Proceedings of the 55th Hawaii International Conference on System Sciences | 2022

Visual Interpretability of Image-based Real Estate Appraisal

Jan-Peter Kucklick
Paderborn University (UPB)
jan.kucklick@upb.de

Abstract

Explainability for machine learning gets more and more important in high-stakes decisions like real estate appraisal. While traditional hedonic house pricing models are fed with hard information based on housing attributes, recently also soft information has been incorporated to increase the predictive performance. This soft information can be extracted from image data by complex models like Convolutional Neural Networks (CNNs). However, these are intransparent which excludes their use for high-stakes financial decisions. To overcome this limitation, we examine if a two-stage modeling approach can provide explainability. We combine visual interpretability by Regression Activation Maps (RAM) for the CNN and a linear regression for the overall prediction. Our experiments are based on 62.000 family homes in Philadelphia and the results indicate that the CNN learns aspects related to vegetation and quality aspects of the house from exterior images, improving the predictive accuracy of real estate appraisal by up to 5.4%.

1. Introduction

In the financial industry, real estate appraisal is one essential application [1] and refers to the price estimation of properties or parcels. Many different stakeholders like real estate customers, agents, financial lenders, and local municipalities are interested in a fast and accurate estimation process supported by computer-assisted mass appraisal (CAMA) [1, 2, 3, 4]. Recently, the data structures got more advanced by including image data in addition to numerical and categorical features, and the algorithms applied got more sophisticated by using deep convolutional neural networks (CNNs). This type of neural network has the ability to spatially decompose images into edges and textures for an analysis. Both resulted in an increased predictive performance [1, 3, 5, 6, 7, 8, 9]. Nevertheless, the use of these algorithms led to more intransparency, limiting the interpretability of CAMA systems. Interpretability refers to providing a human understandable algorithm [10]. Nevertheless, interpretability has been identified as a vital factor in financial machine learning. Three circumstances foster the need for explainability: First, interactions between humans and machines rely on trust. While black-box systems decrease trust because of a missing explanation, interpretability methods can establish open-mindedness about innovative technologies [11]. In a real estate transaction process, many different stakeholders are involved, and trust between parties as well as trust in the CAMA system is essential for a successful real estate sales process. Second, financial transactions are high-stake decisions as a significant monetary loss can be expected when they fail [10, 11, 12]. Therefore, to prevent serious negative effects and to ensure overall quality, CAMA supported by machine learning needs to be controlled by using explainability techniques. Third, governments have regularized the usage of machine learning systems for high-stake decisions by new laws. One example is the right to explanation in the General Data Protection Regulation (GDPR) of the European Union. Therefore, decisions made by machine learning need to be explainable to the customer [11, 13].

Consequently, interpretability is necessary for a real estate appraisal, and there are two ways to establish it: On the one hand, one can make use of inherently interpretable models like linear regressions or decision trees, which the user can read [14] — on the other hand, one can use post-hoc interpretability methods to explain the decision of a black-box algorithm [10]. The choice between inherently and post-hoc interpretable models corresponds to the often stated accuracy-interpretability trade-off because black-box models often perform better in terms of accuracy, but they are opaque in their decision-making [14].

To reduce the stated trade-off, we examine visual interpretability methods for regression tasks to answer the research call of Law et al. [3] for improved explainability in real estate appraisal. We analyze

URI: https://hdl.handle.net/10125/79519
978-0-9981331-5-7

62.000 single-family homes in Philadelphia, PA, and identify visual aspects for price prediction.

While the study by Bin et al. [15] is closely related to ours due to the same geographic location, we go beyond predictive performance by focusing on visual interpretability methods for exterior images. Furthermore, we differ in the choice of dataset, using open data as opposed to a realtor-provided dataset.

Our results show that the predictive performance improves by up to 5.4% when exterior images are used in addition to housing attributes. We notice that the selected CNN architecture has a great influence on the performance and that attention mechanisms do not improve the accuracy in our experiments. Important aspects gathered from the image are information related to the vegetation like trees within the street and yards as well as aspects related to quality like damages on the housing structure.

Our contribution is three-fold: First, to the best of our knowledge, we are the first who use Regression Activation Maps (RAM) and Convolutional Block Attention Module (CBAM) in the field of real estate appraisal for visual interpretability of the exterior image. We derive insights that vegetation in the front yard and housing conditions seem to be important image features. Second, we estimate the value exterior images have for the real estate appraisal process based on a two-stage modeling approach. Finally, we add to the existing body of real estate appraisal research and provide additional empirical evidence that including exterior images increases the predictive performance.

The remaining paper is structured as follows: Chapter 2 summarizes the related work, while chapter 3 introduces the dataset and machine learning model. Chapter 4 discusses the results, while chapter 5 concludes this paper with limitations, implications, and an outlook for future research.

2. Related Work

This chapter will summarize essential research about real estate appraisal, machine learning techniques used for image-based appraisal, and available interpretability methods for regression tasks.

