ibai-publishing
Prof. Dr. Petra Perner
PF 30 11 38
04251 Leipzig, Germany
E-mail: info@ibai-publishing.org

P-ISSN 1864-9734
E-ISSN 2699-5220
ISBN 978-3-942952-83-5

www.ibai-publishing.org

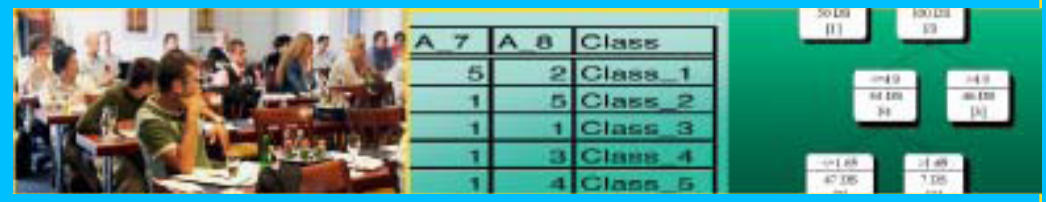
ISBN 978-3-942952-83-5

Advances in Data Mining, Poster Proceedings, ICDM 2021

Petra Perner

Petra Perner (Ed.)

Advances in Data Mining



20th Industrial Conference on Data Mining, ICDM 2021
New York, USA July 14-18, 2021

Poster Proceedings

ibai - publishing

Petra Perner (Ed.)

Advances in Data Mining

Applications and Theoretical Aspects

21th Industrial Conference, ICDM 2021
New York, USA, July 14 – 18 2021
Poster Proceedings

ibai Publishing
www.ibai-publishing.org

Volume Editor

Prof. Dr. Petra Perner
FutureLab Artificial Intelligence IBaI-2

PF 10128
01446 Radeberg
E-mail: pperner@futurelab-ai-ibai-2.de

The German National Library listed this publication in the German National Bibliography.
Detailed bibliographical data can be downloaded from http://dnb.ddb.de.

ibai-publishing
Prof. Dr. Petra Perner
PF 30 11 38
04251 Leipzig, Germany
E-mail: info@ibai-publishing.org
http://www.ibai-publishing.org

P-ISSN 1864-9734
E-ISSN 2699-5220
ISBN 978-3-942952-83-5

Preface

The twenty-one event of the Industrial Conference on Data Mining ICDM was held in New York again (www.data-mining-forum.de) running under the umbrella of the World Congress on "The Frontiers in Intelligent Data and Signal Analysis, DSA 2021" (www.worldcongressdsa.com).

After the peer-review process, we accepted twenty-four high-quality papers for oral presentation. Ten papers are presented in the proceedings by ibai-publishing. http://www.ibai-publishing.org/html/proceedings_2021/proceedings_icdm_2021.php The topics range from theoretical aspects of data mining to applications of data mining, such as in multimedia data, in marketing, in medicine and agriculture, and in process control, industry, and society. Extended versions of selected papers will appear in the international journal Transactions on Machine Learning and Data Mining (www.ibai-publishing.org/journal/mldm).

In all, six papers were selected for poster presentations and two are published in the ICDM Poster and Industry Proceeding by ibai-publishing (http://www.ibai-publishing.org/html/proceedings_2021/proceedings_poster_icdm_2021.php).

The acceptance rate was 20%. The list of accepted papers can be found after the conference under https://www.data-mining-forum.de/past_reports_ov.php. The Corona situation was still in our control after one and a half years.

The tutorial days rounded up the high quality of the conference. Researchers and practitioner`s got an excellent insight in the research and technology of the respective fields, the new trends and the open research problems that we like to study further.
A tutorial on Data Mining, a tutorial on Case-Based Reasoning, a tutorial on Intelligent Image Interpretation and Computer Vision in Medicine, Biotechnology, Chemistry and Food Industry, and a tutorial on Standardization in Immunofluorescence were held before and in between the conferences of DSA 2021.

We would like to thank all reviewers for their highly professional work and their effort in reviewing the papers.

We also thank the members of the Futurelab Artificial Intelligence IBaI-2, Radeberg, Germany (www.futurelab-ai-ibai-2.de), who handled the conference as secretariat. We appreciate the help and understanding of the editorial staff at ibai-publishing house, who supported the publication of these proceedings (http://www.ibai-publishing.org/html/proceeding.php).

Last, but not least, we wish to thank all the speakers and participants who contributed to the success of the conference. We hope to see you in 2022 in New York at the next World Congress on The Frontiers in Intelligent Data and Signal Analysis, DSA

2022 (www.worldcongressdsa.com), which combines under its roof the following three events: International Conferences Machine Learning and Data Mining, MLDM (www.mldm.de), the Industrial Conference on Data Mining, ICDM (www.data-mining-forum.de, and the International Conference on Mass Data Analysis of Signals and Images in Medicine, Biotechnology, Chemistry, Biometry, Security, Agriculture, Drug Discovery and Food Industry, MDA (www.mda-signals.de), the workshops and tutorials.

July 2021 Petra Perner

21th Industrial Conference on Data Mining
ICDM 2021

www.data-mining-forum.de

Chair

Petra Perner	FutureLab Artificial Intelligence IBaI-2,
.....	Germany

Program Committee

Ajith Abraham	Machine Intelligence Research Labs (MIR Labs), USA
Mohamed, Bourguessa	Universite du Quebec a Montreal - UQAM, Canada
Bernard Chen	University of Central Arkansas, USA
Antonio Dourado	University of Coimbra, Portugal
Jeroen de Bruin	Medical University of Vienna, Austria
Stefano Ferilli	University of Bari, Italy
Geert Gins	Glaxo Smith Kline, Belgium
Warwick Graco	ATO, Australia
Aleksandra Gruca	Silesian University of Technology, Poland
Pedro Isaias	Universidade Aberta (Portuguese Open University), Portugal
Piotr Jedrzejowicz	Gdynia Maritime University, Poland
Martti Juhola	University of Tampere, Finland
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Lui Xiaobing	Google Inc., USA
Mehmed Kantardzic	University of Louisville, USA
Eduardo F. Morales	INAOE, Ciencias Computacionales, Mexico
Samuel Noriega	Universitat de Barcelona, Spain
Juliane Perner	Cancer Research, Cambridge Institutes, UK
Moti Schneider	Netanya Academic College, Israel
Rainer Schmidt	University of Rostock, Germany
Victor Sheng	University of Central Arkansas, USA
Kaoru Shimada	Section of Medical Statistics, Fukuoka Dental College, Japan
Gero Szepannek	University Stralsund, Germany

