[Machine Learning and Data Visualisation - MAT3120.3](https://courses.ecu.edu.au/courses/46502)

### Assignment 2

### Report on the analysis and modelling of a dataset

Name: \*\*\*\*\*\*\*\*\*\*

Student ID: \*\*\*\*\*\*\*\*\*\*

Table of Contents

1. Executive Summary 1
2. Background 1
3. Scope and Goals 1
4. Part 1 – General Data Preparation and Cleaning 1
5. Part 2 – Compare the performances of different ML algorithms 2
6. Appendices 5

1. **Executive Summary**

The following report outlines the research into two machine learning algorithms which will be implemented into the findings of security incidents on the network traffic. The two machine learning algorithms integrated for the research includes the Elastic E-Net Regression and the Classification Model. The detailed process will be entailed in this report including the stages of data preparation and cleaning, performing data training and testing of the models along with detailed approach on hyperparameter tuning and comparison of the two models to one another.

1. **Background**

The case scenario involves the producing machine learning model in supporting the plug-in which could successfully detect accurate incident alerts, mainly the excessive false-positives and false-negatives.

1. **Scope and Goals**

The raw data file \*\*\*\*\*\*\*\*\*\* consisting of fields ranging \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* has been provided for utilizing in building machine learning models. The scope of the project is to develop detection algorithms that will provide the most accurate incident detection whereas the goal is to test and evaluate two machine learning algorithms to determine which supervised learning model is best for the task.

1. **Part 1 – General data preparation and Cleaning**
   1. Importing the MLData2023.csv into RStudio

The “\*\*\*\*\*\*\*\*\*\*.csv” is first inserted into RStudio for the purpose of data cleaning, data preparation, training and testing.

* 1. Data cleaning
     1. Clean the whole dataset as per feedback provided from Assignment 1

The data is cleaned by the removing \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* since this column significantly contains proportion of \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*, which could result in inaccurate outcomes when included in training. For categories \*\*\*\*\*\*\*\*\*\* and \*\*\*\*\*\*\*\*\*\*, the outliers having the values \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. Since indicating the outliers could bring out a better chance of receiving accurate outcomes.

* + 1. Filter the data to only include cases labelled with class = 0 or 1

After identifying the outliers from the given dataset, the main target field, ‘Class’ is filtered out \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

* + 1. iv. Merging of the values in categories

The ‘iOS, Linux(Unknown), and Other’ values are merged under a value of ‘Others’ whereas all the Windows values are merged into the value of ‘Windows\_All’. Similar action has been performed on the field of ‘Connection.State’ where the values ‘INVALID, NEW, RELATED’ are merged to be identified as ‘Others’. The merging of the categories involve the usage of the fct.collapse(.) function. Collapsing of the categories could reduce the number of distinct values in a feature by helping in simplifying of the model and reduce the complexity of the learning task. Fewer categories mean \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. Additionally, this could assist in enhancing \*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

v. Select only complete cases using na.omit

The rows containing ‘NA’ values which have been considered as outliers from subpart(i) and (ii) are removed from the dataset using the ‘na.omit(.) feature on RStudio. The removing of these outliers \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

1. **Part 2 – Compare the performances of different ML algorithms**
   1. Randomly select two supervised learning models
   2. Run the ML algorithms in R on two training set

The random selection of the supervised learning models applied in the report are “Logistic Elastic-Net Regression” and “Classification Tree”.

* 1. Perform hyperparameter tuning to optimise the models:

Elastic Net Regression model (unbalanced and balanced dataset)

To maximize the performance of the elastic net regression model, hyperparameter adjustment was done for both the balanced and unbalanced datasets. For the unbalanced dataset, a thorough grid search approach looked at \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. The ideal hyperparameters \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. Through this tuning approach, the model was able to maximize overall accuracy, precision, recall, and F1 score on the test set while achieving a small false positive rate (FPR) and false negative rate (FNR). Comparable steps were taken in the hyperparameter tuning process for the balanced dataset as well, leading to the optimal lambda value \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

It was discovered that by balancing the trade-off between FPR and FNR, these hyperparameters improved the model's \*\*\*\*\*\*\*\*\*\*, \*\*\*\*\*\*\*\*\*\*, \*\*\*\*\*\*\*\*\*\*, and \*\*\*\*\*\*\*\*\*\*. The elastic net regression model showed enhanced efficacy for classification tasks through methodical hyperparameter tweaking, customized to the features of each dataset, guaranteeing stable performance across different data distributions.

Classification Tree model (unbalanced dataset)

The hyperparameter tuning strategy for the Classification Tree model began with \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*to gauge predictive performance using metrics like accuracy and AUC. Subsequently, hyperparameter tuning is engaged via \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* to identify the optimal one. According to figure 1 in Appendix, the \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* is the best value since the \*\*\*\*\*\*\*\*\*\* is the highest as \*\*\*\*\*\*\*\*\*\* could be overfitting and complex. The table which contains the cross-validation(CV) results for various \*\*\*\*\*\*\*\*\*\* values is resulted in the Rcode file.

Classification Tree model (balanced dataset)

Similarly, for the balanced dataset, a comparable hyperparameter tuning process is performed \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* for optimal tuning. The \*\*\*\*\*\*\*\*\*\* value, \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* is considered as the best value for the balanced dataset.

