

A Neural Network-Based Approach in Predicting Consumers' Intentions of Purchasing Insurance Policies

Wen Teng Chang , Kee Huong Lai 

School of Mathematical Sciences, Sunway University, No. 5, Jalan Universiti, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia

Corresponding author: Kee Huong Lai (keehuongl@sunway.edu.my)

Abstract

Insurance is a crucial mechanism used to lighten the financial burden as it provides protection against financial losses resulting from unexpected events. Insurers adopt various approaches, such as machine learning, to attract the uninsured. By using machine learning, a company is able to tap into the wealth of information of its potential customers. The main objective of this study is to apply artificial neural networks (ANNs) to predict the propensity of consumers to purchase an insurance policy by using the dataset from the Computational Intelligence and Learning (CoIL) Challenge 2000. In addition, this study also aims to identify factors that affect the propensity of customers to purchase insurance policies via feature selection. The dataset is pre-processed with feature construction and three feature selection methods, which are the neighbourhood component analysis (NCA), sequential forward selection (SFS) and sequential backward selection (SBS). Sampling techniques are carried out to address the issue of imbalanced class distributions. The results obtained are found to be comparable with the top few entries of the CoIL Challenge 2000, which shows the efficiency of the proposed model in predicting consumers' intention of purchasing insurance policies.

Keywords

Neural network; Feature selection; Classification; Prediction; Consumer targeting.

Citation: Chang, W. T., & Lai, K. H. (2021). A Neural Network-Based Approach in Predicting Consumers' Intentions of Purchasing Insurance Policies. *Acta Informatica Pragensia*, 10(2), 138–154. <https://doi.org/10.18267/j.aip.152>

Special Issue Editor: Chiew Kang Leng, Universiti Malaysia Sarawak, Malaysia

Academic Editor: Zdenek Smutny, Prague University of Economics and Business, Czech Republic

Copyright: © 2021 by the author(s). Licensee Prague University of Economics and Business, Czech Republic.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY 4.0).

1 Introduction

The insurance industry is an indispensable sector of a country. From the economic perspective, it mobilizes the country's domestic savings, allocates capital to productive investments, facilitates trade and commerce, and in turn contributes to the country's sustainable economic growth. Most importantly, the fundamental role of insurance is to protect people against financial losses arising from unexpected events such as death, accidents, fire and theft. Insurance mitigates uncertain risks and provides certainty for the insured in the form of payment to compensate for the financial losses incurred.

In order to achieve the aspirations of increasing the penetration rate and narrowing the protection gap for the benefits of the people (LIAM, 2019), insurers begin to avail themselves of the advancement of technology. For instance, the integration of blockchain and artificial intelligence (AI) in the claim settlement process, and telematics in the usage-based car insurance policies (Nor Shamsiah, 2018).

Specifically, targeting the right customer groups and identifying the factors of buying a policy will be advantageous to insurance companies, in terms of implementing the right strategies to enhance their insurance policies and, in turn, increase the renewal rate. Assessing the factors of customers buying an insurance product can also help insurers to develop new products to cater for the concerns of the uninsured, which will then contribute to an overall increase in the insurance penetration rate.

Although deep learning is commonly used in customer targeting, sales forecasting and data validation, it has yet to be prevalently adopted in the insurance sector. As such, this work will further explore the use of artificial neural networks (ANNs) in the business context of insurance. The contribution of this work is two-fold. The proposed model that is used to predict customers' intention of purchasing a particular insurance policy will enrich the literature in the field of machine learning, applied in the domain of actuarial science. This work also aims to identify the factors affecting the propensity to buy the studied insurance product via feature selection. To evaluate the effectiveness of the proposed model in this work, the results obtained are compared with the results of the submitted entries of the Computational Intelligence and Learning (CoIL) Challenge 2000, which used the same dataset.

The rest of this paper is organized as follows. Section 2 provides an overview of the methods used in consumer targeting. Section 3 describes the dataset and methodology adopted in this work. Section 4 presents the results and discussion. Lastly, Section 5 concludes the paper and provides suggestions for future research.

2 Literature Review

2.1 Consumer targeting

Consumer targeting, or customer targeting, is a business process that defines which subset or portion of customers to focus on, and subsequently, market products to. Direct mailing, which is the problem involved in the CoIL Challenge 2000, is a marketing method that involves directly sending promotional material to a customer without using advertising media. The objective of selecting the subset of customers that shares similar characteristics to market to is to maximize the return made from a marketing investment.

For an insurance company, customer targeting is important for the company to determine the right customer group to market its insurance product to. When carrying out customer targeting, the company is able to recognize the common underlying factors that drive a customer to buy an insurance product. Understanding the needs of customers is of paramount importance. A company could then devise appropriate measures to meet those needs through market research, such as scanning industry reviews, running a focus group or doing a market survey. After identifying the deciding attributes of customers, the company could devise better business strategies to retain its existing customer group. From another

point of view, the insurance company would also be able to recognize the reasons behind the uninsured people's decisions. From there, the company can devote some of its resources to this specific uninsured group by developing new marketing plans or insurance products that meet the needs and abilities of the uninsured group.

There are various techniques and models that have been employed by database marketers to improve the process of consumer targeting in various industries. The following chosen works reported in the literature aim to provide an overview of the state-of-the-art methods used by marketing research across different business domains. Mahmoudi et al. (2020) employed the endogenous attendance attributes (EAA) model and conditional logic model in a proposed choice experiment, to evaluate consumers' preferences on organic products. It is highlighted that the EAA model is suitable for use in datasets where the respondent groups are heterogeneous. Delley et al. (2020) attempted to provide evidence-based recommendations to develop the target product. The fluid milk consumers were analysed and grouped into 3 distinct clusters using the hierarchical cluster analysis. It is noted that redefining the product assortment according to the needs of consumers from different clusters is vital in boosting sales.

In the healthcare industry, Chen et al. (2020) incorporated a casual forests-based approach in a proposed multiperiod, randomized field experiment, to improve cancer outreach. A simulation showed that the promising methodology is able to provide a substantial positive payoff to the healthcare industry. Sentiment analysis is used by Lin et al. (2020) to better formulate marketing strategies. The option mining-based approach successfully established a comprehensive feature keyword library in various fields.

The effectiveness of green marketing is studied and reported by Tsai et al. (2020). The decision-making trial and evaluation laboratory method was used in conjunction with the analytic network process, to evaluate how the introduction of green technologies could affect purchase intentions among consumers. The study identified various dimensions and criteria that influence consumers' purchase intentions.

Summarizing the works above, it could be observed that there is a wealth of developed statistical and econometric methods that could be used to aid consumer targeting. In the following subsection, the integration of artificial intelligence (AI) in such research will be further examined.

