Data Mining Final Report

Introduction:

Fake accounts on social media are so widespread and prevalent now that the majority of profiles that appear on most social media platforms are likely fake. In many ways, the number of fake social media accounts has gotten so out of hand that some users may feel as if their social media platforms have been overtaken by fake accounts and given the rising sophistication of "bot" accounts it is no longer easy to distinguish fake accounts from genuine accounts. Moreover, the problems with fake accounts on social media do not end here as unlike "real" people with "real" accounts many of these "fake" accounts have little issue with engaging in behavior such as soliciting merchandise that would normally earn "real" people scorn. Due to how significant fake social media accounts have become, the social media experience is considerably degraded for many. Motivated by the mounting problems posed by fake social media accounts, our paper examines several popular data mining techniques on the tasks of detecting fake social media accounts as well as uncovering hidden patterns relating to fake and genuine accounts.

Project Setup:

For this project, our group decided we would work on detecting fake accounts on the popular social media platform Instagram. Fake accounts on social media are of particular interest to us as a group because there are countless variations and ways in which people post fake accounts and try to pass them off as genuine. Although it may be hard to comprehend how exactly people are able to keep coming up with new ways of generating fake accounts on social media, through the direct application of data mining techniques, we can more precisely determine what kinds of common patterns and relationships exist in fake accounts and leave out conjecture. In our project, we examined a dataset on fake versus genuine social media accounts and built classification models off of that and compared model performances. Tim tested two types of algorithms on the fake/genuine social media account dataset, namely random forests and logistic regression. Additionally, Tim also analyzed the attributes present in the dataset such as "profile pic, fullname words, private, #posts, #followers" in order to find the most relevant attributes to discover patterns relating to genuine/fake accounts using apriori rule generation and binning. Lastly, Tim tested the predictive abilities of the algorithms on a few real-life Instagram accounts with the aid of a web scraper extension for Instagram and collected some interesting results. *Add a summary of what you did here:*

Logistic Regression and Random Forest.

        In this project, the logistic regression and random forest algorithms were tested on the social media dataset and compared against each other. The reason these algorithms were selected in particular is because logistic regression and random forest represent two entirely different ways of interpreting the data and prediction making, though it should be noted that despite their differences, these two algorithms were both perfectly capable of achieving 100% accuracy in their own right at least once during K-folds cross validation.

        Of the two algorithms, logistic regression is probably the most intuitive to understand when envisioning how a classification algorithm should perform as it produces linear decision boundaries to separate classes similar to support vector machine. As one may guess, logistic regression is more or less a completely deterministic process as it is conceptually related to maximum likelihood estimation from statistical theory. In contrast, random forests, is a stochastic process and looks at things from an entirely different angle. Unlike most traditional classification methods, random forests is an ensemble learner with random initialization that aggregates the results of several decision trees.

        Ultimately, of the two algorithms tested, random forests proved to be superior to logistic regression at the task of social media account classification although it should be noted again that both algorithms have the same performance ceiling of perfect classification rate. From the confusion matrix of random forests, one can see that random forests has higher accuracy and lower error rate. Moreover, since the odds of random forests making a false positive versus a false negative are the same, we do not have to worry about random forests being biased towards making either type 1 or type 2 errors.
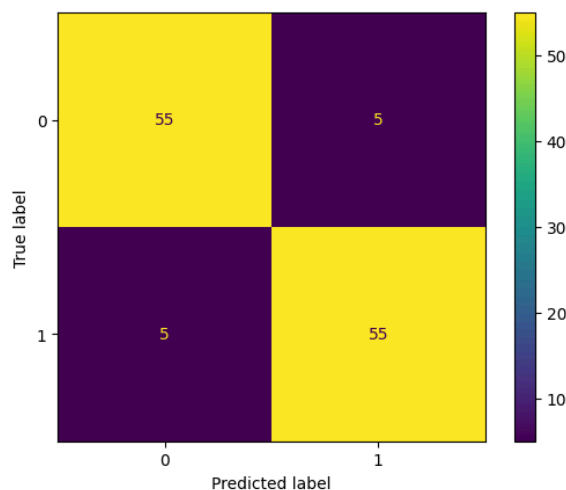
A-priori rule generation and binning:

From the correlation heat map shown below, one can see that there are a fair number of attributes in the original dataset though it is unclear just how meaningful each one these attributes are and in fact some of these attributes may in fact be redundant. Another point that needs to be raised about the number of attributes in our datasets is that there are more attributes present than can be realistically extracted from a given Instagram profile page due to the way private profiles and the like work. Taking all of this into consideration, reducing the number of attributes in our dataset down to a set of only the most meaningful attributes might be a good idea especially if we plan on applying our models to the real world. Therefore, to accomplish this task apriori association analysis was applied to uncover the most meaningful attributes and hidden patterns.
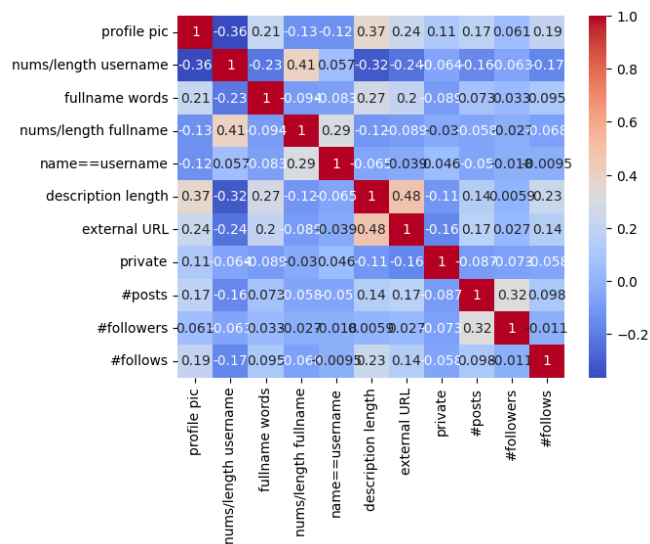
	support	itemsets	length
0	0.812500	(low nums/length username)	1
1	0.815972	(low fullname words)	1
2	0.909722	(low nums/length fullname)	1
3	0.861111	(low description length)	1
4	0.951389	(low #posts)	1
5	0.991319	(low #followers)	1
6	0.934028	(low #follows)	1
12	0.803819	(low #followers, low nums/length username)	2
18	0.810764	(low fullname words, low #followers)	2
21	0.862847	(low nums/length fullname, low #posts)	2
22	0.901042	(low #followers, low nums/length fullname)	2
23	0.847222	(low #follows, low nums/length fullname)	2