Can we use SHAP values to increase the explainability of machine learning techniques in the valuation of the real estate in the Dutch market?

Seminar Case Studies in Applied Econometrics

Felix den Breejen - 455204

Laurence Geenen - 428224

Corné Vriends - 440435

Willemijn Brus - 484505

March 2020

Abstract

In this paper we use Shapley values to increase the interpretability of property value predictions produced by a Neural Network (NN) and Random Forest (RF) model. We use data on property valuation provided by Ortec Finance. It is of great importance for Ortec Finance that their predictions are explainable to individuals such as a judge or an assessor. We show that it is difficult for the currently used Hierarchical Trend Model (HTM) to explain the difference in predictions using the HTM coefficients. That is due to the complex modelling structure. Shapley values can be interpreted as the change in the predicted value of a home when the home characteristic of other homes are changed to match this home. We compute the Shapley values for the HTM model and show that they are qualitatively similar to the Shapley values of the machine learning models. Moreover, we show that it is important to only use Shapley values to compare houses in the same neighbourhood and of the same house type. We conclude that the Shapley values are a better tool than the HTM coefficients to explain differences in predicted house prices between a few similar houses listed at the same time. Hence, we show that the NN and RF combined with the Shapley values give well-explained predictions with high accuracy. Therefore, we advise Ortec Finance to use the Shapley values.

Contents

1. Introduction

In this paper we use the Shapley values to increase the explainability of house predictions made by a Neural Network (NN) and Random Forest (RF) model. Predicted prices need to be interpretable for individuals such as a judge or an assessor. The Hierarchical Trend Model (HTM) is a model that is currently used by Ortec Finance to estimate the value of houses. An advantage of the HTM is that the resulting output can be reliably traced to the inputs for explainability purposes. However, NN and RF often achieve higher levels of accuracy than the HTM. A drawback of the NN and RF is the lack of interpretability of the resulting predictions.

We compute the Shapley values corresponding to each house for each variable. These values are estimated by the SHapley Additive exPlanations (SHAP) method. Shapley values can be interpreted as the change in the predicted value of a house when the characteristics of other houses are changed to match this house. Moreover, we analyse the potential drawbacks of this method. We use the HTM model as a benchmark by computing and comparing the Shapley values of all three models. By comparing the HTM with the other models,we can verify whether the interpretation of the SHAP method and the HTM align. The purpose of our research is stated in the following research question and sub-questions:

To what extent can Shapley values increase the explainability of machine learning techniques utilized for valuation purposes in the Dutch real estate market?

1. To what extent can Shapley values provide interpretable explanations for predictions of house prices?
2. To what degree are the Shapley values comparable between the algorithmic models (NN & RF) and the traditional statistical model (HTM)?
3. To what degree are the Shapley values of the HTM model comparable with the traditional interpretation of this model?
4. Is it problematic that the estimation of Shapley values may utilize unrealistic combinations of variables?

The data used is provided by Ortec Finance and contains information on housing valuations in Amsterdam and the surrounding area. Our findings show that the Shapley values are able to increase the interpretability of predictions of a NN and a RF model. Our paper contributes to making predictions of NN and RF models more interpretable. We focus on the real estate market. However, Shapley values might also be a very valuable tool for many other machine learning applications. More research is required to generalise our findings to other fields.

In the next section, we discuss relevant literature on the interpretability of machine learning techniques, Shapley values and the SHAP method in particular. In the third section, we briefly discuss our data. In the fourth section, we describe the methodology used in this paper. In addition, we explain which potential drawbacks arise with computing these values and how they can influence

the interpretability. In our fifth section, we discuss the results and we complete this paper with some concluding remarks in the sixth section.

2. Literature

2.1. Valuation

The HTM (Francke, 2009) is a model used by Ortec Finance for the valuation of real estate. Different studies on housing valuation show different results for which type of model performs best in for this purpose. Nguyen and Cripps (2001) compare the predictive performance of the artificial neural networks (ANN) with that of the multiple regression analysis (MRA) using data on housing valuation. They point out that the predictive performance of the ANN increases as the size training set increases. They find that the performance of the MRA increases as the functional specification of the model improves. Therefore, the model which has the best predictive performance highly depends on these characteristics. Different conclusions with respect to the best model are thus to be expected when comparing studies. Also, they point out that the performance of the ANN model highly depends on hyper-parameter decisions such as the number of hidden layers and the number of neurons per layer.

However, due to the absence of a methodological approach to optimize the aforementioned hyper-paramaters, experimentation is required in practice for such choices. Hence, the performance of the ANN fluctuates highly and depends on sufficient experimentation. They conclude that the ANN model outperforms the MRA when a sufficiently large test set is used, ranging between 13% and 39% of the whole data set. The MRA model has a better predictive performance in the case of a smaller training set. However, while the predictive performance of the MRA model remains constant as the size of the training set increases, the performance of the ANN significantly increases as the sample size increases.

McCluskey et al. (2013) find that ANN outperforms MRA in terms of valuation and prediction accuracy. However, due to the lack of transparency and interpretability of the ANN model structure and outputs, they prefer the MRA models. Tabales et al. (2013) prefer ANN over classical regression models for predicting house values due to the high flexibility ANN have. In addition, they find that the predictive performance improves in terms of lower values for the error measures when a large enough sample size is used. Mimis et al. (2013) also demonstrate that neural networks outperform an extension of the more traditional linear regression model for the purposes of housing valuation in Athens. Antipov and Pokryshevskaya (2012) compare RF with NN. They show that RF can outperform NN for valuation purposes if there are few missing values and a large number of explanatory variables, which lead to over-fit neural networks.

Several studies thus find that NN and RF have an improved prediction accuracy for housing valuation than more traditional models. However, the interpretability of such models is quite poor (McCluskey et al., 2013).

2.2. Approaches to extract interpretation

In the class of algorithmic models, there is a continuum of interpretation, from models that have an inherent interpretation, such as linear regression and decision trees, to models that have no interpretation like NNs. Several techniques exist to make the latter type more interpretable. Techniques that can be used regardless of the choice of model are described as model-agnostic (Molnar, 2019). We will first discuss a group of model-agnostic techniques that are all visualization techniques. The *Partial Dependence Plots* (PDP) of Friedman (2001) show the (average) partial dependence, or (average) marginal effect, of a subset of variables on the predictions of the model. The *Individual Conditional Expectation* (ICE) plots of Goldstein et al. (2013) extend the idea of PDP. They address the issue of possible interactions and advance the interpretation from the global level to the local individual level. There are two major drawbacks of PDP and ICE. Firstly, both techniques numerically approximate the integral of the marginal distribution. The sampling inherent in the approximation might generate unrealistic samples. This might bias the visualisations. *Accumulated Local Effects* (ALE) plots of Apley and Zhu (2016) are an alternative to PDP and circumvent this drawback. They do so by using a novel approach that limits the approximation to the available data. The second drawback of these visualisation techniques is that they are limited to interpreting a small subset of the variables. This limitation is due to the difficulties that arise when trying to visualise in more than three dimensions. Therefore, these techniques are not able to provide a comprehensive interpretation. However, they can give an indication of the importance of certain variables (Friedman, 2001).

Two techniques that attempt to provide a more comprehensive interpretation are Local Interpretable Model-agnostic Explanations (LIME) of Ribeiro et al. (2016) and SHapley Additive exPlanations (SHAP) of Lundberg and Lee (2017). LIME is a local surrogate approach. It attempts to approximate the predictions of the non-interpretable model at a local level. It is thus able to explain individual predictions. LIME samples observations around an individual observation and weighs each sampled observation with its distance to the individual observation. These weights are then used to estimate an interpretable or local surrogate model on this sampled dataset. An example of such a local surrogate model is a linear regression. This technique suffers from a similar issue as the visual techniques. The sampled observations might be unrealistic and this may in turn bias the explanation. Moreover, the interpretation is not unique, the contribution of a variable is sensitive to the choice of parameters (e.g. different constant σ in the exponential kernel results in a different variable contribution). The second technique, SHAP, is an approximation of the Shapley value. It has a model-agnostic approximation (KernelSHAP) and model-specific approximations (TreeSHAP for RFs and DeepSHAP for NNs (Lundberg et al., 2018)). Using the Shapley value as a comprehensive interpretation is quite attractive due to its uniqueness property (Lundberg and Lee, 2017), that LIME lacks. Furthermore, it provides both a local and a global interpretation. As our paper focuses on the SHAP technique, we will explain it in more detail in the methodology section.

3. Data

In this research project we use data on housing valuations in Amsterdam and the surrounding area, provided by Ortec Finance. This data contains information on 58,216 homes with transaction prices and characteristics over the period from January 2010 to December 2018. The variable of interest in this project is the list price of the houses. Figure 1 shows the monthly median and the rolling monthly mean of the list price over the whole time period. It can be observed that there are sharp peaks in certain months in the monthly median. The sharpest peak occurs in November 2010, in which only four homes were sold. To avoid these kind of peaks, we take the rolling monthly mean of the list prices, with the window size equal to six months. This provides a more smoothed view. From 2016 onwards there is a clear upward trend, which could reflect the increasing scarcity in the Dutch real estate market.

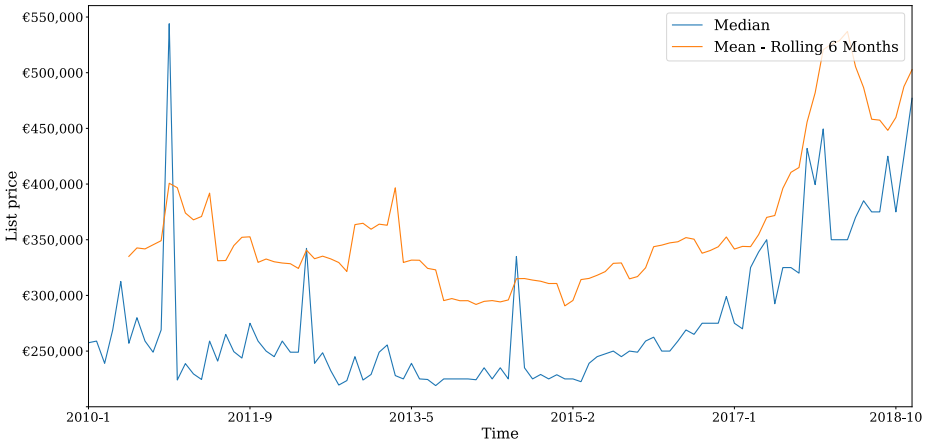

Figure 1: List Price Monthly Median and Rolling Monthly Mean

The data set contains six numerical variables and six categorical variables. The categorical variables are transformed into dummy variables. Our numerical variables are: year built, lot size, square meters, house volume, year listed and list price. All apartments in the data set are given a lot size of 0. Before we insert the data into the model we transform lot size by adding 1 to each value and taking the natural logarithm. This is to limit the impact of houses with a large lot size on the outcome. This is necessary because the neural network converges to a sub-optimal local optimum if the logarithm is not used. Taking the logarithm of lot size or removing the variable altogether reduces the value of the loss function from 13.5% to less than 11%. For the HTM the lot size is not initially transformed. The categorical variables can be divided into variables describing characteristics of the house and variables related to the location of the house. The variables describing the characteristics are housing type, presence of an indoor or outdoor garage, a variable indicating whether the house is a monument or not, and a variable indicating the presence of an indoor or outdoor storage unit. The first variable providing information on the location of

the house is one that indicates in which neighbourhood the house is located. The second variable describing the location contains the digits of the postal code. The values of the postal code all start with 10 or 11 as the first two digits due to the fact that all of the homes in the subset are located in Amsterdam or the surrounding area.

Descriptive statistics for the continuous variables list price, lot size, house volume and square footage can be found below.

	volume	square meters	lot size	list price
mean	272.37	91.77	86.52	340,835.20
stdev	159.05	49.89	1,475.84	287,189.70
min	46.00	15.00	0.00	64,500.00
25th percentile	170.00	60.00	0.00	194,500.00
75th percentile	329.00	110.00	0.00	375,000.00
max	3662.00	2857.00	111,111.00	7,250,000.00
skewness	3.08	8.35	62.50	5.38
kurtosis	21.43	340.14	4,233.18	52.61

Table 1: Descriptive Statistics

Of interest is that the 75th percentile of lot size has a value of 0, indicating that a majority of the homes in this subset are apartments. This may be reasonable, given that Amsterdam is a large, densely populated, city, where space comes at a premium. The list price has a 75th percentile of €375,000, whilst the maximum value is €7,250,000. Similar differences between the 75th percentile and the maximum can be observed in the other three variables, indicating that atypically large and valuable homes are present in the dataset.

Next, we examine the number of unique categories for each of the categorical variables in our data and their respective frequencies. There are 101 unique postal codes, 32 different types of homes, 9 municipalities, 126 quarters and 522 unique neighbourhoods (this hierarchical structure is present in the neighbourhood variable). There are three categories for the garage and storage unit variable: not present, present indoor, present outdoor. The monument variable has two categories: yes and no. The bar plots that contain the five (or less) most frequent values for each category for each variable related to the housing characteristics can be observed in the Appendix. It can be observed that four of the five most common housing types are a form of an apartment. Only a small number are monuments and most lack a garage. Roughly half of the observations have a storage room.

	list price	volume	square meters	lot size	year built
list price	1	0.78	0.75	0.27	-0.17
volume	0.78	1	0.92	0.52	0.04
square meters	0.75	0.92	1	0.54	0.05
lot size	0.27	0.52	0.54	1	0.15
year built	-0.17	0.04	0.05	0.15	1

Table 2: Correlations

We estimate Pearson correlations for the numerical variables in the dataset in order to increase

our understanding of the linear relationships between them. As can be observed in Table 2, both the volume of the home and the square meters of living space have a strong positive correlation with the list price. They are also almost perfectly correlated with each other, indicating that there may be multicollinearity. Lot size has a positive correlation with the list price, but it is not as strong as that of the other size variables. This may be due to the previously mentioned prevalence of apartments in the dataset. Lot size has a strong positive correlation with the volume and square meters. Year built is negatively correlated with list price. The remaining correlations between the variables are weak. For example, volume and year built have a correlation of 0.04.