2.1. Real estate appraisal

Lancaster [16] and later on Rosen [17] proposed the hedonic pricing model, where the overall price of an object is the total sum of the value contributions of its utilities. The hedonic model is often mathematically based on a linear regression and has been the predominant model for real estate appraisal due to its economic interpretability [3]. Through this approach, technical characteristics like size, age, location, condition, and different amenities like parking spaces, fireplaces, and pools were used to estimate the total price of a house [2, 3, 18]. These features often capture hard facts about the property and are stored in numerical or categorical variables [19]. However, soft information relating to the house's appearance also has a considerable influence on the price [3]. These factors include the perceived safety and bustle of the neighborhood, social-economic status and luxuriousness, privacy of the parcel or factors indicating the possibilities for relaxation [3, 20, 21, 22, 23]. A house is often sold when the buyer's characteristics match with the real estate style [21]. Nevertheless, the soft information are implicit and context dependent and therefore often enclosed in unstructured data [19]. Recent research used convolutional neural networks (CNNs) to extract soft information from different image types ranging from floorplans [24], streetmaps [25], and satellite images [3, 5, 8, 26] to interior [7, 23] or exterior images [1, 9, 15, 26]. Most information extracted from exterior images relates to the real estate appeal. The property's aesthetics can be split into two subcategories related to the house and the landscape [20]. Within both subcategories, the style and quality perception are of interest for the price estimation. A higher level of greenness within the area indicates a larger yard and therefore more possibilities for relaxation. This factor indicates more expensive houses [26]. Other variables corresponding to vegetation are trees on the lot and within the street, both having a positive price influence [27] because they can reduce traffic noise and spend shadow in the summer [20]. Also, the lawn quality [28] or trimmed vegetation on the parcel [29] are factors that should be considered. They indicate that the house owner has taken care of a high maintenance yard. The absence of a yard and many concrete areas combined with broken sideways are signs for a lower-priced home [29]. One possible explanation could be that these visual signs are often present in unpleasing and less prestigious environments [20]. Likewise to style and quality aspects of the yard, aspects about the aesthetics of the house and the condition of the dwelling can influence the real estate value. Style aspects such as the type of roof affect the aesthetics and increase the price [1]. Similar to the quality signs of the landscape, factors indicating the house's maintenance (e.g., fresh paint, no broken objects) correspond to a higher real estate value [6]. In general, the condition and first impression of the dwelling seem to correlate with the appeal and the price [15]. All these soft factors can be extracted from image data and enhance the feature space about real estate. As previously omitted in the pricing process, they can now increase the predictive performance [29].

2.2. Real estate modeling strategies

Different machine learning approaches have been used for integrating image data into the appraisal process. Multi-view neural networks are models that handle different data types (e.g., tabular and image data) in an end-to-end fashion. In a multi-view neural network, a CNN for the image data and a fully connected network for the tabular data are combined [1, 3, 8, 9].

In contrast to this one-stage approach, tabular and image data can also be analyzed in separate models forming a multi-step approach [8]. In this modeling strategy, the CNN is used as a feature extractor, transforming the image data to a tabular output, which can be a final prediction (classification or regression) or an intermediary output of the CNN. For example, the CNN can be a place classifier trained on the Places365 dataset [30] with the aim to distinguish different places (i.a. apartment block, street, church, yard, waterfall). The exterior image is classified into these categories and the probabilities of the images belonging to the different classes are then used as additional variables in a downstream hedonic model [6]. Alternatively, one can label an own dataset as Poursaeed et al. did [23]. The authors gathered interior and exterior images and labeled them into different luxury classes. The prediction was then later on used to estimate the overall real estate price in a separate machine learning model. Other authors did not use explicit features like categories but implicit features by using the dense feature representation of the image from the CNN, which was previously trained to predict the price or categories of prices (e.g., below average, average, above average) [5, 15, 31]. Again other authors use a boosting approach, first training a hedonic house price regression. Then they train a CNN to predict the residuals of this regression with the house images. Last, they repeat the hedonic pricing model on the house price and complement the house attributes with the residuals predicted by the CNN [7].

These different approaches have separate strengths and weaknesses. One does not need to select a target variable for the feature extraction in one-stage approaches like multi-view neural networks [3, 8]. However, they are also more intransparent as these methods are black-boxes and no post-hoc interpretability method exists that can deal with multiple data types. Contrasting, multi-stage approaches have the advantage that the complexity of the overall process is split over multiple models. For each model, interpretability can be incorporated by using inherently transparent models for performing the regression [7] or by using post-hoc interpretability methods like Feature Importance [26]. Nevertheless, a suitable target variable is necessary to train the CNN [8].

Besides the different modeling strategies, a recent improvement in machine learning has been the attention mechanism. Loosely speaking, this new type of layer helps to focus the neural network on relevant features. While this technique is often used in machine translation models [32], some authors use it for real estate appraisal [15, 25]. They implement the attention module as a fully connected layer with a softmax activation for the housing attributes [25] or for the globally pooled image features [15]. While the attention mechanism was applied on one dimensional data only, one advancement for CNNs has been CBAM, which is a special attention mechanism for image data. This method helps the CNN to focus on important convolutional channels and essential spatial dimensions in the image [33]. Thus, the CBAM improves the predictive performance [33]. In the next section, we will give an overview of different interpretability techniques for regression tasks.

