Random Forest with Transfer Learning: An Application to Vehicle Valuation

Changro Lee
Department of Real Estate, Kangwon National University, Chuncheon, Republic of Korea
Email: spatialstat@naver.com

***Abstract*—In contrast to their outstanding success in dealing with unstructured data, such as images and natural language, machine-learning models have not shown noticeable achievements in utilizing structured data, i.e., tabular-format data. Part of their excellent performance with unstructured data comes from their capability of transfer learning, which has rarely been utilized in the fields of structured data. In this study, a random forest is used to estimate vehicle prices in the South Korean automobile industry. To enhance the performance of the random forest, when the input variables are structured data and part of them are high-cardinality categorical types, entity embedding vectors are created from a neural network, and reused in the random forest training. This study demonstrates that information in structured data can be efficiently extracted using the entity embedding technique and effectively reused in different but related tasks in the form of transfer learning.**

***Index Terms*—machine learning, transfer learning, structured data, entity embedding, random forest**

I. Introduction

Machine learning is rapidly expanding in various application areas, and has been used with great success in a variety of applications, such as computer vision, voice recognition, and Natural Language Processing (NLP). Machine learning has achieved particularly noticeable advances and commercial success in areas prolific with unstructured data, which are not stored in a structured database format, including image, text, audio, and video data.

In contrast to their outstanding success with unstructured data, machine learning models have not achieved noticeable performance when utilizing structured data (i.e., tabular-format data with rows and columns). Their poor performance can be explained by two factors: First, an artificial neural network, the most popular implementation algorithm for machine learning, is known to be extremely efficient in processing and learning unstructured data, such as images and videos. Structured data can take less advantage of this technique. Second, transfer learning, one of the unique characteristics of machine learning models, has been rarely utilized in the fields of structured data, despite the fact that it enables rapid training and improved performance when data are insufficient.

This study attempts to increase the performance of a machine learning model (i.e., random forest), where the input variables are structured data and most of them are high-cardinality categorical types. First, a random forest is specified and fitted to vehicle registration records to estimate the car prices. Then, the entity embedding vectors are created from neural network training to handle the categorical variables more efficiently. The random forest equipped with the entity embedding vectors is fitted again to the same dataset, and its predictive accuracy is compared to that of an ordinary random forest.

A vehicle valuation service is frequently provided before a car transaction, and an accurate valuation is beneficial for both the buyer and the seller, and a healthy vehicle transaction culture can be achieved through a fair valuation process. It also plays a key role for a variety of stakeholders, including governments for tax assessments of personal property, and banks for the approval of auto loans.

The approach adopted in this study, i.e., enhancing the performance of machine learning models utilizing the entity embedding technique, would significantly promote the adoption of transfer learning in the automotive industry. Application areas dealing with structured data have received less attention compared to those handling unstructured data; thus, this study is expected to boost interest in adopting transfer learning for training structured data.

The remainder of this paper is organized as follows. Section II presents background information on transfer learning and an entity embedding approach for structured data. Section III describes the data used, the architecture of the random forest, and the embedding vectors learned via neural network training. Then, the prediction results of the car prices and implications are discussed in Section IV, and finally, a summary of the study and conclusions are presented in Section V.

II. Literature Review

A. Transfer Learning

Transfer learning refers to the situation in which what has been learned in one setting is exploited to improve generalization in another setting [1]; i.e., a machine learning model developed for one task is reused as the

Manuscript received February 9, 2022; revised June 10, 2022.

doi: 10.12720/jait.13.4.326-331

starting point for a model of a second task. The benefits of transfer learning are twofold. First, it can overcome the data insufficiency problem frequently faced by research practitioners. Machine learning models can be dramatically empowered by a large database; however, for the same reason, they can break down easily when sufficient data are not available for a task. Thus, transfer learning allows practitioners to deal with sparse data problems by leveraging pre-trained models; i.e., by using an already-learned knowledge set. Second, it is generally inadvisable to train a model from scratch, and transfer learning can save the time and resources needed to train a model by reusing components of already-tested models.

