Post-hoc explanation methods for deep neural networks in choice analysis

Niousha Bagheri Khoulenjani*[1], Milad Ghasri[1], Michael Barlow[1]

[1] School of Engineering and IT (SEIT), UNSW Canberra, ACT, 2600, Australia

SHORT SUMMARY

Deep Neural Networks (DNNs) are accurate and powerful tools for modeling travel decisions. Nonetheless, the black-box characteristic of DNNs has decreased their potential implication in discrete choice modeling. In this study, we investigate the potentials of cutting-edge post-hoc interpretation tools in providing behavioral insight into DNN architectures. We evaluate the relationship between the output probabilities and input features using the Shapely Additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). Using SwissMetro dataset, we demonstrate that the outputs of SHAP and LIME are consistent with theory when the architecture of DNN is designed based on the Random Utility Maximization (RUM) theory. However, for a fully connected DNN architecture, SHAP and LIME do not provide behaviorally interpretable outputs. Additionally, the prediction accuracy shows the DNN model based on RUM avoids overfitting.

Keywords: Discrete choice modeling, Deep Neural Networks, Explainable AI

1. INTRODUCTION

DNN models have become ubiquitous in Intelligent Transport System (ITS) due to their powerful predictive and efficient learning algorithms and fixable modelling structure. ITS, as an integrated transport management system, refers to the use of data communication, information processing, traffic management technologies and Artificial Intelligence (AI) in transport (Chen, Liu et al. 2020). All applications in ITS that rely on DNNs are categorized into four groups of computer vision, time series prediction, classification, and optimization (Wang, Zhang et al. 2019). Most studies of DNNs in ITS belong to the time series prediction group to model variables such as travel time, traffic flow, and traffic speed prediction. On the other hand, the number of applications in classification using DNNs, particularly modelling travel mode choice, is fairly limited (van Cranenburgh and Alwosheel 2019). Traditionally, discrete choice modelers have mostly used econometric methods, including discrete choice models. These models are based on a theoretical foundation with predefined assumptions and underlying relationships between dependent and explanatory variables (Train 2009). Econometric methods are inferior to DNN methods in terms of prediction accuracy. This is one of the main reasons that DNNs have become pervasive in modelling individuals' behavior (Golshani, Shabanpour et al. 2018).

DNNs encompass a wide range of architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) (Goodfellow, Bengio et al. 2016), but the applications of DNNs in discrete choice modeling are mainly limited to the most basic DNN's architecture, called Multi-Layer Perceptron (MLP). Even this basic MLP model is shown to achieve higher prediction accuracy in comparison with traditional discrete choice models (Assi, Nahiduzzaman et al. 2018). Accurate modelling of travel behavior is essential, and it is equally important that high accuracy is resulted from an interpretable model. Despite all the advantages of DNNs, they are considered as complex black-box (non-interpretable) models because of the numerous parameters in the model. In other words, the structure of DNNs is not directly interpretable, as hundreds of

parameters need to be described (Zhao, Yan et al. 2020). Nonetheless, a wide range of interpretable tools known as Post-hoc explainability techniques are proposed to extract knowledge from complex DNN models (Arrieta, Díaz-Rodríguez et al. 2020). The post-hoc approaches justify how and why a DNN model has arrived at its prediction (Lipton 2018).

Almost all applications of DNN in discrete choice modeling have applied a post-hoc approach to interpret the DNN models. For example, Sifringer, Lurkin et al. (2020) proposed a DNN architecture inspired by RUM, consisting of an interpretable and a fully connected part. The authors calculated the importance of each input variable to the fully connected part using a traditional post-hoc approach named the saliency map. However, in addition to finding the importance of input variables, the main purpose of post-hoc approaches is evaluating the output of complex machine learning models such as DNNs. In other words, the output of the DNN model needs to be evaluated using post-hoc approaches to increase the trust in DNN's decisions. Although there is a wide variety of post-hoc explanation approaches in the literature, only a few traditional approaches of post-hoc analysis can be found in recent studies e.g. (Wang, Mo et al. 2020, Wang, Wang et al. 2020, Wang, Wang et al. 2021, Wong and Farooq 2021).
This study seeks to evaluate recent DNN models in discrete choice modeling using the state-of-the-art post-hoc approaches. The performance of two novel post-hoc approaches in the literature, Shapely Additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are tested on a fully connected DNN model and a DNN model based on RUM theory. Our study contributes to evaluating the performance and reliability of the recent DNN models in discrete choice modeling. Furthermore, this comparison and evaluation process will help the researcher to select the appropriate DNN architecture and interpretation approach.

