Capstone Project I Proposal

1. Problem

The usernames and passwords we author to protect our data are comprised of the details that uniquely distinguish individuals. These unique details and patterns that we choose, however, tend to oppose randomness and make us predictable. While there is a lot that is known about passwords, the psychological reasons for username and password choice are poorly understood. Therefore an analysis of login credentials will provide invaluable insight into user behavior that may be used to further security. In this study, we will explore 10-million usernames and passwords combinations in order to learn what they reveal about the people who author them. Furthermore, we will use various features found in the username and password data to build a machine learning model to predict the length and character composition of its counterpart as well as understand how people change their passwords over time.

2. Who cares?

This study on username and password data will help to inform anyone who values information security as well as information technology companies such as Google, Apple, Cisco Systems, and ServiceNow. Identifying and understanding user behavior in the context of username and password authorship will inform technology companies to develop stronger security measures and requirements with user identity and access.

3. Data

This project will look at a data set comprised of ten million username and password combinations made publicly available at (https://xato.net/today-i-am-releasing-ten-million-passwords-b6278bbe7495). The data were extracted through a random sampling of thousands of password dumps.

4. Approach

We will approach the data using a supervised classification algorithm approach to build a predictive model. To predict character compositions, we will break down the alphanumeric and symbol features found in the username strings and use this information to explore what might be predictive in their password counterparts. Additionally, we might also perform a regression on their respective character lengths.

5. Deliverables

I. Code (Jupyter Notebooks) for data acquisition, exploratory analysis, and machine learning model development
II. Capstone Project Report
III. Paper on the capstone project