Box plots are used to investigate the relationship between the categorical variables and the list price. These are located in the appendix. We observe that the presence of a garage and being a monument increase the value of a home. The presence of a storage room seems to be negatively associated with the value of a home. This is of note because additional storage room could be expected to increase rather than decrease value. We select the five most common housing types for the box plot for house type. The interquartile range varies widely per house type and terraced homes have the largest observed values in the dataset. This result is not unusual because homes are typically larger than apartments.

4. Methodology

In this section we discuss the methods used during our research. We make use of two types of machine learning models: Random Forests and Neural Networks. The HTM model is used as the benchmark. These three models are explained in section 4.1. The Shapley values are utilized to make these models more interpretable. Section 4.2 discusses the definition of the Shapley value and the SHAP approximation methods. In section 4.3 we discuss the potential drawbacks of SHAP, in particular the unrealistic houses that it can create. The last section discusses the methods applied to evaluate model performance.

4.1. Methodology of used models

4.1.1. Methodology of Neural Networks

Neural Networks are two stage regression or classification models that can be represented by a network diagram, as per Hastie et al. (2009). The network is composed of neurons, see Figure 2. These neurons are split into different layers: the input layer, the "hidden" layers and the output layer, see Figure 3. We illustrate the methodology of the NN using an example with only one hidden layer in order to keep the explanation as simple as possible. The input layer consists of neurons that receive the the raw data as input. The raw data are x_j for $j = 1, ..., p$, where p is the number of explanatory variables. The hidden layer has neurons that receive input from the input layer and combine it to produce a single output per neuron. The output layer produces the final result. For the purpose of this paper that is the predicted value of the home y. Take the hidden layer of our NN, which contains $h = 1, ..., n$ neurons. Each neuron receives the weighted explanatory variables

as input, where w_k are the weights corresponding to the variables. A potential bias b_h is added to each neuron.

$$a_h = \sum_{k=1}^{p} w_{h,k} \cdot x_k + b_h \tag{1}$$

The inputs and the bias combine to form the activation value of a neuron a_h, see (1), in the single hidden layer. Next, this value is fed to an activation function: $z_h = f(a_h)$. The output of this function, z_h, is the output of the concerning neuron. The outputs $z_1, ..., z_n$ are the inputs for the next layer. In this case the next layer is the output layer. In a two layer NN like in Figure 3, the outputs y_i for $i = 1, 2$ are again constructed by

$$y_i = \zeta(\sum_{h=1}^{n} w_{i,h} \cdot z_h + b_i). \tag{2}$$

In (2) ζ is the activation function of the output layer, w_h are the weights corresponding to the output of each neuron in the previous layer, z_h are the output of the previous layer and b_i is the bias corresponding to the final output y_i. We can connect the explanatory variables to the final output by

$$y_i = \zeta(\sum_{h=1}^{n} w_{i,h}^{(2)} \cdot f(\sum_{k=1}^{p} w_{h,k}^{(1)} \cdot x_k + b_h^{(1)}) + b_i^{(2)}). \tag{3}$$

In (3) $w_{h,k}^{(1)}$ represent the weights of the hidden layer and $w_{i,h}^{(2)}$ those of the second layer.

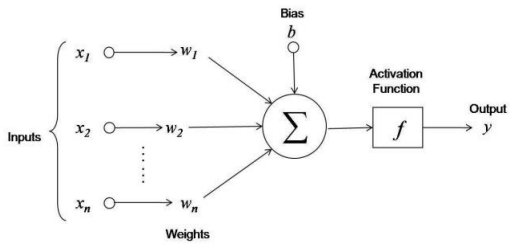

Figure 2: Operations done by a neuron [1]

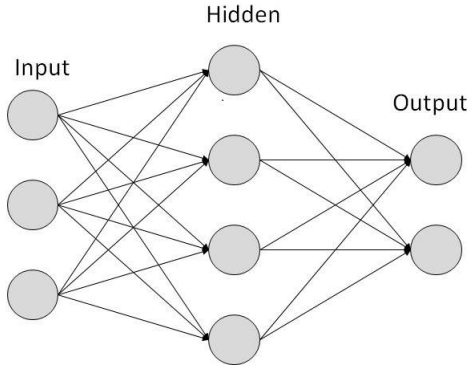

Figure 3: Neural Network structure [2]

The weights that connect the neurons with each other and the bias are the only parameters estimated by the NN. They determine the fit of the model to the data. This is often done by backpropagation. In the learning phase the NN checks whether the resulting output of the NN aligns with the true values of the data. The NN will "go back" and inspect every connection between neurons if this is not the case. More specifically, the NN computes how the resulting

[1] From " First neural network for beginners explained (with code)" by Arnx, A. 2013 (https://towardsdatascience.com/first-neural-network-for-beginners-explained-with-code-4cfd37e06eaf)

[2] From "Neural Network - Databricks" by Databricks (https://databricks.com/glossary/neural-network)

output of the neuron will change when the connecting weights are changed (in other words, the gradient). By doing this backwards the NN can see how the resulting final output reacts to changing the weights. Hence, in this manner the NN can find the weights that minimize the error function. In our project the error function is the Mean Absolute Percentage Error (MAPE). The methodology of the MAPE will be discussed in the subsection of the methodology of model performance.

Several parameters decisions need to be made when constructing a neural network. Firstly, the architecture of the network and the number of hidden layers needs to be determined. In unrelated disciplines such as Computer Vision (CV) and Natural Language Processing (NLP) it is quite common to model a priori knowledge about the nature of the data into the architecture of the NN, as per (Ke et al., 2019)). However, for the purpose of our research there are no apparent a priori structures that we are able to model appropriately into an NN architecture . Therefore, we make the architecture choice a part of our optimization procedure. This is often referred to as Neural Architecture Search (NAS) (Elsken et al., 2019). As this is quite computationally intensive, we restrict the architecture to a Fully Connected Network (FCN) with two hidden layers. Also, after running models with different numbers of hidden layers it was observed that using more than two hidden layers did not increase the predictive performance of the model. Therefore, we choose to use a NN with two hidden layers.

Besides the architecture of the network and the number of hidden layers, the values of several other hyper-parameters needs to be chosen. Firstly, the number of neurons in each hidden layer needs to be determined. This hyper-parameter will be optimized using cross validation. In the next section we explain the methodology of the cross validation in more detail.

Secondly, a suitable activation function needs to be chosen. For the output layer we use the identity function $\zeta_{Identity}(x) = x$. In absence of a predetermined decision rule, we again compare the model performance using several activation functions to select the optimal one for the hidden layers. Often chosen functions are the Sigmoid and ReLU function. We use the ReLU function as the Sigmoid function has a very high computational time and a higher prediction error than the ReLU function in our experiments. The ReLU function is defined as $f_{ReLU}(x) = max(x, 0)$, and can be characterized as a piecewise linear function. The predicted housing value is a piecewise linear function in the housing characteristics because the Neural Network is a combination of linear combinations with piecewise linear activation functions.

Lastly, we use the ADAM optimizer as the optimizer. Kingma and Ba (2014) have designed and tested ADAM on NN and find that it is an efficient optimizer for NN. As this optimizer has an adaptive learning rate, the learning rate of our NN will be optimized automatically by using ADAM as optimizer. For a detailed explanation of ADAM we refer to Kingma and Ba (2014).

The temporal aspect of the data is accounted for via a date variable that is defined as the number of days since a given reference date. We select January 1st, 2000 as the reference date. We add 551 dummies for the 552 neighbourhoods to describe the location aspect of the data. The categorical variables are assigned one dummy for each category as well. We do not take logarithms for the list price, square meters and volume because we would like to interpret the Shapley values

as a change in euros instead of as a percentage.

4.1.2. Methodology of Random Forest

Random Forests for regression purposes is a tree ensemble method proposed by Breiman (2001). As per Hastie et al. (2009), a regression tree is a recursive binary tree for regression purposes. At each node of the tree the best splitting variable and the best splitting point are determined. These are then used to split the data into two regions. The splitting process is then repeated for each sub-region. This process is repeated until the maximum depth is reached, whereafter the process stops in order to prevent overfitting. In the RF method a number of regression trees are combined to produce a prediction. This is done so by averaging the predictions of each of the trees to produce a final prediction. Bootstrap aggregation, introduced by Breiman (1996), is used to produce every new training set by sampling with replacement from the original data set. In the RF technique the variables used for splitting the tree at each node are a random subset of the available variables. Every new split uses a new subset of equal size of the variables. A risk exists that all trees will split using the same strong predictor if a random subset is not used. In that case the resulting trees will be highly correlated. Random Forest is thus used to produce uncorrelated decision trees. The hyper-parameters for RF are the size of the subset used to decorrelate the tree, the depth of the tree and the number of trees. These parameters will be optimized using cross validation.

The temporal aspect of the data is accounted for via a date variable that is defined as the number of days since a given reference date. We select January 1st, 2000 as the reference date. We add the 552 neighbourhoods as one variable to address the location aspect. The categorical variables are not split into dummies (thus, it remains a numerical variable), as tree based methods have difficulties with binary dummy variables that are extremely unevenly distributed. Some ordering is present in the neighbourhood variable as the first few digits refer to the municipality. We do not take logarithms for the list price, square meters and volume, because we would like to interpret the Shapley values as a change in euros instead of as a percentage.

4.1.3. Methodology of hyper-parameter optimization

For the hyper-parameter optimization we use random search in lieu of grid search or manual search. Bergstra and Bengio (2012) show that random trials perform better than trials on a grid (specifically for NNs). The predictive performance of RFs and NNs is quite sensitive to the choice of parameters (Hastie et al., 2009). The optimization procedure is based on an out-of-sample (OOS) approach, where we opt for a (chronological) three-way split of the data. The chronological aspect is essential, as our data has a temporal dimension. If it is ignored the OOS MAPE will be too optimistic (thus biased). One important aspect to keep in mind is that the lower and upper boundaries of the parameter space for the NN is based on an initial experiment to infer where the OOS MAPE did not improve in a meaningful manner. This is in addition to computational considerations.

For RF the values of several hyper-parameters need to be chosen. As noted by Probst et al. (2019) "the literature on RF cruelly lacks systematic large-scale comparison studies on the different

variants and values of hyper-parameters". Therefore, we opt for a random search over a sensible parameter space. Which has been based on heuristics and values observed in an empirical setting. Firstly, m, the number of variables to select from p, needs to be chosen. This determines the degree of shallowness of the tree, drawn uniformly between 1 and $max(\mathrm{p})$). Secondly, the *maxDepth* which controls the number of leaf nodes and hence depth of the tree needs to be chosen. This is considered to be the most consequential regularization parameter. It is drawn uniformly between 5 and 100), Thirdly, the *n_estimators* needs to be determined. It determines the number of trees in a forest. A higher number increases the chance of incorporating all the variables in the model. This in turn increases the complexity of the model. It is drawn uniformly between 10 and 200.

For NNs the following parameter is part of the optimization procedure: the *number of neurons* in each hidden layer (chosen uniformly between $ln(128)$ and $ln(1024)$). The ADAM optimizer retains its default values and is not part of the optimization procedure.

We restrict ourselves a priori to 100 iterations, as for both models there is a trade-off between computation time and parameter optimality. In the end the parameters that achieve the best results on the OOS MAPE for RF as well as NN will be the final model used in this paper.

4.1.4. Methodology of HTM

Vos and Francke (2000) propose the Hierarchical Trend Model (HTM) for real estate valuation. As mentioned previously, this model is used by Ortec Finance for this purpose in the Dutch real estate market. The HTM will serve as our benchmark to contrast with the selected machine learning models. The model is specified as per Francke (2009) as

$$y_t = \iota\mu_t + D_{\vartheta,t}\vartheta_t + D_{\lambda,t}\lambda_t + D_{\gamma,t}\gamma + h(X_t,\beta) + \varepsilon_t, \tag{4}$$

$$h(X_t,\beta) = ln(x_1^{\beta_1} e^{x_2'\beta_2} + x_3'\beta_3), \tag{5}$$

$$\begin{aligned}
\mu_{t+1} &= \mu_t + \kappa_t + \eta_t, & \eta_t &\sim N(0,\sigma_\mu^2 I), \quad \gamma \sim N(0,\sigma_\gamma^2 I),\\
\kappa_{t+1} &= \kappa_t + \zeta_t, & \zeta_t &\sim N(0,\sigma_\kappa^2), \quad \varepsilon_t \sim N(0,\sigma^2 I),\\
\vartheta_{t+1} &= \vartheta_t + \omega_t, & \omega_t &\sim N(0,\sigma_\vartheta^2),\\
\lambda_{t+1} &= \lambda_t + \varsigma_t, & \varsigma_t &\sim N(0,\sigma_\lambda^2).
\end{aligned}$$

The model is estimated on quarterly data. We have the logarithm of the house prices as the dependent variable y_t. The logarithm does not require us to interpret the Shapley values in percentages, in contrast to for the NN and RF-models. See section C.1 in the Appendix. The vector y_t consists of the houses sold in period t and can vary in length by period. The explanatory part of the model can be divided into time-variant and time-invariant components. The μ_t, the general trend component, is set as a local linear trend model where κ_t is the slope of the trend. The ι is a vector of ones. Both ϑ_t, the district time component, and λ_t, the house type time component, are modelled as random walks. The time-invariant components include the neighbourhood level γ and the function of housing characteristics $h(X_t,\beta)$. X_t represents all the explanatory variables

used in NN and RF, except for the neighbourhood variable. $X_t = \{x_1, ..., x_{p-1}\}$ for time period t, where x_p is the neighbourhood variable. The observations in each X_t differ because every period contains different listed houses. However, the corresponding parameters are time-independent. The neighbourhood level γ is set to be a fixed effect. The D matrices are selection matrices. They consist of 0's and 1's to select the proper neighbourhood or house type for each home.

