The Analysis of Synthetic Data Application Peculiarities on Time-Series Forecast Model Selection

Rokas Štrimaitis, Simona Ramanauskaitė, Pavel Stefanovič

Introduction

Time-Series analysis and forecast falls into the artificial intelligence (AI) model area, where constant model adjustment is needed. While concept shift in classification tasks is relevant too, in most of the cases time-series concept shift is faster than in other AI areas. This require additional data analyst work on systematic time-series forecast model tuning or some automated model tuning must be done. As well in some areas (for example accounting, collaboration with different partners and their data forecasting) the variety of time-series data is so high, manual development of data models becomes not an option. Therefore, foundational models for time-series forecasting are developed. In our previous research we investigated possibilities to automate time-series forecasting model selection. The results and similar research papers indicate this task is feasible. At the same time the foundational time-series forecasting model achieved forecast error rate has place to improve. In this research we analyse the effectiveness of synthetically generated data application for more accurate time-series forecasting model selection. The obtained results allow to estimate the synthetic data application peculiarities, highlighting its benefits and potential misuse cases.

Questions:

1. How utilization of synthetic data affects the data forecasting prediction method selection accuracy?
2. does data forecasting prediction method selection model affect the accuracy of the forecast?

VILNIUS TECH
Faculty of Fundamental Sciences

Methods

Dataset:

The **true data** was collected from the accounting system. The company collaborates with multiple clients and stores multiple data fields. For the experiment, the sum of purchase invoices of each client were selected as cases for forecasting. To balance the dataset, “single time” clients were removed, resulting in 424 client records with at least 12 monthly records.

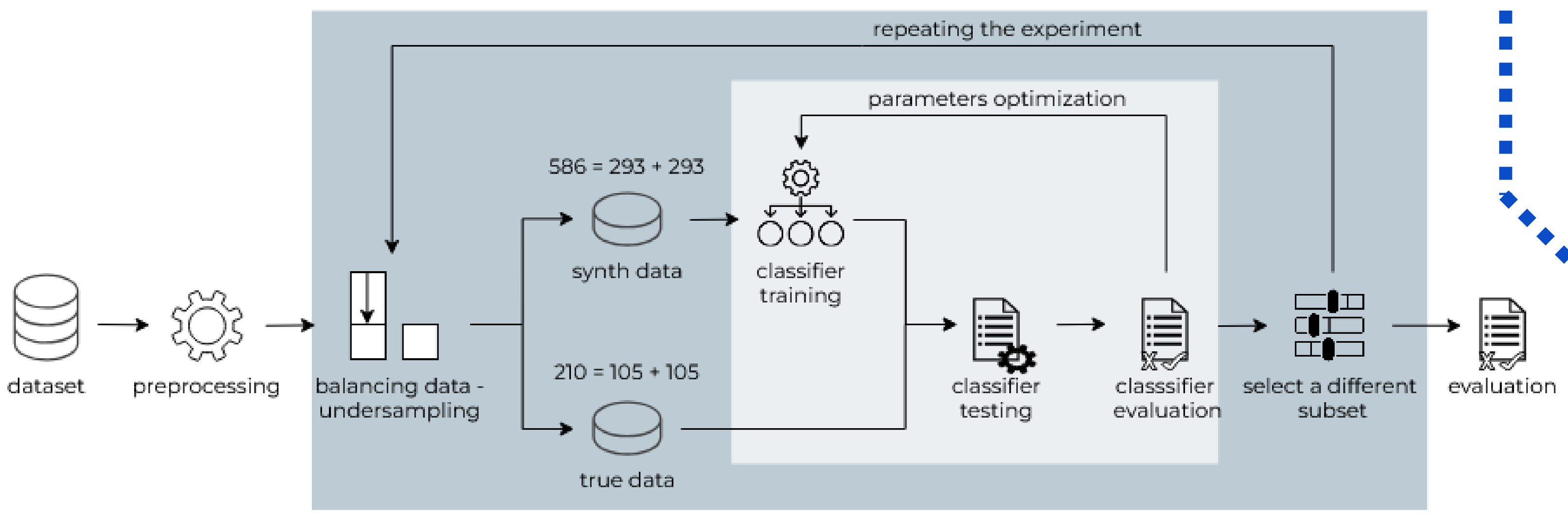
To generate the **synthetic data**, 4 different strategies were chosen, each generating 60-300 records, each including red noise. This resulted in a synthetic dataset of 900 time series.