2.2 Artificial intelligence in consumer targeting

Verma et al. (2021) conducted a comprehensive review on the use of artificial intelligence (AI) in marketing. The main finding of the review work is that AI could be used as a very promising marketing tool in the era of disruptive technology. Ma and Sun (2020) compared the implementation of AI-based approaches with the traditional statistical-based approaches in the context of marketing. The work reports that automated AI agents could proliferate almost every aspect of business and marketing. Thus, it is imperative to adopt AI-based approaches to deepen the understanding of consumers' behaviour to assist companies in making well-informed business decisions.

Arasu et al. (2020) employed machine learning tools in examining the data analytics generated from social media. To predict the consumers' purchasing intention, data mining techniques are also used in the work. The proposed model that integrates the Waikato Environment for Knowledge Analysis (WEKA) machine learning tool is found to outperform other similar algorithms. In the banking sector, Ładyżyński et al. (2019) proposed the use of deep learning and random forests in predicting consumers' willingness in taking up a personal loan. The expert system is reported to be able to extract significant patterns from consumers' historical transaction data. In the context of online group purchases, Leong et al. (2019) employed a neural network-based approach in predicting the group's actual spending. The proposed model successfully identifies the important features or factors that influence the spending of online purchase groups.

To the authors' best knowledge, the research on consumer targeting in the field of actuarial science is limited and this provides the motivation for this research work. The results obtained could also be studied further, to find the underlying factors and important attributes that play major roles in affecting consumers' propensity to purchase insurance policies.

There were numerous submissions of the prediction task of the CoIL Challenge 2000. The winning entry, Elkan (2000), employed a naïve Bayes approach and correctly identified 121 out of 238 policyholders (van der Putten et al., 2000a). The runner up, Kontkanen (2000), scored 115 by using an ensemble of 20 pruned naïve Bayes classifiers. Both Greenyer (2000) and Koudijis (2000) attained 112 correct policyowners by using the JXCS Learning Classifier System, which is a Java-based program developed from accuracy-based XCS classifier systems and evolutionary algorithms. Vesanto and Sinkkonen (2000) employed a radial basis function (RBF) network and attained a score of 110. Jorgensen and Linneberg (2000) correctly identified 108 caravan policyholders by using a random-access memory (RAM)-based neural network. Shtovba and Mashnitskiy (2000) adopted a backpropagation multilayer feedforward neural network. However, Elkan (2001) pointed out the risk of overfitting of the proposed model.

Other published works have analysed the nature of the dataset and the solutions submitted to the competition. Elkan (2001) emphasized the risk of overfitting and the issue of statistical significance when handling this dataset. A bias-variance analysis was performed by van der Putten and van Someren (2004) to investigate the reason behind the wide range of performance in the competition. Their study ascertained the significance of procedures before and after the core modelling step and introduced the concept of bias-variance decomposition. Darzi et al. (2019) highlighted the incompatibility of the dataset due to the fact that the distribution of the target variable classes is highly imbalanced, and proposed a data-level method of sampling and a cost-sensitive learning (CSL) approach.

Wu and Barbará (2002) developed a two-part model combining the logistic and loglinear components to impute missing values into the dataset and scored 94 out of 238 policyholders. Zadrozny and Elkan (2002) applied a calibrated two-class probability estimates on naïve Bayes (NB) and support vector machine (SVM). Kim et al. (2005) observed 120 caravan policy owners by implementing an ANN model with the Evolutionary Local Selection (ELSA) algorithm. Rahangdale et al. (2016) employed a k-nearest neighbour (k-NN) classifier and a naïve Bayes classifier, and obtained 37 and 54 policyholders out of the test set of 4000 respectively. Darzi et al. (2019) achieved a score of 237 out of 238 policyholders by using a Tree Augmented Naïve Bayesian (TAN) method with different sampling approaches, such as undersampling, oversampling and hybrid sampling, which combines both undersampling and oversampling concepts.

Summarizing the aforementioned works, it is noticed that a wide range of methods has been used to solve the problem associated with the dataset, including logistic regression, support vector machines, fuzzy classifiers, naïve Bayes, decision trees and neural networks. It is also worth noting that the low-variance learner naïve Bayes classifiers are very competitive in the CoIL Challenge 2000. In this study, a neural network-based model is applied to the dataset after considering the possible restrictions of the dataset emphasized in the previous works.

3 Methodology

The framework of the methodology is presented in Figure 1.

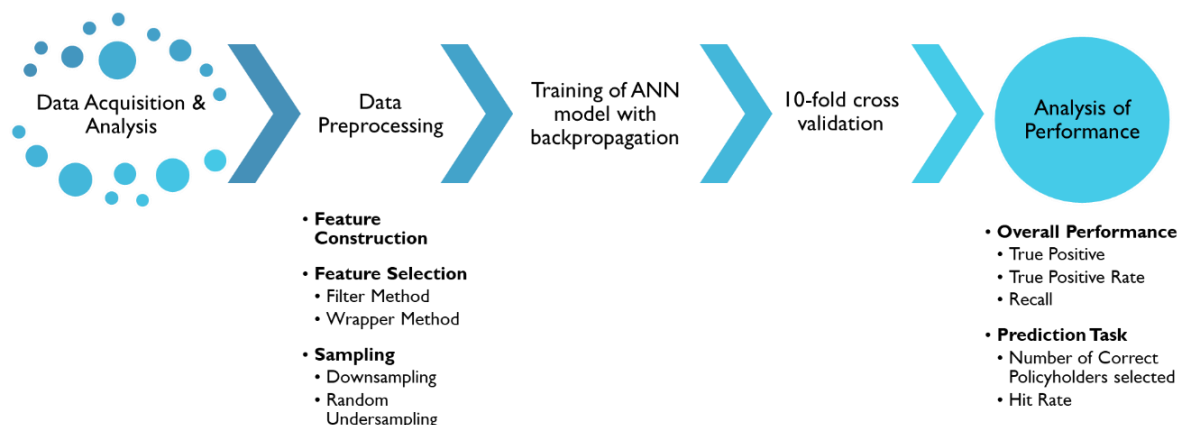


Figure 1: Flow chart of proposed methodology.

3.1 Data acquisition

The publicly available dataset used in this study is provided by the Dutch data mining company Sentient Machine Research and was used for the CoIL Challenge 2000 (van der Putten et al., 2000a). The CoIL Challenge 2000 was a data mining competition organized by the Computational Intelligence and Learning (CoIL) Cluster, a network of excellence funded by the European Union (EU). It was held from 17 March to 8 May 2000 and had attracted 43 submissions in total (van der Putten & van Someren, 2004). The competition aimed at encouraging the technology of computational intelligence and learning to be applied to real-world problems and to stimulate incessant search for new and improved solutions.

