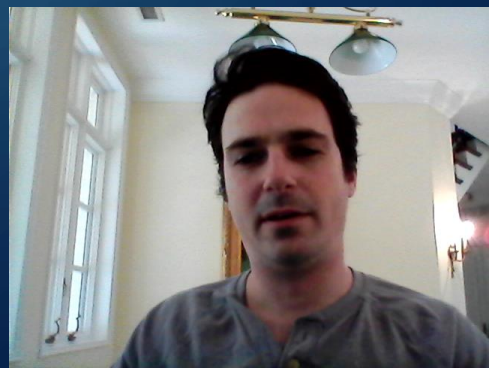
MALWARE DETECTION WITH MACHINE LEARNING

BRETT LUSKIN

DATA 606 – CAPSTONE SPRING 2020



RECAP

- Network Intrusion Detection System (NIDS) using UNSW-NB15 Dataset
- No Domain Knowledge
- Benchmark using **existing research** (Reproducibility)
- Implement new method

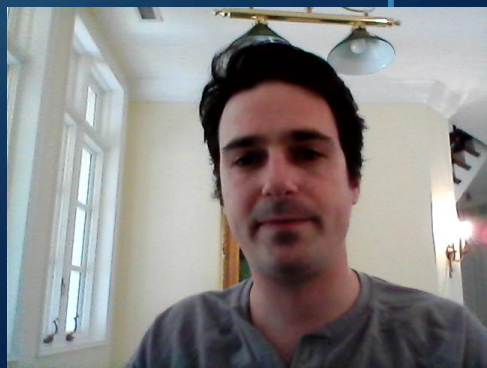
SIGNATURE VERSUS ANOMALY DETECTION

Signature detection: NIDS based on a database of existing known attacks

Anomaly detection: NIDS based on detecting unknown attacks using profile parameters

