Hi everyone! My name is []. I will be presenting our research titled “Using Machine Learning to Identify Anomalous Activities for Data Leakage Detection”.

# Introduction & Background

Data leakage is a critical issue in cybersecurity, and it often goes unnoticed until it’s too late. These incidents often result in significant financial and reputational damage to organizations. Machine learning has emerged as a powerful tool for detecting anomalies that could signal data leakage. This study evaluates multiple machine learning approaches to determine the most effective model for detecting these abnormal activities.

# Dataset

We used a labeled dataset from Kaggle containing various aspects of user interactions with the system. The dataset contains almost 50k records and 15 columns. The features includes activity information, authentication methods, user actions, and the presence of abnormalities in user behavior. The target variable is binary, indicating whether an activity is abnormal. 31% of the activities in our dataset were labeled as abnormal.

# Approach - Procedures

Our approach involves four main steps:

First, data cleaning: we removed missing data, which accounted for 12% of the dataset.

Second, data transformation: categorical features were transformed into one-hot encoded data. The timestamps were converted into seconds after an hourly pattern abnormality was found in the exploratory data analysis.

Third, feature extraction: we used principal component analysis to reduce the dataset from 17 columns to 13 key features.

Last, for modeling, the dataset was split with 8% used for training and 20% for testing the models.

# Approach - Models

We explored various kinds of models. For unsupervised learning, we tried isolation forests. For semi-supervised learning, we used autoencoders. For supervised learning, we trained logistic regression, decision trees, random forests, support vector machines, and XGBoost. All models were tuned using 10-fold cross-validation to optimize hyperparameters. The cutoff thresholds for supervised learning models were chosen to balance the precision and recall rate.

# Evaluation

In our evaluation, we measured model performance using metrics including accuracy, precision, recall, and F1-score. Accuracy measures the overall performance of a model, but it can be misleading, particularly when dealing with imbalanced datasets, where it does not differentiate the cost of misclassifying different classes. Precision is to evaluate the quality of the positive prediction of a model. Recall measures how well a model can identify positive instances in a dataset. Since precision and recall often have an inverse relationship, we used F1 score that combines both precision and recall as our final evaluation metric.

# Results

A significant finding was that supervised learning models, particularly XGBoost which has an F1 score of 0.77, outperformed others in detecting anomalies. The threshold of the XGBoost model was determined by the golden cross, the point where precision equals recall, which is 0.5919.

# Conclusion

To conclude:

Machine learning, especially supervised methods, is effective for detecting data leakages.

Among the models tested, XGBoost delivered the best overall performance.

Future work could focus on deploying these models in real-time systems and exploring ensemble techniques to enhance detection performance.