2.3. Interpretability methods of visual regression tasks

Interpretability is one desired property of a machine learning system and part of explainable artificial intelligence (XAI). Accordingly, XAI aims to create understandable, comprehensive, interpretable, and transparent machine learning systems [12, 34]. Interpretability in this context describes the ability of a human to understand the system. While interpretability approaches exist for image classification, ranging from input permutations [35] to the analysis of gradients [36] or the usage of local approximations [37], methods for continuous outcome variables (regressions) are rare. Two different methods are available, namely sliding-window heatmap (SWH) [38] and Regression Activation Maps (RAM) [39]. Both can generate saliency maps, indicating the most important parts in the image according to the value contribution of the region. In combination with the CBAM layers, explainability should be increased. As CBAM helps to focus on important spatial regions and essential features (channels), derived saliency maps for interpretability tend to be clearer, which was shown in exemplary classification tasks [33].

2.3.1. Regression activation map (RAM) RAM originates from the medical domain, but it is not domain-specific [39]. This method is the counterpart to Class Activation Maps (CAM) [40] and is its version for regression. In a CNN, the last convolutional layer contains k feature maps denoted by g_k, with spatial coordinates (i,j). These feature maps are globally averaged pooled and weighted by w_k for regressing the output Y (see Formula 1). Z denotes the number of pixels in the feature map. Consequently the prediction of the regression output is the weighted sum of the feature maps, where each feature map contributes to the prediction by w_k.

$$\hat{Y} = \sum_{k=1}^{K} w_k \cdot \overbrace{\frac{1}{Z} \cdot \sum_{i,j}}^{global\ average\ pooling} g_k(i,j) \qquad (1)$$

The saliency map G contains the price influence for each spatial coordinate (i,j). For RAM, the price influence is the linear additive combination of each feature map g_k, weighted by the regression coefficient w_k, which can be expressed as:

$$G(i,j) = \sum_{k=1}^{K} g_k(i,j) * w_k \qquad (2)$$

2.3.2. Sliding window heatmap (SWH) SWH is a method based on occlusion sensitivity. Initially proposed by Zeiler and Fergus [35] for classification, it has been adapted for regression problems [38]. By sliding a patch (x') over an image, the difference between the prediction without and the prediction with the patch is calculated [14]. This can be formulated as:

$$G(i,j) = \widehat{f(x)} - f(\widehat{x'_{i,j}}) \qquad (3)$$

with G being the saliency map and $G(i,j)$ the price contribution of the (i,j)'s coordinate in the image. When important aspects of the image are occluded, the final prediction should change significantly.

Both methods, SWH, and RAM have advantages and hurdles. SWH is model-agnostic because it does not need to access the model's weight, therefore it is useable for a large variety of algorithm classes. Nevertheless, its downside is its sensitivity towards the hyperparameters patch size and color [14, 41], leading to inconsistent results. Regression Activation Map is robust[1], however as it relies on global average pooled features for the prediction, architectural limitations arise. The model needs to include a global average pooling layer and can not have any fully connected layers between the pooling and output layers. Because the sensitivity towards the hyperparameters of SWH can significantly change the results [14], we will use RAM for this research paper to get a visual explanation of the exterior images.

3. Dataset and Modeling

This chapter gives an overview of the used datasets and tested models.

3.1. Data preparation

We use appraisal open data from 62,641 real estates in Philadelphia, PA [43] for our experiments. The dataset includes details about the size (lot area, bedrooms, bathrooms), age, condition, technical details like the type of heating system, view from the dwelling, and location details in the form of the zip code. Moreover, the dataset includes amenities like fireplaces and garage spaces. We focus on family homes, excluding condos and apartments from the dataset. We delete observations with unlogical values, i.a. having a size of 0 square feet or a price of 0 USD. We regroup rare values in the categorical variables to the value 'others'. The descriptive statistic is summarized in Table 1. The mean value of a house is 81,303 USD. On average, it has three bedrooms.

We download the exterior images from Bing Maps [44] in a three step approach. As the geographical coordinates are required for the download, we first transform the address of the houses to latitude and longitude with Bing Maps. In a second step, we calculate the bearing from the street to the house, similar to the approach from Johnson et al. [29]. We use the direct neighbors on the opposite side of the street as a reference. In a third step, based on the coordinates and the bearing, we use the Bing Maps API to obtain a 512 by 512 pixels image for each house. We resize the images to 256 by 256 pixels to reduce training times of the CNN.

To pre-process the housing attributes, we use n-1 dummy variables for categorical variables. Numerical variables like the size of the living area in square feet or the age are standardized. We perform a logarithmic transformation of the real estate appraisal value to get a normal distribution of the target variable.