The prediction task of the competition introduced a direct mailing problem in which the participants were required to select a subset of 800 customers of a caravan insurance policy out of 4000 customers in order to achieve the best response rate from this targeted selection. The prediction task can be considered a classification problem that categorizes potential customers of the caravan insurance. The problem is meaningful such that if a company recognizes its potential customers, the company can reduce its direct mailing costs and come up with better marketing strategies (van der Putten et al., 2000a).

The training set and the test set were predetermined by the competition organizer, consisting of 5822 and 4000 samples respectively. The dataset contains 9822 customer records in total, with each record encompassing 86 attributes: 43 socio-demographic data items and 43 product ownership information items. The 86th attribute, which is the target variable, indicates whether the customer owns the caravan insurance policy (van der Putten et al., 2000a). It is worth noting that the socio-demographic information was derived from the particular customer's postal code, which implies that two different customers with the same postal code could have exactly the same socio-demographic data. Even though it is absolutely possible that customers living in the same area have different decisions on purchasing a policy, this assumption is commonly used and reasonable for research purposes as the company may only have access to very general information about its customers (van der Putten & van Someren, 2004).

3.2 Preliminary data analysis

There is no missing value in the dataset. Since all the attributes were discretized by the organizer of the competition, the dataset contains only discrete-valued features where continuous-valued data is categorized into different discrete levels. Furthermore, there exist different customers with the same values in all attributes but behaving differently on purchasing the caravan insurance policy in the dataset.

This is the result of the dataset discretization. Seewald (2000), a participant of CoIL Challenge 2000, removed these inconsistent instances to avoid confusion to his algorithm. However, all the inconsistent data samples are retained in this study because it is not always possible to obtain as detailed customer information as expected.

It is very important to notice that the distribution of classes of the target attribute is extremely skewed to class zero, which means that the majority of the customers do not own a caravan insurance policy. Generally, most of the classification methods are prone to bias towards the majority class in this situation. Moreover, if the prediction model assigns all the customers to class zero without realizing the customers' purchasing pattern, the model can still attain 94% of prediction accuracy because 94% of the customers are grouped in class zero. This implies that a high prediction accuracy does not necessarily indicate that the model is good. Therefore, different evaluation metrics need to be defined to evaluate the ability of the model to recognise the minority class. Data pre-processing steps are also essential to deal with this imbalanced dataset.

3.3 Data pre-processing

3.3.1 Feature selection

Feature selection is the process of selecting a subset of the original variables by eliminating features that are redundant or carry little predictive information. There are two types of feature selection approaches that are used in this study, namely the filter method and the wrapper method. The filter method evaluates the features independently from the model. Each feature is ranked by a score assigned based on a pre-specified evaluation metric and it is then decided whether the feature will be retained in or eliminated from the dataset. The modified dataset is fed into the model to be trained and tested based on the selected features only. Since the filter method is independent of the predictive model, a selected feature subset that possesses a good evaluation score does not guarantee high classification accuracy when these features are trained and tested using the model (Zainuddin et al., 2016).

On the other hand, the wrapper method evaluates the goodness of selected subsets of features according to the performance of the predictive model, which is cross entropy in this study. Unlike the filter method, the wrapper method incurs relatively higher computational costs because the predictive model needs to be retrained for each proposed feature subset (Zainuddin et al., 2016).

The filter method employed in this study is the neighbourhood component analysis (NCA). This is a nearest neighbour-based feature weighting algorithm that learns a feature weighting vector by minimizing the loss of the k-nearest neighbour (k-NN) classifier with a regularization term (Yang et al., 2012). No parametric assumptions pertaining to the distribution of the data is made in this method. A feature with its weight exceeding a pre-specified threshold will be selected (MATLAB, 2020).

Sequential forward selection (SFS) and sequential backward selection (SBS) are selected as the wrapper methods in this study. SFS selects one feature at a time until the addition of another feature does not improve the quality of the subset with the condition that once a feature is selected, it remains selected (Liu & Motoda, 2008). Likewise, SBS removes one feature at a time and a feature removed will never be considered for inclusion again (Liu & Motoda, 2008).

3.3.2 Feature construction

Feature construction is a dimensionality reduction method which generates a new set of features that are more efficient by creating or inferring additional features (Motoda & Liu, 2002). All new features are transformed from the original features. Feature construction is performed to enhance the expressive power of the original features (Motoda & Liu, 2002). Moreover, feature construction could reduce bias

error by changing the search bias of a learning method or relaxing the learning bias of a classifier (van der Putten & van Someren, 2004).

In this study, new features are constructed based on knowledge and experience. Two features, “Customer Subtype” and “Customer Main Type” are broken down into their respective binary variables. This method was performed by Kim et al. (2005). Furthermore, the 32nd to 34th attributes, which are features “1 car”, “2 cars” and “no car”, are combined into a new attribute, named “Mean Number of Cars”, whereas the five features related to income, 37th to 41st attributes, are combined into a new attribute “Mean Income”. This process of combining is performed by averaging the original features of each new attribute. This is because each of these features was originally expressed in a way to provide the proportion of residents living in an area that belongs to that feature. Since the actual number of cars owned or income of each customer cannot be derived from the dataset, the average of these two attributes could provide a better generalization about the number of cars or income of customers that stay in the common area and it is more intuitive for the prediction model to learn.

3.3.3 Sampling approaches

Any classifier model is sensitive to the effect of class imbalance in datasets. To cope with class imbalance, there are two main approaches, namely internal methods, which develop new algorithms specially to address the imbalance in data, and external methods, which modify the dataset to be used for training.

An external method, specifically undersampling, will be used in this study. External methods alter the distribution of rare and frequent patterns in the dataset to improve the detection of the minority class. This operation is known as sampling and its purposes are to increase the sample rate of the minority class and attenuate the skewness in the distribution of the minority class by generating a new dataset from the original dataset (Darzi et al., 2019).

The two sampling techniques are undersampling, which removes samples of the majority class, and oversampling, which adds samples to the minority class. As oversampling requires replicating the existing samples of the minority class and creating new samples in a specific region of the input space, it is rather expensive in terms of computational costs, and therefore only undersampling is performed in this study. Via undersampling, some instances from the majority class of the original dataset are eliminated until a pre-specified balance ratio is achieved.

On the other hand, downsampling is performed to decrease the sample rate of the majority class by an integer factor, called a downsample gap. If the downsample gap is denoted by h , the samples will be removed for every h customer of the majority class in the training set. This will result in a new training set containing n' samples of class zero, as shown in Equation (1), where the original size of class zero is 5474.

$$n' = 5474 - \left\lfloor \frac{5474}{h} \right\rfloor, \quad (1)$$

The numbers of the downsample gap that are used in this study are 2 and 3. Downsample gap 2 implies that the size of the majority class in the training set is halved. Subsequently, random undersampling is performed on the training set. This approach eliminates instances from the majority class at random until it matches the size of the other class. In this study, samples from class zero are removed at random until the ratio of the size of class zero to class one reaches 2:1 and 1:1. It is worth mentioning that this random elimination method poses the risk of removing potentially useful data, leading to great information loss and an increase in variance of the model (Cateni et al., 2013).

