoders (VAEs) combined with a parameter selection technique based on a gaussian mixture model (GMM).

a. The Study

Of the five datasets Bowman uses, four are from the UCI Machine Learning Repository, Knowledge Discovery and Data Mining Competition in 1999 (Statlog Shuttle, Forest Covertype, and Fashion-MNIST), and one is from USAspending.gov. The USAspending.gov dataset encompasses FY 2014 to FY 2018. The author uses the UCI repository datasets to evaluate the VAE performance against ground truth labels. The author then applied the tuned VAE to the unlabeled DoN contract data.

b. Conclusion/Recommendations

Bowman finds that altering of the original objective function was beneficial to increasing anomaly detection effectiveness using VAE relative to the baseline, and that GMM is capable of assigning quantifiable metrics for the author's analysis. Our use of autoencoders to reduce dimensionality is discussed in Chapter III.

5. Utilization of Machine Learning Techniques to Detect Anomalies in Department of Defense Contract Data

Lee (2019) discusses the lack of extensive knowledge in identifying irregular spending patterns that may point out questionable spending practices in the USAspending.gov data. The author addresses this problem by proposing a set of statistical methods to locate those contracts for further scrutiny.

a. The Study

Lee uses the USAspending.gov data set for his research. Specifically, the author examines fixed-price U.S. Army contracts issued between fiscal years 2013 and 2018 in the area of information technology. He uses machine learning to develop metrics for outliers in the amount of contract obligations, and for monitoring awarding offices with respect to the use of set-aside categories and no-bid contract awards.

b. Conclusion/Recommendations

Lee's statistical approach creates avenues for investigators to screen contracts for anomalies. The author recommends that additional details be captured with the

USAspending.gov data, such as a fraudulent adjudication feature. Incorporating this additional feature can make use of data from the Office of Justice Programs (2019).

III. DATA AND METHODOLOGY

In this chapter, we describe the USAspending.gov data set and presents our methodology. In the methodology section, we address data preparation, dimensionality reduction, anomaly detection techniques, and research design.

A. DATA

USAspending.gov award data is pulled daily from the Federal Procurement Data System Next Generation database. We focus on data from FY 2017 to 2020 for fixed price US Army contracts in the area of computing and electronic equipment, which comprises 22,491 contracts.

1. Structure

The contract spending data in its original form consists of compressed comma-separated values (CSV) files retrieved from USAspending.gov. Each compressed file represents an entire fiscal year, unless a fiscal year is being downloaded mid-year, at which point only those available contract data are available. Each CSV file consist of no more than one million observations and 276 features per observation. USAspending.gov adds or removes features based on requested new data or due to unexpected national-level expenditures, such as those related to Coronavirus Disease 2019 (COVID-19). For the first two quarters of FY 2020, the dataset contains 276 features, but for the final two quarters the data contains 282 features. The additional six features added by USAspending.gov are COVID-19 related.

2. Description of Features

USAspending.gov data features provide information in four primary areas:

- *who*, as it relates to contractor and awarding agencies;
- *what*, identifying the contract in the North American Industrial Classification System (NAICS) (see Appendix A);

- *where*, as it relates to the contractor and performance location;
- *when*, as it relates to a contract timeline.

This information is primarily accessible by the overseeing agencies, Army Contracting Command, DCMA, and DCAA. For a full list of the feature names see Appendix B.

3. Scope of Data

The scope of the data selections for this study are the following: Department of the Army contracts; NAICS prefix 334 (computer and electronic product manufacturing), and fixed contact pricing. Our research approach is applicable to contracts with other attributes as well.

B. METHODOLOGY

In this section, we discuss our methodological approach to benchmarking anomaly detection techniques. Specifically, we review the data preparation of USAspending.gov, techniques used for dimensionality reduction, techniques used for anomaly detection, and the design of our study.

1. Data Preparation

After scoping the data set, 18.8 million observations, with 276 features, remain. Preprocessing further reduce the data set to 22,491 observations, with 39 features (see Appendix C). When converting the results into a numerical data matrix, a data set with dimensions of 22,491 observations and 86 features is created. We then split the data in to a training and test set. The training set is the portion of the data set that covers FY 2017 to the second-quarter of FY 2020. The training set consist of 21,193 observations. The test set is the portion of the data that covers the second-half of FY 2020, April to September. The test set consist of 1,310 observations, which includes the 12 prepended synthetic anomalies. Appendix C contains a list of the 39 features retained from the original 276 features.

The 12 synthetic anomalies are derived from three observations. The original data set we picked three at random as a base. Using the three observations as a basis, 12 anomalies were created based on sampling and replacing the binary features of the observations and manually changing the federal action obligation. Appendix D contains a list of these 33 binary features and one continuous value, the federal action obligation.

To distinguish between degrees of anomalous behavior, two sets of the 12 synthetic anomalies were created, "extremely anomalous" and "moderately anomalous." Extremely anomalous was designated this name based on randomly changing all 33 binary features, while moderately anomalous involved randomly changing five of the 33 features. Changing the federal action obligations to normalize values of 2.00, 1.00, 0.75, 0.50, and 0.25 was applied to both data sets.

2. Dimensionality Reduction

In this section, we identify the techniques used for the reduction of the data set's dimensionality. The objective of dimensionality reduction is to remove noise from the data and extract essential information with a smaller number of variables. We look closer at this objective by studying the use of no dimensionality reduction and dimensionality reduction using principal component analysis (PCA) and an autoencoder.

a. No Dimensionality Reduction

No dimensionality reduction serves as an additional test group to assess benchmarking results. We address whether or not using dimensionality reduction provides better capability for identifying anomalies in the USAspending.gov data. As explained in Section 3.B.1, the final data set has 86 dimensions.

b. Principal Component Analysis

Principal component analysis (Faraway 2016) serves as one of two dimensionality reducing techniques being assessed in this study. The findings from this research will identify if the PCA approach to using linear relationship is more effective on the USAspending.gov data set compared to the non-linear approach of an autoencoder, or a

non-dimensional approach. The number of dimensions for PCA will be discussed in Chapter IV, results and analysis.

c. Autoencoders

The use of the R software (R Core Team 2019) package **keras** to build an autoencoder (Allaire 2020) serves as the second dimensionality reducing technique being assessed in this study. Much like the PCA objective, assessing the effects of an autoencoder will identify if its non-linear approach to the USAspending.gov data set will show to be the most effective of the three techniques. The structuring and parameter selection of our autoencoder is addressed in Chapter IV. In this section, we describe the fundamentals of an autoencoder.

Autoencoders are neural networks where the input and output values are the same. Dertat describes how autoencoders serve as a dimensionality reducing algorithm (2017). Figure 1 gives a visualization of an autoencoder.

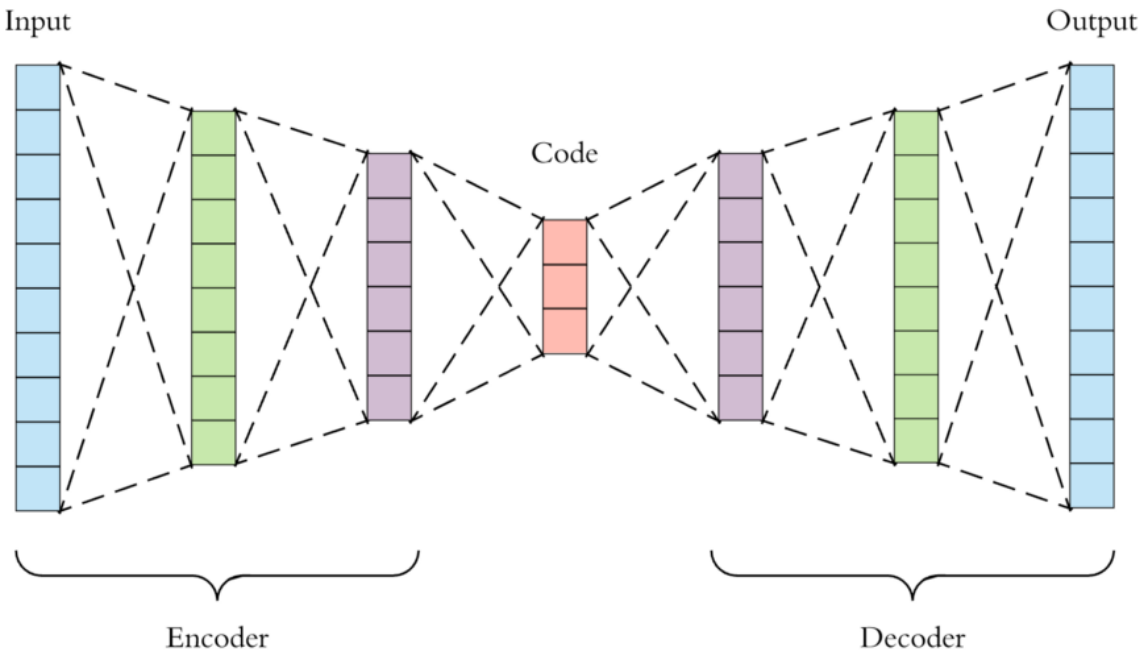

Figure 1. Visualization of an autoencoder. Source: Dertat (2017).

In Figure 1, the blue cubes, on the encoder half, represent the original input data values, while the green and purple represent compression of the data. At the code block, hyperparameters, set by the user, defines how the autoencoders process the data before passing the data forward to the decoder. As noted by Dertat, four hyperparameters are selected when training an autoencoder:

- Code size: Also known as latent size, it represents the most compressed state of the data set.

- Layers: The number of layers represents the depth of the autoencoder. Depth plays a role in data compression and training set size requirement. Figure 6 depicts an autoencoder with four layers, two in the encoder and two in the decoder.

- Neurons per layer: The numbers of neurons for each layer are set by the user. Selecting these parameters is computationally intensive as it involves measuring the predictive performance of the autoencoder for each possible configuration. Typically, the number of neurons decreases with each layer in the encoder, and increases in the decoder.

- Loss function: Calculating loss is computed with either one of two classes, binary classification and multiclass classification (Dertat 2017). In our research, we calculate loss using the binary classification, specifically the binary cross entropy.

3. Anomaly Detection

In this section, we briefly review anomaly detection and identify the techniques used to identify anomalies within the USAspendings.gov data set. The objective of using anomaly detection is to obtain a subset of records that are the best candidates for further inspection due to their unusual characteristics. The anomaly detection techniques that we consider are isolation forests (IF), clustering (density-based spatial clustering of applications with noise [DBSCAN] and k-means), and k-nearest neighbors (kNN).

a. Anomaly Detection Fundamentals

In this section, we review fundamental concepts of anomaly detection focusing on input data, types of anomalies, the labeling of the data set, and the output values of identified anomalies.

(1) Input Data

The data on which anomaly detection is applied typically consists of a large number of instances (observations), each of which contains measurements on a number of attributes (variables) (Tan et al. 2020). Attributes can be either binary, numerical, or categorical.

(2) Types of Anomalies

Anomalies can be partitioned into three types: point, contextual, and collective (Chandola et al. 2009). Figure 2 illustrates these anomaly types in a two-dimensional space. Non-anomalous clusters are designated N_1 and N_2, while points O_1, O_2, and O_3 are the three types of anomalies.

