Network Intrusion Detection using UNSW-NB15 Dataset

By-

Avantika Deshmukh
Daniel Cersosimo
Harshini Akuleti
Himachand Pasupuleti
Purna Jadhav

Introduction

Importance of network security in today's digital landscape:

- In today's interconnected digital landscape, evolving cyber threats pose significant risks to organizations and individuals, making the security of computer networks paramount.
- Network security breaches can result in severe consequences, including data breaches, financial losses, and reputational damage for businesses.

Objectives:

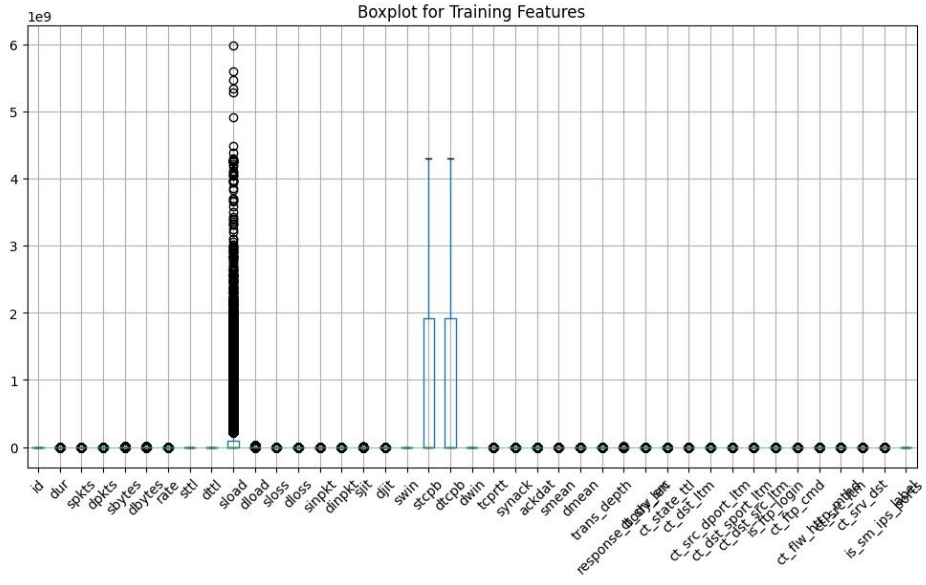
- Conduct comprehensive data exploration, preprocessing on the UNSW-NB15 dataset to utilize and assess multiple models for network intrusion detection.

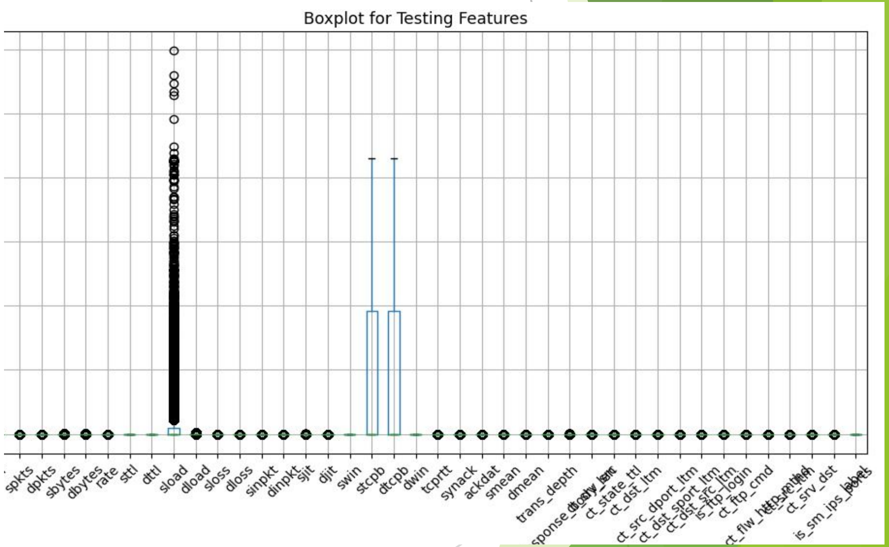
Data Acquisition and Preprocessing

- Data Acquisition: Secure the UNSW-NB15 dataset

- Preprocessing:
  - Identify nulls (none in this dataset)
  - Assess statistical distributions and properties for outliers and noise
  - Handle discrepancies in the categorical features
  - Transforming categorical data to numerical data using one-hot encoding