3.4 Numerical experiment

3.4.1 Neural network pattern recognition (NNPR)

Conceptually, artificial neural networks (ANNs), or neural networks, are created to emulate the way a human brain carries out a specific task such as recognition and categorization (Haykin, 2010). A neural network, similar to the human brain, acquires knowledge from its environment through a learning process and stores the acquired knowledge via its synaptic weights (Haykin, 2010).

The numerical experiments are performed by using the mathematical software MATLAB® R2020b (MATLAB, 2020). The Neural Network Pattern Recognition (NNPR) application in MATLAB is used to run the dataset to obtain the preliminary results. The cross entropy is the default performance function. The activation functions of the neurons in the hidden layer and output layer are the hyperbolic tangent function and the sigmoid function, respectively. The default data processing procedure applied to the dataset is mapping of each attribute's minimum and maximum value to a range $[-1,1]$ using Equation (2):

$$x_{\text{new}} = \frac{2(x-x_{\min})}{x_{\max}-x_{\min}} - 1, R_f \in [-1,1], \quad (2)$$

where x is the value of an attribute of a particular customer, x_{\max} and x_{\min} denote the maximum and minimum value of the attribute respectively, and x_{new} represents the value after data processing.

3.4.2 ANN model

In order to compare the results of the competition submissions and other published works, the dataset is separated into a training set and a test set as specified in the competition. The training set contains 5822 customer records, whereas the test set consists of 4000 customer samples. A multilayer perceptron (MLP) neural network with a hidden layer is used, with the number of hidden neurons set to be equal to the square root of the number of input variables, which is 85 if no feature selection or feature construction is applied. This is a rule of thumb when determining the number of hidden neurons in a hidden layer (Kim et al., 2005).

The hyperbolic tangent function is the activation function of the neurons in the hidden layer as it has better properties for learning purposes which will result in a faster training (Netšajev, 2016). The hyperbolic tangent function scales the output to a range between -1 to 1. For the output layer, the non-linear sigmoid function is chosen as the activation function because it gives an output that ranges between 0 and 1.

Given a prediction value from the ANN model \hat{y}_n and a target value y , a performance function measures the difference between these two values. The performance function in this ANN model is the cross entropy because it is proven to speed up the learning process and provide satisfactory overall network performance with a comparatively short stagnation period (Joost & Schiffmann, 1998). The cross entropy function is given by Equation (3), where N is the total number of samples, \hat{y}_n is the n th predicted value from the proposed model, and y_n is the n th known target value from the dataset.

$$\text{Cross entropy} = -\frac{1}{N} \sum_{n=1}^N [y_n \log \hat{y}_n + (1 - y_n) \log (1 - \hat{y}_n)]. \quad (3)$$

The scaled conjugate gradient (SCG) backpropagation is the learning algorithm of the proposed ANN model. The commonly used backpropagation algorithm is a gradient descent-based approach which adjusts the weights in the direction of the steepest descent (negative of the gradient) to minimize the performance function of the prediction model. For the conjugate gradient backpropagation, the synaptic weights are adjusted along conjugate directions and faster convergence is observed compared to the steepest descent directions (MATLAB, 2020). Hence, the SCG backpropagation is adopted to shorten the

computational time, especially for the model that employed the wrapper method in its feature selection step.

Furthermore, each of the features in the dataset is normalized so that the normalized features have a common mean of 0 and a standard deviation of 1. The z-score normalization algorithm is shown in Equation (4):

$$x_{\text{new}} = \frac{(x - x_{\text{mean}})}{x_{\text{std}}}, \quad (4)$$

where x is the value of an attribute of a customer in the dataset, x_{mean} and x_{std} denote the mean and standard deviation of the attribute, respectively, and x_{new} represents the value after normalization. This normalization step is required because the features of the CoIL Challenge 2000 dataset have different ranges. The purpose of normalization is to change the values across different features in the dataset to a common scale, without distorting the differences in the ranges of values.

3.4.3 Prediction task

In order to compare the results obtained in this work with those from the submissions of the CoIL Challenge 2000, the algorithm of the numerical experiment is amended to deal with the prediction task of the competition, in which a group of 800 customers is required to be chosen out of the 4000 customers. The subset of 800 potential customers is selected from the test set by changing the classification threshold. Prediction results are determined according to a classification threshold (Zou et al., 2016). Typically, the threshold is set to 0.5 in which any output that is greater than 0.5 will be assigned to class one. In order to select the top 800 customers who are most likely to buy the caravan insurance policy, the classification threshold is adjusted to be the output value of the 800th largest output. Hence, the 800 customers with prediction outputs greater than the adjusted classification threshold are assigned to class one.

3.4.4 Ten-fold cross validation

To evaluate the performance of the ANN model with different data pre-processing methods, a 10-fold cross validation (CV) is performed. The dataset is partitioned into 10 disjoint subsets. The first subset is used as the test set and the remaining subsets are used to train the ANN model. The trained ANN model is then tested on the test set. This procedure is repeated until each of the 10 subsets is used as the test set. The purpose of the 10-fold cross validation is to ensure that each data item has an equal chance of being trained and used as a test set.

3.4.5 Evaluation metrics

Two sets of evaluation criteria are defined for the overall performance of the model on the test set containing 4000 samples and the performance of the prediction task of the CoIL Challenge 2000. All the evaluation metrics are presented in the form of “mean \pm standard deviation” as a result of the 10-fold cross validation. The minority class, which is class one of the target variable, is considered a positive class, whereas class zero is regarded as a negative class. To evaluate the overall performance on a test set, precision and recall will be used, as described in Equations (5)–(6), where TP, FP, and FN represent the cases of true positive, false positive, and false negative, respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

In order to select the 800 customers, the number of the true caravan policy owners identified is used as the primary measurement to compare the results of this study with those of other works. A higher number of correctly identified policy owners implies a better model. To resemble the direct mailing evaluation,

the hit rate is also presented. The higher the hit rate, the more response a company would receive from the marketing strategy. The hit rate is calculated as the percentage of correct policyholders out of the selected 800 customers, as shown in Equation (7). Since the maximum number of caravan policyholders in the test set is 238, the best hit rate that can be achieved for the dataset is 29.75%.

$$\text{Hit rate} = \frac{\text{Number of true caravan policy owners correctly identified}}{800} \times 100\% \quad (7)$$

4 Results and Discussion

4.1 Overall performance

The overall performance of the different models is presented in Table 1. After performing the z-score normalization on the data, it is obvious that the results of the model without feature selection were better than the results of the NNPR application, with an average true positive value of 2.90. This proves that besides accelerating the convergence of backpropagation, dataset normalization will improve the performance of ANN models.