Transfer learning is performed most actively in computer vision applications. In general, imagery data are not easy to collect in a massive manner, both in terms of the resources required and legal issues, such as copyright problems. Various pre-trained models have already been developed in the form of Convolutional Neural Networks (CNNs), and are provided through convenient Application Programming Interfaces (APIs). Representative pre-trained CNNs include VGG, ResNet, and Inception. They were often trained on more than one million images, and demonstrated state-of-the-art performance in a range of image recognition and object detection tasks: VGG [2], [3], ResNet [4], [5], and Inception [6], [7].

Transfer learning is also being adopted universally in NLP. Machine learning language models require a vast amount of text corpora to learn the meanings contained in written documents or spoken languages. In NLP, pre-trained language models are mainly developed in high-resource settings, and word embedding plays a vital role in the realization of transfer learning. The logic behind word embeddings is that words are represented as low-dimensional vectors that capture both the syntax and semantics of the text corpus. Popular pre-trained models of word embeddings include Word2Vec and Glove. These word embeddings are frequently used as the first text-processing layer in machine learning models in sentiment analysis and text classification: Word2Vec [8], [9] and Glove [10], [11].

In contrast to the abundant presence of popular pre-trained models in the field of unstructured data, such as images and text, few studies have been reported on transfer learning in the areas of structured data, let alone the existence of popular pre-trained models specialized for structured data.

B. Entity Embedding Approach to Structured Data

Although transfer learning is frequently utilized for unstructured data, it has rarely been employed for structured data, and this study was motivated to fill this research gap. While unstructured data are commonly used in computer vision and NLP, the dominant data type found in the social sciences is structured data. Business performance records, service-history records of products, and real estate registries are certain examples frequently observed in a variety of businesses. This study focuses on the exploitation of structured data.

The variable types in structured data comprise continuous and categorical variables. Continuous variables, such as vehicle price and energy capacity (cc), can be represented by real numbers. Categorical variables, such as vehicle body color and type, can be represented by integers; however, the integers are only used for convenience to label the different states and have no meaningful information in and of themselves.

Categorical variables are abundant in the automotive industry. For instance, vehicles are often classified as sedans, station wagons, or sport-utility vehicles. Car dealers promote their listings by providing a variety of categorical information: desirable body color, absence of physical damage, absence of broken parts such as a seat warmer, presence of post-factory modifications such as multimedia systems, and numerous presence-absence type variables (with or without alloy wheels, and with or without sunroofs). Categorical scales are pervasive in business practices, including those of the automotive industry, and it is not surprising that a variety of categorical variable-specialized methods have been developed, ranging from the simplest form of contingency tables to matched-pair analysis to logistic regression models.

These categorical variables cannot be directly fed into a model; they must first be converted into numerical representations. The simplest technique is to convert each element in a categorical variable into a separate binary dummy variable, which is often called the one-hot encoding approach. This method has been most frequently used in valuation literature [12]-[14]. However, creating dummy variables may not be the most effective method of extracting information from a categorical variable. This approach has two shortcomings. First, when there are several high-cardinality variables, such as a vehicle model number or ZIP code, one-hot encoding often places excessive demands on computational resources. In addition, it treats different values of categorical variables completely independently of each other and does not take into account the informative relations between them.

This study utilizes an entity embedding approach for transfer learning in structured data. Entity embedding is a technique for mapping categorical values to a multi-dimensional space with fewer dimensions than the original number of levels, where values with similar function outputs are close to each other [15], [16]. As explained earlier, this approach is widely used in NLP fields because words can be viewed as an agglomeration of high-cardinality categorical variables. This representation allows the intrinsic properties of each categorical value to be extracted, and more efficiently used in a quantitative model.

