ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

COMPERATIVE EVALUATION OF UNSUPERVISED FRAUD DETECTION ALGORITHMS WITH FEATURE EXTRACTION AND SCALING IN PURCHASING DOMAIN

M.Sc. THESIS

Yiğit Can TAŞOĞLU

Data Engineering and Business Analytics

Big Data and Business Analytics Programme

AUGUST 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

COMPERATIVE EVALUATION OF UNSUPERVISED FRAUD DETECTION ALGORITHMS WITH FEATURE EXTRACTION AND SCALING IN PURCHASING DOMAIN

M.Sc. THESIS

Yiğit Can TAŞOĞLU
(528211079)

Data Engineering and Business Analytics

Big Data and Business Analytics Programme

Thesis Advisor: Assist. Prof. Dr. Mehmet Ali ERGÜN

AUGUST 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SATIN ALMA ALANINDA ÖZELLİK ÇIKARMA VE ÖLÇEKLEME İLE DENETİMSİZ SAHTEKARLIK TESPİT ALGORİTMALARININ KARŞILAŞTIRMALI DEĞERLENDİRMESİ

YÜKSEK LİSANS TEZİ

Yiğit Can TAŞOĞLU
(528211079)

Veri Mühendisliği ve İş Analitiği

Büyük Veri ve İş Analitiği Programı

Tez Danışmanı: Dr. Öğr. Üyesi Mehmet Ali ERGÜN

AĞUSTOS 2024

Yigit Can TAŞOĞLU, a M.Sc. student of İTU Graduate School student ID 528211079, successfully defended the thesis/dissertation entitled "COMPERATIVE EVALUATION OF UNSUPERVISED FRAUD DETECTION ALGORITHMS WITH FEATURE EXTRACTION IN PURCHASING DOMAIN", which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assist. Prof. Dr. Mehmet Ali ERGÜN** İstanbul Technical University	.....
Jury Members :	**Assoc. Prof. Dr. Ömer Faruk BEYCA** Istanbul Technical University	.....
	Prof. Dr. Selim ZAİM Ibn Haldun University	.....

Date of Submission : 23 May 2024
Date of Defense : 21 August 2024

FOREWORD

As I begin this academic journey, I want to express my gratutude to people who supported me during this process. Their support was really important to me.

Firstly I want to thank my family, my mom, Nuran Tasoglu, and my dad, Sezar Tasoglu. Their belief and love towards me was and incredible support for me. Without their support I wouldnt think that I will be able to have this academic success.

To my friend, Nuri Orhan, thank you for always being supportive to me. Your support was always boosting my life.

I am grateful to my university advisor Doc.Dr. Ömer Faruk Beyca, for his guidence and dedication was really important to me. This support helped me a lot to grow academically.

Lastly, I want to thank my friend and manager, Zeliha Öztürk. She was supporting me all the way long. Thank you for your understanding the balance between the work and study and supporting me along the way all the time.

May 2024

Yiğit Can TAŞOĞLU
(Data Engineering and Business Analytics)

TABLE OF CONTENTS

Page

FOREWORD ..... **v**
TABLE OF CONTENTS ..... **vii**
ABBREVIATIONS ..... **ix**
LIST OF TABLES ..... **xii**
LIST OF FIGURES ..... **xiv**
SUMMARY ..... **xvi**
ÖZET ..... **xviii**
1. INTRODUCTION **1**
1.1 Background and Motivation 1
1.2 Problem Statement 2
2. LITERATURE REVIEW **5**
3. METHODOLOGY **13**
3.1 Overview of Statistical Outlier Detection Techniques 13
3.1.1 Z-score method 13
3.1.2 Modified z-score method 13
3.1.3 Interquartile range / Tukey's Fence method 14
3.1.4 Standardized residuals 14
3.1.5 Cook's distance 14
3.2 Distance Based Methods 15
3.2.1 Mahalanobis distance 15
3.2.2 Local outlier factor (LOF) 15
3.2.3 DBSCAN (density-based spatial clustering of applications with noise) 15
3.3 Machine Learning Based Methods 16
3.3.1 Isolation forest 16
3.3.2 One-class supportv machines (SVM) 16
3.3.3 Autoencoders 17
3.4 Feature Based Methods 17
3.4.1 Histogram-based outlier detection (HBOS) 17
3.4.2 Angle-based outlier detection (ABOD) 17
3.4.3 Principal component analysis (PCA) 18
3.5 Evaluation Metrics for Outlier Detection 18
3.6 Data Collection 19
3.7 Experimental Setup 21
4. CASE STUDY **23**
4.1 Model Comparison 23

4.2 Data Cleaning and Transformations 24
4.2.1 Data cleaning 25
4.2.2 Data transformations 25
4.2.2.1 Data scaling 26
4.2 Feature Extraction 36
5. EVALUATION STRATEGY 43
5.1 Visual Inspection 43
5.2 Business Impact Analysis 43
5.3 Anomaly Confirmation 43
5.4 Comparison with Baseline Methods 43
5.5 Evaluation with Synthetic Data 44
6. RESULTS AND ANALYSIS 45
6.1. Limitations 45
7. CONCLUSION 47
REFERENCES 49
CURRICULUM VITAE 51

ABBREVIATIONS

ABOD : Angle-based Outlier Detection
AE : Autoencoder
AFRL : Air Force Research Laboratory
AUC : Area Under Curve
CBLOF : Cluster-Based Local Outlier Factor
CIFAR : Canadian Institute For Advanced Research
CMGOS : Clustering-based Multivariate Gaussian Outlier Score
COF : Connectivity-Based Outlier Factor
COP : Copula
DAGMM : Deep Autoencoding Gaussian Mixture Model
DARPA : Defense Advanced Research Projects Agency
DBSCAN : Density-based Spatial Clustering of Applications with Noise
DNN : Deep Neural Network
DSEBM : Deep Structured Energy Based Models
EM : Expectation Maximization
ERP : Enterprise Resource Planning
EUR : Euro
FN : False Negative
FP : False Positive
FSA : Finite State Automata
GBP : British Pound Sterling
GMM : Gaussian Mixture Model
GMUS : GMM-based Undersampling
GUMM : Gaussian Uniform Mixture Model
HBOS : Histogram-based Outlier Score
HMM : Hidden Markov Models
INFLO : Influenced Outlierness
IT : Information Technology
NSL : Network Security Laboratory
KDDCUP : Knowledge Discovery in Databases Dataset
KNN : K-Nearest Neighbour
LDCOF : Local Density Cluster-based Outlier Factor
LOCI : Local Correlation Integral
LOF : Local Outlier Factor
MIT : Massachusetts Institute of Technology
MNIST : Modified National Institute of Standards and Technology
NN : Neural Network
NVI : Neural Variational Inference
OC : One Class
OCSVM : One Class Support Vector Machine
ODIN : Out-of-Distribution Detector for Neural Networks
PCA : Principle Component Analysis
PDF : Probability Density Function

ABBREVIATIONS (CONTINUED)

PST **:** Probabilistic Suffix Trees
ROC **:** Receiver Operating Characteristic Curve
SAP **:** Systems, Applications and Products in Data Processing
SRM **:** Supplier Relationship Management
SSC **:** Subspace Clustring
SVM **:** Support Vector Machine
TN **:** True Negative
TP **:** True Positive
UCF **:** University of Central Florida
USD **:** United States Dollar
VAE **:** Variational Autoencoder

LIST OF TABLES

Page

Table 4.1: Performance metrics for each detection algorithm. **2424**
Table 4.2: Scaling effect comparison for Local Outlier Factor algorithm**. 3030**
Table 4.3: Scaling effect comparison with another dataset for Local Outlier Factor algorithm. **..... 344**
Table 4.4: Feature Extraction effect comparison for Local Outlier Factor algorithm. **..... 399**

LIST OF FIGURES

Page

Figure 4.1: 3D Scatter Plot for Comparison Data (Normal and Outlier Data Points) **23**
Figure 4.2: 3D Scatter Plot for Local Outlier Factor Algorithm Predictions **24**
Figure 4.3: Gauss Distribution for Feature 1(Normal and Outlier Distributions)... **277**
Figure 4.4: Gauss Distribution for Feature 2(Normal and Outlier Distributions) ..... Error! Bookmark not defined.**8**
Figure 4.5: Gauss Distribution for Feature 3(Normal and Outlier Distributions)... **288**
Figure 4.6: Gauss distributions of Features are really different due to magnitude difference. **288**
Figure 4.7: 3D representation of Dataset. **299**
Figure 4.8: Graph represents the predictions of Local Outlier Factor without scaling.. ..... **299**
Figure 4.9: Graph represents the predictions of Local Outlier Factor with scaling... **30**
Figure 4.10: ROC Curves without scaling. **31**
Figure 4.11: ROC Curves with scaling. **31**
Figure 4.12: Confusion Matrix without scaling. **32**
Figure 4.13: Confusion Matrix with scaling. **32**
Figure 4.14: 3D representation of 2nd Dataset. **33**
Figure 4.15: Predictions of Local Outlier Factor without scaling. **33**
Figure 4.16: Predictions of Local Outlier Factor with scaling. **34**
Figure 4.17: ROC Curves with scaling for 2nd Dataset. **345**
Figure 4.18: ROC Curves with scaling for 2nd Dataset. **355**
Figure 4.19: Confusion Matrix without scaling for 2nd Dataset. **366**
Figure 4.20: Confusion Matrix with scaling for 2nd Dataset **366**
Figure 4.21: 3D Visualisation of Predictions for Existing Dataset without feature extraction. **38**
Figure 4.22: 3D Visualisation of Predictions for Existing Dataset with feature extraction. **39**
Figure 4.23: ROC Curves without extraction. **40**
Figure 4.24: ROC Curves with extraction. **40**
Figure 4.25: Confusion Matrix without feature extraction **41**
Figure 4.26: Confusion Matrix with feature extraction. **41**

COMPERATIVE EVALUATION OF UNSUPERVISED FRAUD DETECTION ALGORITHMS WITH FEATURE EXTRACTION AND SCALING IN PURCHASING DOMAIN

SUMMARY

The main aim of the research is to evaluate and compare various unsupervised outlier detection methods that do not require labeled data, making them suitable for real-world purchasing data where labels are often unavailable. The thesis highlights the challenges of fraud detection in large datasets, particularly in industries like finance and purchasing, where fraudulent activities can cause significant financial losses if not identified early.

The motivation behind the research lies in the limitations of traditional, rule-based detection methods, which often fail to capture complex fraud patterns. Unsupervised algorithms, which can detect anomalies based on their deviation from the general behavior of the dataset, offer a proactive approach to fraud detection by identifying unseen fraud concepts.

This study applies various methods, including distance-based, machine learning-based, and feature-based models, and focuses on enhancing these models through feature extraction and scaling techniques. The thesis evaluates several algorithms, such as Local Outlier Factor (LOF), DBSCAN, and Isolation Forest, using performance metrics like accuracy, precision, recall, and F1 score. LOF was identified as the most effective model, achieving the highest accuracy and demonstrating a robust ability to detect irregular patterns in the purchasing data. However, the effectiveness of all algorithms was significantly enhanced by data transformations, particularly scaling. Scaling ensures that features with differing magnitudes, such as quantities and prices, do not distort the results, allowing for more accurate anomaly detection. The importance of feature extraction is also emphasized, as it helps identify intricate patterns between data points. Extracted features, such as the frequency of purchase orders, vendor categories, and purchase amounts, provide deeper insights into potential fraud indicators. Additionally, the study recognizes that the integration of multiple models can reduce the limitations inherent in individual algorithms, thus creating a more comprehensive fraud detection framework. By combining different unsupervised methods and leveraging feature extraction, the research offers a more adaptive and reliable approach to identifying fraudulent activities. In conclusion, this study proves that employing a combination of unsupervised outlier detection methods, along with appropriate data preprocessing techniques, significantly improves fraud detection in purchasing systems.

These methods not only enhance accuracy but also help businesses reduce financial risks and improve operational efficiency, ensuring a more secure and effective fraud prevention strategy.

SATIN ALMA ALANINDA ÖZELLİK ÇIKARMA VE ÖLÇEKLEME İLE DENETİMSİZ SAHTEKARLIK TESPİT ALGORİTMALARININ KARŞILAŞTIRMALI DEĞERLENDİRMESİ

ÖZET

Çalışmanın temel amacı, denetimsiz aykırı değer tespit yöntemlerini karşılaştırarak, etiketli veriye ihtiyaç duymayan bu algoritmaların gerçek dünya satın alma verileri üzerindeki etkinliğini değerlendirmektir. Etiketlenmiş verinin genellikle mevcut olmadığı durumlarda bu tür yöntemler oldukça avantajlıdır. Tez, finans ve satın alma sektörleri gibi büyük veri setlerine sahip alanlarda dolandırıcılık tespitinin zorluklarını vurgulamaktadır. Dolandırıcılık erken tespit edilmediğinde, bu sektörlerde ciddi mali kayıplara yol açabilmektedir.

Geleneksel, kurallara dayalı tespit yöntemleri genellikle karmaşık dolandırıcılık desenlerini yakalamakta yetersiz kalmaktadır. Bu yöntemler, dolandırıcılığın sürekli değişen doğası karşısında etkinliğini kaybetmekte ve yeni dolandırıcılık tekniklerini tespit etmekte başarısız olabilmektedir. Dolandırıcılar, kurallara dayalı sistemlerdeki boşlukları tespit edip bunlardan yararlanabilir. Bu noktada denetimsiz algoritmalar devreye girer. Denetimsiz öğrenme algoritmaları, veri setindeki genel davranıştan sapmaları tespit ederek, daha önce görülmemiş dolandırıcılık yöntemlerini ortaya çıkarabilmektedir.