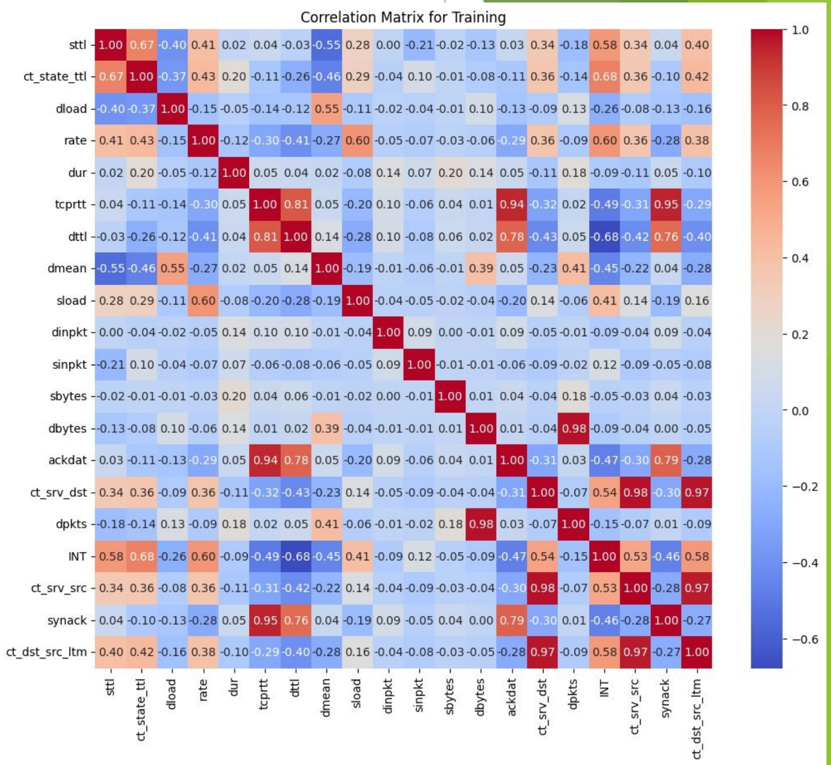
Data Acquisition and Preprocessing





Exploratory Data Analysis

- Performed Correlation Analysis
  - variables with very high positive correlation
    - ct_state_ttl and sttl
    - ackdat and smean
  - variables with very high negative correlation
    - dmean and sttl



