Arghyajit Debnath

Member of AI/ML Intern of Team 74

1. Introduction

This project involves building a regression model to predict financial losses caused by global cybersecurity threats from 2015 to 2024. The model uses machine learning techniques to learn patterns from historical data.

2. Dataset Details

- Source: Internal dataset (cleaned by Sudip)
- File: Global_Cybersecurity_Threats_2015-2024.csv
- Time Range: 2015 to 2024
- Key Columns: Attack Type, Financial Loss (in Million $), Country, Sector

3. Tools & Libraries

- Language: Python
- IDE: Google Colab
- Libraries: pandas, scikit-learn, matplotlib, seaborn

4. Methodology

Data Preprocessing:

- Loaded dataset using pandas.
- Label encoded categorical features using LabelEncoder.

Anomaly Detection:

- Used IsolationForest (contamination=0.05).
- Detected 150 anomalies out of 3000+ entries.

Regression Modeling:

- Used RandomForestRegressor.

- Target variable: Financial Loss (in Million $)

- Train-test split: 80/20

- Model trained and predictions made on test data.

5. Evaluation

Evaluation Metrics:

- Mean Squared Error (MSE): 860.72

- R^2 Score: -0.064 (indicating poor fit; worse than mean prediction)

6. Interpretation & Suggestions

- Try OneHotEncoding instead of LabelEncoding.

- Remove anomalies before training regression models.

- Consider feature engineering for better representation.

- Use advanced models like XGBoost.

- Apply feature scaling where appropriate.

7. Conclusion

Despite the poor R^2 score, this project successfully implemented a complete ML pipeline for financial loss prediction using cybersecurity data. Improvements in preprocessing and model selection can enhance predictive performance.

8. Appendix

Sample Output:

Anomaly counts:

1 2850

-1 150

Random Forest Regression Results:

MSE: 860.72

R^2 Score: -0.064

9. Visual Output Screenshots