Bu çalışma, mesafe tabanlı, makine öğrenimi tabanlı ve özellik tabanlı modeller gibi çeşitli yöntemleri uygulamakta ve bu modelleri özellik çıkarma ve ölçekleme gibi veri işleme teknikleriyle iyileştirmeyi hedeflemektedir. Local Outlier Factor (LOF), DBSCAN ve Isolation Forest gibi yaygın olarak kullanılan yöntemler bulunmaktadır. Bu algoritmalar, doğruluk, hassasiyet, geri çağırma ve F1 skoru gibi performans metrikleri kullanılarak karşılaştırılmıştır. Yapılan analizlerde, LOF algoritması en yüksek doğruluğa ulaşarak, satın alma verilerinde riskli talepleri tespit etme açısından en etkili model olarak öne çıkmıştır. LOF algoritması, komşuluk ilişkilerini inceleyerek, her bir veri noktasının lokal yoğunluğunu hesaplar ve yoğunluktan sapmaları tespit eder. Bu sayede, dolandırıcılık faaliyetleri gibi aykırı değerlerin tespitinde başarılı olur. DBSCAN, yoğunluk tabanlı bir kümeleme algoritmasıdır ve küme dışındaki noktaları aykırı değer olarak tanımlar. Bu algoritma, özellikle kümelenemeyen ve izole kalan noktaları dolandırıcılık açısından incelemek için uygundur. Ancak, DBSCAN'ın performansı, parametrelerin doğru ayarlanmasına bağlıdır. Isolation Forest ise, veri noktalarını bölerek aykırı değerleri tespit eder. Aykırı değerler genellikle daha az bölünme gerektirdiği için, bu model dolandırıcılık tespitinde etkili olabilir. Ancak, diğer algoritmalara göre bazı karmaşık dolandırıcılık desenlerini yakalamakta yetersiz kalabilir.

Özellikle veri dönüşümleri ve ölçekleme teknikleri ile model verimi önemli ölçüde artırılmıştır. Ölçekleme, veri setindeki sayısal özelliklerin farklı büyüklükteki değerlerinin analiz sonuçlarını bozmasını engellemek için kullanılan önemli bir tekniktir. Özellikle, satın alma verilerinde miktar ve fiyat gibi değişkenlerin farklı büyüklüklerde olması, ölçekleme yapılmadığında bazı özelliklerin diğerlerine göre daha fazla etkili olmasına neden olabilir. Yapılan deneylerde, ölçekleme uygulanmadığında LOF algoritmasının doğruluğu %88 iken, ölçekleme sonrası bu oran %98'e yükselmiştir. Ölçekleme ayrıca algoritmaların hesaplama süresini ve doğruluğunu artırarak, verilerin daha hızlı işlenmesini sağlamaktadır. Özellikle, yüksek boyutlu veri setlerinde mesafe tabanlı algoritmaların performansı, ölçekleme ile büyük ölçüde iyileştirilmektedir.

Özellik çıkarma, model performansını artıran kritik bir adımdır. Çıkarılan özellikler, modelin daha karmaşık dolandırıcılık kalıplarını tespit etmesini sağlamaktadır. Örneğin, bir kullanıcının belirli bir tedarikçiye yönelik alışılmadık derecede sık taleplerde bulunması, potansiyel bir dolandırıcılık göstergesi olarak değerlendirilebilir. Benzer şekilde, tedarikçi kategorileri, satın alma tutarları ve onay süreleri gibi çıkarılan özellikler, dolandırıcılık tespitinde derinlemesine analizler yapılmasına olanak tanır.

Tezde yapılan deneylerde, özellik çıkarma işlemi uygulanmamış bir veri seti ile çıkarılan özelliklerin kullanıldığı bir veri seti karşılaştırılmıştır. Sonuçlar, çıkarılan özelliklerin model performansını önemli ölçüde artırdığını göstermiştir. Özellik çıkarma uygulanmayan veri setinde LOF algoritmasının doğruluğu %36 iken, çıkarılan özelliklerle bu oran %54'e yükselmiştir. Bu, özellikle veri setindeki önemli ilişki ve kalıpların yakalanmasına olanak sağlamış ve dolandırıcılık tespitini daha güvenilir hale getirmiştir. Tezde yapılan vaka çalışmasında, 934 satın alma işlemi incelenmiş ve denetimsiz aykırı değer tespit algoritmaları kullanılarak her bir işlem için risk puanı hesaplanmıştır. Risk puanlarına göre en riskli %10'luk dilimde yer alan işlemler, satın alma uzmanları tarafından manuel olarak incelenmiştir. Bu incelemeler sonucunda, seçilen 95 işlemden 58'inin dolandırıcılık faaliyetleri içerdiği tespit edilmiştir. Bu bulgu, modelin yaklaşık %61 oranında bir tespit doğruluğuna sahip olduğunu göstermektedir. Bu oran, iş süreçlerinde önemli bir iyileşme sağlamaktadır; çünkü manuel incelemelerde genellikle işlemlerin yalnızca %1 ila %3'ü dolandırıcılık faaliyetleri içerir.

Tezde denetimsiz dolandırıcılık tespit algoritmalarının etkinliğini değerlendirmek için çeşitli stratejiler kullanılmıştır. Görsel inceleme, model performansını değerlendirmek için kullanılan önemli bir yöntemdir. Farklı boyutlardaki veriler grafiksel olarak incelenerek, dolandırıcılık kalıpları görsel olarak tespit edilmeye çalışılmıştır. Ayrıca, ticari etki analizi ve anomali doğrulaması gibi yöntemlerle modellerin iş süreçlerine olan etkisi değerlendirilmiştir. Bununla birlikte, çalışmanın bazı sınırlamaları bulunmaktadır. Çalışmada incelenen veri seti yalnızca belirli bir sektör ve coğrafi bölge ile sınırlıdır. Bu nedenle, geliştirilen modellerin genelleştirilebilirliği konusunda sınırlamalar olabilir. Ayrıca, algoritmaların parametre optimizasyonuna yönelik daha

kapsamlı bir çalışma yapılmamıştır. Dolayısıyla, farklı veri setleri üzerinde yapılan incelemelerde parametrelerin ayarlanması gerekebilir.

Sonuç olarak, denetimsiz aykırı değer tespit algoritmalarının, özellikle özellik çıkarma ve ölçekleme gibi ön işleme teknikleriyle desteklendiğinde, satın alma sistemlerinde dolandırıcılığı tespit etmede etkili bir şekilde kullanılabileceğini kanıtlamaktadır. Bu yaklaşımlar, şirketlerin dolandırıcılık faaliyetlerini daha hızlı ve etkili bir şekilde tespit etmelerine olanak tanımakta, böylece finansal kayıpları ve operasyonel aksaklıkları en aza indirmektedir. Tezin bulguları, denetimsiz öğrenme algoritmalarının iş süreçlerinde önemli bir katma değer sağladığını ve bu yöntemlerin gelecekte daha da yaygın bir şekilde kullanılabileceğini göstermektedir.

1. INTRODUCTION

1.1 Background and Motivation

Fraud detection, outlier detection or anomaly detection is one of the most important areas of machine learning and data mining. Basically, these methods try to identify observations which are not having the same behavior as general population within a dataset. Depending on the application target, outliers can be just detected to clean the dataset, or it can also be the main scope. Outliers can tell many things about the dataset itself. Generally, outliers are rare events or errors that can occur in datasets. Outlier evaluation plays an important role for model performance and stability. In addition to that, outlier detection is also such a crucial operation in many industries like finance, cybersecurity, quality control, and healthcare. Many industries use outlier detection models to be able to prevent fraud, time loss, reputation loss, money, and life.

There are many different approaches to detect outliers. In the real world, most of the use cases have no label to group observations within a dataset. Also, data labeling is most of the time costly, requires manual effort and can take time. Unsupervised outlier detection methods do not need any label to identify outliers and work without any prior domain or product knowledge. Main purpose of unsupervised outlier detection algorithms is to identify observations which have different behavior than the rest of the dataset. Many outlier detection models are focused on each observation one by one. However, outlier or fraud cases can occur with repetitive instances as well.

Motivation of the research is to compare existing unsupervised outlier detection models and analyze the performance changes by extracting features from the existing data fields. By applying feature extraction techniques, the target is to identify recurring fraud cases. Many outlier detection algorithms are not performing well to detect recurring instances by just evaluating each single instance one by one.

Finance and purchasing applications require accurate outlier detection models. Many companies and customers lose high amounts of money due to fraudsters. Minimizing losses and protecting customers and reputation is critical for financial institutions. It is also important to identify outliers and frauds in the early stage of proceses. These techniques also play a crucial role in manufacturing to detect defective products in the early production process to ensure high quality and low-cost standards. Health care applications also outlier detection models to identify unusual patterns for patients.

Outlier detection models have been so popular since decades due to their ability to prevent losses. Thanks to this popularity there are a wide range of techniques and algorithms. Distance based, density based, clustering based, and statistical approaches

are the most common outlier detection models and techniques. All the outlier detection models perform differently on different datasets. These models also have different limitations and strengths. Target of the research is to combine different outlier detection models into one superior model by applying all models at the same time to the same dataset. In addition to that research aims to increase the performance of outlier detection models by extracting features from existing dataset to detect recurring anomalous instances.

In summary, the main objective of the research is to combine existing outlier detection models into one superior model to have a robust outlier detection model. By combining different outlier detection models, the target is to eliminate limitations from each model. Additionally, feature extraction techniques will be applied to extract all the possible information from existing Purchasing data which includes recurrent purchasing orders. By extracting instance connection information, the model can perform better for recurrent fraud instances.

1.2 Problem Statement

Purchasing is one of the cores and most important activities for companies. Also having control over purchasing is playing a crucial role especially for big companies to stay profitable. When purchasing volumes get higher and higher, companies might start to lose control for every small purchase. There are many rules, controls, and standards however, fraudsters can find many ways to steal money from companies. With the increasing purchasing volumes, it is almost impossible to check each purchasing request one by one. Many companies developed measures to avoid losses. Many of the measures are conventional methods and sometimes these measures are not effective to detect fraudulent activities. Effectiveness of these measures for detecting fraudulent activities is often limited. To solve this problem, many companies are applying statistical methods in order to automate the process control and ensure system security. However conventional methods often fail to detect the complex nature of patterns. It can cause a high rate of false positives and missed fraud cases.

To address false positives and missed fraud cases, unsupervised fraud detection techniques can be applied. This helps companies to detect complex patterns of fraudsters. Also unsupervised outlier detection algorithms can adapt to changes pretty quickly. Due to the nature of fraud, fraudsters always seek a new way to steal money from companies. Especially for new and unseen observations, unsupervised outlier detection models are a good solution. Unsupervised outlier detection algorithms can analyze unusual patterns, customer and employee behavıour, outliers, or anomalies that differ significantly from expected behavior. These models can be much more advanced alternatives to the conventional methods.

Using unsupervised outlier detection models has many advantages compared to conventional methods. First, these models offer a more general approach to fraud detection instead of specifically determining the hard coded rules. Since there are no

hard coded rules, these models are giving opportunity to have a proactive approach. This enables companies to be one step ahead of the fraudsters. Every new created fraud concept might be detected via unsupervised outlier detection algorithms. Since the new concept will have different behavior from the existing ones, these algorithms come very handy. Another big advantage compared to conventional methods is, conventional methods require manual effort and tıme. Purchasing experts must analyze the fraud patterns for each fraud concept. Rules must be specialized for each attempt. Manual investigation and analysis can also be wrong due to human factors. Unsupervised anomaly detection algorithms can increase operational efficiency and also companies can focus on analyzing and investigating confirmed fraud cases.

Companies which integrate unsupervised outlier detection models into the statistical and conventional methods increase fraud detection capabilities. It also increases the effectiveness of existing fraud detection systems.
Especially by using statistical methods and unsupervised outlier detection systems together, companies create a more robust and accurate approach to identify and prevent fraudulent activities. These models help to create a safer environment and prevent financial loss for the companies.

2. LITERATURE REVIEW

Outlier detection is crucial especially for big organizations. In the last decades many studies have been completed in the field of outlier detection. This section provides an overview for some of the important studies related with outlier detection in the purchasing field.

Study 1: A Comparison of Outlier Detection Techniques for High-Dimensional Data by
Xiaodan Xu, Huawen Liu, Li Li, Minghai Yao (2018)
This study focused on overview of the outlier detection methods for multivariate data. It specifically focuses on complex datasets with high number of features. This study evaluates the outlier detection algortihms mainly in 3 groups. It divides algorithms such as ensemble-based methods, subspace methods and neighbor methods. Ensemble methods used are out of distribution detection algorithm, copula-based outlier detection algorithm and Gaussian Uniform Mixture algorithm. Subspace method algorithms which is evaluated are Subsapce Outlier detection and high contrast sub space algorithms. Last group as neighbor methods and algorithms used to evaluate are kNN, weighted distance kNN, local outlier factor algorithm, fast angle-based outlier detection algorithm and Local Outlier Probability algorithm. These algorithms studied with some of the most popular public datasets.

Result of the comparison is neighbor based algorithms has a better performance result in general. Gaussian based algorithms are performing lower than expectations especially for the complex high dimensional data. Since Gaussian models can be easily affected by the data distribution, this study also shows the weak side of Gaussian models. Since complex real-world datasets are used, and their distributions are difficult to presume, gaussian methods can not perform well.

Study 2: A Geometric Framework for Unsupervised Anomaly Detection by Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, Sal Stolfo
This paper is focusing on representing each observation of the population on a feature space. Feature space mapping is up to selected algortifm with in the geometric framework for unsupervised anomaly detection. Outliers are mapped due to transformations to distant and sparse positons in the feature space.
Also, three different algorithms for detection outliers have been compared. These 3 algorithms are Cluster-based, k-nearest neighbor-based and the last algorithm is a Support Vector Machine based approach.

Study 3: A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data by Markus Goldstein, Seiichi Uchida (2016)
This study offers a evaluation results with 19 different unsupervised anomaly detection algorithm. These algorithms are evaluated with 10 different datasets which is coming from different application fields. Results are grouped and evaluated in 3 anomaly detection method groups which are nearest neighbor based, clustering based and

remaining from these 2 groups. As a result, this paper suggests two main things. First, when there are global anomalies, local anomaly detection algorithms are not performing well due to their nature. Second, global anomaly detection algorithms are also not performing well when there are local anomalies. As an overall results, nearest neighbor, and local outlier algoritms are recommended to use in first place.

