Using Machine Learning to Identify Anomalous Activities for Data Leakage Detection

SAN DIEGO STATE UNIVERSITY

Sheng-Chun Lim and Hunter Paul
Computer Science Department
San Diego State University

Introduction

- Data leakage has become a critical concern for modern organizations, posing risks such as financial losses, reputational damage, and legal liabilities. The consequences of exposing sensitive information can be severe and far-reaching. With the increasing reliance on digital systems and the rapid growth of data, there is an urgent need for efficient and accurate methods to detect data leakage.
- This research aims to investigate the effectiveness of machine learning techniques in identifying data leakage by focusing on anomaly detection in user activities within a computer network or system.

Background

- Anomaly detection techniques have traditionally focused on single paradigms, such as unsupervised learning. Semi-supervised and supervised methods are often underutilized in the context of data leakage detection.
- Lack of comprehensive studies that explore diverse machine learning paradigms for anomaly detection

Dataset

- Data Leakage Detection Dataset from Kaggle
- The dataset captures various aspects of user interactions with the system and the presence of abnormalities in user behavior
- 49,500 records x 15 columns (43,560 records x 11 columns are used)

Activity Info	Authentication methods	Actions	Target variable
Authority levels*	Password	Data modification	Abnormality (31% are abnormal)
Activity timing+	PIN	Confidential data access	
Data sensitivity*	Multi-Factor Authentication (MFA)	File transfer	
		File operation*	
		External interactions	

All variables are binary, except for:
* Categorical variables + Continuous variables

Approach

Procedures

- **Data Cleaning**
  - Remove missing data (12%)
- **Data Transformation**
  - Transform categorical features into one-hot encoded
  - Transform datetime data into seconds
- **Feature Extraction**
  - Extract13 features from 17 columns using principal component analysis (PCA)
- **Modeling**
  - 80% for training set, 20% for testing set

Models

Paradigm	Model	Description
Unsupervised Learning	Isolation Forest	Detect anomalies by isolating data points in a high-dimensional space
Semi-Supervised Learning	Autoencoder	Learn to reconstruct normal patterns and flag deviations as anomalies
Supervised Learning	Logistic Regression	A generalized linear model that predicts the probability of a binary outcome based on input features
	Decision Tree	A hierarchical model that splits data into branches based on feature thresholds to classify or predict outcomes
	Random Forest	An ensemble method using multiple decision trees to improve accuracy and reduce overfitting
	Support Vector Machine (SVM)	Finds an optimal hyperplane to classify data points, particularly effective in high-dimensional spaces
	XGBoost	An efficient gradient-boosting algorithm designed for speed and performance

All models were tuned using 10-fold cross-validation to optimize hyperparameters.

Model Evaluation

Metrics	Formula	Meaning
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	An overall performance measure
Precision	$\frac{TP}{TP+FP}$	The model's reliability in detecting anomalies
Recall	$\frac{TP}{TP+FN}$	The model's ability to capture all anomalies
F1-Score	$\frac{2 \cdot precision \cdot recall}{precision + recall}$	The harmonic mean of precision and recall, balancing false positives and false negatives

Results

Model Performance on the test dataset

	Accuracy	Precision	Recall	F1-Score
Isolation Forest	.63	.59	.63	.61
Autoencoder	.67	.56	.67	.57
Logistic Regression	.70	.67	.70	.67
Decision Tree	.69	.69	.69	.69
Random Forest	.77	.76	.77	.76
SVM	.72	.76	.72	.73
XGBoost	**.77**	**.77**	**.77**	**.77**