Joao Miguel Costa Sousa	Technical University of Lisbon, Portugal
Markus Vattulainen	Tampere University, Finnland
Zhu Bing	Sichuan University, China

Table of Content

Variational-Autoencoder Architectures for Anomaly Detection in Industrial Processes

Felix Neubürger, Yasser Saeid, and Thomas Kopinski

South Westphalia, University of Applied Sciences, Lindenstr 53, 59872 Meschede, Germany neubuerger.felix@fh-swf.de saeid.yasser@fh-swf.de kopinski.thomas@fh-swf.de

Abstract. In this paper we describe the use of Variational-Autoencoder architectures for the unsupervised detection of anomalies in industrial processes. To this end we implement a Variational Long Short-Term Memory (LSTM) Autoencoder and a Convolutional Variational Attention Autoencoder with TensorFlow and TensorFlow-Probability and train it on a variation of the Tennessee-Eastman dataset. We then construct an anomaly score from the Variational Autoencoder's output making use of the Bayesian properties of the trained model. Using the anomaly score and the autoencoder output we then perform a simple binary classification in order to evaluate the improvement of our method over generic classifications. This model is benchmarked against supervised AutoML and a Convolutional Variational Attention Autoencoder. We find that Variational-LSTM-Autoencoder and Convolutional Variational Attention Autoencoder yield promising results in unsupervised anomaly detection and lay groundwork for more complex use cases in industry applications.

Keywords: Bayesian Deep Learning · Autoencoder · Anomaly Detection · Tennessee-Eastman · Attention · Predictive Maintenance · Machine Learning in Industry · Intelligent Control Systems.

1 Introduction

This work is embedded in the general field of predictive maintenance (PM). In PM a system of control is created that surveys the industrial process. A control system can infer a defined machine state. This machine state can be defined as normal or one of various faulty machine states. This information can be used to stop a system before a critical, possibly damaging, event occurs. Other control systems can also predict the future machine states. This prediction can be used to plan maintenance schedules efficiently. Since industry processes become more and more complex over the years data driven PM systems gain importance to achieve maximum efficiency. In this contribution we present complex autoencoder architectures for anomaly detection with the goal of utilizing them in predictive maintenance scenarios. Autoencoders are architectures for unsupervised learning capable of learning a representation of the training data [1].

Unsupervised learning in the context of anomaly detection can be used in predictive maintenance scenarios because anomalies are rare events. Utilizing autoencoders for anomaly detection mean they can be trained on the "normal state" of a system. When an anormal state occurs the autoencoder cannot reconstruct the input and a high deviation of output to input is to be expected. Anomaly metrics can then be constructed from this approach. Using Variational-Autoencoders an anomaly score can be constructed from the probabilities given by the learned probability distributions. In this paper we show how this methodology can be used for anomaly detection on a variation of the Tennessee-Eastman dataset [2]. The structure of this contribution is laid out as follows: First, we present an overview of existing methods in time series analysis and anomaly detection and contrast them with our approach in chapter 4. After the introduction we describe the used dataset. Finally we present the results of our analysis in chapter 5 and discuss the results in chapter 6.

2 Related Work

The detection of anomalies is well studied in computer science due to the enormous impact it has on industrial applications [3]. During the last years, various well-established approaches for time series analysis have been applied dating back almost 100 years [4]. They include Holt-Winters exponential [5] and linear-quadratic estimates [6] among others. Long short-term memory (LSTM) encoders [7] are used as a trend predictor of their predominant performance in collecting time-related data with long-term patterns. Recurrent Neural Networks (RNN) can be trained to be sensitive to noise and anomalies. These anomalies rarely occur in real-world applications in contrast to normal conditions. Also in real-world applications there might not be annotated data available. For this reason, the focus of research in the literature is on the unsupervised detection of anomalies. Since LSTMs are sensitive to noise and anomalies, this leads to the detection of instability performance. Therefore one can say that worse performance is observed when the time series are mixed with noise. Detection approaches found with Variational Autoencoders (VAE) are largely due to the different distribution between normal and abnormal states. Since the noise and anomalies also flow into the model training, these signals are unexpectedly reconstructed like normal signals [8]. To model time-series data with its temporal dependencies, we use an LSTM network [7], which is a type of RNN. An LSTM network can make use of long-term dependencies and avoid the vanishing gradient problem. Researchers have used LSTM networks for prediction in anomaly detection domains such as the following: radio anomaly detection and electroencephalogram (EEG) signal anomaly detection [9]. Malhorta et al. introduced an LSTM-based anomaly detector (LSTM-AD) that measures the distribution of prediction errors [10]. However, the method may not predict time-series under unpredictable external changes such as manual control and load on a machine. Alternatively, researchers have introduced RNN- and LSTM-based autoencoders for reconstruction-based anomaly detection [10]. In this paper, different complex

autoencoder architectures are embedded in a Bayesian framework and compared in the context of anomaly detection in an exemplary industrial process. We construct a network that generates representations of the data that can be used as a classification input. This work can be seen as groundwork for further research in the field of unsupervised anomaly detection with Variational Autoencoders (VAEs).