Correlation Matrix for Training

	sttl	ct_state_ttl	dload	rate	dur	tcprtt	dttl	dmean	sload	dinpkt	sinpkt	sbytes	dbytes	ackdat	ct_srv_dst	dpkts	INT	ct_srv_src	synack	ct_dst_src_ltm
sttl	1.00	0.67	-0.40	0.41	0.02	0.04	-0.03	-0.55	0.28	0.00	-0.21	-0.02	-0.13	0.03	0.34	-0.18	0.58	0.34	0.04	0.40
ct_state_ttl	0.67	1.00	-0.37	0.43	0.20	-0.11	-0.26	-0.46	0.29	-0.04	0.10	-0.01	-0.08	-0.11	0.36	-0.14	0.68	0.36	-0.10	0.42
dload	-0.40	-0.37	1.00	-0.15	-0.05	-0.14	-0.12	0.55	-0.11	-0.02	-0.04	-0.01	0.10	-0.13	-0.09	0.13	-0.26	-0.08	-0.13	-0.16
rate	0.41	0.43	-0.15	1.00	-0.12	-0.30	-0.41	-0.27	0.60	-0.05	-0.07	-0.03	-0.06	-0.29	0.36	-0.09	0.60	0.36	-0.28	0.38
dur	0.02	0.20	-0.05	-0.12	1.00	0.05	0.04	0.02	-0.08	0.14	0.07	0.20	0.14	0.05	-0.11	0.18	-0.09	-0.11	0.05	-0.10
tcprtt	0.04	-0.11	-0.14	-0.30	0.05	1.00	0.81	0.05	-0.20	0.10	-0.06	0.04	0.01	0.94	-0.32	0.02	-0.49	-0.31	0.95	-0.29
dttl	-0.03	-0.26	-0.12	-0.41	0.04	0.81	1.00	0.14	-0.28	0.10	-0.08	0.06	0.02	0.78	-0.43	0.05	-0.68	-0.42	0.76	-0.40
dmean	-0.55	-0.46	0.55	-0.27	0.02	0.05	0.14	1.00	-0.19	-0.01	-0.06	-0.01	0.39	0.05	-0.23	0.41	-0.45	-0.22	0.04	-0.28
sload	0.28	0.29	-0.11	0.60	-0.08	-0.20	-0.28	-0.19	1.00	-0.04	-0.05	-0.02	-0.04	-0.20	0.14	-0.06	0.41	0.14	-0.19	0.16
dinpkt	0.00	-0.04	-0.02	-0.05	0.14	0.10	0.10	-0.01	-0.04	1.00	0.09	0.00	-0.01	0.09	-0.05	-0.01	-0.09	-0.04	0.09	-0.04
sinpkt	-0.21	0.10	-0.04	-0.07	0.07	-0.06	-0.08	-0.06	-0.05	0.09	1.00	-0.01	-0.01	-0.06	-0.09	-0.02	0.12	-0.09	-0.05	-0.08
sbytes	-0.02	-0.01	-0.01	-0.03	0.20	0.04	0.06	-0.01	-0.02	0.00	-0.01	1.00	0.01	0.04	-0.04	0.18	-0.05	-0.03	0.04	-0.03
dbytes	-0.13	-0.08	0.10	-0.06	0.14	0.01	0.02	0.39	-0.04	-0.01	-0.01	0.01	1.00	0.01	-0.04	0.98	-0.09	-0.04	0.00	-0.05
ackdat	0.03	-0.11	-0.13	-0.29	0.05	0.94	0.78	0.05	-0.20	0.09	-0.06	0.04	0.01	1.00	-0.31	0.03	-0.47	-0.30	0.79	-0.28
ct_srv_dst	0.34	0.36	-0.09	0.36	-0.11	-0.32	-0.43	-0.23	0.14	-0.05	-0.09	-0.04	-0.04	-0.31	1.00	-0.07	0.54	0.98	-0.30	0.97
dpkts	-0.18	-0.14	0.13	-0.09	0.18	0.02	0.05	0.41	-0.06	-0.01	-0.02	0.18	0.98	0.03	-0.07	1.00	-0.15	-0.07	0.01	-0.09
INT	0.58	0.68	-0.26	0.60	-0.09	-0.49	-0.68	-0.45	0.41	-0.09	0.12	-0.05	-0.09	-0.47	0.54	-0.15	1.00	0.53	-0.46	0.58
ct_srv_src	0.34	0.36	-0.08	0.36	-0.11	-0.31	-0.42	-0.22	0.14	-0.04	-0.09	-0.03	-0.04	-0.30	0.98	-0.07	0.53	1.00	-0.28	0.97
synack	0.04	-0.10	-0.13	-0.28	0.05	0.95	0.76	0.04	-0.19	0.09	-0.05	0.04	0.00	0.79	-0.30	0.01	-0.46	-0.28	1.00	-0.27
ct_dst_src_ltm	0.40	0.42	-0.16	0.38	-0.10	-0.29	-0.40	-0.28	0.16	-0.04	-0.08	-0.03	-0.05	-0.28	0.97	-0.09	0.58	0.97	-0.27	1.00

