VYTAUTO DIDŽIOJO UNIVERSITETAS
Informatikos fakultetas

TO MERGE OR NOT TO MERGE DATASETS? WHAT DO THE EXPERIMENTS SHOW?

INTRODUCTION

"To be or not to be" is a legendary question. We face a similar question when we have limited and hard-to-access data. In this case, using synthetically generated data, we do not know whether the assumptions used in generating it are correct and similar in reality.

DATA:

This study uses three different synthetically generated, publicly available datasets of money laundering cases. 11 attributes were extracted from each dataset for training and testing machine learning methods. Mahootika dataset is referred to as the 1st dataset, AMLSim as the 2nd dataset and Paysim as the 3rd dataset.

MONEY LAUNDERING DETECTION:

Steps to detect money laundering in this research:

1) Creating additional attributes;
2) Selecting attributes;
3) Standardizing data;
4) Splitting training and testing subsets;
5) Balancing data;
6) Training machine learning methods;
7) Testing models;
8) Evaluating results;
9) Merging datasets.

MODELS:

- Linear Regression;
- Random Forrest;
- XGBoost;
- Support Vector Machine;
- Isolation Forrest;
- Ensemble.

EVALUATION METRICS:

- Accuracy;
- Sensitivity;
- Specificity;
- F1;
- AUC.

AUTHORS:

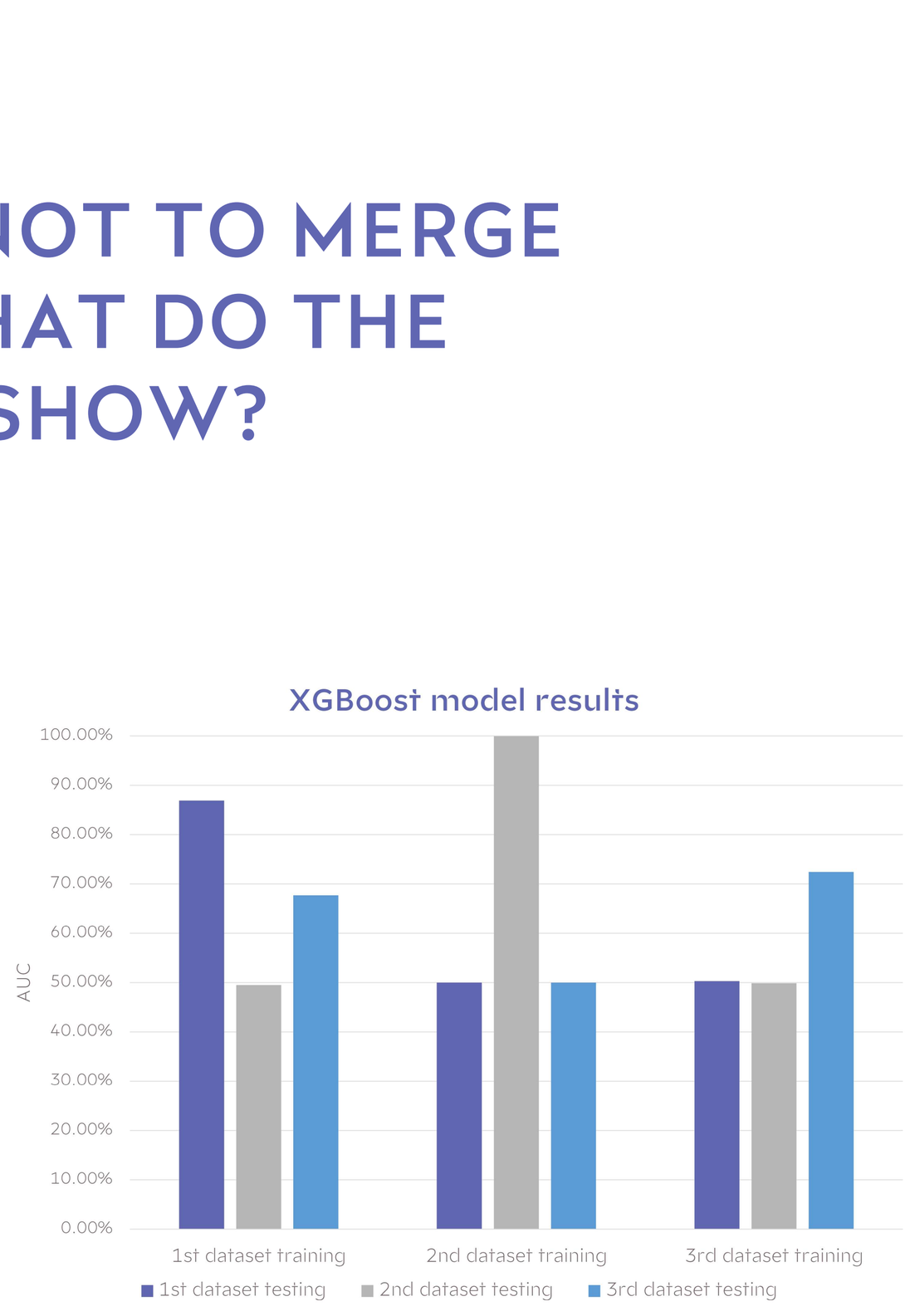
Paulius Savickas
paulius.savickas@vdu.lt