Study 4: DEEP AUTOENCODING GAUSSIAN MIXTURE MODEL FOR UNSUPERVISED ANOMALY DETECTION by Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, Haifeng Chen (2018)

Study focuses on the problem of the density estimation in the original feature space during anomaly detection. Especially when the dimension of the dataset increase it is harder to implement and find outliers with density estimation since any instance can be a rare event with a low probability. Focus of the study is to overcome this issue by first implementing the dimensionality reduction and later density estimation applied in low dimensional space. But the problem with the dimensionality reduction is during the reduction, key indicators for anomalies can be removed. To overcome this problem this study suggest to use Deep autoencoding gaussian mixture model to save all related indicators in low dimensional space. Second Gaussian mixture model learned how to apply densitiy estiomation tasks for low dimensional input data for complex structures. In this study 12 different outlier detection methods compared over 4 sample datasets.

Evaluation shows on many public datasets that deep autoencoder gaussian mixture models performing much better than state-of-the-art anomaly detection techniques. It performs up to 14% improvement based on the standard F1 score.

Study 5: A Hybrid Unsupervised Clustering-Based Anomaly Detection Method by Guo Pu, Lijuan Wang, Jun Shen and Fang Dong (2021)

This study shows another technique to combine conventional clustering techniques with SVM supervised models. Basically, first it targets the problem of robustness for convetional clustering techniques. Generally, result of these algorithms is really connected with parameter initialization and parameters of the models. To overcome this issue paper suggests a combination of SVM with conventional clustering algorithms. First paper suggests to partion the data in to many and various low dimensional sub spaces. This can help to improve within cluster similarity. Later SVM, is implemented for all subspaces to understand the data characteristics to identify outliers. A dissimilarity or similarity vector represents distances between outliers and inliers from each created subspace. Later anomalies detected based on a threshold of the vector. This increases the robustness and accuracy of detection via using many subspaces. Evaluation of the idea and technique was done by comparing three other clustering algorithms for unsupervised detection. Algorithms selected as base line are SSC, DBSCAN, K-means algorithms.

The experimental shows that SSC-OCSVM algorithm is performing better than baseline algorithms.

Study 6: Unsupervised Clustering Approach for Network Anomaly Detection by Iwan Syarif, Adam Prugel-Bennett, Gary Wills

This paper is comparing and evaluation study. Study compared anomaly detection algorithms over misuse detection algorithms with DARPA/Lincoln dataset. During evaluation distance-based outlier detection algorithm, k means standard algorithm with its variations used such as improved k-Means, k-Medoids and EM clustering.
To compate anomaly detection algorithms, also misuse detection algorithms are also evaluated such as naïve Bayes, rule induction, decision tree and nearest neighbor. Study showed that misuse detection algorithms are only performing well for know anomaly cases and fails to detect with many unseen outliers.

Study 7: Deep Unsupervised Anomaly Detection by Tangqing Li, Zheng Wang, Siying Liu and Wen-Yan Lin (2021)

Study focuses specifically on anomaly detection in large datasets. This approach is using clustering techniques to identify the normal distribution and later an autoencoder trained on normal data observations. Autoencoder does not use any label, it is just getting the support of clustering algorithms to find normal population behaviour. Study tested on 5 different datasets and proposed method is performing much better than unsupervised techniques.

Study 8: Auto insurance fraud detection using unsupervised spectral ranking for anomaly by Ke Nian, Haofan Zhang, Aditya Tayal, Thomas Coleman, Yuying Li (2016)

This paper suggests a method to detect fraudulent activities via detecting anomalies in the connection with features. Capturing these relationships with kernel similarity, via using spectral analysis of Laplacian of the kernel similarity. This study evaluated the results with both real-world data and syntetic data.

Study 9: Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection by Jiong Zhang and Mohammad Zulkernine

This research employs the Random Forest algorithm, a high performant technique, within anomaly-based Network Intrusion Detection Systems (NIDSs).
Study compares results agaınst SVM, KNN. First method learns patterns from the network and and labeled by services. Later frameworks assumes each services has its own pattern for normal activities. System is really dependend on number of outliers. However unsupervised method performs better.

Study 10: Insurance Fraud Detection with Unsupervised Deep Learning by Chamal Gomes, Zhuo Jin, Hailiang Yang (2021)

Thıs study evaluates the two deep learning models, variational autoencoders and autoenoders. It proposes a new technique to understand the feature importance. It is more focused to get insights, why and how fraud occurs. Three different datasets used in this paper to evaluate results. This study gives insights to understand and interpret from probability distributions.

Study 11: Anomaly Detection: A Survey by Varun Chandola, Arindam Banerjee, and Vipin Kumar (2007)
This paper provides a well content for many anomaly detection techniques. It is not just explaining techniques but also explains the type of anomalies as well. It covers many different domains and discussing for different fraud detection problems. It has a holistic approach for the literature.

Study 12: A Comparative Evaluation of Anomaly Detection Techniques for Sequence Data by Varun Chandola, Varun Mithal, and Vipin Kumar (2008)
This paper investigates many techniques on various datasets for anomaly detection. During the evaluations, kernel, and window-based techniques and markovian techniques presented. New KNN technique which performs better introduced. For Markovian technique, new finite state automation-based technique proposed which is better than the existing finite state automation technique.

Study 13: Anomaly Detection for Discrete Sequences: A Survey by Varun Chandola, Arindam Banerjee, and Vipin Kumar (2009)
This survey aims to offer a thorough and organized summary of the current research addressing the challenge of identifying anomalies in discrete sequences. Kernel Based, Window Based and Markovian techniques applied to compare the results.

Study 14: Understanding Anomaly Detection Techniques for Symbolic Sequences by Varun Chandola, Varun Mithal, and Vipin Kumar (2009)
This study tries to extend the results of the previous experimental investigations by evaluating the performance of different outlier detection techniques by using diverse types of sequence datasets. The analysis conducted in this paper enables a relative assessment of these techniques, showing their respective strengths and weaknesses. In this survey, kernel, and window-based techniques and markovian techniques shown to evaluate results.

In conclusion, analysis of various studies on unsupervised anomaly detection techniques provides valuable insights into the various fields. These studies highlight the importance of robust outlier detection methods for high-dimensional data and multivariate datasets. The use of geometric frameworks and deep learning approaches, such as autoencoders and Gaussian mixture models, has proven to be effective in detecting anomalies in different domains, including network intrusion detection and insurance fraud detection. The evaluation and comparison of these algorithms have demonstrated their varying performances, emphasizing the need for careful selection based on specific data characteristics and application requirements. Furthermore, the significance of unsupervised clustering-based approaches for anomaly detection in complex systems, such as network traffic and symbolic sequences, has been acknowledged. Overall, this comprehensive survey offers a foundation for researchers and practitioners to choose suitable techniques and methodologies for unsupervised anomaly detection tasks. These related studies show the importance of outlier detection

in various domains. Generally finding files, the most appropriate method should be selected based on specific characteristics of the dataset and based on objective.

Study 15: A robust EM clustering algorithm for Gaussian mixture models by Miin-Shen Yang, Chien-Yo Lai, Chih-Ying Lin (2011)

The target of this paper is clustering-based probability models. Especially, the paper proposes much more robust Expectation and Maximization algorithm for Gaussian Mixture models. In this paper a robust Expectation and Maximization clustering algorithm which will be a good performant as well for different cluster volumes is shown. As a result, a new structure for the Expectation Maximization without initialization was suggested. The proposed well performant Expectation Maximization algorithm for GMM evaluates all instances as initials to find out right initial values. Number of clusters defined based on penalty term. Paper also supports the finding with several datasets.

Study 16: Outlier detection via multiclass deep autoencoding Gaussian mixture model for building chiller diagnosis by Viet Tra, Manar Amayri, Nizar Bouguila (2021)

Paper focuses on two important problems for the Chiller diagnosis namely outlier detection and the insufficient amount of correct labeled data. This paper suggests a multiclass deeo auto encoding gaussian mixture model to distinguish anomalies in normal data. Extensive trial findings within this study have demonstrated a superior performance. These results were compared with existing models.

Study 17: Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection by Peng An, Zhiyuan Wang, Chunjiong Zhang (2022)

The paper suggests employing unsupervised ensemble autoencoders coupled with the Gaussian mixture model to apply in multiple domains. This will help irrespective of the characteristics of each field. Attention based latent representations and reconstructed variables are levared within the ensemble autoencoder. To estimate the population density expectation maximization algorithm used together with GMM. Once estimated population density surpasses the learing threshold, the population instance classified as outlier. Experiments evaluated with three datasets and it confirms that the results of the proposed model competes significantly with the chosen anomaly detection baselines.

Study 18: Using Gaussian Mixture Models to Detect Outliers in Seasonal Univariate Network Traffic by Aarthi Reddy, Meredith Ordway-West, Melissa Lee, Matt Dugan, Joshua Whitney Ronen Kahana, Brad Ford, Johan Muedsam, Austin Henslee, Max Rao (2017)

This article introduces a method for identifying outliers in seasonal, univariate network traffic data using Gaussian Mixture Models. Algorithm is exclusively assessed on time series data sourced from network traffic However, it can be readily adapted for application to various other types of seasonal univariate big datasets.

The results are contrasted with traditional outlier detection methods, which typically assume that all data within a set of single probability density function. In conclusion, this paper underlines the distinct advantage of Gaussian Mixture Models in identifying not just 'peaks' within datasets, but also very low valleys.

Study 19: Outlier Detection Algorithm Based on Gaussian Mixture Model by Wenbo Liu, Delong Cui, Zhiping Peng, Jihai Zhong (2019)

This paper presents a new method for high-dimensional complex datasets to address the challenge of outlier detection. A Gaussian mixture model-based outlier detection algorithm is developed. The proposed method utilizes the global optimization expectation maximization algorithm to fit a Gaussian mixture model to the dataset. Outliers are detected based on the three standard deviation threshold applied to each Gaussian component. This algorithm effectively detects outliers in high-dimensional and complex data, as verified by experimental validation.

Study 20: GMM-based Undersampling and Its Application for Credit Card Fraud Detection by Fengjun Zhang, Guanjun Liu, Zhenchuan Li, Chungang Yan, Changjun Jiang (2019)

This article proposes a novel Gaussian Mixture Undersampling. Initially, a Gaussian Mixture Model (GMM) is employed to fit the major group of samples. Later by examining the probability density function of predicted minority samples using the well-fitted GMM, the maximum PDF value serves as the boundary between two classes. As last step, undersampling is executed on the majority samples near the boundary identified by the cross-edge.

Group 1: Comparative Evaluation of Unsupervised Anomaly Detection Algorithms

Study 1: A Comparison of Outlier Detection Techniques for High-Dimensional Data
Study 3: A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data
Study 14: Understanding Anomaly Detection Techniques for Symbolic Sequences
Study 15: A robust EM clustering algorithm for Gaussian mixture models
Study 18: Using Gaussian Mixture Models to Detect Outliers in Seasonal Univariate Network Traffic
Study 19: Outlier Detection Algorithm Based on Gaussian Mixture Model

Group 2: Geometric Framework and Deep Learning Approaches

Study 2: A Geometric Framework for Unsupervised Anomaly Detection
Study 4: DEEP AUTOENCODING GAUSSIAN MIXTURE MODEL FOR UNSUPERVISED ANOMALY DETECTION
Study 7: Deep Unsupervised Anomaly Detection
Study 10: Insurance Fraud Detection with Unsupervised Deep Learning
Study 16: Outlier detection via multiclass deep autoencoding Gaussian mixture model for building chiller diagnosis

Group 3: Clustering-Based Anomaly Detection
Study 5: A Hybrid Unsupervised Clustering-Based Anomaly Detection Method
Study 6: Unsupervised Clustering Approach for Network Anomaly Detection
Study 12: A Comparative Evaluation of Anomaly Detection Techniques for Sequence Data
Study 13: Anomaly Detection for Discrete Sequences: A Survey
Study 17: Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection.

Group 4: Specific Application-Based Anomaly Detection
Study 8: Auto insurance fraud detection using unsupervised spectral ranking for anomaly.
Study 9: Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection
Study 11: Anomaly Detection: A Survey
Study 20: GMM-based Undersampling and Its Application for Credit Card Fraud Detection

3. METHODOLOGY

The methodology applied in this report focuses on unsupervised outlier detection in the purchasing area. In purchasing, detecting abnormalities plays a crucial role. Since data has no labels to predict, supervised machine learning algorithms can't be used. By using unsupervised learning algorithms, we aim to detect potential abnormalities within purchasing datasets. Our target is to detect unusual purchasing behaviors, irregularities, and potentially fraudulent activities. This methodology section explains step by step all processes including data collection, data cleaning and transformations, feature extraction, model selection, and evaluation.

3.1 Overview of Statistical Outlier Detection Techniques

Statistical outlier detection techniques are one of the first and widely used detection algorithms. These methods are widely used to identify outliers or anomalies within the sample dataset by using mathematical calculations. Target of these algorithms is to identify instances which deviate significantly from the general behavior of the sample dataset.

3.1.1 Z-score method

Z-Score method is one of the fundamental approaches for outlier detection by using standard deviation. It calculates the standard deviation of each observation that deviates from the mean of the dataset. Observations which have more standard deviation distance from the mean are considered as outliers or anomalies. This is one of the most basic statistical outlier detection algorithms and generally works well when sample data follows a Gaussian distribution. It is suitable for datasets which are normally 13istributed. Many of the Purchasing data unfortunately does not follow Gaussian distribution. This algorithm is simple yet effective on limited datasets.

3.1.2 Modified z-score method

Since Z-Score has a downside of not working well on skewed datasets or datasets which do not follow Gaussian distribution, Modifed Z-Score methods address this problem. It calculates Median Absolute Deviation to be able to have more robust calculations on
extreme values. It is more resistant to the extreme values and to use Mean Absolute Deviation enables Modified Z-Score method to be more effective in skewed or heavy tailed distributions. However Modified Z-Score method also has a downside. This method can be so sensitive especially when the sample size is so low. In small datasets, calculation of Mean Absolute Deviation can be not reliable so it can create inaccuracies in outlier detection. Z score algorithms also require a threshold to divide instances as outlier and normal instances. While this easy parameter can be advantageous in some cases, due to the simplicity of the method it has a lack of skill to identify complex patterns. Especially local densities can be wrong estimated.

