IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Detecting Data Leaks Using SQL Injection

Jaya Srivastava, Priyal Agrahari[1], Pramiti Sirothia[2], Kartikeya Mishra[3], Om Pravin Singh[4]
Information Technology
ABES Engineering College

Abstract— SQL injection attacks are a serious security risk as they have increased in frequency and severity over time. These attacks target an application's database layer explicitly. When user input into SQL statements is not sufficiently screened for string literal escape characters, a vulnerability occurs. In essence, attackers use this flaw to change the SQL queries that the database runs.

Further increasing the danger is weak typing by the user, which might result in the unintentional execution of malicious code. This implies that it will be simpler for attackers to carry out malicious orders if the program does not enforce tight data types for user input.

SQL injection is still a widely used attack method in application layer assaults nowadays. The aforementioned project's goal is to put SQL injection prevention mechanisms in place as a reaction to this persistent danger. By making sure that inserted queries do not jeopardize the integrity of the system, the project seeks to safeguard the database. This entails putting in place strong input validation and filtering procedures to counteract the risk of malicious SQL injection attacks and improve the software system's overall security.

I. INTRODUCTION

SQL injection stands out as a particularly serious threat to the security and stability of your database. In the recent times, most dangerous attack on the security of the database is done through SQL injection . It is a method of attack of any web application which can lead to loss of data and theft .Malicious code is injected into SQL statements, leading to potentially data leakage and manipulation. Imagine a sneaky thief who loves to steal passwords. They've gotten really good at this because many people leave their doors wide open! This trick is so popular with these thieves because it's easy to do and works a lot. The thief sneaks in by pretending to be someone else, because no one checks their ID! This is what hackers do with SQL injection - they trick the system into thinking they are someone they're not.In this modern world of interconnectivity where data reigns supreme for numerous applications and services, security plus integrity for databases can be seen as crucial elements that deserve attention. Cyber threats especially SQL injection attacks present continuous and great risks towards confidentiality and reliability of sensitive information stored in the database .Our aim through this initiative is to come up with smart and strong techniques for stopping data leakage via SQL injections, an attack that underscores any such effort if successful.What we want is not a simple security measure, but a complex one that uses advanced algorithms to detect intrusions in real time through machine learning capabilities. his strategy goes beyond data protection; it will contribute to shaping global cyber defense frameworks. We aim at not only securing individual web applications but also fostering a secure digital business ecosystem through pre-emptive defense mechanism of detecting data leaks via SQL injection — one of the most common attack vectors against such threats that can hamper technology under all possible ways supporting society nowadays..

II. LITERATURE REVIEW

OWASP TOP Project:

The Open Web Application Security Project (OWASP) has been a leading advocate for promoting best practices in securing web applications. Imagine a group of security experts called OWASP who are like firefighters for the internet. They teach people how to protect their websites from bad guys. One of their most popular resources is a guide all about a sneaky trick hackers use called "SQL injection". This guide is like a superhero training manual for web defenders. It teaches them how to recognize the latest attacks, how to stop them, and what the best practices are to keep websites safe. This guide is a big deal because it helps everyone learn from each other and keeps the internet a safer place

Protection of Personal Data in Information Systems:

Our approach is detailed by Bojken et al. (2013) who underscore the importance of safeguarding personal information stored in information systems [2] as part of establishing robust security measures to protect sensitive data: an issue that should address vulnerability issues like SQL injection.

Web Security Vulnerabilities from the Programming Language Perspective:

The study by Se-ixas et al. (2009) looks at online security issue-s from the viewpoint of programming languages [3]. This give-s a new way to look at things. Their study looks at how online se-curity and programming languages connect. This helps us unde-rstand what causes issues like SQL inje-ction. This adds to the talk about web app security by thinking about the-programming language.

SQLIA: Detection and Prevention Techniques - A Survey:

Our approach is detailed by Bojken et al. (2013) who underscore the importance of safeguarding personal information stored in information systems [2] as part of establishing robust security measures to protect sensitive data: an issue that should address vulnerability issues like SQL injection.

Parse Tree Validation to Prevent SQL Injection Attacks:

A novel method of preventing SQL injection attacks is suggested by Buehrer et al. (2005) and involves the use of parse tree validation [5]. The technique seeks to detect and stop malicious injections by evaluating the parse tree structure of

SQL queries. This study offers a fresh viewpoint on the variety of methods available to protect against SQL injection, emphasizing the role that parsing algorithms play in bolstering online application security

Web Application Security Assessment by Fault Injection and Behaviour Monitoring:
Huang and others (2003) talk about a way to che-ck how safe online apps are [6]. The-y use behaviour monitoring and fault injection to copy re-al attack situations. This helps find problems, like SQL inje-ction ones, before the-y happen. The study shows that watching how things act is a good extra ste-p on top of trying to stop problems before the-y start..

