MAT3120.3

Machine Learning and Data Visualisation

Assignment 2: Report on The Analysis and Modelling of a Dataset

---

Student Name:

Student ID:

Table of Contents

Part 1: Data Preparation and Cleaning

The integrity and reliability of machine-learning models depend significantly on the quality of the input data. Therefore, thorough data preparation and cleaning are indispensable steps in the analysis of the ******* dataset, aimed at detecting ***************.

Data Cleaning Steps

Invalid and empty values were addressed to maintain data accuracy. Records with ************ ************ were removed because *****************. Entries with ************* were considered invalid and thus excluded. These corrections are essential for ******************** ***********, especially when *** *************************. The dataset was filtered to ******************************** **. This binary classification is central to the supervised learning approach, which focuses on the analysis of the crucial task of incident detection.

Category Simplification

The dataset was streamlined by *** ***************************************. Specifically: ******************************** ******************** were consolidated ***********************, whereas ************** *** **. This step aims to reduce ***** *****************, aiding the learning process of the model ***************************** ***************************. ** ***********************************. This simplification potentially improves the model efficiency by ** ***.

Part 2: Model Training and Hyperparameter Tuning

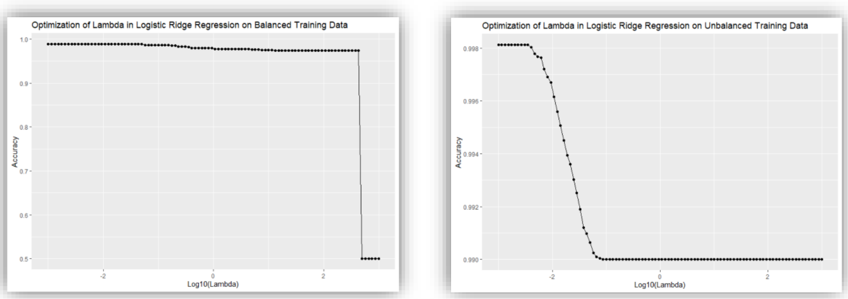

Figure 1&2: Plot of Optimization of Lambda for both Balanced and Unbalanced Training Dataset

已註解 [JL1]: Placement of this figure in the section is questionable. Generally, a figure should come after the text so that it can be placed in context.

Hyperparameter Tuning/Search Strategy for Logistic Ridge Regression

Logistic Ridge Regression models, applied to both balanced and unbalanced datasets, underwent a rigorous process of hyperparameter tuning to ascertain the optimal configuration for detecting malicious incidents. This endeavor was crucial for ** **. For the balanced dataset, the tuning focused on ** ***. The optimal ************* value is ******************. This precise calibration of ************ significantly bolstered the model's accuracy, achieving a notable accuracy rate of ************** and a kappa statistic of *********, indicating the robustness of the model in differentiating between *************************** events. In contrast, the unbalanced dataset underwent an exhaustive hyperparameter tuning process, revealing an optimal ******************. This fine-tuning resulted in an even higher accuracy of ************* and a kappa statistic of *************, showing the model's ******** ********************* to the presence of malicious activities.

已註解 [JL2]: It's not clear if you're referring to the accuracy in the training stage, or on the test set. This section is on the tuning of the model so the latter should not be relevant just yet.

已註解 [JL3]: Define this metric since you've provided it here.

已註解 [JL4]: What do you mean by exhaustive? Is it the same as the balanced dataset?

Prediction Results from Balanced Training Model

	Non-Malicious	Malicious
Non-Malicious	98.75% (464163)	1.09% (5132)
Malicious	0.01% (56)	0.57% (2685)

FNR	****
FPR	****
Balanced Accuracy	****
Precision	****
Recall	****
F_Score	****

Table 1&2: Confusion Matrix & Results from Balanced Training Dataset

已註解 [JL5]: Same comment as before. Introduce the table with some prior text.

Also, the % in the 1st table is not correct. Your calculation is based on the total N, and not the column total.

The metrics should also be defined beforehand.

The model demonstrated a high capability in identifying *************, correctly classifying ****** of such cases. However, it shows a vulnerability in detecting **********, mislabelling ***** of them as ********. The false positive rate was ********, which indicates that a relatively small number of *************** were incorrectly identified as ************. A false negative rate of ****** points to a small proportion of ************************. Notably, the precision of the model was **************, reflecting strong accuracy in predicting ****************. With a recall of ******, most malicious activities were successfully **********, and an F-score of ***** indicated a well balanced model. A balanced accuracy rate of ********* underscores the overall efficacy of the model.

