Through trial and error, each of the three models (Support Vector Machines, Logistic Regression, and Decision Tree) now has an accuracy of 80%. The 11 features that I selected are a combination of both strings and numbers; whereas numbers are readily transformed into the array for the algorithms, the strings had to be vectorized such that each unique string became a subset of the main feature. For example, one of my features was the protocol of the URL. This feature would be split up into multiple features since different protocols (http, https, etc.) would represent different values. These different features were then encoded in binary to indicate which feature belonged to the URL. The features that I used were: protocol, last URL token (as split by dots), number of tokens (as split by non-alphanumeric characters), average token length (as split by non-alphanumeric characters), number of subdomains, length of domain, top-level domain, whether or not an IP address is domain, delta between registered date and last seen date, number of dots, and number of slashes. I selected each of these because I felt they were useful in identifying malicious URLs. These malicious URLs tend to be longer, and feature more symbols, different top-level domains, seen soon after being registered, and have deep domain names. The accuracy could most likely be increased if able to gain more data about the WHOIS information or by querying services that maintain databases of malicious URLs.

Support Vector Machines

- [http://appleethnic.com/] Should have been marked benign, was marked malicious due to time delta being too small.
- [http://www.rubayetenterprise.com/wp-includes/certificates/hsbcred.html] Should have been marked malicious, was marked benign due to time delta being too large.
- [http://ilangaijeyaraj.org/urch/wellsfargo/] Should have been marked malicious, was marked benign due to time delta being too long, length is fairly short, and top-level domain is typically benign.
- [http://c.trackkbr.com/?E=SkD1EyPQsHn5mLJIDBEUBxc5UzPBsBVW-7DftSGWV...] Should have been marked benign, was marked malicious due to length.
- [http://mirror.mailings.emailer3.com/?e=abuse%40priestlegal.com.au&s=...] Should have been marked benign, was marked malicious due to length and the number of symbols.

Logistic Regression

- [http://outtlookweb.tripod.com/] Should have been marked malicious, was marked benign due to short length and large time delta.
- [http://www.startcupsardegna.it/%3Flpon%3Derin-cummins-fappening&...] Should have been marked benign, was marked malicious due to long length and foreign top-level domain.
- [http://www.matunisie.com/forum/index.php?/topic/159-nabeul/] Should have been marked benign, was marked malicious due to relatively long length and the number of symbols in the URLs.
- [http://crystals-car.com.es/] Should have been marked benign, was marked malicious due to top-level domain, time-delta of 0.

- [http://download-mp3-song.net/download-mp3/171302658/mitti-di-khushboo-...l] Should have been marked benign, was marked malicious due to length, symbols.

Decision Tree

- [http://www.benchmarkemail.com/c/u?9tSN8Jgb1CaQ1JQj2f1UsrkzYtJRKs0w1cn...] Should have been marked benign, was marked malicious due to length.
- [http://reversewhois.domaintools.com/?email=faef0d1860645e875e48cad06cd67d] Should have been marked benign, was marked malicious due to length.
- [http://74.122.161.50//ql.html?r=uem01*insbulkw4nutra1056oth=oth.1ck1.1cb...] Should have been marked benign, was marked malicious due to the fact that the domain is an IP address, the length of the URL, and the number of symbols in the URL.
- [http://ftpshodloer.free.fr/] Should have been marked benign, was marked malicious due to top-level domain.
- [http://www.letsdnd.com/mymail/8028/c515c7edc6b24b0c46f40537599bf2da/aHR0cDovL3d3dy5sZXRzZG5kLmNvbS9yZXNwb25zaXZlLXdlYi1kZXNpZ24tdXRpbGl0eS1idXNpbmVzc2VzLw/1] Should have been marked benign, was marked malicious due to length.

Of the three visualizations, the decision tree is by far the easiest to understand. The other two visualizations are simply too abstract to provide any meaningful representation of the data. Besides being easy to understand, it is also able to represent all of the features used in the machine learning process and how it arrived at the decision for each URL. Overall, the program runs a reasonable pace; while not as long as I had expected, it certainly is not the fastest program. The logistic regression seems to take the shortest amount of time, followed by the decision tree, and finally the support vector machines. Each of the following visualizations was made using the first two features of the set, with the exception of the decision tree which used all of the features. The decision tree would not fit in the document, so it has been placed along with the rest of the code.