For ANN models with different combinations of feature construction, feature selection and sampling approaches, the effects of these data pre-processing methods are analysed as follows. In general, feature construction improved the results of all the models listed in Table 1, except the case of SFS. Most of the models obtained a higher mean of true positive, precision and recall after the feature construction step was applied. Hence, it is proven that feature construction is effective in solving classification problems by creating a new set of features that are more efficient for the classification task. This advantage can be attained only if the newly derived features are able to discover the hidden information about the relationships among the original features (Motoda & Liu, 2002). Therefore, the new set of features obtained from knowledge-based feature construction in this study is reasonable and fruitful. The exception of SFS might be attributed to a subset of different features that was selected when the model was fed with the constructed feature set.

Table 1: Overall performance.

Data pre-processing	Feature construction	True positive (TP)	Precision	Recall
Without any feature selection	No	2.90 ± 4.51	0.19 ± 0.17	0.01 ± 0.02
	Yes	3.60 ± 2.41	0.37 ± 0.13	0.02 ± 0.01
Neighbourhood component analysis	No	1.70 ± 1.64	0.19 ± 0.17	0.01 ± 0.01
	Yes	2.70 ± 0.95	0.33 ± 0.14	0.01 ± 0.00
Sequential forward selection	No	1.40 ± 1.35	0.20 ± 0.17	0.01 ± 0.01
	Yes	1.20 ± 1.48	0.14 ± 0.17	0.01 ± 0.01
Sequential backward selection	No	2.50 ± 2.42	0.23 ± 0.12	0.01 ± 0.01
	Yes	4.40 ± 5.34	0.37 ± 0.13	0.02 ± 0.02
Downsampling (gap = 3)	No	5.40 ± 2.37	0.32 ± 0.08	0.02 ± 0.01
	Yes	6.20 ± 5.55	0.29 ± 0.13	0.03 ± 0.02
Downsampling (gap = 2)	No	9.70 ± 4.08	0.26 ± 0.06	0.04 ± 0.02
	Yes	11.30 ± 5.56	0.27 ± 0.11	0.05 ± 0.02
Random undersampling (Ratio = 2:1)	No	35.30 ± 35.00	0.13 ± 0.09	0.15 ± 0.15
	Yes	53.90 ± 38.01	0.13 ± 0.06	0.23 ± 0.16
Random undersampling (Ratio = 1:1)	No	81.90 ± 63.71	0.12 ± 0.07	0.34 ± 0.27
	Yes	108.40 ± 58.62	0.09 ± 0.04	0.46 ± 0.25

Analysing the results of the models that do not employ feature construction, none of the three feature selection methods seemed to improve the performance. This might be due to the fact that the features selected were not effective enough to solve the classification problem. On the other hand, the sampling

approaches provided better results compared to the basic model without any feature selection. As the size of class zero matched the size of class one in the training set, the mean true positive increased significantly to 81.90. The value of recall increased but the value of precision decreased as the ratio of class zero to class one approached 1:1, because as the number of true positives rose, the number of false negatives decreased but the number of false positives increased. This is reasonable because a prediction model is not able to increase the number of true positives without increasing the cases of false positives (Zou et al., 2016). Therefore, a trade-off between precision and recall is required to obtain the optimal result.

It is also important to notice that as more samples of the majority class were eliminated from the training set, the performance of the prediction model improved at the expense of its stability, as depicted in Figure 2. This phenomenon is further illustrated in Table 2, where the four models that applied the sampling approaches had increasingly greater standard deviations as the balancing effect of the sampling approaches on the uneven training set increased. This is because the undersampling approach usually comes with the risk of discarding useful or important samples. Hence, loss of information in the training set brought great instability to the prediction model.

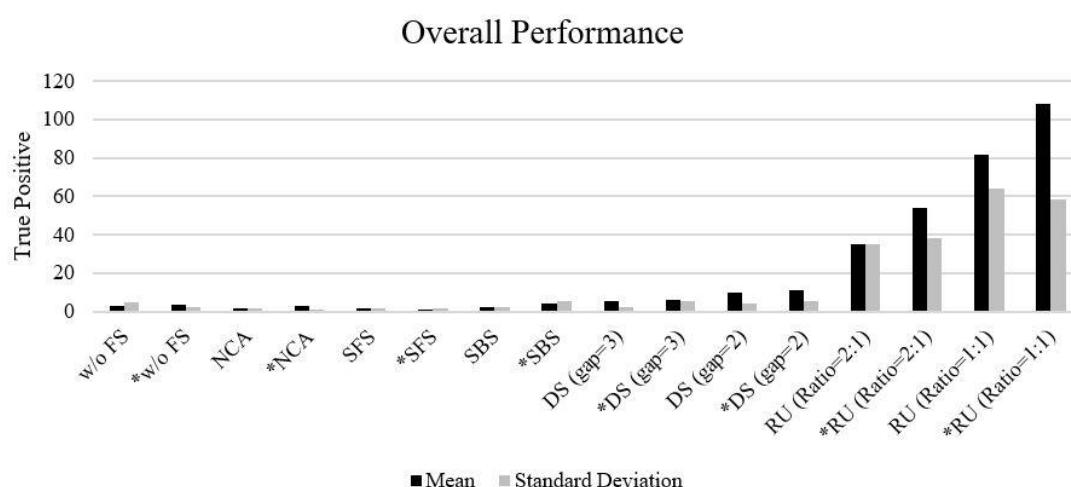


Figure 2: Overall performance of ANN models with different data pre-processing steps.

Note: *: with feature construction; FS: feature selection; DS: downsampling; RS: random undersampling.

Table 2: Effect of undersampling on model stability.

Data pre-processing	Size of class zero in training set	Size of class one in training set	Approximate ratio of class zero to class one	Standard deviation of true positive
Downsampling (gap = 3)	3650	348	10:1	2.37
Downsampling (gap = 2)	2737	348	7:1	4.08
Random undersampling (Ratio = 2:1)	696	348	2:1	35.00
Random undersampling (Ratio = 1:1)	348	348	1:1	63.71

4.2 Prediction task

It can be observed from Table 3 that the basic ANN model without any data pre-processing procedures yielded a mean score of 107. The ANN model that incorporated the SFS identified the highest number of correct caravan policyholders among all the models, with a mean value of 114.50 and a hit rate of 14.31%. This model also had the lowest standard deviation of 2.32 among all the models. Comparing to the basic

ANN model, the NCA did not improve the results but the SBS gave a slight improvement on identifying the caravan insurance policy owners. This shows that the wrapper methods, which evaluate the selected feature subsets based on the performance of the intrinsic ANN model, are important in the prediction task. The wrapper methods demonstrated a relatively lower standard deviation compared to the other models listed in Table 3. This finding corroborated the ultimate goal of feature selection, which is to reduce the variance of the classifier (van der Putten & van Someren, 2004).