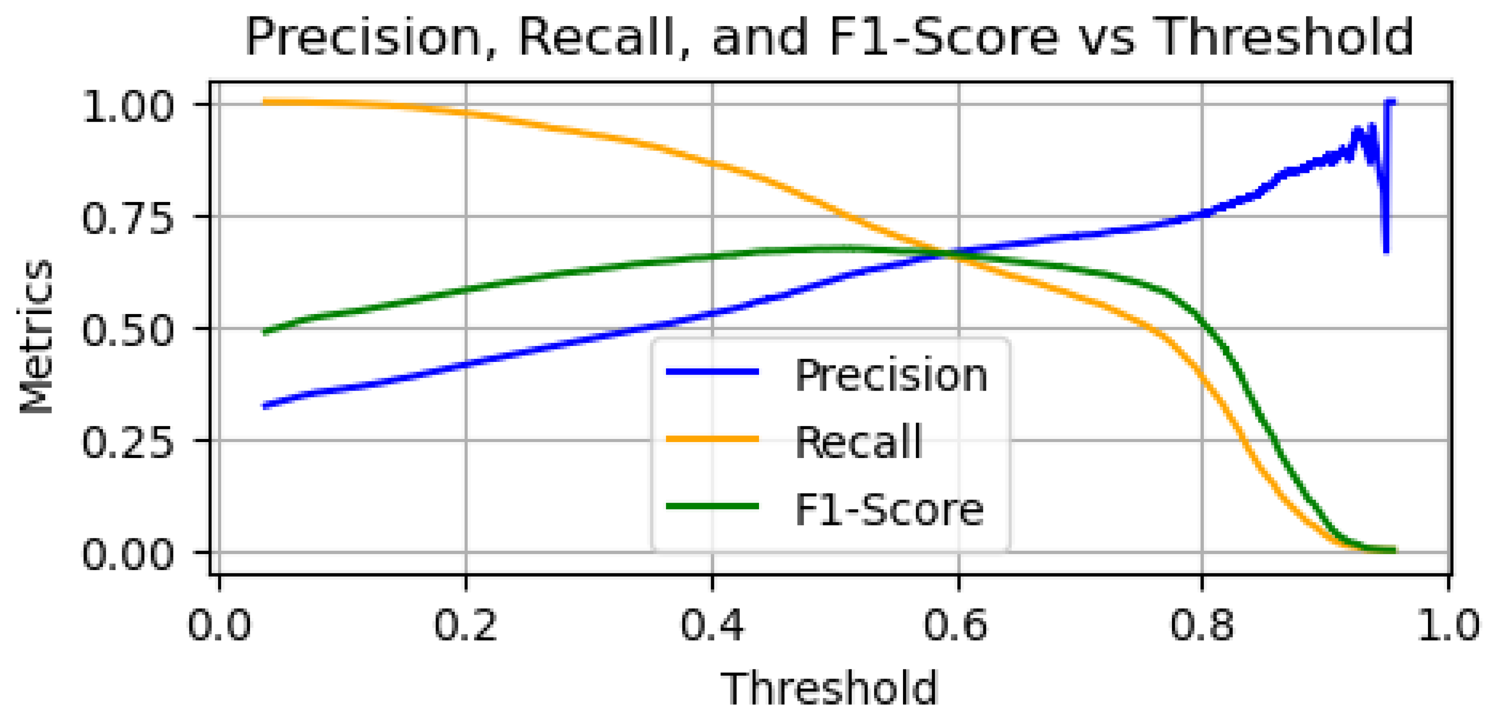

Figure. Precision/Recal/F1 Curve of XGBoost Model on training set
The optimal threshold was determined at the "golden cross"–the point where precision equals recall, which is 0.5919.

Conclusions

- Key Findings:
  - XGBoost demonstrated the best performance, making it the most suitable for anomaly detection for data leaks in the given dataset
- Future Work:
  - Integrate models into real-time monitoring systems
  - Explore ensemble methods combining the strengths of multiple paradigms

References

Nassif, A. B., Talib, M. A., Nasir, Q., Albadani, H., & Dakalbab, F. M. (2021). Machine learning for cloud security: a systematic review. *IEEE Access*, *9*, 20717-20735.
Naseer, S., Saleem, Y., Khalid, S., Bashir, M. K., Han, J., Iqbal, M. M., & Han, K. (2018). Enhanced network anomaly detection based on deep neural networks. *IEEE access*, *6*, 48231-48246.