2. METHODOLOGY

Deep Neural Networks and Random Utility Maximisation

The architecture of DNNs models used in discrete choice analysis can be divided into two groups. The first group are those that use a fully connected architecture e.g. (Assi, Nahiduzzaman et al. 2018, Zhao, Yan et al. 2020), and the second group are those who use customized architectures to make the model consistent with behavioral theories such as RUM (Wang, Mo et al. 2020, Wong and Farooq 2021). The Fully connected DNN (F-DNN) connects input variables to the output probabilities through several layers with hundreds of parameters (Goodfellow, Bengio et al. 2016). In F-DNN, the utility of each alternative is connected to the attributes of all alternatives. The second group encompasses specific types of DNNs developed based on RUM which computes the utility of each alternative based on its corresponding attributes.
A recent DNN architecture with alternative-specific utility functions (ASU-DNN) proposed by Wang, Mo et al. (2020), could achieve high accuracy levels in modeling discrete choice data. ASU-DNN contains an input layer, two hidden layers and the output layer. Assume input variables to be a vector of $\boldsymbol{x}$, and the input variables are divided into a vector of alternative specific variables denoted by x_{ik} and a vector of individual specific variables denoted by x_i where $i \in \{1,2,\dots,n\}$ and $k \in \{1,2,\dots,K\}$. Then, consistent with RUM (Train 2009), the utility of each alternative is defined as a function of individual specific variables and its corresponding alternative specific variables as indicated in equation (1).

$$V_k = V(x_i, x_{ik}) = w_k(g_1 \circ g_2 \dots g_{M_2})\left(\left(g_1^{x_{ik}} \circ g_1^{x_{ik}} \dots g_{M_1}^{x_{ik}}\right)(x_{ik}), \left(g_1 \circ g_2 \dots g_{M_1}\right)(x_i)\right) \quad (1)$$

In this equation M_1 and M_2 are the number of neurons in the first and second hidden layers respectively; $g(t) = Max(t, 0)$ is the RELU activation function, and w_k is the vector of parameters

to be estimated. In the last layer, the Softmax activation function, shown in equation (2) is applied to the utilities in order to calculate the output probabilities (Goodfellow, Bengio et al. 2016):

$$S(V_{ik}) = \frac{e^{V_{ik}}}{\sum_j e^{Vij}} \tag{2}$$

Interpretation methods

The demands for interpretability in DNNs have increased in recent years (Arrieta, Díaz-Rodríguez et al. 2020). Therefore, many approaches as the post-hoc expainability methods for DNNs, have been developed. The existing post-hoc explainability approaches fall into six categories of text explanation, visual explanation, local explanation, explain by examples, explain by simplification, and feature relevance explanation. For further details about each category, the reader is referred to Arrieta, Díaz-Rodríguez et al. (2020). In this study, we apply two approaches, Shapely Additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) from the local explanation category.