Overview of Machine Learning Models

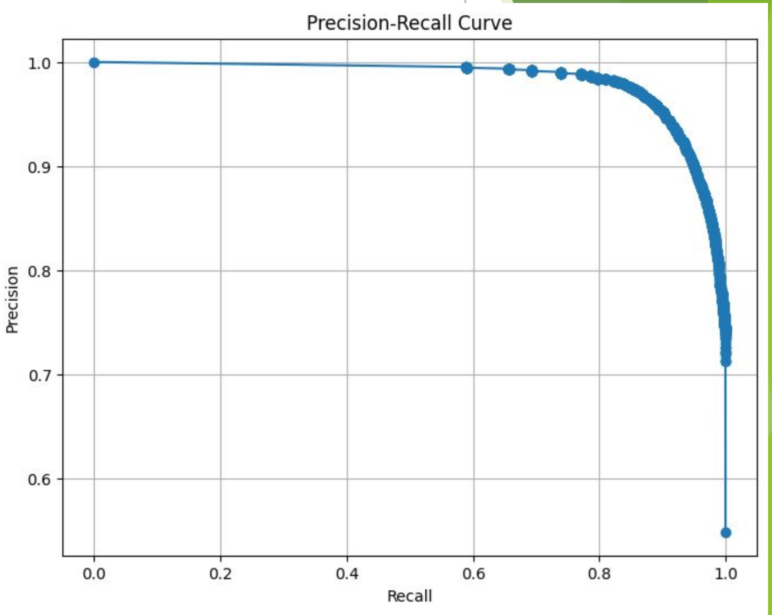
- **Machine Learning for Network Intrusion Detection**
  - Comprehensive exploration of various algorithms to identify effective classifiers.
  - Systematic approach considering each algorithm's strengths and limitations.
  - Aim for optimal performance using appropriate tools and techniques.
  - Conduct a run of each model for both high dimensional and PCA reduced feature inputs. Top accuracies from overall runs details below

- **Machine Learning Algorithms:**
  - k-Nearest Neighbors (k-NN) | top accuracy: .79 | Poor
  - Logistic Regression | top accuracy: .71 | Poor
  - SVM | top accuracy: .77 | Poor
  - Random Forest | top accuracy : .92 | Strong
  - XGBoost | top accuracy: .93 | Strong
  - LightGBM | top accuracy: .94 | Strongest

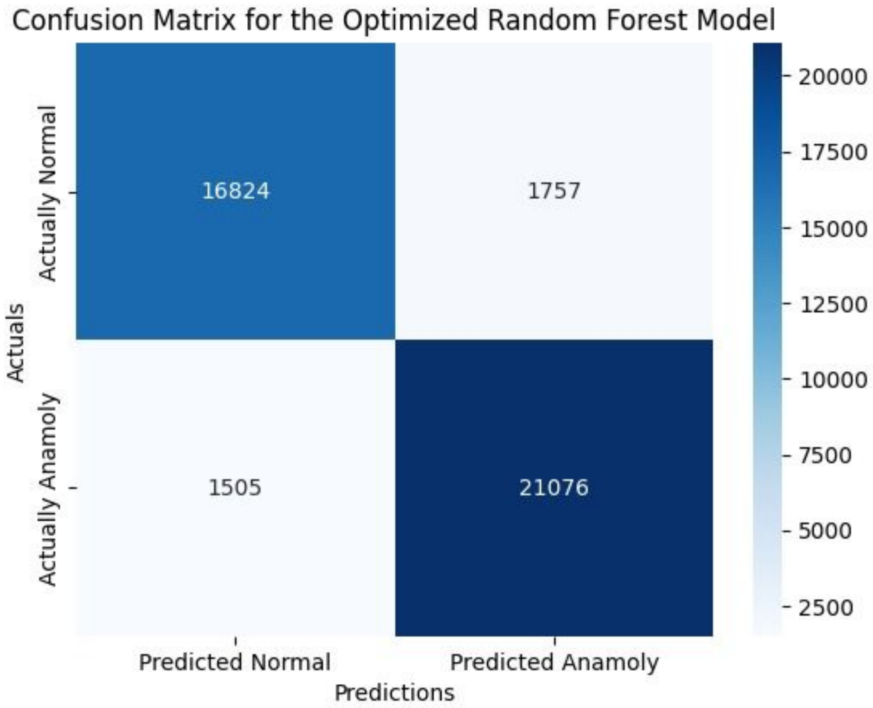
Random Forest Optimization

- Three primary methods used to optimize the Random Forest
  - Feature Selection
  - n trees hyperparameter tuning
  - Anomaly classifying threshold optimization
- Results

	precision	recall	f1-score	support
0	0.92	0.91	0.91	18581
1	0.92	0.93	0.93	22581
accuracy			0.92	41162