Setting our experiment in the context of related work, instead of using a one-stage modeling approach [3, 8], we use a two-stage modeling approach like in [6]. The reasons for using a two-stage approach is that visual interpretability techniques like RAM can be applied on the CNN to increase its interpretability. Additionally, to explain price effects from the structured data, we use a Linear Regression for the price estimation, following the traditional hedonic pricing model [17]. By this

[1] We test the robustness of RAM by using the Sanity Checks proposed by Adebayo et al. [42] in chapter 4.3

Table 1. Summary description of the dataset

Variable	Mean	Standard Deviation	Minimum	Maximum
Price	81,302.82	58,554.22	3,400.00	1,004,300.00
Garage Space	0.12	0.33	0.00	3.00
Total Living Area	1,140.04	275.30	500.00	7,262.00
Full bath	1.03	0.19	1.00	4.00
Bedrooms	3.03	0.40	2.00	6.00
Age	90.74	11.43	1.00	132.00

combination, we ensure interpretability for both data types in the multi-view learning model, which is hard to achieve in a one-stage model. Instead of using an intermediary target variable for training the CNN [5, 6], we use the real estate price to ensure a meaningful feature extraction. Therefore, we model the real estate appraisal process as:

$$y = \alpha + \beta \hat{f}_\theta(I) + \gamma X + \epsilon \qquad (4)$$

with X being the housing attributes and $\hat{f}$ being the price estimate based on the exterior image I. In a first step, the CNN f is trained to predict the overall real estate price by using the exterior image. We do not model this step as a classification task, as price classes like above and below average are often too generic to capture fine granular image features. The same holds true for using a place classifier as these models can often only distinguish between places like park, apartment block or house, however, are not specific enough to capture factors like the quality and appearance of the house and the lot as described in chapter 2.1. The overall real estate price is estimated in a second step (Formula 4), by combining the house characteristics X and the price estimate of the image $\hat{f}$. Our modeling approach has the advantage that the coefficients α (Intercept), β (influence of the price estimate based on the image) and γ (price influence of the house characteristics) remain interpretable.

As a baseline, we use a hedonic pricing model without interaction effects [2].

For the CNN f, we use different architectures. Model 1a and 1b are based on the VGG-16 architecture [45] following previous related research [6, 7]. We customize the VGG-16 model by deleting its fully connected layers, to be able to apply RAM. Model 2a and 2b use the ResNet50v2 architecture [46]. We select the ResNet model because it has previously been stated to be a better-performing CNN architecture due to its robustness against vanishing gradients [46]. Despite the two different architectures, we also test whether additional CBAM layers can improve the real estate appraisal performance. Therefore, after each convolutional block in the VGG-16 and ResNet50v2 architecture, we implement a CBAM layer with the parameters described in [33]. Our tested models are Model 1a (VGG), Model 1b (VGG with CBAM), Model 2a (ResNet), Model 2b (ResNet with CBAM)[2]. To the best of our knowledge, we are among the first to use CBAM attention mechanisms for real estate appraisal.

We train our models with a batch size of 32, Adam optimizer with a learning rate of 0.001, a maximum of 80 epochs with an early stopping applied on the validation loss. We perform five-fold cross-validation to check the robustness of the model performance. We split the data randomly into 80% training, 10% validation, and 10% test set. The next chapter will report the results in detail.

Table 2. Mean performance of the tested models from the 5-fold cross validation. Standard Deviations are reported in brackets. Best model results marked in bold.

Modelname	MAE	RMSE
Baseline	18,533 [342]	33,413 [2,909]
Model 1a	18,540 [342]	33,521 [2,923]
Model 1b	18,533 [342]	33,412 [2,909]
Model 2a	**17,894** [344]	**31,586** [2,185]
Model 2b	18,115 [416]	32,186 [2,783]

4. Results

In this section, we summarize our results concerning the metrical performance and interpretability of the tested models.

4.1. Metrical evaluation

Based on our metrical results stated in Table 2, we conclude that adding additional information from the

[2] Figure available via: `bit.ly/visual_XAI`

exterior image of a real estate improves the appraisal performance, which supports previous research [6, 9, 26, 29]. However, we find varying performance between the used architectures. Model 1a and 1b based on VGG-16 could not outperform the baseline estimate of the linear regression without an image based price estimate. Model 2a and 2b based on ResNet50v2 are 5.4%, respectively 3.6% better in RMSE than the baseline. Surprisingly, the additional CBAM layers could not improve the performance compared to the standard ResNet50v2 architecture.

4.2. Regression coefficients

By using a two-stage modeling approach, the interpretability of coefficients is maintained. In the following, we perform the coefficient analysis for the best model, Model 2a. For example, increasing the living area by one standard deviation, c.p., raises the real estate value by $\exp^{0.15} - 1 \approx 16\%$. Increasing the age by one standard deviation, c.p., reduces the price by approximately 1.4%. When an air condition system is added to the real estate, the estimated price increases by 12%. Compared to the exterior condition above average, the condition average results in a price decrease of approximately 9.4%. When the house is located in the Zip Code Area 19119 (Mount Airy), the price increases by 46% (Figure 1). This effect seems natural, when taken in consideration that Mount Airy was awarded of one of the most attractive big-city neighborhoods to live in [47].