The time-independent portion of the model is related to the characteristics of a house, see (5). The x_1 variable is the house size in square footage. The x_2 is a vector containing the characteristics directly related to the house, in our case the house volume and the year of construction. The x_3 vector consists of variables not related to the building: the garage, storage room, monument and lot size. The specific choice for $h(X_t, \beta)$, the individual characteristics of the house, is due to the requirement (a demand set by real estate appraisers according to Francke (2009)) that the value of the land and the building should be separable.

It is a time series model that can be written in state space form to allow for parameter estimation and optimization via the Kalman Filter. The state space form of the model consists of the measurement equation and transition equation. Our notation is similar to Francke and Vos (2004).

$$\begin{aligned} y_t &= Z_t\alpha_t + D_{\gamma,t}\gamma + h(X_t,\beta) + \varepsilon_t, \quad \varepsilon_t \sim N(0, R_t), \quad R_t = \sigma^2 I_t \\ \alpha_{t+1} &= T\alpha_t + \xi_t, \quad \xi_t \sim N(0, Q), \quad Q = diag(\begin{bmatrix}\sigma_\mu^2, & \sigma_\kappa^2, & \sigma_\vartheta^2 ... \sigma_\vartheta^2, & \sigma_\lambda^2 ... \sigma_\lambda^2\end{bmatrix}) \\ Z_t &= \begin{bmatrix}\iota & 0 & D_{\vartheta,t} & D_{\lambda,t}\end{bmatrix}, \quad \alpha_t = \begin{bmatrix}\mu_t & \kappa_t & \vartheta_t & \lambda_t\end{bmatrix}^T \end{aligned} \tag{6}$$

In (6), T is a time-independent matrix representing the transition of the state variables over time. In our model this T corresponds to an identity matrix with an extra 1 on position row 1, column 2, accounting for the fact that κ_t should be added to μ_{t+1} in the general trend component. Our choice for districts and neighbourhoods, which determines the resolution of our HTM model, is "*municipalities*" (9 respectively) and "*quarters*" (126 respectively). The above implementation fails to account for the variation in the number of observations over time. The (standard) Kalman Filter requires the number of observations in every time period to be the same. We apply the following procedure, as has been suggested to us by Ortec Finance, in order to achieve this.

Split $Z_t'Z_t$ into $P_tH_tP_t'$ via eigenvector decomposition for each time period. Where P_t is the eigenvector matrix and H_t is the diagonal matrix with positive eigenvalues ordered from large to small. Take r as the smallest number of positive eigenvalues over all periods. Define A_t as the first r columns of $Z_tP_tH_t^{1/2}$. Pre-multiplying (4) with A_t' transforms the dimension of y_t from a varying number to r. This dimension reduction technique preserves all information if r equals the column length of Z_t. If r is smaller, only the first r eigenvectors are used, sorted (in descending order) according to their eigenvalue.

The identification issues that arise when estimating the trend in levels and having a full classification for districts and house types are resolved by setting the $\mu_0 = 0$, $\vartheta_{0,1} = 0$ & $\lambda_{0,1} = 0$ (Francke and Vos, 2004). In other words, we set the first period of the general trend, the first period of the

first house type and the first period of the first district to zero.

$$A_t' y_t^R = A_t' y_t - A_t' h(X_t, \beta) - A_t' D_{\gamma,t} \gamma = A_t' Z_t \alpha_t + A_t' \varepsilon \tag{7}$$

Estimation is done with maximum likelihood, optimizing over the variables β, γ, and σ's. The first step is to create y_t^R and to remove the time-independent components. Then y_t^R, $h(X_t, \beta)$, $D_{\gamma,t}$, Z_t and ϵ_t are pre-multiplied by A_t' for the dimension reduction, see (7). These variables which have been reduced in dimension are used to retrieve the state space from the Kalman Filter and to calculate the likelihood.

It is important to have a sufficient number of observations in each time period, because the lowest number of observations in a time period strongly influences the choice of r. In the data set there is a month which has 3 observations, resulting in a r of 2. This causes too much information loss to our liking, so we decide to use quarters instead of months. The minimum number of observations of the quarters is 27, resulting in an r of 10.

4.2. Methodology of Shapley values

In this section the methodology of the Shapley values is discussed. The Shapley value is a way to quantify the contributions of the characteristics of a home its predicted value. We first focus on the prediction function in the case of unknown housing characteristics. We then introduce the Shapley value and give its interpretation. We discuss the approximation with SHAP using the Kernel SHAP in Appendix C.1, Deep SHAP in C.2 and Tree SHAP in C.3.

4.2.1. Value Function

Let f be the prediction function of a house of an underlying model, in our case a HTM, RF or NN. In general, this function can only return a prediction if the full set of variables F is used as input. In order to calculate the Shapley values, we need to define the behaviour of the prediction function in case some variables are omitted from the prediction. As notation we have: variable set $S \subseteq F$, $S^C := F \backslash S$, housing characteristics x for all variables, $x(S)$ for a subset of variables and v is the final predicted value. The random vector X is a house drawn from our theoretical housing probability distribution and the x is a known house. The expectation towards X is denoted by $\mathbb{E}_X$.

$$v_{remodel}(S) = f(x(S)) \tag{8}$$

$$v_{conditional}(S) = \mathbb{E}_X[f(X(S), X(S^C)) \mid X(S) = x(S)] \tag{9}$$

$$v_{marginal}(S) = \mathbb{E}_X[f(x(S), X(S^C))] \tag{10}$$

In (8), (9) and (10) three options for estimating the predicted value $v(S)$, when only the variables in S are known, are shown. The first equation (8) shows a prediction function $v_{remodel}$ that can accept less than all variables F. In practice, this means that the model is re-estimated for every possible subset of variables $S \subseteq F$. This method is only feasible for very fast models, such as a linear

regression. In our case, using the value function in (8) is computationally too expensive. The value function $v_{conditional}$ in (9) is seen most often in the literature. It has a very nice interpretation. For example, given that we know the house is a villa but we do not know any other housing characteristics, the predicted value is the average value of all villas. The big caveat here is that we are required to know the housing density conditional on $x(S)$ to estimate $v_{conditional}(S)$. For the value function $v_{marginal}$ in (10) it is relatively easy to estimate the expectation as

$$v_i(S) = \frac{1}{n}\sum_{j=1}^{n} f(x_i(S), x_j(S^C)). \tag{11}$$

For this equation $x_i(S)$ are the housing characteristics x of house i and variables S. The interpretation is as follows: given that we know the house is a villa but we do not know anything else, take all houses and change them to be villas. Predict their values and take the average. That is $v_i(S)$. Where S is a set with only the index for the type of house and house i a villa for the purpose of this example.

From an interpretation perspective, the value function $v_{conditional}$ is the most preferred. However, $v_{marginal}$ is usually used in practice. The estimation of $v_{conditional}$ is still a matter of ongoing research. For the Python SHAP package by Lundberg and Lee (2017), the authors propose to use $v_{conditional}$, but estimate their values with v_i from (11). They do so by assuming the variables are independent, effectively equating $v_{conditional}$ to $v_{marginal}$. Aas et al. (2019) offer new ways to approximate $v_{conditional}$. The debate on whether $v_{conditional}$ is better than $v_{marginal}$ is also not concluded, with Janzing et al. (2019) arguing that $v_{marginal}$ is conceptually better than $v_{conditional}$. On the other hand, in the field of housing prediction, using the value function in $v_{marginal}$ can be problematic. For example, apartments of size 1000m2 are rare to nonexistent, creating problems with unrealistic houses. We further investigate this issue in section 4.3. In our research, we will primary use $v_{marginal}$ because this one can be estimated for all models. The interpretation of $v_{conditional}$ will also be presented. For more exotic choices as value function in the context of Shapley values, see the overview of Sundararajan and Najmi (2019).

4.2.2. Shapley Value

The Shapley value measures the contribution of a variable to the predicted value of a house. Now that we have an approximation of the predicted value v_i, given that we do not know all characteristics, we will define the Shapley value. Let the function $\sigma : F \to \mathbb{N}$ be one possible ordering of the variables F and let Π be the set of all possible σ. The set Π contains $p!$ different orderings σ. The function σ has as input the index of a variable and as output the position in the ordering. The set $P_\sigma(k)$ contains all variables that precede variable $k \in F$ in the order σ. In the definition of this set, the r value is a position before the position of variable k. The inverse function σ^{-1} maps the position back to the variable. The set $P_\sigma(k)$ is defined as

$$P_\sigma(k) := \{\sigma^{-1}(r) \mid r \in \mathbb{N}, r < \sigma(k)\}. \tag{12}$$

The Shapley value ϕ_{ik} for house i and variable k is defined as

$$\phi_{ik} = \frac{1}{p!} \sum_{\sigma \in \Pi} v_i(P_\sigma(k) \cup \{k\}) - v_i(P_\sigma(k)), \tag{13}$$

where p is the total number of variables $|F|$. Given an arbitrary subset of variables $S \subseteq F \backslash \{k\}$, the Shapley value ϕ_{ik} is the average increase in predicted value v_i by adding variable k. In this mathematical form it is difficult to give a good interpretation of the Shapley value in the context of housing prediction. That is because the concept of missing characteristics is difficult to apply to homes. The Shapley value will be easier to interpret if we fill in the approximation formula (11) for v_i as

$$\phi_{ik} = \frac{1}{n} \frac{1}{p!} \sum_{\sigma \in \Pi} \sum_{j \in \mathbb{N}} f(x_i[P_\sigma(k) \cup \{k\}] \cup x_j[(P_\sigma(k) \cup \{k\})^C]) - f(x_i[P_\sigma(k)] \cup x_j[P_\sigma(k)^C]). \tag{14}$$

This formula is complex, so we write out the expression in words. Pick an ordering σ and draw a house j from the set of all house indices $\mathbb{N}$. Split the ordering of σ on the variable k. Compute the prediction using the housing characteristic of house i for the variables preceding k and including k and use the housing characteristics of house j for the omitted variables. Then subtract the prediction using the housing characteristic of house i for the variables preceding k but excluding k and use the housing characteristics of house j for the omitted variables including k. Average over all the houses j in the dataset and all possible orderings σ.

The fact that we sum over all the houses in the dataset is both a curse and a blessing. It is a curse because of the computational overhead. The number of possible orderings σ is $p!$ which means that the sum consists of $np!$ terms. One way to tackle this is to sample σ and j. However, the SHAP methods as explained in the following sections are preferred over this idea. The blessing comes from the fact that we gain an additional tool. By sampling over a subset of houses, we gain knowledge about the increase in prediction with respect to only a subset of houses instead of the whole dataset. In formula (14) this corresponds to changing $\mathbb{N}$ to a subset of house indices. We can even compare one house with another house to see the change in predicted value compared to only that house. With this new perspective, we create the summarized explanation for Shapley in explanation 4.1.

Explanation 4.1. You want to compare one house with a few other houses. The Shapley value is the change in predicted value when the other houses are reconstructed to match your house by changing one housing characteristic at the time, averaged over all houses and reconstruction orderings.

To add a small example: your house and the house of your neighbour are identical, except for the fact that your house has a larger garden and a garage. The predicted value of your house is €320,000 and that of your neighbour is €300,000. The house of your neighbour can be reconstructed to match your house in two orderings: first the garden then the garage or vice versa. When first adding the garden, the predicted value rises by €6,000 and an additional €14,000 by adding a

garage. The other way around: adding the garage first adds €12,000 and the garden adds €8,000. The Shapley value of your house with respect to your neighbour is €7,000 for the garden and €13,000 for the garage.

The explanation of the Shapley value in 4.1 is lengthy. For the sake of simplicity, we can assume that the reconstruction orderings do not matter. If the model would be linear, this is actually the case. The simplified, incomplete explanation is presented in explanation 4.2. When the technical details do not have priority, this is an easier description to provide.

Explanation 4.2. You want to compare one house with a few other houses. The Shapley value of a housing characteristic is the change in predicted value when the other houses are reconstructed by changing their housing characteristic to match yours, averaged over all houses.

To further establish our point of choosing the $v_{marginal}$ value function, we now discuss the $v_{conditional}$. The Shapley value based on the $v_{conditional}$ value function has a different interpretation. For this value function we do not have an expression for the sample conditional expectation, so we create the explanation from combining equations (9) and (13). The order σ can be seen as the order in which the variables are made known. Without knowing any characteristic of a house, the best prediction is the average house value. If one variable is presented, for example that the house is a villa, how does the prediction change? In this example, the housing type is the first variable in the presentation order. The other variables can be added next. The Shapley value is the average over all presentation orderings. The summary explanation is given in explanation 4.3.

Explanation 4.3. You want to know how the predicted value of one house came about. The Shapley value is the change in predicted value when the housing characteristics are made known to you one at a time, averaged over all presentation orderings.

Similarly to the $v_{marginal}$, we create an additional simplified explanation as depicted in explanation 4.4. Overall, the interpretation of the Shapley value with the $v_{conditional}$ is less useful than the Shapley value with the $v_{marginal}$ value function. First of all, with the $v_{conditional}$ value function, houses cannot be compared. The Shapley value should be seen as a value unique to each house. Secondly, the interpretation itself is not intuitive in practice. In the property valuation industry it is not common that we have a scenario in which we want to compare the prediction of a house to that of the same house with a variable unknown.

Explanation 4.4. You want to know how the predicted value of one house came about. The Shapley value is the difference in predicted value when a housing characteristic is known versus when it is unknown.