3 Description of the dataset

Twenty-first-century Industrial Control Systems (ICS) consist of sensors and other control systems that generate multivariate time series data. Tennessee Eastman is simulated data for anomaly detection in Industrial Control Systems. The Tennessee Eastman data is referenced by [2] and encompasses "faulty free" and "faulty" datasets. Each data frame contains 55 columns ('faultNumber', 'simulationRun', 'sample', the other columns contains the process variables). The variables is all numerical data and can be downloaded here [11]. This dataset is commonly used for benchmarking anomaly detection models. The complexity of this dataset is comparable to that found in industrial processes on a sensor level. In more complex control systems this work can be transferred to that specific case. A diagram of the chemical process is displayed in figure 1

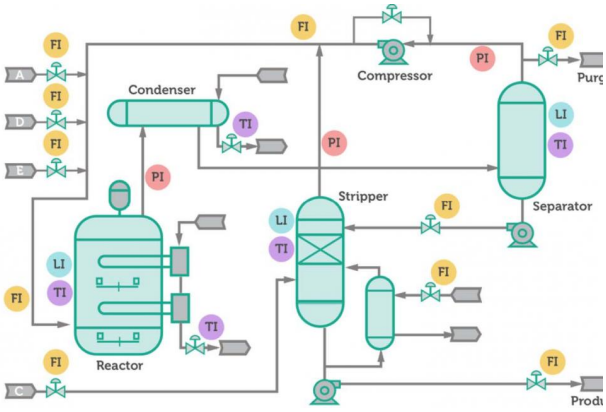


Fig. 1. Processdescription of the Tennessee-Eastman Process [11]

4 Methodology

The neural networks were implemented in python using the packages NumPy, Pandas, TensorFlow, scikit-learn and matplotlib. For the implementation of Bayesian neural networks the extension package tensorflow-probability was utilized. As a simple preprocessing step the data was first loaded and merged into

a single file. For the LSTM-VAE model the data has to be preprocessed so the model can interpret the temporal structure of the data. To this end, the data is extended by a dimension via a sliding window approach. For that each row in the data is shifted by one time step backwards, so that in each row the information of the k earlier data points is available. The dimensionality of the whole dataset is increased by k for a temporal lookback of k events. This preprocessing step increases the total physical memory needed for training the model. In real-time applications this could lead to problems depending on the computing capacities and length of lookback. For the other models used in this work this data augmentation is not needed which saves computing time and resources. Before being fed into the model a training/testing split is performed to evaluate the model's performance. For evaluation purposes the evaluation data is sampled as such that the anomaly:normal state ratio is 50:50. The training and testing data is normalized prior to being used in the model. Because the data is normalized a Gaussian prior can be used in the output nodes of the Bayesian neural network. Autoencoders belong to the family of unsupervised learning algorithms. The autoencoder is trained to learn to map the input back to itself. To this end, it is first mapped into a low-dimensional latent space in an encoder. From this low-dimensional space, it is mapped back to the original dimensionality with the help of the decoder. This property of the autoencoder is exploited in anomaly detection. Here, the normal state of a system is used to train the autoencoder. It can therefore be assumed that, given a sufficient amount of data, the autoencoder is able to map normal states of the system well onto itself. For the quantisation of this mapping, the mean squared error between input and output can be used, for example. Heuristically, an overtraining to the normal state of a system takes place. Overtraining can be explained as the model learning the given dataset by heart. This leads to the – for this approach positive – effect that it can be assumed that an anormal state cannot be reconstructed well and that a greater deviation between input and output is seen. The variational autoencoder is a method of Bayesian deep learning. Here, the architecture of the autoencoder is extended by the autoencoder learning posterior distributions of the attributes in the low-dimensional latent space and in the output nodes. For these distributions, an assumption must be made about the true distribution of the attributes. A diagram of this architecture is shown in 2. In this work, it is assumed that the distributions of the individual attributes and the distribution in the latent space are uncorrelated standard normal distributions. In more complex applications, a multivariate correlated distribution can also be assumed. When feeding a feature vector through the variational autoencoder the output is sampled out of the learned distribution in the output nodes. Repeating this inference step n times one can do statistical inference to compare the output to the input.

The autoencoders used in this work minimize the Evidence Lower BOund (ELBO) loss function.

$$\mathrm{ELBO}(x) = \int \mathrm{d}z q(z|x) \log p(x|z) + \int \mathrm{d}z q(z|x) \frac{q(z|x)}{p(z)}, \tag{1}$$

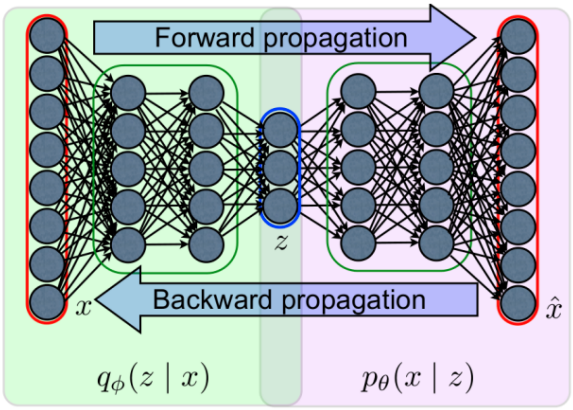


Fig. 2. Structure of a Variational Autoencoder [12]

with $p(z)$, the prior on the latent representation z, $q(z|x)$, the variational encoder, $p(x|z)$, the decoder — how likely is the feature vector x given the latent representation z. The first integral in the ELBO equation is called the reconstruction term. It asks how likely we are to start at a feature vector x, encode it to z, decode it, and get back the original x. It is parametrized as the log-likelihood in this case. The second term is the Kullback-Leibler-Divergence which is a measure for how different two probability distributions are [13] . As can be seen here, the exact mapping of the encoders and decoders is not fixed, therefore it is possible to use different neural network architectures. Usually the encoder and decoder networks are fully connected layers. In the case of image processing two-dimensional convolutional neural networks (2D-CNNs) are used [14]. In our contribution, we used a Variational-LSTM-Autoencoder to properly capture the causal effect of time series anomalies. The LSTM-Autoencoder functions analogously to the fully connected autoencoder. After the preprocessing of the input data the latent space representation is created by the LSTM-Encoder. From that latent space the LSTM-Decoder tries to reconstruct the input. It has been shown over the years that LSTM Networks perform better on sequential and time series data than fully connected networks. Since LSTM networks are resourceful to train and the preprocessing does not scale well with long lookbacks [15] and large amounts of data we use another encoder-decoder architecture. This architecture makes use of the attention mechanism in conjunction with one-dimensional convolution. This network architecture is shown to work very well on sequential data. [16] One advantage is the faster training process since the gradient does not have to be computed as often as in LSTM networks. Also the data can be fed into the network without any preprocessing because the one-dimensional convolution acts as the sliding window that has been used for the LSTM. For the comparison we construct an LSTM-Autoencoder and a CNN-Attention-Autoencoder with similar amount of trainable parameters. After the