We estimate a price increase between 15% (Model 2b) and 23% (Model 2a) for each standard deviation increase of the value estimate of the CNN. Consequently, the exterior image has a stronger price influence than adding an air condition or equally or stronger effect than increasing the living area by one standard deviation (approx. 275 square feet). The price estimation of the exterior image is obviously one important feature in the final price estimation. The coefficients show that our model extracts latent aspects related to the price. Nevertheless, without visual interpretability methods, the CNN remains a black box. In the next section, we apply RAM to identify important image features which are influencing the price.

4.3. Visual interpretability results by RAM

RAM highlights important aspects in the image (see Figure 2), where price increasing areas are indicated in red, areas decreasing the price in blue and neutral areas in green. We find that trees within the street (Image A), as well as a nice front yard including a lawn and vegetation seem to increase the price (Image B). Both factors have been previously identified as important aspects in real estate appraisal literature [20, 26, 27, 29]. Additionally, the quality and aesthetics of the house can have a price influence. For example, the poor condition of neighboring houses (broken down front, no roof etc.) lowers the price of a house (Image D). Not only the pure condition, but also the dwelling aesthetics can influence the price, where a marquee is raising the price, while the empty entrance of the neighbor decreases the price (Image C). These factors match previous research on price influencing factors [1, 6, 15, 20]. However, we notice that the influence of aesthetics might go beyond the house and the yard because the neighboring houses and their condition also seem to influence the real estate price.

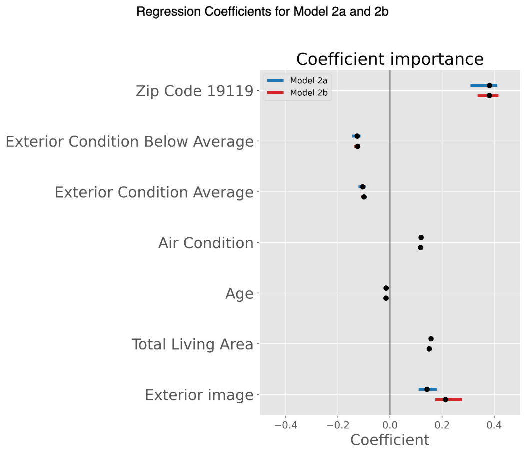

Figure 1. Selected regression coefficients for the linear regression with either using Model 2a or 2b for the CNN

Despite the content wise analysis of the saliency maps, we noticed that for some examples the sky (Model 2b) or cars (Model 2a) were activated as an influential factor (e.g. Image B, D). It seems that the CNN (mostly Model 2b) might have learned noise from the image like clouds, or doors of a car.

In the past, post-hoc explainability methods got also criticized because they do not necessarily explain anything about the model. A truthful interpretation is dependent on the model and the data - in other words, if the model or the data changes, the explanation should change too. Adebayo et al. [42] detected that different interpretability methods violate this assumption and are invariant to the dataset or to weights of the neural network. Therefore, the authors formulated two sanity checks, to assess the explainability. The first one inspects the explanation with respect to the model weights, while the second one tests the explanation against the data invariance. Testing the

