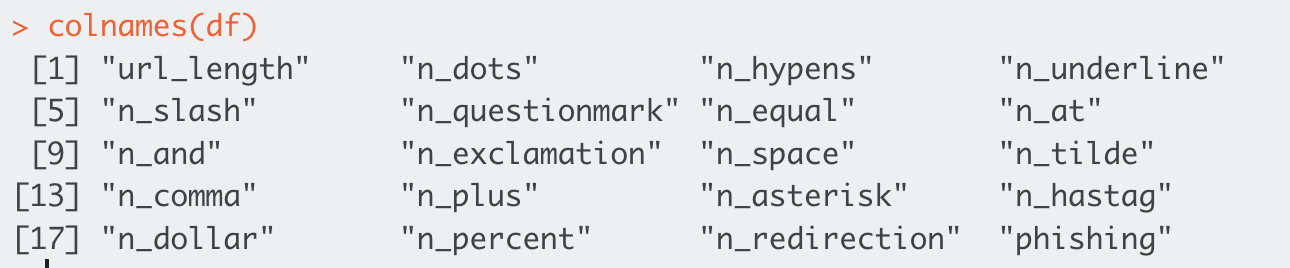
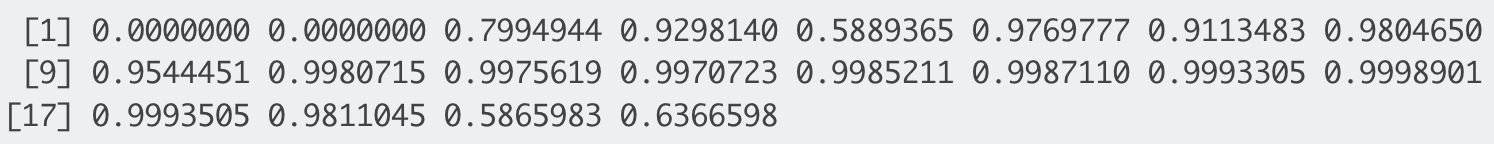
# BIOS 735 Project Plan & Proposal

## **Choice of dataset**

* Web Page Phishing data (classification)
  + https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset

## **Proposal**

### Web Phishing dataset summary

* 1. Cross-sectional dataset on web page phishing. It contains the extracted feature from URLs and a binary outcome variable (whether it is a phishing website). Original URLs are not available in this data. It comes from Kaggle: https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset
  2. 100,077 observations and 20 variables.
  3. Response variable phishing: Binary, is this webpage phishing or not?
  4. Explanatory variables: 19 features extracted from the original URL including:
     + “url\_length”: The length of the URL
     + “n\_dots”: The number of dots in the URL
     + “n\_hypens”: The number of hyphens in the URL, etc.
     + All columns:
     + 
  5. The data is very clean. There is no missingness in this data, with no NA values on inspection.
  6. Primary concern with the data: there is a lot of zero-inflation in most of the covariates. Specifically, below is the percentage of zeros in each column. Note that it makes sense for there to be no zeros in url\_length and n\_dots. However, there are nearly 98% zeros in a lot of these covariates. The response variable phishing is not zero-inflated; it is fairly balanced in terms of labels with 63% zeros and 37% ones.
     + We are still looking into the literature on how best to handle this, but we don’t think that we need to do anything specific other than conduct feature selection after model-fitting to see if these features are informative. However, our suspicion is that all of these zero-inflated covariates may get thrown out due to the extremely small variance of these covariates, so we are not sure if this will even be useful. If you have any other suggestions about if we need to do anything specific for this, please let us know.

### Research question(s) of interest

* 1. Model the probability of phishing using extracted features of URLs
     + Comparing likelihood-based approach (GLM) with random forest (ML)
       1. Split the data into training (80%), and test set (20%)
       2. Utilize the normalized training set to build the models
          1. For GLM:

Fit a generalized linear models with L1 penalty (LASSO) to select features (use 5-fold CV on normalized training set to select the hyperparameter for the model)

Utilize coordinate descent to optimize LASSO

Compare different link functions: logistic, probit, and other link functions

Reference: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/>

* + - * 1. For ML:

Construct a random forest classifier for and classification (5-fold CV will be applied on training set for hyperparameter tuning and feature selection)

* + - 1. Comparison metrics:
         1. Since our outcome of interest is binary, we will compare metrics below of the two models on both the training and test data

Evaluation metrics: AUC, precision, recall, accuracy, F1 score (if imbalanced data)

### Specific tasks

|  | Coding | Report writing | Who to claim |
| --- | --- | --- | --- |
| Data exploratory analysis | Summarization (summary statistics & figures, univariate and bivariate)  Preprocessing (missing data, transformation, etc.) | Introduction | Madhuri Raman |
| Likelihood-based model | - Model implementation using optimization method from Model 2 (TBD)  - Feature selection | Method (writing out likelihood equations, optimization algorithm)  Result (Figures) | Jiawen Du  Sophie Shan |
| Random Forest (ML model) | - Model implementation with cross-validation  - Tuning hyperparameters  - Feature selection | Method  Result (Figures) | Will Tang  Xuejun Sun |
| Result gathering | Rmarkdown writeup  Package creation  Repository management | Result  Discussion | Madhuri Raman |

### To-do next

* 1. Job splitting - assign specific tasks to specific people
     + Code
       1. 2 people for GLM
       2. 2 people for ML
       3. 1 person for EDA and package creation/repository management
     + Report writing
       1. TBD
  2. Start working on data preprocessing, EDA, and likelihood based method (from scratch)
  3. Schedule next meeting
  4. Create a timeline for project progression and completion

## **Weekly meeting time**

* Next meeting: week of March 25th or April 1st
* Future meetings: as necessary (weekly date/time TBD)