INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

INTERIM PROJECT REPORT
GROUP 17

VASISHT DUDDU 2015137

SHUBHAM KHANNA 2015179

ANUBHAV JAIN 2015129

MOTIVATION

- Increasing malware complexity and sophistication
- Polymorphic and metamorphic malware change form dynamically and cannot be detected by traditional anti-virus
- Hard Coded and Rule Based IDS not applicable
- Require to adapt defences to detect attacks based on past data
- ML models can help to create robust defences

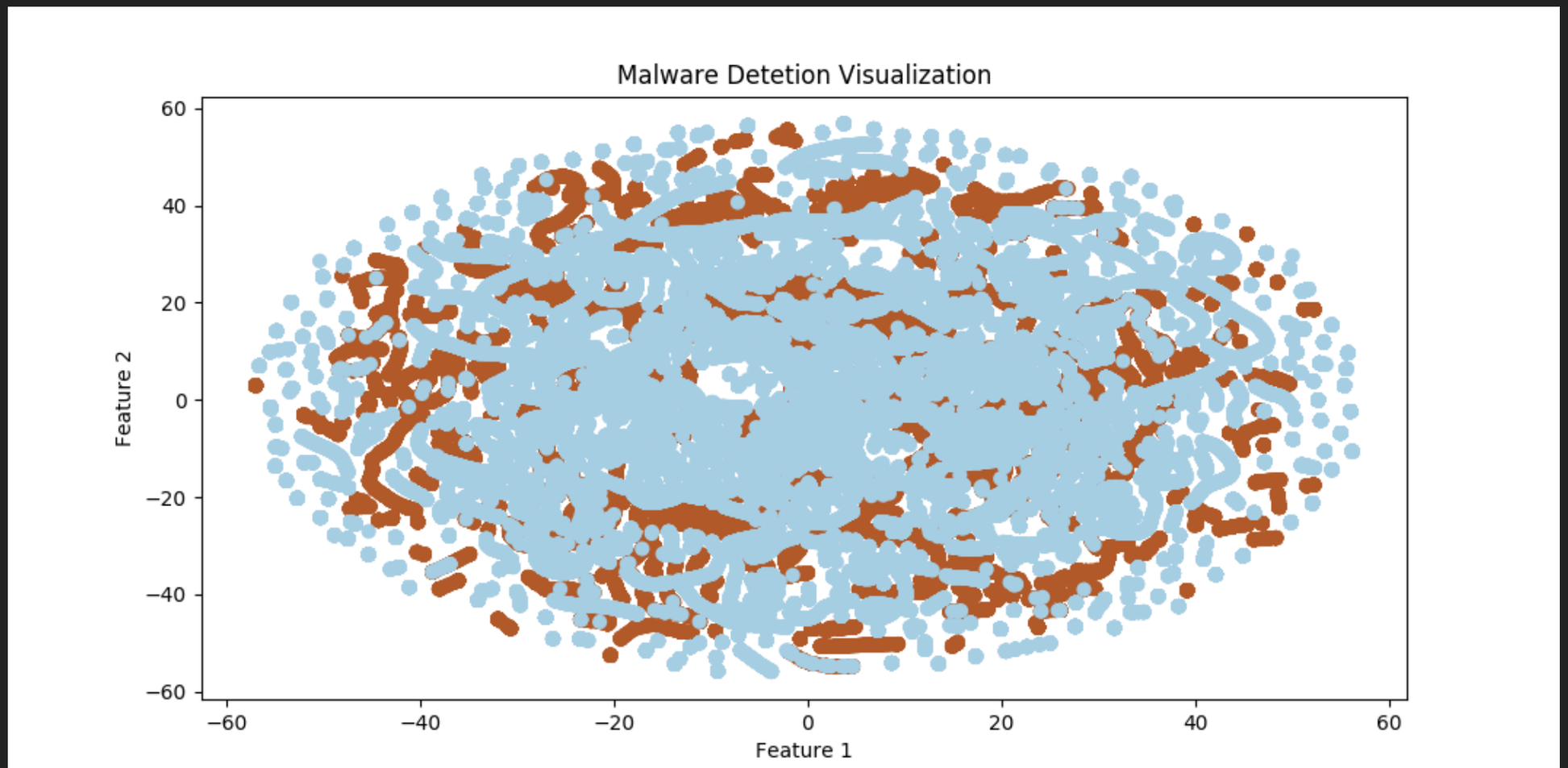
PREVIOUS WORK AND RESULTS

Source	Model	Accuracy	False Positives
Tek	Random Forest	99.35	0.56
Adobe	Random Forest	98.21	6.7