By adopting the entity embedding approach, especially for structured data with dominant high-cardinality categorical variables, several problems can be mitigated. First, it avoids an unrealistic amount of computational resource consumption owing to the traditional one-hot encoding of high-cardinality variables. Second, it treats different values of categorical variables in a meaningful manner, instead of processing them completely independently of each other. Third, it avoids the feature engineering step because embeddings, by nature,

intrinsically group similar values together, removing the need for domain experts to learn the relationships between values in the same categorical variable. Finally, the learned embeddings can be visualized using a dimensionality reduction technique, which can provide additional insights to business practitioners.

III. Random Forest and Embedding Vectors

A. Dataset

The dataset used is the registration records of vehicles manufactured from January 2016 through March 2019. All cars must be registered in the vehicle registry at the time of shipping from the factory, and these registration records include information, such as the factory invoice price, manufacturer, and vehicle type. The dataset includes vehicle attributes of 3,090,818 automobiles,[1] and Table I lists the descriptive statistics of the dataset.

TABLE I. Descriptive Statistics of 3,090,818 Automobiles (Manufactured Jan. 2016–Mar. 2019)

Variable	Min.	Mean	Median	Max.
Invoice price (KRW)	4,269,000	22,526,000	20,446,000	500,000,000
Passenger capacity (number of passengers)	2	5.1	5	15
Engine capacity (cc)	998	1,846	1,995	15,983
Car name	Morning: 300,731 (10%), Porter II: 276,899 (9%), Avante: 259,733 (8%), Spark: 230,526 (7%), Sonata: 201,638 (7%)			
Model number	DC487: 190,496 (6%), TA51BG-S:104,751 (3%), AD4DB6-C-O: 88,227 (3%), T4N20-4D: 76,600 (2%), ZA69S: 62,959(2%)			
Manufacturer	Hyundai: 1,222,288 (40%), KIA: 768,190 (25%), GM: 380,142 (12%), Samsung: 346,122 (11%), Mercedes Benz: 49,086 (2%)			
Vehicle type	Passenger car: 2,621,909 (85%), Truck: 440,931 (14%), Van: 27,978 (1%)			
Usage	Non-business: 2,727,154 (88%), Business: 358,443 (12%), Governmental use: 5,221 (0%)			
Fuel type	Gasoline: 1,693,445 (55%), Diesel: 1,117,028 (36%), LPG: 186,113 (6%), Hybrid: 83,069 (3%), Miscellaneous: 11,163 (0%)			
Year of manufacture	2016: 1,177,734 (38%), 2017: 827,746 (27%), 2018: 703,894 (23%), 2019: 381,444 (12%)			

Note: Only primary levels are presented in the car name, model number, and manufacturer for readability. For the entirety of levels in the car name, see Fig. 1.

The median invoice price for vehicles in the dataset is 20,446,000 KRW (approximately 18,500 USD). The median passenger capacity and engine capacity are 5 person and 1,995 cc, respectively. Car name includes 43 levels; the relatively common names are Morning and Porter II. All the primary car names presented in Table I are from vehicles produced domestically, that is, by South Korean manufacturers. Model number refers to a vehicle model code, which is used to identify the engine type, body shape, and thus, can vary, even within the same car name. It includes 142 levels, and expert knowledge is needed to fully understand them. Manufacturer includes nine values, with the most frequent makers being Hyundai, KIA, GM, and Samsung, as presented in Table I.

The dominant vehicle type is a passenger car (85%), most vehicles are for non-business purposes (88%), and gasoline (55%) is the most frequent fuel type. Year of manufacture includes four values from 2016 and 2019 and was treated as a continuous variable.

B. Specification of a Random Forest

A random forest is used in this study. It is widely used in machine learning [17], and has the excellent advantage of being able to natively process categorical variables. It fits data by splitting them based on an input variable; if the input variable is categorical, the split is accomplished using the levels or elements of the variable. This approach has been used frequently in real estate valuation [18]-[20].

