Introducción a la Ciencia de datos

“La Longitud De Las Contraseñas Y Su Influencia En Su Seguridad”

Simon Martin Sposito

Link Al Proyecto: DataCience.ipynb

Introducción

Este informe detalla el análisis del proyecto realizado en el marco del curso de ciencia de datos. El objetivo central de este proyecto era determinar la influencia de la longitud de las contraseñas en la seguridad de las mismas. Los datos proceden de una muestra representativa de contraseñas. Se aplicaron análisis descriptivos junto con técnicas de visualización de datos para identificar los patrones y relaciones clave que existen.

La relación entre la longitud de la contraseña y la variedad de caracteres proporcionará información valiosa para establecer las mejores prácticas para la creación de una contraseña segura.

El análisis implicó el preprocesamiento de los datos, la generación de métricas y gráficos y la interpretación de los resultados, todo ello en relación a los objetivos del curso. En las siguientes secciones se presentan los pasos metodológicos seguidos, los principales hallazgos y las conclusiones alcanzadas en este proyecto.

Descripción

El conjunto de datos analizado contiene información sobre contraseñas y varias características asociadas, está compuesto por 10,000 filas y 9 columnas con los siguientes datos:

Password: La contraseña en texto plano.
Length: Longitud total de la contraseña.
Num Chars: Número de caracteres alfabéticos (excluyendo dígitos y símbolos).
Núm Digits: Número de dígitos presentes en la contraseña.
Num Upper: Número de caracteres alfabéticos en mayúscula.
Num Lower: Número de caracteres alfabéticos en minúscula.
Num Special: Número de caracteres especiales (ejemplo: !@#$%^&*).
Num Vowels: Número de vocales presentes en la contraseña.
Num Syllables: Estimación del número de sílabas en la contraseña.

Análisis

Primeramente como mis datos no poseían datos considerables inútiles o nulos, directamente realice el cálculo de las medidas de posición(Media,Mediana,Moda) y dispersión(Desvío,Mad,Etc) para cada de las 8 Variables siendo Length nuestra variable objetivo, una vez que obtuve las medidas para cada variable se realizó un paneo general de las mediciones obtenidas, La longitud media de las contraseñas está en 6.65, donde en su mayoría resultan caracteres y en minúsculas,además los dígitos también están en muy baja cantidad por lo tanto la mayoría de contraseñas resultan alfabéticas, por último las variables dígitos, mayúsculas, y especiales tienen valores altos de CV (mayores a 1), lo que indica que estas características están muy dispersas entre las contraseña (algunas las usan mucho, pero no la mayoría).

Luego realice una tabla de frecuencia para identificar posibles patrones entre la longitud y el resto de variables, observando los resultados de la tabla obtuvimos que la mayoría de las contraseñas(6.325 precisamente) son no numericas, solo 85 contienen al menos una mayúscula (es decir gran parte son minúsculas aproximadamente 7.832), en cuanto a los caracteres especiales se encuentran en escasa cantidad, por último en cuanto a las vocales y sílabas se encuentran entre una frecuencia de 0 a 5 en su mayoría.

Siguiendo con el análisis del comportamiento realice 8 gráficos, un principal en el que compare la cantidad de caracteres y dígitos dependiendo de la longitud y el resto comparando independientemente la longitud con cada una del resto de variables, por un lado caracteres,dígitos y minúsculas resultan en cierta rango(longitud mayor a 4 caracteres) proporcionales en parte a la longitud, esto también se ve reflejado en el primer gráfico, si bien hay algunas diferencias se observa que en general el resto de variables son proporcionales al tamaño.

En relación a los pasos previos para seguir analizando las relaciones con mi variable objetivo, realice el cálculo de la correlación de pearson (valores entre 1 y -1) para el resto de variables y en relación a esta misma, los resultados obtenidos para caracteres (1), dígitos (0,35) y sílabas (0,23) arrojan una relación positiva y en menor medida o insignificantes son mayúsculas (0,019), vocales (0,057) y especiales (-0,0067) respectivamente.

Para continuar realize tanto una categorización de las contraseñas (Débil, Media, Fuerte) como también el cálculo de su entropía que estima la seguridad de una contraseña considerando su longitud y el espacio de caracteres posibles y agregando ambas medidas al data frame original, la categorización gráficamente arrojó valores desbalanceados por lo tanto realice la codificación de las variables categóricas (0-Débil, 1-Media, 2-Fuerte) y aplique un smote al conjunto de datos pasando de 10007 a 47358 contraseñas

Por último y no menos importante realice el modelo predictivo, dividi mi conjunto de de datos (Entrenamiento y Prueba), los clasifique, genere un reporte, una matriz de confusión y un gráfico donde se observa la relevancia en cuanto a la seguridad de las contraseñas, por una lado la clasificación de las contraseñas es muy buena, por otro lado las variables más importantes son la longitud, las especiales y las mayúsculas, la entropía también resulta importante dado que es proporcional a la longitud y la cantidad de caracteres, en cuanto al

modelo predictivo me pareció adecuado utilizar Random Forest para generar una predicción más exacta y estable, logrando que el grado de seguridad esperado

Conclusión

En conclusión, se logró analizar la relación entre la longitud de las contraseñas y sus características asociadas para evaluar su influencia en la seguridad. Los resultados muestran que, si bien la longitud es un factor clave para la seguridad de las contraseñas, su efectividad se ve notablemente mejorada al incluir caracteres especiales, mayúsculas y dígitos.

El análisis descriptivo mostró que la mayoría de las contraseñas analizadas son alfabéticas y en minúsculas, con poca presencia de caracteres especiales o mayúsculas, lo que demostró una baja seguridad. Las métricas de correlación y la categorización de contraseñas que gracias a su facilidad de interpretación, corroboraron que, aunque la longitud tiene un impacto positivo en la seguridad, la variedad en el tipo de caracteres es esencial para incrementar la entropía y, por ende, la fortaleza de las contraseñas.

La implementación del modelo predictivo basado en Random Forest permitió clasificar contraseñas con buena precisión y destacar la relevancia de las variables más significativas, como la longitud, los caracteres especiales y la entropía.