training the model can be evaluated and an anomaly score is constructed. For the evaluation the held out dataset is passed through the VAE 10 times. From this the mean prediction together with the standard deviation of the outputs is obtained. It is expected that the VAE trained on normal state data is able to reconstruct the normal state data well and the mean squared error (MSE) [17] from input to output and the standard deviation are small. From that theorem we construct an anomaly score to be:

$$\mathrm{A}(x_{\mathrm{out}}, x_{\mathrm{in}}) = \frac{1}{n_{\mathrm{features}}} \sum_{k=1}^{n_{\mathrm{features}}} (x_{\mathrm{in},k} - \mu_{x_{\mathrm{out}},k})^2 \cdot \sigma_{x_{\mathrm{out}}}, \tag{2}$$

with $\mu_{x_{\mathrm{out}}}$ the mean of the predicted outputs and $\sigma_{x_{\mathrm{out}}}$ the standard deviation of the outputs. We expect this anomaly score to be a small value when the VAEs output resembles the input. Since the input is passed through the VAE multiple times the standard deviation of the outputs is computed. We expect a small standard deviation for normal state inputs because the model is trained to reconstruct these cases. In contrast we expect a higher MSE and standard deviation in anormal states because the VAE cannot properly reconstruct the input. To analyze the performance of our method compared to other algorithms, different types of neural networks are deployed. One of those methods is based on the Attention mechanism. Attention mechanisms attempt to mimic human perception, selectively focusing attention on parts of the target areas in order to gain more detail of the targets and suppress other useless information. The calculation of the weight coefficient is the focus of attention. There is a proportional relationship between weight and attention. That is, the weight represents the importance of information and the value is the relevant information [18]. A diagram of the Attention network architecture is shown in 3.

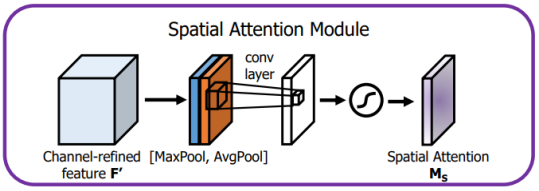


Fig. 3. Diagram of each attention sub-module [19].

Another benchmarking model is AutoML [20]. AutoML selects the most suitable architecture in an autonomous way without human intervention. We would like to discuss whether AutoML performs better on the time series dataset to allow a better understanding of AutoML in this new domain. Finally, we will use these experiences to discuss how future AutoML tools can improve the user experience for machine learning methods.

5 Experimental Results

In the evaluation of the results, the familiar labels *normal state* and anormal state are reintroduced. For the evaluation, all the different faults in the Tennessee-Eastman dataset are combined into the anormal state. The error states 3, 9 and 15 were excluded from the evaluation because they are very hard to detect without deeper analysis of the dataset [21]. The evaluation is thus based on a binary classification with only one variable. Here, the classification is done via a cut to the introduced anomaly score 2. If an input event is reconstructed with an anomaly score above that cut it is classified as an anormal event. If the anomaly score is calculated to be lower than that cut the state is classified as normal. Figures 4 and 5 show the Recall, Precision and F1-Scores for different cuts on the anomaly score for the binary classification. The metrics used are defined via the confusion matrix with the true positive tp, false positive fp, false negative fn and true negative tn as:

$$\text{precision} = \frac{tp}{tp + fp} \tag{3}$$

$$\text{recall} = \frac{tp}{tp + fn} \tag{4}$$

$$\text{F1} = \frac{2tp}{2tp + fp + fn} \tag{5}$$

$$\text{accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \tag{6}$$

It can be seen that different thresholds for the anomaly score lead to different evaluation metrics down the line. The shape of the precision, recall and F1-score curves is as expected. A local maximum for the F1-score can be found. The cut for further evaluation is made at the maximum of the F1-Score. This evaluation approach shows how powerful the anomaly score can be as an attribute for classification. The confusion matrices for a classification threshold of $\theta_A = 1.65$ for the different models can be seen in Figures 6 and 7. The confusion matrices display how many correct classifications (true positive and true negatives) occur in the evaluation dataset in conjunction with wrong classifications (false positives and false negatives). In the industrial process the confusion matrix can be used to estimate the cost of blind trust towards the model. The confusion matrix can be multiplied by a cost matrix containing the costs for false alarms (prediction break, real state normal) or unexpected failures (prediction normal, real state break). The classification threshold can then be optimized to minimize the running costs of the machine. In this work the cost of the cases is not known. In general however it can be assumed that the prediction of a normal state while there is a anormal state is very costly since critical damage can occur. False alarms are less of a problem but also cost maintenance and lead to machine downtime. At the chosen classification threshold both autoencoder models produce a false negative rate below 4% with the false negative rates being: $\text{FNR}_{\text{LSTM-VAE}} = 2.9\%$ for the LSTM-VAE and $\text{FNR}_{\text{CNN-Attention-VAE}} = 3.3\%$ for the CNN-Attention-VAE. Table 1 shows the final evaluation scores for straight cuts on the anomaly

score of $\theta_A = 1.65$ for a binary classification in *normal state* and *anormal state*. It can be seen that despite the simplicity of the binary classification via a cut on a single variable a high accuracy and precision can be obtained. The Area under ROC Curve (AUC) can be used to quantify the general expressivity of a classifier. The high scores show that the anomaly score is a useful feature for classification. The results for the autoencoders are compared to a supervised AutoML classifier. The supervised classifier naturally works better because it has been trained with more information. The representation of the data obtained by the autoencoder can however be used as input to other supervised models. This may improve the classification results of supervised anomaly detection models in the future.

Table 1. Evaluation scores for the different autoencoder model architectures in comparison to the supervised benchmark created with Keras AutoML.