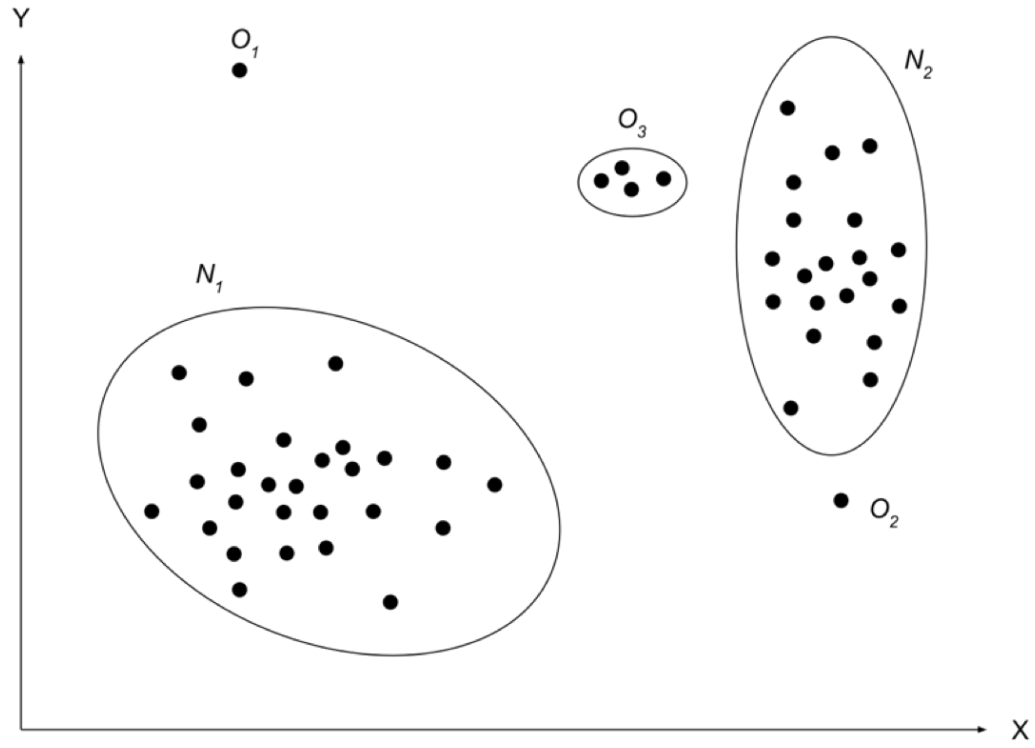

Figure 2. Illustration of simple anomalies in two-dimensional space. Source: Kibish (2018).

(3) Point Anomalies

Point anomalies are individual observations not adhering to an expected pattern, such as a single large financial transaction compared to historical purchases (Perera 2015). In Figure 2, O_1 and O_2 represents this type of anomaly (Kibish 2018).

(4) Contextual Anomalies

Contextual anomalies are observations that can be anomalous in one state and normal in another, such as an otherwise typical observation occurring outside an expected time frame (Perera 2015). Figure 3 illustrates an example of this type. In this example, O_1 would be labeled as an anomaly (Kibish 2018).

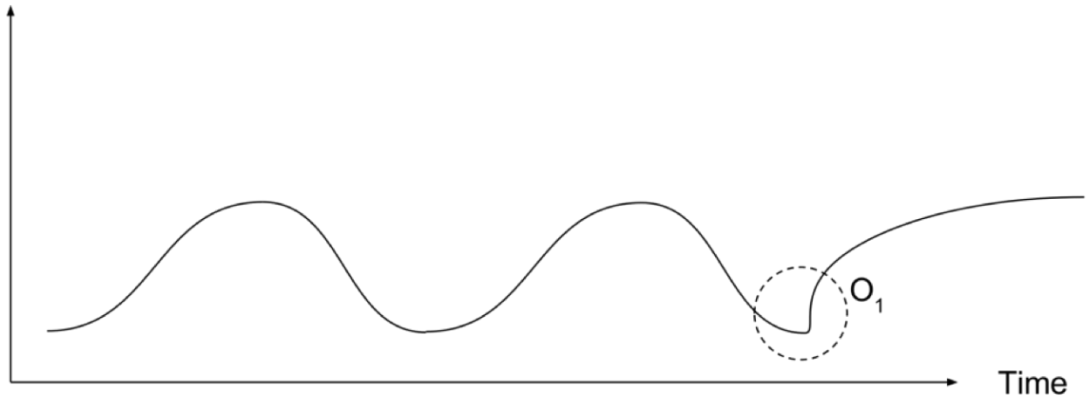

Figure 3. Contextual anomaly example. Source: Kibish (2018).

(5) Collective Anomalies

Collective anomalies are a group of similar observations that are all anomalous compared to the data set, and not based on individual value (Perera 2015). In Figure 2, O_3 represents a collective anomaly (Kibish 2018).

(6) Data Labels

The labeling of data identifies if a particular data object within the data set is either anomalous or not. Based on this knowledge, data set labeling falls into one of three

categories: supervised, semi-supervised, and unsupervised (Chandola et al. 2009). In Figures 4–6, anomalies are labeled as black circles, normal observations as white circles, and unknowns are grey circles.

(7) Supervised Anomaly Detection

In a supervised data set, all observations are labelled either normal or anomalous. At this point, a classifier can be trained (Kibish 2018). A classifier is an algorithm that sorts data based on their categories are either normal or anomalous (DeepAI 2020). Examples includes logistic regression, decision trees, and neural networks. Figure 4 depicts a supervised anomaly detection setup.

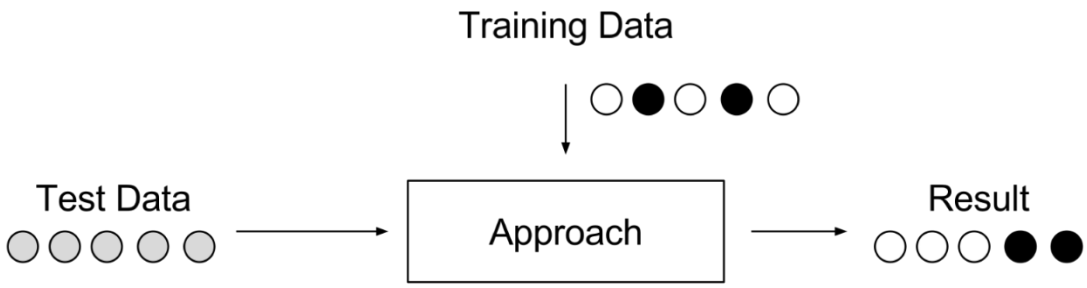

Figure 4. Supervised anomaly detection. Source: Kibish (2018).

(8) Semi-Supervised Anomaly Detection

In a semi-supervised data set, the training set consist of normal labelled observations. The idea is that the classifier trained on a data set where all observations are labelled as normal would, on new data, identify those observations that deviate from a normal pattern. (Kibish 2018). Figure 5 depicts a semi-supervised anomaly detection setup.

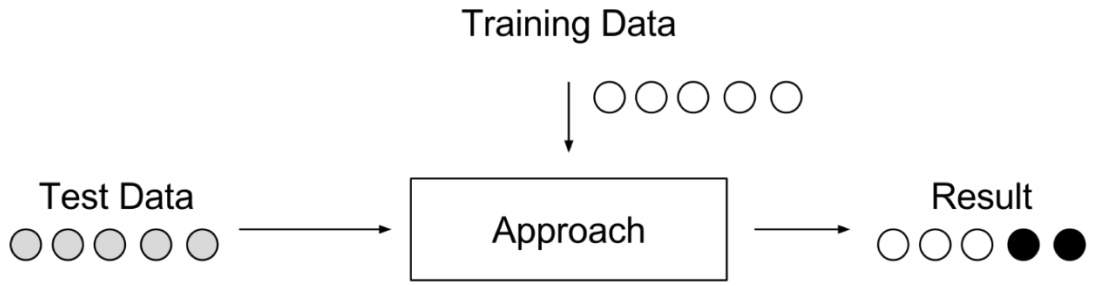

Figure 5. Semi-supervised anomaly detection. Source: Kibish (2018).

(9) Unsupervised Anomaly Detection

In an unsupervised anomaly detection data set, none of the observations are labelled as normal or anomalous. Anomalies are identified based solely on their deviations from the data set natural features (Kibish 2018). Figure 6 depicts an unsupervised anomaly detection setup.

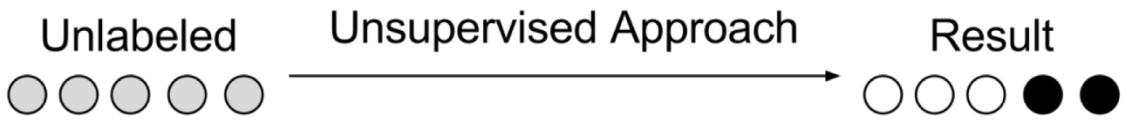

Figure 6. Unsupervised anomaly detection. Source: Kibish (2018).

(10) Output of Anomaly Detection

The output of an anomaly detection algorithm falls into one of two categories, scores or labels (Chandola et al. 2009). For this study, our focus is on a ranked list, which identifies the degree to which observations are anomalous or labeled as anomalous, in the case of DBSCAN.

(11) Scores

Scoring techniques, as noted by Chandola et al., is when each observation is assigned a score based on the degree in which they are considered an anomaly (2009). With the compilation of all anomaly scores, a ranked list can be generated.

(12) Labels

Chandola et al. further explain that classifiers used on a data set, that does not use a scoring system, label observations as either normal or anomalous (2009). In this research, labels are used when the algorithm being reviewed does not generate scores.

b. Isolation Forest

Like random forests an IF (David 2020) is built on the idea of decision trees. The main difference between the two is that IF scores the unusualness of observation relative to others. This process occurs through the algorithm's random selecting of features and values for the split (Liu 2020).

c. Clustering

Clustering is an unsupervised learning method with the goal of grouping like observations within a data set. The intent is to let a clustering algorithm take in inputs that have no labels and place them into similar groups based on the similarity of their features. McGregor (2020) notes that they are different types of clustering algorithms:

- Density-based: Data is grouped by concentration of instances, with higher being surrounded by areas of lower instances. DBSCAN (Hahsler et al. 2019) is an example of a density-based algorithm.

- Distribution-based: Data is grouped based on the probability that they belong to a given cluster. The higher the probability, the closer the data to the center, and the lower the probability the further the data from the center. Expectation-maximization is an example of a distribution-based algorithm.

- Centroid-based: Data is grouped based its squared distance from the centroid, or cluster center. K-means clustering is an example of a centroid-based algorithm.

- Hierarchical-based: Data is grouped based on the dataset predominant ordering. Balance Iterative Reducing and Clustering using Hierarchies is an example of a hierarchical-based algorithm (McGregor 2020).

Figure 7 depicts the use of DBSCAN as a clustering algorithm. In Figure 7, two clusters are present, blue and yellow rings of data points. The black points are anomalies or outliers. Figure 7 also illustrates one of the benefits on DBSCAN, the ability to find non-linear clusters.

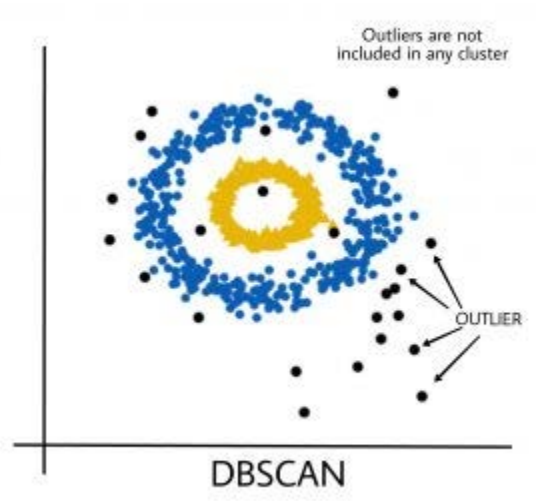

Figure 7. Illustration of a DBSCAN output. Source: Dey (2019).

