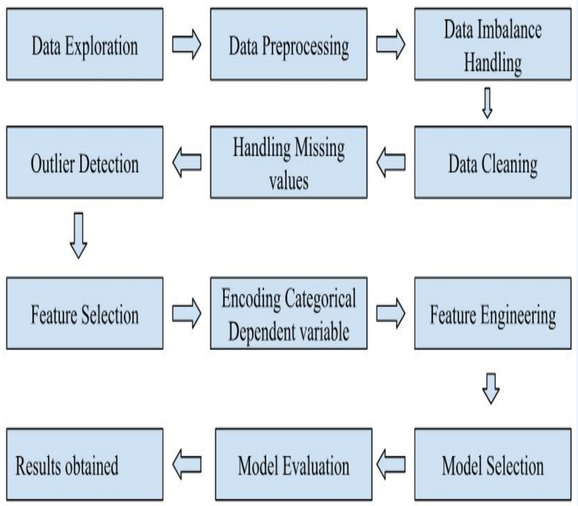
SQL INJECTION DETECTION

Asmita Deshpande, Gayathri Gurram, Jahnavi Reddy Reddy Reddy
IS Department University of Maryland Baltimore County

ABSTRACT

Cybersecurity is becoming increasingly important as more and more sensitive information is being stored and transmitted online.SQL Injections are a common way for cyber attackers to gain access to systems steal data, and launch attacks.Creation of an efficient algorithm designed to pinpoint IP addresses that carry out SQL injection attacks, with a particular focus on those executing Union and Blind Queries is necessary. The algorithm is trained and tested on a dataset comprised of NetFlow data, indicative of malicious activity, and gathered via the DOROTHEA framework. This dataset is split into a training component and a testing component to ensure the model's robustness in different settings. Initial data analysis indicated a numerical dataset with a binary outcome variable, guiding the choice of classification models such as logistic regression, random forest, and XGBoost for the task at hand. The project's intention is to refine these models to accurately generalize across various network data, conforming to the NetFlow V5 protocol.

METHODOLOGY



In exploring a large dataset for SQL injection detection, we discerned a limited number of influential features amidst a noisy background, leading us to enhance our feature engineering approach. We standardized numerical features and encoded categorical variables, then conducted feature importance analysis to refine the predictors for our model. This process was essential for understanding each feature's impact on the target variable and optimizing the dataset for better model training.

Business Understanding of the Problem

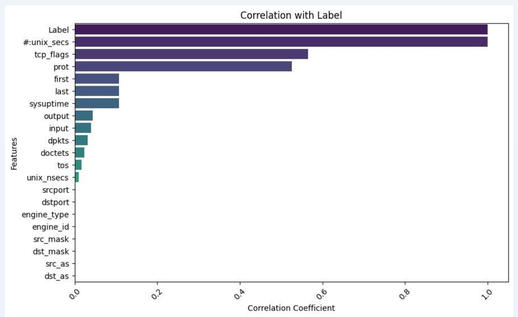
From business perspective protecting sensitive data especially for organizations handling sensitive consumer data becomes paramount to maintaining trust of the customers while in compliance with the data protection regulations which intern helps with elimination in disruption of daily business operations which increases productivity, reducing financial loss.

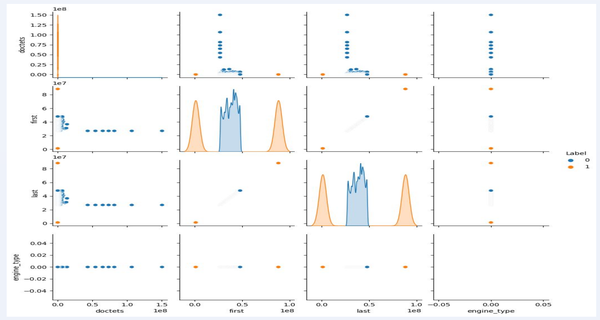
Data understanding and preparation

The dataset we have chosen is from a research data which is divided into 2 datasets : one for training(D1) and the other one for testing purpose (D2). Since it is a mid-sized data and has non-null records the main focus has been on improving the quality of the data rather than on data cleaning preparing it to better fit the model reducing any bias possible. Part of feature engineering, feature importance analysis has been performed to understand the correlativity of each feature with respect to the target feature. Only the top 15 strong attributes have been considered to reduce dimensionality and any overfitting possible. Going forward with the selection of the model that best fits this data has been based on the linearity of the features with respect of the target feature which is indeed helpful in selection of the model.