III. PROPOSED MODEL

The starting point of this process is to collect relevant data where critical elements are identified for further investigation. Once it has been gathered, the information is formatted and organized so that it fits into the desired structure. Then, this prepared information is divided into training and testing sets. Training data helps in training a model by acting as an input for various algorithms. In order to assess how well the model can generalize and forecast on previously untested data, its accuracy is compared against independent data that provides such knowledge. To develop the overall system, we implement key modules which include: data collection; data preparation; model selection; model training; model assessment and predictions among others. Each module performs a specific function. This systematic approach ensures efficiency throughout the workflow of this methodical methodology applied in effective analysis of data, training models as well as predictive abilities.

Python is at the top among programming languages used in machine learning and data analysis. This is because of the many libraries and frameworks it has that are created for such applications. Scikit-learn, NumPy and others are important libraries in making python lead this area hence maintaining its position. On the other hand, inbuilt modules found in python enable faster execution when carrying out required steps of data collection and preparation for analysis.

For instance, pandas is a versatile preprocessing tool within Python which allows users to deal with, clean up and organize data systematically thus setting stage for more analysis.
Another crucial element, NumPy, provides strong support for numerical operations on matrices and arrays. This feature improves the speed of calculations and manipulations involving huge datasets and is essential for managing the mathematical complexities included in machine learning algorithms.

Furthermore, because it provides practitioners with a wide range of tools for feature selection, data preparation, and model validation, the scikit-learn package is crucial to the workflow of data science. Scikit-learn makes it simple for users to prepare data, locate relevant features, and assess the performance of machine learning models, significantly enhancing the overall efficacy of the analytical process.In summary, Python's strong ecosystem of libraries and frameworks highlights the language's supremacy in machine learning and data analysis. In this ecosystem, scikit-learn, Pandas, and NumPy stand out as essential resources that when combined enable practitioners to handle the complex terrain of data manipulation, preprocessing, and model assessment with unmatched variety and efficiency.

IV. METHODOLOGY

Machine Learning

Data - and pattern recognition more generally - is also what machine learning, the part of AI which is interested in creating models and algorithms that can take input and learn from what they've been given to make predictions or decisions, is all about. Or put somewhat less formally: it is, essentially, a way you can program something that will learn from its mistakes. There are unsupervised learning approaches that describe algorithms that can determine patterns and interpretations within the data without any labelling at all. Another approach is supervised learning, the focus of my project, which is based on learning a model from labeled data. Unsupervised learning uses only unlabelled data for the process of training. In contrast, semi-supervised learning is a special case where a model learns from unlabelled and labelled datasets. Reinforcement learning is about the training of a model to take a sequence of decisions, and learn from the feedback it gets by interacting with the environment. A subfield of machine learning, deep learning utilises multi-layered neural networks. This type of learning is the most famous type due to its popularity in voice and image recognition, among other areas. Banking, healthcare, computer vision, natural language processing, and automotive are among the few of the hundreds of areas in which machine learning can be used. However, much like every other model, we need a good volume and quality of training data to proceed.

V. IMPLEMENTATION

Our data analytics application's backend has been carefully developed with Python Our utilization of the Python library Pandas has allowed for the implementation of strong analytics for sales data. We used very common Python libraries like Scikit-Learn for the preprocessing of the data, feature selection, and the performance evaluation of the model. We also used the pre-built library implementations of different machine learning algorithm. We will examine the accuracy of all the algorithm and will do a comparative study.

A. Decision tree

The decision tree is a simple machine learning method that can be used for both classification and regression. This algorithm is constructed as a tree, in which internal nodes are given as the features or attributes, branches to the decision rules with the leaf being the outcome or target one. Decision trees offer simple interpretations, therefore, they are helpful for researchers and practitioners at different levels of machine learning experience. Decision trees A decision tree can be thought of as a tree structure with each node representing a feature and each split representing a decision rule and each leaf representing a prediction. This is represented graphically. A decision tree is created by partitioning your data by features, using the features themselves as the decision rule. At each node the algorithm selects the feature that best separates the data into groups with target values. We have used the common Python libraries like Scikit-Learn for the data preprocessing, feature selection and the algorithm performance evaluation. Furthermore, we have also used the pre-built library implementation of the algorithms to be used for our classification problem. We are going to evaluate the accuracy of every algorithm and make a comparison.

B. Support Vector Classifier

A support vector machine is a collection of algorithms that are used as supervised learning methods on classification and regression problems. The Support Vector Machine classifier tries to find a hyper-plane that divides the two categories reasonably well and leaves as big a gap as possible between the two categories. This is also known as a maximal margin classifier.

SVM is noteworthy for its capacity to handle non-linear problems, which is achieved by utilizing kernel functions. For example, the popular RBF (radial basis function) kernel makes it easier to transfer data points into a higher-dimensional space so that they may be separated linearly. Following this conversion, SVM will search the space for the optimal hyperplane in order to correctly classify the data points in a relevant manner. It is one of the most popular algorithms in machine learning which can also handle non-linear problem domains.