To conclude this section, we want to highlight that the explanations presented above depend crucially on the specific value function and approximated value function, as given in the previous section. The SHAP approximations methods in Appendix C all estimate the Shapley value with value function $v_{marginal}$. An exact estimation of the Shapley value with the $v_{conditional}$ exists only for the RF models.

4.3. Drawbacks of marginal Shapley

Shapley values with the $v_{marginal}$ value function are computed in part by adding the variables of a different house to the house in question. This computation thus allows for the creation of unrealistic combinations of variables, if variables are correlated. For example, we could have two houses that have the same characteristics except for the square footage and volume. The second house has triple the area and volume of the first house. As part of the Shapley formula, an unrealistic house will be created using the volume of the first house and the area of the second house. This unrealistic house has one third of the height of an average house. It is possible that the model values this unrealistic house less than the first house because of the unrealistic height. The estimation of the Shapley values thus depends on the creation of unrealistic houses.

There are two cases in which the prediction of the model might cause distortions in the Shapley value. The first case is a matter of the underlying sample. When a model is trained, it is fitted to the data that is presented. In our sample, we do not have unrealistic houses. Houses with a height of 0.80m, which can be created as intermediate house, do not exist in our dataset. Therefore, the prediction our model gives is questionable. Due to non-linearities, the model might predict negative values or values above one billion dollars. The second case is interpretation focused. If the Shapley is calculated based on, for example, the change in predicted value when a garage is added for an unrealistic house with height 0.80m, what does it mean? Not only do these houses not exist in the dataset, but they also do not exist in reality either. The interpretation can thus become very difficult if unrealistic houses are created.

Our primary concern is the influence of the unrealistic houses on the Shapley value. It could be the case that the influence of the unrealistic houses on the Shapley value will average out. On the other hand, there is the possibility that it will systematically bias the results. If the influences average out, we can ignore the problems it causes with the interpretation: when their influence is negligible, we do not have to interpret the difference in Shapley value caused by the unrealistic houses. In order to analyse this problem, we have constructed a method to distinguish between unrealistic houses and realistic houses. We also discuss how their influence can be measured.

To identify the possible creation of unrealistic homes, we focus on which characteristics are present for properties grouped by neighbourhood or house type. We determine the minimum and maximum values for lot size, square meters, volume and the height (volume/square meters) of homes in each neighbourhood and house type. If the characteristics of a semi-constructed house do not fit between the minimum and maximum values, we classify the house as unrealistic. We also check which house types are present in a neighbourhood, and classify any house with a house type that is not present as unrealistic. In this manner we create a function $h : \mathbb{R}^p \to \mathbb{B}$ that takes the value of 1 if the house is realistic and the value of 0 if the house is unrealistic. The $\mathbb{B}$ denotes the set of the binary numbers. All houses in the data set are classified as realistic by construction.

$$W = \{(i, j, \sigma) \mid i, j \in \mathbb{N}, i \neq j, \sigma \in \Pi\} \tag{15}$$

$$R = \{(i, j, S) \mid i, j \in \mathbb{N}, i \neq j, S \subseteq F, h(x_i(S) \cup x_j(S^C)) = 1\} \tag{16}$$

$$V_{k,all} = \{(i, j, \sigma) \in W \mid x_{ki} \neq x_{kj}\} \tag{17}$$

$$V_{k,real} = \{(i, j, \sigma) \in V_{k,all} \mid (i, j, P_\sigma(k) \cup \{k\}) \in R, (i, j, P_\sigma(k)) \in R\} \tag{18}$$

$$V_{k,unreal} = V_{k,all} \setminus V_{k,real} \tag{19}$$

Let W as shown in (15) be a set of tuples including all pairs of houses i and j and all possible orderings σ. When we compute the Shapley value, we reconstruct house j as to become house i by changing variables in order σ, see also section 4.2.2. To judge if a house is real, we need to know which variables of the semi-reconstructed house are of house i and which are of house j. Define $S \subseteq F$ as the variables that come from house i. The semi-reconstructed house has characteristics $x_i(S) \cup x_j(S^C)$. By plugging these characteristics into the function h, we can split the houses in realistic and unrealistic. The set R consists of all houses, both from the dataset and semi-reconstructed, that are realistic.

We are interested in the average change in predicted value when adding a variable. When the different houses have the same value for a housing characteristic, we exclude this pair for houses from the sample for this variable. That is because the difference in predicted value for a change of the same value is zero. The set of the remaining houses is denoted as $V_{k,all}$. During the reconstruction process, the house can change from realistic to unrealistic due to the change of variable k. We determine a transition to be unrealistic if either the house before or after the transition is unrealistic. The set of houses which have a realistic transition for variable k and ordering σ is given by $V_{k,real}$. Finally, the set of unrealistic transitions for variable k is given by $V_{k,unreal}$.

$$\tau_k(i, j, \sigma) = f(x_i(P_\sigma(k) \cup \{k\}) \cup x_j((P_\sigma(k) \cup \{k\})^C)) - f(x_i(P_\sigma(k)) \cup x_j(P_\sigma(k)^C)) \tag{20}$$

The change in predicted value for including variable k is denoted by τ_k, see (20). This is a part from the formula used in the Shapley approximation, see (14).

$$\phi_{ik} = \frac{1}{n} \frac{1}{p!} \sum_{\sigma \in \Pi} \sum_{j \in \mathbb{N}} \tau_k(i, j, \sigma) \tag{21}$$

$$\text{numerical:} \quad \rho_{k,\lambda} = \frac{1}{|V_{k,\lambda}|} \sum_{(i,j,\sigma) \in V_{k,\lambda}} \frac{\psi_k(i, j, \sigma)}{x_{ik} - x_{jk}} \tag{22}$$

$$\text{categorical:} \quad \rho_{k,\lambda} = \frac{1}{|V_{k,\lambda}|} \sum_{(i,j,\sigma) \in V_{k,\lambda}} |\psi_k(i, j, \sigma)| \tag{23}$$

The Shapley value can be calculated based on this τ_k, see (21). Using this τ_k, we create the variables $\rho_{k,\lambda}$ that can tell us the influence of the unrealistic houses. The λ is a placeholder for *all*, *real* or

unreal. If a variable is numerical then $\rho_{k,\lambda}$ is defined as in (22). If the variable is categorical then $\rho_{k,\lambda}$ is defined as in (23).

The $\rho_{k,\lambda}$ is interpreted as the average change in the prediction per unit x_{ik}, for the numerical variables k that differ between a pair of houses. The categorical variables have a similar interpretation. The $\rho_{k,\lambda}$ is then the absolute change in value for a pair of houses that differ in the variable k. To estimate $\rho_{k,\lambda}$ we randomly draw a subset W_{drawn} from the set W. If the bias resulting from the unrealistic houses averages out then we should observe no difference between $\rho_{k,real}$ and $\rho_{k,unreal}$. We investigate three different groups of houses for differences in $\rho_{k,\lambda}$: the whole dataset, houses from only one neighbourhood, and houses of only one house type. The set of housing indices $\mathbb{N}$ is set accordingly. For every group we sample 10,000 pairs of houses with a random order σ.

4.4. Model performance evaluation

The predictive performance of the three selected models will be measured via the mean absolute percentage error (MAPE). The MAPE is defined as

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{24}$$

The actual value is denoted by y_t, the predicted value by $\hat{y}_t$ and n is the number of houses in the set. For the HTM we take e to the power of the prediction in order to retrieve the true predicted price. We aim to achieve a MAPE of 12% or lower, as this is a benchmark that has been set by Ortec Finance.

We consider a result to be interpretable for the purpose of this paper if it fits two criteria. It must be possible to determine which variables are used by a model for a prediction and secondly it is a prerequisite that we can clearly separate and measure the influence of said variables. Shapley values could meet these criteria.

As discussed previously, the HTM shows which variables contributed to its prediction via the estimated coefficients of these variables. The magnitude of the coefficients describes the contribution of each respective variable. In order to see if the Shapley values are comparable with this traditional interpretation of the HTM model the Shapley values of the HTM model are computed as well. If the Shapley and HTM interpretation are comparable, they should provide the same insights regarding the magnitude and the sign of the variables.

Moreover, as the NN and RF lack this traditional interpretation of the HTM, we compare the two machine learning approaches to the HTM. We will investigate if the variables that are assigned large Shapley values for the HTM are also of influence in the Shapley values estimated for the NN and RF.

In addition, when explaining the appraised value of a home, the home is often compared to two or three slightly different houses in the area. The predictive model is then used to explain why the value of this house deviates from that of the comparable houses. Hence, we explore if we can explain the difference in the model predictions of three or four slightly different houses using the

corresponding Shapley values.

5. Results

In this section we present our results for the estimated models. The computed Shapley values of the three models are discussed. We investigate if the estimated Shapley values of the most accurate model can be viewed as "realistic" and if the potential problems adversely affect the estimated Shapley values. The Shapley results for the HTM are also compared to the traditional interpretation of the estimated HTM coefficients.

5.1. Predictive Performance

In this section we present the predictive performance of our three estimated models. See Table 3 for the MAPE results.

MAPE (%)	In sample	OOS 2018	OOS 2019
NN	8.96	10.38	10.40
RF	6.10	14.31	16.99
HTM	19.64	19.10	19.04

Table 3: Predictive performance of the estimated NN, HTM & RF models

As can be observed, the estimated NN model has the lowest MAPE: 10.38% for 2018 and 10.40% for 2019, indicating that it has the best OOS predictive performance. The RF performs second best with an out of sample fit of 14.31% and 16.99%. However, it has a large discrepancy between its in-sample and out-of-sample performance which is an indication of potential overfitting. The difference in OOS between 2018 and 2019 is also concerning. The benchmark HTM model performs the worst with a MAPE of 19.10% for 2018 and 19.04%, but it does not show signs of overfitting. We thus continue with NN as the primary model for the Shapley analysis.

5.2. Shapley analysis

In this section the estimated Shapley values of the three models are discussed and compared. The Shapley values shown in the following figures belong to the houses from the test sample calculated in comparison to the randomly selected subset of homes. This subset consists of 1000 homes for both the NN and RF, while 100 are selected for the HTM. These homes are randomly sampled and we assume that using a subset instead of the whole data set does not produce a quantitatively different result. The Shapley values are the deviation from the average prediction of these selected houses. We compute the Shapley values for the houses from the year 2018.

The Shapley values for NN are estimated using Deep SHAP. See Figure 4 for the SHAP summary plot for the NN. The SHAP summary plot shows each Shapley value for each variable for each home. The y-axis has the variables ordered from high to low with respect to average absolute Shapley value. This reflects the average absolute importance of each variable. The colours show the variable

value for each observation as explained in the right part of the figure. The x-axis measures the Shapley value for each observed value of each variable. For the type of home, neighbourhood code, garage dummies, monument dummy and storage dummy variables the feature value colouring scheme should be ignored as these variables have been given a numerical variable for estimation purposes but are dummy variables.

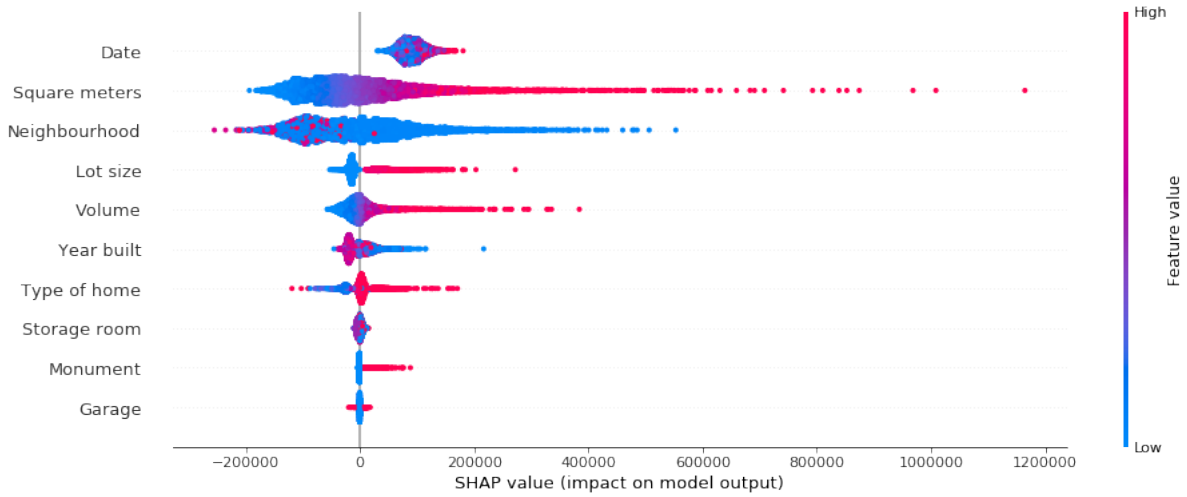

Figure 4: SHAP summary plot of NN

As observed in the figure, the **date** variable has the highest importance with respect to its Shapley value. It is always attributed a positive Shapley value. This can be due to the fact that the plot has been estimated using the houses from previous years. The average home prices are the higher in 2018 than in the years in the period 2010-2017. The **square meters** of living space has the second-largest impact with respect to the Shapley value. The Shapley values for this variable skew to the right. The colour of the Shapley values slowly shifts from blue to red, which means that a higher number of square meters leads to a larger valuation. The **neighbourhood** variable has the third-largest average Shapley values, reflecting that the location matters for the value of a home. The distribution of the Shapley values for **lot size** is skewed to the right, but most observations have a value close to 0. This observed pattern can be attributed to the prevalence of apartments which lack a lot. Nevertheless, it shows that the presence of a lot typically increases the value of homes for non-apartments as per the Shapley values. The **volume** variable has a similarly skewed distribution of the Shapley values as for the square meters variable, but more concentrated around 0. As the square meters are ranked higher in terms of average absolute Shapley value than the volume, we can ascertain that the NN considers the square meters to contain more information than the volume. **Year built** is skewed to the right, but most observations have a very low Shapley value. Older buildings are the observations with the largest positive Shapley values. This may seem counter-intuitive but can possibly be explained by the fact that very old homes may often be monuments. **House type** has an almost symmetric distribution. Certain home types result in a lower than average valuation while others are associated with larger prices. The **garage**,

monument and **storage** values have the lowest Shapley impact. For the garage and monument variables this can possibly be explained by the observed patterns in the data. It can be ascertained from the data that the vast majority of the observations are apartments that lack a garage and status as a monument. This means that generally the Shapley value will be low for these variables. However, it can be observed in the plot that the Shapley values for monument are skewed to the right when not equal to 0, indicating that the variable has a positive effect on home value if "present". The garage and storage values show that on average the Shapley value for those variables does not deviate greatly from zero. For the garage variable having a garage can even produce negative values.