Feature construction and sampling approaches that performed well on the overall test set were not able to improve the results for the prediction task. For sampling techniques, the deterioration of performance might be attributed to the greater instability that incurred when selecting a subset of 20% from the original 4000 customers in the test set. The use of feature construction is not significant and this might be due to an increase of variance error, which is a risk of adding constructed features (van der Putten & van Someren, 2004).

According to Elkan (2001), all socio-demographic attributes except "Purchasing power class" were removed because significance testing commonly showed that demographic attributes do not add extra predictive power to models in commercial data mining. After applying his methodology to the best-performing model, the mean number of correct policyholders increased to 117.60 as shown in Table 4. Hence, this shows that industrial experience is beneficial in increasing the effectiveness of the prediction model.

Table 3: Results from selection of 800 customers.

Data pre-processing	Feature construction	Number of correct policyholders	Hit Rate (%)
Without any feature selection	No	107.00 \pm 9.66	13.38 \pm 1.21
	Yes	107.70 \pm 3.68	13.46 \pm 0.46
Neighbourhood component analysis	No	104.20 \pm 6.03	13.03 \pm 0.75
	Yes	104.70 \pm 8.50	13.09 \pm 1.06
Sequential forward selection	No	114.50 \pm 2.32	14.31 \pm 0.29
	Yes	110.90 \pm 5.70	13.86 \pm 0.71
Sequential backward selection	No	110.40 \pm 11.06	13.80 \pm 1.38
	Yes	110.10 \pm 5.38	13.76 \pm 0.67
Downsampling (gap = 3)	No	111.40 \pm 6.10	13.93 \pm 0.76
	Yes	106.20 \pm 9.96	13.27 \pm 1.25
Downsampling (gap = 2)	No	111.70 \pm 3.33	13.96 \pm 0.42
	Yes	107.60 \pm 6.93	13.45 \pm 0.87
Random undersampling (Ratio = 2:1)	No	86.50 \pm 27.41	10.81 \pm 3.43
	Yes	90.70 \pm 24.69	11.34 \pm 3.09
Random undersampling (Ratio = 1:1)	No	86.00 \pm 20.20	10.75 \pm 2.53
	Yes	84.00 \pm 24.64	10.50 \pm 3.08

Table 4: Results after removal of socio-demographic attributes.

Data pre-processing	Removal of socio-demographic attributes	Number of correct policyholders	Hit rate (%)
Sequential forward selection	No	114.50 \pm 2.32	14.31 \pm 0.29
	Yes	117.60 \pm 2.99	14.70 \pm 0.37

4.3 Comparison with others' works

Comparing to the submissions of the CoIL Challenge 2000, the best model of this study managed to identify approximately 117 caravan policyholders from the selection of 800 customers. The results generated from the ANN model with SFS and the removal of socio-demographic information are comparable with the submissions of the competition. The results could be positioned in second place among the participants, as shown in Figure 3.

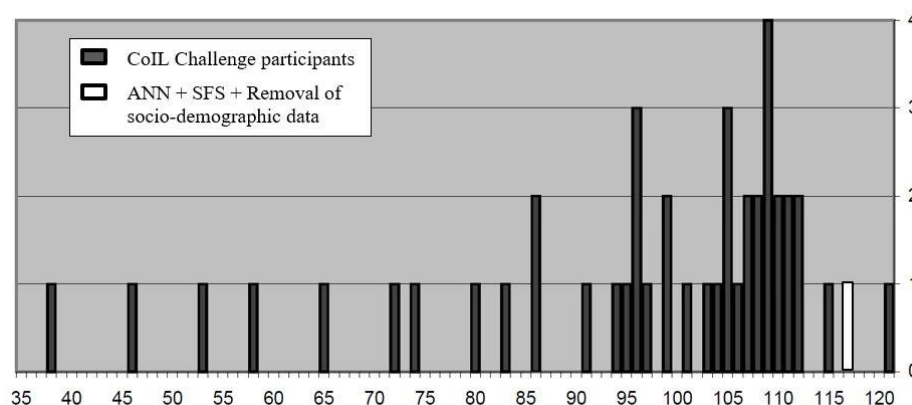


Figure 3: Result comparison of current work with those submitted to CoIL Challenge 2000. Source: adapted from (van der Putten et al., 2000b).

Table 5: Comparison with others' works.

Authors	Year	Methods proposed	Results
Wu and Barabara	2002	Two-part model combining logistic and loglinear components to impute missing values	94
Zadrozny and Elkan	2002	(i) Naïve Bayes (NB) (ii) Linear kernel support vector machine (SVM)	(i) 0.10818 (ii) 0.11122
Kim et al.	2005	ANN with Evolutionary Local Selection (ELSA)	120
Rahangdale et al.	2016	(i) k -Nearest Neighbour (k -NN) classifier (ii) Naïve Bayes (NB) classifier	(i) * 37 (ii) * 54
Darzi et al.	2019	Tree Augmented Naïve Bayesian (TAN) + sampling approach	237
This work	2021	(i) ANN with SFS and removal of socio-demographic information (ii) **ANN with random undersampling and feature construction	(i) 117 (ii) *108

Since Zadrozny and Elkan (2002) only reported the results in mean square error (MSE), a proper comparison with this work could not be made. In terms of the overall performance on the test set (marked as * in Table 5), where a total of 4000 cases was used, the ANN models with random undersampling and feature construction are more competitive than both the models of Rahangdale et al. (2016). However, the model of this study suffers from great instability, judging from the relatively high values of standard deviation, as shown in Table 1 and Figure 2.

Comparing to the results of the competition prediction task, the ANN model with SFS and the removal of socio-demographic information outperformed the model proposed by Wu and Barabara (2002). However, the performance of the models of Kim et al. (2005) and Darzi et al. (2019) were better than that of the best model in this study. This shows that the Evolutionary Local Selection (ELSA) algorithm used in Kim et al. (2005) is a more effective feature selection method. The combination of the Tree Augmented Naïve Bayesian (TAN) model and the sampling approaches that was adopted by Darzi et al. (2019) is an impressive methodology as only one caravan policyowner was not recognized.

