Password Strength Prediction

Amrutha Karuturi

Abstract:

This project endeavors to create a machine learning model with the objective of predicting password strength, categorizing user-generated passwords into weak, normal, and strong classes based on semantic analysis. The study employs techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to transform password text into a machine-readable format. Logistic regression serves as the chosen classifier in this endeavor. The aim is to enhance password security through the accurate classification of password strengths.

Data:

The data consists of 100,000 rows and 2 columns contains the passwords and their corresponding strengths with 0,1,2 as the possible values that indicate weak,normal and strong respectively. The data does not contain any duplicate values and Null values.

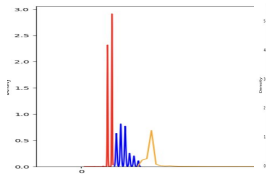
Data preprocessing:

To asses the password more clearly semantic analysis is done on the data and observed that 26 are total numeric, 1506 passwords are totally in uppercase, 97203 passwords are alpha numeric, 932 passwords have first letter in upper case,2663 passwords poses special characters.

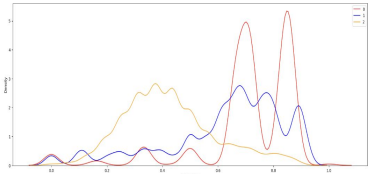
Feature Engineering: New features lowercase, uppercase, digit, special character frequency are created to indicate the fraction of the corresponding features in a given password.

Data Analysis:

Conducted descriptive statistics on all features to assess their distributions across varying strength values. The primary objective was to discern the significance of each feature in predicting password strength. A key aspect of our analysis involved examining the extent of overlap in distribution curves for different strength categories. Features demonstrating minimal overlap were identified as having higher importance in predicting password strength. 'Length' feature does not have any overlap among its distribution curves for different strengths, and lower_case_frequency has next least overlap in its distribution across the 3 strength categories, remaining features are observed to have a significantly large amount of overlaps. .This methodology provides valuable insights into the distinctiveness of feature distributions and aids in understanding their relevance to the overall predictive model.



1.Distribution of length



2. Distribution of lower case frequencies

strength	length min	length max	length mean	length median
0	1	7	6.550947	7.0
1	8	13	9.611074	9.0
2	14	220	15.953421	16.0

strength	lower_freq min	lower_freq max	lower_freq mean	lower_freq median
0	0.0	1.000	0.708050	0.714
1	0.0	0.923	0.630067	0.667
2	0.0	0.917	0.424679	0.400

Feature Transformation:

Now the data has to be processed to be machine readable. The given string is converted to a vector representation with Term Frequency-Inverse Document Frequency (TF-IDF) vectorization.

- TF is a measure of how frequently a term t, appears in a document d. We calculate the char frequencies for all the characters and all the passwords in this manner.
- IDF is a measure of how important a term is. We need the IDF value because computing just the TF alone is not sufficient to understand the importance of words.
- We can now compute the TF-IDF score for each character in the passwords dataset. Character with a higher score are more important, and those with a lower score are less important.

At last the two feature columns 'Length' and lower_case_frequency are added to the above data frame created by vectorization.

$$tf_{t,d} = \frac{n_{t,d}}{Number\ of\ terms\ in\ the\ document} \qquad (tf\_idf)_{t,d} = tf_{t,d} * idf_t$$

$$idf_t = \log \frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

Modelling:

For modeling purpose Logistic regression model is chosen. The data is split into train and test sets with test_size=0.2 .The Logistic regression model is fitted on the training dataset and an accuracy score of 0.801 is obtained by comparing the y_prediction(obtaining by inputting x_test into the model) and y_test sets.