For our research, the two clustering techniques assessed are the centroid-based algorithm, k-means clustering, and the density-based algorithm, DBSCAN. Optimum parameters selection for k-means clustering and DBSCAN is addressed in Chapter IV.

d. K-Nearest Neighbors

K-nearest neighbor is a non-parametric classification method that identifies an unknown observation based on the surrounding observations. Simply, the observations that are in the majority are used to label the unknown observation. The parameter used in this labeling process is user defined. Specifically, the user defines the number of neighbors to assess in identifying a label for the unknown observation. In our research, we use distance as a measure of the degree in which an observation is anomalous. K-nearest neighbor serves as the last of the four anomaly detection techniques being reviewed for their effectiveness on the USAspending.gov data set. The parameter value for kNN is discussed in Chapter IV.

4. Design

In this subsection, we describe the study design that we use to address our research questions.

a. Design Overview

The design of our study follows three steps in benchmarking anomaly detection on contract data from USAspending.gov. The first step is data preparation and separation for testing. The second step is model selection and parameter identification for the studied techniques. The third step is implementation and analysis of results.

(1) Step 1: Data Preparation and Separation

This step involves preprocessing the data, identifying training and test sets, and creating synthetic anomaly data. Preprocessing of the USAspending.gov data set follows the process described in Chapter III, Section B.1.

(2) Step 2: Model selection and Parameter Identification

This involves identifying the structuring and parameters that are used with dimensionality reduction and anomaly detection algorithms. For dimensionality reduction, PCA and autoencoder, the number of dimensions to retain and hyperparameters are derived, respectively. With the anomaly detection algorithms, k-means clustering,

DBSCAN, and kNN, parameters for the number of clusters, the minimum number of points clustered together, and the number of nearest neighbors, respectively, are identified.

(3) Step 3: Research Design and Evaluation Metrics

The design of the research is focused on comparing three dimensionality reduction techniques, on which each anomaly detection algorithm is applied. This setup is the basis for the benchmarking comparison between the type of dimensionality reduction, and the anomaly detection algorithm that is the more effective combination. Figure 8 provides a description of the research design.

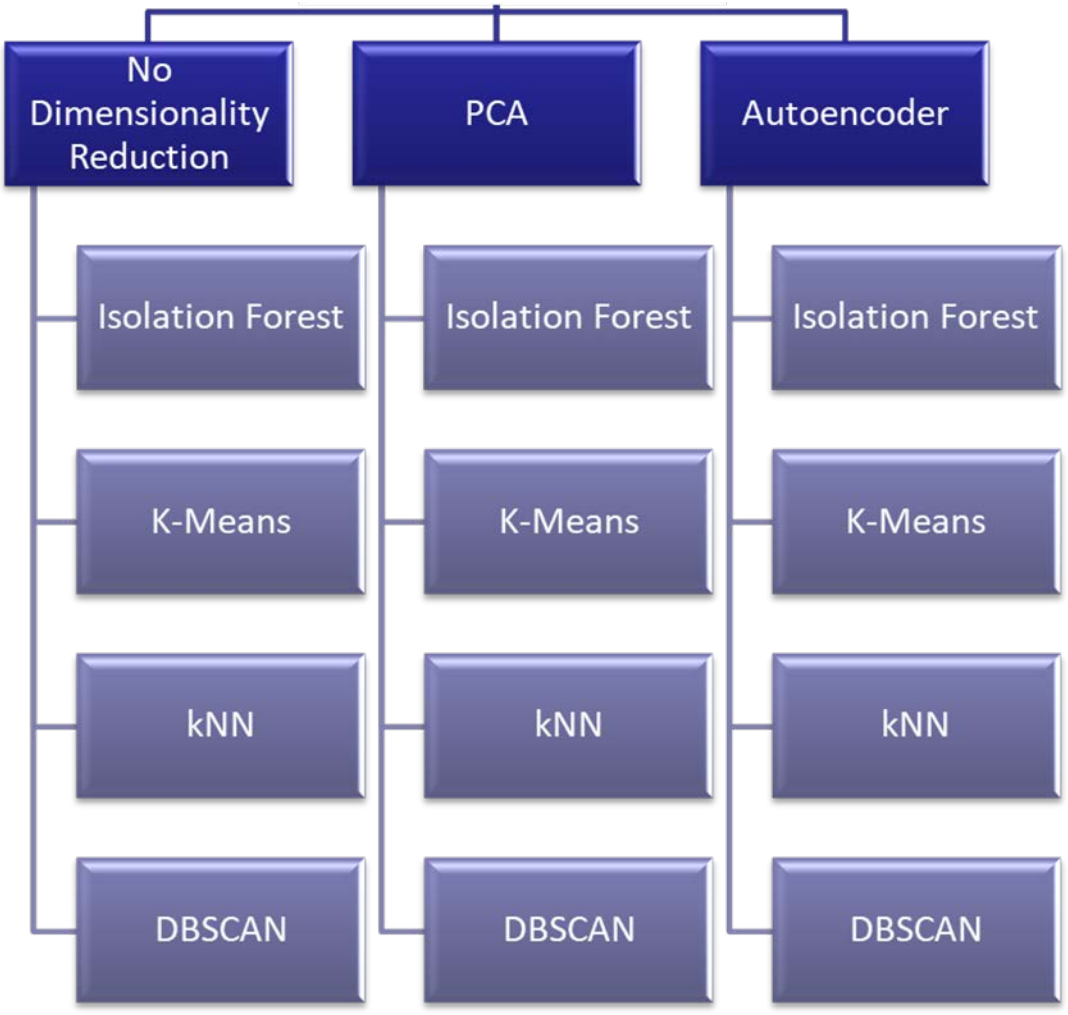

Figure 8. Visual depiction of research design.

We employ two evaluation metrics that we call “Top 100” and “Depth to Half.” For Top 100, we count the number of the 12 synthetic anomalies that we placed in the test data, that occurred among the 100 observations that have the highest anomaly scores, or in the case of DBSCAN, that are not assigned to a cluster. The measure of effectiveness is based on the number of the 12 synthetic anomalies, that we placed in the test data, that appears in the Top 100. For Depth to Half, we take the ranking (largest to smallest) of anomaly scores, at which the sixth seeded anomaly appears. For example, if the sixth seeded anomaly appears at number 328 in the test set, then the Depth to Half score is 328.

IV. RESULTS AND ANALYSIS

In this chapter, we present the results of our research. We use the R programming language for all calculations. First, we identify the parameters for the dimensionality reduction and anomaly detection algorithms. Then, we compare and discuss the results.

A. PARAMETER SELECTION

In this section, we identify the parameters used for the dimensionality reduction and anomaly detection algorithms. We begin by identifying parameters of the two dimensionality reduction techniques, PCA and autoencoder. Then, we identify the parameters of the anomaly detection techniques, k-means clustering, DBSCAN, and kNN.

1. Dimensionality Reduction

a. Principal Components

We retain 20 principal components, which explain approximately 80% of total variability. As depicted in Figure 9, this choice is based on comparing the cumulative proportion of variability with the number of principal components. The elbow of the graph can be seen at around 20 principal components mark.

b. Autoencoder

We use the following parameter settings for an autoencoder applied to the USAspending.gov learning data: three encoder and decoder layers, 2048 neurons per layer, and a latent size of 20. The latent size is comparable to the number of dimensions to which the data are reduced. Figure 10 depicts the various autoencoder models evaluated, using a latent size of 10. The reconstruction error is near its minimum with three layers, and 2048 neurons per layer. Reconstruction error is the average distance from each row in the data matrix to its autoencoder reconstruction.

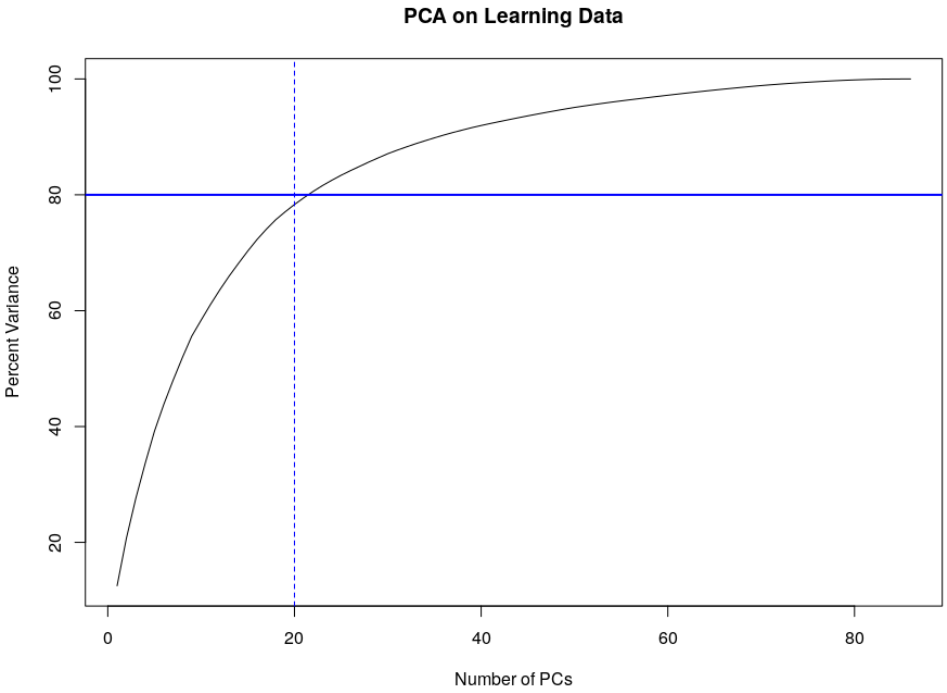

Figure 9. Selection of number of principal components.

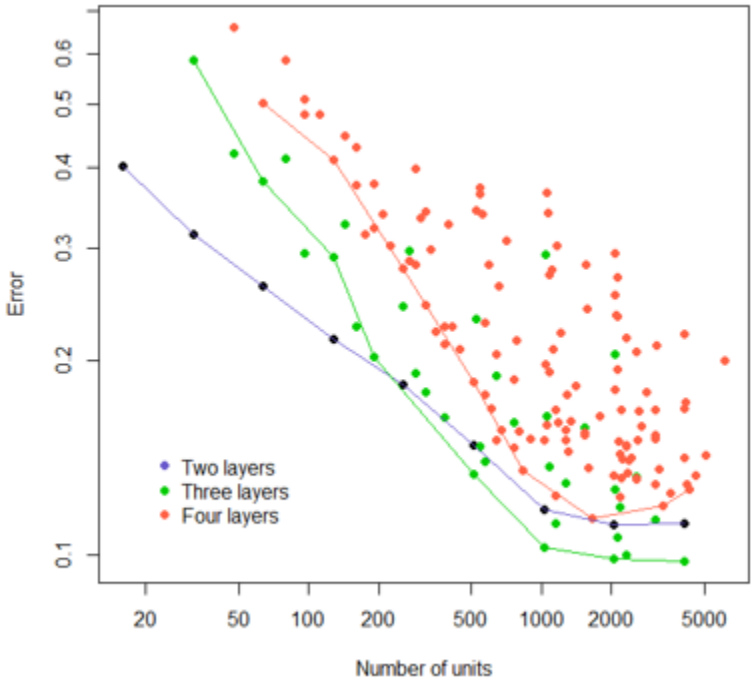

Figure 10. Selection of number of encoder layers and neurons per layer.