4.4 Discussion

Table 6 shows the feature subsets that were selected by NCA, SFS and SBS. It is worth noting that the subsets of the features selected from both the wrapper methods vary for each run of the model because the features are selected based on the performance of the basic ANN model, which has different initial values of weights and biases in each run. The backpropagation algorithm of the ANN model will converge to a different minimum point if the initial values of weights and biases are different.

There are some common features that were frequently selected by NCA, SFS and SBS, namely the number and contribution of car policies. The number feature has a high degree of correlation with the contribution feature. Furthermore, the more a person spends, the more likely he would purchase a caravan insurance. This may be due to the fact that both cars and caravans belong to the same category of vehicles on the road and these two types of vehicles are typically bought for private use. Customers with more car policies and contributions may think that buying the caravan insurance is essential for protecting themselves and their family. Therefore, customers with more car policies and contributions on car policies are more willing to purchase the caravan insurance policy because caravans also serve a similar purpose as a vehicle for private and recreational use.

Additionally, the customer income class is another important factor that determines who would purchase the caravan policy. The target population contains most of the upper quartile of income classes, but not the highest levels. A caravan is used when the owner goes on holidays and it is intuitive that the owner will sleep either in the caravan or in a tent. People who belong to the highest income class may prefer the hotel as their accommodation on holidays. Hence, it is reasonable that customers from the upper quartile of income classes, who can afford to purchase a caravan, but not from the highest income class, are more likely to purchase a caravan insurance policy.

Table 6: Features selected by proposed feature selection methods.

Methods	NCA	SFS	SBS
Features selected	<ul style="list-style-type: none"> - Customer Subtype - High level education - Middle management - Social class A - Income 75-122,000 - Contribution private third party insurance - Contribution family accidents insurance policies - Contribution disability insurance policies - Contribution social security insurance policies - Number of car policies - Number of life insurances - Number of boat policies - Number of bicycle policies - Number of social security insurance policies 	<ul style="list-style-type: none"> - Contribution car policies - Contribution agricultural machine policies - Contribution fire policies - Number of third-party insurances (agriculture) - Number of car policies 	All features except : <ul style="list-style-type: none"> - unskilled labourers - contribution private third-party insurance

5 Conclusion

The results obtained in this study are comparable with other relevant works. In this study, three data pre-processing approaches, namely feature selection, feature construction and undersampling, were proposed to cope with the imbalanced dataset. Feature construction and undersampling are useful in predicting

consumers' intention of purchasing the caravan insurance policy. The ability to extract the underlying relationship between the original features in feature construction is proven in this study. The results of implementing SFS and eliminating socio-demographic features are comparable with other submissions of the CoIL Challenge 2000.

Though the dataset used in this study was from the year 2000, the feature selection methodology proposed could be adopted and implemented in investigating the propensity of consumers to purchase other identified insurance products, such as crop insurance (Fahad et al., 2018) and earthquake disaster insurance (Xu et al., 2018).

For future works, newer neural network models, such as convolutional neural networks (Zhang et al., 2018), and more advanced learning algorithms, such as deep learning (Shrestha & Mahmood, 2019), could be considered. The approach of finding the best classification threshold suggested by Zou et al. (2016) could be applied to solve the issue of imbalanced datasets. A feature selection method guided by a metaheuristic algorithm, such as the Evolutionary Local Selection (ELSA) method, proposed by Kim et al. (2005) could also be used to further improve the hit rate of the direct mailing problem. The more sophisticated sampling approaches proposed by Darzi et al. (2019), for instance, the hybrid sampling and random oversampling examples (ROSE), could also be applied with feature construction to achieve a higher score in the prediction task.

Additional Information and Declarations

Funding: This work was funded by Sunway University's Individual Research Grant GRTIN-IRG-100-2021.

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: W.T.C.: Data curation, formal analysis, investigation, methodology, writing – original draft; K.H.L.: Conceptualization, funding acquisition, methodology, project administration, supervision, writing – review & editing.



Data Availability: The dataset used in this study is provided by the Dutch data mining company Sentient Machine Research and was used for the CoIL Challenge 2000 (van der Putten et al., 2000a).

References

- Arasu, B. S., Seelan, B. J. B., & Thamaraiselvan, N. (2020). A machine learning-based approach to enhancing social media marketing. *Computers & Electrical Engineering*, 86, 106723. <https://doi.org/10.1016/j.compeleceng.2020.106723>
- Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135, 32-41. <https://doi.org/10.1016/j.neucom.2013.05.059>
- Chen, Y., Lee, J. Y., Sridhar, S., Mittal, V., McCallister, K., & Singal, A. G. (2020). Improving cancer outreach effectiveness through targeting and economic assessments: Insights from a randomized field experiment. *Journal of Marketing*, 84(3), 1-27. <https://doi.org/10.1177/0022242920913025>
- Darzi, M. R. K., Niaki, S. T. A., & Khedmati, M. (2019). Binary classification of imbalanced datasets: The case of CoIL challenge 2000. *Expert Systems with Applications*, 128, 169-186. <https://doi.org/10.1016/j.eswa.2019.03.024>
- Delley, M., & Brunner, T. A. (2020). A segmentation of Swiss fluid milk consumers and suggestions for target product concepts. *Journal of Dairy Science*, 103(4), 3095-3106. <https://doi.org/10.3168/jds.2019-17325>
- Elkan, C. (2000). CoIL Challenge 2000 entry. <http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/ELKANP~1.pdf>
- Elkan, C. (2001). Magical thinking in data mining: lessons from CoIL challenge 2000. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 426-431). ACM. <https://doi.org/10.1145/502512.502576>
- Fahad, S., Wang, J., Hu, G., Wang, H., Yang, X., Shah, A. A., Nguyen, T. L. H. & Bilal, A. (2018). Empirical analysis of factors influencing farmers crop insurance decisions in Pakistan: Evidence from Khyber Pakhtunkhwa province. *Land Use Policy*, 75, 459-467. <https://doi.org/10.1016/j.landusepol.2018.04.016>
- Haykin, S. (2010). *Neural networks and learning machines*. Pearson Education India.