Overall, the SHAP results clearly distinguish which variables have played large and small roles in the predictive valuation of homes and quantifies their influence. The resulting Shapley values do not produce results that are counter-intuitive or cannot be explained by the nature of the data used for estimation and prediction.

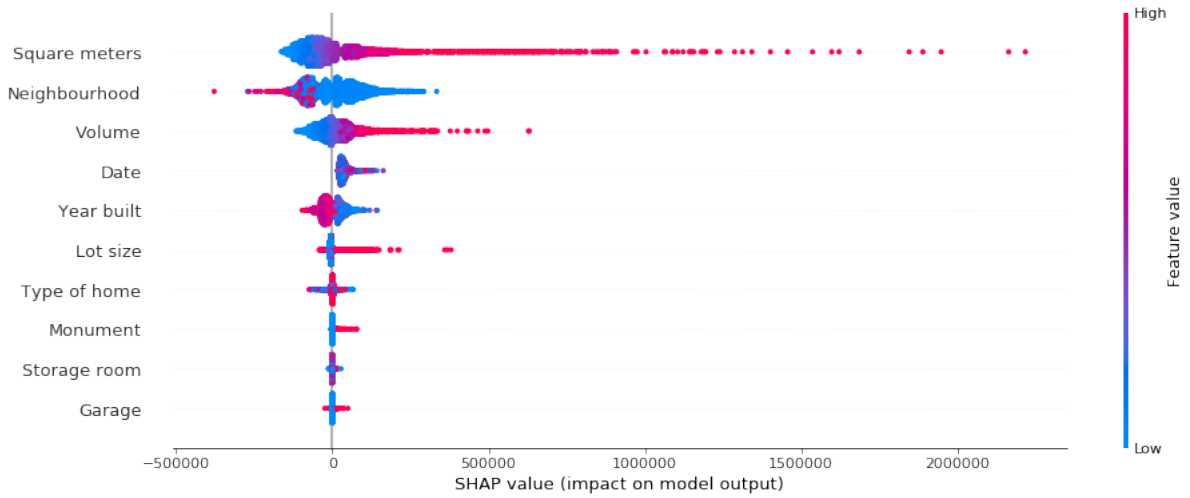

Figure 5: SHAP summary plot of Random Forest

We next examine the Shapley values estimated for the Random forest model, as seen in Figure 5. In contrast to the estimated Shapley values for the NN, **date** is not the most important variable with respect to Shapley values. This can be explained by the different methods both models use to construct a relationship between the explanatory variables and the listed price. As mentioned in the methodology, the NN constructs piecewise linear relations between the input variables and prediction output. The RF on the other hand uses a splitting mechanism in order to link the variable to the predicted value. For each split, the RF determines a threshold value for a variable. Observations with a value below the threshold go to one branch and the observations with a value above go to the other branch. This results in a step function for the date variable. However, a step function is less straightforward than a piecewise linear function as a model for date due to the present trend. Hence, the NN better captures the behaviour of the date than the RF does. This results in the lower importance on the date variable for RF.

The **square meters** of living space has the status of the most important variable in the Random Forest model. The attributed Shapley values also skew further to the right for the RF than for the NN. The second most important variable, **neighbourhood**, is less skewed than for the NN. The Shapley values for **volume** skew to the right, similarly to square meters. Date is attributed as the fourth most important variable for the RF, with a distribution resembling that of the NN. **Year built** skews slightly to the right, but has almost symmetrically distributed Shapley values. **Lot size** has Shapley values that skew right but is attributed a much lower importance in the RF than for in the NN. **House type** has an almost symmetrical distribution of its Shapley values, but it is not attributed a large degree of importance. The **garage**, **storage** and **monument** variables are the variables of least importance. The presence of a garage or storage room results in small Shapley values. The monument variable skews right as being a monument is valuable. However, there are few monuments in the sample so its average absolute Shapley importance is low given that the non-monuments have very low Shapley values for this variable. Overall, the distributions of the Shapley values of the variables for the RF resemble that of those of the NN. However, certain variables have importance rankings that differ, such as the date.

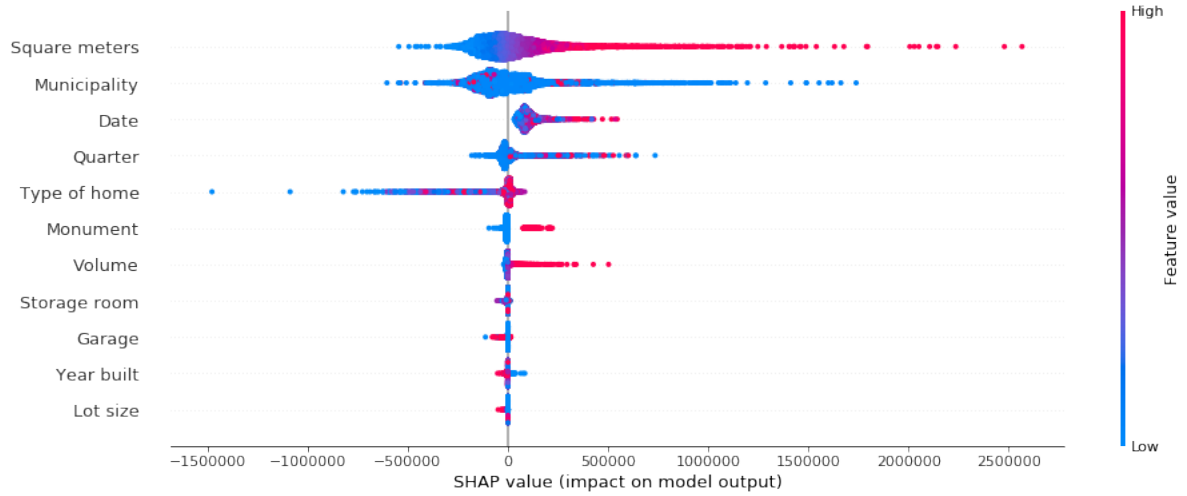

Figure 6: SHAP summary plot of HTM

The SHAP summary plot for the HTM is shown in Figure 6. There are two important changes regarding the variables between the HTM and the machine learning models. First, the geographical location variables are different. NN and RF have the neighbourhood variable, while the HTM has the municipality and quarter variables. Secondly, the temporal aspect is split into different trends for the HTM: a general trend, a house type trend and a municipality trend. The date variable in the SHAP summary plot for the HTM only refers to the general trend. This modelling choice leads to a higher importance for the house type and municipality and a lower importance for the date in the SHAP summary plot of the HTM.

As we can see in the plot, the **square meters** of living space is the variable deemed to have the highest importance for its Shapley values. This result aligns with that of the RF. Furthermore, its

Shapley value distribution is also skewed to the right. The **municipality** and **quarter** variables are also of great importance as per the average Shapley values. The **date** variable has a Shapley value distribution that is skewed to the right. Its Shapley values are always positive, in line with the results for the NN and RF. The **house type** variable Shapley values skew left, in contrast to the result for the NN and RF. The **monument** variable is assigned much greater importance than for the NN and RF. In some instances when a home is not a monument, the variable is attributed negative Shapley values. **Volume** has Shapley values that skew to the right, matching the pattern seen for the square meters. The **garage** and **storage** Shapley values skew to the left and are assigned a greater importance than seen previously. **Lot size** and **year built** generally are assigned the lowest Shapley values. This is of interest as these two variables have been assigned much larger values by the machine learning models.

The HTM results contrast with those of the NN and RF. The variables differ in importance between the NN and RF, but the Shapley value distributions are not different to a large degree. However, the HTM not only assigns a very different importance ranking than the two machine learning methods, but it also has noticeably different Shapley value distributions in certain instances.

5.3. Interpretation of the estimated HTM parameters

Table 5 shows the estimated β coefficients of the HTM. For the interpretation of the estimated coefficients it is important to keep in mind that the coefficients β_1 and β_2 are associated with the value of the building itself, while part of β_3 captures the effect on the value of the land that is not directly linked to the building itself. The estimated β_1, which captures the effect of the square meters on the list price, has a value of approximately 0.83. That means that on average a 1% increase in the number of square meters increases the predicted list price of a home by 0.83%. The β_2 estimates the effect of the direct characteristics of the home, which are the year built and volume. The estimated value for the volume is 0.00016, indicating that an increase in volume by one cubic meter leads to an 0.016% increase in the value of a home. An increase in the year built by one year decreases the list price on average by -0.017%. The estimated value for the β_3 reflects the effect of variables not related to the building itself. The interpretation of its value is however linked to the value of the building. For each home, the components for building B and land L are determined as

$$B = x_1^{\beta_1} e^{x_2' \beta_2},$$
$$L = x_3' \beta_3.$$

The two variables B and L only have an interpretation relative to each other. We can say that $\frac{B}{B+L}$ per cent of the value of the house is determined by the building itself, and the other $\frac{L}{B+L}$ is because of the land. The value B cannot be interpreted as the value of the building itself, because the trends are not included in B. The effect of a one unit increase in x_3 on the list price is β_3/B.

This thus means that for example an additional 1 m2 in lot size increases the total value of a property less when the building itself is more valuable. It increases it by $(7/B)\%$.

The fact that B only has a relative interpretation complicates a general interpretation of the exact value of the estimated coefficients. Nevertheless, it can be observed that an increase in lot size, having an outdoor storage or outdoor garage and having monument status have positive effects on the value of a home. Having an indoor garage or indoor storage can decrease the value of a home. For the garage, storage and monument dummy variable coefficients they are evaluated with respect to the dummy variables of not having a storage, garage or monument status. We can also do comparisons between the dummies. For example, being a monument results in a percentage increase in value that is four times larger than that of adding an outdoor storage.

The betas thus provide the time-invariant portion of the predicted value of a home. The remaining portion of the prediction is derived from the house type, district and general trend. To help quantify these trends we examine the estimated σ of these components of the HTM, see Table 6.

	β_1
square meters	0.83085

	β_2
volume	0.00016
year built	-0.00017

Table 4: estimated β HTM

	β_3
lot size	0.07
garage indoor 1	-1.89
garage outdoor 2	0.77
storage indoor 1	-1.02
storage outdoor 2	0.20
monument	8.24

Table 5: estimated β HTM

σ_μ	0.000236
σ_κ	0.000062
σ_λ	0.000044
σ_ϑ	0.000700
σ_ε	0.007263

Table 6: estimated σ HTM

In Table 6 the estimated σ variables of the HTM can be observed. The σ_ε is the standard deviation of the residuals of the measurement equation. It has a value of 0.0073, which is a measure of the expected error of the residuals. The σ_μ is the standard deviation of the general trend and has a value of 0.00024. The σ_λ is the standard deviation of the house type time component. Its value of 0.00004 shows that the price trends for the house types show much less variation than the general trend μ. The σ_ϑ is the standard deviation of the district time component. Its estimated value of 0.00070 indicates that the variation in price trends across districts is approximately three times as large as that of the variation in the general trend.

The trends of the HTM can also be interpreted. The estimated trend for apartments over the period 2010 - 2018 is shown in Figure 7. The y-axis shows the percentage change in the value of apartments in comparison to the first quarter of 2010. We can see that the apartments increase in value over time in comparison to other house types, accounting for other house characteristics and trends. The Shapley values cannot be used to acquire an equivalent interpretation. The trends are rarely useful for explaining the differences in valuation between houses because we are often interested in comparing houses on the same date.

The SHAP summary plot previously estimated for the HTM showed that square meters has on average the largest Shapley value of all variables. From the estimated HTM coefficients it is not

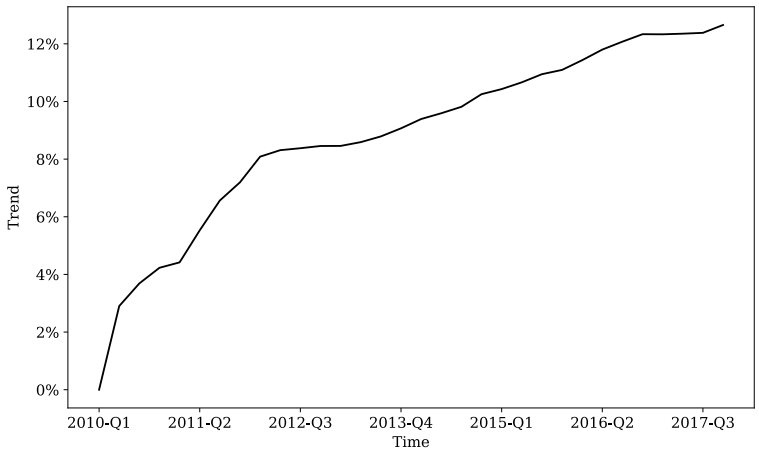

Figure 7: Estimated trend for appartments by the HTM (Q1 2010 - Q4 2018)

possible to conclude that β_1 is the most important because of the different interpretations of the β's. The three next most important variables as per the average Shapley value are the district, neighbourhood and house type. The district and house type effects are modeled as trends in the HTM. To compare the value of two homes that have the same characteristics except for district, the difference in the estimated γ for their respective districts quantifies the difference in predicted value. The neighbourhood is modeled as a fixed effect in lieu of a trend, but its interpretation in the case of comparing two homes would be similar: calculate the difference in the estimated γ coefficient. This is more easily done than via the Shapley values, where for each comparison we estimate a new set of Shapley values.