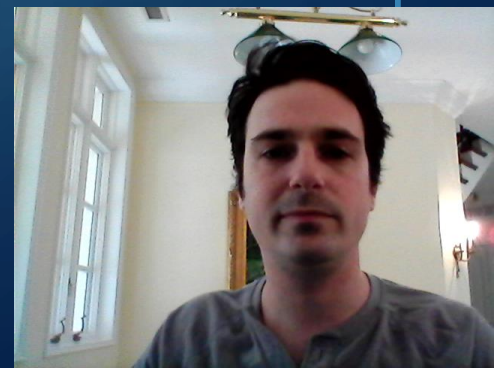
MACHINE LEARNING METHODS

- Logistic Regression
- SVM (RBF kernel)
- PCA
- Random Forest
- ADABoost



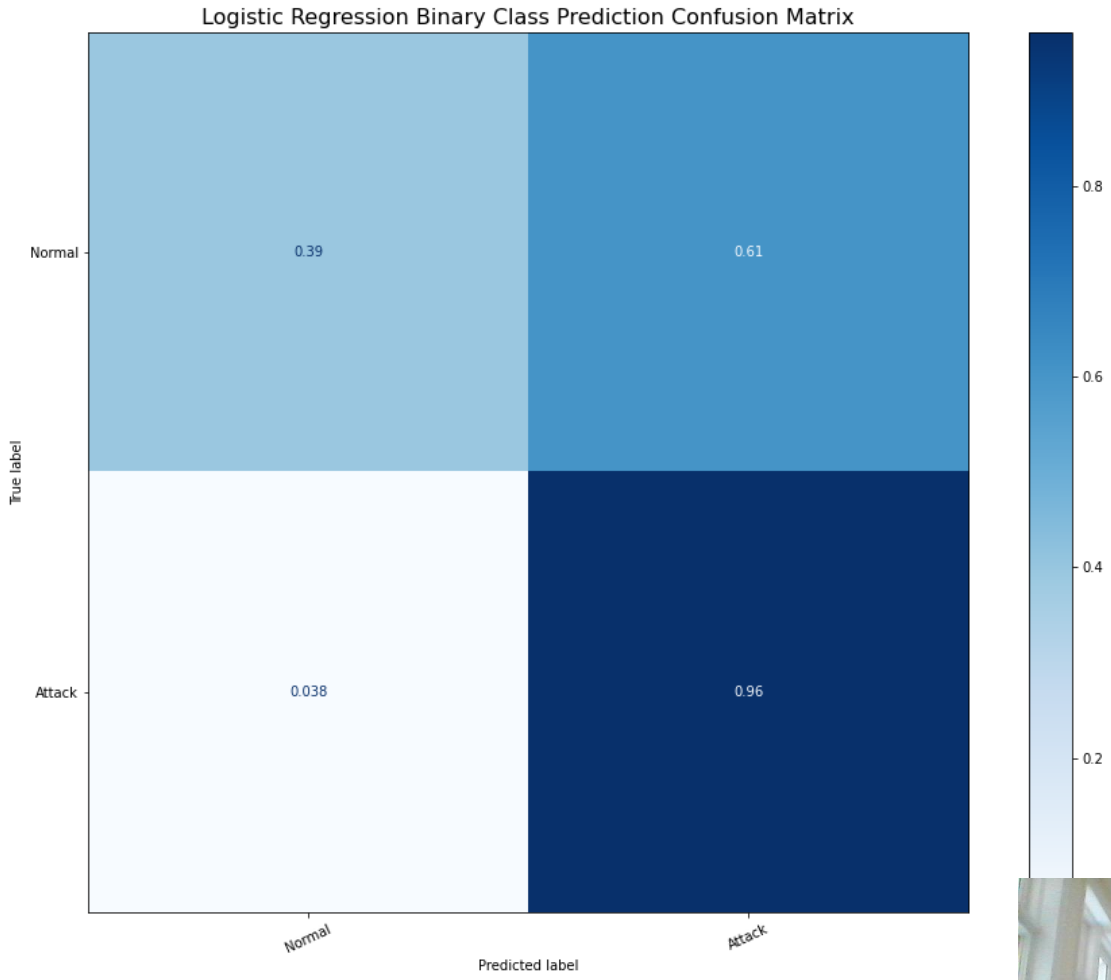
MACHINE LEARNING RESULTS

- Logistic Regression – Mediocre, linear method
- SVM (RBF kernel) – Coin flip, compute intensive
- PCA – Coin flip
- Random Forest – Strong
- ADABoost – Strong