3.1.3 Interquartile range / Tukey's Fence method

This method calculates the range of upper and lower quartiles of a sample dataset. Observations which fall below lower quartile or upper quartile with defined range identified as outliers. This method assumes that we have a symmetric distribution. When data is skewed or follows different distributions with extreme tails, the Interquartile method can not effectively detect outliers. In addition to that, the Interquartile Range method does not consider the relationship within the neighboring clusters. This limits the ability of detection for complex datasets. While detecting outlier's threshold must be selected. This can be subjective and may not capture the outliers correctly. Due to the simplicity of the algorithm, it can not detect the high dimensional relationships.

3.1.4 Standardized residuals

Standardized Residual is a statistical measure used in regression analysis for outlier detection. Standardized residuals can be calculated by dividing the residual to the estimated standard deviation of the residuals. It gives a measure for each observation how observations deviate from the regression line. This algorithm can work well when there is a regression relationship between the variables. If the relationship is nonlinear then standardized residuals might not be able to detect outliers accurately. In addition to this downside, residual analysis is sensitive for extreme outliers. These observations can extremely affect the calculation of standardized residuals. So, it can be biased towards instances that have a strong influence on the regression model. Additionally, the standard residual method is focusing the distance from the regression line. This can be a problem for detecting the outliers which have high dimensional anomalies. It does not consider the relationship between the neighboring clusters or observations. This can limit the effectiveness of outlier detection especially on complex datasets.

3.1.5 Cook's distance

Cook's Distance is one similar method to Standardized Residuals. It uses regression analysis to identify outliers. The difference from Standardized Residuals is, it calculates the impact of each observation on the regression model by calculating the change in the parameters when instance removed. Cook's distance calculates the distance between predicted values with and without the instance and normalizes with standard residual error. It also has the same downsides as Standardized Residuals. This method also relies on a regression model which is used. When there is no linear relationship between the feauteres then Cook's distance might not accurately detect outliers.

3.2 Distance Based Methods

3.2.1 Mahalanobis distance

Mahalanobis distance is using a covariance matrix to calculate the distance for each observation from the mean of the sample dataset. Mahalanobis distance takes correlations between the variables into account by using covariance matrix. It gives a distance measure of each observation from the multivariate distribution of the sample dataset. Considering all variables while calculating the distance is one of the advantages compared to other statistical outlier detection algorithms. However, when the number of variables increases, calculation of the covariance matrix becomes more complex and the accuracy of Mahalanobis distances can be inaccurate. Mahalanobis distance method can be less reliable especially in high dimensional sample datasets. Since it checks multivariate effects, Mahalanobis distance works well when sample datasets have a multivariate normal distribution. When there is a skewed or tailed multivariate distribution, Mahalanobis distances might not detect outliers accurately. In addition to that, Mahalanobis distance does not take local densities or neighboring clusters into account. It calculates the distance for each instance independently from other instances. This can affect the accuracy of detecting outliers which are grouped within clusters.

3.2.2 Local outlier factor (LOF)

Local outlier factor is in the density-based model group. It calculates the local density deviation of each instance to its neighboring instances. It assumes that outliers have low local density compared to other instances. The local outlier factor calculates the local outlier score by comparing each instance's average distance to local density. Higher local outlier factor score indicates that instance is much more far than other instances to the local density. While calculating the score, it checks only k-nearest neighbors' distance to the local density. K is a parameter of neighboring instance count and defined by the user. Local outlier factor has an advantage of detecting outliers with irregular densities and complex patterns. It is robust to any kind of distribution and can identify accurately high and low dimensional sample datasets. One negative side of Local outlier factor is defining the appropriate value for k parameter. It can affect the sensitivity of detection.

3.2.3 DBSCAN (density-based spatial clustering of applications with noise)

The DBSCAN method separates outliers from normal instances based on density connectivity. It has clustering capabilities and identifies outliers as instances which do not belong to any cluster or cluster with a small group of instances. This algorithm runs based on two parameters which are epsilon and minPts. Epsilon parameter limits the rule for maximum distance between two instances to consider these as a neighbor. In addition to that, minPts also defines the minimum number of instances required to create a dense region. DBSCAN selects each instance randomly and expands the cluster by iteratively having connections with neighbors with respect to defined

parameters. This process is recursively repeated until all instances and dense regions are identified.

3.3 Machine Learning Based Methods

Unsupervised outlier detection algorithms used for several decades to identify more complex patterns to detect outliers. By increased computational power of the computers, popularity of unsupervised models is increased. One of the most important advantages is to identify complex patterns and capability of considering relationships between the instances.
Also, these algorithms are designed to detect outliers without labeling the instances. These methods have many different principles to detect outliers. Unsupervised outlier detection methods can be split into 4 different categories as Proximity/Density Based Models, Distribution Based Models, Ensemble Based Models and Deep Learning models. Some of the algorithms use density estimations and some of them use distance-based measures. Ensemble methods are also widely used within this category to combine multiple outlier detection models to achieve better accuracy and robustness. In addition to that, deep learning techniques are also widely used in recent decades such as autoencoders to identify outliers.

Unsupervised algorithms are generally superior alternatives to statistical outlier detection techniques due to their capability of capturing more complex patterns. Also, these approaches enable us to have exploratory approaches to identify anomalies in complex datasets.

3.3.1 Isolation forest

Isolation forest algorithm uses decision trees to isolate outliers. It creates random splits to the data until each outlier is isolated with its own split. It selects each feature randomly within the range of each feature. This process repeats itself till each instance is isolated or till the algorithm reaches its user-defined parameter limits. Outliers are defined as the instances which require less split than the other instances. Outliers will have less path in the dataset compared to the general behavior of the normal instances. That means outliers are not following the typical patterns and split points compared to most of the dataset. These algorithms can also work efficiently when the dataset has complex patterns and high dimensions. Random splitting also contributes to fast computation and robustness even if there is a high dimension and not normal gaussian distribution.

3.3.2 One-class supportv machines (SVM)

One-Class Support Vector Machines method uses a binary classification algorithm. Target of the algorithm to detect outliers by separating normal instances in a high-dimensional feature space by building a hyperplane which covers most of the instances within the sample dataset. Created hyperplane helps to define a boundary and instances outside of the boundary considered as outliers. Algorithm seeks to find a possible high-

dimensional hyperplane to maximize the margin between the instances. While maximizing the margin it also tries to decrease the number of the instances which fall outside of the hyperplane. Advantage of the One-Class Support Vector Machine algorithm is its ability to capture complex hyperplane boundaries. It can handle nonlinearly separable data via using kernel trick. Kernel trick helps to transform the data into a higher-dimensional space which helps also to identify more complex patterns to detect outliers accurately.

3.3.3 Autoencoders

Autoencoders are models that use neural network architecture to detect outliers. It is designed to learn a representation of each instance by encoding to much lower dimension space and later decoding back to its original form. Since outliers are rare events, autoencoders can effectively learn the normal instances better than outliers. Due to better fit to normal instances, reconstruction error for normal instances is much less than outliers. Based on the reconstruction error, autoencoders can split each instance by defined threshold or percentage of expected outlier. Neural networks can capture complex patterns and non-linear relationships in the sample dataset. This feature makes autoencoders effective to identify outliers especially in high-dimensional datasets.

3.4 Feature Based Methods

Feature based methods in anomaly detection process the data around the transformations and manipulation of features in the dataset. Especially they transform the same feature by different techniques into new dimensions. Some of them analyzes each feature independently and some of them uses various techniques to combine features.

3.4.1 Histogram-based outlier detection (HBOS)

Histogram-based outlier detection method uses histograms to identify anomalies in sample dataset. First, the method divides data into equal width bins in the range of feature values. Each bin represents a specific range of values. After binning the data, the method counts the instances in each bin range to determine bin density. It is equivalent to the frequency in each bin. After analyzing each density for all bins, outliers exist when there is significant deviation from overall expected density distribution. Method compares each bin density to the overall density of the sample dataset. It can efficiently detect outliers in large datasets without having the assumption of normal gaussian distribution.

3.4.2 Angle-based outlier detection (ABOD)

Angle-Based outlier detection method calculates the angles for each instance pairwise in multidimensional space to detect anomalies. It calculates variances of the angles for each instance pair by comparing other instances. Outliers have more variance in angle

values than normal instances within the sample dataset. This method particularly works well with high-dimensional datasets.

3.4.3 Principal component analysis (PCA)

Originally principal component analysis method used to reduce dimensionality in sample dataset while maximizing the variance. Outlier detection with PCA can be applied with the same logic that Autoencoders use. PCA first identifies the linear combinations of features that can preserve the maximum variance called as principal components. Then the algorithm projects the data to lower-dimensional space by principal components. However, while reducing the dimension, there is reconstruction error. This error is calculated by measuring the discrepancy between the original instance and projected space instance. Outliers identified with reconstruction error. Instances which have more reconstruction error than the overall dataset behavior are called outliers.

3.5 Evaluation Metrics for Outlier Detection

Evaluation of outlier detection methods is as important as outlier detection model development. Without having the right evaluation, assessment results might be not representing the real world and it can be biased. To have the right metrics and evaluation method is a really important step of outlier detection activities.

Evaluation metrics provide many measures to understand the performance of outlier detection models. It can measure the effectiveness and accuracy of the models. Also, these metrics help identify which outlier detection model performs better than the others. Since each dataset has their own characteristic, it is always beneficial to develop multiple models and compare.

Parameter tuning is also an important part of model development to achieve better accuracies to detect outliers. Evaluation metrics can help to find out the best parameter settings for each model. Threshold selection is one of the good examples for the parameters. Most of the outlier detection algorithms require threshold selection to separate outliers from normal instances. Threshold parameters can be selected with different values and evaluation metrics can be compared. In addition to that, evaluation metrics also keep the control of the models during production. Performance monitoring of the algorithms are also important due to the nature of outliers. Behavior of data can change any time and without a good evaluation mechanism, models can fail to identify accurately.
Effectiveness of the models can be calculated on outliers' real status. After the outlier detection, results can be assessed by the experts. Most used evaluation metrics are True Positive, True Negative, False Positive, False Negative counts. Also, it is possible to derive other metrics from these 4 quantitative measures which is, Accuracy, Precision, Recall and F1 Score. These measures can be derived as follows.

Accuracy

$$(TP + TN) / (TP + TN + FP + FN) \quad (3.1)$$

Precision

$$TP / (TP + FP) \quad (3.2)$$

Recall

$$TP / (TP + FN) \quad (3.3)$$

F1 Score

$$2 * (Precision * Recall) / (Precision + Recall) \quad (3.4)$$

These metrics give insights about the performance of the outlier detection models.

3.6 Data Collection

Purchasing data used for the analysis mainly comes from 2 main systems of SAP. SAP ERP (Enterprise Resource Planning) and SAP SRM (Supplier Relationship Management). These systems are commonly used by many organizations to manage and control purchasing activities.

Data collected directly internal SAP ERP and SAP SRM databases by collaborating with the IT department and system administrators to have the access rights. SAP ERP and SAP SRM systems have a lot of tables regarding material flows, purchasing, invoice, customer and logistics. In order to capture the relevant information, purchasing experts are involved in the data collection phase as well.

First, we tried to analyze which data fields can be important to detect fraudulent activities. Second, we captured all purchasing related data from SAP ERP and SAP SRM databases. These relevant purchasing data from both SAP ERP and SAP SRM systems includes information such as purchase orders with respective purchase order number as identifier of each unique purchase order, invoices, supplier details, item descriptions, quantities, price information, requester, approver, cost center number,

account numbers used to categorize types of financial transactions, currency in purchase order, local currency, tax code, company code, plant code, material type, request date, delivery date, last change date, date of purchasing document, purchasing document category, posting date, number of items in purchasing
document, purchasing group, purchasing organization, supplying vendor, other involved users, status of purchasing document, logical system, exchange rate, current goods receipt, number of deliveries, reason for ordering, material number, target quantity, net price, vendor name, supplier type, type of purchase order, workflow type and related workflow systems. All these data fields can have an impact to identify fraud or abnormalities. Since SAP has complex data relations, first we gathered all these Purchasing Order related data fields and then tried to merge these fields into 1 dataset. More than 15 datasets has been combined into 1 dataset in order to have all data fields in one place. Specific dataset operations have been made by using merge and combine operations based on common key fields.

SAP data includes many different categories in the same tables. Due to the operational complexity it must be stored in the same tables. Filtering these data after collecting from SAP plays a crucial role. To analyze and detect outliers accurately this step must be handled carefully. Fraud and abnormalities can have different characteristics for different types of transactions, such as service orders and material orders have different fraud characteristics. Even data fields can be different by transaction type. By filtering data unsupervised outlier detection models can focus on specific subsets which have unique fraud patterns and specifically can be developed accordingly. For instance, filtering sub datasets based on country information is also essential. Different countries mıght have different regulations, limits, and specific types of SAP rules. Due to these differences, there are different types of country specific fraud behaviors. Filtering based on country also allows us to capture country specific fraud patterns and detect outliers more accurately. By doing so, we can capture deeper insight into fraudulent activities that may vary across different regions. Filtering the data allows us to apply specialized outlier detection models which are more accurate and resulting in mode precise comprehension of fraud risks within specific categories or geographical contexts.

In addition to SAP, we collected the data from further internal company specific systems. These additional data fields can enrich the dataset to capture all fraud or abnormality patterns and risk indicators. These additional sources enable us to evaluate non-SAP factors, such as verifying if the approver assigned in SAP aligns with the appropriate cost center approver. Other important risks can be determining if requests are made by individuals from their own manager or other managers, identifying frequent purchase requesters for all departments, detecting bot users and more. Therefore, during the data collection and preparation step, data from these internal company data sources is gathered to have a holistic approach to understand all possible fraud and abnormality patterns. By doing so, algorithms now can capture a comprehensive understanding of purchasing activities and potential outliers.

3.7 Experimental Setup

The experimental setup is one of the most crucial steps of the analysis. In order to ensure the reproducibility dataset, data preprocessing steps, dataset, and models explained in detail.