Dovilė Kuizinienė
dovile.kuiziniene@vdu.lt

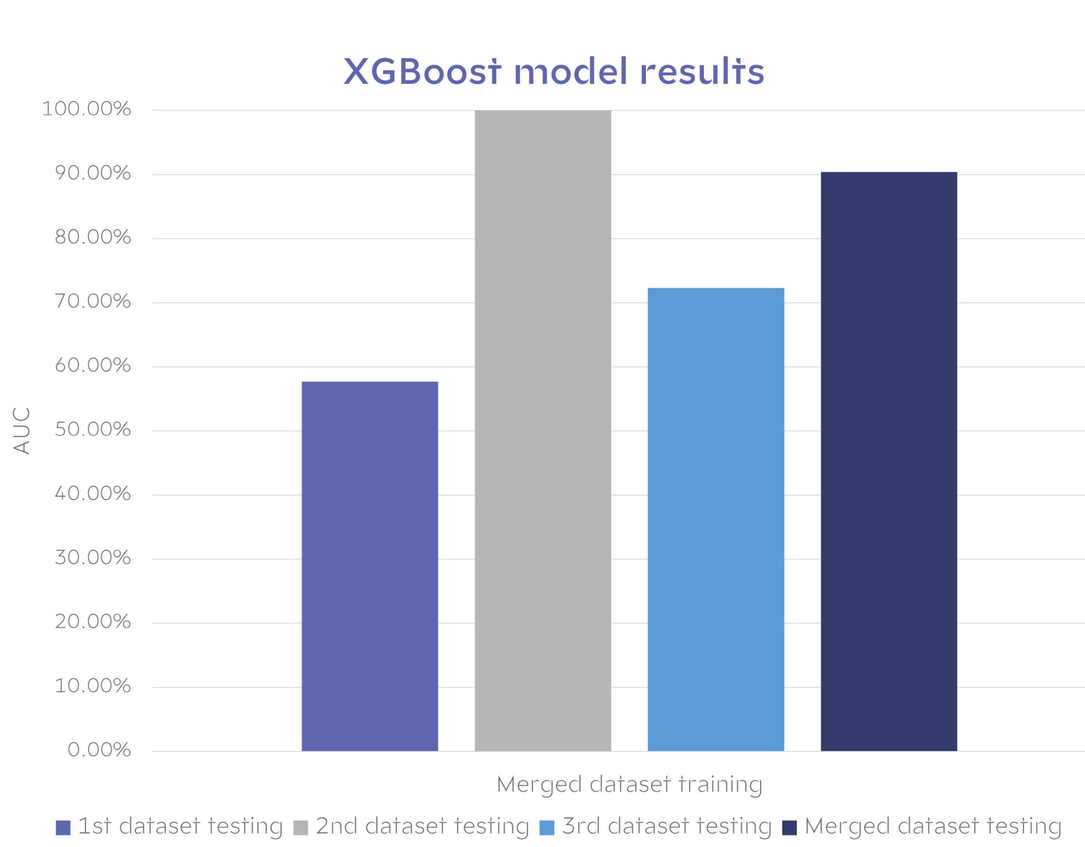
Tomas Krilavičius
tomas.krilavicius@vdu.lt

CARD
CENTRE FOR APPLIED RESEARCH AND DEVELOPMENT

XGBoost model results



Methods were trained using 70% of the datasets and tested on the remaining 30% of the same dataset and 100% of other datasets.



RESULTS:

After training the models on one dataset, the methods performed well when tested on the rest of the same set, but when tested on other sets, the models classified most of the values into one class, and the AUC score was ~50%. This showed that the datasets are generated based on different assumptions and cannot be verified. When the datasets were merged, and the models were tested on other datasets as well as the remaining test sample from the merged dataset, they performed more accurately. XgBoost correctly identified 84.4% of all money laundering cases and 96.3% of legitimate payments.

Importance of attributes

Importance	Random Forest	Linear regression	XGBoost
1	Amount CV	Action count	Amount CV
2	Max amount	Time difference	Same amount
3	Same amount	Same amount	Max amount
4	Amount	Same amount time difference	Min amount
5	Median amount	Min amount	Same amount time difference