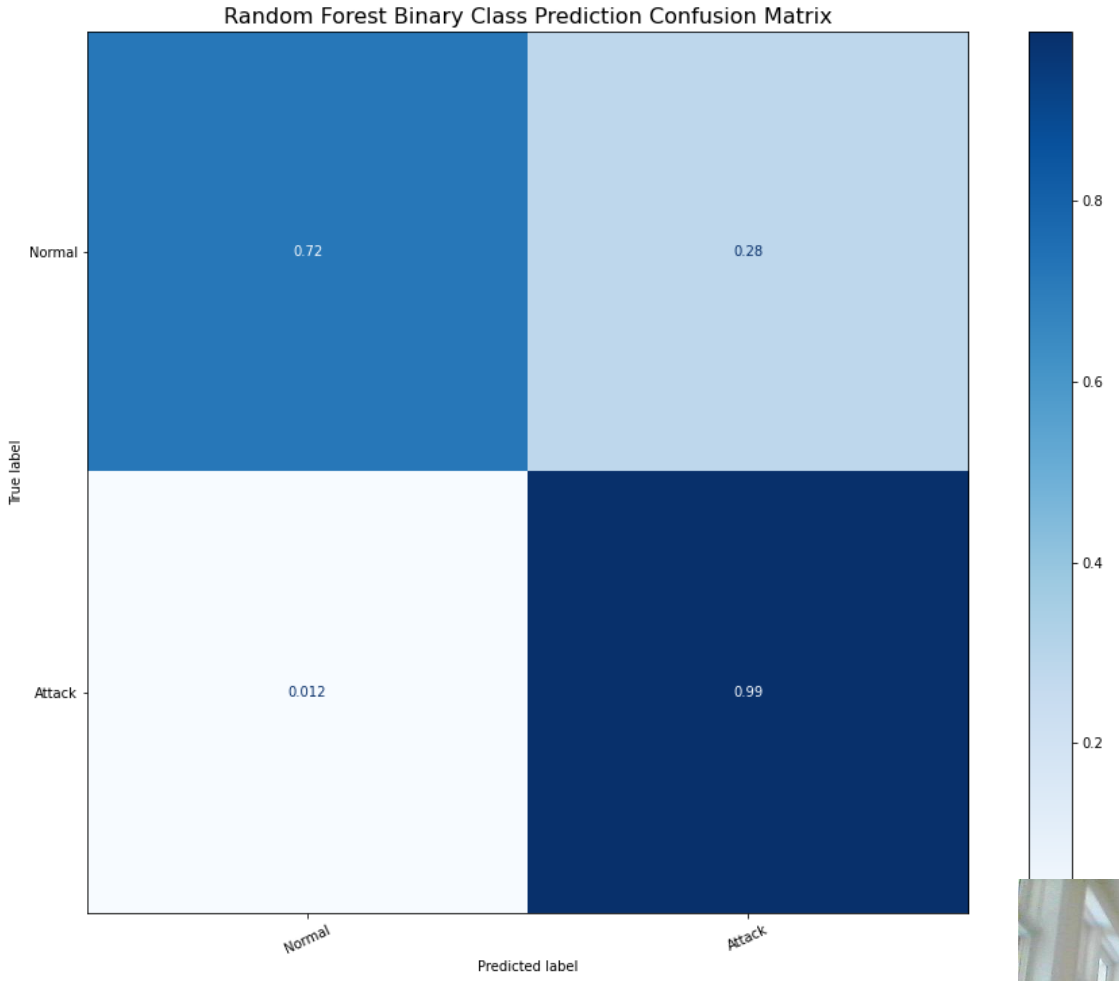
LOGISTIC REGRESSION

- Good binary classifier
- Linear
- Implemented with default sklearn parameters



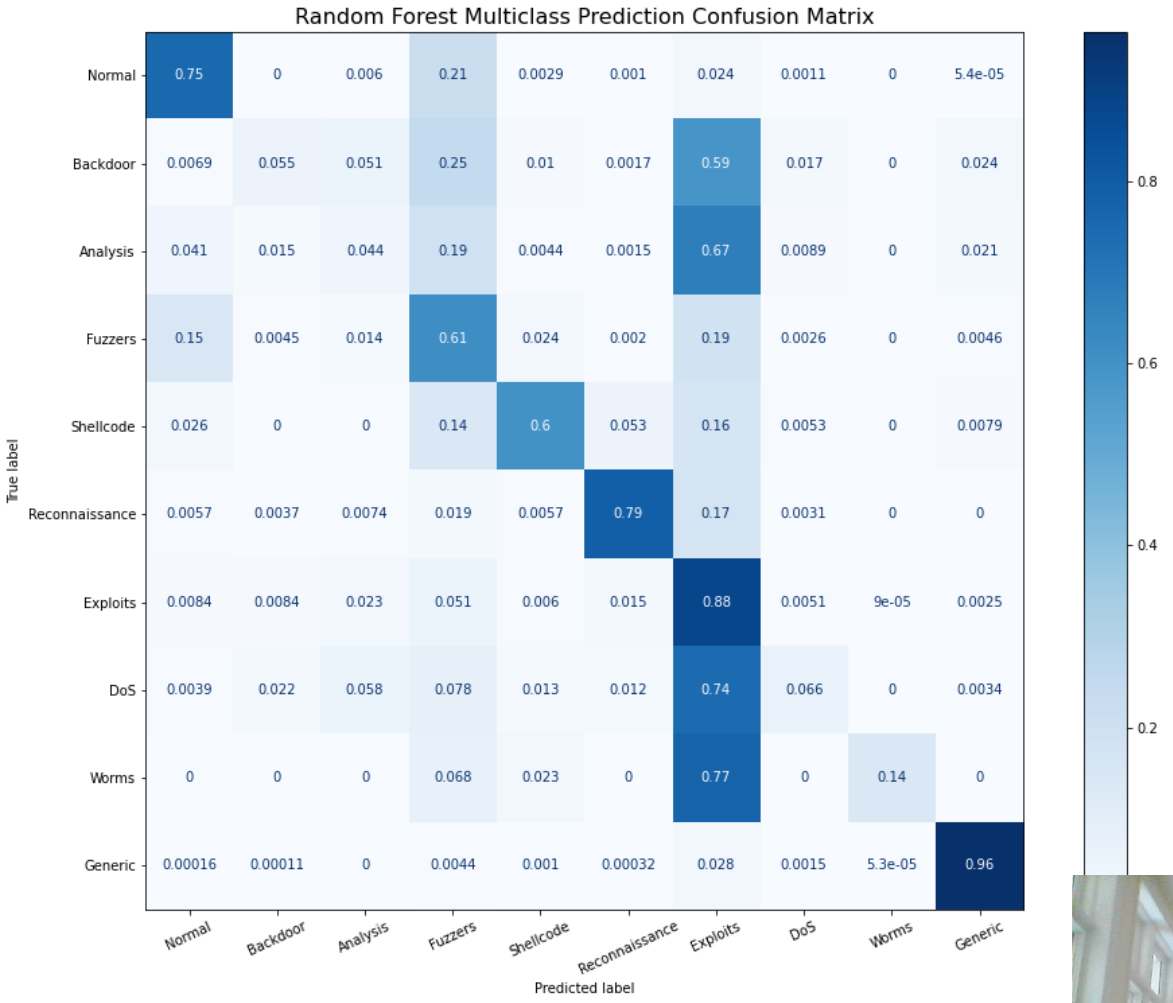
RANDOM FOREST

- Decision Tree Classifier
- Multiple weak learners mitigates overfitting
- N_estimators = 18
- Max_depth = 15



RANDOM FOREST

- Decision Tree Classifier
- Multiple weak learners mitigates overfitting
- N_estimators = 18
- Max_depth = 15