$\theta_A = 1.65$	Accuracy	Precision	AUC
LSTM-VAE	0.9665	0.9615	0.9903
CNN-Attention-VAE	0.9647	0.9645	0.9902
AutoML	**0.9814**	**0.8339**	**0.9980**

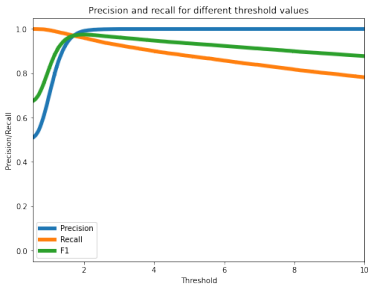


Fig. 4. Precision, Recall and F1-Score for the binary classification with a straight cut on the anomaly score for the LSTM-VAE.

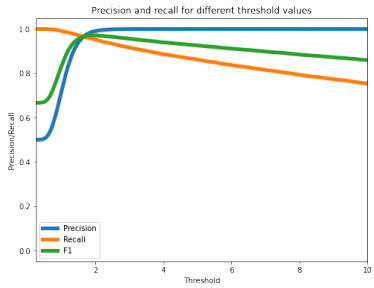


Fig. 5. Precision, Recall and F1-Score for the binary classification with a straight cut on the anomaly score for the CNN-Attention-VAE.

The results obtained in this study show that methods of unsupervised learning, when combined with bayesian deep learning, can lead to useful representations of data. These techniques will be further used in other complex industrial research processes. This work lays the groundwork for the usage in the upcoming data analysis tasks in the WiTraPres project [22]. The results obtained from a strictly data driven analysis will be combined with expert knowledge to gain more insights as presented in [23].

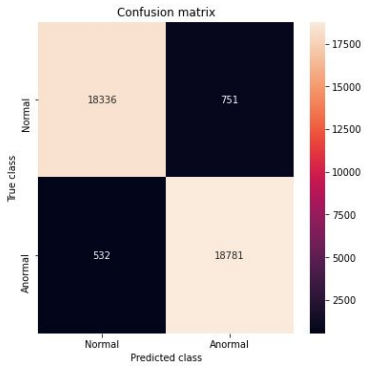


Fig. 6. Confusion Matrix for the LSTM-VAE with a straight cut on the anomaly score at $\theta_A = 1.65$

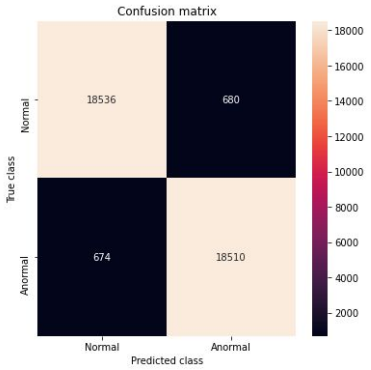


Fig. 7. Confusion Matrix for the CNN-Attention-VAE with a straight cut on the anomaly score at $\theta_A = 1.65$

6 Conclusion

The results obtained in this work show that more complex variational autoencoder architectures can successfully be used to generate useful representations time series data. As discussed in Chapter 5 the inherent probability measures of the variational autoencoder can be used to generate powerful features for simple binary classification or further supervised classification and regression tasks. The techniques used here can be easily applied to other use cases in anomaly detection and predictive maintenance. The representation of time series data can also be used in combination with other prediction methods to investigate complex industrial processes with heterogeneous data sources. We benchmarked unsupervised autoencoder models against a supervised AutoML framework with the introduction of labelled data after the training process. The main advantage of this approach however is the transferability of the methods. The methods can be quickly applied to processes where either no labels are available or where no detailed information about faults is known. The results from this purely data driven approach can in practice be used to work out fault scenarios with engineers.

References

1. Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. _eprint: https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690370209.
2. Cory A. Rieth, Ben D. Amsel, Randy Tran, and Maia B. Cook. Additional Tennessee Eastman Process Simulation Data for Anomaly Detection Evaluation. 2017.
3. Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *CoRR*, abs/1901.03407, 2019.

4. Nikolaos Kourentzes, Fotios Petropoulos, and Juan Trapero. Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, 30:291–302, 04 2014.
5. Prajakta Kalekar. Time series forecasting using holt-winters exponential smoothing. *Time Series Forecasting Using Holt-Winters Exponential Smoothing*, 01 2004.
6. Andrew C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1990.
7. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
8. Run-Qing Chen, Guang-Hui Shi, Wan-Lei Zhao, and Chang-Hui Liang. Sequential VAE-LSTM for anomaly detection on time series. *CoRR*, abs/1910.03818, 2019.
9. Tolga Ergen and Suleyman Serdar Kozat. Unsupervised anomaly detection with lstm neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):3127–3141, Aug 2020.
10. Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. *CoRR*, abs/1607.00148, 2016.
11. Xiaolu Chen. Tennessee eastman simulation dataset, 2019.
12. Jinwon An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015.
13. S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.
14. Keiron O'Shea and Ryan Nash. An Introduction to Convolutional Neural Networks. *arXiv:1511.08458 [cs]*, December 2015. arXiv: 1511.08458.
15. Fakultit Informatik, Y. Bengio, Paolo Frasconi, and Jfirgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. *A Field Guide to Dynamical Recurrent Neural Networks*, 03 2003.
16. Shun-Yao Shih, Fan-Keng Sun, and Hung-Yi Lee. Temporal pattern attention for multivariate time series forecasting. *CoRR*, abs/1809.04206, 2018.
17. Mean Squared Error. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 653–653. Springer US, Boston, MA, 2010.
18. Yang Liu, Lixin Ji, Ruiyang Huang, Tuosiyu Ming, and Chao Gao. An attention-gated convolutional neural network for sentence classification. *CoRR*, abs/1808.07325, 2018.
19. Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. *CoRR*, abs/1807.06521, 2018.
20. Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *CoRR*, abs/1908.00709, 2019.
21. M. Grbovic, W. Li, N. A. Subrahmanya, A. K. Usadi, and S. Vucetic. Cold Start Approach for Data-Driven Fault Detection. *IEEE Transactions on Industrial Informatics*, 9(4):2264–2273, November 2013. Conference Name: IEEE Transactions on Industrial Informatics.
22. Prof. Dr. M. Hermes Prof. Dr. T. Kopinski, Prof. Dr. A. Schwung. Witrapres.
23. Fernando Arévalo, Cristhian Tito, Mochammad Rizky Diprasetya, and Andreas Schwung. *Fault Detection Assessment using an extended FMEA and a Rule-based Expert System*. July 2019. Pages: 745.

