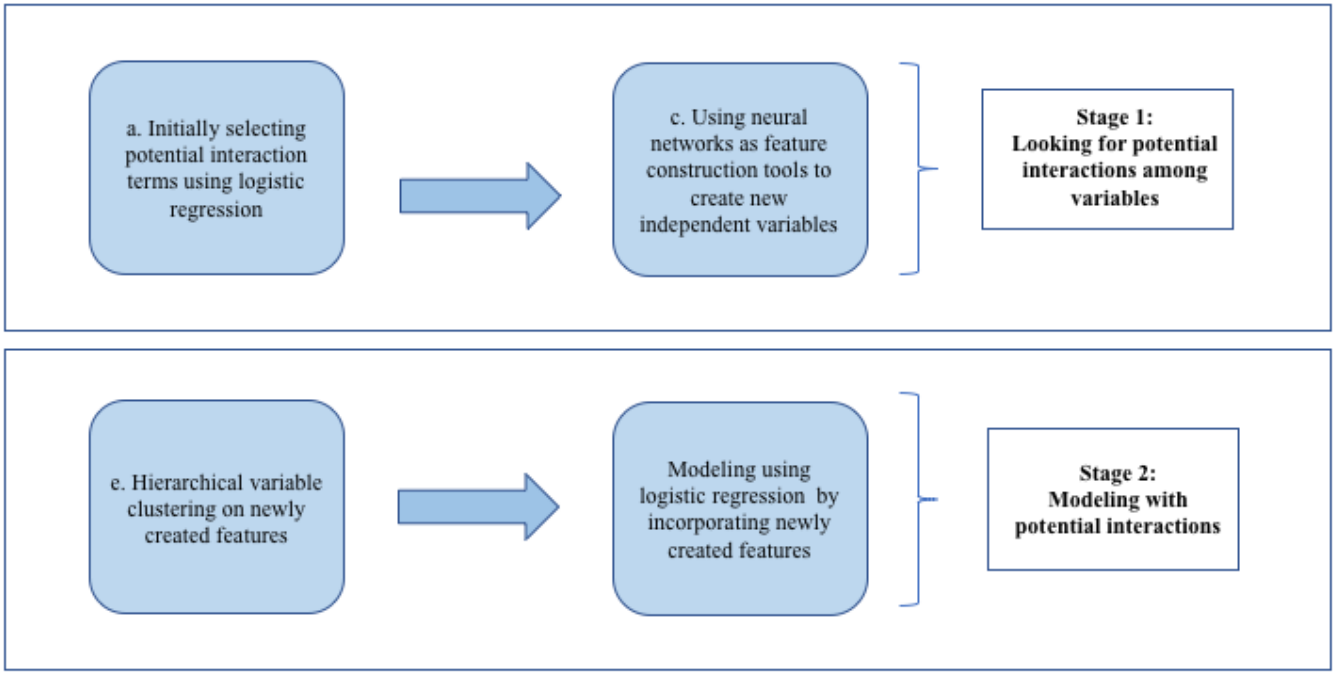
Data Science Research Series

Two-stage hybrid model



BACKGROUND

This paper aims at exploring the performance of the proposed two-stage hybrid model using feature construction algorithms based on artificial neural networks and hence improving the performance of traditional logistic regression in bankcard response modeling. The rationale under the analyses is firstly to use simple neural network structures as new feature construction tools, then the newly created features are used as the additional input variables in logistic regression. To demonstrate the effectiveness of the proposed two-stage hybrid model, its performance is compared with the traditional model by using the credit customer response dataset through cross-validation. It is observed that the proposed model outperforms the traditional logistic regression in terms of classification accuracy, the area under the receiver operating characteristic curve, and KS$ statistics. By creating new features using neural network technique, the underlying nonlinear relationships between variables are identified. Furthermore, when using the neural network as the feature construction tool in the proposed model, we only use a

very simple neural network structure based on some subsets of the data. Consequently, it could overcome the drawbacks of the neural network in terms of its long training time and too complex topology. Therefore, the proposed approach in this paper provides an efficient advance in bankcard response modeling and credit modeling, where traditionally variable selection through logistic regression was the pervasive approach.

APPROACH

As shown in Figure 1, the two-stage hybrid model contains main stages and each stage has three steps included. In the first stage, the main purpose is to look for new features or potential interactions by using the original variables. After data pre-processing, 178 independent variables remain after variable clustering. These variables form 15,753 different pairs of variables.

