Assignment 3
MAT3120

Jordan Farrow (10653054)
Edith Cowan University, Joondalup
Dr. Johnny Lo

Table of Contents

1. Data Preparation

1.1 Data Cleaning

To ensure models accurately classify APT and non-APT attacks, cleaning of the dataset was performed. Entries with 99999 and less than 0 values were considered unrealistic in the context of network traffic and were excluded. Entries containing ‘???’ values also did not aid classification and were excluded. In addition, the ‘IP range trust score’ feature was excluded due to the excessive number of invalid values which would cause models to inaccurately interpret its impact on the attack types. Furthermore, the features ‘hits’, ‘attack source IP address count’, ‘average ping to attacking IP’ and ‘individual URLs requested’ underwent either square-root or log transformations to reduce the severity its outliers had on model predictions (Carrascosa, 2025; Lee, 2020; Osborne, 2002).

1.2. Combining Categories

By combining similar sub-categories together into a larger one, the modes efficiency increases as less parameters are considered. In addition, it makes interpreting model predictions easier as less parameters impact the final result. Finally, it could reduce overfitting, allowing the model to focus on generalised outcomes of one large sub-category , rather than minute details within various smaller sub-categories (Google, 2024). Table 1 notes the sub-categories which were combined into larger categories.

Table 1

List of Combined Sub-Categories

Categorical feature	Combined sub-categories	New sub-category
Source OS detected	Windows 10 , Windows Server 2008	Windows_All
Target Honey Port Server OS	Windows (Desktops) , Windows (Servers)	Windows _DeskServ
	Linux , MacOS (All)	MacOS_Linux

2. Model Tuning & Prediction Performance Analysis

2.1. Elastic Net-Regression Model

2.1.1. Hyperparameter Tuning

The elastic net-regression model underwent hyperparameter tuning through a grid search, a strategy which checked all possible alpha and lambda combinations to produce the most optimal model (Coroiu, 2016). To populate this grid, alpha values ranged from 0.1 to 0.9 and incremented by 0.1, with 100 lambda values ranging between 10^{-3} and 10^{3}. Within each combination, 10-fold cross-validation was used to apply each combination within 10 split data cases to determine the optimal alpha and lambda values to find the combination that provides the greatest accuracy across various data scenarios. Seen in Table 2, the model’ s optimal values include 0.8 alpha and 0.01873817 lambda providing a mean accuracy in the training data across its 10 k-folds of 80.59%.

Table 2

Highest Elastic-Net Regression Model Accuracies of Alpha Lambda Combinations

Alpha	Lambda	Accuracy (mean across 10 k-folds)
0.8	0.01873817	0.8059
0.9	0.01629751	0.8058
0.7	0.02154435	0.8058
0.9	0.01417474	0.8058
0.8	0.01629751	0.8057

2.1.2. Results

Shown in table 3, the model has a moderate capability in identifying APT known attacks, with 80.6% of APT cases identified correctly however, shows weaknesses, incorrectly identifying 19.4% of APT known attacks as otherwise. Similarly, the model also has a moderate capability in identifying non-APT attack, with 80.9% of cases correctly identified and 19.1% classified otherwise. With an overall accuracy of 80.7%, it indicates that either attack types will be incorrectly classified around 20% of the time. Due to the severity of APT attacks, it is recommended to seek another model which can better identify the attack types for instance, classification tree and bagging tree, which see a greater classification performance. If implemented, many non-APT attacks will be treated otherwise, causing extensive mitigation strategies to be implemented for an attack which does not require it. On the other hand, APT attacks incorrectly identified could be treated less severe, in turn providing greater success chances for attackers.

Table 3

Elastic Net-Regression Model Confusion Matrix

Predicted / Actual	Yes	No
Yes	51187 (80.6%)	12373 (19.1%)
No	12301 (19.4%)	52244 (80.9%)

2.2. Classification Tree

2.2.1. Hyperparameter Tuning

This model had been tuned using the grid search strategy to find the optimal complexity parameter (CP) value (Coroiu, 2016). A range of 15 CP values chosen by the R were trained using the 10-fold cross validation, providing each CP 10 data splits to be tested with. This number of CP values ensures the model is neither overfitted or underfitted, ensuring the most optimal APT classification model is chosen for classifying attack types (Hastie et al., 2009). According to table 4, a CP value of 0.001139287 was most optimal, providing the greatest mean accuracy (90.12%), in addition to the value which prevents the most overfitting.