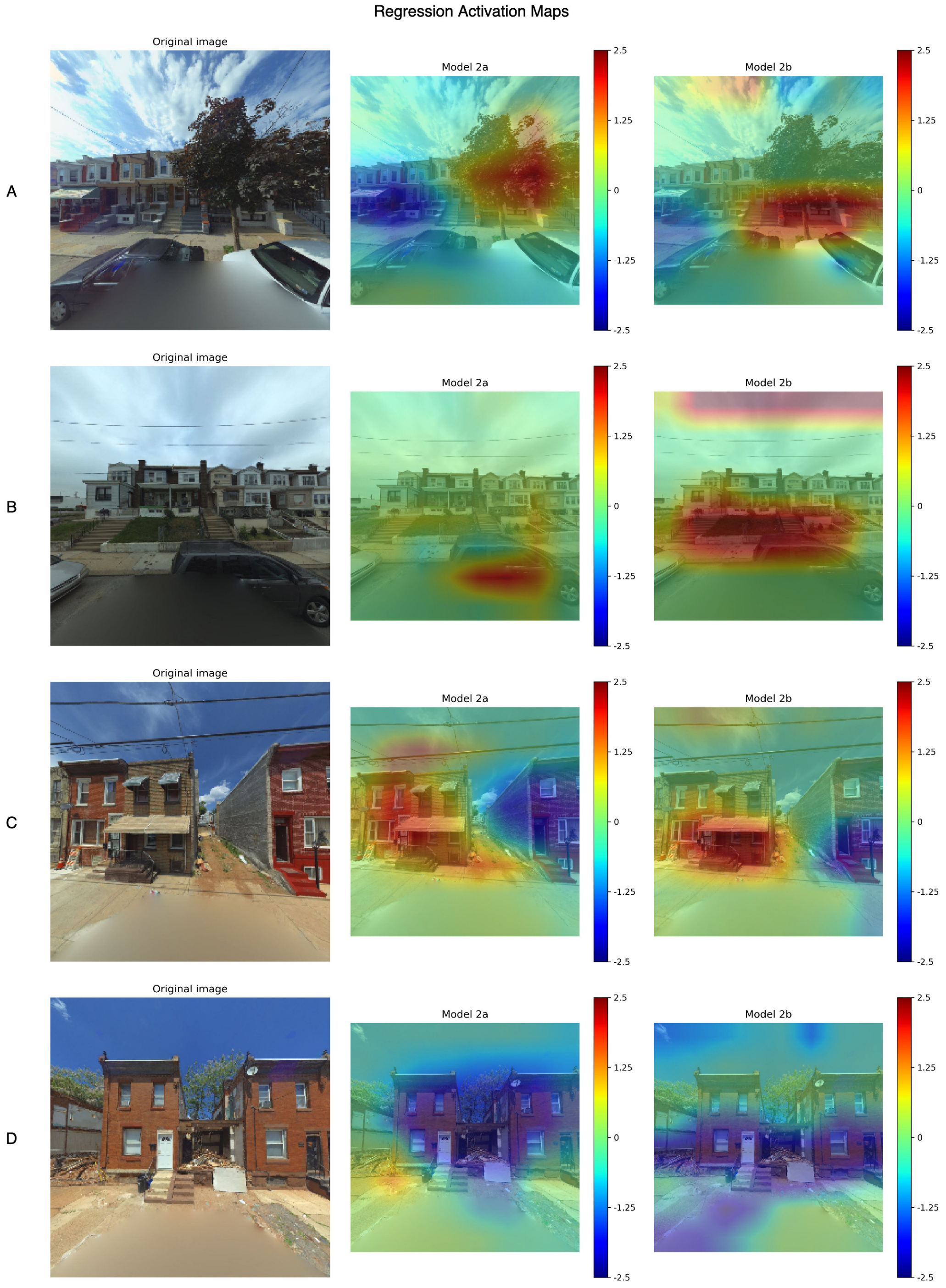

Figure 2. Regression activation maps for Model 2a and 2b. Red/blue indicates areas increasing/decreasing the price. The scale is measuring the price change in percent. Original images: ©Microsoft Inc. [44].

explainability method against weight invariance can be performed by calculating the Spearman correlation, the absolute Spearman correlation and the Structured Similarity Index Measure (SSIM) between two sets of explainations. The first ones are gathered from the original model. For the second set, the model weights are randomly reinitialized step by step from output layer to input layer. The post-hoc explainability method passes the test when the explanation changes for a changing model, leading to a sharp drop in the correlation and SSIM, beginning in the top layers. To the best of our knowledge, we are the first performing this sanity check for RAM. For RAM for both models 2a and 2b, the correlation, absolute correlation and SSIM drop with the start of the weight randomization (black dotted line) (Figure 4). Therefore, RAM is variant to the model's weights and passes the test.

For checking the model's explanation against invariance to the data, a new model is trained with the same architecture and input data, however, the training labels are randomly shuffled [42]. Many modern machine learning algorithms still perform very well on the training set with shuffled data as these models can remember the true label. However, they will perform poorly on the test data. Similar to the sanity check of the weights, the Spearman correlation and absolute Spearman correlation between the explanations for the test data of the original model and the new model which was trained on shuffled training data are calculated. If the explainability method is invariant to the data, in other words, if the explanation does not change when the label changes, the explanation method does not explain anything about the true label. It fails the check if both correlation measures are high, indicating no difference between the explanations. For both models, we see only a weak (absolute) correlation (Figure 3), indicating that the RAM is sensitive to the data and thus passes the test. Concluding from the sanity checks, RAM seems to be a suitable explainability method for regression problems.

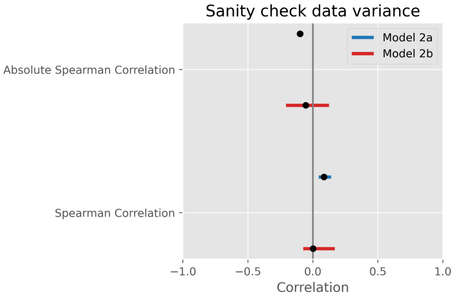

Figure 3. Sanity checks for data variance for Model 2a and 2b based on Adebayo et al. [42].

5. Discussion

Our results indicate that by using a two-stage modeling approach, interpretability of the regression coefficients can be maintained, while through using post-hoc explainability methods like RAM, visual features detected by the CNN can be identified. This answers the call from Law et al. [3] for image based interpretability in real estate appraisal. Moreover, visual interpretability helps to use sophisticated CAMA systems because the user's trust can be generated. Additionally, the application can be made GDPR compliant, because it supports the right to explanation. Finally, CAMA systems can be inspected and debugged by using RAM. These results support the research of [10, 13] for advancing on XAI in financial high-stakes decision.