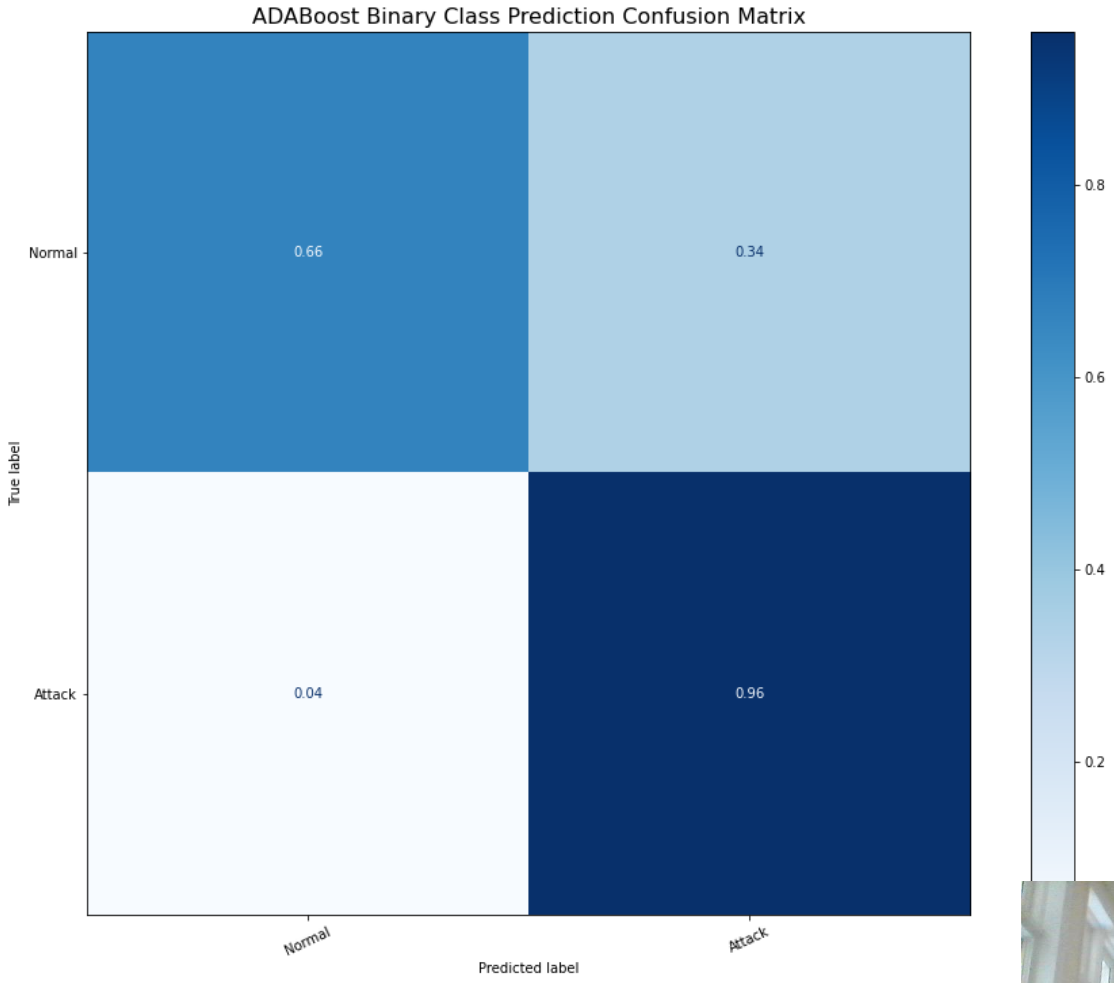


RANDOM FOREST BENCHMARK

- K. Hassine, A. Erbad and R. Hamila, "Important Complexity Reduction of Random Forest in Multi-Classification Problem," 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 2019, pp. 226-231.
- Their Multiclass Accuracy: 75%
- My Multiclass Accuracy: 76%

ADABOOST

- Decision Tree Classifier
- Copy classifiers have weight adjusted by previous iterations
- Adaptive Boosting
- N_estimators = 18

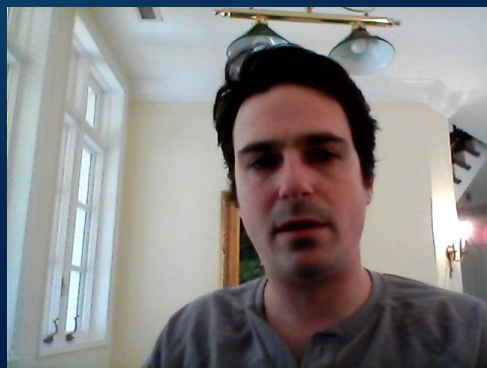


MACHINE LEARNING SUMMARY

- Decision Trees perform
- Logistic Regression surprises
- SVM disappoints
- PCA exists

NEURAL NETWORK

- L. Zhiqiang, G. Mohi-Ud-Din, L. Bing, L. Jianchao, Z. Ye and L. Zhijun, "Modeling Network Intrusion Detection System Using Feed-Forward Neural Network Using UNSW-NB15 Dataset," 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, 2019, pp. 299-303.
- Ten hidden layers, 10 neurons each, Stochastic Gradient Descent, 10 epochs
- Activation function? Batch normalization?