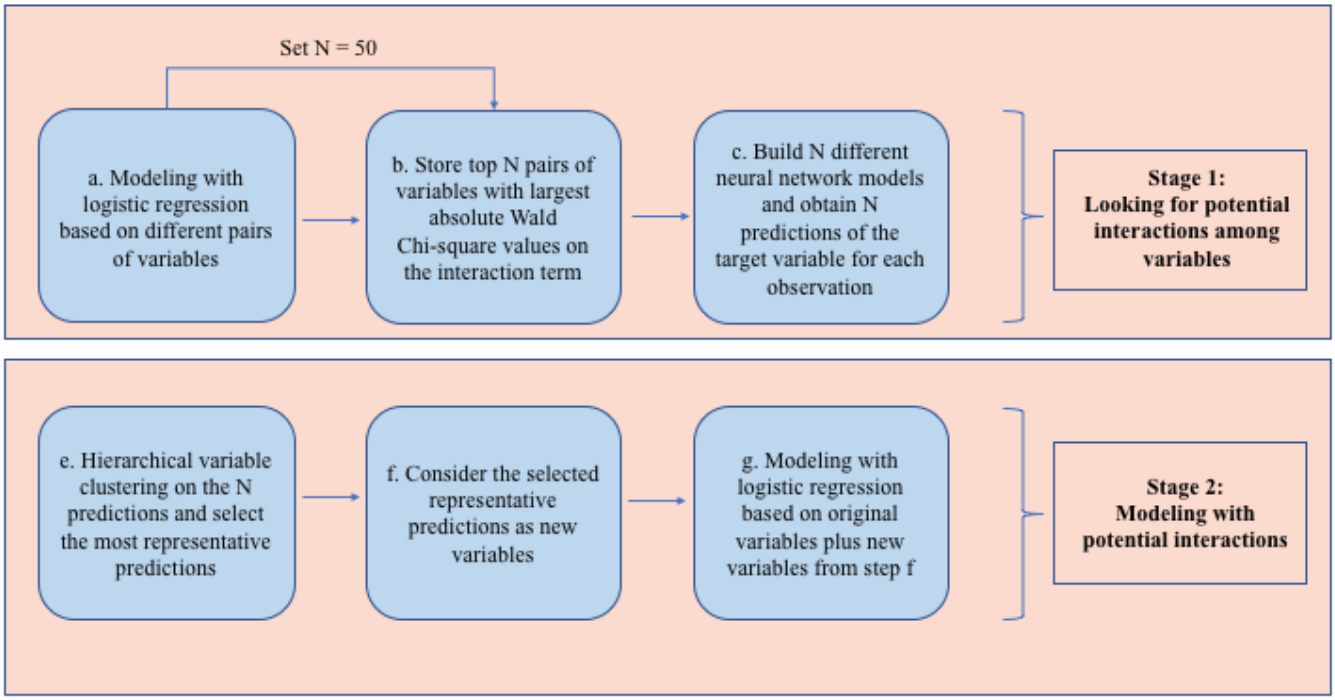


Figure 1. The diagram of the two-stage hybrid model

In step A of Figure 1, the proposed model starts from modeling with logistic regression based on these 15,753 different pairs of variables. Based on the 15,753 logistic regressions, 15,753 Wald Chi-square tests are individually implemented to test the significance of the interaction terms. As a result, the absolute Wald Chi-square values (or, corresponding p values) of the 15,753 interaction terms are recorded. In step B, the top N pairs of variables with highest absolute Wald Chi-square values (corresponds to lowest p values) from their interaction terms in the logistic regression are stored. In this paper, the value of N is set to 50 via experiments. In step C, the selected 50 pairs of variables are used to construct 50 different neural network models on the training set. The predictions yhat on the target variable from each neural network model can be obtained. In this paper, these predictions are denoted as yhat_0, yhat_1, …, yhat_49 and are considered to be the 50 potential new features created based on neural network models.

In the second stage, the main purpose is to build logistic regression models after adding newly created features. In step D, hierarchical variable clustering is applied on the 50 newly created features yhat_0, yhat_1, …, yhat_49 to reduce potential multicollinearity problems. In step E, variables with lowest 1-R^2ratio in each cluster is selected as the representative variable in the current cluster. These representatives are finally used as the new features. In step F, logistic regression is implemented using the 178 variables plus the newly created features.

RESULTS

The performance of the proposed two-stage hybrid model is compared with the traditional logistic regression using accuracy, AUC and KS-statistics on both training and validation sets. The results of traditional logistic regression and the proposed two-stage hybrid model are shown in Table 1 and 2, respectively.

Table 1. Performance of logistic regression. No. feature denotes the number of features used in the model. Accuracy(T) and Accuracy(V) denotes classification accuracy on training and validation sets, respectively. AUC(T) and AUC(V) denotes AUC on training and validation sets, respectively. KS(T) and KS(V) denotes KS statistics on training and validation sets, respectively.