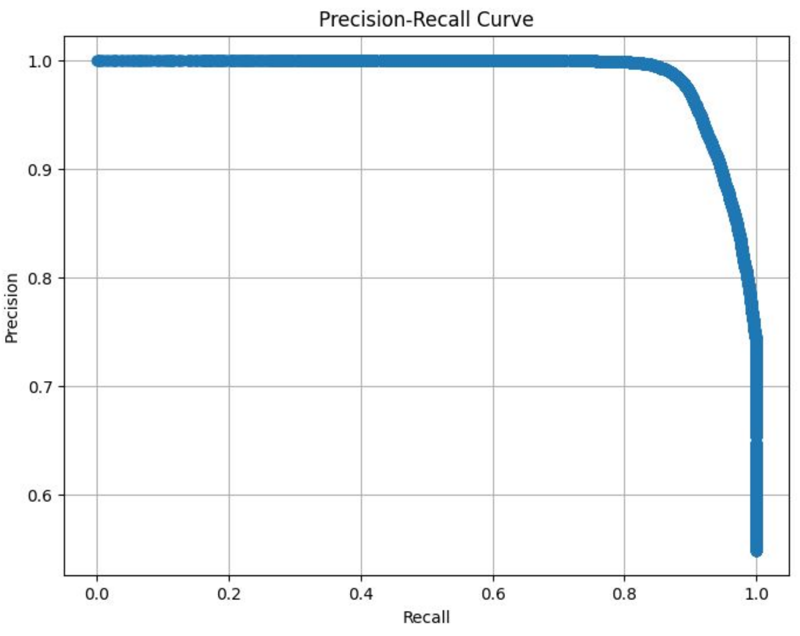
Random Forest Visualization



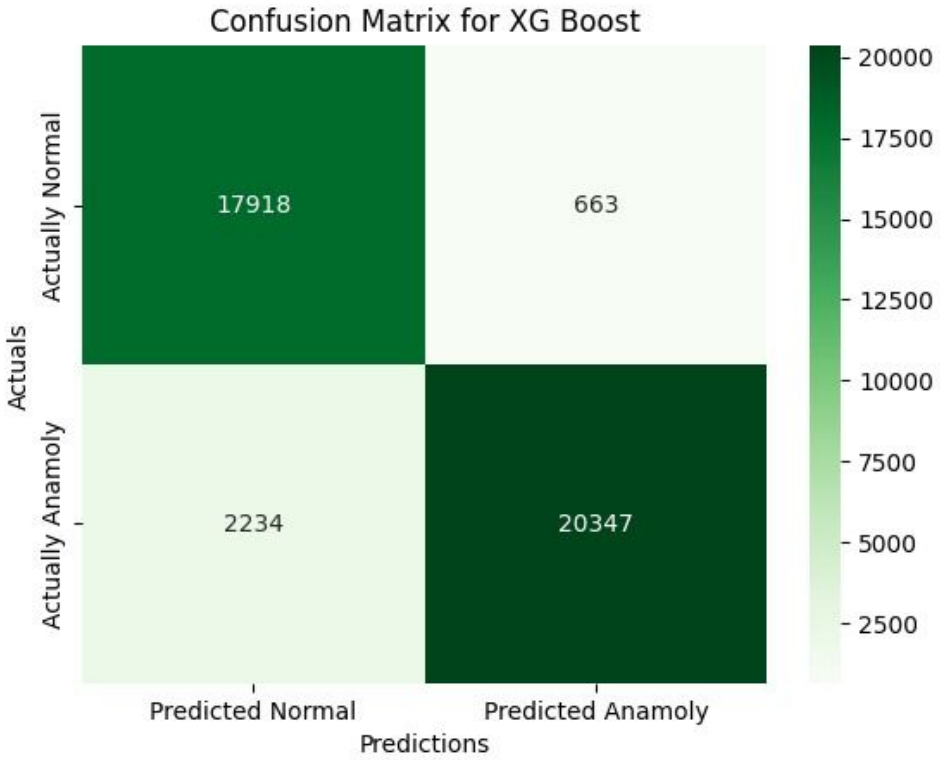
XG Boost Optimization

- Optimized the model via similar anomaly classifying threshold optimization
- Results

	precision	recall	f1-score	support
0	0.89	0.96	0.93	18581
1	0.97	0.90	0.93	22581
accuracy			0.93	41162