SHAP is a game theory interpretation method of machine learning methods that evaluates the negative and positive impact of input variables (Lundberg and Lee 2017). For a given input X and a DNN model $f(X)$, SHAP utilizes an Explanation Model (EM) to evaluate the contribution of each input variable x_i to the model f. EM sets up a relationship between x_i and the model outputs. The parameters of this model are called SHAP value denoted by φ_i. SHAP values are defined as the weighted average of the marginal contributions over all possible coalitions $|F|!$ and are calculated as indicated in equation (3).

$$\varphi_i(f) = \sum_{\{S \subseteq F\} \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f(x_{S \cup \{i\}}) - f(x_S) \right] \tag{3}$$

In this equation, F is the total number of features, and S is a subset of F. $f(x_{S \cup \{i\}})$ is the model output using feature i and features in S, and $f(x_S)$ is the model output using features in S but without feature i. As computing the exact value of φ_i is challenging, several methods have been introduced to approximate SHAP values (Lundberg and Lee 2017). In this study, we apply Kernel SHAP as it is a model-agnostic method that can be used for all types of machine learning models, and it is a reasonable approach when number of input variables are small in the dataset (Lundberg and Lee 2017).

Similar to SHAP, the LIME belongs to the local interpretation category that measures the impact of input variables on the variations of model output (Ribeiro, Singh et al. 2016). LIME generates new datasets around an observation x consisting of the corresponding outputs of the model. Then, an explainable model g is trained on the new dataset that is weighted by the proximity of the sample observations. With the new explainable model g and trained DNN model f, it is possible to provide a rough estimate of the contribution of input variable x to the model f. To accomplish this, the following objective function is minimized:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g) \tag{4}$$

Where G denotes the set of all interpretable models, and $\Omega(g)$ is the complexity of model g. π_x is the proximity measure between generated data to sample x. L measures how unfaithful g is in the approximation of f in the locality defined by π_x.

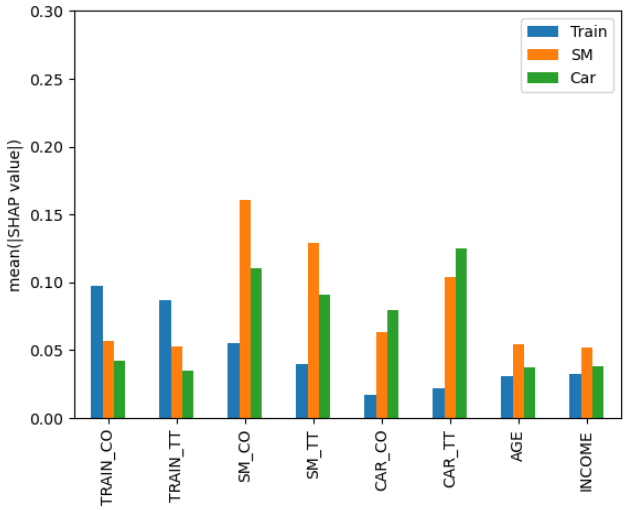
Although it is expected LIME and SHAP yield similar results, they have different structures in interpreting models. While LIME generates a perturbed dataset to fit an explainable model, SHAP requires an entire sample to approximate SHAP values.

3. RESULTS AND DISCUSSION

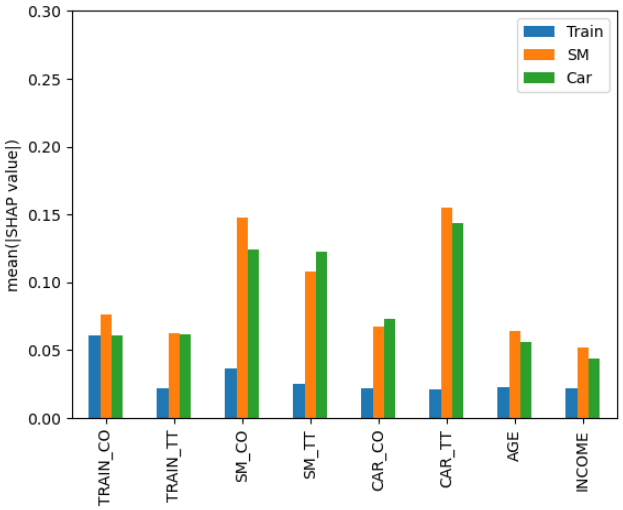
This study uses the Swissmetro dataset for evaluating the interpretation of DNNs. This dataset was compiled in Switzerland in 1998 (Bierlaire, Axhausen et al. 2001). It contains 1192 respondents who were asked to choose their preferred transportation mode among three alternatives of train, Swissmetro and car. This dataset contains 9,036 observations after cleaning. In the current study, Travel Time (TT), Travel Cost (CO), AGE and INCOME are selected among available variables for choice analysis.