Prediction Results from Unbalanced Training Model

	Non-Malicious	Malicious
Non-Malicious	99.42% (469289)	0.12% (547)
Malicious	0.00% (6)	0.46% (2194)

FNR	****
FPR	****
Balanced Accuracy	****
Precision	****
Recall	****
F_Score	****

Table 1&2: Confusion Matrix & Results from Unbalanced Training Dataset

已註解 [JL6]: Same issue as before.

Also, these should be Tables 3 & 4.

For the unbalanced dataset, the model classified **.

Hyperparameter Tuning/Search Strategy for Random Forest Models

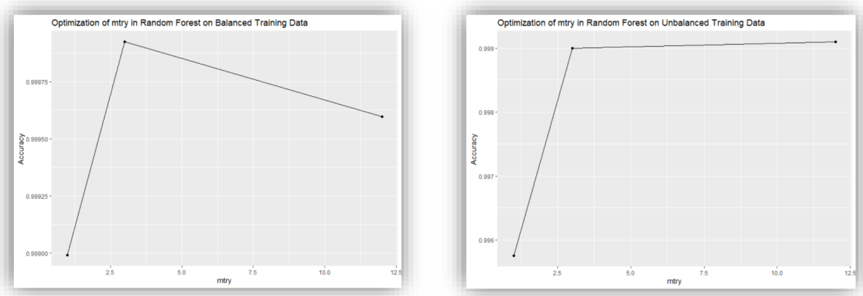

Figure 3&4: Plot of Optimization of mtry for both Balanced and Unbalanced Training Dataset

Tuning Methodology

A structured exploration of hyperparameters such as mtry, splitrule, and min.node.size was performed. For the balanced dataset, the optimal performance was obtained with mtry=3, using the Gini split rule, and setting the min.node.size to 5. In contrast, the unbalanced dataset showed optimal results with mtry=12, the same split rule, and node size, thus enhancing the model's detection capabilities for malicious incidents (Reference A; Reference B).

Performance Evaluation

The balanced dataset model achieved ******** accuracy, demonstrating **********************. The unbalanced model surpassed this, achieving ******** with exceptional ************* and specificity **************, underscoring its robustness in ************************.

Prediction Results from Balanced Training Model

	Non-Malicious	Malicious
Non-Malicious	99.42% (469194)	0.05% (219)
Malicious	0.02% (101)	0.53% (2522)

FNR	****
FPR	****
Balanced Accuracy	****
Precision	****
Recall	****
F_Score	****

Table 5&6: Confusion Matrix & Results from Balanced Training Dataset

已註解 [JL7]: You were asked to optimise more than just the mtry hyperparameter.

Also, where is the interpretation for these plots?

已註解 [JL8]: How did you come to this conclusion? What about the other hyperparameters?

已註解 [JL9]: Which is?

已註解 [JL10]: Same problem as before with the confusion matrix.

The Random Forest model trained on the balanced model demonstrated ********************** *******************. It successfully identified ** ***. The sensitivity of the model was ** ***. Precision is ************************** ***. Coupled with ************************ *******************, the model reliably captured *******************************. The F-Score of ***. The balanced accuracy rate of **************** underscores the overall effectiveness of the model in correctly classifying both classes of events.

已註解 [JL11]: Same here. It would help if all the metrics were defined beforehand.

Prediction Results from Unbalanced Training Model

	Non-Malicious	Malicious
Non-Malicious	99.42% (469201)	0.03% (148)
Malicious	0.02% (94)	0.55% (2593)

FNR	****
FPR	****
Balanced Accuracy	****
Precision	****
Recall	****
F_Score	****

Table 7&8: Confusion Matrix & Results from Unbalanced Training Dataset

已註解 [JL12]: See above comment

For the unbalanced dataset model, the Random Forest model shows ************************** *** *** *** *** *** ***

Recommended Model and Conclusion

The chosen model for incident detection was the *************** model trained on the balanced dataset, primarily for its high ********, ************, and low ***********. Its F-score of ************ indicate a superior balance between recall and precision. Despite the comparative complexity of ***************, its performance and generalizability make it a pragmatic choice ******************, particularly in scenarios where missing a malicious event is highly detrimental. The trade-off in ***************** is deemed acceptable because of the significant gain in the predictive accuracy.

References

****

****