To prepare the **dataset** for the model that determines the most suitable time series forecasting model, the records of each client were processed by defining their features (mean, min. and max. values, std. dev., quartiles, p-value, correlation coefficient,...). Each client and the synthetic time series data were trained and tested with SARIMA, LSTM and DLT forecasting models. The best model was selected as the class for the record.

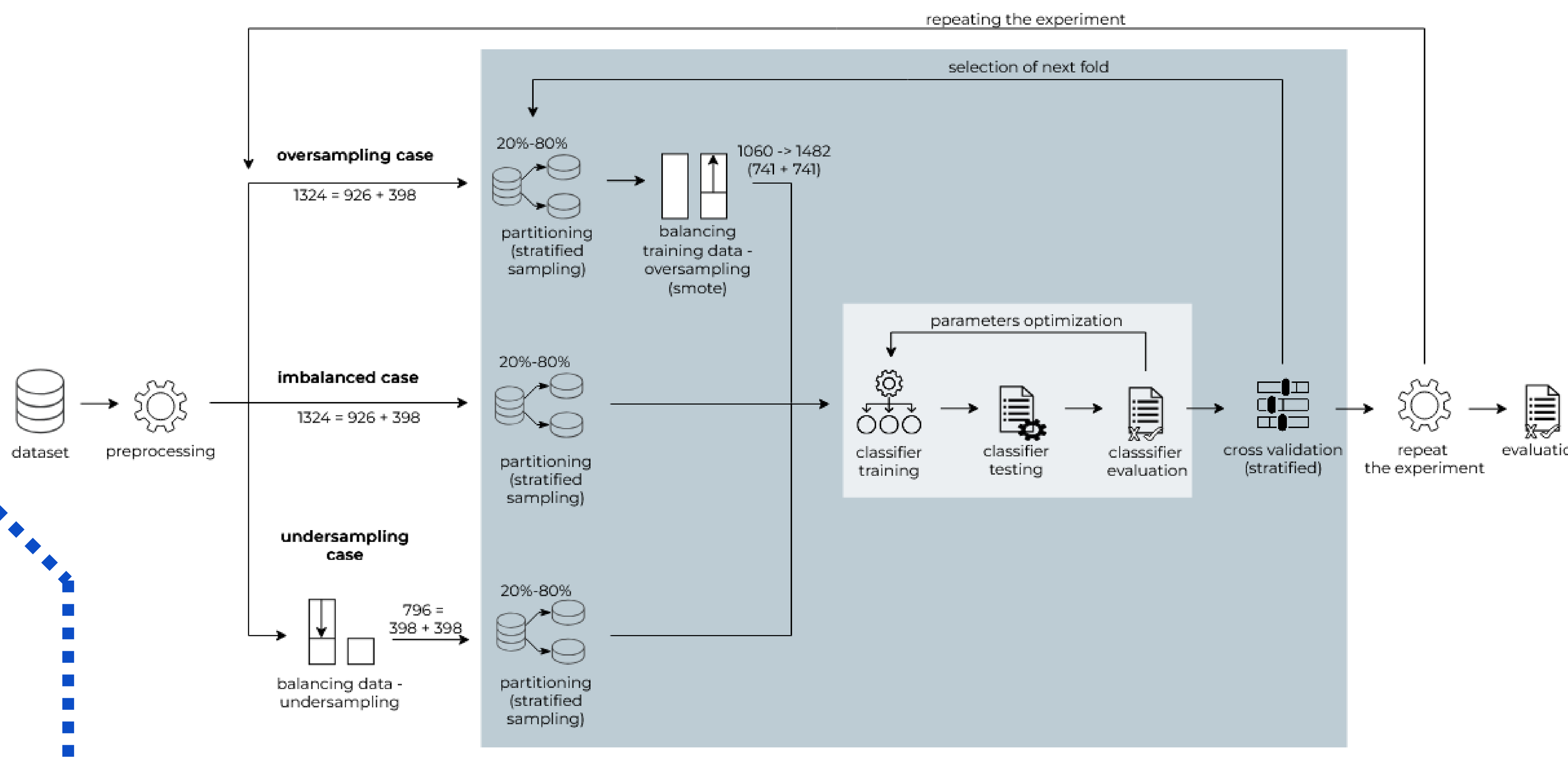
For **classification model** development, cross validation, TBPE and hill climbing methods were used to find how accurate the time-series forecast model can be estimated, based on the time-series defining properties, not the data itself.