In Figure 11 we evaluate autoencoders with latent sizes ranging from four to 39, fixing the number of layers at three and the number of neurons per layer of 2048. We average the results of four runs of this analysis. Figure 11 shows that the reconstruction error nearly achieves a minimum at a latent size of 20.

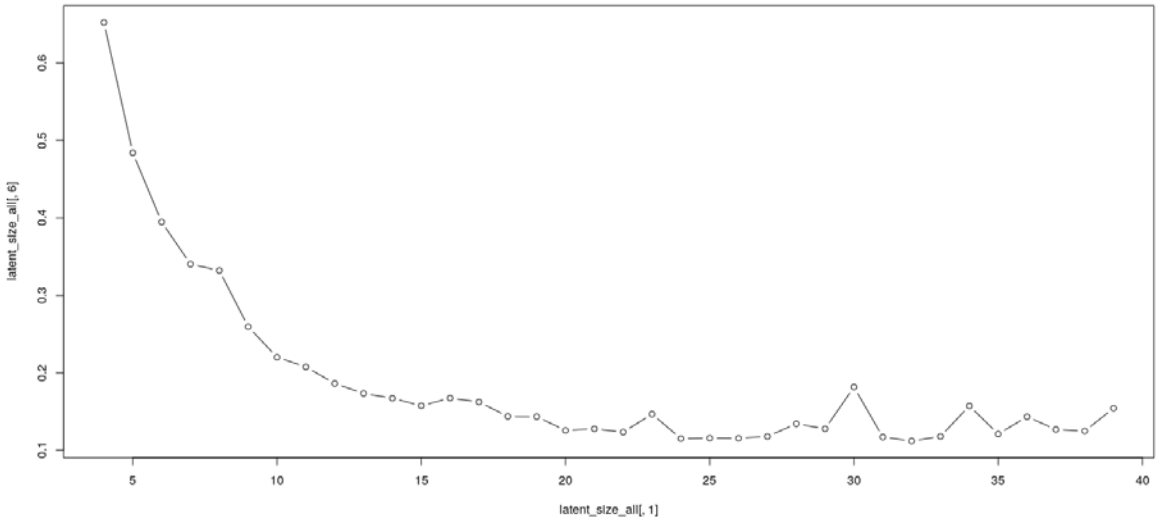

Figure 11. Selection of number of latent variables

2. Anomaly Detection

a. K-Means Clustering

The number of clusters selected for k-means clustering is 30. Five approaches were used to identify the number of clusters for the USAspending.gov data set: sum of squares of errors (SSE), the function **pamk** of the flexible procedures for clustering (**fpc)** package (Hennig 2020) in the R programming language; the function **kmeansrun** of the **fpc** package (Hennig 2020); the function **nclcust** of the **NbBlust** package (Charrad et al. 2014); and the function **mcclust** of the **McClust** package (Fritsch 2012). Table 1 gives the results from applying all five methods. The **mcclust** method suggests three clusters, which is unusually small compared to the others. We decided to use 30 clusters, which is more typical of the methods evaluated.

Table 1. Number of cluster determination for k-means clustering

Algorithms	Number of Clusters
SSE	30
Pamk	24
Kmeansrun	30
NbClust	34
McClust	3

b. DBSCAN

The minimum number of observations to form a region for USAspending.gov is 87 for the no dimensionality reducing technique, and 21 for both of the dimensionality reducing techniques. These values are based on the rule of thumb for selection, "dimensionality plus one" (Hahsler et al. 2019).

c. K- Nearest Neighbors

The number of nearest neighbors for classification in theUSAspending.gov data set is 145. This value is derived from the rule of thumb (Hassanat et al. 2014) of taking the square root of the number of observations in the training data set. The final value is rounded down to 145.

B. RESULTS

In this section, we discuss the results of our study. We begin by analyzing both sets of anomalous data, extremely and moderately. Recall that "extremely anomalous" refers to 33 binary features assigned values randomly, and "moderately anomalous" refers to five randomly selected binary features assigned values randomly. In both cases, 12 anomalies are generated using three actual records from USAspending.gov with four generated records in each case. With both sets of anomalous data, we further examine the effects of adjusting the federal action obligation to the following values: 2.00, 1.00, 0.75, 0.50, 0.25. Table 2 list the federal action obligation (FAO) range for both the training and test sets.

The manually adjusted values reflects values out side the maximum range, near the lower minimum range, and various levels in between.

Table 2. Federal action obligation normalized range for training and test sets

Range	Training Set	Test Set
Minimum	0.0000	0.4018
First Quartile	0.3665	0.6329
Median	0.4405	0.6710
Mean	0.4520	0.6837
Third Quartile	0.5206	0.7205
Maximum	1.0000	0.9962

1. Federal Action Obligation Value of 2.00

Figure 12 compares extremely and moderately anomalous data, with an FAO value of 2.00, on the two metrics consider: Top 100 and Depth to Half. With an FAO of 2.00, the value is outside the maximum quartile for the training and test sets. In all, IF using no dimensionality reduction is the most effective combination of methods and techniques. Focusing on the extremely anomalous cases, both IF and the autoencoder are equally effective with the Top 100 and Depth to Half metrics. With the moderately anomalous cases, a separation in performance is observed between the dimensionality reduction methods. For the Depth to Half results, PCA is the least effective. When considering k-means clustering using PCA, the final result is 493 with Depth to Half. This number represents the number of records an investigator would have to examine for half the anomalies. In this case, the investigator would have to examine nearly 38% of their records.

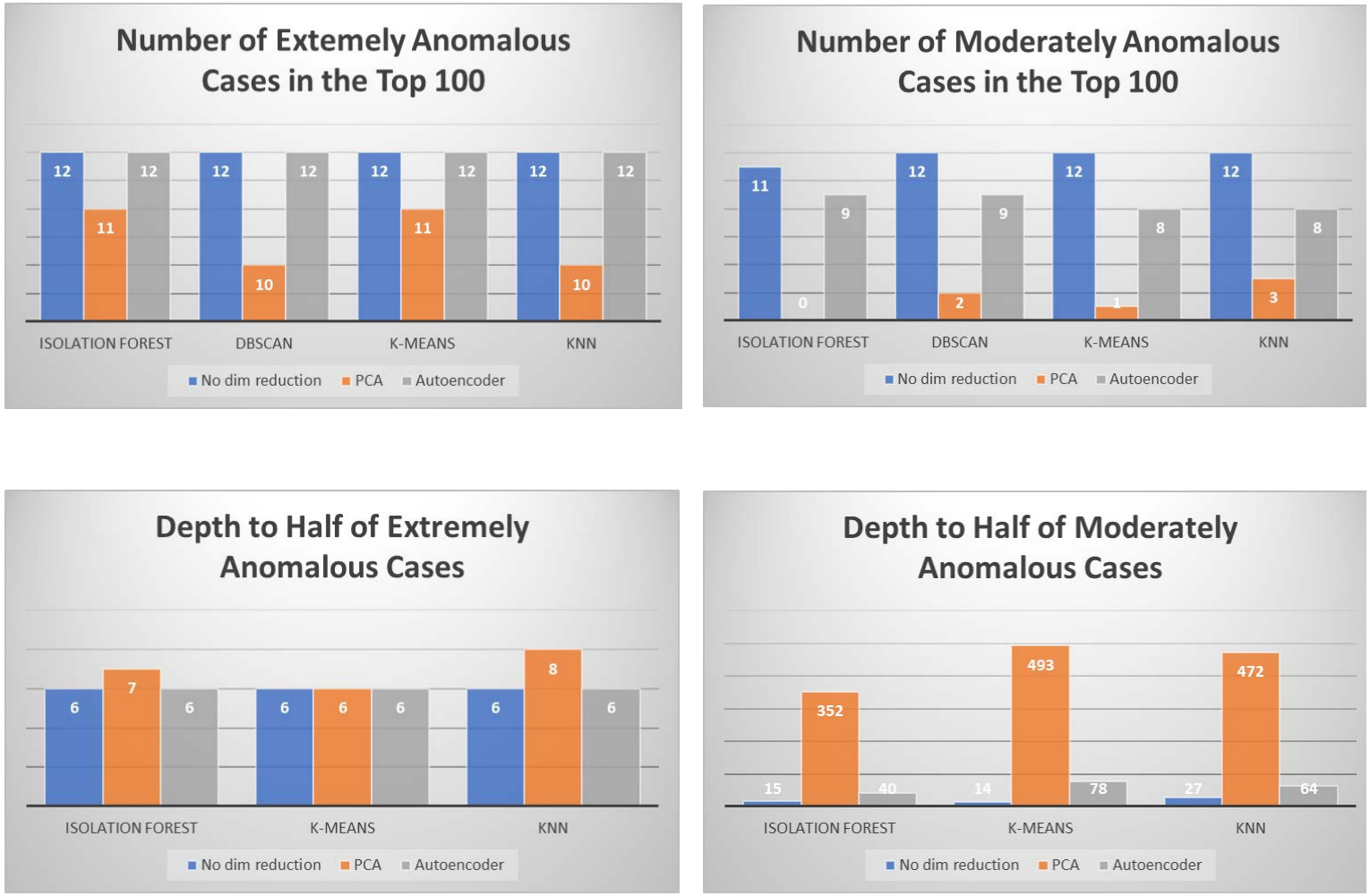

Figure 12. Comparison of Top 100 and Depth to Half results with an FAO value of 2.00.

2. Federal Action Obligation Value of 1.00

In this section, we review Figure 13, which compares extremely and moderately anomalous data with an FAO value of 1.00. At 1.00, the FAO value is at the maximum quartile for both the training and test sets. Even with the FAO value decreasing by 50% from the original 2.00, in extremely anomalous cases, IF and the autoencoder perform nearly equally as well as they did in Figure 12. With Depth to Half, a division in performance between no dimensionality reduction and dimensionality reduction is seen. For moderately anomalous cases, IF out performs the dimensionality reduction methods. Notably, both k-means clustering and kNN using the autoencoder shows improvement in the Depth to Half metrics compared to Figure 12. The expected result for k-means clustering and kNN using the autoencoder is a decrease in effectiveness, similar to the anomaly detection techniques using PCA. This expectation is based on the FAO value not being as extreme as 2.00.