- Joost, M., & Schiffmann, W.** (1998). Speeding up backpropagation algorithms by using cross-entropy combined with pattern normalization. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(2), 117-126. <https://doi.org/10.1142/S0218488598000100>
- Jørgensen, T. M., & Linneberg, C.** (2000). Subspace projections—an approach to variable selection and modeling. In *CoLL challenge 2000: The insurance company case*. Leiden Institute of Advanced Computer Science.
- Kim, Y., Street, W. N., Russell, G. J., & Menczer, F.** (2005). Customer targeting: A neural network approach guided by genetic algorithms. *Management Science*, 51(2), 264-276. <https://doi.org/10.1287/mnsc.1040.0296>
- Ładyżyński, P., Żbikowski, K., & Gawrysiak, P.** (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*, 134, 28-35. <https://doi.org/10.1016/j.eswa.2019.05.020>
- Leong, L. Y., Hew, T. S., Ooi, K. B., & Tan, G. W. H.** (2019). Predicting actual spending in online group buying—An artificial neural network approach. *Electronic Commerce Research and Applications*, 38, 100898. <https://doi.org/10.1016/j.elerap.2019.100898>
- LIAM.** (2019). LIAM re-elects Anusha Thavarajah and Rangam Bir as president and vice-president for the term 2019/2020. <https://www.liam.org.my/index.php/newsmedia-room/media-releasepress-statements/english/1067-liam-re-elects-anusha-thavarajah-and-rangam-bir-as-president-and-vice-president-for-the-term-20192020>
- Lin, H. C. K., Wang, T. H., Lin, G. C., Cheng, S. C., Chen, H. R., & Huang, Y. M.** (2020). Applying sentiment analysis to automatically classify consumer comments concerning marketing 4Cs aspects. *Applied Soft Computing*, 97, 106755. <https://doi.org/10.1016/j.asoc.2020.106755>
- Liu, H., & Motoda, H.** (Eds.). (2008). *Computational methods of feature selection*. CRC Press.
- Ma, L., & Sun, B.** (2020). Machine learning and AI in marketing—Connecting computing power to human insights. *International Journal of Research in Marketing*, 37(3), 481-504. <https://doi.org/10.1016/j.ijresmar.2020.04.005>
- Mahmoudi, H., Farajpour, M., & Afrasiabi, S.** (2021). The preferences of consumers for organic tea: evidence from a stated choice experiment. *Journal of the Saudi Society of Agricultural Sciences*, 20(4), 265-269. <https://doi.org/10.1016/j.jssas.2021.02.006>
- MATLAB.** (2020). *Matlab R2020b*. The MathWorks Inc.
- Motoda, H., & Liu, H.** (2002). Feature selection, extraction and construction. *Communication of IICM*, 5, 67-72.
- Netšajev, E.** (2016). Motor insurance clients risk level evaluation using artificial neural networks and deep learning. https://a-lab.ee/edu/sites/default/files/Netsajev_Bsc.pdf
- Nor Shamsiah, M. Y.** (2018). Governor's remarks at the Malaysian Insurance Institute (MII) Summit - "Innovation in a disruptive era". <https://www.bnm.gov.my/-/governor-s-remarks-at-the-malaysian-insurance-institute-mii-summit-innovation-in-a-disruptive-era>
- Rahangdale, G., Ahirwar, M., & Motwani, M.** (2016). Application of k-NN and Naive Bayes Algorithm in Banking and Insurance Domain. *International Journal of Computer Science Issues*, 13(5), 69. <https://doi.org/10.20943/01201605.6975>
- Seewald, A. K.** (2000). CoLL Challenge 2000 - submitted solution. <http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/SEEWAL~1.pdf>
- Shrestha, A., & Mahmood, A.** (2019). Review of deep learning algorithms and architectures. *IEEE Access*, 7, 53040-53065. <https://doi.org/10.1109/ACCESS.2019.2912200>
- Shtovba, S., & Mashnitskiy, Y.** (2000). *The backpropagation multilayer feedforward neural network based competition task solution*. Vinnitsa State Technical University.
- Tsai, P. H., Lin, G. Y., Zheng, Y. L., Chen, Y. C., Chen, P. Z., & Su, Z. C.** (2020). Exploring the effect of Starbucks' green marketing on consumers' purchase decisions from consumers' perspective. *Journal of Retailing and Consumer Services*, 56, 102162. <https://doi.org/10.1016/j.jretconser.2020.102162>
- Van der Putten, P., & Van Someren, M.** (2000a). CoLL Challenge 2000: The insurance company case. In *Technical Report 2000-09*. Leiden Institute of Advanced Computer Science, Universiteit van Leiden. <https://www.kaggle.com/uciml/caravan-insurance-challenge>
- Van der Putten, P., de Ruiter, M., & van Someren, M.** (2000b). CoLL challenge 2000 tasks and results: Predicting and explaining caravan policy ownership. CoLL Challenge, 2000. <http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/PUTTEN~1.pdf>
- Van der Putten, P., & Van Someren, M.** (2004). A bias-variance analysis of a real world learning problem: The CoLL challenge 2000. *Machine learning*, 57(1), 177-195. <https://doi.org/10.1023/B:MACH.0000035476.95130.99>
- Verma, S., Sharma, R., Deb, S., & Maitra, D.** (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights*, 1(1), 100002. <https://doi.org/10.1016/j.ijimei.2020.100002>
- Vesanto, J., & Sinkkonen, J.** (2000). Submission for the CoLL Challenge 2000. <http://liacs.leidenuniv.nl/~puttenpwhvander/library/cc2000/VESANT~1.pdf>
- Wu, X., & Barbará, D.** (2002). Modeling and imputation of large incomplete multidimensional datasets. In *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery* (pp. 286-295). Springer.
- Xu, D., Liu, E., Wang, X., Tang, H., & Liu, S.** (2018). Rural households' livelihood capital, risk perception, and willingness to purchase earthquake disaster insurance: Evidence from southwestern China. *International Journal of Environmental Research and Public Health*, 15(7), 1319. <https://doi.org/10.3390/ijerph15071319>

-
- Zadrozny, B., & Elkan, C.** (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 694-699). ACM.
<https://doi.org/10.1145/775047.775151>
- Zainuddin, Z., Lai, K. H., & Ong, P.** (2016). An enhanced harmony search based algorithm for feature selection: Applications in epileptic seizure detection and prediction. *Computers & Electrical Engineering*, 53, 143-162.
<https://doi.org/10.1016/j.compeleceng.2016.02.009>
- Zhang, W., Du, Y., Yoshida, T., & Wang, Q.** (2018). DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network. *Information Processing & Management*, 54(4), 576-592.
<https://doi.org/10.1016/j.ipm.2018.03.007>
- Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y.** (2016). Finding the best classification threshold in imbalanced classification. *Big Data Research*, 5, 2-8. <https://doi.org/10.1016/j.bdr.2015.12.001>
-

Editorial record: The article has been peer-reviewed. First submission received on 15 April 2021. Revisions received on 24 May 2021, 21 June 2021 and 28 June 2021. Accepted for publication on 28 June 2021. The editors coordinating the peer-review of this manuscript were Chiew Kang Leng  and Zdenek Smutny . The editor in charge of approving this manuscript for publication was Zdenek Smutny.

Special Issue: Inspiring Technologies for Digital Inclusivity. Selected papers from the 12th International Conference on Information Technology in Asia (CITA'21).

Acta Informatica Pragensia is published by Prague University of Economics and Business, Czech Republic.

ISSN: 1805-4951