Enhancing User's Income Estimation with Super-App Alternative Data

Gabriel Suárez[1], Juan Raful[1], Maria A. Luque[1], Carlos F. Valencia[2], and Alejandro Correa-Bahnsen[1]

[1] Rappi, Cl. 93 #19-58, Bogotá, Colombia.
[2] Industrial Engineering, University of Los Andes, Cra 1 Este 19-40, Bogota, Colombia

Abstract. This paper presents the advantages of alternative data from Super-Apps to enhance user's income estimation models. It compares the performance of these alternative data sources with the performance of industry-accepted bureau income estimators that takes into account only financial system information; successfully showing that the alternative data manage to capture information that bureau income estimators do not. By implementing the TreeSHAP method for Stochastic Gradient Boosting Interpretation, this paper highlights which of the customer's behavioral and transactional patterns within a Super-App have a stronger predictive power when estimating user's income. Ultimately, this paper shows the incentive for financial institutions to seek to incorporate alternative data into constructing their risk profiles.

Keywords: Fintech · Income Estimate · Alternative Data · Credit Risk.

1 Introduction

Technology-based companies are disrupting traditional business lines while they show an unprecedented ability to capture, manage, and analyze large volumes of customer data, previously unavailable to the organizations that traditionally operated in such lines, and therefore becoming increasingly competitive within them [2,11,4]. Given this, the business model of Super-Apps has emerged. Those are, mobile applications that seek to satisfy a various number of customer's daily needs, all within the same marketplace-based application. Whereas traditionally, users could only found those different services and solutions in a mobile application specifically designed for providing each one of them, like Uber for transportation and Uber Eats for delivery services.

One of the traditional sectors Super-Apps is venturing into is the financial sector, and so, it is logical then to raise the question on whether this Super-App data adds value within this field. This publication will explore how this alternative data might help financial institutions address one of the fundamental questions in credit risk management; that is, how much it should be lent to a customer, by building the most accurate user income estimate. This paper will then try to determine the following research questions:

1. Is there a significant improvement in income estimation statistical performance when using alternative data sources, retrieved from a Super-App, compared to industry-accepted Income Estimates?
2. What behaviors these Super-App features reveal, and how do they differ from traditional financial information resources?
3. Which of these behaviors appear to offer more predictive power?

2 Credit Risk and Income Estimation

Financial Institutions have to effectively manage credit risk to lend money to their customers, being this the potential that their counter party will fail to meet its obligations under agreed terms [1]. Hence, the challenge for these companies relies on the fact that each customer has a different probability of failing his particular obligations (PD) (as this is the reflection of each customer individual psychological characteristics [12] together with its financial knowledge, demographic characteristics and situational factors [5]) , as well as an individualized credit loss amount if the debtor of the loan defaults or so called loss given default (LGD) [12,5,1,10]. Therefore, financial institutions have to answer two fundamental questions for their accurate credit risk management, being one to which customers should these institutions lend money to and the second one, how much money should they lend to them. To address the second question, it is necessary for financial institutions to estimate each of its customers' payment capability. Properly calculating the amount of money they should lend to its customers allows financial institutions to accurately estimate their exposure at default (EAD) [1], which ultimately leads to organizations to operate within their desired expected loss (EL) in correspondence with their risk appetite. $EL = PD * LGD * EAD$.

Traditionally, financial institutions take user's income as a starting point to answer this second question. Nevertheless, the willingness of consumers to provide this information is relatively low, as user's income, being part of the user's financial information, is the most sensitive data to provide from their personal identifiable information [6]. This, leads financial institutions to try estimate user's income from their available financial information, and in some cases having to require it from third party institutions that gather this type of data. This estimate is, for some organizations, the best proxy for knowing this relevant user feature. However, having an income estimator that relies absolutely on financial information marginalizes the opportunity to have a complete profile of those users who do not have information within the financial system.

3 Users Interactions with Super-Apps

Alternative Data sourced from a Super-App can be retrieved from the various ways users of the Super-App interact and navigate through it. Moreover,

each transaction carried out within them different solutions (Delivery of restaurants/groceries/goods, transportation, travel, e-commerce and many more) generates different types of data. The features that can be collected then variate in their values in unique ways for each user as they reflect their individual behavioral and consumption patterns, and can be grouped into four categories:

Personal Information: The identifiable demographic attributes of the user such as age, place of residence, the brand of cell phone as well as personally identifiable information such as wealth estimators or address from where the orders where placed.

Consumption Patterns: User's consumption patterns are retrieved from the delivery vertical within the Super-App, which are all the services and solutions related to the purchase and delivery of groceries, food, clothing, technology, pharmaceutical products, and others.

Payment Information: This is the information that can be retrieved from all of the transactions the user makes in the Super-App. The features that can be extracted from this type of interaction allow identifying user's favorite payment method, tendencies of installments when paying regular orders versus more expensive orders, number of times a credit card is declined and even the number of credit cards the user has available to pay within the app and the level of them.

Financial services: This last set of features comes from the fintech functionality within the Super-App. These features collect users' behavior towards the financial services or products delivered via technology ranging from e-wallets and digital cards to loan services, on- and offline payments, and money transfers [8].

4 Experimental Setup

4.1 Data

Our data set consists of the transactional historical information from 43.270 users within a Super-App. This transactional information is collected directly from the orders and interactions with more than 15.000 restaurants and 2.000 grocery stores. From these orders, it was possible to retrieve several other features such as the latitude and longitude from which each order was placed, time features such as the day of the week and the hour where the order was made, the used payment method and - when applicable - its credit card information, the device and the operating system used to interact with the super-app and many more. All this data allowed us to understand each customer's consumption patterns and build the variables that might be indicators of each user's income Moreover, we had access to the real-validated income of the users, who built the training and testing sets for the model, together with their Bureau Income Estimates.