DATASET

- Prepared training set: 175,341 records
- Full dataset: 2.5 million records

The UNSW-NB15 Dataset Description

The UNSW-NB15 source files (pcap files, BRO files, Argus Files, CSV files and the reports) can be downloaded from HERE.

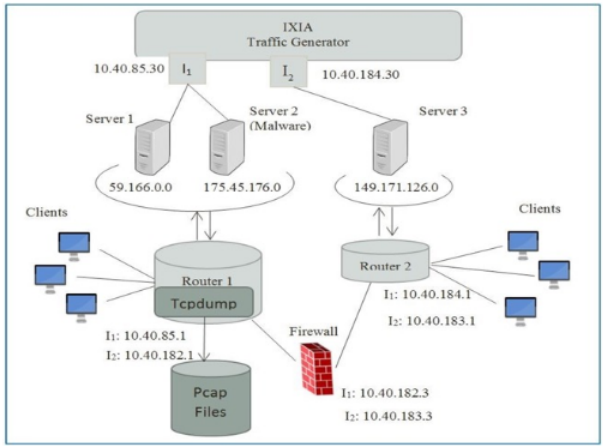


Figure 1: UNSW-NB15 Testbed

The raw network packets of the UNSW-NB 15 dataset was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours.

Tcpdump tool is utilised to capture 100 GB of the raw traffic (e.g., Pcap files). This dataset has nine types of attacks, namely, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. The Argus, Bro-IDS tools are used and twelve algorithms are developed to generate totally 49 features with the class label.

These features are described in UNSW-NB15_features.csv file.

The total number of records is two million and 540,044 which are stored in the four CSV files, namely, UNSW-NB15_1.csv, UNSW-NB15_2.csv, UNSW-NB15_3.csv and UNSW-NB15_4.csv.

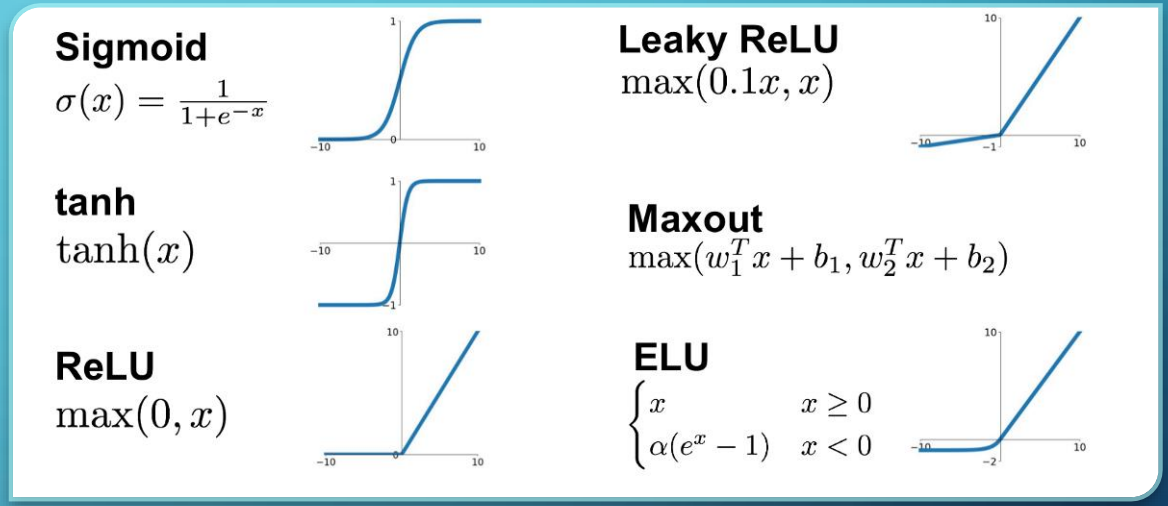
The ground truth table is named UNSW-NB15_GT.csv and the list of event file is called UNSW-NB15_LIST_EVENTS.csv.

A partition from this dataset is configured as a training set and testing set, namely, UNSW_NB15_training-set.csv and UNSW_NB15_t

The number of records in the training set is 175,341 records and the testing set is 82,332 records from the different types, attack and configuration dataset and the method of the feature creation of the UNSW-NB15, respectively.

