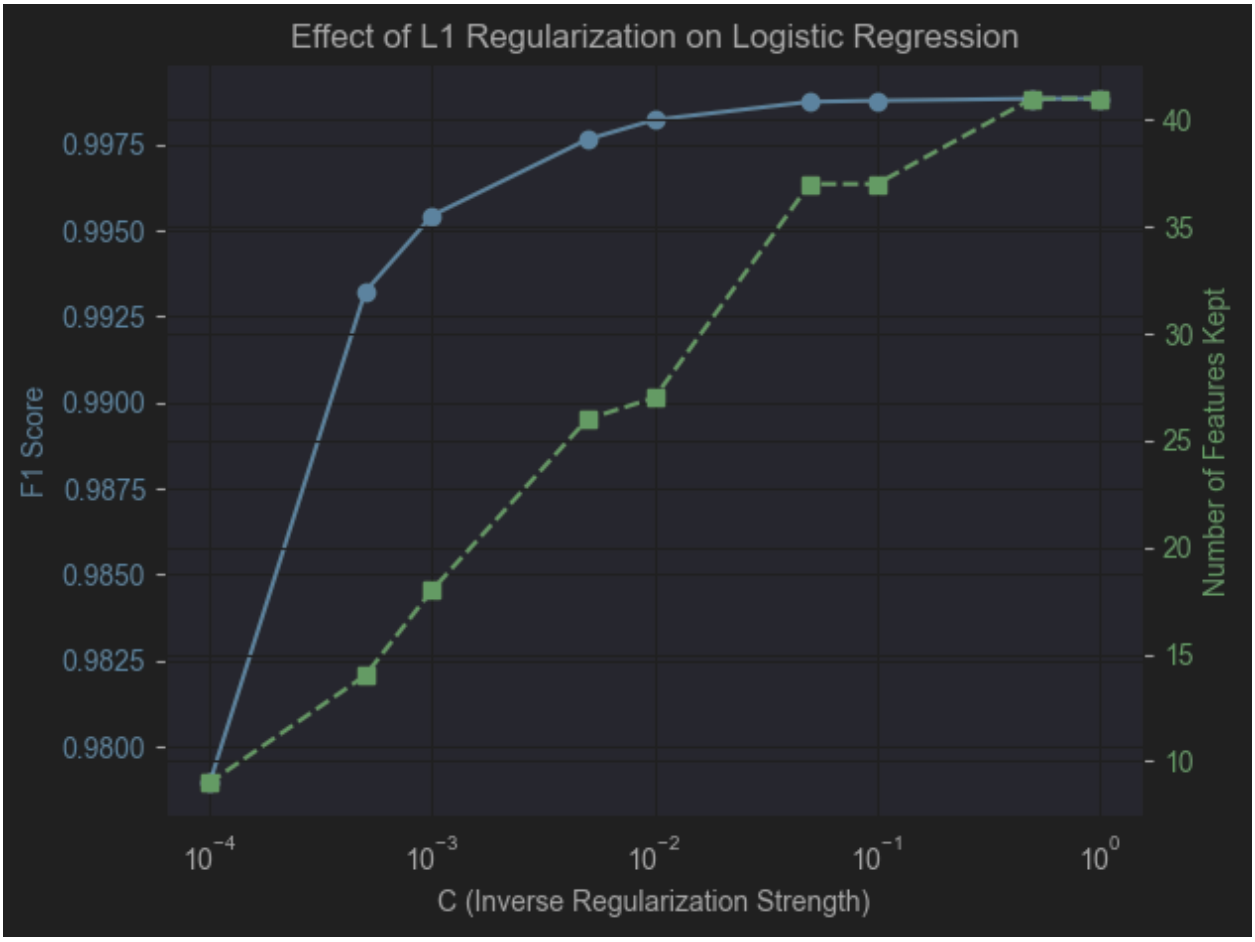
CS4403 – Assignment #5

Chris Higgins

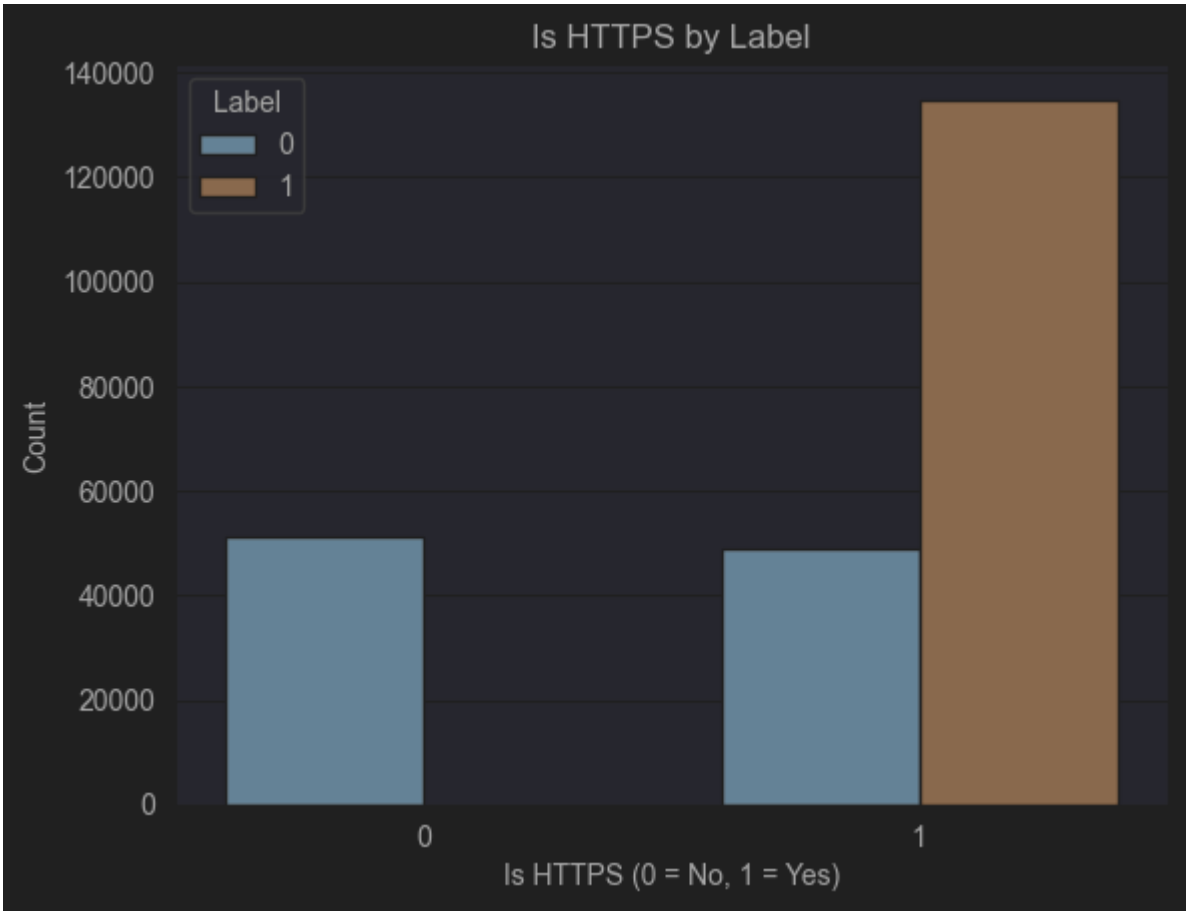
April 3rd, 2025

Effect of L1 Regularization on Logistic Regression (Dual-Axis Line Plot)