Table 4

Highest Classification Tree Model Accuracies With Various Complexity Parameters

Complexity Parameter	Accuracy (mean across 10 k-folds)
0.001139287	0.9012
0.001176038	0.9008
0.001194414	0.9004
0.001359794	0.8993
0.001653804	0.8985

2.2.2. Results

This model has shown a strong capability in classifying APT attacks, with 91.1% of APT attacks being correctly identified in table 5. Despite having a higher percentage than the tuned elastic-net regression model, it still has weaknesses, with 8.9% of APT attacks also seen in the table being identified as false negatives. The model had also correctly classified 89.1% of non-APT attacks, with 10.9% being identified incorrectly. Between the two attack types, the accuracy in identifying either was 90.13%. This indicates that there is a 10% possibility that the model incorrectly identifies an attack.

Table 5

Classification Tree Model Confusion Matrix

Predicted / Actual	Yes	No
Yes	57868 (91.1%)	7020 (10.9%)
No	5620 (8.9%)	57597 (89.1%)

2.3. Bagging Tree

2.3.1. Hyperparameter Tuning

Tuning the bagging tree model involved the grid search strategy finding the optimal nbagg, cp and minsplit value combination by tuning each simultaneously in order to produce the lowest misclassification error (Coroiu, 2016). Each parameter was optimised using 3 different values. Nbagg used 25, 150 and 25. Complexity parameter used 0, 0.5 and 0.1. Finally, minsplit used 5, 20 and 5. The choices of these CP values ensures the model is neither overfitted or underfitted, ensuring the most optimal APT classification model is chosen for classifying attack types (Hastie et al., 2009). The optimal values for this model are noted in table 6 using a 150 nbagg, 0 CP and 15 minsplit, resulting in the lowest OOB misclassification error of 7.39%.

Table 6

Lowest Bagging Tree Model Misclassification Error With Different Parameter Combinations

nbagg	cp	minsplit	OOB misclassification error (%)
150	0	15	7.39
150	0	5	7.4
100	0	15	7.4
125	0	15	7.4
100	0	10	7.41

2.3.2. Results

Seen in table 7, the tuned model correctly classifies 92.3% of APT attacks, in addition to 92.7% of non-APT attacks, indicating a strong performance in classifying attack types. However, it is to be noted that 7.7% of APT defined attacks were miscategorised as false negatives, in addition to 7.3% of non-APT attacks as false positives, indicating there are weaknesses present within the model. The model's 92.48% overall accuracy further supports this, indicating the model's capability at classifying either attack types with only 7.52% of misclassifications occurring.

Table 7

Bagging Tree Model Confusion Matrix

Predicted / Actual	Yes	No
Yes	58588 (92.3%)	4736 (7.3%)
No	4900 (7.7%)	59881 (92.7%)

3. Recommended Model

The bagging tree model was chosen for APT classification due to having the highest capability in detecting APT attacks (92.3%), being 11.7% and 1.38% higher than the elastic-net regression (80.6%) and classification tree (91.1%) models respectively. Due to the severity of APT attacks and the negative impacts it can have on the targeted organisation, it was important to choose the model which had the least false negatives. In addition, the model has the highest percentage for correctly identifying non-APT attacks (92.7%), being 11.8% and 3.6% higher than the elastic-net regression and classification tree model respectively. A low false positive rate is important since it ensures less non-APT attacks will be treated as APT attacks, allowing resources to be given to those which require it most. Although the bagging tree also has the highest overall accuracy, it must be noted that 7.52% of either APT or non-APT attacks will be misclassified.

4. References

Carrascosa, I. P. (2025, March 11). *Dealing with Outliers: A Complete Guide*. KDnuggets. https://www.kdnuggets.com/dealing-with-outliers-a-complete-guide

Coroiu, A. M. (2016). Tuning model parameters through a Genetic Algorithm approach. *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*. https://doi.org/10.1109/iccp.2016.7737135

Google. (2024, October 9). *Overfitting*. https://developers.google.com/machine-learning/crash-course/overfitting/overfitting

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Lee D. K. (2020). Data transformation: a focus on the interpretation. *Korean journal of anesthesiology*, 73(6), 503–508. https://doi.org/10.4097/kja.20137

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, 16(2). https://openpublishing.library.umass.edu/pare/article/1455/galley/1406/view/