Finance and Economics Discussion Series

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

Explaining Machine Learning by Bootstrapping Partial Marginal Effects and Shapley Values

Thomas R. Cook, Zach D. Modig, Nathan M. Palmer

2024-075

Please cite this paper as:
Cook, Thomas R., Zach D. Modig, and Nathan M. Palmer (2024). "Explaining Machine Learning by Bootstrapping Partial Marginal Effects and Shapley Values," Finance and Economics Discussion Series 2024-075. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2024.075.

Explaining Machine Learning by Bootstrapping Partial Marginal Effects and Shapley Values

Thomas R. Cook[*][†] Zach Modig[†][‡] Nathan M. Palmer[†][‡]

August 6, 2024

Abstract

Machine learning and artificial intelligence are often described as "black boxes." Traditional linear regression is interpreted through its marginal relationships as captured by regression coefficients. We show that the same marginal relationship can be described rigorously for any machine learning model by calculating the slope of the partial dependence functions, which we call the partial marginal effect (PME). We prove that the PME of OLS is analytically equivalent to the OLS regression coefficient. Bootstrapping provides standard errors and confidence intervals around the point estimates of the PMEs. We apply the PME to a hedonic house pricing example and demonstrate that the PMEs of neural networks, support vector machines, random forests, and gradient boosting models reveal the non-linear relationships discovered by the machine learning models and allow direct comparison between those models and a traditional linear regression. Finally we extend PME to a Shapley value decomposition and explore how it can be used to further explain model outputs.

JEL Classifications: C14, C18, C15, C45, C52

1 Introduction

Machine learning (ML) and artificial intelligence (AI) methods are often regarded as a black box: they may capture useful interactions and nonlinearities in data, but the shape and

[*] Federal Reserve Bank of Kansas City Email: thomas.cook@kc.frb.org

[†] The views expressed in this article are those of the authors and do not necessarily reflect the views of the Federal Reserve Board, the Federal Reserve Bank of Kansas City or the Federal Reserve System.

[‡] Federal Reserve Board of Governors Email: nathan.m.palmer@frb.gov

nature of the relationships are difficult to ascertain. There is a growing appetite to use ML models in finance and economics for purposes ranging from academic study to credit underwriting. Simultaneously, machine learning interpretability is of growing interest to financial regulators. In 2021 five US financial agencies jointly issued a request for information on financial institutions' use of ML and AI that included a specific sub-section discussing ML explainability[1]. Likewise, Brainard (2021) notes that the lack of interpretability is one of the key problems facing the use of ML methods for financial services, and outlines several ways these difficulties are manifest.

Traditional regression models are often interpreted through their marginal effects, both through point estimates and the uncertainty in those point estimates. In a simple linear model the marginal effects are captured in the coefficient parameters, and the point estimates and variances for each coefficient are typically displayed in a regression table. By contrast, an ML model has no simple relationship between deep model parameters and the model's marginal relationships. For example, a deep neural network may have thousands of parameters related to any single marginal relationship.

This paper proposes a solution: directly construct marginal effects for any ML model as the slope of Friedman's (2001) partial dependency function (also know as the partial dependency plot or PDP). Traditionally in the ML literature the PDP is displayed in levels, and only the point estimate of the PDP is calculated. However, it is straightforward to demonstrate that when applied to linear regression, the slope of the PDP directly replicates the regression coefficients, and bootstrapping produces standard errors comparable to the analytical results of OLS. When applied to non-linear ML models, this approach generalizes the concept of the regression coefficient in a model-agnostic way. We refer to the slope of the PDP as the Partial Marginal Effect, or PME. This approach allows direct comparison between a regression coefficient from a linear model and the PME of a non-linear ML model. We further examine Shapley values to extend the logic of the PDP and PME and assess variable importance.