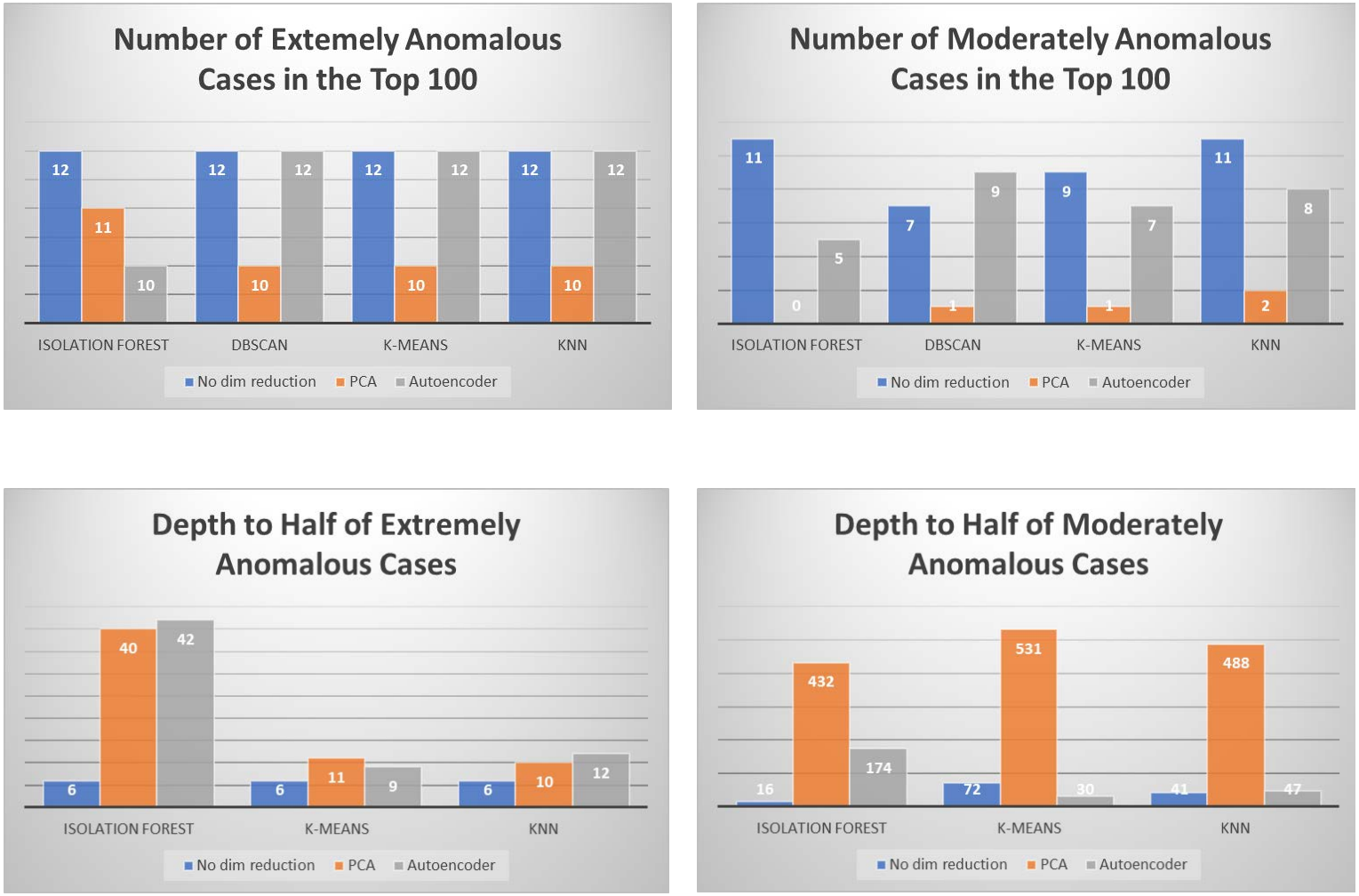

Figure 13. Comparison of Top 100 and Depth to Half results with an FAO value of 1.00.

3. Federal Action Obligation Value of 0.75

Figure 14 present results of the extremely and moderately anomalous data with an FAO value of 0.75. At 0.75, the seeded anomalies value remains near the maximum quartile for the training set, and third quartile for the test set. The FAO value represents a 62.5% decrease from 2.00. Again, for the extreme anomalous cases, IF and the autoencoder perform equally well, while PCA performs the worst for Top 100 and Depth to Half. In the moderate anomalous cases, the trend of no dimensionality reduction continues as the most effective, and PCA the least. A notable occurrence, k-means clustering using PCA shows an increase in effectiveness compared to Figure 12 and Figure 13.

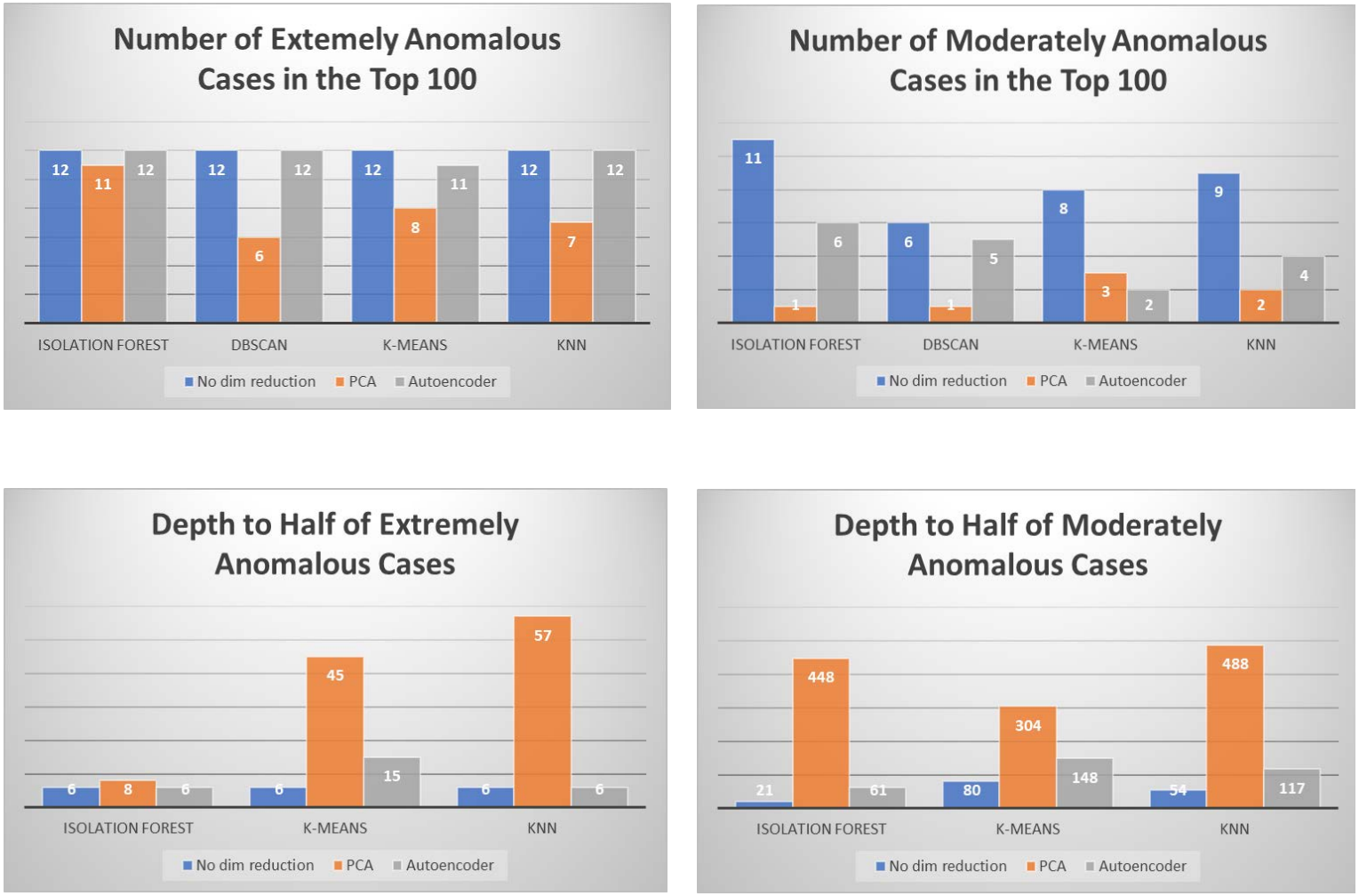

Figure 14. Comparison of Top 100 and Depth to Half results with an FAO value of 0.75.

4. Federal Action Obligation Value of 0.50

Figure 15 exhibits results of extremely and moderately anomalous data with an FAO value of 0.50. A FAO of 0.50 reflects the third quartile for the training set and the first quartile for the test set. At 0.50, this value represents a 75% decrease from 2.00. With a decrease in value of 75%, IF using no dimensionality reduction is the most effective in all four graphs, closely followed by kNN using no dimensionality reduction. Considering overall trends, Figure 15 moderately anomalous results, not including DBSCAN, are close in value with Figure 14. This behavior is likely due to the FAO values being on the tail ends with the maximum and minimum quartiles. This observation also includes Figure 16, FAO value of 0.25. However, the autoencoder with a FAO 0.50, Figure 15, is not representative of this pattern.

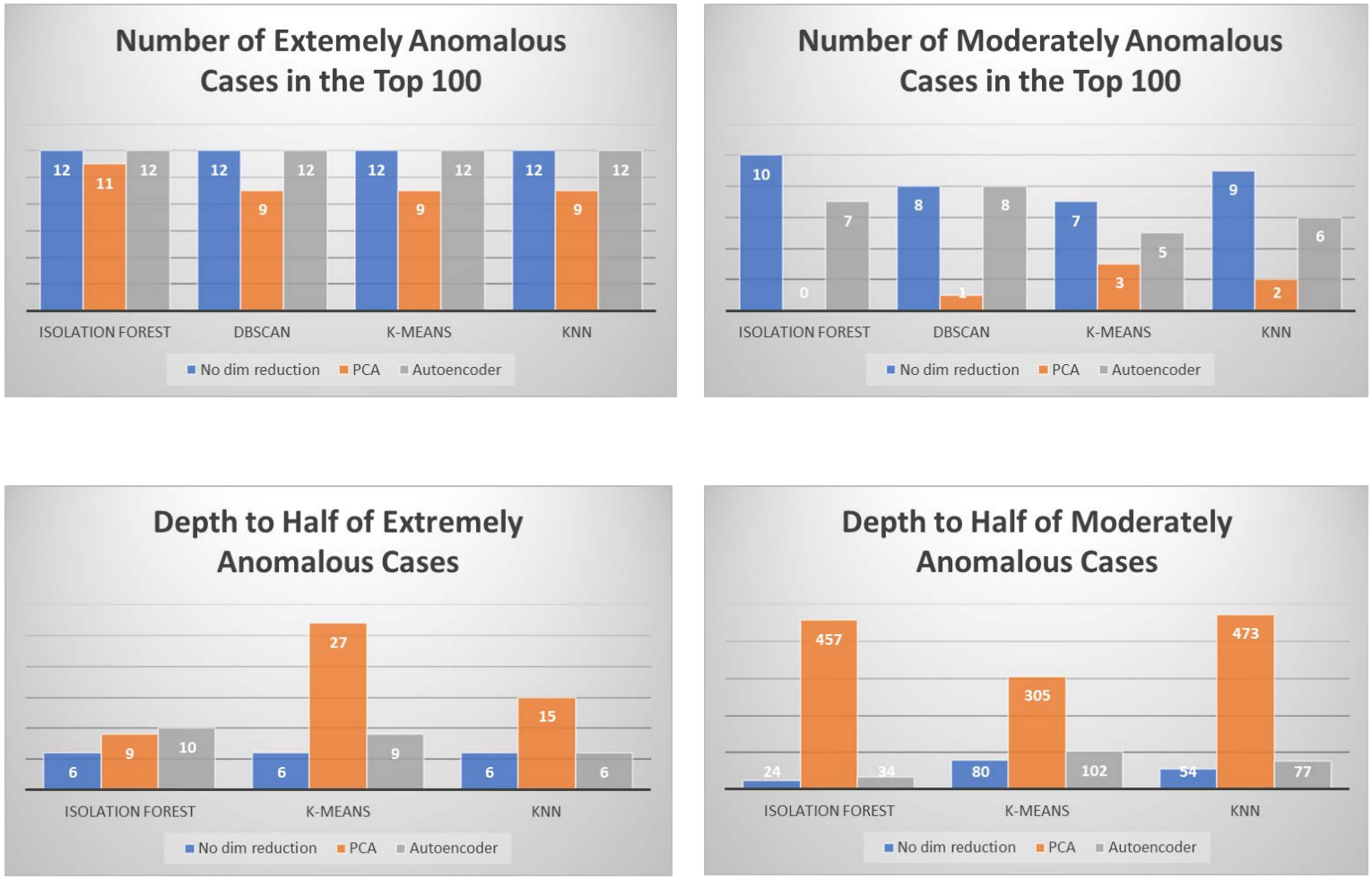

Figure 15. Comparison of Top 100 and Depth to Half results with an FAO value of 0.50.