The first important thing for experimental setup for the unsupervised outlier detection models is data. Dataset contains 30 different features and 1700 purchasing transactions from Automotive Industry company. The dataset preprocessed in order to remove missing values. Numerical features are normalized due to the importance of magnitude. Scale of the numerical features were so diverse. Categorical variables also preprocessed with one-hot encoding. In addition to that, feature extraction techniques are applied to enrich the data. Feature extraction also increases the accuracy by establishing association between similar transactions. Dataset has no labels available and because of that there is no reason to split dataset as train or test set. All data has been fed into anomaly detection models. 4 different anomaly detection algorithms applied and combined to have more robust results. In this phase there is no parameter setup, and algorithms are fitted with default values due to the lack of knowledge about data characteristics. Algorithms first calculate the anomaly score for each observation separately. After having anomaly scores from each algorithm for each observation, algorithm-based anomaly scores are first normalized and then combined. After the score combination of each observation, the top 1% identified as anomaly. To evaluate the model performance, Anomaly purchasing transactions are shared with purchasing experts to have evaluation. Based on the investigation from purchasing experts, algorithm accuracy was evaluated.

4. CASE STUDY

During case study 4 different anomaly detection algorithm compared with their standard parameters. Each of these algorithms has a unique approach to detect anomalies. Through analysis and experimentations, target is to find out which algorithm is performing best with standard parameters for the same data characteristics. Later, best anomaly detection method selected, and further transformations and techniques applied with the selected most performant anomaly detection model. Experiment done with Gaussian Mixture Model, Local Outlier Factor Model, Isolation Forest Model, DBScan Model.

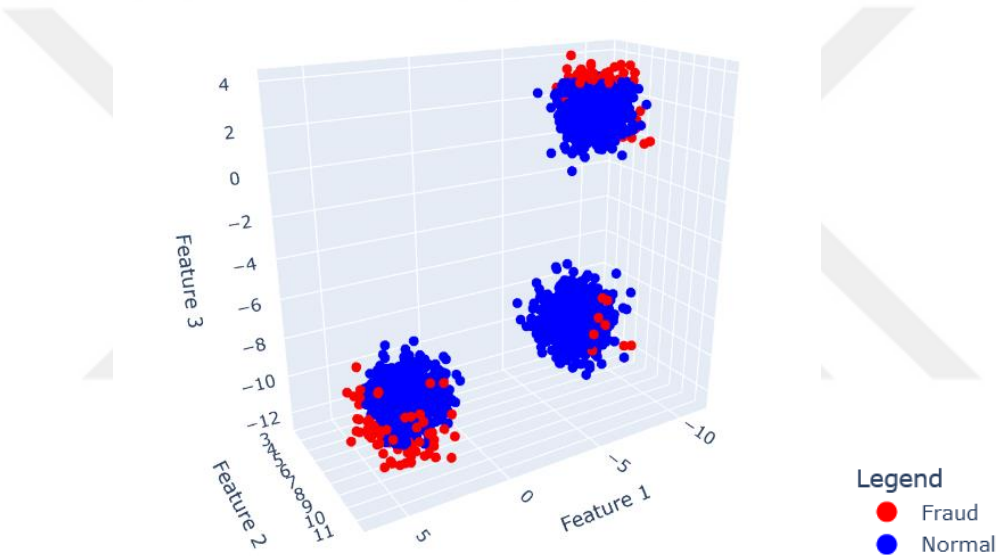

Figure 4.1: 3D Scatter Plot for Comparison Data (Normal and Outlier Data Points)

4.1 Model Comparison

Comparison of the models are curicial to select the base model for further analysis. Sınce we would like to understand the general model behaviour for sample data, each model trained with default parameters. During the training real status of the transactions not fed in to models. After training, each detection model predict results for each transaction. These prediction results used to calculate performance metrics with real status of the transactions. Performance of different detection models can be compared with classification metrics.

Table 4.1: Performance metrics for each detection algorithm.

Metric	GMM	DBSCAN	Isolation Forest	LOF
Accuracy	0.66	0.88	0.80	0.91
Precision	0.12	0.36	0.29	0.35
Recall	0.5	0.58	0.98	0.11
F1 Score	0.19	0.45	0.45	0.17
Specificity	0.68	0.91	0.79	0.98

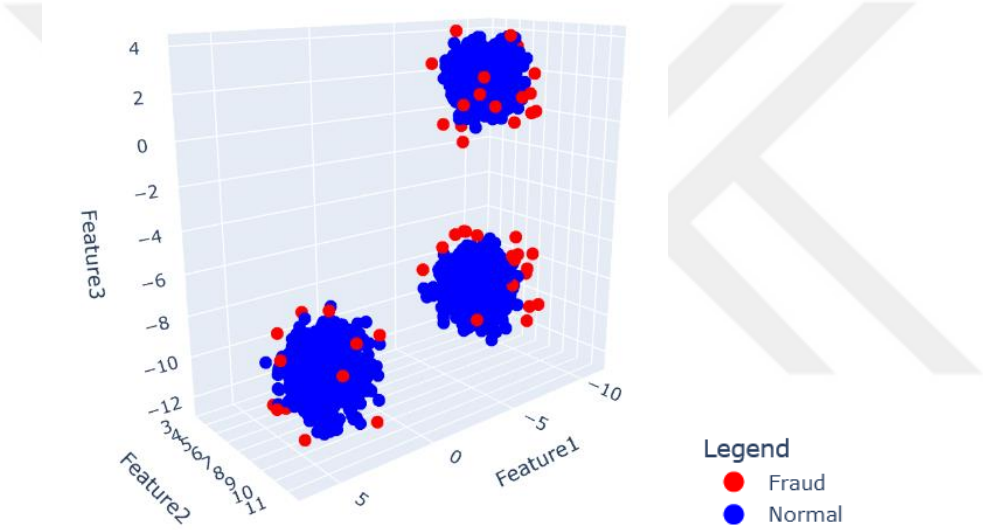

Figure 4.2: 3D Scatter Plot for Local Outlier Factor Algorithm Predictions

As a result of first comparison Local Outlier Factor algorithm selected because it has the best Accuracy. In this use-case the most important parameter is accuracy. Precision and Recall parameters are less important. Since Local outlier factor has the best accuracy, it is selected as base algorithm for further analysis.

4.2 Data Cleaning and Transformations

Data cleaning and transformations are one of the most important steps of using data from SAP and other data sources to detect outliers and frauds with unsupervised models. These steps help us to ensure that data is correct and accurate, consistent, suitable, and fit for the algorithms to use.

4.2.1 Data cleaning

First step of data cleaning is to detect missing values in the dataset. This step includes identifying missing values for each feature and implementing related appropriate strategies to fill with mean, median, regression imputation, special input or drop rows. Some transactions from SAP systems in the purchasing domain can have missing or incomplete supplier details, such as missing supplier names, addresses, or contact information. These fields can be empty due to not properly entered information available at the time of recording the transaction. Another fact is SAP is a live system and purchasing activities never stop. Due to the ongoing process steps at a time, some of the Purchasing orders which have been created on the system might be not finalized. Due to this, some Purchasing orders might hav missing approvers due to the status of purchasing order. In the same way some of the purchasing orders can be canceled and other workflow information could be created. This creates another field for Purchasing documents. Certain fields in the purchasing data may be left blank or undefined, such as purchase requisition numbers, delivery dates or payment terms. In some cases, due to data privacy regulations or company policies, certain fields or attributes may be intentionally masked leading to missing values in those specific fields. As a result, due to the special needs of each purchasing document, some fields might be empty. All of the empty fields have to be evaluated specifically one by one with purchasing experts. By evaluation of missing values, the dataset becomes more accurate, complete and ready for related analysis

Data standardization is a critical step in preparing SAP purchasing data for analysis. It involves ensuring consistency across the dataset by standardizing data formats, units, and variables. For example, currency values from different sources can be converted to a common currency (e.g., converting USD, EUR, and GBP to a unified currency). Numerical values can be normalized to a consistent scale, such as scaling quantities or amounts between 0 and 1. Additionally, dates can be converted into a standardized format (e.g., YYYY-MM-DD) to enable accurate date-based calculations and comparisons. These standardization techniques promote data coherence and enable meaningful analysis across the dataset.

4.2.2 Data transformations

Data transformations have such a critical role while using data from SAP systems to detect outliers and frauds using unsupervised outlier detection models. Data transformation is crucial because these steps can change the nature of the data. Wrong transformations can lead to expect a change in characteristics behavior. Due to the changed characteristics in the dataset, outlier detection models can capture the wrong behaviors which will then lead to inaccurate outlier detection algorithms. Since there are different types of data in SAP like int, float, string, date, and time all fields must be evaluated separately and with care in order to have accurate results. In this section we need to apply two important data transformation processes which are Variable Encoding and Data Scaling.

4.2.2.1 Data scaling

Data scaling is necessary for many statistical methods and important because scaling operations can change the behavior dramatically. Data scaling, also known as feature scaling, helps us to transform numerical values in a data set to the same level of range. Data scaling is not just important to have efficiency for model development but also is necessary to have comparable measurement units. Features of SAP systems can have different scales and units. While some purchase orders can have 1000 kilograms of some specific raw material and have 10 euros as price. To have a better comparison between the variables to detect outliers more accurately, scaling comes handy. Scaling helps us to ensure that features with different units as in the given example like kilograms and euros or percentages are on the same numerical scale. This allows for no biased and meaningful comparison between the variables. When unsupervised outlier detection algorithms are developed on unscaled data, variables with larger magnitudes or ranges can have more impact on the result of analysis. Features with high magnitude can influence the result more than the features with low magnitude. On the other hand, many unsupervised outlier detection models run on distance and similarity calculations. Data scaling will improve the performance of these algorithms. When there is a larger range of features, algorithms calculate the distance with bias means larger the feature more the impact on outlier. Scaling ensures that the algorithm considers all variables equally. Also scaling the data helps us to have a better interpretation. Variables on a similar scale are easier to interpret.

To show the effect of scaling, a small experiment can be developed by using the same algorithm and parameters on scaled and unscaled dataset. This experiment has been studied with a sample dataset which has 100 observations with three numerical features. Dataset created to have 10 known outliers with labels. We need labels to compare the accuracies. During the model implementation, labels are not fed into the algorithm. Local outlier factor algorithm applied for the sake of simplicity. The Local Outlier Factor (LOF) model is an unsupervised anomaly detection model which calculates the local density deviation of a given data point with respect to its neighbors. It considers as outliers the samples that have a substantially lower density than their neighbors. Distances and effects of scaling are important for the algorithm since it creates dense areas accordingly in feature space. To have reusability the local outlier factor algorithm parameters selected as default parameters. Default parameters are used as follows; {'algorithm': 'auto', 'contamination': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None,
‘'n_jobs': None, 'n_neighbors': 20, 'novelty': False, 'p': 2}

There are several metrics used to evaluate the performance of the models. The choice of which metrics to use depends on the specific problem and the importance of different evaluation criteria. In this experiment accuracy, precision, recall, f1 score, specificity and roc curve were compared on the same model but with a scaled and unscaled dataset. Accuracy measures the proportion of correct predictions out of the total number of predictions. However, accuracy alone may not be sufficient if the classes are imbalanced. Precision calculates the proportion of true positives (correctly

predicted positives) out of all positive predictions (true positives + false positives). It represents the model's ability to avoid false positives. Recall (also called sensitivity or true positive rate): Recall calculates the proportion of true positives out of all actual positives (true positives + false negatives). It represents the model's ability to identify all positive instances. F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of precision and recall. F1 score is often used when there is an uneven class distribution. Specificity (also called true negative rate): Specificity calculates the proportion of true negatives out of all actual negatives (true negatives + false positives). It represents the model's ability to identify all negative instances. ROC curve (Receiver Operating Characteristic curve): The ROC curve is a plot that illustrates the performance of a binary classification model across various classification thresholds. It shows the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity). The Area Under the ROC Curve (AUC-ROC) is also commonly used as a summary metric, where a higher value indicates better performance.
Formulation of these metrics will be explained in detail in the Evaluation Metrics section.

Below sample dataset to compare models with known outliers visualized. This multivariate dataset contains 100 instances with 3 features and 10 of the instances known as outliers.

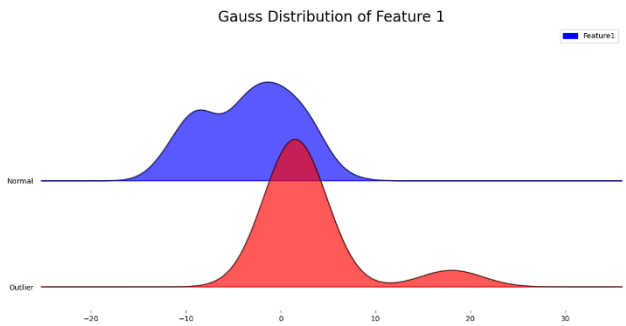

Figure 4.3: Gauss Distribution for Feature 1(Normal and Outlier Distributions)

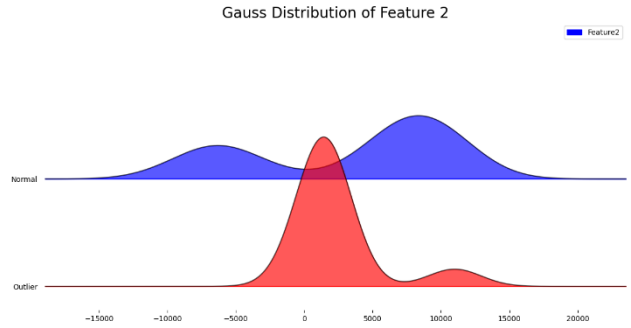

Figure 4.4: Gauss Distribution for Feature 2(Normal and Outlier Distributions)

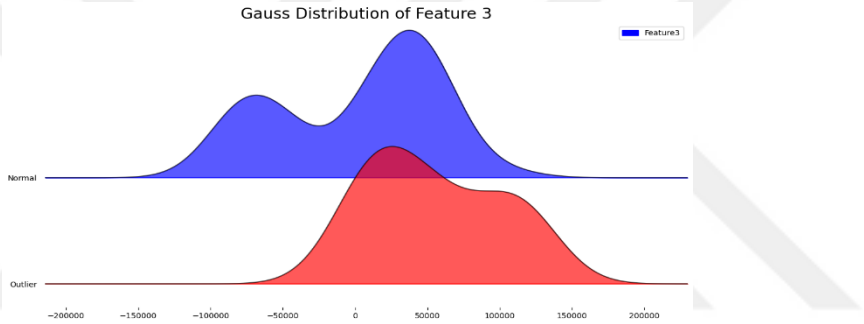

Figure 4.5: Gauss Distribution for Feature 3(Normal and Outlier Distributions)

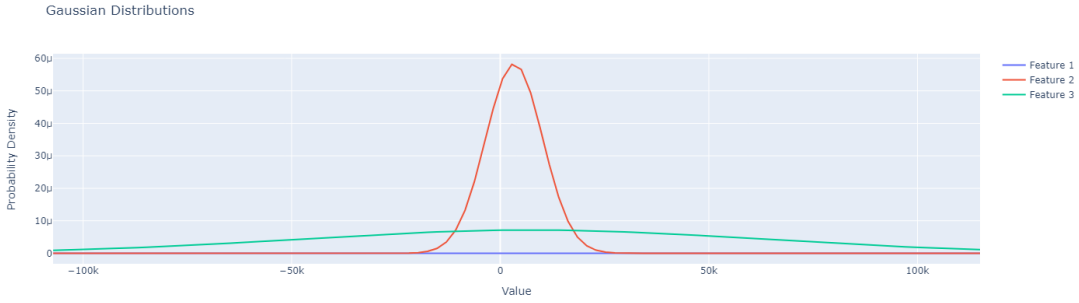

Figure 4.6: Gauss distributions of Features are different due to magnitude difference.