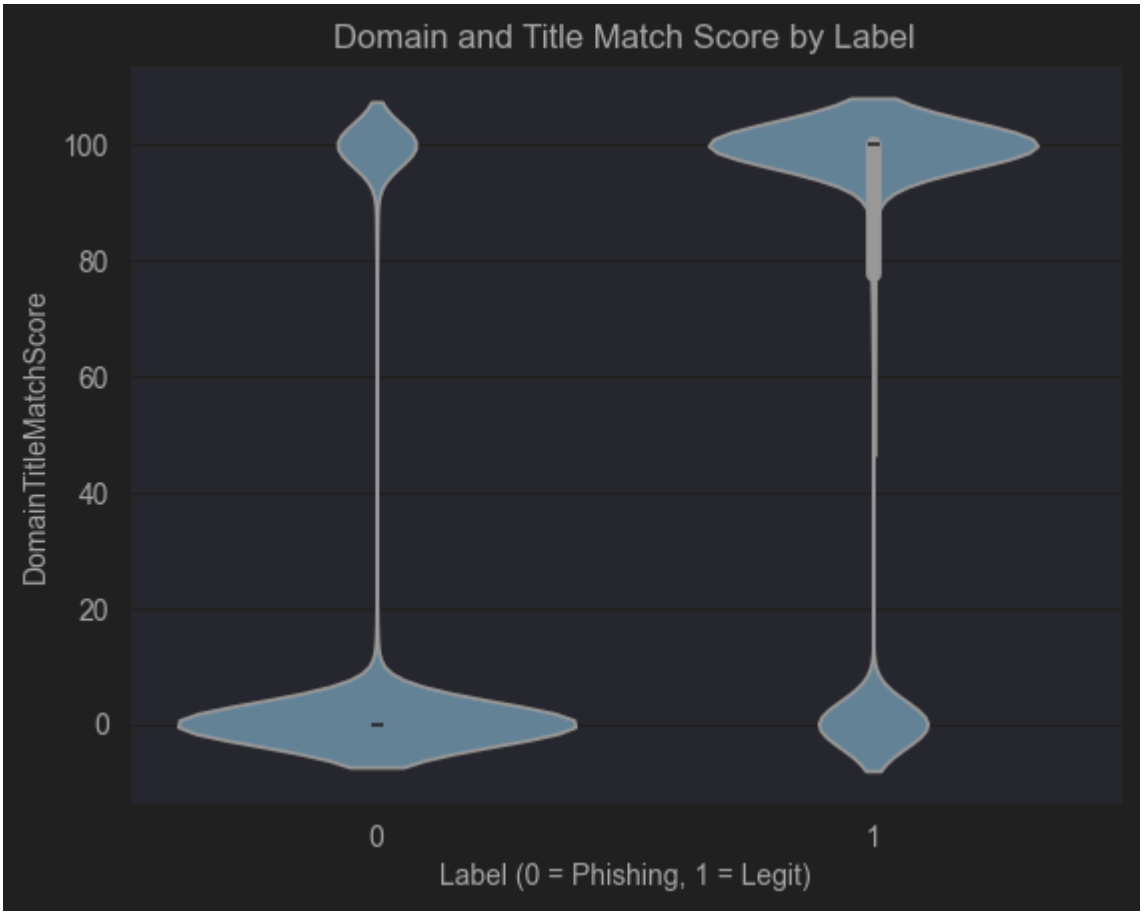
This plot visualizes the effect of changing the c parameter (the inverse regulation strength) on a logistic regression model trained with L1 regularization. It tracks how model performance and feature count evolve. I used L1 regularization to help identify the most important features, and in the case of my dataset – with over 50 features – I chose to reduce the features quite aggressively. After some preliminary feature elimination, I was left with 41 features, and this chart shows that model performance only slightly improves as the model retains more features. In my case L1 regularization was effective in reducing complexity whilst preserving performance. I used a dual-axis line plot because it effectively captured the relationship between c, performance, and model complexity. This analysis helped me to select c parameter value that balanced performance and model simplicity.

HTTPS Usage and Label Distribution (Dual Bar Plot)



This plot visualizes the distribution of HTTPS usage amongst legitimate (1 – orange) and phishing websites (0 – blue). The goal in this plot was explore how HTTPS adoption differs across legitimate and phishing websites. HTTPS encrypts the data exchanged between a user's browser and a website, and is generally seen as a sign of trustworthiness. Phishing websites, however, have increasingly started using HTTPS to appear legitimate, so it was important to see if this feature still held value. The chart shows that while both phishing and legitimate websites use HTTPS, legitimate websites in dataset overwhelmingly use it. The distribution amongst phishing websites is a near-even split. This shows that while HTTPS isn't a guarantee of legitimacy, its absence is still a strong indicator of a phishing website. I chose to use a bar plot because it simply captures the distributions of HTTPS usage between the labels.

Domain and Title Match Score by Label (Violin Plot)



This plot explores the match score between a website's domain and its HTML title. Legitimate websites often have similar titles and domains, whereas phishing websites often spoof domain names but can lack a matching HTML title. This plot shows the spread and concentration of match scores amongst phishing and legitimate websites. I found this to be an interesting visualization, as each plot was almost perfectly opposite of the other. The plot shows that legitimate websites typically have higher match scores, clustering near 100, whereas phishing websites tend to have lower match scores. This separation suggests that this could be a strong indicator. I chose the violin plot because it effectively shows the density and spread of the scores, making it easy to examine the difference in distribution between phishing and legitimate websites.