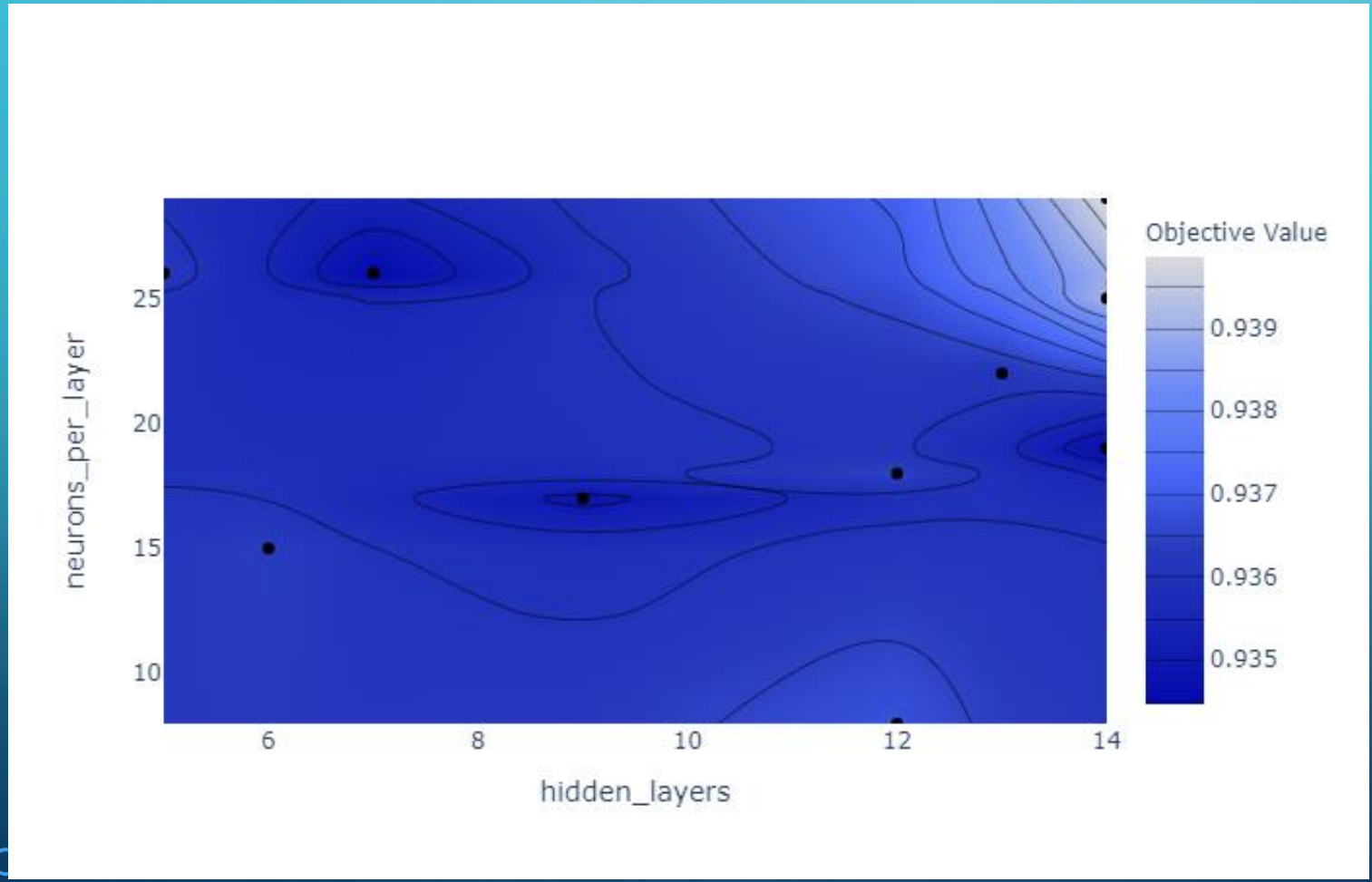
FIRST NEURAL NETWORK

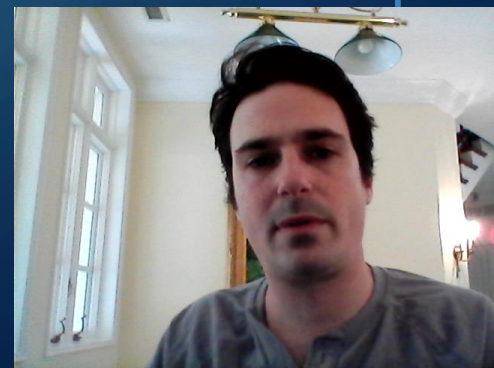
- Sequential network
- Linear layers with Batch normalization
- ReLU activation
- Sigmoid activation in last hidden layer
- 10 epochs
- Stochastic Gradient Descent