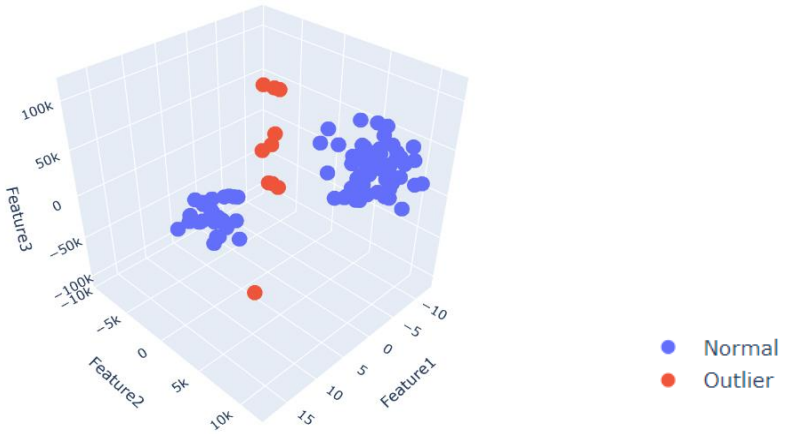

Figure 4.7: 3D representation of Dataset.

After visualizing the dataset, the Local Outlier Factor algorithm was implemented. Feature1, Feature2 and Feature3 fed into the model. Due to the inaccurate detection as we can see Local Outlier Factor model without scaling shows poor performance and misdetect the outliers.

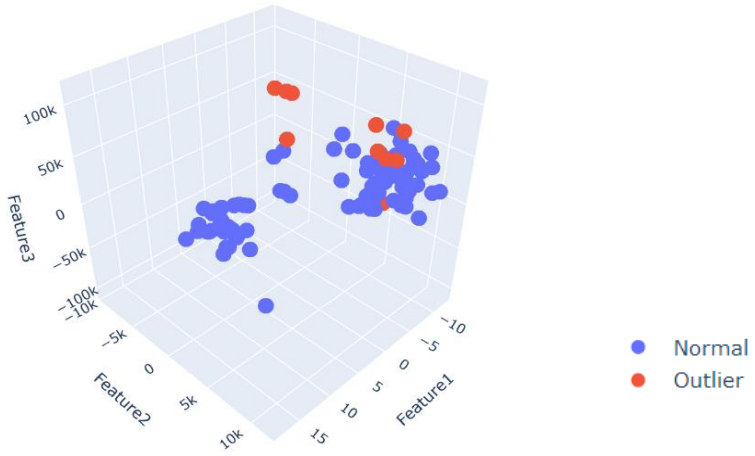

Figure 4.8: Graph represents the predictions of Local Outlier Factor without scaling.

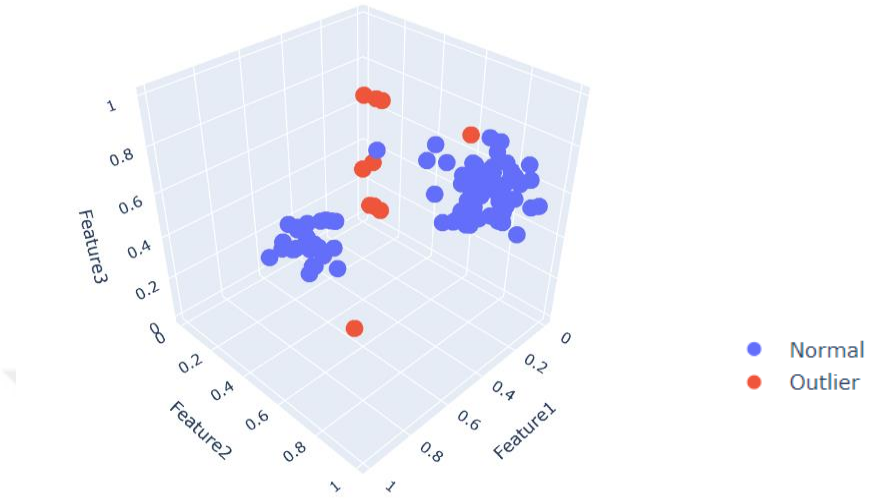

Figure 4.9: Graph represents the predictions of Local Outlier Factor with scaling.

Secondly the dataset scaled into the range 0 and 1 and then Local outlier Factor model applied to scaled features. Scaled Feature1, Feature2 and Feature3 fed into the model. Compared to the previous model, scaling helped a lot to increase the performance and effectiveness of the Local Outlier Factor model. As we can also observe from graphs, distance based unsupervised outlier detection algorithms can be affected from the magnitude of the features.

In addition to graphical representation, we check the evaluation metrics to be able to have subject results.

Table 4.2: Scaling effect comparison for Local Outlier Factor algorithm.

Metric	LOF without Scaling	LOF with Scaling
Accuracy	0.88	0.98
Precision	0.4	0.9
Recall	0.4	0.9
F1 Score	0.40	0.9
Specificity	0.93	0.98

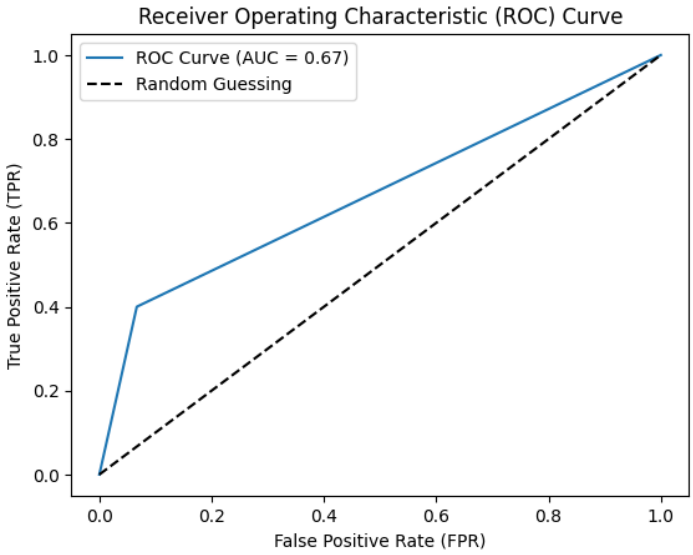

Figure 4.10: ROC Curves without scaling.

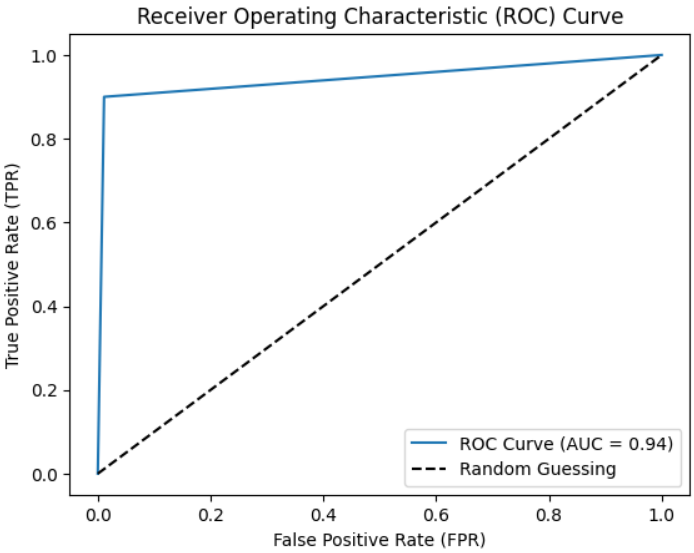

Figure 4.11: ROC Curves with scaling.

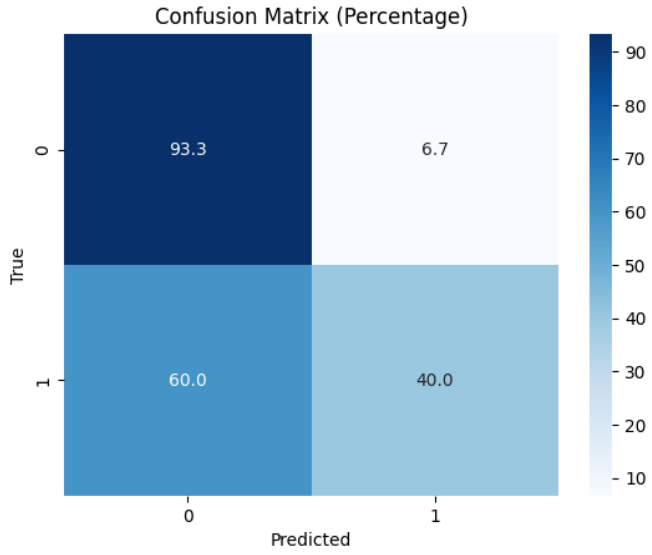

Figure 4.12: Confusion Matrix without scaling.

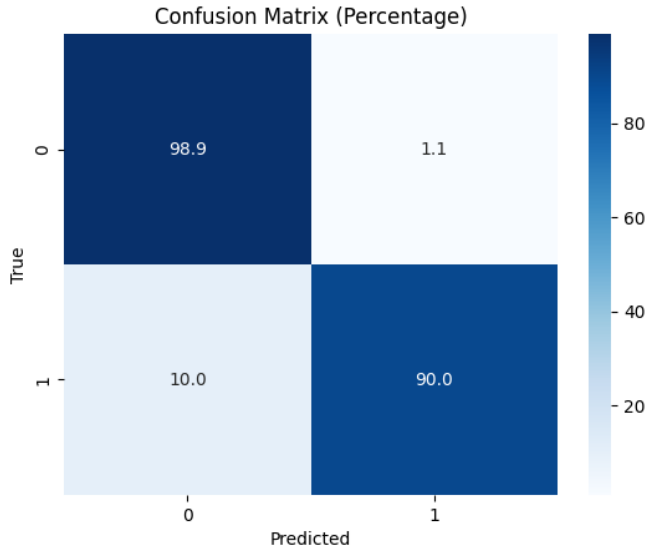

Figure 4.13: Confusion Matrix with scaling.

Using scaling improves all the evaluation metrics for LOF model. To challenge the scaling, another experiment has been designed. Again, for sake of simplicity three random features will be used. Clusters created with more standard deviation to have a more complex dataset.

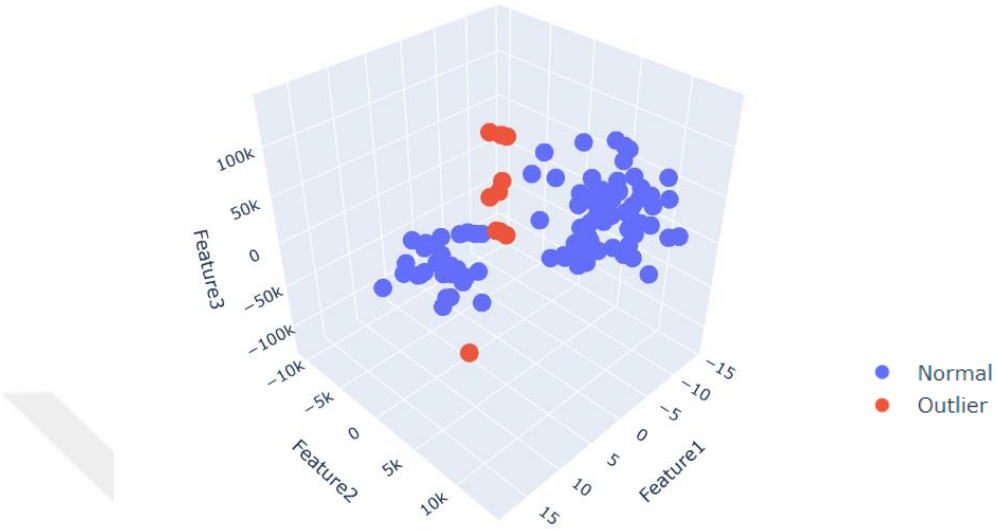

Figure 4.14: 3D representation of 2nd Dataset.

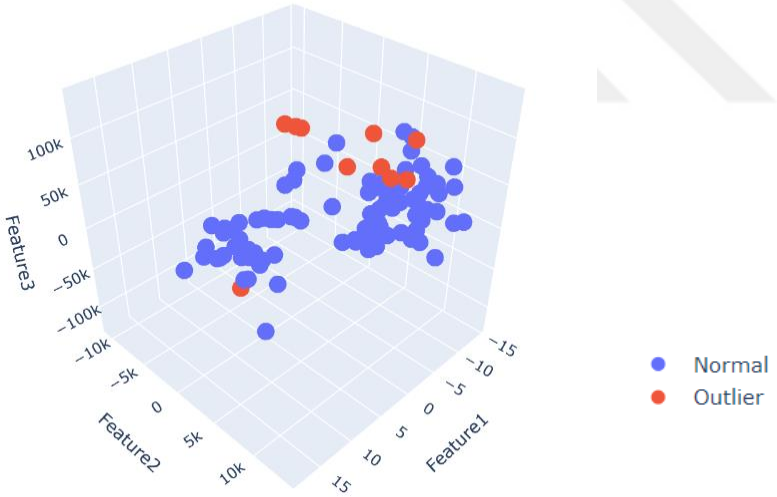

Figure 4.15: Predictions of Local Outlier Factor without scaling.