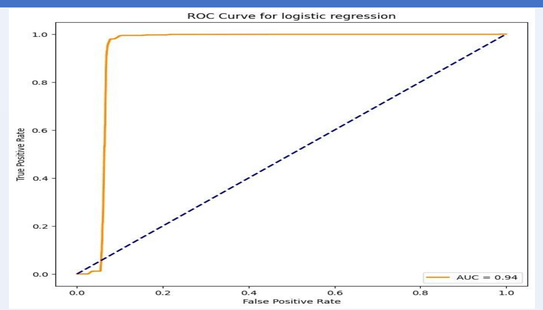
Algorithm - Supervised Learning

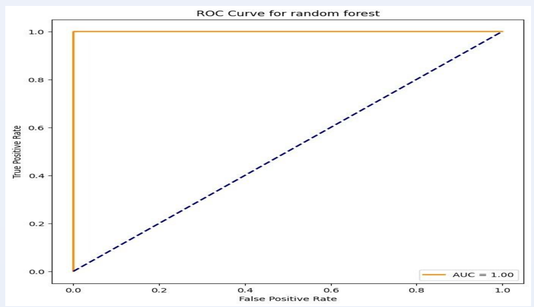
Since the data is pretty clear along with the target attribute which is a binary variable supervised algorithms like classification are utilized in predicting the benign and malignant ones. For this problem logistic regression with L1 regularization and random forest algorithms are utilized.These algorithms are chosen based on linear relationship between features and target while keeping the size of the data into consideration. We have utilized c-statistic and accuracy score as a measure to assess the overall performance of the model. Logistic regression with L1 regularization also known as Lasso regression attributes to a simpler and more interpretable model which helps with overfitting of the data which encourages the model to the generalize new, unseen data. Random Forest algorithm is also tested to make sure the model is not overfitting. SQL injection often involves complex, non-linear relationships between input features and likelihood of the attack. Random Forests are quite robust and capable to capture such non-linearities making them suitable for tasks where decision boundary is intricate than basic decision trees.





RESULTS





CONCLUSION

The graphs obtained illustrate the data analytics process in identifying SQL injection threats. They show the relationships and distributions of network data features with respect to SQL injection labels. Scatter plots indicate a clear distinction between normal and anomalous data points, while histograms exhibit the data distribution for specific features. The bar chart highlights the correlation strength between each feature and the SQL injection label, guiding feature selection for the predictive model. Collectively, these visualizations underscore the effectiveness of data preprocessing and feature engineering in enhancing model accuracy for cybersecurity threat detection.The project effectively crafted an algorithm to identify SQL injection attacks, particularly Union and Blind Queries, using a comprehensive NetFlow data collection. Through initial data analysis and meticulous feature engineering like normalization and encoding, we enhanced the precision of predictive models including logistic regression, random forest, and XGBoost. This integrated methodology of data preparation and strategic feature determination laid a solid groundwork for advanced cybersecurity analysis.

FUTURE WORK

Since the data is mid sized so there can be modeled using machine learning and when highly advanced models like deep learning models(CNN,RNN etc) used leads to overfitting. So when the data sized is increased then there's a scope of utilization of advanced models for more effective detection of complex attacks.

Another drawback of the data is lack of timestamp which can make it the result a little ambiguous. Adding timestamp attributes greatly to overall classification process while utilizing geospatial models which helps in identification of specific location of attack.

REFERENCES

Ignacio Crespo, & Adrián Campazas. (2022). SQL Injection Attack Netflow [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6907252

Aabc/IPT-netflow, 2022. Ipt-netflow: netflow iptables module for linux kernel. https://github.com/aabc/ipt-netflow (accessed July 28, 2022).

A. Campazas-Vega, I.S. Crespo-Martínez, Á.M. Guerrero-Higueras, C. Álvarez-Aparicio, V. Matellán

Analysis of netflow features' importance in malicious network traffic detection

Computational Intelligence in Security for Information Systems Conference, Springer (2021), pp. 52-61

ACKNOWLEDGEMENT

https://drive.google.com/file/d/1EGOgQ1vP1Bf4tc6q2V5uffrW2NyzsuA8/view?usp=drive_link
https://drive.google.com/file/d/1EMI1nWXmKCVSu8VFRIiQ7seoPb0Rv447/view?usp=drive_link