Data set used by Adobe is more extensive and contains more features for complex malware

- Malware classification of PE32 executables
- Siddiqui et al. Accuracy : 94%
- Schultz et al. Accuracy : 97.76%
- Shafiq et al. Accuracy : 99%
- Ye et al. Accuracy : 92%
- Ye et al. Accuracy : 93.8%

DATA SET



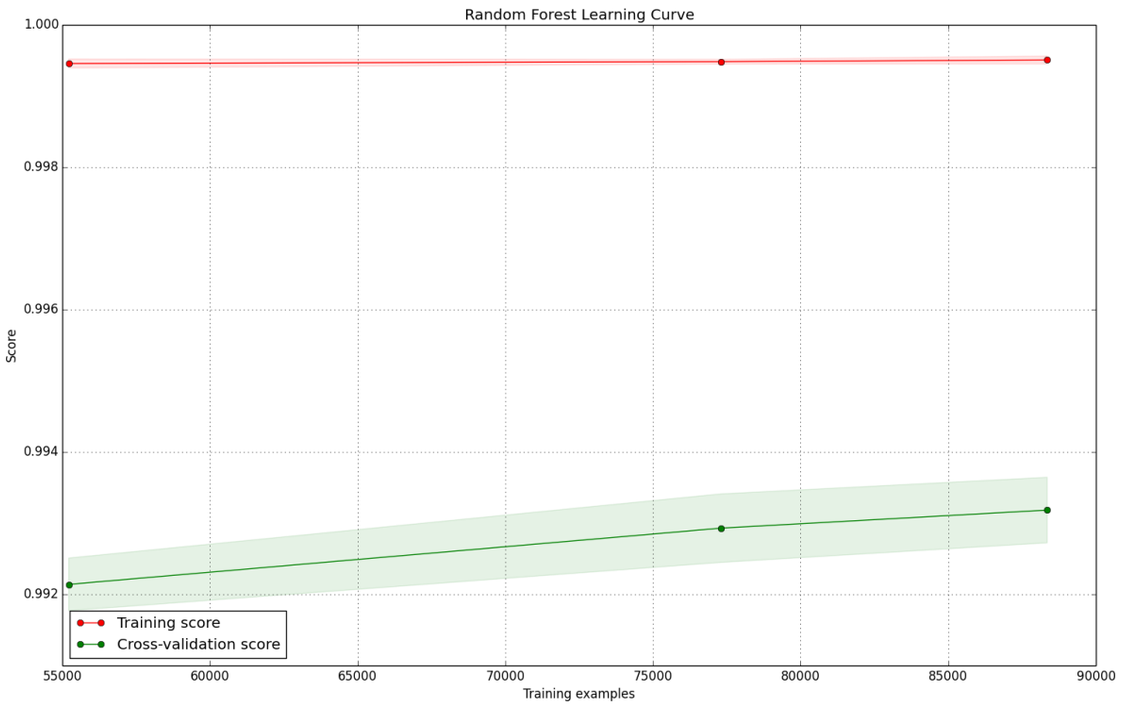
- Total Size : 138047 (Data is Non-Separable)
- Training Size : 96632 Testing Size : 41415
- Features: 54 After Feature Extraction : 14