Important factors derived from the image relate to the vegetation and quality aspects of the house, which supports previous related research [6, 15, 20, 26, 27, 29]. However, while these factors relating to soft information were previously identified by manually labeling all observations or at least some images to train a CNN [3, 20, 29], we show as first authors that these variables can be extracted automatically by the CNN with training only on the real estate price. No additional labeling of vegetation or house quality is required. While the performance gain in RMSE is similar compared to related research [6, 26], we found that the used CNN architecture strongly influences the predictive power. A possible explanation why VGG-16 did not outperforme the baseline could be that by omitting the last two fully connected layers to apply RAM lowered the performance significantly. Otherwise, it is possible that this algorithm had issues with the vanishing gradient problem, while the ResNet50v2 architecture is less prone to this phenomenon [46]. Additionally, it remains unclear why using additional CBAM layers did not improve the accuracy of the models. Finally, performing the sanity checks for saliency maps [42] revealed that RAM is variant to the model's parameters and the label and thus being a reliable post-hoc interpretability methods for regression problems.

6. Conclusion

The performed study progresses on visual interpretability for image-based real estate appraisal by answering the research call from Law et al. [3] for more explainability. Nevertheless, it does not come without limitations. Our analysis is based on one geographical region (Philadelphia) and one image data type (exterior images) only. Thus, ablation and replication studies should be performed across cities and image data types like interior images or floorplans to examine the results in other settings. Additionally, it should be examined, why using CBAM did not boost the accuracy of the models. Moreover, advances should be made in the field of visual explainability methods for regression tasks. Currently, only SWH and RAM are available for this application, where both have different pros and cons. One large limitation of RAM is its architectural

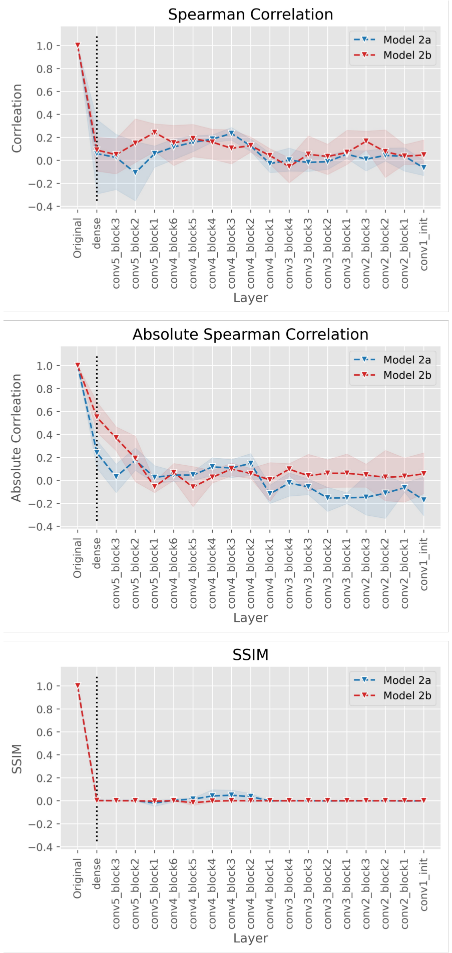

Figure 4. Sanity checks for Model 2a and 2b based on Adebayo et al. [42], repeated 5 times for the selected models. The black dotted line indicates the layer from which the randomization was performed.

constraints, limiting the use for more advanced neural network architectures with multiple fully connected layers after the global average pooling as well as the use for other multi-view learning strategies like multi-input neural networks. Finally, while we used the sanity checks to technically evaluate RAM, it still remains open, whether the different stakeholder groups within real estate appraisal evaluate the saliency maps as helpful and trustworthy. Therefore, future research should perform user-group based experiments. Despite these limitations, implications for research and practice are that by using the right modeling approach in combination with suitable post-hoc interpretability methods, the combination of hard information and soft information in form of images can be made interpretable. All stakeholders within the appraisal process can benefit from the increased explainablility.

References

[1] X. Liu, Q. Xu, J. Yang, J. Thalman, S. Yan, and J. Luo, “Learning Multi-Instance Deep Ranking and Regression Network for Visual House Appraisal,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 8, pp. 1496–1506, 2018.

[2] V. Limsombunchai, “House price prediction: hedonic price model vs. artificial neural network,” in *New Zealand Agricultural and Resource Economics Society Conference*, pp. 25–26, 2004.

[3] S. Law, B. Paige, and C. Russell, “Take a Look Around: Using Street View and Satellite Images to Estimate House Prices,” *ACM Trans. Intell. Syst. Technol.*, vol. 10, nov 2019.

[4] W. McCluskey, W. Deddis, A. Mannis, D. McBurney, and R. Borst, “Interactive application of computer assisted mass appraisal and geographic information systems,” *Journal of Property Valuation and Investment*, vol. 15, no. 5, pp. 448–465, 1997.

[5] A. J. Bency, S. Rallapalli, R. K. Ganti, M. Srivatsa, and B. S. Manjunath, “Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery,” in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 320–329, 2017.

[6] Z. Bessinger and N. Jacobs, “Quantifying curb appeal,” in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 4388–4392, IEEE, 2016.

[7] C. Naumzik and S. Feuerriegel, “One picture is worth a thousand words? the pricing power of images in e-commerce,” in *Proceedings of The Web Conference 2020*, WWW ’20, (New York, NY, USA), p. 3119–3125, Association for Computing Machinery, 2020.