The advantage of the traditional interpretation of the HTM results is that the coefficients are applicable across all homes. Given the values for each variable and the respective estimated coefficients, we can in general terms explain the value of the home as the sum of the values multiplied by the coefficients. For the Shapley values on the other hand, the values estimated for each home are home specific. For example, the Shapley value per square meter will differ across homes. Because the HTM model use logarithms, the interpretations of the β's as percentages only hold for small deviations of x. This means that the coefficients cannot be used to do a comparison between houses, as differences are often bigger than a few percents. On top of that, the β_3 has a different building value B for each home, so a comparison between homes is not possible. In conclusion, the interpretation of the HTM can tell us how the value of a house changes in terms of small percentage changes in x, but does not provide additional information on why two houses have different predicted values.

5.4. Evaluation of Unrealistic Houses

We next investigate whether the Shapley values estimated previously contain many "unrealistic" cases where the chosen subset of variables produces an improbable home during the computation of the values. As previously described in the methodology, we randomly draw pairs of homes and transform one home into the other using an ordering of variables. We calculate the change in predicted values and check if they are different for the realistic houses in comparison to the unrealistic houses. The Neural Network is used for this analysis.

Perc(%)	sq. m.	volume	lot size	year b.	date	garage	storage	m.ment	h. type	n.hood
All	38.6	39.4	49.5	67.7	67.4	67.1	67.4	67.2	47.1	49.3
N.hood	79.7	79.6	72.4	87.8	88.1	87.6	87.9	88.1	72.0	87.6
H. type	59.8	61.7	81.0	84.4	84.3	83.8	84.4	83.7	83.4	71.6

Table 7: Percentage realistic

Table 7 shows the percentage of constructed homes that are classified as real per variable for houses drawn from the three different datasets. These data sets are the whole data set, a subset of one neighbourhood and a subset of one house type. We choose the neighbourhood with code 3620401 because it is the most common neighbourhood from the most occurring four-digit postal code in our data. Terraced homes are selected as the house type for the house type subset. We illustrate the interpretation of the results using the year built from the whole dataset as an example. When the year built of a house is replaced by another value of year built, the house before and after the replacement are classified as realistic houses in 67.7% of the cases. In the case that the change in year built is classified as unrealistic, this is not necessarily due to the replaced variable itself. It might be that one of the other explanatory variables already caused the semi-constructed house to be unrealistic.

As can be observed, the percentage of realistic houses classified per variable range from about 39% to 68% for the whole dataset. Variables with a very high percentage of real houses are the three dummy variables and the date variable. For the three dummy variables this high percentage of realistic houses can be explained by them being independent of most of the other variables. Because of this independence, there is a small probability of an unrealistic combination arising from these variables. The same independence holds for the date variable. Few variables have a distribution that is dependent on the date of the house. The probability of getting an unrealistic combination of a certain date and another variable is small.

On the other hand, only 38.6% and 39.4% of homes are classified realistic given the bounds for square meters and volume respectively. This can be explained by their interdependence. Changing one by a large amount without changing the other often makes the house unrealistic. For house type and neighbourhood less than 50% are labelled as realistic. This can be explained by the variation in characteristics per neighbourhood and house type. For some house types, variables as square meters and volume might depend more on each other than for other house types. Hence, when replacing those variables by other random values, the probability of creating an unrealistic

combination between house type and for example square meters is quite high. However, for other house types, these characteristic might vary more over the sample. Hence, the probability of getting an unrealistic house for these types will be quite low. The same holds for the neighbourhoods.

We research whether the Shapley values estimated for the whole data set can be trusted, given the many different neighbourhoods and house types present in the data. Table 7 shows that the percentages of realistic houses increase for the two selected subsets in comparison to the whole dataset. It is possible that value estimation should be restricted to more specific subsets.

	sq.m.	volume	lot size	year b.	date	garage	storage	monument	h. type	n.hood
Real	2281	202	8315	-1	48	16453	8835	13097	18513	79770
Unreal	2240	202	10750	-23	48	18078	10060	18112	28458	105024
All	2255	202	10511	-8	48	17095	9233	14865	24836	92615

Table 8: Estimated $\rho_{k,\lambda}$ for whole dataset

Table 8 shows the resulting values for each variable and house classification. We illustrate the interpretation of the values using the value for square meters for the houses classified as realistic. On average, for houses that have different values for the variable and are classified as realistic both before and after the transformation of square meters, an additional square meter results in a deviation of €2,281 from the reference list price.

As Table 8 shows, the difference in average value is not large for square meters, volume, and date sold. The values for garage and storage differ slightly. Nevertheless, for monument, house type, year built and neighbourhood we do observe large differences. The average values of all the homes are biased upwards for these two variables. However, the year built variable has a similar percentage real to the date sold variable, but a much larger difference in average value. This can potentially be attributed to the fact that the year built has an effect that may vary dependent on the exact date and type of home. For monuments an increase in year built may be associated with a negative value, while it would be attributed a positive value for a terraced home built in the last 10 years. Hence, in the subsets that we select for further analysis, the average effect of year built may switch between being positive and negative.

Given the previous results for all the data, we select a subsample for the neighbourhood level. Table 7 shows the majority of the generated homes are classified as realistic.

	sq meters	volume	lot size	year built	date	garage	storage	monument	house type
Real	1831	184	11831	818	56	16935	9313	-	9523
Unreal	2091	196	19038	444	56	18790	10931	-	25106
All	1884	186	18508	763	56	17297	9500	-	15011

Table 9: Estimated $\rho_{k,\lambda}$ for neighbourhood subset

We can see in Table 9 that the square meters, volume, date sold, garage and storage variable do not have large differences in the average value between real and not-real homes. House type, year built and lot size are variables for which the values do diverge. There are no monuments in this neighbourhood. The house type and lot size results may be associated with each other, as

for example a terraced home is more likely to have a non-zero lot size than an apartment. As previously mentioned, the effect of year built may also depend on the nature of a home. This gives cause to investigate what happens if we isolate one specific house type.

	sq meters	volume	lot size	year built	date	garage	storage	monument	n.hood
Real	1793	151	6761	-136	47	18982	10936	24852	84229
Unreal	1976	177	9331	-214	50	20273	12367	34798	144947
All	1868	161	7252	-148	47	19188	11217	27774	101549

Table 10: Estimated $\rho_{k,\lambda}$ for house type subset

A subsample was also analyzed for a specific house type: terraced homes. Table 10 shows a very large difference in the average value between realistic and unrealistic homes for the neighbourhood variable. The lot size, year built and monument variables also have noticeably different average values. The remaining variables do not differ to a large degree.

Isolating either house type or neighbourhood thus fails to prevent the realization of large differences in values for some variables between realistic and unrealistic homes. However, the variables square meters, volume, year built, date, garage and storage do not have a big difference in average change of prediction value between unrealistic and realistic house. The variables that could be an issue, lot size and monument, are very related to the house type and neighbourhood. It is perhaps wise to isolate both house type and neighbourhood simultaneously when determining values to help explain house valuation predictions. Hence, we advise limiting the use of Shapley values for inter-home comparison to homes located in the same general location and of the same type.

5.5. Comparison with 3 homes

We also give an example of using Shapley values to explain the differences in valuation for a small group of homes. We select four homes that are gallery apartments from the neighbourhood with code 3620401. We set the date listed to that of the first home such that the date does not affect the prediction. We then predict the values of these homes using the NN and compute the Shapley values of the first home in comparison with the other three. This example serves to match the procedure that an appraiser uses to value a home, where three similar homes are used for the valuation.

	sq meters	volume	year built	garage	storage	monument	predicted price
Home 1	95	250	1998	Indoor	Outdoor	No	334,606
Home 2	77	241	1961	None	None	No	232,509
Home 3	100	270	1962	None	Outdoor	No	290,864
Home 4	65	175	1960	None	Indoor	No	193,683

Table 11: Variable values and predicted list price

As can be observed in Table 11, the selected homes are of the same house type and from the same neighbourhood but differ with respect to characteristics such as the square meters of living space. The differences in characteristics can be linked to the estimated Shapley values in Table 12.

Home 1 is estimated to have a larger price than the other three homes, and the Shapley values help explain how its predicted price deviates from the average predicted price of the other three homes. The year built, the square meters, the volume and the storage have helped to increase its predicted price. However, the presence of an indoor garage has been attributed a negative effect. The garage variable can be seen to have a negative effect in some instances, as seen previously in the SHAP summary plot for the NN. Also, the coefficient in the HTM model is negative.

square meters	volume	year built	garage	storage	monument
38404	1688	54027	-5522	6991	0

Table 12: Shapley values

Table 12 shows the computed Shapley values for four houses. We illustrate the interpretation of our results using the methodology explained in section 4.2.2 for Home 1 and characteristic square meters. The Shapley value is the change in predicted value when Home 2, 3 and 4 are reconstructed to match Home 1 by changing one housing characteristic at the time, averaged over all three houses and reconstruction orderings. The average predicted value of Homes 2, 3 and 4 is €239,019. The value of €38,404 for square meters means that the 95 square meters of living space have contributed €38,404 to the deviation of the predicted price from the average predicted price. Likewise, the year built of 1998 has contributed €54,027 to the predicted price. The outdoor storage unit has added €6,991 in value, while the volume of 250 cubic meters has increased the predicted price by €1,688. Lastly, the indoor garage has decreased the predicted price by €5,522. Adding up all changes, the Shapley value has explained how the predicted price of Home 1 €334,606 differs from the average of the other three houses €239,019.

6. Conclusion

In this paper we use the Shapley values to increase the interpretability of predictions of house values done by a NN and RF model. The Shapley values are computed and visualized using methods called SHAP. Achieving a high prediction accuracy while keeping the predicted outcome interpretable is of high importance in the real estate market. Not only for the municipalities raising taxes on properties, but also for the homeowners themselves. As NN and RF generally give accurate predictions, they could be of use for predicting the list prices. However, the interpretability and therefore explainability of the resulting predictions is difficult. To address this we can estimate the Shapley value for each house for each explanatory variable. We use the HTM as a benchmark for the interpretability of our resulting Shapley values. The advantage of the HTM is that unlike the NN and RF it estimates coefficients representing a relationship between the explanatory variables and dependent variable. Therefore, the predicted output of the HTM is more explainable than the predicted output of a NN or RF model in the absence of Shapley values.

We use the $v_{marginal}$ value function to estimate the Shapley values. We define the Shapley value as the change in predicted value when all other houses are reconstructed to match the house

by changing one housing characteristic at the time, averaged over all houses and reconstruction orderings. Using the $v_{marginal}$ function creates the problem that unrealistic houses might arise if variables are dependent on each other. We construct a method to detect the cases in which unrealistic houses are problematic for the computed Shapley values. Using this method we can distinguish between Shapley values computed for realistic and unrealistic houses. We find that unrealistic houses are indeed created when computing the Shapley values in instances where the house type or neighbourhood differs for the houses that are being used for the comparison. Therefore, our advice is to use Shapley values to explain differences in listed prices of houses of the same type, in the same neighbourhood. It is often of most interest for taxation purposes to compare one house with three similar houses. Hence, the usage of Shapley values to make machine learning models more interpretable can be very helpful in this setting.

When comparing the SHAP plots of the NN and RF, we see that the computed Shapley values are very similar. The distributions of the Shapley variables for each variable are similar and in general terms the importance rankings do not greatly differ. When comparing the computed Shapley values of the HTM with those of the NN and the RF the ranking as well as the distributions of variables differ. However, most of these changes can be explained by the functional form through which the HTM depends on the variables. Hence, even though the ranking and distribution of the Shapley values of the HTM do not exactly coincide with those of the machine learning models, the contributions of each variable are qualitatively similar.

We compare the Shapley values of the HTM with the HTM coefficients to evaluate the interpretability of both measures. Both measures indicate how much a variable contributes to the predicted list price. However, there are some important distinctions between the two interpretations. Firstly, unlike the HTM coefficients, The Shapley values differ for each house. Hence, dissimilarly to the HTM coefficients, the Shapley values cannot be used to give general explanations to variables that hold for all houses. Secondly, because the HTM uses logarithms, the interpretations of the β's can only be interpreted when the change in the explanatory variables is less than a few percents, while the Shapley values can be used to interpret changes of any size in the explanatory variables. Thirdly, for a certain house, each variable is represented by a Shapley value that can be interpreted in the same way. However, not all the estimated parameters of the HTM can be interpreted in the same way. Altogether, we argue that in the context of explaining why the value of a certain house differs from the value of another house, the Shapley values give a more comprehensive explanation than the HTM coefficients.

In general, we conclude that the Shapley values can be a good tool to increase the interpretability of the predicted house prices by NN and RF models. We highlight the finding that Shapley values are most reliable in the case of comparing houses of similar house type and neighbourhood. Therefore, we recommend the usage of the Shapley values to increase the interpretability and explainability of predicted list prices to Ortec Finance.

References

Aas, K., Jullum, M., Løland, A., 2019. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. arXiv preprint arXiv:1903.10464 .

Antipov, E. A., Pokryshevskaya, E. B., 2012. Mass appraisal of residential apartments: An application of random forest for valuation and a cart-based approach for model diagnostics. Expert Systems with Applications *39*, 1772 – 1778.