	True	Synth	Sum
LSTM	319	607	926
SARIMA	105	293	398

Synthetic data training, true data testing



Synthetic + true data training and testing



Results

Models	Avg. acc.	Best accuracy		
		LSTM	SARIMA	Parameters
k-NN	0.57	0.77	0.39	k =7
SVM	0.56	0.85	0.32	c = 87.027
MLP	0.59	0.80	0.42	22 iter, 4 layers, 17 neurons
PNN	0.56	0.88	0.29	theta plus 0.920, theta minus 0.372
random forest	0.61	0.80	0.49	199 trees, 17 tree depth
gradient boosted trees	0.58	0.69	0.49	173 trees, 5 tree depth

	Models	Avg. acc.	Best accuracy		
			LSTM	SARIMA	Parameters
Imbalanced	k-NN	0.76	0.88	0.49	k = 5
	SVM	0.77	0.89	0.48	c = 19.982
	MLP	0.78	0.89	0.52	171 iter, 4 layers, 48 neurons
	PNN	0.78	0.95	0.37	theta plus 0.661, theta minus 0.107
	random forest	0.79	0.91	0.49	139 trees, 27 tree depth
	gradient boosted trees	0.77	0.89	0.49	54 trees, 6 tree depth
Oversampled	k-NN	0.71	0.71	0.68	k = 9, knn smote = 44
	SVM	0.75	0.98	0.19	c = 0.049, knn smote 50
	MLP	0.75	0.77	0.69	143 iter, 5 layers, 40 neurons, knn smote 5
	PNN	0.77	0.89	0.49	theta plus 0.817, theta minus 0.362, knn smote 24
	random forest	0.72	0.75	0.65	28 trees, 6 tree depth, knn smote 27
	gradient boosted trees	0.75	0.80	0.61	82 trees, 5 tree depth, knn smote 10
Undersampled	k-NN	0.68	0.70	0.71	k = 9
	SVM	0.64	0.69	0.64	c = 57.519
	MLP	0.73	0.74	0.73	95 iter, 1 layer, 67 neurons
	PNN	0.70	0.77	0.68	theta plus 0.427, theta minus 0.138
	random forest	0.74	0.78	0.72	187 trees, 22 tree depth
	gradient boosted trees	0.72	0.71	0.75	21 trees, 6 tree depth

Conclusions

Experiments have shown that training models only on synthetic data is not a suitable option for selecting the best model for forecasting time series data.

True data can improve the accuracy of the classifier. In the case of undersampling, the models showed the best accuracy.

Experiments on data sampling strategies show that the highest average accuracy can be achieved with an imbalanced dataset, but the accuracy of the classes becomes unbalanced as well. While undersampling reduces the average accuracy, also guarantees both classes will be equally reflected in the model.

Primary references:

1. F. Petropoulos, et al., “Forecasting: theory and practice”, International Journal of Forecasting, vol. 38, iss. 3, pp. 705–871, 2022.
2. W. Li; and J. Liao, “A comparative study on trend forecasting approach for stock price time series”, Proceedings of the 2017 11th IEEE International Conference on Anti-counterfeiting, Security, and Identification (ASID), 27-29 October 2017, Xiamen, China
3. S. Elsworth, and S. Guttel, “Time Series Forecasting Using LSTM Networks: A Symbolic Approach”.