The two models of ASU-DNN and F-DNN are developed using the Swissmetro dataset. Then the two methods are SHAP and LIME are used to provide insight into how these models make predictions. SHAP and LIME also show which features are most important in both models separately. In the end, the performance of ASU-DNN and F-DNN is compared through accuracy and log-likelihood. In this experiment, ASU-DNN includes two layers with $M_1 = M_2 = 100$. Similar to ASU-DNN, F-DNN includes 2 layers with 100 neurons in each layer.

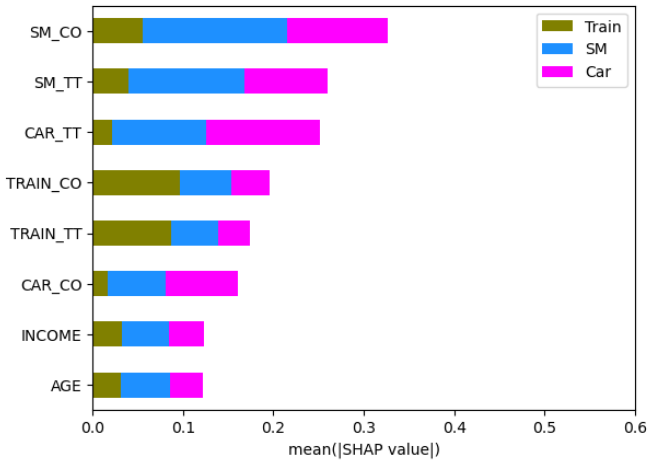
In this study, the impact of an input variable on each class will be calculated using SHAP and LIME. In contrast to an image input with a 2D set of pixels (which is the common application for post-hoc explanation methods), the position of the input variables always has the same meaning. For example, if the first input variable of all observations is age, then the first LIME and SHAP values will always be the impact of a passenger's age on each class. The average of absolute SHAP values for all features in each class is reported in Figure 1. The first line illustrates the SHAP values in ASU-DNN and F-DNN corresponding to each class, and the second line ranks the summation of SHAP values for each feature. As shown in plot (a), features related to each utility have the most contribution to probability of that utility. For example, Swissmetro Cost (SM_CO) and Swissmetro Travel Time (SM_TT) have the highest impact on Swissmetro. However, from (b), there is no specific connections between input variables and utilities. For example, Train Travel Time (Train-TT) has the highest impact on Swissmetro and Car. Plots (c) and (d) demonstrate the overall importance of input variables on the output probabilities in ASU-DNN and F-DNN. Both models show that SM_CO, SM_TT and Car Travel Time (CAR_TT) have the highest impact on the mode choice decision. Also, INCOME and AGE have the least feature importance in this classification task.