Random forest is an ensemble learning method that constructs multiple decision trees at training time and outputs the mean prediction of individual decision trees for regression tasks.[2] It applies a *bagging* procedure to decision tree building models. It not only randomly chooses samples for each tree's training, but also randomly selects a subset of variables, when choosing each split in each decision tree [21]. In this study, the random forest was implemented with 50 trees, its training criterion was the mean squared error, and the maximum depth of a tree was six.[3]

Nine input variables were used to estimate the vehicle price in this study, as presented in Table II.[4] The target variable was the invoice price, which was log-transformed to alleviate a right-skewed distribution and then standardized to have a mean of zero and a standard deviation of one. Then, the 3,090,818 vehicles were randomly divided into training data (80%, 2,472,654 vehicles) and test data (20%, 618,164 vehicles) for subsequent analysis.

TABLE II. Input Variables

Variable	Variable type	Number of levels
Car name	Categorical	43
Model number	Categorical	142
Manufacturer	Categorical	9
Vehicle type	Categorical	3
Usage	Categorical	3
Fuel type	Categorical	5
Year of manufacture	Continuous	-
Passenger capacity	Continuous	-
Engine capacity	Continuous	-

[1] The initial dataset included over four million automobiles. The following records were removed during the data pre-processing: records with missing values, redundant records, and records related to extremely rare vehicles (less than 30 in frequency), in terms of the car name, model number, and manufacturer.

[2] For classification tasks, it outputs the class given by the mode of the individual decision trees.

[3] These hyperparameters were determined after reviewing the performances produced by the 5-fold validation dataset.

[4] In used-car markets, the mileage and the physical damage are important factors affecting a vehicle price. However, these factors were not considered in this study because the invoice price at the time of shipping from the factory, i.e., the price for a new car, was used as a target variable.

C. Embedding Vectors Learned from Neural Network Training

As presented in Table II, a few categorical variables are characterized by high-cardinality, such as the car name, model number and manufacturer. The one-hot encoding technique processes these variables very inefficiently, although the random forest is capable of natively processing categorical variables. Hence, they must be represented in the form of an embedding vector. The embedding vector utilized in this study was obtained from the training results of a neural network and reused in the fitting of the random forest.

This network is a fully connected layer neural network, with the following architecture: Nine input variables were created, and then embedding layers for the three input variables corresponding to the high-cardinality categorical variables were additionally created and added to the architecture. A suitable number of dimensions had to be determined for each embedding layer, and the prediction performance for various dimensional sizes was reviewed using the usual cross-validation process. The numbers of dimensions given to each categorical variable through this cross-validation are 20, 40, and 4 for the car name, model number, and manufacturer, respectively. Finally, three hidden layers were added to the end of the architecture to include more parameters to capture minor data nuances. The final architecture of the neural network used to obtain the embedding vectors is presented in Table III.

TABLE III. ARCHITECTURE OF THE NEURAL NETWORK USED TO OBTAIN EMBEDDING VECTORS

Nine input variables	Car name, Model no., Manufacturer, Vehicle type, Usage, Fuel type, Year of manufacturer, Passenger capacity, Engine capacity
Three entity embedding layers	Car name (20 dimensions), Model no. (40 dim.), Manufacturer (4 dim.)
Dense layer 1	64 neurons
Dense layer 2	32 neurons
Dense layer 3	4 neurons
Output layer	1 neuron (vehicle price)

The resultant embedding vectors take the following form: a 43×20 matrix for the car name, a 142×40 matrix for the model number, and a 9×4 matrix for the manufacturer. Table IV demonstrates what the embedding vectors of the car name look like, and Fig. 1 illustrates the embedding represented in a 2D space, where the x-axis and y-axis are the first and second vectors of the 20 embedding vectors.