5. Federal Action Obligation Value of 0.25

Figure 16 compares extremely and moderately anomalous data, with an FAO value of 0.25. With an FAO value of 0.25, this value is near the first quartile for the training set, and outside the minimum quartile for the test set. At 0.25, this value represents an 87.5% decrease from 2.00. In all, the most effective dimensionality reduction method is no dimensionality reduction, followed by the autoencoder, and then PCA. In the extremely anomalous case, both no dimensionality reduction and the autoencoder performed similarly. In the moderately anomalous case, no dimensionality reduction is the only method that shows improvement with all anomaly detection techniques compared to Figure 15. Again, this similarity is likely due to FAO values of 0.50 and 0.25 being near the minimum quartile.

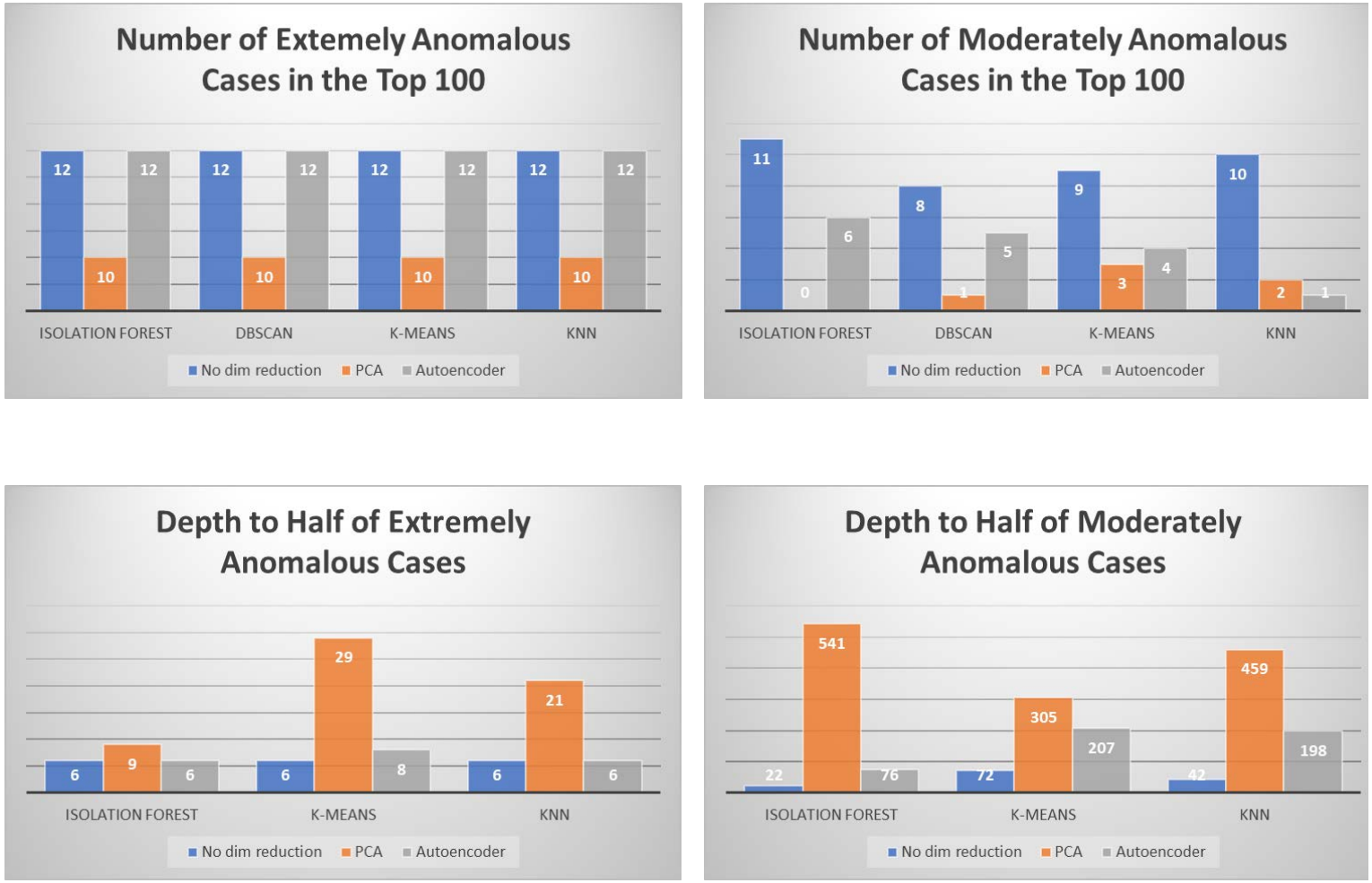

Figure 16. Comparison of Top 100 and Depth to Half results with an FAO value of 0.25.

6. Summary of Results

In this section, we provide a side-by-side comparison of the Top 100 and Depth to Half. This section focuses on the moderately anomalous results due to the extremely anomalous showing very little overall change between FAO values. Each anomaly detection technique is displayed in its own table. Tables 3 to 6 list the results of each anomaly detection technique in the order of: IF, DBSCAN, k-means clustering, and kNN.

Table 3. Combined isolation forest results for moderately anomalous

Isolation Forest	Federal Action Obligation Values				
	2.00	**1.00**	**0.75**	**0.50**	**0.25**
Top 100 - No Dim Reduction	11	11	11	10	11
- PCA	0	0	1	0	0
- Autoencoder	9	5	6	7	6
Depth to Half - No Dim Reduction	15	16	21	24	22
- PCA	352	432	448	457	541
- Autoencoder	40	174	61	34	76

Table 4. Combined DBSCAN results for moderately anomalous

DBSCAN	Federal Action Obligation Values				
	2.00	**1.00**	**0.75**	**0.50**	**0.25**
Top 100 - No Dim Reduction	12	7	6	8	8
- PCA	2	1	1	1	1
- Autoencoder	9	9	5	8	5

Table 5. Combined k-means clustering results for moderately anomalous

K-Means Clustering	Federal Action Obligation Values				
	2.00	**1.00**	**0.75**	**0.50**	**0.25**
Top 100 - No Dim Reduction	12	9	8	7	9
- PCA	1	1	3	3	3
- Autoencoder	8	7	2	5	4
Depth to Half - No Dim Reduction	15	72	80	80	72
- PCA	352	531	304	305	305
- Autoencoder	40	30	148	102	207

Table 6. Combined k-nearest neighbor results for moderately anomalous

K-Nearest Neighbor	Federal Action Obligation Values				
	2.00	**1.00**	**0.75**	**0.50**	**0.25**
Top 100 - No Dim Reduction	12	11	9	9	10
- PCA	3	2	2	2	2
- Autoencoder	8	8	4	6	1
Depth to Half - No Dim Reduction	27	41	54	54	42
- PCA	472	488	488	473	459
- Autoencoder	64	47	117	77	198

Table 7 gives the averages of Figure 12 to Figure 31. The averages are calculated based on using the number of anomalies detected within the Top 100, and the Depth to Half of the anomalies, against the number of runs for the FAO values. For Top 100, higher values are more efficient, while for Depth to Half, lower values are more efficient.

Table 7. Averages of anomaly detection results

Dimensionality Reduction Techniques	Anomaly Detection Algorithms			
	IF	DBSCAN	K-Means	kNN
Top 100 - No Dim Reduction	10.8	8.2	9.0	10.2
- PCA	0.2	1.2	2.2	2.2
- Autoencoder	6.6	7.2	5.2	5.4
Depth to Half - No Dim Reduction	19.6	--	63.8	43.6
- PCA	446.0	--	359.4	476.0
- Autoencoder	77.0	--	105.4	100.6

Figure 17 compares the total number of anomalies for each dimensionality reduction method averaging over the five values of federal action obligation for synthetic anomalies. Again, no dimensionality reduction dominates PCA and autoencoder overall. Between the two dimensionality reduction methods considered, autoencoder performs somewhat better than PCA. Differentiation the three methods is greater when the degree of anomaly is less severe. These observations are reflected both in the average percentage of anomalies encountered in the top 100 records, and in the number of records that must be inspected to reach half of the anomalies.

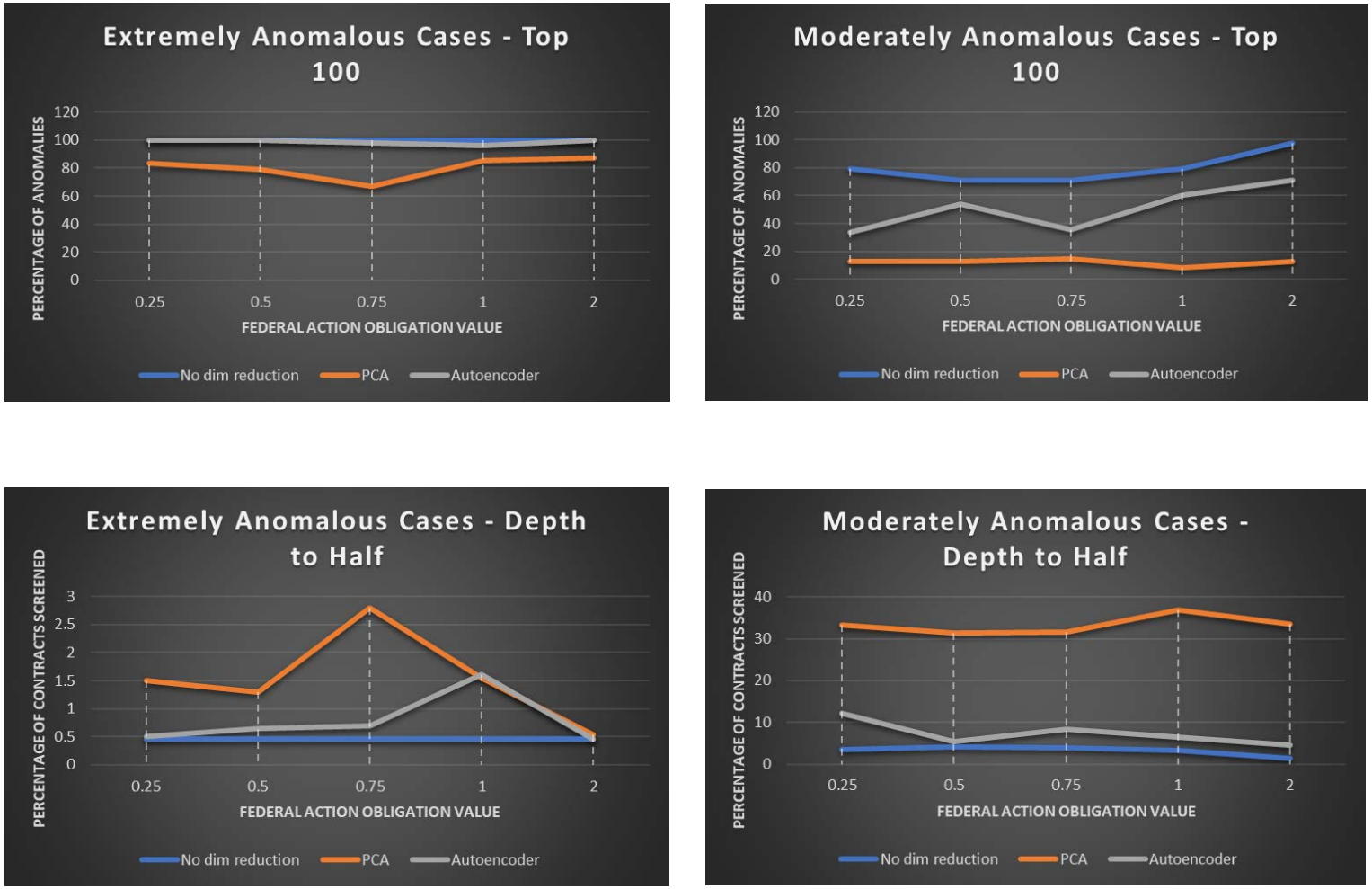

Figure 17. Comparison of Top 100 and Depth to Half total number of anomalies for all FAO values

V. CONCLUSION

From the beginning of fiscal year 2020, until the end of August, the Army obligated nearly $77 billion for defense contracts. In an effort to provide visibility of expenditures, not only for the Army, but the federal government as a whole, the Federal Funding Accountability and Transparency Act of 2006 was passed into law. With its enhancement, the Digital Accountability and Transparency Act of 2014, the USAspending.gov website was created. The purpose of this website is to provide data on spending expenditures for the U.S. public. To ensure compliance with statutory and regulatory requirements, oversight agencies employ techniques, to include anomaly detection, to assist them.

