Password Strength Prediction

Avinash Raval, Freya Modi, Freyal Shah, Zalak Shah

School of Engineering and Applied Science, Ahmedabad University, Ahmedabad, Gujarat 380013 Email: [avinash.r, freya.m, freyal.s, zalak.s]@ahduni.edu.in

***Abstract*—In this paper, we try to reproduce the machine learning-based models to predict the strength of a password. The objective of the project is to develop a tool that can accurately classify a password into one of three categories: weak, medium, or strong. The model is trained on a large data-set of passwords, with labels indicating their strength, and various features such as length, complexity, and usage of special characters, numbers, and capital letters. We use multiple machine learning algorithms to identify the best performing model. Models including Logistic Regression and Decision Tree Classifier are used for analysis. Thus, we even figure out the balance between accuracy and efficiency by implementation of big data.**

***Keywords*—Accuracy, Strength, Regression, Classification, Password strength**

1. INTRODUCTION

Passwords are widely used for securing online accounts and sensitive information. They are designed to protect user data from unauthorized access by requiring an authentication process to gain access to a particular account. However, the use of weak passwords can make accounts vulnerable to hacking, posing a significant risk to users' privacy and security. Weak passwords are generally easy to guess or crack using various methods, including brute force attacks or dictionary attacks. As such, there is an urgent need to develop an effective and reliable technique for measuring the strength of a password to ensure that users are not using weak passwords that can be easily guessed or cracked. The development of an efficient password strength measuring tool can assist users in generating more robust passwords that are more difficult for attackers to break. This, in turn, can help to reduce the likelihood of security breaches, thereby safeguarding users' confidential data. A reliable password strength measuring technique can also help organizations to enforce stronger password policies, ensuring that their employees use robust passwords that are less susceptible to attacks. Therefore, the development of an efficient password strength measuring tool is crucial in enhancing online security and ensuring that users' privacy and security are not compromised.

1. LITERATURE SURVEY

“A Comparative Study of Password Strength Estimation Techniques” by Julien Freudiger et al., which compares different methods for password strength estimation and identifies their strengths and weaknesses.

“A Machine Learning Approach to Password Strength Meters” by Ingmar Baumgart et al., which proposes a machine learning-based approach to password strength estimation and evaluates its performance against other methods.

“Password Strength Meters: Do They Even Work?” by Stuart Schechter et al., which evaluates the effectiveness of existing password strength meters and identifies areas for improvement.

“Using Machine Learning to Predict the Strength of Passwords” by Ravi Sahita et al., which proposes a machine learning- based approach to password strength prediction and evaluates its performance on a large dataset of passwords.

These papers provide valuable insights into password security, password strength estimation techniques, and the effectiveness of machine learning-based approaches for password strength prediction.

1. IMPLEMENTATION

The proposed model is implemented using Python and various libraries such as scikit-learn, pandas, and NumPy. The accuracy of the model is evaluated using the CLF estimator.

**Finding the dataset:** The first task involves finding the appropriate dataset from a reliable source so that the model accuracy is not at stake.

**Checking for missing values:** It is possible that data samples have a missing value (Not Applicable-NA) in the DataFrame in which they are loaded. This may happen when the data is not captured properly.

**Removing the missing values:** Missing values can lead to erroneous predictions and therefore it is very important to remove them. This involves implementation of functions which removes all the rows with missing values and hence makes the dataset error free.

**Conversion of Dataset into an array:** Since arrays are much easier to work with, the DataFrames are converted into arrays.This is done by first importing the NumPy library and using its function array() to get an array of data samples.

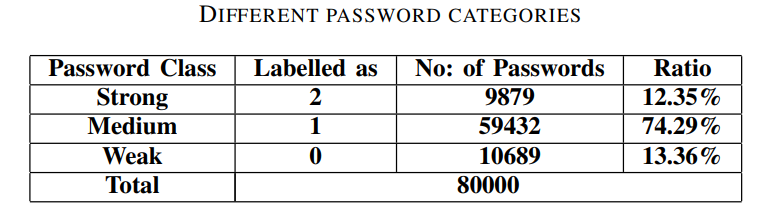
**Tokenization:** It is a process of dividing the text data into word tokens that will serve as the inputs of the model. These words are then converted into numeric data with the help of tfidf vectorizer.

**Training:** The model is a Decision Tree Classifier in which the intermediate nodes represent the features and the leaf node represents the output.The model first extracts features namely password length and complexity in terms of number of uppercase, lowercase letters, special characters and numerals in the password. It then learns from the mapping of the input features and output labels, which is provided by the training set.

**Testing:** After the model has been successfully trained, it is tested against the test data in order to check for its accuracy.

1. RESULTS

The current model for prediction is Decision Tree Classifier. The results show that our model can predict the strength of a password with an accuracy of 97%, thereby helping users create stronger passwords and improve their online security. We are still in process to implement the Logistic Regression model which uses different features such as length, complexity, and usage of special characters, numbers, and capital letters. We would like to present a detailed comparative analysis between the Regressor and Classifier as our end result. This result can either be in the form of an extension or a Web Dev application.

1. CONCLUSIONS

To conclude the discussion, it can be said that the Classifier with an accuracy rate of 97% can serve as an effective solution for predicting password strength, which is a crucial factor in ensuring online security. However, it is important to note that relying solely on one method is not advisable, as there may be limitations to its effectiveness in certain situations.

Moreover, while the decision tree algorithm exhibits the capability to process both categorical and numerical data, it falls short in accurately predicting data with minimal variability. As a result, in order to overcome this issue, the subsequent algorithm to be employed is logistic regression.

VI. REFERENCES

[1] Bansal, B. (2019, june 27). *password strength classifier dataset*. Kaggle.com[.https://www.kaggle.com/datasets/bhavikbb/password-strength-classifier-dataset](https://www.kaggle.com/datasets/bhavikbb/password-strength-classifier-dataset)

[2] *Building a Password Strength Classifier Model Using Machine Learning*. (2022, January 26). Section.io. Retrieved March 11, 2023, from [https://www.section.io/engineering-education/building-a-password-strength-classifier-model-using-machine-learning/#adding-features-and-labels](https://www.section.io/engineering-education/building-a-password-strength-classifier-model-using-machine-learning/" \l "adding-features-and-labels)

[3] AdityaBagad/Password-Strength-Analysis: Password Strength Analysis using Machine Learning Classification Algorithms. (n.d.). GitHub. Retrieved March 11, 2023, from <https://github.com/AdityaBagad/Password-Strength-Analysis>

[4] U. Farooq, “Real Time Password Strength Analysis on a Web Application Using Multiple Machine Learning Approaches,” International Journal of Engineering Research & Technology (IJERT), ISSN: 2278- 0181 Vol. 9 Issue 12, December 2020

[5] Sarkar, S., & Nandan, M. (2022). Password Strength Analysis and its Classification by Applying Machine Learning Based Techniques. 10.1109