TABLE IV. EMBEDDING VECTORS OF THE CAR NAME LEARNED FROM NEURAL NETWORK TRAINING

Car name	vector1	vector2	vector3	…	vector18	vector19	vector20
BMW 520d	-0.0023	-0.0029	0.0000	…	-0.0002	-0.0133	0.0237
CHEVROLET IMPALA	-0.0788	-0.0036	0.0019	…	-0.0316	0.2470	-0.2797
E300	-0.0088	0.0424	-0.0400	…	0.0385	0.0309	0.0047
…	…	…	…	…	…	…	…
Tivoli	0.0017	-0.0440	-0.0256	…	-0.0069	0.0375	-0.0175
Tivoli Air	-0.0010	-0.0400	-0.0298	…	-0.0341	-0.0429	0.0084
PORTER 2	-0.0416	-0.0052	-0.0213	…	0.0117	0.0052	-0.0230

Note: Shown partially for readability

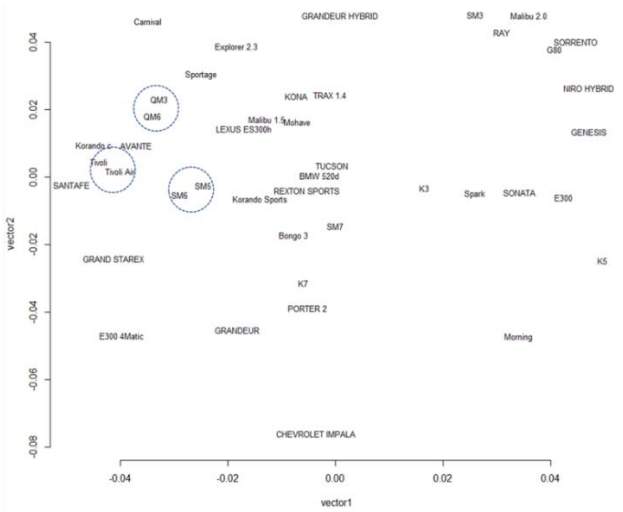

Figure 1. Car name embedding in a 2D space using the first and second vectors of the 20 embedding vectors.

The first and second embedding vectors shown in Fig. 1 may be interpreted as follows: SM5 and SM6, indicated by the dotted circle, are variants of the same vehicle model produced by the same car manufacturer. Other dotted circles also denote car names that are generally recognized as similar vehicles by both public and industry experts. However, interpreting the learned embedding involves subjective judgments, and we do not attempt to interpret its meanings because our primary goal is to reuse it in a subsequent model and achieve a higher performance than a baseline model.

IV. RESULTS AND DISCUSSION

A. Results

The Root Mean Squared Error (RMSE) was used to compare the model performance, as shown in the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2} \quad (1)$$

where y denotes the observed invoice price and $\hat{y}$ denotes the estimated price from the random forest.

The dataset was randomly split into a training set (80%) and a test set (20%). This split was repeated 100 times, and the random forest was fitted to the corresponding split dataset 100 times. Table V lists the model performance based on the test dataset. One-hot encoding was adopted for the categorical variables in the ordinary random forest. For the random forest with transfer learning, embedding vectors trained by a neural network were used as input variables, instead of the original variables in a categorical form. As shown in Table V, the RMSE of the random forest with transfer learning was significantly reduced, demonstrating the potential capability of transfer learning.

TABLE V. COMPARISON OF MODEL PERFORMANCE

	Random forest	Random forest with transfer learning
RMSE	0.40–0.41	0.26–0.27

The square of the RMSE, that is, $RMSE^2$, can be considered the variance, and the F-statistics can be generated accordingly. The performance difference in the table can be evaluated with the alternative hypothesis H_a: $RMSE^2$ without transfer learning ≠ $RMSE^2$ with transfer learning. The F-statistics that were obtained ranged from 2.2 to 2.5, which are greater than the critical value (≈1.0) at the significance level of 0.05, and the alternative hypothesis can be accepted. The corresponding p-values were smaller than 0.001. Based on the magnitude of the F-statistics and p-values, it can be concluded that the random forest with transfer learning outperforms the ordinary random forest.