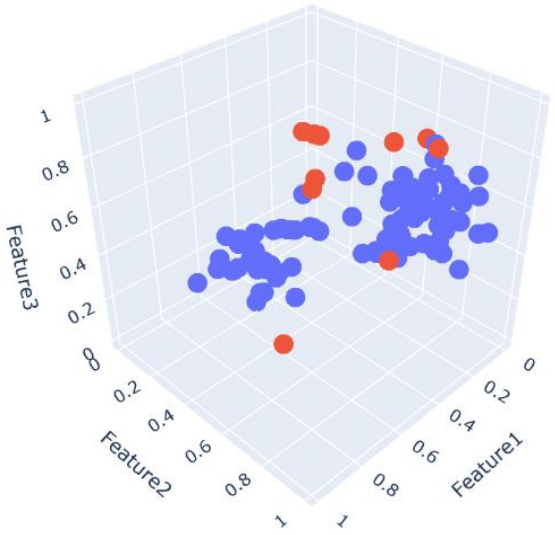

Figure 4.16: Predictions of Local Outlier Factor with scaling.

In this dataset evaluation metric for both models are lower compared to previous experiment due to the more complex dataset and noise.
In addition to graphical representation, we check the evaluation metrics to be able to have subject results for the second experiment too.

Table 4.3: Scaling effect comparison with another dataset for Local Outlier Factor algorithm.

Metric	LOF without Scaling	LOF with Scaling
Accuracy	0.86	0.92
Precision	0.3	0.3
Recall	0.3	0.6
F1 Score	0.3	0.6
Specificity	0.92	0.95

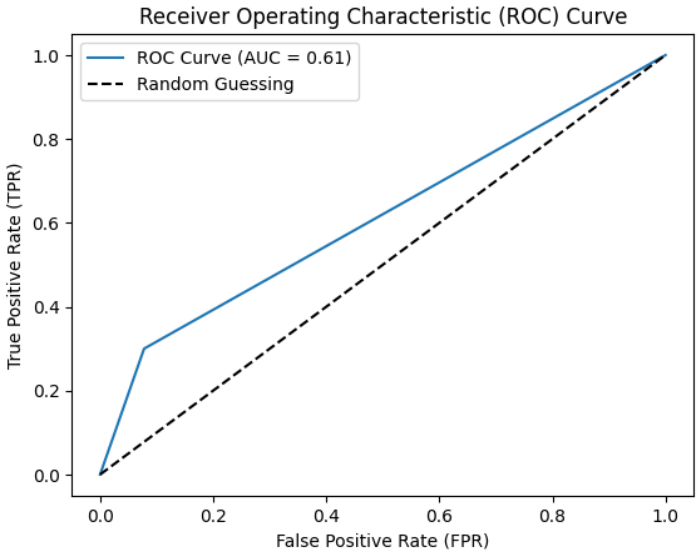

Figure 4.17: ROC Curves with scaling for 2nd Dataset.

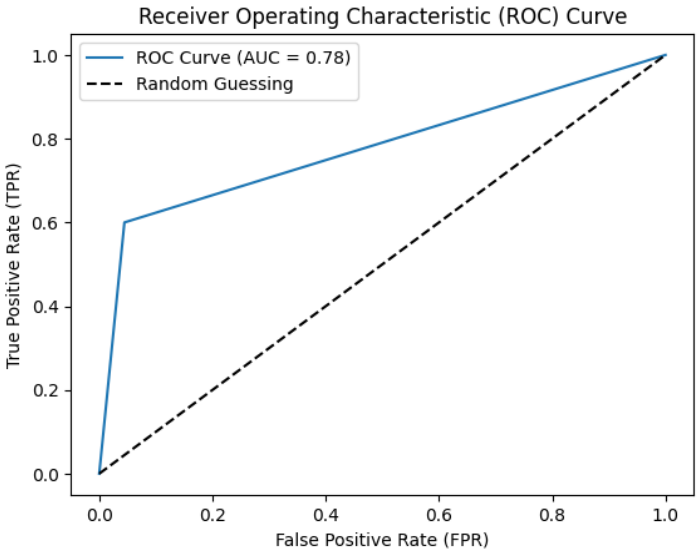

Figure 4.18: ROC Curves with scaling for 2nd Dataset.

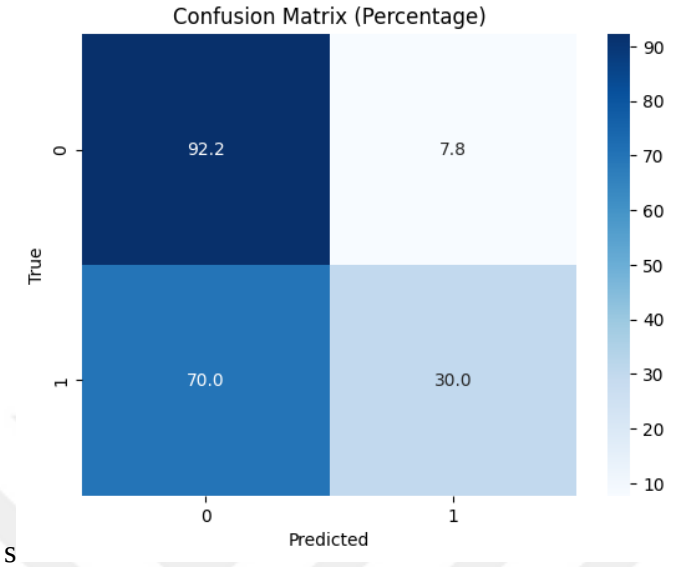

Figure 4.19: Confusion Matrix without scaling for 2nd Dataset.

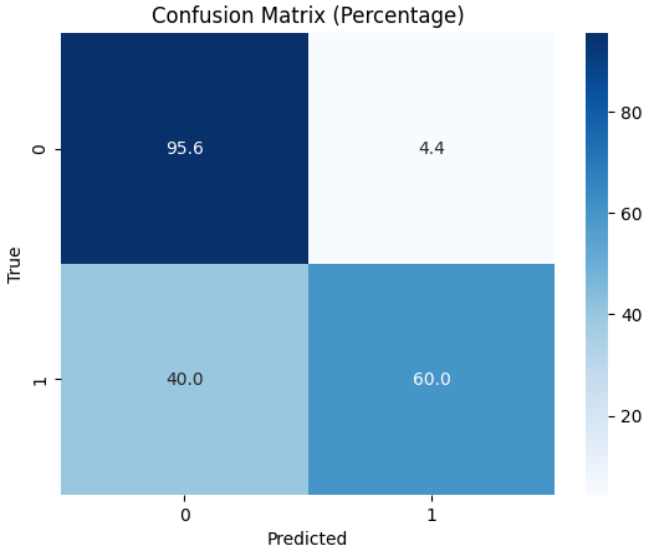

Figure 4.20: Confusion Matrix with scaling for 2nd Dataset**.**

These experiments show us, when there is a magnitude difference between the variables, it is important to scale the data. Scaling increases the efficiency of unsupervised outlier detection models.

4.2 Feature Extraction

Feature extraction has an important effect on model efficiency. To enrich the data this step should be handled with care. The target of the feature extraction is to transform

raw data to meaningful and representative features that can capture the complex patterns and characteristics of the purchasing processes.

Two main aspects exist for this study. First aspect of the feature extraction is using the domain knowledge. In the context of purchasing area for fraud detection, domain specific knowledge can be very useful to extract features. This can include creating special rules, thresholds and using expert knowledge for the feature extraction process. For example, each country has some specific set of purchasing limits per concept or vendor categories can be used to extract important features to capture deviations from expected behaviors. By using domain knowledge also, we can create additional columns by using existing data with some specific calculations. These extracted features will help to capture complex rules and will create some associations between the instances. Some of the extracted columns regarding to domain knowledge is given below.

Purchase Order Amount: The total value of purchase orders by Vendor, by User, by Approver and by Cost Center can be a significant feature for identifying outliers. Unusual high or low values may indicate risks in purchasing activities. In addition to that pairwise calculations can also be an important risk indicator. Total Purchase Order Amount from User to specific Vendor, User to Specific Approver, Approver to Specific Vendor and User to Approver and then to specific Cost center is some of the pairwise calculations that can identify risks accurately.

Purchase Order Item Quantity: Quantity of items requested in a single purchase order can be also a risk indicator. High or low quantities can suggest irregularities.

Vendor Category: Vendor types are also an important risk indicator. Based on the vendor types, such as location, size, order amount per year and relations in years with companies can help us to identify fraudulent activities. For example, a company which has been delivering goods for 10 years has relatively low risk compared to a company which started to deliver goods 1 year back.

Purchasing Frequency: The frequency at which specific user or department requests purchase order can be informative as well. Dramatically increasing demand can also be an important risk indicator. Spikes or drops which are not common for department or user behavior can be a good feature to control to detect outliers.

Purchase Order Approval Time: Generally, the time taken for purchase order to be approved by manager can provide insights as well. Unusual short times or unusual long times can have an important role. Normally this information is in a dataset as two separated columns as created date and approved date. But with just basic calculations we can also gather this valuable information.

Purchase Order Line-Item Descriptions: Text descriptions can be also important for identifying anomalies. It is possible to extract many additional features from descriptions like character count, company name check, routing information and many more. When the general behavior of the user is to write 30 letters and for specific Purchase Order if the user writes just 3 letters this can indicate some information, not necessarily fraud but something to control. Also, the routing purchasing department to specifically some companies can be identified as a risk factor. Normally purchasing departments decides for the company and users shouldnt involve. If some user

specifically mentions some specific vendor name, this can be also extracted as a feature. Many more features which are not mentioned here are extracted by discussing with Purchase experts to get the full potential of the data.

Second aspect is extracting features by using statistical methods like Principal Component Analysis, Partial Least Square Discriminant Analysis, and Independent Component Analysis. In addition to that feature selection also aims to identify the most relevant subset of the features from the original dataset.

To understand the importance of feature extraction, the same model applied to datasets. First dataset has just original features and the second dataset has original features and extracted features from the same dataset.

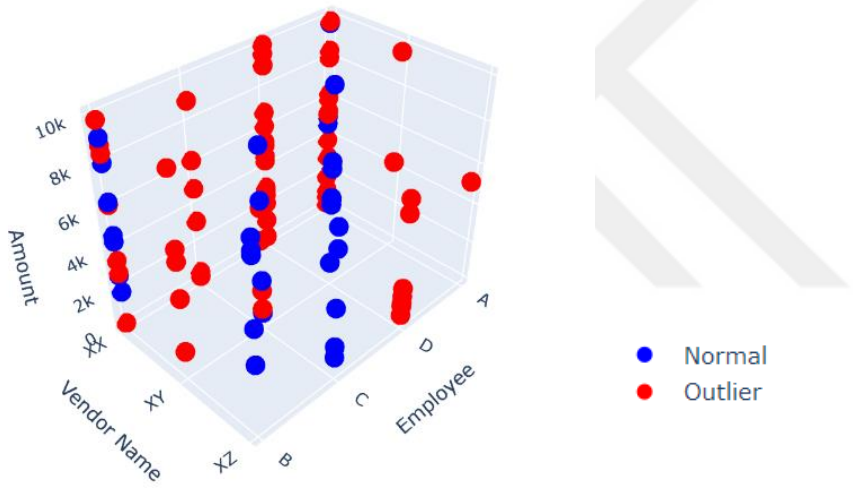

Figure 4.21: 3D Visualisation of Predictions for Existing Dataset without feature extraction.

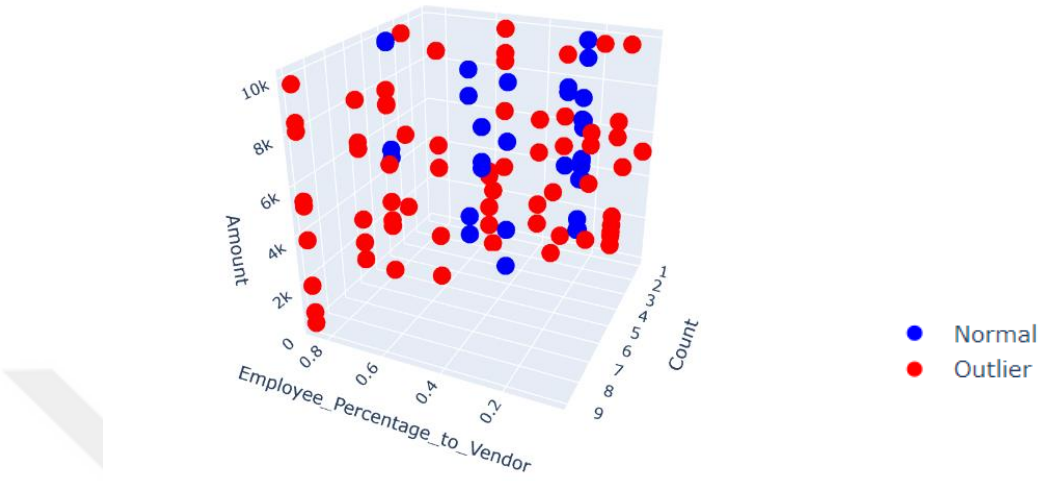

Figure 4.22: 3D Visualisation of Predictions for Existing Dataset with feature extraction.

Since performance of the model for this sample dataset is not good enough, we need to also compare model evaluation metrics.

Table 4.4: Feature Extraction effect comparison for Local Outlier Factor algorithm.

Metric	LOF without Extraction	LOF with Extraction
Accuracy	0.36	0.54
Precision	0.25	0.38
Recall	0.6	0.86
F1 Score	0.36	0.53
Specificity	0.25	0.4

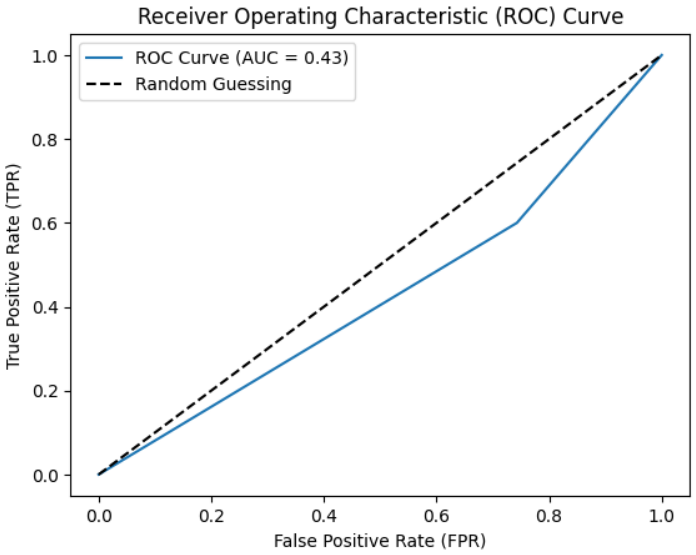

Figure 4.23: ROC Curves without extraction.