Apley, D. W., Zhu, J., 2016. Visualizing the effects of predictor variables in black box supervised learning models.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization *13*, 281–305.

Breiman, 2001. Random forests. Machine Learning *45*, 5–32.

Breiman, L., 1996. Bagging predictors. Machine Learning *24*, 123–140.

Elsken, T., Metzen, J. H., Hutter, F., 2019. Neural architecture search: A survey. Journal of Machine Learning Research *20*, 1–21.

Francke, M. K., 2009. The Hierarchical Trend Model, John Wiley Sons, Ltd, chap. 8, pp. 164–180.

Francke, M. K., Vos, G. A., 2004. The hierarchical trend model for property valuation and local price indices. The Journal of Real Estate Finance and Economics *28*, 179–208.

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics *29*, 1189–1232.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2013. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation.

Hastie, T., Tibshirani, R., Friedman, J. H., 2009. John Wiley Sons, Ltd.

Janzing, D., Minorics, L., Blöbaum, P., 2019. Feature relevance quantification in explainable ai: A causality problem. arXiv preprint arXiv:1910.13413 .

Ke, G., Zhang, J., Xu, Z., Bian, J., Liu, T.-Y., 2019. TabNN: A universal neural network solution for tabular data.

Kingma, D. P., Ba, J., 2014. Adam: A method for stochastic optimization.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2019. Explainable ai for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610 .

Lundberg, S. M., Erion, G. G., Lee, S., 2018. Consistent individualized feature attribution for tree ensembles.

Lundberg, S. M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pp. 4765–4774.

McCluskey, W., McCord, M., Davis, P., Haran, M., McIlhatton, D., 2013. Prediction accuracy in mass appraisal: a comparison of modern approaches. Journal of Property Research *30*, 239–265.

Mimis, A., Rovolis, A., Stamou, M., 2013. Property valuation with artificial neural network: the case of athens. Journal of Property Research *30*, 128–143.

Molnar, C., 2019. Interpretable Machine Learning.

Nguyen, N., Cripps, A., 2001. Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. Journal of Real Estate Research *22*, 313–336.

Probst, P., Wright, M. N., Boulesteix, A.-L., 2019. Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery 9, e1301.

Ribeiro, M. T., Singh, S., Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier.

Shrikumar, A., Greenside, P., Kundaje, A., 2017. Learning important features through propagating activation differences.

Sundararajan, M., Najmi, A., 2019. The many shapley values for model explanation. arXiv preprint arXiv:1908.08474 .

Tabales, J. M. N., Caridad, J., Carmona, F. J. R., 2013. Artificial neural networks for predicting real estate prices. Revista De Metodos Cuantitativos Para La Economia Y La Empresa *15*, 29–44.

Vos, A., Francke, M., 2000. Efficient computation of hierachical trends. Journal of Business and Economic Statistics 18, 51–57.

Appendices

Appendix A Figures data section

A.1 Barplots

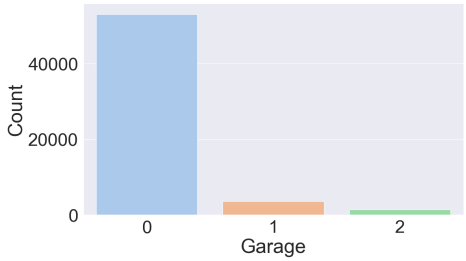

Figure 8: Frequency of garage (0 = none , 1 = indoors, 2 = outdoors)

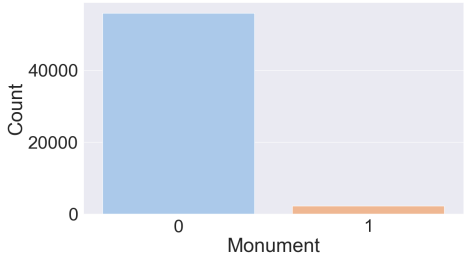

Figure 9: Frequency of monuments. (0 = no, 1 = yes)

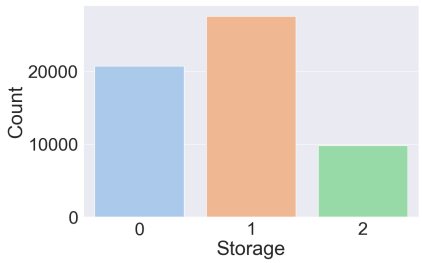

Figure 10: Frequency of storage (0 = none, 1 = indoors, 2 = outdoors

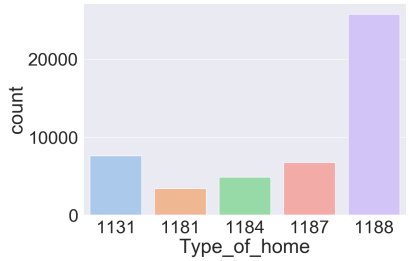

Figure 11: Frequencies house types (1188 = upstairs apartment, 1131 = terraced house, 1187 = ground floor apartment, 1184 = portico apartment and 1181 =gallery apartment)

A.2 Boxplots

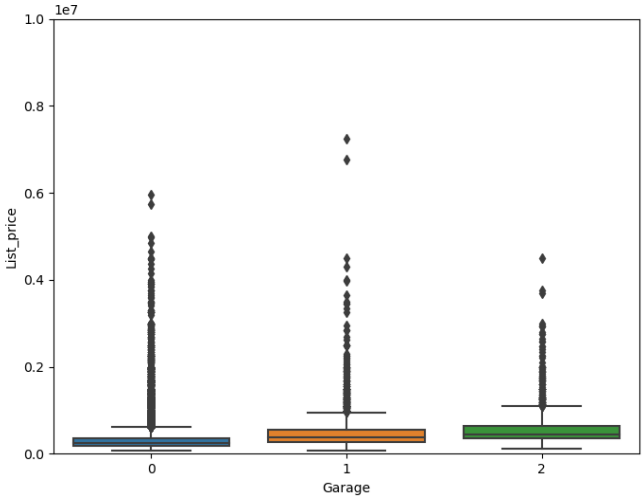

Figure 12: Boxplot of garage variable vs. list price

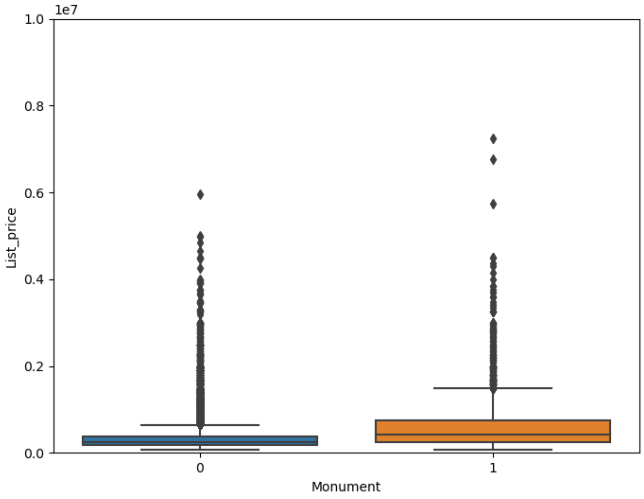

Figure 13: Boxplot of monument variable vs. list price

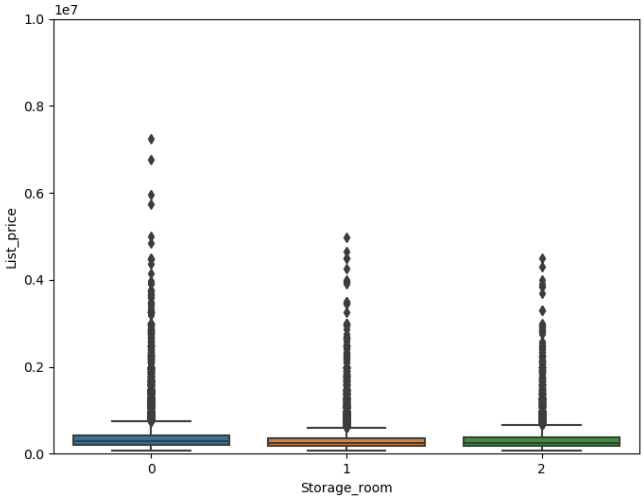

Figure 14: Boxplot of storage room variable vs. list price

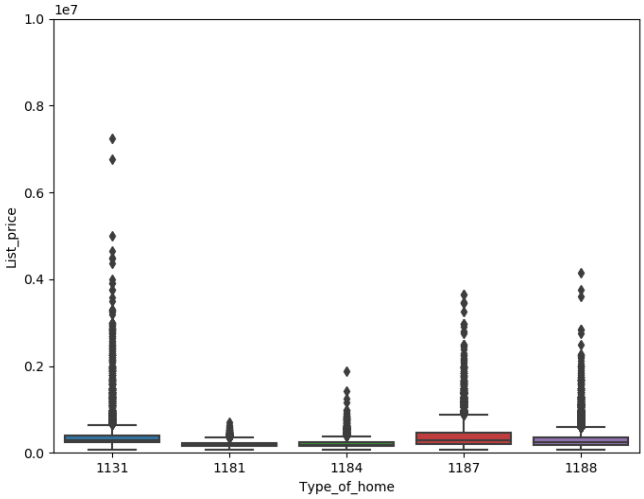

Figure 15: Boxplot of house type variable vs. list price

Appendix B Certainty rank

B.1 Methodology

Moreover, we attempt to ascertain a degree of confidence and reliability for the houses in our dataset using the Shapley values. We propose a confidence/irregularity measure based on the Shapley values computed for each home individually. The aim of this certainty rank is to single out homes that are assigned substantially different Shapley values as a result of the predictions of the model. This measure is computed as

$$\psi_i = \underset{i}{rank}(\sum_k [\underset{i}{rank}(\frac{\phi_{ik}}{x_{ik} - \bar{x}_k}) - \frac{1}{2}]^2), \tag{25}$$

where ϕ_{ik} is the Shapley value of variable k and house i. The variable ψ_i is the resulting certainty rank. The operator $rank_i(a_i)$ ranks the values a_i from lowest to highest. It assigns the number 0 to the lowest value and the number 1 to the highest value. The other values are monotonically spread over the interval between 0 and 1 with a distance of $\frac{1}{n-1}$ between each pair of consecutive values. Given that the ϕ_{ik} differs per home, we standardize it by $x_{ik} - \bar{x}_k$. A large home potentially has a large Shapley value for volume, while a small home may have a small Shapley value for this variable. This information is not particularly useful in determining which houses are treated differently by the model. Dividing by the housing characteristic ensures we filter out the houses for which the Shapley value per cubic meter is different from the average. The Shapley value is only divided by $x_{ik} - \bar{x}_k$ if the variable is numerical, such as area or volume of the house. For dummy variables, $x_{ik} - \bar{x}_k$ is set to 1 because e.g. the neighbourhood code minus the average neighbourhood code has no interpretation.

To provide intuition for (25), we give an example. Let x_i be a house with a much larger than average square meters of living space, and let k be the index indicating the square meters. We expect the Shapley value ϕ_{ik} to be much larger than the other Shapley values for square meters as well. Dividing by the square footage in excess of the average value $x_{ik} - \bar{x}_k$, gives the increase in predicted value for house i per extra square meter in comparison to the average home. This house could be treated in a special way by the model, for example for a big house the extra square meter may be worth less. In that case, the fraction $a_{ik} = \phi_{ik}/(x_{ik} - \bar{x}_k)$ is different in comparison to other houses. For this reason we rank all fractions a_{ik}. The values close to 0 and 1 are the strange values. By subtracting 1/2 and squaring the resulting value, the strange values are close to 1/4 and the average values are close to 0. By summing over the house characteristics and then ranking again, we get a measure ϕ_i which is close to 0 when the house has very average Shapley values, and close to 1 when the house has very strange Shapley values.

B.2 Results

In this section we analyse the confidence/certainty ranking based on the Shapley values. We use the results of the NN for the creation of this ranking. As explained in the methodology section, we

rank the homes in the data by how much the Shapley values deviate. The top 1% of the homes in the ranking, those ranked as deviating the most, consists primarily of high-value homes with an average value of €1,580,163. These homes are also of various types, with 14 different types of homes present with varying lot sizes. The bottom 1% of homes, those that deviate the least, have an average value of €344,988.20. These homes represent only 5 types of homes and all share the same lot size, that of 0. It can thus be observed that the most trustworthy homes are of a much lower value than those with the largest deviations from the average. They are also all a certain type of apartment. The homes that deviate the most on the other hand include a more varied group of homes with respect to type, but also with respect to the presence of a lot. These exploratory results indicate that the more atypical a home with respect to the average home in the dataset is, the more its Shapley values will deviate. This could suggest that an atypical home is special and needs special Shapley values, or it could mean that the prediction of the model might fail for atypical houses.

To further explore the differences between the houses with a low and a high certainty rank, we fit a decision tree through the data. The certainty rankings of the houses are split in two groups: the 90% most certain (group name: certain) and the 10% least certain (group name: uncertain). The decision tree will split the houses based on its characteristics to create the most homogeneous groups possible. Figure 16 shows the result of the decision tree. In every leaf we can see a measure of homogeneity (gini), the percentage of houses included in the leaf (sample), the percentage of houses in the leaf belonging to category uncertain and certain respectively (value) and which of the two groups has the majority (class). Leaves with relatively more uncertain houses are coloured more orange. From the tree we can draw conclusions on the characteristics of the uncertain houses. For example, all houses with a square footage above 190 m2 are categorized as uncertain 78% of the time. This value increases to 94% for the houses that are also built before 1958. Houses that are smaller than 190 m2 and with a volume below 300 m2, but are classified as monuments, are also seen as uncertain 59% of the time.