B. Effects of Transfer Learning

The goodness-of-fit of the two models for the test dataset is depicted in Fig. 2 (one example result of the 100 fittings). Both the predicted and observed prices were normalized, and the predicted prices appear to follow the observed prices approximately well.

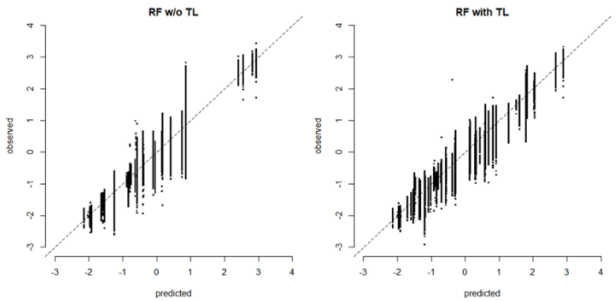

Figure 2. Goodness-of-fit in models for the test dataset (one result of the 100 fittings). Note: RF and TL stand for random forest and transfer learning, respectively.

The predicted values of both models show a common pattern: The models predict vehicle prices discretely, forming distinct clusters for each predicted value. This pattern is understandable because categorical variables are dominant in the input dataset, and the random forest is fitted to the data by creating a split, based on the levels in a categorical variable.

This discontinuous pattern was significantly mitigated for the prediction values estimated by utilizing transfer learning, as shown in the right panel of Fig. 2. This improvement could be attributed to the embedding vectors employed in the random forest. The embedding approach assigned similar numerical values to similar elements in a categorical variable, elevating the continuous relationships between those elements.

Transfer learning in the form of entity embedding can be exploited for various purposes. First, entity embedding vectors can be learned in advance for large datasets consisting of high-cardinality variables, and reused for tasks that suffer from a small data problem. Second, the entity embedding approach can be adopted in algorithms that can primarily use continuous variables; a clustering algorithm or a principal component analysis can efficiently handle continuous variables. By accepting the entity embedding approach, categorical values can be converted into continuous values and fed into a continuous value–friendly algorithm.

V. CONCLUSION

Although machine learning models have achieved brilliant success when utilizing unstructured data, such as images or text, the same does not apply to the exploitation of structured data. In addition, the presence of categorical variables with high-cardinality makes it more difficulty for machine learning models to efficiently process structured data. The most frequent approach to such categorical data has been to use one-hot encoding in a quantitative model; however, this approach is inefficient for high-cardinality categorical variables.

This study attempted to enhance the performance of a random forest through transfer learning. The entity embedding vectors were learned through neural network training, and were reused in the form of transfer learning for a subsequent model. First, the registration records of vehicles were chosen for the analysis, and a random forest was specified with 50 individual decision trees, and fitted to the records. Then, a random forest empowered by the entity embedding vectors was fitted to the same dataset. The random forest with transfer learning outperformed the ordinary random forest. This could be attributed to the capability of the entity embedding to extract meaningful relationships between elements in each categorical variable.

In real-world applications of machine learning models, sparse data problems are common. Although the rise of the Internet has simplified the collection of large datasets in real time, small and medium-sized datasets are still frequently used. When sufficient data are lacking, data-oriented techniques, such as machine learning models, inevitably have limited capabilities. This study offers a way to alleviate this problem by adopting transfer learning, that is, by leveraging an already-learned knowledge set through the entity embedding approach. Transfer learning via entity embedding can be used in a wider area. For example, entity embeddings would be obtained in advance from sales records of residential properties, and they can be used for the valuation of less frequently traded properties in the market, such as commercial or special-purpose properties. This study is expected to promote the

rapid adoption of transfer learning approaches in fields where structured data are dominant.

