MVP Tuwaiq Final Project

Graphical user interface, text, application, email

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, bar chart

Description automatically generatedThis MVP explains the data analysis regarding the dataset obtained by Kaggle. Figure 1 shows that several information about the data. For example, there are 322 missing title tag in the HTML source code. Additionally, it checks how many forms containing the contend of URL without HTTPS protocol. In addition, there are 2153 out of 10,000 matches of the most frequent domain hostname in the HTML source code with the URL of the provided website. Moreover, there are 140 JavaScript command in the HTML source code that disable the right click. Thus, this section illustrates a general insight of what is happening after analyzing the data and get as much information as possible. Figure 2 visualizes the relationship between the number of underscores in a random email with the URL length using bar plot from matplotlib library. Figure 3 shows the relation between urlLength feature and the output. It is observed that it is difficult to know the class label by using only the urlLength which indicates that there are lots of urlLength that share the same number with both binary classes. Therefore, this MVP gives the reader the one of the main objectives of the final project which is understanding the data and some of the relation between features and output. Moreover, this project aims to build supervised machine learning algorithm to predict the class label with better accuracy above 70% or obtain the best accuracy by implementing different classification algorithms.