Revisiting the purpose of this thesis, which is to identify algorithms that are effective for generating a recommenders list of contracts for further review, two questions were posed:

Question 1: How do dimensionality reduction methods compare in their performance?

Question 2: How do anomaly detection algorithms compare in their performance?

To answer the research questions, we consider 120 distinct settings to make comparisons:

- Three dimensionality reduction methods
- Four anomaly detection algorithms
- Two types of anomalous data sets
- Five values of federal action obligation for synthetic anomalies

The criteria for evaluating performance is based on a scenario in which 100 contracts can be inspected by an investigator, with the objective of providing the investigator those contracts that standout as most unusual with respect to others. In this sense, we are using an anomaly detection method as a means of generating a recommender's list. The test data that we use is seeded with 12 artificial records that are truly anomalous. For performance metrics, we calculate the number of those records that

appear in the recommender's list, and the number of records that must be inspected, in decreasing order of anomaly, to reach the sixth seeded anomaly.

Question 1. We compare three dimensionality reduction methods applied to the USAspending.gov data: no dimensionality reduction, implying 86 dimensions; PCA, implying 20 dimensions; and autoencoder, implying 20 dimensions. We observe no benefit from the use of PCA or an autoencoder relative to no dimensionality reduction, holding all other factors fixed. This surprising result may be explained by the fact that the USAspending.gov data is highly categorical, with only one continuous variable. We do not imply that our results would apply to data set that contains many continuous variables. Of the two methods that significantly reduced dimension, the autoencoder outperforms the PCA.

Question 2: We compare four anomaly detection methods: isolation forest, DBSCAN, k-means clustering, and k-nearest neighbors. Isolation forest, with no dimensionality reduction, outperforms the other techniques across a wide range of circumstances. With no dimensionality reduction and across a range of settings, on average IF found 10.8 out of 12 seeded anomalies. Dimensionality reduction, however, adversely affects the performance of IF on the USAspending.gov data to a greater extent than the other anomaly detection methods.

Based on these results, the most efficient combination of algorithms to use with the USAspending.gov data set is isolation forest without reducing the dimensionality. In another data set, one that may have more continuous values, the results would likely be different then what this research has revealed.

APPENDIX A. NAICS PREFIX 334 SUBCATEGORIES

NAICS	Category
334	**Computer and Electronic Product Manufacturing**
3341	**Computer and Peripheral Equipment Manufacturing**
33411	Computer and Peripheral Equipment Manufacturing
334111	Electronic Computer Manufacturing
334112	Computer Storage Device Manufacturing
334118	Computer Terminal and Other Computer Peripheral Equipment Manufacturing
3342	**Communications Equipment Manufacturing**
33421	Telephone Apparatus Manufacturing
334210	Telephone Apparatus Manufacturing
33422	Radio and Television Broadcasting and Wireless Communications Equipment Manufacturing
334220	Radio and Television Broadcasting and Wireless Communications Equipment Manufacturing
33429	Other Communications Equipment Manufacturing
334290	Other Communications Equipment Manufacturing
3343	**Audio and Video Equipment Manufacturing**
33431	Audio and Video Equipment Manufacturing
334310	Audio and Video Equipment Manufacturing
3344	**Semiconductor and Other Electronic Component Manufacturing**
33441	Semiconductor and Other Electronic Component Manufacturing
334412	Bare Printed Circuit Board Manufacturing
334413	Semiconductor and Related Device Manufacturing
334416	Capacitor, Resistor, Coil, Transformer, and Other Inductor Manufacturing
334417	Electronic Connector Manufacturing
334418	Printed Circuit Assembly (Electronic Assembly) Manufacturing
334419	Other Electronic Component Manufacturing
3345	**Navigational, Measuring, Electromedical, and Control Instruments Manufacturing**
33451	Navigational, Measuring, Electromedical, and Control Instruments Manufacturing
334510	Electromedical and Electrotherapeutic Apparatus Manufacturing
334511	Search, Detection, Navigation, Guidance, Aeronautical, and Nautical System and Instrument Manufacturing
334512	Automatic Environmental Control Manufacturing for Residential, Commercial, and Appliance Use
334513	Instruments and Related Products Manufacturing for Measuring, Displaying, and Controlling Industrial Process Variables

334514	Totalizing Fluid Meter and Counting Device Manufacturing
334515	Instrument Manufacturing for Measuring and Testing Electricity and Electrical Signals
334516	Analytical Laboratory Instrument Manufacturing
334517	Irradiation Apparatus Manufacturing
334519	Other Measuring and Controlling Device Manufacturing
3346	**Manufacturing and Reproducing Magnetic and Optical Media**
33461	Manufacturing and Reproducing Magnetic and Optical Media
334613	Blank Magnetic and Optical Recording Media Manufacturing
334614	Software and Other Prerecorded Compact Disc, Tape, and Record Reproducing

APPENDIX B. USASPENDING.GOV DATA COLUMN LABELS

Columns	Labels
1	contract_transaction_unique_key
2	contract_award_unique_key
3	award_id_piid
4	modification_number
5	transaction_number
6	parent_award_agency_id
7	parent_award_agency_name
8	parent_award_id_piid
9	parent_award_modification_number
10	federal_action_obligation
11	total_dollars_obligated
12	base_and_exercised_options_value
13	current_total_value_of_award
14	base_and_all_options_value
15	potential_total_value_of_award
16	action_date
17	action_date_fiscal_year
18	period_of_performance_start_date
19	period_of_performance_current_end_date
20	period_of_performance_potential_end_date
21	ordering_period_end_date
22	solicitation_date
23	awarding_agency_code
24	awarding_agency_name
25	awarding_sub_agency_code
26	awarding_sub_agency_name
27	awarding_office_code
28	awarding_office_name
29	funding_agency_code
30	funding_agency_name
31	funding_sub_agency_code
32	funding_sub_agency_name
33	funding_office_code
34	funding_office_name
35	treasury_accounts_funding_this_award
36	federal_accounts_funding_this_award
37	foreign_funding
38	foreign_funding_description
39	sam_exception

40 sam_exception_description
41 recipient_duns
42 recipient_name
43 recipient_doing_business_as_name
44 cage_code
45 recipient_parent_duns
46 recipient_parent_name
47 recipient_country_code
48 recipient_country_name
49 recipient_address_line_1
50 recipient_address_line_2
51 recipient_city_name
52 recipient_state_code
53 recipient_state_name
54 recipient_zip_4_code
55 recipient_congressional_district
56 recipient_phone_number
57 recipient_fax_number
58 primary_place_of_performance_country_code
59 primary_place_of_performance_country_name
60 primary_place_of_performance_city_name
61 primary_place_of_performance_county_name
62 primary_place_of_performance_state_code
63 primary_place_of_performance_state_name
64 primary_place_of_performance_zip_4
65 primary_place_of_performance_congressional_district
66 award_or_idv_flag
67 award_type_code
68 award_type
69 idv_type_code
70 idv_type
71 multiple_or_single_award_idv_code
72 multiple_or_single_award_idv
73 type_of_idc_code
74 type_of_idc
75 type_of_contract_pricing_code
76 type_of_contract_pricing
77 award_description
78 action_type_code
79 action_type
80 solicitation_identifier
81 number_of_actions
82 inherently_governmental_functions
83 inherently_governmental_functions_description

84 product_or_service_code
85 product_or_service_code_description
86 contract_bundling_code
87 contract_bundling
88 dod_claimant_program_code
89 dod_claimant_program_description
90 naics_code
91 naics_description
92 recovered_materials_sustainability_code
93 recovered_materials_sustainability
94 domestic_or_foreign_entity_code
95 domestic_or_foreign_entity
96 dod_acquisition_program_code
97 dod_acquisition_program_description
98 information_technology_commercial_item_category_code
99 information_technology_commercial_item_category
100 epa_designated_product_code
101 epa_designated_product
102 country_of_product_or_service_origin_code
103 country_of_product_or_service_origin
104 place_of_manufacture_code
105 place_of_manufacture
106 subcontracting_plan_code
107 subcontracting_plan
108 extent_competed_code
109 extent_competed
110 solicitation_procedures_code
111 solicitation_procedures
112 type_of_set_aside_code
113 type_of_set_aside
114 evaluated_preference_code
115 evaluated_preference
116 research_code
117 research
118 fair_opportunity_limited_sources_code
119 fair_opportunity_limited_sources
120 other_than_full_and_open_competition_code
121 other_than_full_and_open_competition
122 number_of_offers_received
123 commercial_item_acquisition_procedures_code
124 commercial_item_acquisition_procedures
125 small_business_competitiveness_demonstration_program
126 simplified_procedures_for_certain_commercial_items_code
127 simplified_procedures_for_certain_commercial_items

128 a76_fair_act_action_code
129 a76_fair_act_action
130 fed_biz_opps_code
131 fed_biz_opps
132 local_area_set_aside_code
133 local_area_set_aside
134 price_evaluation_adjustment_preference_percent_difference
135 clinger_cohen_act_planning_code
136 clinger_cohen_act_planning
137 materials_supplies_articles_equipment_code
138 materials_supplies_articles_equipment
139 labor_standards_code
140 labor_standards
141 construction_wage_rate_requirements_code
142 construction_wage_rate_requirements
143 interagency_contracting_authority_code
144 interagency_contracting_authority
145 other_statutory_authority
146 program_acronym
147 parent_award_type_code
148 parent_award_type
149 parent_award_single_or_multiple_code
150 parent_award_single_or_multiple
151 major_program
152 national_interest_action_code
153 national_interest_action
154 cost_or_pricing_data_code
155 cost_or_pricing_data
156 cost_accounting_standards_clause_code
157 cost_accounting_standards_clause
158 government_furnished_property_code
159 government_furnished_property
160 sea_transportation_code
161 sea_transportation
162 undefinitized_action_code
163 undefinitized_action
164 consolidated_contract_code
165 consolidated_contract
166 performance_based_service_acquisition_code
167 performance_based_service_acquisition
168 multi_year_contract_code
169 multi_year_contract
170 contract_financing_code
171 contract_financing