Model	No. feature	Accuracy(T)	Accuracy(V)	AUC(T)	AUC(V)	KS(T)	KS(V)
Full Model	178	0.841	0.827	0.845	0.802	0.499	0.441
1	20	0.834	0.825	0.825	0.792	0.499	0.438
2	18	0.834	0.823	0.824	0.792	0.493	0.439
3	16	0.833	0.823	0.821	0.790	0.490	0.431
4	14	0.831	0.822	0.819	0.787	0.486	0.426
5	12	0.830	0.824	0.817	0.785	0.479	0.421
6	10	0.825	0.824	0.804	0.775	0.474	0.415
7	8	0.821	0.823	0.800	0.768	0.458	0.415
8	6	0.815	0.815	0.777	0.756	0.417	0.395

Table 2. Performance of the proposed two-stage hybrid model. No. feature denotes the number of features used in the model. Accuracy(T) and Accuracy(V) denotes classification accuracy on training and validation sets, respectively. AUC(T) and AUC(V) denotes AUC on training and validation sets, respectively. KS(T) and KS(V) denotes KS statistics on training and validation sets, respectively.

Model	No. feature	Accuracy(T)	Accuracy(V)	AUC(T)	AUC(V)	KS(T)	KS(V)
Full Model	200	0.846	0.831	0.847	0.816	0.529	0.471
1	20	0.840	0.830	0.825	0.801	0.504	0.457
2	18	0.836	0.823	0.824	0.801	0.504	0.457
3	16	0.836	0.825	0.822	0.800	0.503	0.455
4	14	0.835	0.825	0.820	0.800	0.496	0.452
5	12	0.833	0.824	0.818	0.800	0.493	0.451
6	10	0.831	0.823	0.814	0.792	0.484	0.449
7	8	0.827	0.822	0.809	0.790	0.467	0.447
8	6	0.823	0.817	0.801	0.787	0.458	0.442

By comparing results summarized in Tables 1 and 2, it can be concluded that, the two-stage hybrid model has the best bankcard response capability in terms of the classification accuracy, AUC, and KS statistics when the same number of features are selected in the model. Even though the proposed model uses only 6 features, the obtained KS statistics valued 0.442 on validation set in Table 2 is still higher than that from the full model valued 0.441 based on logistic regression in Table 1. Consequently, we can conclude that the proposed two-stage hybrid model outperforms the commonly utilized logistic regression and hence provides efficient alternatives in conducting bankcard response tasks. Neural networks are shown to be a good support to logistic regression as it can efficiently construct new independent variables which may provide valuable information for further managerial and related decision makings.

CONCLUSIONS

Logistic regression and LDA are the most commonly utilized statistical techniques in the credit research domain. However, these techniques only focus on exploring linear relationship among variables and sometimes produce poor bankcard response capabilities. In this situation, neural network, which could handle the non-linear relationship among the variables, is becoming a very popular and powerful choice in dealing with bankcard response problems due to its outstanding bankcard response capability. However, in the meanwhile, neural network is also being criticized for its long training process as well as the complex topological structure. In this paper, the purpose is to propose a two-stage hybrid approach by using neural network as a supporting tool for logistic regression to improve the performance of bankcard response model. The rationale underlying the analyses is firstly using the neural network as a new feature construction tool, then in the second stage, the newly created features are added into the logistic regression to improve the performance.

To demonstrate the effectiveness of the proposed two-stage hybrid bankcard response model, both traditional logistic regression and the proposed model are applied in the credit card customer response dataset using hold-out cross validation approach. The results demonstrate that by identifying new features, the hybrid two-stage model outperforms the commonly utilized logistic regression in terms of classification accuracy, AUC and KS statistics. Furthermore, the neural network structure used in the proposed model is very simple. This can overcome the shortcomings of neural network in terms of its complex topology. More importantly, when using neural network as the feature construction tool in this study, only the subset of the dataset is used. This could reduce the training time in comparison with building neural networks on the entire dataset. Considering that the major drawbacks of neural network are its complex topology and long training process, the proposed model is a good alternative than traditional logistic regression and neural network modeling directly on the entire dataset. The two-stage model demonstrates the capability in creating new while important independent variables and hence can improve the performance of logistic regression while reduce the training time of neural network modeling on the whole data. The framework proposed in this paper and the research findings provide efficient alternatives for future researchers in conducting bankcard response problems.

CITATION FOR FULL ARTICLES

Yan Wang, Sherry Ni, Brian Stone. A two-stage hybrid model by using artificial neural networks as feature construction algorithms. Submitted to the 14th International Conference on Advanced Data Mining and Applications (ADMA 2018).