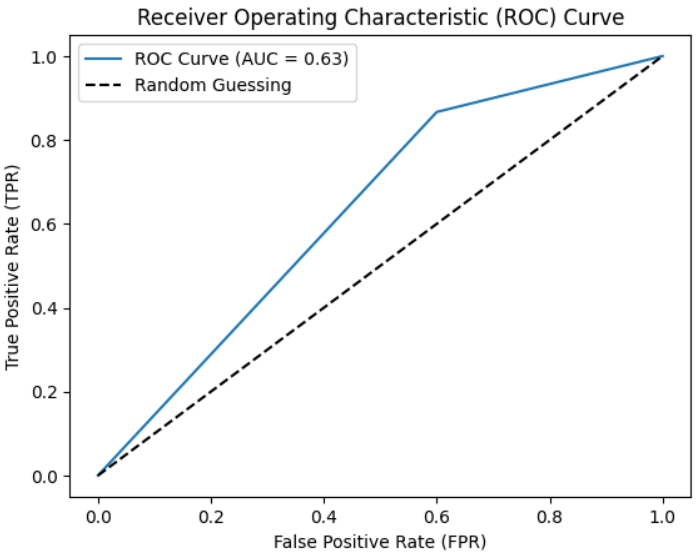

Figure 4.24: ROC Curves with extraction.

After ROC curves also confusion matrix shows clear advantage of feature extraction as well.

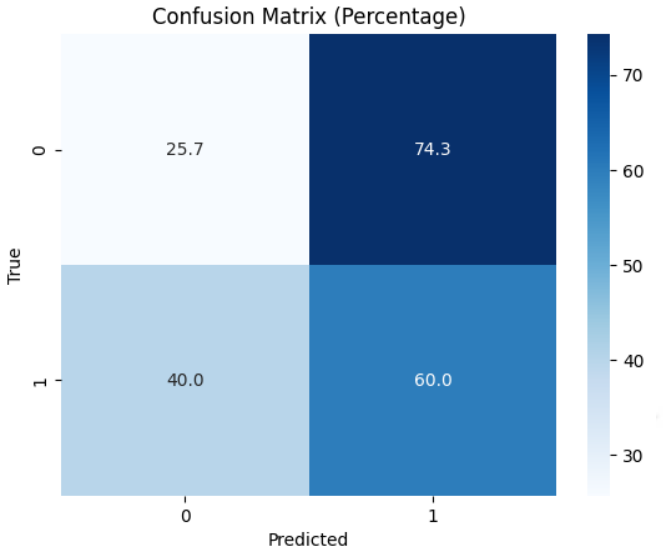

Figure 4.25: Confusion Matrix without feature extraction

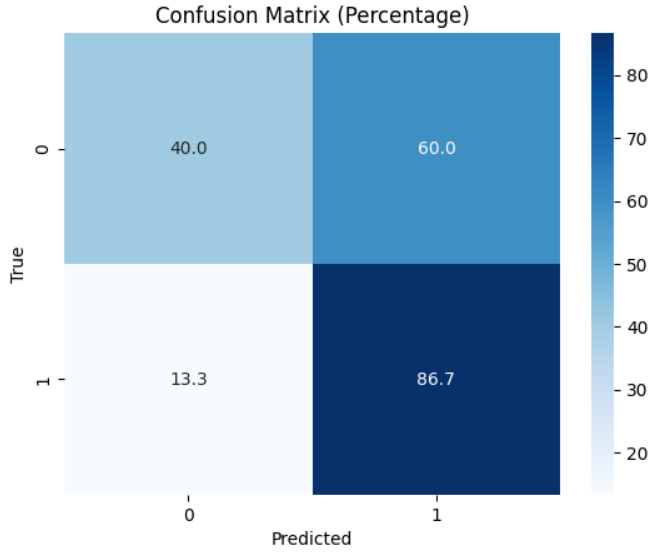

Figure 4.26: Confusion Matrix with feature extraction.

5. EVALUATION STRATEGY

There are various approaches to evaluate unsupervised outlier detection techniques in the purchasing domain using data from SAP. Due to the lack of labels, there is a major difference for evaluation strategy. Conventional evaluation metrics like accuracy or precision are not directly suitable to use. These metrics require labeled data and can not be directly used. However, there are other strategies to evaluate the performance of the models.

5.1 Visual Inspection

Visual inspection is one of the most used techniques to understand the outlier or fraud behavior. In various dimensions data can be plotted and different patterns can be observed. This allows domain experts to examine results. Based on these visualizations' domain experts can provide qualitative feedback. They can identify if the related features are working well with the model to detect outliers.

5.2 Business Impact Analysis

This method allows us to understand the performance of the model just when the model is deployed. Domain experts can evaluate the impact of the detected outliers on the purchasing process with relevant business metrics. They assess if the detected outliers are indicators for potential fraud cases or irregularities. This can have an impact on financial performance. We can consider the potential impact of the detected outliers on key business indicators to provide models' efficacy.

5.3 Anomaly Confirmation

This method is the most significant way to assess models' performance. Domain experts manually check the outliers which are detected by unsupervised outlier detection models. By reviewing and confirming a subset of outliers, it is possible to verify whether predictions are correct or not. In this phase, based on the evaluation of domain experts, we use conventional evaluation metrics as well to assess the model performance.

5.4 Comparison with Baseline Methods

Baseline methods are simple yet sometimes also effective in order to detect anomalies in the purchasing domain. There are many rule-based controls in order to mitigate the risks of fraud. Some of the rules compared with unsupervised outlier detection algorithms predictions. This comparison can provide feedback also about model performance.

5.5 Evaluation with Synthetic Data

Unsupervised outlier detection algorithms evaluated with synthetic data which has been prepared by domain experts to assess model performance. This data was prepared with domain experts and fed into unsupervised outlier detection models to understand the model behavior for known characteristics of fraud. In addition to that, by generating synthetic data specialized for business needs, this method ensures and controls the model performance for expected fraud characteristics which is occurred, or which can happen in the future.

6. RESULTS AND ANALYSIS

To compare and understand the accuracy of the unsupervised outlier detection model, a study was executed. In this study information of 934 transactions collected from related SAP sources. This data covers one specific manufacturing plant for 1 year. Unsupervised outlier detection algorithm ran on this population to calculate risk score of each observation. After risk score calculation ~%10 of the highest risk transactions are manually investigated by purchasing experts. 95 transactions are selected from manual investigation.

Manual investigation results show that out of 95 selected transactions 58 of them were having some fraudulent activity in different categories. This means the developed model have detection rate of ~%61. This is a significant improvement for the business. In daily business, generally %1 / %3 of all transactions can have fraudulent activity. By using this outlier detection model, purchasing experts can focus on the riskiest transactions.

6.1. Limitations

This study is offering limited results within a specific area of purchasing. Due to the different characteristics of each purchasing area, the same algorithms might not work well in other areas. Also, the dataset which was evaluated is having purchasing transactions just from one specific country. It is not logical to apply the same model to other countries or purchasing domains. Each specific case should be handled separately. Due to the characteristics of the data, anomalies can be in different structures for each specific area. Data preprocessing steps may also vary data quality might be different for each purchasing domain. Country specific rules and features also makes it harder to have one unique model. This study also does not focus on anomaly detection model parameter fine tuning or selection. This study was executed for a specific area of purchasing domain to have general understanding of how to use unsupervised outlier detection algorithms. Due to the complexity of the fraud, data selection also might have some problems. Biased data could be evaluated for some time range. Due to these limitations and realities any assumptions made during analysis can have potential validity problems in the future. Currently algorithms are working on a really limited data volume and in the further studies, models will be tested on other datasets as well. Generalizability capability is important, but it is too early to prove it.

7. CONCLUSION

In this study, we implemented different fraud detection algorithms to secure the business transactions which is vital especially for big companies. As purchasing volume increases, companies must deal with detecting fraudulent activities which can lead to financial losses. Conventional methods and controls are still important yet not effective for complex fraud patterns. To be able to solve this issue, the study focused on many different unsupervised outlier detection algorithms to automate the system and enhance business security.
The motivation of this research was to develop suitable, robust, and scalable outlier detection models that can adapt the process and improve the current detection rates. Many unsupervised outlier detection models compared and aimed to improve their performance by using various feature extraction methods.

The findings indicate that the currently implemented model which uses purchasing data from SAP and additional extracted features, achieved a detection rate ~%61.
This is a significant improvement for business as it allows also purchasing experts to focus on their efforts on the riskiest transactions.

In addition to that, the advantage of using unsupervised outlier detection models is that these models can offer a more general and proactive approach to fraud detection. These algorithms don't rely on hard-coded rules and can adapt directly to new and unseen fraud concepts. It can increase operational efficiency and minimize human error.

As a conclusion, integration of unsupervised outlier detection algorithms with traditional statistical methods increases the company's fraud detection capabilities and the effectiveness of existing fraud prevention systems. This can help companies for a safer environment, protect their financial assets and reputation.

REFERENCES

Aarthi Reddy, Meredith Ordway-West, Melissa Lee, Matt Dugan, Joshua Whitney Ronen Kahana, Brad Ford, Johan Muedsam, Austin Henslee, Max Rao (2017). Using Gaussian Mixture Models to Detect Outliers in Seasonal Univariate Network Traffic. *2017 IEEE Symposium on Security and Privacy Workshops*, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8227312.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, Haifeng Chen (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. *ICLR 2018 Conference, from* https://openreview.net/pdf?id=BJJLHbb0-

Chamal Gomes, Zhuo Jin, Hailiang Yang (2021). Insurance Fraud Detection with Unsupervised Deep Learning. DOI:10.1111/jori.12359.

Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, Sal Stolfo. A Geometric Framework for Unsupervised Anomaly Detection. DOI: 10.1007/978-1-4615-0953-0_4.

Fengjun Zhang, Guanjun Liu, Zhenchuan Li, Chungang Yan, Changjun Jiang (2019). GMM-based Undersampling and Its Application for Credit Card Fraud Detection. Conference: *International Joint Conference on Neural Networks (IJCNN2019) At: Budapest, Hungary*, from https://www.researchgate.net/publication/336412402_GMM-based_Undersampling_and_Its_Application_for_Credit_Card_Fraud_Detection

Guo Pu, Lijuan Wang, Jun Shen and Fang Dong (2021). A Hybrid Unsupervised Clustering-Based Anomaly Detection Method. DOI:10.26599/TST.2019.9010051.

Iwan Syarif, Adam Prugel-Bennett, Gary Wills. Unsupervised Clustering Approach for Network Anomaly Detection. DOI:10.1007/978-3-642-30507-8_13.

Jiong Zhang and Mohammad Zulkernine. Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection. DOI: 10.1109/ICC.2006.255127.

Ke Nian, Haofan Zhang,, Aditya Tayal, Thomas Coleman, Yuying Li (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. DOI:10.1016/j.jfds.2016.03.001.

Markus Goldstein, Seiichi Uchida (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. DOI: 10.1371/journal.pone.0152173.

Miin-Shen Yang, Chien-Yo Lai, Chih-Ying Lin (2011). A robust EM clustering algorithm for Gaussian mixture models from https://doi.org/10.1016/j.patcog.2012.04.031

Peng An, Zhiyuan Wang, Chunjiong Zhang (2022). Ensemble unsupervised autoencoders and Gaussian mixture model for cyberattack detection. DOI:10.1016/j.ipm.2021.102844.

Tangqing Li, Zheng Wang, Siying Liu and Wen-Yan Lin (2021). Deep Unsupervised Anomaly Detection. DOI: 10.1109/WACV48630.2021.00368.

Varun Chandola, Arindam Banerjee, and Vipin Kumar (2007). Anomaly Detection: A Survey. DOI:10.1145/1541880.1541882.

Varun Chandola, Arindam Banerjee, and Vipin Kumar (2009). Anomaly Detection for Discrete Sequences: A Survey. DOI:10.1109/TKDE.2010.235.

Varun Chandola, Varun Mithal, and Vipin Kumar (2008). A Comparative Evaluation of Anomaly Detection Techniques for Sequence Data. DOI:10.1109/ICDM.2008.151.

Varun Chandola, Varun Mithal, and Vipin Kumar (2009). Understanding Anomaly Detection Techniques for Symbolic Sequences. Technical report from https://www.academia.edu/324784/Understanding_Anomaly_Detection_Techniques_for_Symbolic_Sequences?uc-sb-sw=34272414

Viet Tra, Manar Amayri, Nizar Bouguila (2021). Outlier detection via multiclass deep autoencoding Gaussian mixture model for building chiller diagnosis. DOI:10.1016/j.enbuild.2022.111893

Wenbo Liu, Delong Cui, Zhiping Peng, Jihai Zhong (2019). Outlier Detection Algorithm Based on Gaussian Mixture Model. DOI: 10.1109/ICPICS47731.2019.8942474

Xiaodan Xu, Huawen Liu, Li Li, Minghai Yao (2018). A Comparison of Outlier Detection Techniques for High-Dimensional Data. DOI:10.2991/ijcis.11.1.50

CURRICULUM VITAE

Name Surname **:** Yiğit Can TAŞOĞLU

EDUCATION **:**

- B.Sc. : 2017, Uludag University, Electrical and Electronics Engineering Department

PROFESSIONAL EXPERIENCE AND REWARDS

- 2016 - 2017 Robert Bosch A.S. Intern
- 2017 - 2019 Robert Bosch A.S. Production Engineer and I4.0 Responsible
- 2019 - 2022 Robert Bosch A.S. Data Scientist
- 2022 - 2024 Robert Bosch GmbH Data Scientist
- 2024 Robert Bosch GmbH Data Scientist and Product Owner