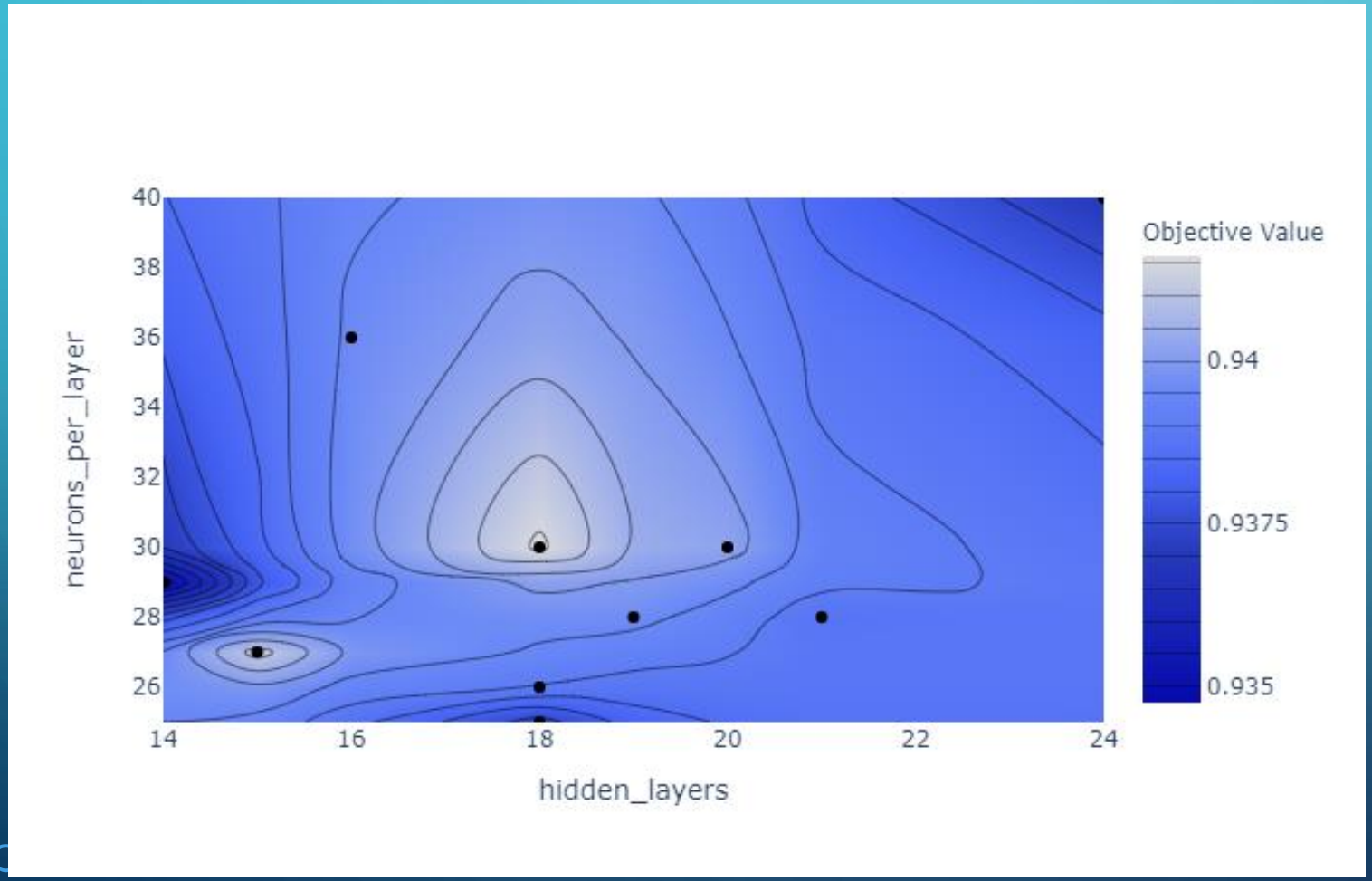
https://towardsdatascience.com/complete-guide-of-activation-functions-34076e95d044