172 purchase_card_as_payment_method_code
173 purchase_card_as_payment_method
174 contingency_humanitarian_or_peacekeeping_operation_code
175 contingency_humanitarian_or_peacekeeping_operation
176 alaskan_native_corporation_owned_firm
177 american_indian_owned_business
178 indian_tribe_federally_recognized
179 native_hawaiian_organization_owned_firm
180 tribally_owned_firm
181 veteran_owned_business
182 service_disabled_veteran_owned_business
183 woman_owned_business
184 women_owned_small_business
185 economically_disadvantaged_women_owned_small_business
186 joint_venture_women_owned_small_business
187 joint_venture_economic_disadvantaged_women_owned_small_bus
188 minority_owned_business
189 subcontinent_asian_asian_indian_american_owned_business
190 asian_pacific_american_owned_business
191 black_american_owned_business
192 hispanic_american_owned_business
193 native_american_owned_business
194 other_minority_owned_business
195 contracting_officers_determination_of_business_size
196 contracting_officers_determination_of_business_size_code
197 emerging_small_business
198 community_developed_corporation_owned_firm
199 labor_surplus_area_firm
200 us_federal_government
201 federally_funded_research_and_development_corp
202 federal_agency
203 us_state_government
204 us_local_government
205 city_local_government
206 county_local_government
207 inter_municipal_local_government
208 local_government_owned
209 municipality_local_government
210 school_district_local_government
211 township_local_government
212 us_tribal_government
213 foreign_government
214 organizational_type
215 corporate_entity_not_tax_exempt

216 corporate_entity_tax_exempt
217 partnership_or_limited_liability_partnership
218 sole_proprietorship
219 small_agricultural_cooperative
220 international_organization
221 us_government_entity
222 community_development_corporation
223 domestic_shelter
224 educational_institution
225 foundation
226 hospital_flag
227 manufacturer_of_goods
228 veterinary_hospital
229 hispanic_servicing_institution
230 receives_contracts
231 receives_financial_assistance
232 receives_contracts_and_financial_assistance
233 airport_authority
234 council_of_governments
235 housing_authorities_public_tribal
236 interstate_entity
237 planning_commission
238 port_authority
239 transit_authority
240 subchapter_scorporation
241 limited_liability_corporation
242 foreign_owned
243 for_profit_organization
244 nonprofit_organization
245 other_not_for_profit_organization
246 the_ability_one_program
247 private_university_or_college
248 state_controlled_institution_of_higher_learning
249 X1862_land_grant_college
250 X1890_land_grant_college
251 X1994_land_grant_college
252 minority_institution
253 historically_black_college
254 tribal_college
255 alaskan_native_servicing_institution
256 native_hawaiian_servicing_institution
257 school_of_forestry
258 veterinary_college
259 dot_certified_disadvantage

260 self_certified_small_disadvantaged_business
261 small_disadvantaged_business
262 c8a_program_participant
263 historically_underutilized_business_zone_hubzone_firm
264 sba_certified_8a_joint_venture
265 highly_compensated_officer_1_name
266 highly_compensated_officer_1_amount
267 highly_compensated_officer_2_name
268 highly_compensated_officer_2_amount
269 highly_compensated_officer_3_name
270 highly_compensated_officer_3_amount
271 highly_compensated_officer_4_name
272 highly_compensated_officer_4_amount
273 highly_compensated_officer_5_name
274 highly_compensated_officer_5_amount
275 usaspending_permalink
276 last_modified_date

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C. LIST OF DATA SET 39 FEATURES RETAINED

Columns	Name
10	federal_action_obligation
17	action_date_fiscal_year
27	awarding_office_code
82	inherently_governmental_functions
90	naics_code
108	extent_competed_code
110	solicitation_procedures_code
112	type_of_set_aside_code
118	fair_opportunity_limited_sources_code
122	number_of_offers_received
176	alaskan_native_corporation_owned_firm
181	veteran_owned_business
182	service_disabled_veteran_owned_business
183	woman_owned_business
184	women_owned_small_business
185	economically_disadvantaged_women_owned_small_business
188	minority_owned_business
189	subcontinent_asian_asian_indian_american_owned_business
190	asian_pacific_american_owned_business
191	black_american_owned_business
192	hispanic_american_owned_business
193	native_american_owned_business
194	other_minority_owned_business
196	contracting_officers_determination_of_business_size_code
215	corporate_entity_not_tax_exempt
216	corporate_entity_tax_exempt
217	partnership_or_limited_liability_partnership
227	manufacturer_of_goods
230	receives_contracts
232	receives_contracts_and_financial_assistance
240	subchapter_scorporation
241	limited_liability_corporation
242	foreign_owned
243	for_profit_organization
259	dot_certified_disadvantage
260	self_certified_small_disadvantaged_business
261	small_disadvantaged_business
262	c8a_program_participant
263	historically_underutilized_business_zone_hubzone_firm

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. LIST OF DATA SET 33 BINARY FEATURES

Columns	Names
34	awarding_office_codeW91ZLK
47	extent_competed_codeTRUE
50	type_of_set_aside_codeTRUE
51	fair_opportunity_limited_sources_codeTRUE
58	alaskan_native_corporation_owned_firmt
59	veteran_owned_businesst
60	service_disabled_veteran_owned_businesst
61	woman_owned_businesst
62	women_owned_small_businesst
63	economically_disadvantaged_women_owned_small_businesst
64	minority_owned_businesst
65	subcontinent_asian_asian_indian_american_owned_businesst
66	asian_pacific_american_owned_businesst
67	black_american_owned_businesst
68	hispanic_american_owned_businesst
69	native_american_owned_businesst
70	other_minority_owned_businesst
71	contracting_officers_determination_of_business_size_codeS
72	corporate_entity_not_tax_exemptt
73	corporate_entity_tax_exemptt
74	partnership_or_limited_liability_partnershipt
75	manufacturer_of_goodst
76	receives_contractst
77	receives_contracts_and_financial_assistancet
78	subchapter_scorporationt
79	limited_liability_corporationt
80	foreign_ownedt
81	for_profit_organizationt
82	dot_certified_disadvantaget
83	self_certified_small_disadvantaged_businesst
84	small_disadvantaged_businesst
85	c8a_program_participantt
86	historically_underutilized_business_zone_hubzone_firmt

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Aggarwal C, Yu P (2008) Outlier detection with uncertain data. Cite Seer X Pennsylvania State University, PA. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.297.7372&rep=rep1&type=pdf.

Allaire J, Chollet F (2020) Keras: R interface to 'Keras'. R package, version 2.3.0.0. https://CRAN.R-project.org/package=keras.

Bowman B (2019) Anomaly detection using a variational autoencoder neural network with a novel objective function and gaussian mixture model selection technique. Master's thesis, Graduate School of Operational and Information Sciences, Naval Postgraduate School, CA. http://hdl.handle.net/10945/62853.

Campos G, Zimek A, Sander J, Campello R, Micenkova B, Schubert E, Assent I, Houle M (2016) On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Min. Knowl. Disc.* 30, 891–927 https://doi.org/10.1007/s10618-015-0444-8.

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust: An R package for determining the relevant number of clusters in a data set. *J.l of Stat. Software* 61(6), 1–36. http://www.jstatsoft.org/v61/i06/.

Dertat A (2017) Applied deep learning—Part 3: Autoencoders. Towards data science. Accessed November 1, 2020, https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798.

Chandola, V, Banerjee A, Kumar V (2009*)* Anomaly detection: A survey. *ACM Comp. Surv.* 41(3), https://doi.org/10.1145/1541880.1541882.

David C (2020) Isotree: Isolation-based outlier detection. R package, version 0.1.18. https://CRAN.R-project.org/package=isotree.

DCAA (2020) Frequently asked questions. Defense Contract Audit Agency, Accessed November 4, 2020, https://www.dcaa.mil/Guidance/FAQs/.

DCMA (2020) About the agency. Defense Management Contract Agency, Accessed November 4, 2020, https://www.dcma.mil/About-Us/.

DeepAI (2020) Classifier. Deep AI, Accessed November 16, 2020, https://deepai.org/machine-learning-glossary-and-terms/classifier.

Dey D (2019) Dbscan clustering in ML: Density based clustering. Geeks for Geeks. Accessed November 9, 2020, https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/.

Emmott A, Das S, Dietterich T, Fern A, Wong W (2013 August) Systematic construction of anomaly detection benchmarks from read data. *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Descriptions* (Corvallis, OR), https://dl.acm.org/doi/10.1145/2500853.2500858.

Faraway J (2016) *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (2nd ed.). Boca Raton FL: CRC Press.

Fritsch A (2012) Mcclust: Process an MCMC sample of clusterings. R package, version 1.0. https://CRAN.R-project.org/package=mcclust.

Hahsler M, Piekenbrock M, Doran D (2019) Dbscan: fast density-based clustering with R. *Journal of Statistical Software* 91(1), 1–30. https://doi.org/10.18637/jss.v091.i01.

Hennig C (2020) Fpc: Flexible procedures for clustering. R package, version 2.2-8. https://CRAN.R-project.org/package=fpc.

Hassanat A, Abbadi M, Altarawneh G, Alhasanat A (2014) Solving the problem of the k parameter in the KNN classifier using an ensemble learning approach. *International Journal of Computer Science and Information Security* 12(8), 33-39. https://ia800700.us.archive.org/25/items/JournalOfComputerScienceIjcsisAugust2014/JournalOfComputerScienceIjcsisVol.12No.8August2014.pdf.

Kibish S (2018) A note about finding anomalies. Towards Data Science. Accessed November 1, 2020, https://towardsdatascience.com/a-note-about-finding-anomalies-f9cedee38f0b.

Koufakou A, Georgiopoulos M (2010) A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. *Data Min. Knowl. Disc.* 20, 256–289, https://doi.org/10.1007/s10618-009-0148-z.

Lee N (2019) Utilization of machine learning techniques to detect anomalies in Department of Defense contract data. Master's thesis, Graduate School of Operational and Information Sciences, Naval Postgraduate School, CA. http://hdl.handle.net/10945/62268.

Liu F, Ting K, Zhou Z (2020) Isolation forest. H_2O.ai. Accessed December 6, 2020, https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/if.html.

Luebke, M (2020) "Normal" data vs. an anomaly: Detecting the difference. Accessed October 29, 2020, https://sciencelogic.com/blog/what-is-anomaly-detection.

McGregor M (2020) 8 clustering algorithms in machine learning that all data scientist should know. Free Code Camp. Accessed November 9, 2020, https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/.

Perera S (2015) Introduction to anomaly detection: Concepts and techniques. My Views of the World and Systems. Accessed November 1, 2020, https://iwringer.wordpress.com/2015/11/17/anomaly-detection-concepts-and-techniques/.

R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

SBIR (2020) The roles of DCMA and DCAA with Department of Defense awards. Small Business Innovation Research. Accessed November 4, 2020, https://www.sbir.gov/tutorials/accounting-finance/tutorial-5.

Steinbuss G, Bohm K (2020) Benchmarking unsupervised outlier detection with realistic synthetic data. Karlsruhe Institute of Technology, Germany. https://arxiv.org/pdf/2004.06947.pdf.

Tan P, Steinbach M, Karpatne A, Kumar V (2020) Introduction to data mining (second edition). Computer Science & Engineering User Home Pages. Accessed November 3, 2020, https://www-users.cs.umn.edu/~kumar001/dmbook/index.php.

USAspending.gov (2020) Award data archive. Accessed October 30, 2020, https://www.usaspending.gov/download_center/award_data_archive.

Wang H, Bah M, Hammad M (2019) Progress in outlier detection techniques: A survey. *IEEE Access*, Piscataway, NJ. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8786096

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California