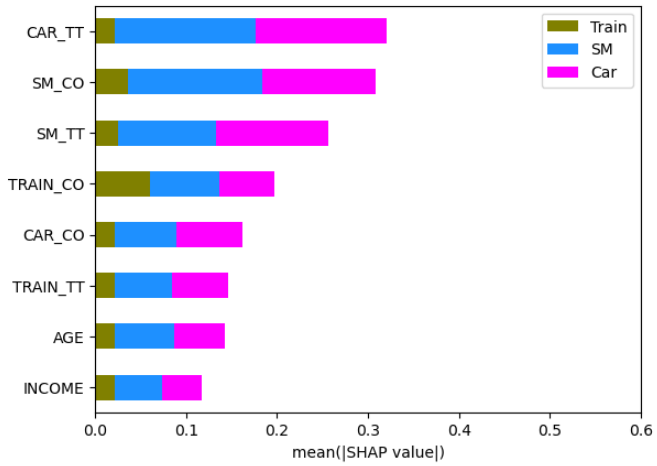
(a) ASU-DNN SHAP values



(b) F-DNN SHAP values



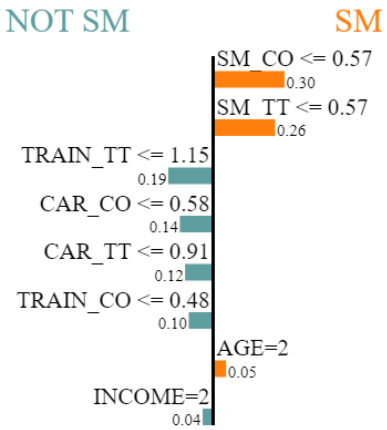
(c) Ranking SHAP values of ASU-DNN



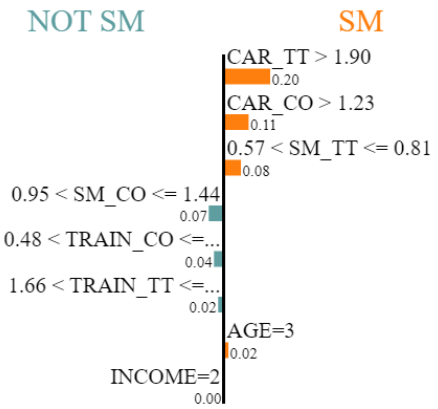
(d) Ranking SHAP values of F-DNN

Figure 1: Interpretation results of ASU-DNN and F-DNN using SHAP

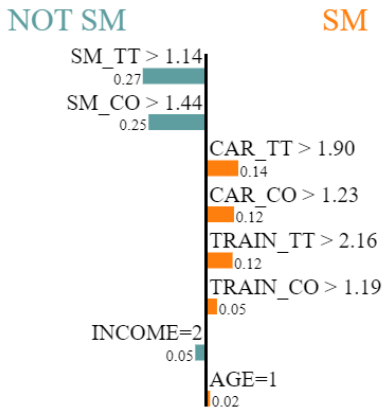
Figure 2 demonstrates the local interpretability analysis with LIME for instance numbers 130 and 2137 (randomly selected). All four bar graphs reflect the contribution of each feature to the classification of respective 130 and 2137 instances. The true classes of both instances are Swiss-metro. For ASU-DNN, both (e) and (g) showed that SM_TT and SM_CO have the most contribution in the output probability SM. On the contrary, (f) and (h) shows CAR_TT has the most impact on the probability of Swissmetro in F-DNN model.



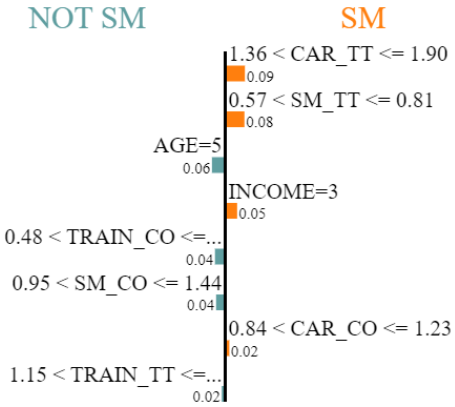
(e) ASU-DNN LIME values
Instance = 130



(f) F-DNN LIME values
Instance = 130



(g) ASU-DNN LIME values
Instance = 2187



(h) F-DNN LIME values
Instance = 2187

Figure 2: Interpretation results of ASU-DNN and F-DNN using LIME

For training the DNN models, the dataset is divided into 70% training dataset and 30% test dataset. Table 1 shows the number of parameters, loglikelihood and accuracy of F-DNN and ASU-DNN for the test and train datasets. Although the performance of F-DNN in training is pressive, ASU-DNN outperforms it in terms of accuracy and loglikelihood on the test dataset. This indicates that overfitting is less likely when the RUM theory is implemented in the DNN architecture. In ASU-DNN, the number of parameters is reduced from 21,403 to 1,803 which means in this architecture, connections that are not supported by the theory are removed from the model. Therefore some spurious correlations that potentially could cause overfitting are avoided in this architecture (Yang, Chou et al. 2022).