OPTIMIZING WITH OPTUNA

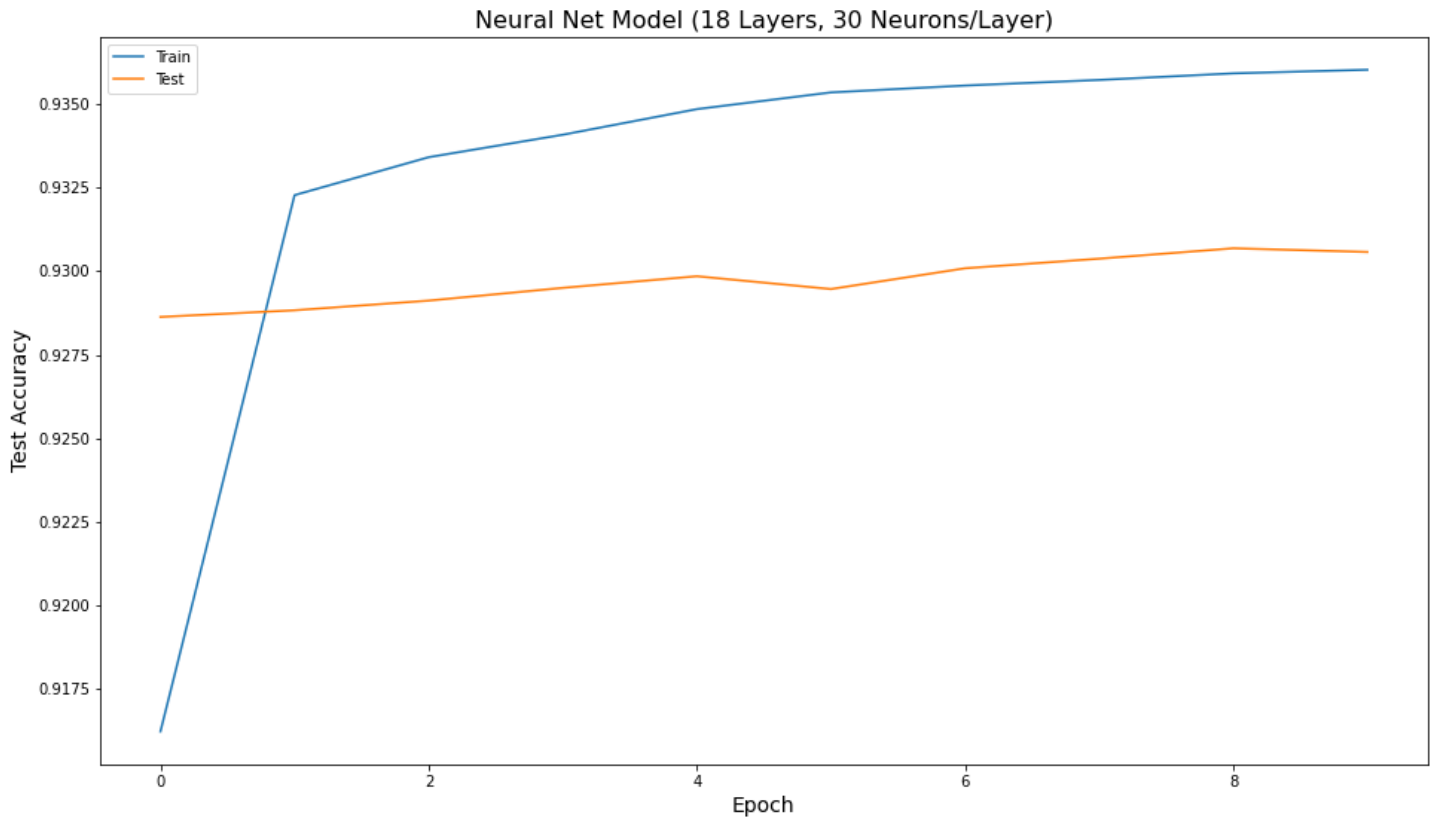




OPTIMIZING WITH OPTUNA

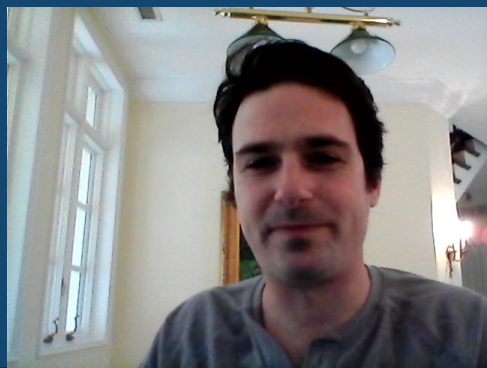


OPTIMIZED MODEL



NEXT STEPS

- Optimize linear model with new parameters
- Other Neural Network options
- Revisit Machine Learning methods?



REFERENCES

K. Hassine, A. Erbad and R. Hamila, "Important Complexity Reduction of Random Forest in Multi-Classification Problem," 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 2019, pp. 226-231.

T. Janarthanan and S. Zargari, "Feature selection in UNSW-NB15 and KDDCUP'99 datasets," 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, 2017, pp. 1881-1886.

D. Jing and H. Chen, "SVM Based Network Intrusion Detection for the UNSW-NB15 Dataset," 2019 IEEE 13th International Conference on ASIC (ASICON), Chongqing, China, 2019, pp. 1-4.

L. Zhiqiang, G. Mohi-Ud-Din, L. Bing, L. Jianchao, Z. Ye and L. Zhijun, "Modeling Network Intrusion Detection System Using Feed-Forward Neural Network Using UNSW-NB15 Dataset," 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, 2019, pp. 299-303.

UNSW-NB15 Dataset: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/

My github: https://github.com/blusk44/Capstone606/tree/master/Delivery-3