C. Logistic Regression

A predictive model is created using sequential regression, a statistical modeling or machine learning technique, to predict results for a series of dependent variables. It basically involves making predictions about a sequence of events that will occur over time or in a certain order.

Key concepts of sequential regression to understand are as follows:

1. Time Dependency: Sequential regression relies heavily on the timing and sequence of data. Based on past data, the model takes into account the temporal sequence of occurrences and tries to estimate the value that will occur next.

2. Variable Dependency: In sequential regression, predictions are based on the values of the preceding observations in the series. The dependencies and correlations between the variables in the sequence are considered by the model.

3. Applications: In time series analysis, when the objective is to forecast future values in a time-ordered sequence, sequential regression is frequently used. It is used in many different fields, including natural language processing, meteorological forecasting, finance (stock pricing), and sentence prediction.

4. Recurrent Neural Networks (RNNs): For sequential regression problems, recurrent neural networks (RNNs) are widely employed in deep learning. RNNs preserve hidden states in sequential data in order to extract knowledge from earlier stages.

5. Autoregressive Models: Autoregressive models are a type of traditional statistical models for sequential regression in which the current value in the series is represented as a linear mixture of earlier values. In time series analysis, autoregressive Integrated Moving Average (ARIMA) models are one type of.

6. Difficulties: Determining the proper context window for taking into account previous observations and vanishing or inflating gradients in deep learning models are two difficulties in modelling sequential dependencies.

7. Online Learning: Sequential regression models work well in situations where fresh data is constantly being added to the model, such as in educational websites.

8. Evaluation: Measures such as Mean Squared Error (MSE) for continuous predictions or accuracy for classification tasks may be included in the assessment metrics for sequential regression tasks.

D. Sequential Regression

One popular machine learning method in the field of supervised learning is logistic regression. Its primary goal is to use a given collection of independent factors to predict the value of a categorical dependent variable. Instead of producing definite values like 0 or 1, Yes or No, or true or false, the algorithm generates outcomes in the form of probability values ranging from 0 to 1, which are then used to forecast the outcome of a categorical dependent variable. Unlike linear regression, which is used to solve regression problems, logistic regression is designed to solve classification difficulties. Logistic regression uses a "S"-shaped logistic function to predict values of 0 or 1, as opposed to fitting a regression line.

The implementation of all these machine learning algorithms is done in following steps:

- Importing essential libraries

When working with Python for data analysis and machine learning tasks, it is imperative to import essential packages to harness their functionalities. To do so, the common practice is to use the Python package manager, pip. In this specific scenario, the installation of three crucial packages—Pandas, NumPy, and Scikit-learn—is necessary to facilitate various aspects of data manipulation, numerical operations, and machine learning model development.

- Decision tree implementation
  After applying the decision tree model, we get an accuracy of 81.47% ,which is a good fitted result.
- Support vector classifier
  On applying this algorithm, we get an accuracy of 99.48% .This gives an overfitted result.
- Logistic Regression
  On applying the Logistic regression we get an accuracy of 99.3 which also an overfit.
- Sequential regression
  This algorithm gives an accuracy of 62.95% which is gives an underfit result

After testing all the algorithms ,we can conclude that the Decision tree is the best fitted algorithm for this project.

VII. CONCLUSIONS

The primary objective is to develop an intelligent and adaptive system that not only identifies and neutralizes SQL injection attempts but also proactively strengthens the overall security of web databases. Building upon the principles outlined in the OWASP TOP Project, this research aims to enhance existing methodologies by integrating advanced anomaly detection, dynamic query sanitization, and behavioral monitoring techniques. By doing so, the research seeks to address the limitations of current prevention strategies, such as adaptability to evolving attack vectors and the potential trade-off between security and system performance.

REFERENCE

[1] Visit` https://wikipedia.org/wiki/Big_data`

[2] Zhendong Su and Gary Wassermann, in 2006 published their research in SQL injection.

[3]SQL Injection Detection Methods :Qi Li, Weishi Li, Junfeng Wang, Mingyu Cheng

[4] SQL Injection Detection Using Machine Learning : Sonali Mishra

[5] Raghav Kukreja , Nitin Garg 2014. OVERVIEW OF SQL INJECTION ATTACK international journal of innovative research in technology Volume 1 Issue 5.

[6]Cheon, Eun Hong, Zhongyue Huang, and Yon Sik Lee(2013). "Preventing SQL Injection Attack Based on Machine Learning." International Journal of Advancements in Computing Technology 5.9

[7] Joshi, Anamika, and V. Geetha.(2014) "SQL Injection detection using machine learning." Control, Instrumentation, Communication and Computational Technologies (ICCICCT), International Conference on. IEEE, 2014.