University of Maryland, Baltimore County

Detecting Insider Threats: A Comparative Analysis of Supervised and Unsupervised Machine Learning Approaches

BY: Aswin Kumar Janakiraman, Kush Patel, Pavan Singampalli, Vinay Krishna Kumar

Abstract

Insider threats pose a significant cybersecurity challenge, accounting for approximately 60% of all cyberattacks. This research investigates applying supervised and unsupervised machine learning techniques to detect anomalous user behaviors that may indicate potential security risks. Utilizing the CERT Insider Threat Test Dataset from Carnegie Mellon University, comprising over 850,000 log entries including login events, emails, and device information, the study employs a comprehensive approach to anomaly detection. The research develops and evaluates multiple machine learning models, including Logistic regression, Support Vector Machines (SVM), K-nearest neighbors (KNN), and an ensemble approach of KNN and Random Forest. Key challenges addressed include data imbalance and model overfitting, with novel strategies such as feature engineering, rule-based suspicious activity labeling, and planned data augmentation techniques like SMOTE. Preliminary results demonstrate exceptional performance with Logistic Regression, KNN and SVM models achieving up to 99% precision and 96% recall calls for addressing the issue of overfitting and dataset imbalance. That's where the Ensemble model of KNN & RF which produces better results. On the other side, unsupervised learning models like DBSCAN and Hierarchical Clustering are implemented with advanced regularization techniques. This research contributes to the critical field of insider threat detection by showcasing the potential of machine learning in identifying subtle behavioral anomalies that may indicate potential security risks.

Methodology

Data Preprocessing:

- **Cleaning:** Removed duplicates, handled missing values (e.g., CC, BCC), and standardized timestamps for consistency.
- **Feature Engineering:** Created key features to capture user activity patterns, such as email size, attachment presence, and recipient count. Anomaly indicators were derived for detecting unusual behavior.
- **Scaling:** Applied Z-score standardization and Min-Max scaling to prepare data for models like Logistic Regression, KNN, and SVM.
- **Imbalanced Data:** Used resampling techniques (SMOTE) and adjusted class weights to address class imbalance.
- **Feature Selection:** Removed highly correlated features based on Info. gain and applied PCA for dimensionality reduction.

Modeling:

- **Supervised Models:** Logistic Regression, KNN, and SVM were used to classify normal vs. anomalous activity.
- **Ensemble Model:** A hybrid of KNN and Random Forest was employed to improve detection accuracy.
- **Unsupervised Models:** DBScan and Hierarchical Clustering were used to identify outliers and group abnormal user behavior.

Evaluation:
Model performance was evaluated using accuracy, precision, recall, and F1-score, with cross-validation to ensure robustness.

Related Works

Mylrea, M., Reed, S., & Sheldon, F. (2018). Insider threat cybersecurity framework webtool & methodology: Defending against complex cyber-physical threats. IEEE Security and Privacy Workshops (SPW).

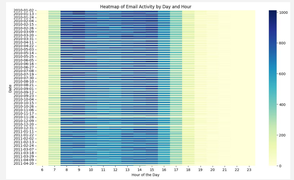
Georgiadou, A., Mouzakitis, S., & Askounis, D. (2021). Detecting insider threat via a cyber-security culture framework. Journal of Computer Information Systems, 62(4), 706-716.

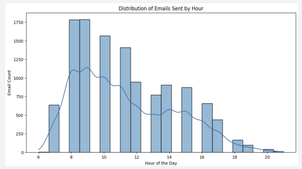
Federal Bureau of Investigation. (2020). Making Prevention a Reality: Identifying, Assessing, and Managing the Threat of Targeted Attacks. https://www.fbi.gov/file-repository/making-prevention-areality.pdf

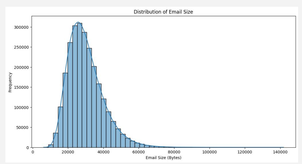
Exploratory Data Analysis

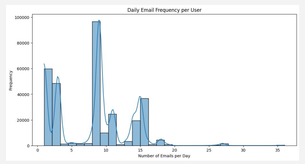
- **Data Distribution:** Visualized feature distributions (e.g., email size, number of recipients) using histograms and boxplots to identify skewness and outliers.
- **Correlation Analysis:** Generated heatmaps to identify correlations between features, helping to eliminate redundant variables and prevent multicollinearity.
- **Temporal Patterns:** Plotted email activity by time of day and day of the week to uncover trends (e.g., unusual activity during non-business hours).
- **Anomaly Detection:** Visualized user behavior using scatter plots and pairwise plots to spot unusual patterns in email interactions (e.g., high email volume or large attachments).

