(i) Project title and the names of the team members

- Title: Machine Learning Model Generation for Fraudulent Email Detection
- Names: Zhixi Lin, Linwei Jiang, Haiping Yuan

(ii) Detailed description of the research question

Our intuition came from the fact that ordinary spam email that contains useless information generally doesn't do much harm other than creating inconvenience for the receiver. Fraudulent email, on the other hand, could cause potential loss other than making the recipient's inbox inconvenient to use. Taking phishing emails as an example, it has the potential of directing the recipient to accidentally give out their own credentials, where the recipient might lose possession of not only their own privacy information but potentially take huge financial losses or other threats.

Thus, if a model could be generated to identify these potentially dangerous fraudulent emails, in addition to putting them inside the spam email folder, we can also report the senders to an authoritative database to track down these senders or avoid other people from receiving emails from these accounts.

(iii) A thorough literature survey with appropriate references

The lack of creating a dynamic fraud detective system is one of the disadvantages of traditional ways of email filters. As fraudsters always keep updating their mode of phishing to escape discovery (Akinyelu & Adewumi 2014). In the research, the author claimed that random forest is the best approach with 99.7% accuracy on the best eight feature classification (Akinyelu & Adewumi, 2014).

With phishing attacks changing, the firefly algorithm (FFA) with SVM based become a successful classifier to detect phishing and harmful emails. (Akinyelu & Adewumi, 2016) In the study, FFA with SVM based is used for feature selection and parameter selection, which is considered the critical part of the model. By using this model, the accuracy of the test is 99.94% and the FP rate and FN rate are lower than 6% by classifying more than 4000 harmful emails.

Natural language processing (NLP) is applied for data pre-processing and SVM, and Probabilistic Neural Network (PNN) is combined for data analysis on the detected system (Kumar et al.2020). Using NLP and machine learning is an effective way to catch phishing emails (Saleem 2020). In Saleem's research of classifying non-phishing and phishing emails after the data process, the random forest got 98% of accuracy, while the decision tree and Naïve Bayesian in gaussian got the highest accuracy machine learning model (Saleem 2020).

Feature selection in three machine learning methods like SVM, Naïve Bayesian, and Multi-Layer Perceptron is an effective model to mark spam email (Jayasimha, 2020).

(iv) Details of the algorithm(s) to be implemented

The following are a few algorithms that we've found would suit our purpose. We will select one out of them to use for our initial implementation. If the selected algorithm is inefficient and/or we have a sufficient amount of time, we will continue with another algorithm just to compare the efficiency/accuracy between the two.

Random Forest (RF algorithm): An learning method for classification, regression, and other tasks. Operates by constructing multiple decision trees and then later merging them together for a more accurate prediction. The algorithm will take the average of output from various trees to make a prediction.

Support Vector Machine (SVM algorithm): A machine learning algorithm used for classification and regression. This algorithm uses an N-dimensional space to find a hyperplane that distinctly classifies data points into its own class. Depending on the number of features, the number of dimensions of the hyperplane will also increase.

Bayesian Network: A probabilistic graphical model that represents a set of nodes. Each of which represents a variable and its edge represents the conditional probability for the variables

(v) Expected experiments and analysis to be performed, including a list of datasets to be used for the experiments

In reference (2) we found a list of email datasets with phishing/fraud, spam, and normal email types. We can start from there to see what we can use, then proceed to look for other datasets if necessary.

Once we have successfully implemented a program with one of our algorithms and come up with trained models, if the results are not ideal and there is still sufficient time to proceed with a new algorithm, we will try another algorithm to compare the prediction accuracy and/or efficiency.

(vi) A timeline of significant steps in the project

1. Oct. 10 - Oct. 16
    - Dataset collection and filtration for later usage.
2. Oct. 17 - Oct. 25
    - Learn about PyTorch, machine learning APIs, machine learning algorithms, data processing APIs, and all other necessary components to start the implementation step. Since we already have Dataset ready, we are able to start processing data while learning this step.
3. Oct. 26 - Nov 14
    - Data Pre-processing, code implementation.
4. Nov. 15 - Nov. 25
    - Model generation, testing, debugging.
5. Nov. 25 - Dec. 5
    - Final report.

References:

1. Adewumi, O.A. and Akinyelu, A.A. (2016), "A hybrid firefly and support vector machine classifier for phishing email detection", Kybernetes, Vol. 45 No. 6, pp. 977-994. https://doi.org/10.1108/K-07-2014-0129
2. Gangavarapu, T., Jaidhar, C.D. & Chanduka, B. Applicability of machine learning in spam and phishing email filtering: review and approaches. Artif Intell Rev 53, 5019–5081 (2020). https://doi.org/10.1007/s10462-020-09814-9
3. Andronicus A. Akinyelu, Aderemi O. Adewumi, "Classification of Phishing Email Using Random Forest Machine Learning Technique", Journal of Applied Mathematics, vol. 2014, Article ID 425731, 6 pages, 2014. https://doi.org/10.1155/2014/425731
4. Saleem, Babar M. The P-Fryer: Using Machine Learning and Classification to Effectively Detect Phishing Emails, Marymount University, Ann Arbor, 2021. ProQuest, https://www.proquest.com/dissertations-theses/p-fryer-using-machine-learning-classification/docview/2572551978/se-2.
5. Kumar, A., Jyotir, M. C., & Díaz, V. G. (2020). A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing. International Journal of Electrical and Computer Engineering, 10(1), 486-493. https://doi.org/10.11591/ijece.v10i1.pp486-493