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge: MIT Press, 2016, vol. 1, no. 2

[2] S. Tammina, "Transfer learning using VGG-16 with deep convolutional neural network for classifying images," *International Journal of Scientific and Research Publications*, vol. 9, no. 10, pp. 143-150, 2019.

[3] L. Wen, X. Li, X. Li, and L. Gao, "A new transfer learning based on VGG-19 network for fault diagnosis," in *Proc. IEEE 23rd International Conference on Computer Supported Cooperative Work in Design*, May 2019, pp. 205-209.

[4] C. A. Ferreira, T. Melo, P. Sousa, M. I. Meyer, E. Shakibapour, P. Costa, and A. Campilho, "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2," in *Proc. International Conference Image Analysis and Recognition*, June 2018, pp. 763-770.

[5] A. S. B. Reddy and D. S. Juliet, "Transfer learning with ResNet-50 for malaria cell-image classification," in *Proc. International Conference on Communication and Signal Processing*, April 2019, pp. 945-949.

[6] C. Lin, L. Li, W. Luo, K. C. Wang, and J. Guo, "Transfer learning based traffic sign recognition using inception-v3 model," *Periodica Polytechnica Transportation Engineering*, vol. 47, no. 3, pp. 242-250, 2019.

[7] C. Wang, D. Chen, L. Hao, X. Liu, Y. Zeng, J. Chen, and G. Zhang, "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146533-146541, 2019.

[8] O. Abdelwahab and A. Elmaghraby, "UofL at SemEval-2016 task 4: Multi domain word2vec for Twitter sentiment classification," in *Proc. the 10th International Workshop on Semantic Evaluation*, 2016, pp. 164-170.

[9] M. Aydoğan and A. Karci, "Turkish text classification with machine learning and transfer learning," in *Proc. International Artificial Intelligence and Data Processing Symposium*, 2019, pp. 1-6.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532-1543.

[11] Z. Chen, Y. Huang, Y. Liang, Y. Wang, X. Fu, and K. Fu, "RGloVe: An improved approach of global vectors for distributional entity relation representation," *Algorithms*, vol. 10, no. 2, p. 42, 2017.

[12] P. L. Goffe, "Hedonic pricing of agriculture and forestry externalities," *Environmental and Resource Economics*, vol. 15, no. 4, pp. 397-401, 2000.

[13] C. F. Chen and R. Rothschild, "An application of hedonic pricing analysis to the case of hotel rooms in Taipei," *Tourism Economics*, vol. 16, no. 3, pp. 685-694, 2010.

[14] T. E. Panduro and K. L. Veie, "Classification and valuation of urban green spaces—A hedonic house price valuation," *Landscape and Urban Planning*, vol. 120, pp. 119-128, 2013.

[15] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," arXiv preprint arXiv:1604.06737, 2016.

[16] V. Efthymiou, O. Hassanzadeh, M. Rodriguez-Muro, and V. Christophides, "Matching web tables with knowledge base entities: From entity lookups to entity embeddings," in *Proc. International Semantic Web Conference*, October 2017, pp. 260-277.

[17] T. C. Au, "Random forests, decision trees, and categorical predictors: the 'absent levels' problem," arXiv preprint arXiv:1706.03492, 2017.

[18] E. A. Antipov and E. B. Ponkryshevskaya, "Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics," *Expert Systems with Applications*, vol. 39, no. 2, pp. 1772-1778, 2012.

[19] J. Hong, H. Choi, and W. S. Kim, "A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea," *International Journal of Strategic Property Management*, vol. 24, no. 3, pp. 140-152, 2020.

[20] S. Yilmazer and S. Kocaman, "A mass appraisal assessment study using machine learning based on multiple regression and random forest," *Land Use Policy*, vol. 99, p. 104889, 2020.

[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

Changro Lee is a professor at the Department of Real Estate, Kangwon National University (KNU), South Korea. Before joining KNU, Lee was a researcher at Korea Institute of Local Finance. He has worked in the fields of real estate management, machine learning, and local financing of real estate.