XGBoost Visualization

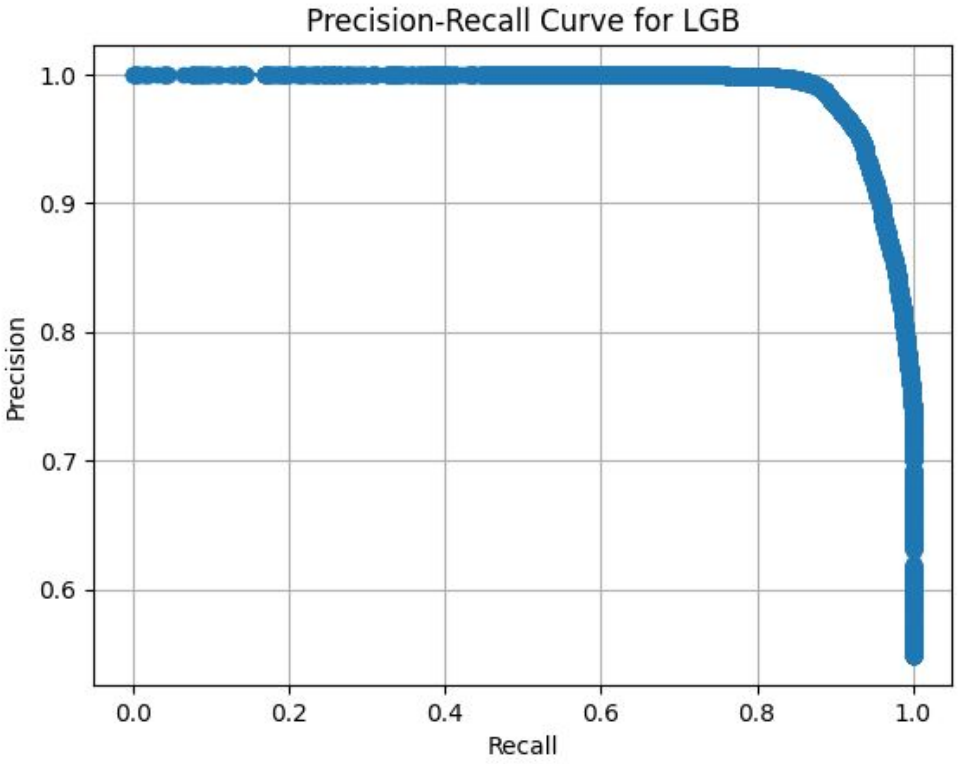


LightGBM Optimization

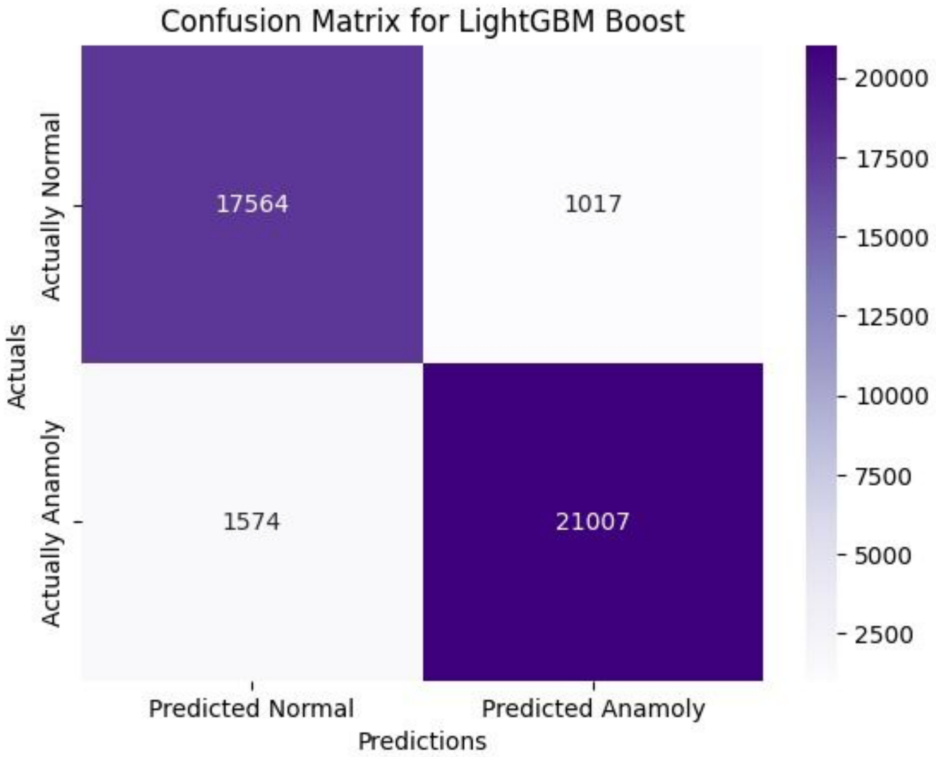
- Optimized the model via similar anomaly classifying threshold optimization
- Results

	precision	recall	f1-score	support
0	0.92	0.95	0.93	18581
1	0.95	0.93	0.94	22581
accuracy			0.94	41162