4.2 Setup

Observing the existing research in terms of credit risk and income estimation, and in order to make the most of the available data and the computational power

of artificial intelligence, it was decided to implement an XGBoost model for the scope of this research as it has proven to outperform other algorithms such as neural networks, support vector machines, bagging-NN and many others with regard to structured data [13,9]. And furthermore has showed predictive power for credit risk assessment models using alternative data. [8]

Considering this research is being carried out within a business environment, evaluating statistical model performance has to come in handy in terms of interpretability to discuss its possible business implications effectively. Hence, we will evaluate the proposed model according to the Mean Average Percentual Error (MAPE). This metric is a measure of prediction accuracy and a widely accepted indicator within the organization environment when considering the utilization of a prediction method [14]. This metric is also a variation of the Mean Absolute Error used to evaluate previous Income Estimate models' performance in literature [7]. We implemented cross-validation with five splits for evaluating the model, keeping a data proportion of 80% -20% for training and testing in each iteration, respectively. Furthermore,we conducted statistical tests such as Levene for homoscedasticity of variables and Mann-Whitney to evaluate of significance in the MAPE.

The Mean Average Percentual Error (MAPE) used for the performance evaluation, is no more that the average of each observation of the percentual error of it, MAPE can be expressed as :

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \frac{A_i - F_i}{A_i}$$

Where A_i indicates the actual value of the user's income for the observation i while F_i indicates the forecasted value of the user's income for the observation i.

5 Results

5.1 Statistical model performance

The results obtained in this study show a significant improvement in performance in terms of the MAPE for the model that incorporates Super-App alternative data compared with the Bureau Income Estimates by themselves as can be seen in Fig. 1 and is confirmed by the result for the p-value of the Mann Whitney non-parametric test for both Bureau vs. Super-App and Bureau + Super-App vs Bureau comparison in Table 1. Moreover, an equivalent performance of the combined model versus the Super-App Only is obtained, according to the p-value of the Mann-Whitney non-parametric test results, presented in Table 1. The p-values of the Levene test conducted to verify homoscedasticity and hence the applicability of the Mann Whitney test, are also presented in Table 1.

Therefore, our results suggest that financial institutions can build better and more accurate estimates solely with this alternative data, without requiring a particular bureau income estimate. This, as the model that is built with both

alternative data and the Bureau Income Estimate (Bureau + Super-App) does not outperforms the alternative data-only model (Super-App) Fig. 1 supported by equivalent result of the Mann Whitney non-parametric test Table 1. Which ultimately, indicates that the bureau income estimate subject of this study does not manages to capture any significant information that the Super-App alternative data is not already capturing in terms of the user's creditworthiness. It is worth mention that building a particular model only with the retrievable features that the interaction of a user with a Super-App generates theoretically does not line up with the stated objective of credit risk management, that is to incorporate all the available information of a user to build the most detailed and truthful profile possible. However, the results indicate that the bureau income estimate does not add value when building this profile to assets income estimate, and the Super-App features are showed to be sufficient to complete said profile without having to resort to the financial history of the users. Ultimately, pointing out that financial institutions, that manage to incorporate this alternative data when building their credit risk profiles, can reduce the risk in their decision-making process of addressing how much money they should lend to their customers, operating under a more accurately estimated expected loss (EL).

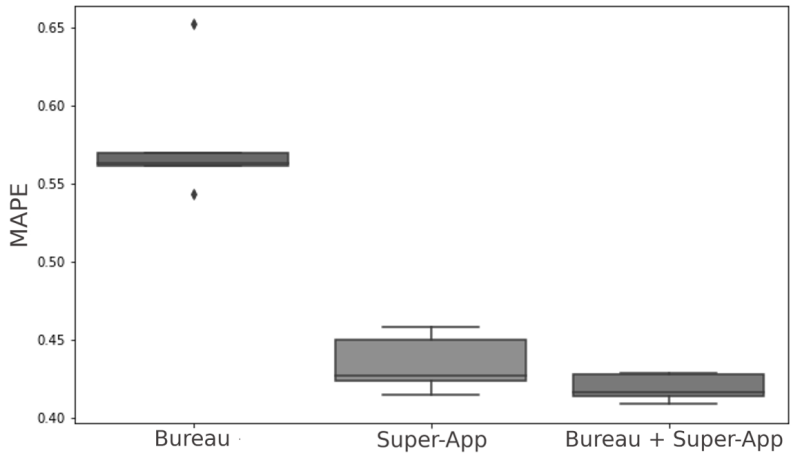


Fig. 1. MAPE performance by model

5.2 Feature Importance

In order to be able to determine which of the behavioral patterns identified have more predictive power when estimating users income, as previously mention, we implemented SHapley Additive exPlanations [3] in order to give some degree of explainability to the results of the executed XGBoost model.

Table 1. Mann Whitney and Levene test P-Value for MAPE performance

Model Comparison	Levene	Mann Whitney
Bureau, Super-App	0.96364	0.00609
Bureau + Super-App, Bureau	0.40492	0.00609
Bureau + Super- App, Super-App	0.29555	0.14813

