Project Report

Microsoft Cybersecurity Incident Classification with Machine Learning

Introduction:

The rise in cyber threats has significantly increased the workload for Security Operations Centers (SOCs), which are responsible for monitoring, detecting, and responding to cybersecurity incidents. SOCs deal with a vast number of alerts daily, many of which are false positives or benign incidents that do not require immediate action. This creates a challenge in efficiently triaging and prioritizing real threats, leading to alert fatigue and slower response times.

This project aims to address these challenges by developing a machine learning model capable of classifying cybersecurity incidents into three categories: True Positive (TP), False Positive (FP), and Benign Positive (BP). By automating the classification process, the model will assist SOC analysts in identifying true threats faster, improving the overall security posture of organizations while reducing the manual effort involved in managing false alarms.

Problem Statement:

SOCs face the challenge of managing an overwhelming number of alerts daily. Most of these alerts are not critical, but manually identifying which ones require attention is time-consuming and prone to errors. This project seeks to solve this problem by creating a machine learning model that automatically classifies incidents as True Positive, False Positive, or Benign Positive. The model will reduce manual efforts, improve response times, and help SOC teams focus on actual threats.

Data Exploration:

- Data Loading: The dataset was loaded in chunks to handle its large size.
- Summary Statistics: The data was analyzed to check its structure, data types, and missing values.

- Visualizations: The distribution of key features, including the target variable (IncidentGrade), was visualized to understand class imbalances.
- Class Imbalance: The Benign Positive class is significantly overrepresented compared to the other classes.

Data Preprocessing:

- Handling Missing Data: Missing values were imputed using forward fill and mean imputation. Columns with more than 50% missing values were dropped.
- Feature Engineering: Derived timestamp-based features and removed redundant columns.
- Encoding Categorical Variables: Categorical features were converted into numerical formats using encoding technique.
- Scaling: Standardized numerical features to ensure equal contribution during model training.

Data Splitting:

Split the data into training and validation sets to evaluate model performance.

Train-Validation Split: Data was split into 80% for training and 20% for validation, while maintaining the balance between classes.

Model Selection and Training:

- **Logistic Regression:** A simple model used as a baseline for comparison.
- **Decision Tree:** A non-linear model that works well for small datasets and easy interpretability.
- **Random Forest**: An ensemble of decision trees that provides more accuracy and stability.
- **XGBoost:** A powerful algorithm that handles large datasets efficiently**.**

- ✓ **Random Forest** was the best-performing model, achieving high accuracy and macro-F1 scores.
- ✓ **XGBoost** also performed well, though slightly lower than Random Forest.

Model Evaluation and Tuning:

Evaluate the model's performance using cross-validation and optimize it using hyperparameter tuning.

Metrics Used

- **Accuracy**: Measures overall correctness.
- **Precision**: Measures how many positive predictions were correct.
- **Recall**: Measures how well the model identifies actual positives.
- **Macro-F1 Score**: A balanced metric that treats all classes equally.

Hyperparameter Tuning: RandomizedSearchCV was used to find the best settings for Random Forest and XGBoost

Key Outcomes

- **Random Forest** performed best, achieving an accuracy of 99% and a macro-F1 score of 98%.

```
Comparison Table:
              Model  Accuracy  Macro-F1 Score  Precision  Recall
Logistic Regression      0.91            0.76       0.80    0.73
      Decision Tree      0.97            0.91       0.90    0.92
      Random Forest      0.99            0.98       0.98    0.98
            XGBoost      0.99            0.97       0.97    0.97

Best Model Based on Macro-F1 Score:
Model             Random Forest
Accuracy                   0.99
Macro-F1 Score             0.98
Precision                  0.98
Recall                     0.98
Name: 2, dtype: object
```

Evaluation on Test Set:

- Test the final model on unseen data to ensure it generalizes well.
- The Random Forest model was evaluated on the test set, achieving high precision, recall, and macro-F1 scores.

Classification Report (Test Data)

The classification report below provides detailed performance metrics for each category on the test dataset

```
Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.75      0.94      0.83     24124
           1       0.88      0.83      0.86     21252
           2       0.99      0.98      0.99    303765

    accuracy                           0.97    349141
   macro avg       0.87      0.92      0.89    349141
weighted avg       0.97      0.97      0.97    349141


Macro-F1 Score: 0.89
Macro Precision: 0.87
Macro Recall: 0.92

Confusion Matrix on Test Data:
[[ 22582   1177    365]
 [  2342  17701   1209]
 [  5373   1174 297218]]
```

DATA SET LINK: DATA