- This is our top performing model



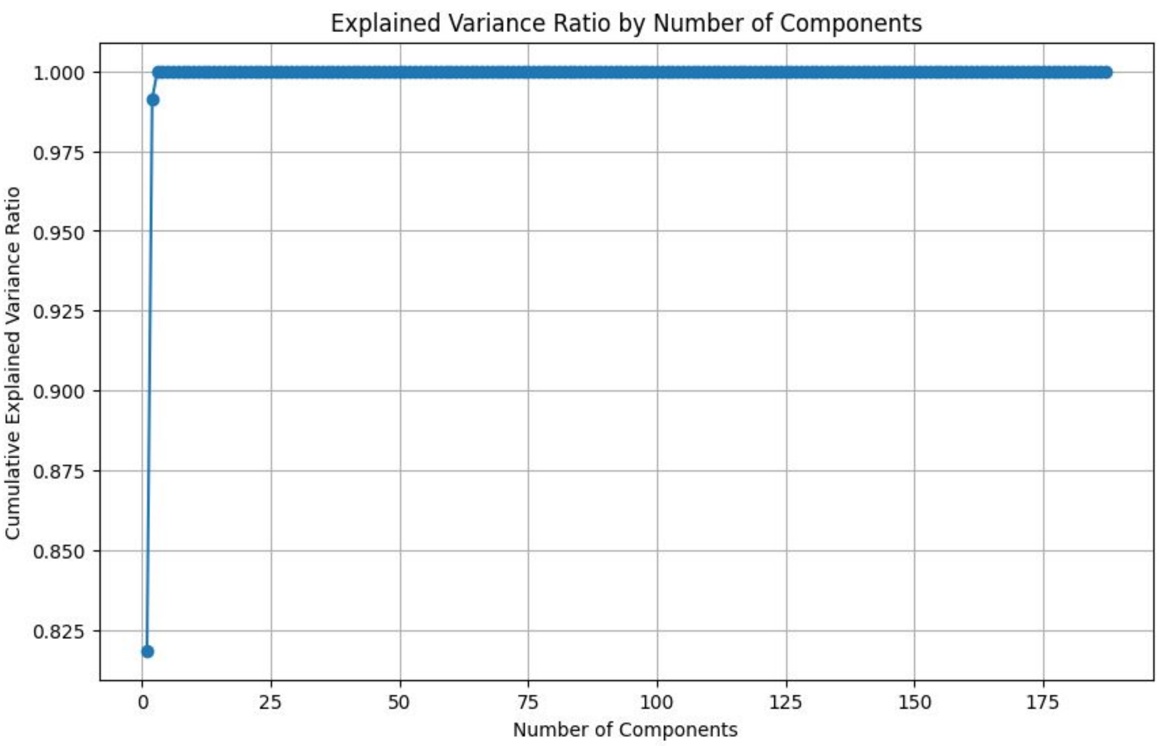
LightGBM Visualization



Principal Component Analysis

- Performance Comparison: Conducted a side-by-side analysis of models trained with original vs. PCA-transformed features.

- Sought to consider PCA to potentially enhance viability of non decision tree based models such as k-NN and even be able to attempt SVM
- Utilized the elbow method to determine the optimal amount of principal components as 3
- Gridsearch employed for k-NN however improvement was minimal and no PCA input models saw strong performance



Conclusion

- Identified three robust models with efficient accuracy, recall, precision, and F1-score metrics, showcasing their effectiveness in anomaly detection.

- LightGBM model outperformed others marginally followed by XGBoost and Random Forest. These models could be recommended for Intrusion Detection System to support Security Information and Event Management (SIEM) systems especially for systems which operate in real time due to the combination of lightweight overhead and performance.