Figure 2 is the visualization for the Bureau Income Estimates combined with Super-App alternative data model feature importance, where features appear from top to bottom in order of importance; Figure 2 shows that the distance of the industry-accepted Bureau Income Predictor of the user and the average of the same estimate for the users within the same age range presents the highest predictive power from the considered features. In contrast, as for alternative data features, the number of times the user has placed orders at one of the most expensive restaurants within the Super-App (Where the average plate has a value higher than the average plate of 70% of the other restaurants) has the most robust predictive value. Furthermore, Figure 2 shows each feature's impact on the combined features model. Notably, users with a taste for expensive (top) restaurants (Delivery_COUNT_Orders_Top_Restaurants), a higher delivery consumption (Delivery_Total_Consumption), a higher amount of money debited through financial services (Financial_Services_AVG_Debit_Amount and Financial_Services_Debit_Perc_80), users with high-level or premium credit cards registered in the app (Payment_MAX_CC_Score), users who differ to a large number of instances orders paid by credit card (Payment_Info_AVG_Instances_CC_Ord) and users with a preference for the Credit Card as payment method (Payment_Info_PCT_CC_As_Payment_Method) are estimated to have a higher income than their counterparts. Moreover, users with a high engagement towards the Super-App different verticals were also predicted with higher incomes than users with a lower engagement (Delivery_Count_Vertocales). It can also be highlighted that users with a taste for discounts are predicted to have lower incomes than users who do not use discounts in their delivery orders. As can be seen with the performance of variables like Delivery Discount Level' and 'Delivery AVG Paid,' both proxies of the user's behavior towards discounts, where a high values of the variable indicates high adoption towards discounts and, as can be seen in Figure 2, are associated with low income users. The tipping behavior results show that users who on average tip less in their delivery orders are estimated to have lower incomes than users that leave higher tips. Hence our results indicate that the propensity of tipping high, the taste for expensive restaurants and products, credit card preference, short term period installments for credit card payments and overall a high monetary consumption and high amount of money debited through financial services are all strong positive income estimators. On the other hand, our results show that the preference to place orders with discounts is a robust negative income estimator.

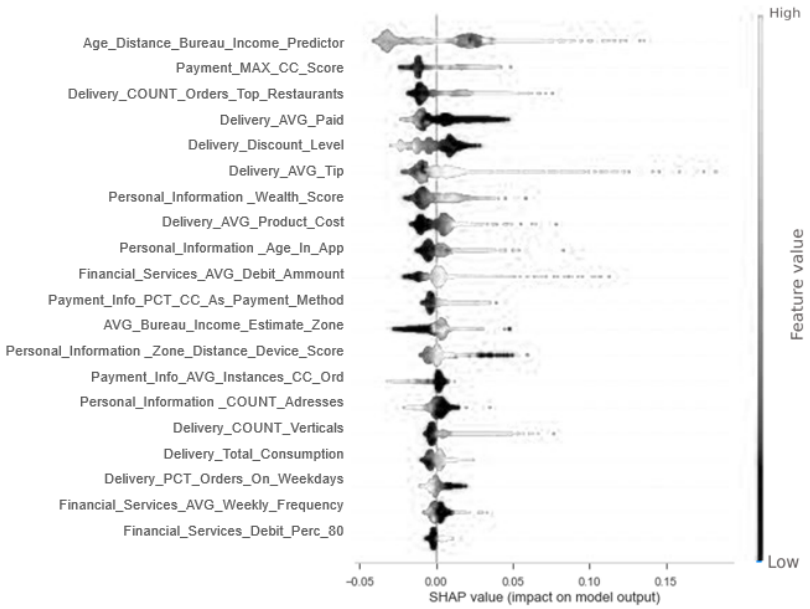


Fig. 2. Local interpretability with SHAP of Bureau + Super-App model

6 Conclusions

This paper determined that the proposed Super-App alternative data have predictive power towards income estimation. Moreover, there is a significant improvement in statistical performance when taking these alternative data sources into account compared to the performance of industry-accepted Bureau Income Estimates by themselves. It was possible to identify consumption patterns, personal information, payment behaviors_preferences, engagement towards financial products and engagement towards the Super-App verticals from extracting data from the users' interactions with the Super-App. These behaviours presented by alternative data showed predictive power when estimating user's income, being all this information non-available to traditional bureau estimates, that incorporate only past financial information of the user.

Moreover, we presented how financial institutions that possess this kind of information should include it into building their users' credit risk profiles.; as this would implicate operating within a more accurate expected loss. Furthermore, this paper layout how financial institutions that manage to incorporate Super-App sourced alternative information into their credit risk profiling will be able to assess the income of customers who do not have previous financial information which should represent a benefit towards the bankarization of this population, which is outside the reach of traditional bureau income estimates.

This study's results present the incentive for financial institutions to seek to incorporate this type of information into constructing their risk profiles. Doing so will allow them to keep up with these technology-based financial institutions that, in the meantime, will have a competitive advantage if they manage to align their strategical operation with the insights of the data they can collect. Additionally, our findings should motivate further research in this field to determine what other sources of alternative information can provide value towards building complete and accurate profiles for financial institutions' users.

References

1. Basel Committee on Banking Supervision: The internal ratings-based approach (2001), Consultative Document
2. Einav, J.L.: The data revolution and economic analysis. Innovation Policy and the Economy, **14**, 1–24 (2014)
3. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
4. McAfee, E.B.: Big data: The management revolution. Harvard Business Review (2012)
5. Perry, V.G.: Giving credit where credit is due: the psychology of credit ratings. The Journal of Behavioral Finance **9**(1), 15–21 (2008)
6. Phelps, J., Nowak, G., Ferrell, E.: Privacy concerns and consumer willingness to provide personal information. Journal of Public Policy & Marketing **19**(1), 27–41 (2000)
7. Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., Aletras, Studying user income through language, behaviour and affect in PLOS ONE **10**(9), 1–17 (2015)
8. Roa, L., Correa-Bahnsen, A., Suarez, G., Cortés-Tejada, F., Luque, M.A., Super-app behavioral patterns in credit risk models: Financial, statistical and regulatory implications, Expert Systems with Applications **169** (2021)
9. Salvaire, P.: Explaining the predictions of a boosted tree algorithm : application to credit scoring. Master's thesis, Universidade Nova de Lisboa (2019)
10. Schuermann, T.: What do we know about loss given default? (2004)
11. Shaw, J.: Why "big data" is a big deal (Aug 2016), `https://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal`
12. Tokunaga, H.: The use and abuse of consumer credit: Application of psychological theory and research. Journal of economic psychology **14**(2), 285–316 (1993)
13. Xiaa, Y., Liu, C., Li, Y., Liu, N.: Boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications **78**, 225–241 (2017)
14. LNCS Homepage, `http://www.springer.com/lncs`. Last accessed 4 Oct 2017

Table of Content

adfa, p. 1, 2011.