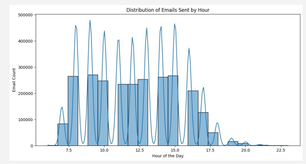
Unusual Behavior: Identified spikes in activity, particularly in off-hours or external communications, as potential indicators of insider threats.
Feature Relationships: Found strong correlations between email frequency, attachment size, and potential anomalies, guiding feature selection for modeling.

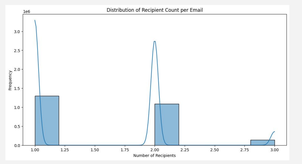




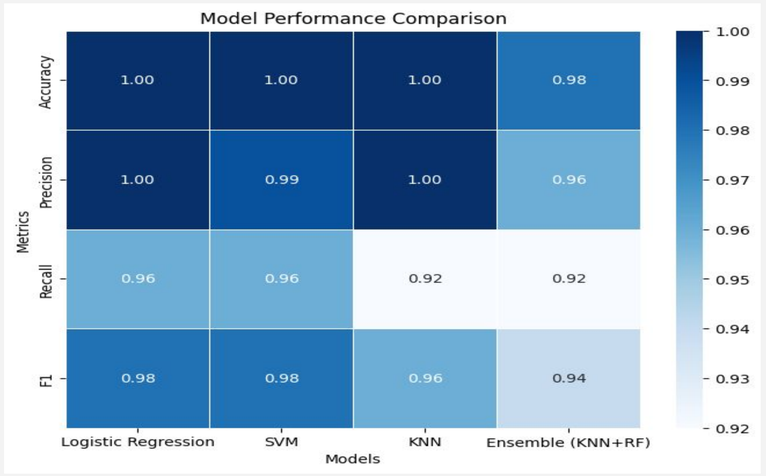




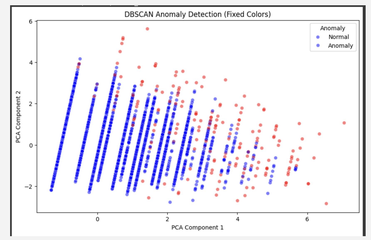


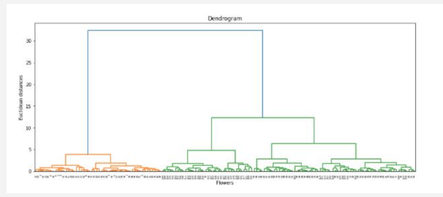


Supervised Models



Unsupervised - Hierarchical Clustering & DBScan





Conclusion

Our study demonstrates the strengths and limitations of supervised and unsupervised learning for insider threat detection. Supervised models like Logistic Regression provided high accuracy but were limited by reliance on labeled data and challenges with overfitting and imbalanced datasets. Unsupervised methods, such as DBSCAN and Hierarchical Clustering, effectively identified anomalies without prior labels and excelled in handling complex patterns, though they required careful parameter tuning.

Future Works

These models will be pipelined in the application where one can upload the logs data and check if they are prone to insider threats. This application will also give you suggestions based on which hour it found the anomaly making it easy to track down the insider person involved in the act.

Results

	Logistic Regression	SVM	KNN	Ensemble (KNN+RF)	DBSCAN	H. Clustering
Accuracy	1	1	1	0.98	0.95	0.89
Precision	1	0.99	1	0.96	0.94	0.91
Recall	0.96	0.96	0.92	0.92	0.99	0.90
F1	0.98	0.98	0.96	0.94	0.96	0.90