Table 1: The loglikelihood and accuracy of F-DNN and ASU-DNN for the test and train datasets

		Train dataset		Test dataset	
	Number of parameters	**Loglikelihood**	**Accuracy**	**Loglikelihood**	**Accuracy**
F_DNN	21,403	1330.17	98.79	8861.69	67.20
ASU_DNN	1,803	3519.97	74.92	1831.45	71.78

4. CONCLUSIONS

In this study, DNN models for choice modelling are analyzed using state-of-the-art interpretation techniques. It is crucial to clarify how predictions are formed by DNN models when it comes to

artificial intelligence in discrete choice modeling. The contribution of this research is the use of recent post-hoc approaches to uncover new insights from the recently developed DNN model based on RUM theory. We apply SHAP and LIME, two of the most recent interpretation approaches, to evaluate the performance of ASU-DNN and F-DNN. The contributions of each input variable in both models F-DNN and ASU-DNN are retrieved using SHAP and LIME. The interpretation analysis of DNN models shows that DNNs with a theory-based architecture (ASU-DNN) have more consistency with the RUM theory, in contrast with conventional DNN models (F-DNN). Additionally, the results revealed that ASU-DNN could reduce overfitting by eliminating unsupported connections.

This research indicates a new research direction of using post-hoc analysis in discrete choice modeling. Future studies can concentrate on extracting information from DNN models using other post-hoc approaches such as DeepLift.

REFERENCES

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina and R. Benjamins (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion **58**: 82-115.

Assi, K. J., K. M. Nahiduzzaman, N. T. Ratrout and A. S. Aldosary (2018). "Mode choice behavior of high school goers: Evaluating logistic regression and MLP neural networks." Case studies on transport policy **6**(2): 225-230.

Bierlaire, M., K. Axhausen and G. Abay (2001). The acceptance of modal innovation: The case of Swissmetro. Swiss Transport Research Conference.

Chen, C., B. Liu, S. Wan, P. Qiao and Q. Pei (2020). "An edge traffic flow detection scheme based on deep learning in an intelligent transportation system." IEEE Transactions on Intelligent Transportation Systems **22**(3): 1840-1852.

Golshani, N., R. Shabanpour, S. M. Mahmoudifard, S. Derrible and A. Mohammadian (2018). "Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model." Travel Behaviour Society **10**: 21-32.

Goodfellow, I., Y. Bengio and A. Courville (2016). Deep learning, MIT press.

Lipton, Z. C. (2018). "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue **16**(3): 31-57.

Lundberg, S. M. and S.-I. Lee (2017). "A unified approach to interpreting model predictions." Advances in neural information processing systems **30**.

Ribeiro, M. T., S. Singh and C. Guestrin (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Sifringer, B., V. Lurkin and A. Alahi (2020). "Enhancing discrete choice models with representation learning." Transportation Research Part B: Methodological **140**: 236-261.

Train, K. E. (2009). Discrete choice methods with simulation, Cambridge university press.

van Cranenburgh, S. and A. Alwosheel (2019). "An artificial neural network based approach to investigate travellers' decision rules." Transportation Research Part C: Emerging Technologies **98**: 152-166.

Wang, S., B. Mo and J. Zhao (2020). "Deep neural networks for choice analysis: Architecture design with alternative-specific utility functions." Transportation Research Part C: Emerging Technologies **112**: 234-251.

Wang, Y., D. Zhang, Y. Liu, B. Dai and L. H. Lee (2019). "Enhancing transportation systems via deep learning: A survey." Transportation research part C: emerging technologies **99**: 144-163.

Wong, M. and B. Farooq (2021). "ResLogit: A residual neural network logit model for data-driven choice modelling." Transportation Research Part C: Emerging Technologies **126**: 103050.
Yang, Y.-Y., C.-N. Chou and K. Chaudhuri (2022). "Understanding rare spurious correlations in neural networks." arXiv preprint arXiv:.05189.
Zhao, X., X. Yan, A. Yu and P. Van Hentenryck (2020). "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models." Travel behaviour society **20**: 22-35.