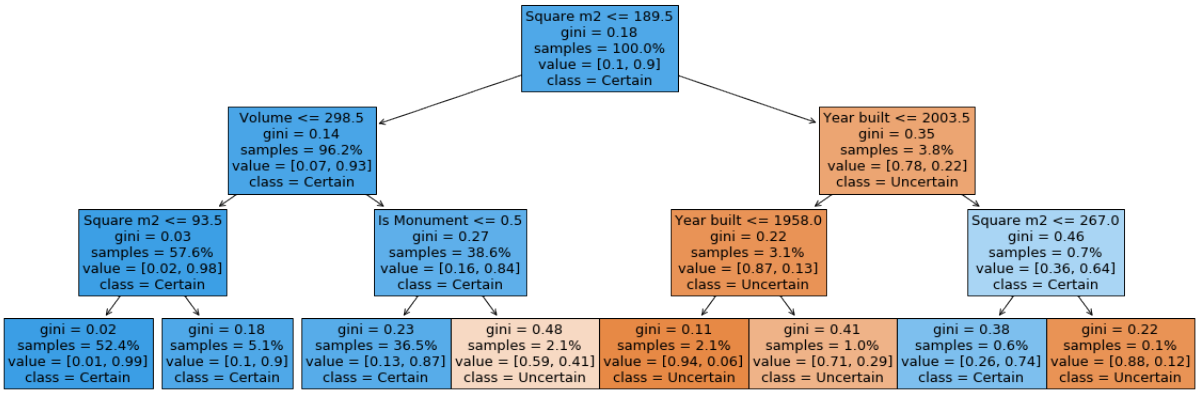

Figure 16: Decision Tree

Using Shapley values and calculating certainty ranks can therefore provide valuable information

when exploring when the model predicts accurately. We like to know which types of houses are problematic in terms of their prediction, so that we can decide whether to assign a human appraiser or account for them in other ways. A standard approach is to do a type of performance analysis, where houses with a high average test error are seen as unreliable. One disadvantage of this method, is that it is difficult to classify groups of houses that are difficult to predict. Test errors have a high variability, because errors can be small and big due to chance. Therefore, a lot of data is needed to actively distinguish low and high error housing groups. At the same time the number of houses in these special groups can be relatively small, making average prediction errors measures unreliable. The certainty rank analysis can also classify housing groups for which the prediction might be uncertain. In contrast to the performance analysis with test errors, the certainty rank analysis makes it easier to identify types of houses that need special attention, as has been done above. In addition, the certainty measures can be calculated even for houses that do not have a known value. It cannot be concluded from the certainty rank analysis that the model performance is lower for uncertain houses than certain houses. The main idea behind this analysis is to identify where the model makes different decisions in calculating the predicted value in comparison with the average house. It might very well be that different from averages houses should get a different from average treatment. It does however provide housing groups whose test errors may be explored. This way, the certainty rank analysis can play a key role in the identification of difficult to predict houses.

Appendix C SHAP

C.1 Kernel SHAP

The Kernel SHAP method is the first approximation of the Shapley values that we discuss. The Shapley values here are based on the $v_{marginal}$ value function for any underlying model. The basic idea behind this procedure is to make the approximation in equation (14) more efficient. Shapley values of house i can be added together to compute the difference in prediction value for a home in comparison to the average prediction: $\sum_{k \in F} \phi_{ik} = v_i(F) - v_i(\emptyset)$. This additivity property is exploited to approximate the Shapley value using linear regression. Adding weights to create a weighted least squares estimation finalizes the Kernel SHAP procedure.

We first define the linear regression without weights. For the Shapley values for house i, draw variables $S_r \subset F$, $1 \leq |S_r| \leq |F| - 1$ for $r = 1, ..., M$ with M the maximum number of iterations. The variable z_{kr} keeps track of which variables are in S_r and k is a variable index $k = 1, ..., p$. Set z_{kr} to 1 if $k \in S_r$ and set z_{kr} to 0 otherwise. The prediction y_r is the prediction with the variables in S_r minus the average prediction: $v_i(S_r) - v_i(\emptyset)$. The regression is given by

$$y_r = \phi_{i1} z_{1r} + ... + \phi_{ip} z_{pr} + \varepsilon_r, \tag{26}$$

where ϕ_{ik} is the Shapley value for house i and variable k. The intuition behind the regression comes from the additivity property of the Shapley values. The difference in prediction by adding the variables in S_r should sum approximately to the Shapley values that correspond to the variables in S_r. It is an approximation, because the Shapley value is the average over all different orderings and the S_r represents only a part of these orderings. Lundberg and Lee (2017) prove that with correct weighting, the $\hat{\phi}_{ik}$ from the weighted least squares are consistent estimators for the Shapley values.

$$w_r = \binom{p}{|S_r|} \frac{|S_r|(p - |S_r|)}{p - 1} \tag{27}$$

The provided weights are shown in (27). The intuition behind these weights is linked to the orderings σ from section 4.2.2. When ordering p variables and the variable of interest is the first in the order, the other variables can be ordered in $(p-1)!$ different ways. When the variable of interest is in the kth position, the number of ways the variables before k can be ordered is $(k-1)!$ and after k is $(S-k)!$. This means that some subsets S_r should receive more weight than others. The Shapley weight would be $\binom{p}{|S_r|-1}$. The first factor in the regression weight (27) is very similar to this Shapley weight, so this gives us some intuition on why the weights w_r follow this formula. The estimation procedure can further be sped up by sampling S_r that have a high weight w_r more often, and correct the weights accordingly.

In comparison with the two model-specific Shapley value estimators in the two upcoming sections, the Kernel SHAP method can be used to estimate the Shapley values for all models. In turn, Kernel SHAP is a slower algorithm than Deep SHAP and Tree SHAP. There is an additional benefit of using Kernel SHAP instead of the other two. The y_r variable can be freely transformed before

the Shapley values are calculated. This allows us to first reverse the logarithms of the house prices, as used in the HTM model. The Deep SHAP and Tree SHAP have to conform to the functional form of the predicted value as given by the model. If we want to calculate the Shapley values in euros, this means that we cannot take the logarithm of the list prices. In order to model the housing prices in the Neural Network and Random forest with logarithms and still get the Shapley values in euros, Kernel SHAP can be used instead.

C.2 Methodology of DeepLIFT and Deep SHAP

The DeepLIFT method, introduced by (Shrikumar et al., 2017) for explaining the output of neural networks, analyses why the output of a certain observation deviates from a given reference value. More specifically it explains this deviation by how the activation of this observation deviates from the activation of the reference input. The reference value depends on the purpose of the model and needs to be set by the designer of the model. The Deep SHAP method is equal to the DeepLIFT method but with a specific reference value. We will discuss this reference value more in-depth after discussing the methodology of the DeepLIFT method.

By using backpropagation, the DeepLIFT method analyses these deviations of the reference value for every neuron in each layer. In this case the output layer contains only one neuron. The output of this neuron y is the value of the house. The DeepLIFT method thus analyses how the deviation of a list price from a set reference value is related to how the corresponding explanatory variables deviate from their reference values. The deviation of y from the reference value y_{ref} is defined as $\Delta y = y - y_{ref}$. The deviation of x from the reference value x_{ref} is defined as $\Delta x = x - x_{ref}$. The y is the final output, but each neuron in the hidden layers has output z_h for $h = 1, ..., n$, where n is the number of neurons in a layer. We define the deviation of z_h from the reference value $z_{ref,h}$ as $\Delta z_h = z_h - z_{ref,h}$. The change in Δz_h caused by the change in Δx_k is called the contribution score $C_{\Delta x_k \Delta z_h}$. In other words the interpretation of $C_{\Delta x_k \Delta z_h}$ is: "How much does the output z_h of a neuron change, when the input x_k is changed with Δx_k?" Consistent with the previous sections we have $k = 1, ..., p$ explanatory variables. The previous layer gives the input for the consecutive layer. The DeepLIFT method computes the contribution scores for each neuron in each layer. The sum of the contribution scores over all inputs is equal to the deviation of the output Δz_h see (28). Shrikumar et al. (2017) describe this as the summation-to-delta property. For the first hidden layer, with the inputs being the explanatory variables, the summation-to delta property for each neuron z_h is defined as

$$\sum_{k=1}^{p} C_{\Delta x_k \Delta z_h} = \Delta z_h. \tag{28}$$

For the output layer, with the inputs being the outputs of the previous layer, the summation-to delta property for y is defined as

$$\sum_{h=1}^{n} C_{\Delta z_h \Delta y} = \Delta y. \tag{29}$$

They define define the multiplier of Δx_k to Δz_h, $m_{\Delta x_k \Delta z_h}$ as

$$m_{\Delta x_k \Delta z_h} = \frac{C_{\Delta x_k \Delta z_h}}{\Delta x_k}. \tag{30}$$

The multiplier captures the effect that the deviation from the reference value for an input neuron has on the deviation of the output neuron, proportional to the magnitude of the input deviation. For the last layer, this is the contribution of an input neuron to the deviation of the value of the house. For the first layer, an example is how much one additional square meter deviation contributes to the deviation of the output of a neuron in the first layer. This multiplier is thus defined for each connection between neurons of two consecutive layers. To analyse the effect of the actual input variable on the final output, Shrikumar et al. (2017) use the chain rule for multipliers. The ultimate goal of the DeepLIFT here is to connect the value of the house to the explanatory variables. For explanatory purposes we give an example with a NN with one hidden layer. Our output layer only has only one output neuron which captures the value of a house. This value is defined as y. The inputs of the hidden layer are the p explanatory variables $x_1, ..., x_p$. We define this layer with n neurons: $z_1, ..., z_n$. To compute the multiplier for a given explanatory variable k to the value of a house i, $m_{\Delta z_h \Delta x_k}$ the multiplier chain rule is be defined as

$$m_{\Delta x_k \Delta y_i} = \sum_{k=1}^{p} m_{\Delta x_k \Delta z_h} m_{\Delta z_h \Delta y_i}, \tag{31}$$

where $m_{\Delta z_h \Delta y_i}$ is the multiplier for the input of the neurons of the hidden layer to the final output and $m_{\Delta x_k \Delta z_h}$ is the multiplier for the explanatory variables to the outputs of the neurons in the hidden layer. When adding a second hidden layer, the same methodology for computing the multiplier for each neuron applies. DeepLIFT uses backpropagation to compute the multiplier of each explanatory variable for the value of the house. Thus, first the multipliers for the inputs of the second hidden layer to the final output are computed. Using these multipliers, the multipliers of the inputs of neurons in previous layers can then be computed.

The Deep SHAP method is an adapted version of the DeepLIFT. Lundberg and Lee (2017) calculate the Shapley value by setting the reference value of the explanatory variables to their average value. They define the approximate Shapley value as

$$\phi_{ik} = m_{\Delta x_k \Delta y_i}(x_k - E[x_k]). \tag{32}$$

This method works due to the fact that the neural network is approximated as a linear function by making use of the multipliers as described above. The term $m_{\Delta x_k \Delta y_i}$ has a similar interpretation as the derivative of y_i to x_k. This linearisation has as a consequence that the variables are independent: the order σ in which the variables are added becomes irrelevant. Hence, the Shapley formula given in (13) reduces to (32). This means that this approximated Shapley value is close to the actual Shapley value if the house characteristics are close to the average house characteristics.

This is an argument for using the Deep SHAP to only compare a house with other houses that are similar. Whether the approximation is of sufficient quality for houses that diverge far from the average is unknown.

C.3 Methodology of Tree SHAP

Lundberg et al. (2018) introduce the Tree SHAP method to estimate Shapley values for the Random Forests and other tree-based methods. Using the the Tree SHAP method to compute exact the Shapley values highly reduces the computational time needed. We define the following for the purpose of the following explanation: T represents the number of trees, L is the maximum number of leaves in each tree, M is the number of variables, D is the maximum depth of any tree and R is the number of houses to compare against.

They first introduce a slower, less complex algorithm that can estimate Shapley values exactly in $O(TLM2^M)$ time. It computes Shapley values with value function $v_{conditional}$, see section 4.2.1. This algorithm estimates following the standard procedure of the definition of Shapley as shown in (13). If S contains all variables then the prediction from the node where x lands is the prediction. If S contains no variables, the weighted average of all the prediction nodes is used. If S contains a subset of the variables that is not empty, predictions of unreachable nodes are excluded. An unreachable node is a node that includes a variable that is not present in the specific subset. The remaining branches are weighted by the number of samples per branch to produce the prediction. This procedure is run for every possible subset in order to compute the Shapley values. Therefore, the computational time of this algorithm is high.

Lundberg et al. (2018) have introduced an algorithm that computes the exact Shapley values as well. However it reduces the computational time from $O(TLM2^M)$ to $O(TLD^2)$. Instead of following the procedure mentioned above for each possible subset one by one, this algorithm pushes all possible coalitions down the tree simultaneously. For example, if the split on the dummy outdoor garage is done in a specific node, all subsets containing houses with an outdoor garage need to go to the same branch. In order to do this, the algorithm needs to oversee which subsets go down each branch and it does so via recursion. The weighting of these splits differs from the simple algorithm as the size of the subsets also need to be taken into account, given that the weighting of the Shapley values depends in part on $|S|$. To take that into account, all the possible sizes of the subsets are tracked during the recursion. Therefore, by keeping track of which subsets go down on which branch, the algorithm needs to follow the above-explained procedure only once. This reduces computational time considerably in comparison to the one by one algorithm.

A third algorithm has been proposed by Lundberg et al. (2019). In contrast to the methods described above, the estimated Shapley value is based on the value function $v_{marginal}$. The computation time for the Independent Tree SHAP is $O(TRL)$. Instead of pushing the variables through the tree, this algorithm takes a weighted average of the leaf values. The weights are calculated using a background sample of houses in such a way that the features are independent of each other. Their procedures lay out how these weights are calculated efficiently. Because we are interested

in the Shapley values based on the $v_{marginal}$ value function, this algorithm will be used in our research.