[8] J.-P. Kucklick and O. Müller, “A comparison of multi-view learning strategies for satellite image-based real estate appraisal,” in *The AAAI-21 Workshop on Knowledge Discovery from Unstructured Data in Financial Services*, 2021.

[9] J.-P. Kucklick, J. Müller, D. Beverungen, and O. Müller, “Quantifying the impact of location data in real estate appraisal - a gis-based deep learning approach,” in *ECIS, 2021*, 2021.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[11] W. Samek and K.-R. Müller, *Towards Explainable Artificial Intelligence*, pp. 5–22. Cham: Springer International Publishing, 2019.

[12] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[13] M. E. Kaminski, “The right to explanation, explained,” *Berkeley Tech. LJ*, vol. 34, p. 189, 2019.

[14] M. Du, N. Liu, and X. Hu, “Techniques for Interpretable Machine Learning,” *Commun. ACM*, vol. 63, pp. 68–77, dec 2019.

[15] J. Bin, B. Gardiner, E. Li, and Z. Liu, “Multi-source urban data fusion for property value assessment: A case study in Philadelphia,” *Neurocomputing*, vol. 404, pp. 70–83, 2020.

[16] K. J. Lancaster, "A New Approach to Consumer Theory," *Journal of Political Economy*, vol. 74, no. 2, pp. 132–157, 1966.

[17] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34–55, 1974.

[18] N. Kok, E.-L. Koponen, and C. A. Martínez-Barbosa, "Big Data in Real Estate? From Manual Appraisal to Automated Valuation," *The Journal of Portfolio Management*, vol. 43, no. 6, pp. 202–211, 2017.

[19] J. M. Liberti and M. A. Petersen, "Information: Hard and soft," *Review of Corporate Finance Studies*, vol. 8, no. 1, pp. 1–41, 2019.

[20] E. Elam and A. Stigarll, "Landscape and house appearance impacts on the price of single-family houses," *Journal of Environmental Horticulture*, vol. 30, no. 4, pp. 182–188, 2012.

[21] S. M. Quiring, "Package Your Home to Sell.," *Leaflet/Texas Agricultural Extension Service; no. 2256.*, 1987.

[22] J. Chen, J. S. Evans-Cowley, R. C. Rutherford, and B. W. Stanley, "An empirical analysis of effect of housing curb appeal on sales price of newer houses," *Applied Finance*, vol. 1407, 2013.

[23] O. Poursaeed, T. Matera, and S. Belongie, "Vision-based real estate price estimation," *Machine Vision and Applications*, vol. 29, no. 4, pp. 667–676, 2018.

[24] K. Solovev and N. Pröllochs, "Integrating floor plans into hedonic models for rent price appraisal," in *Proceedings of the Web Conference 2021*, WWW '21, (New York, NY, USA), p. 2838–2847, Association for Computing Machinery, 2021.

[25] J. Bin, B. Gardiner, Z. Liu, and E. Li, "Attention-based multi-modal fusion for improved real estate appraisal: a case study in Los Angeles," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 31163–31184, 2019.

[26] Z. Kostic and A. Jevremovic, "What image features boost housing market predictions?," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1904–1916, 2020.

[27] G. H. Donovan and D. T. Butry, "The effect of urban trees on the rental price of single-family homes in Portland, Oregon," *Urban Forestry & Urban Greening*, vol. 10, no. 3, pp. 163–168, 2011.

[28] J. Ho, "Machine learning for causal inference: An application to air quality impacts on house prices," *arXiv: 1808.02547 v1*, 2016.

[29] E. B. Johnson, A. Tidwell, and S. V. Villupuram, "Valuing Curb Appeal," *The Journal of Real Estate Finance and Economics*, vol. 60, no. 1, pp. 111–133, 2020.

[30] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[31] P. Helber, B. Bischke, Q. Guo, J. Hees, and A. Dengel, "Multi-Scale Machine Learning for the Classification of Building Property Values," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4873–4876, IEEE, 2019.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[34] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[36] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, oct 2017.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 1135–1144, Association for Computing Machinery, 2016.

[38] R. R. Yang, S. Chen, and E. Chou, "AI Blue Book: Vehicle Price Prediction using Visual Features," *arXiv preprint arXiv:1803.11227*, 2018.

[39] Z. Wang and J. Yang, "Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation," *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 514–521, 2018.

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[41] N. Bansal, C. Agarwal, and A. Nguyen, "Sam: The sensitivity of attribution methods to hyperparameters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8673–8683, 2020.

[42] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[43] City of Philadelphia, "Metadata cataloug," 2018. `https://metadata.phila.gov/`.

[44] Microsoft Inc., "Get a static map," 2020. `https://docs.microsoft.com/en-us/bingmaps/rest-services/imagery/get-a-static-map`.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, jun 2016.

[47] S. Smith, "Morning headlines: Cnn money says philly's best neighborhood is mt. airy," *Philadelphia Magazine*, 08 2013.