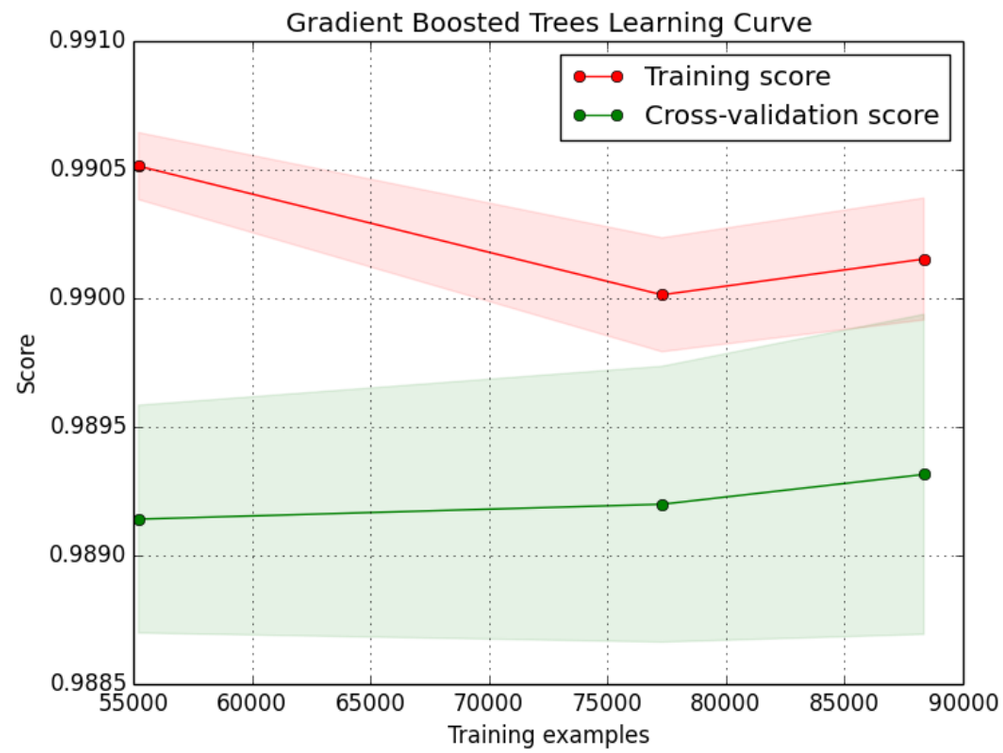
EVALUATION METRICS

- Classification Accuracy
- False Positive Percent (Evaluated using Confusion Matrix)
- AUC Score from ROC Curve
- Threshold tuning for better specificity vs sensitivity

ANALYSIS

- Objective: Reduce False positives maintaining a good accuracy
- Model Selection: Tried SVM, Adaboost, Logistic Regression, Random Forest and Gradient Boosted Trees
- Feature Selection: Using Tree Classifier to reduce to 14 features
- Tuned each classifier for better results





RESULTS

Model	Accuracy	False Positives	AUC Score
Logistic Regression	30.06	1	0.5
Random Forest	98.22	0.91	0.987
Gradient Boosted Trees	98.45	0.71	0.986

- Gradient Boosted Trees better than Random Forest due to lower false positives
- Logistic Regression and other linear models not suitable for the data
- Ensemble approaches are give better results than other classifiers
- Very close to state of the art (Tek)

FUTURE WORK

- Use UNSW-NB15 data set for IDS
- Training: 175,341 samples; Testing set : 82,332 samples ; 49 features with multiple output classes
- Train deep neural networks(NN) to predict network attacks
- Use different architectures and parameters for NN
- Tuning and Analysis: Grid search, cross validation, randomisation of data set
- Evaluation metrics: Classification Accuracy, False Negative, AUC Score, ROC curve for further adjustments

REFERENCES

- Machine learning for malware detection: https://www.randhome.io/blog/2016/07/16/machine-learning-for-malware-detection/
- Towards Classification of Polymorphic Malware: https://www.blackhat.com/docs/webcastTowardsClassificationofPolymorphicMalware-Final.pdf
- M. Siddiqui, M. C. Wang, and J. Lee. Detecting trojans using data mining techniques. In D. M. A. Hussain, A. Q. K. Rajput, B. S. Chowdhry, and Q. Gee, editors, IMTIC, volume 20 of Communications in Computer and Information Science, pages 400-411
- M. G. Schultz, E. Eskin, E. Zadok, and S. J. Stolfo. Data mining methods for detection of new malicious executables. In Proceedings of the 2001 IEEE Symposium on Security and Privacy
- M. Z. Shafiq, S. M. Tabish, F. Mirza, and M. Farooq. Pe-miner: Mining structural information to detect malicious executables in realtime. In Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection
- Y. Ye, L. Chen, D. Wang, T. Li, Q. Jiang, and M. Zhao. Sbmds: an interpretable string based malware detection system using svm ensemble with bagging. Journal in Computer Virology
- Y. Ye, D. Wang, T. Li, and Ye. Imds: Intelligent malware detection system. In Proccedings of ACM International Conference on Knowlege Discovery and Data Mining (SIGKDD 2007