* 1. Evaluating the predictive performance of two ML models (FPR, FNR, Accuracy,…)

Elastic Net Regression model with balanced and unbalanced data on testing set

The following table shows the summarized tested results of E-Net Regression model trained with balanced and unbalanced data

|  |  |
| --- | --- |
| Tested on unbalanced data | Tested on balanced data |
| Metric Value  1 False Positive Rate (FPR) \*\*\*\*\*\*\*\*\*\*  2 False Negative Rate (FNR) \*\*\*\*\*\*\*\*\*\*  3 Overall Accuracy (Test Set) \*\*\*\*\*\*\*\*\*\*  4 Precision \*\*\*\*\*\*\*\*\*\*  5 Recall \*\*\*\*\*\*\*\*\*\*  6 F1 Score (F-score) \*\*\*\*\*\*\*\*\*\* | Metric Value  1 False Positive Rate (FPR) \*\*\*\*\*\*\*\*\*\*  2 False Negative Rate (FNR) \*\*\*\*\*\*\*\*\*\*  3 Overall Accuracy (Test Set) \*\*\*\*\*\*\*\*\*\*  4 Precision \*\*\*\*\*\*\*\*\*\*  5 Recall \*\*\*\*\*\*\*\*\*\*  6 F1 Score (F-score) \*\*\*\*\*\*\*\*\*\* |
| Confusion Matrix  pred.class.elnet NonMal Mal  NonMal \*\*\*\*\* \*\*\*\*\*  Mal \*\*\*\*\* \*\*\*\*\* | Confusion Matrix  pred.class.elnet NonMal Mal  NonMal \*\*\*\*\* \*\*\*\*\*  Mal \*\*\*\*\* \*\*\*\*\* |

Table 1.Table containing testing results including confusion matrices of E-Net Regression trained data

There are obvious differences between the performance evaluation of two machine learning models trained on balanced and unbalanced samples. With a \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*, the unbalanced dataset model exhibits \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. It exhibits good categorization with high recall, precision, and F1 score. However, when compared to the data of the balanced dataset, it exhibits a FPR of \*\*\*\*\* and FNR of \*\*\*\*\* with an overall accuracy of \*\*\*\*\*. It performs well in recall with decent precision even if its total accuracy is deficient. These differences highlight the trade-offs between model performance metrics, emphasizing the importance of selecting the most suitable model based on project.

Classification model with balanced and unbalanced data on testing set

The following table shows the summarized tested results of Classification model trained with balanced and unbalanced data.

|  |  |
| --- | --- |
| Tested on unbalanced data | Tested on balanced data |
| Metric Value  1 False Positive Rate (FPR) \*\*\*\*\*  2 False Negative Rate (FNR) \*\*\*\*\*  3 Overall Accuracy (Test Set) \*\*\*\*\*  4 Precision \*\*\*\*\*  5 Recall \*\*\*\*\*  6 F-score \*\*\*\*\* | Metric Value  1 False Positive Rate (FPR) \*\*\*\*\*  2 False Negative Rate (FNR) \*\*\*\*\*  3 Overall Accuracy (Test Set) \*\*\*\*\*  4 Precision \*\*\*\*\*  5 Recall \*\*\*\*\*  6 F-score \*\*\*\*\* |
| Confusion Matrix    pred.class NonMal Mal  NonMal \*\*\*\*\* \*\*\*\*\*  Mal \*\*\*\*\* \*\*\*\*\* | Confusion Matrix  pred.class NonMal Mal  NonMal \*\*\*\*\* \*\*\*\*\*  Mal \*\*\*\*\* \*\*\*\*\* |

Table 2.Table containing testing results including confusion matrices of Classification trained data

When assessing the performance of machine learning models trained on both unbalanced and balanced datasets, noticeable differences in their predictive abilities have become apparent. The model trained on the unbalanced dataset shows \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. Despite this, it \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*, the model trained \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*.

* 1. Brief statement on final recommended model

The final recommendation leans toward the \*\*\*\*\* model applied to the unbalanced dataset \*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. In contrast to other model, \*\*\*\*\* achieves a significantly greater \*\*\*\*\* with less \*\*\*\*\*and \*\*\*\*\* while maintaining \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. The decision to favor \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* model on the unbalanced dataset is primarily driven by its superior \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. The preference for the \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*model over the unbalanced dataset stems from its notable \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. Additionally, \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*. In summary, the \*\*\*\*\* model on the \*\*\*\*\* dataset achieves a blend of \*\*\*\*\*and \*\*\*\*\*, making it the recommended option for this project.

Appendices

A graph of a function

Description automatically generated with medium confidenceA graph with a line and numbers

Description automatically generated

Figure 1. Results from classification tree model on unbalanced dataset

A graph with numbers and lines

Description automatically generatedA graph with blue dots and numbers

Description automatically generated

A diagram of a computer

Description automatically generatedFigure 2. Results from classification tree model on balanced dataset

Figure 3.Classification tree model on unbalanced dataset

A screenshot of a computer

Description automatically generated

Figure 4. Classification tree model on balanced dataset