To demonstrate the method we apply it in two hedonic house pricing exercises. We select the hedonic housing problem for a number of reasons. First, the pricing of housing is an important part of the credit-extension decision. Currently, housing appraisal are often conducted by human specialists, but it is not hard to imagine a future world in which part

[1] The five agencies were the OCC, the Federal Reserve Board, the FDIC, the CFPB, an the NCUA. The RFI was titled "Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning," and more details can be found at this link: 86 FR 16837.

of that process involves machine learning. Second, house prices themselves are an important channel of economic activity, particularly during business cycles (see Leamer et al. (2007), Leamer (2015), Glaeser and Sinai (2013), or Piazzesi and Schneider (2016) for an excellent overview). Understanding these channels can directly and indirectly aid policy practitioners. Finally, pricing a house is perhaps the oldest financial situation in which a machine learning method has been applied. We refer, of course, to nearest-neighbor regression: predicting the expected value for an observation as the average of the N most similar observations[2].

1.1 Literature

This paper contributes to the extensive and fast-growing interpretability literature in machine learning. Breiman (2001) provides an introduction to interpretability vs prediction in machine learning. Semenova, Rudin, and Parr (2019) and Molnar (2021) are two modern overviews of ML interpretability that provide a wide survey of the field. This paper in particular extends the partial dependence plot (PDP) described in Friedman (2001)[3]. Section 3 provides similar discussion with respect to Shapley values (Shapley, 1953) and in response to two additional aspects of model interpretation: the effect of feature inclusion and feature importance. Discussion of both the PDP and Shapley values demonstrate their equivalence to parameter estimates in the context of a linear model.

This paper is most closely related to Joseph (2019), in which the author constructs a regression table using a Shapley-Taylor decomposition of an arbitrary model. The current paper differs from Joseph (ibid.) in terms of what is examined; the current paper directly examines the slope of the partial dependency function directly to characterize the non-linear marginal relationships of arbitrary models, as well as extensions of Shapley values that include partial dependence operations, while Joseph (ibid.) employs a Shapley-Taylor decomposition to summarize model properties.

This paper also contributes to the literature on house pricing with machine learning. Machine learning models have been widely explored for house pricing models, however, these studies either focus on the accuracy of the ML model (See Limsombunchai, 2004; McCluskey

[2] For completeness, we include PMEs of nearest-neighbor regression of our main model in the suplemetary appendix.

[3] The PDP is closely related to the "observed-value" approach described in Hanmer and Ozan Kalkan (2013). A related method are the ICE plots developed in Goldstein et al. (2015), which are a visualization of the marginal distribution discussed in Section 2.2, Equation 3. Apley and Zhu (2020) derive an alternative to PDP functions. We discuss the relationship between their approach and PMEs in the online Appendix A.

et al., 2013) or focus on model-specific interpretation of less-complex ML type models (See, for example, Čeh et al., 2018; McMillen and Redfearn, 2010). The current paper examines hedonic house pricing to conduct general inference and describe the underlying non-linear relationships discovered in the micro-level house pricing data.

The illustrative hedonic house pricing model employed is inspired by the meta-analysis in Sirmans, Macpherson, and Zietz (2005) and Zietz, Zietz, and Sirmans (2008), and the data used is described in De Cock (2011).

The rest of this paper is organized as follows: Section 2 outlines how PDP and the corresponding PME can be constructed as a generalization of linear model coefficients. Section 3 provides a similar discussion with regard to Shapley values. Section 4 applies our results to the hedonic house-pricing exercise, Section 5 extends the analysis to changes in preferences driven by the COVID housing boom in Boise Idaho. Section 6 concludes.

2 Model Agnostic Inference via Partial Marginal Effects

There are two common purposes for constructing statistical models that relate a left-hand side (target) variable to a right-hand side (input) variable. In machine learning parlance, this is a supervised learning problem.

The first purpose is *prediction*: given a new observation of inputs, predict the associated target. This is a common use case, and improved predictive performance is an often-cited reason for employing ML and AI models in place of traditional models.[4]

The second purpose is *inference*: rather than predict the target, the emphasis is on describing the world by examining the relationship between the target and the inputs that is captured when the model is fit to the data, as well as describing the statistical properties of that relationship. For example, does a particular input variable have a positive or negative relationship with the target? Is that relationship statistically significant?

Inference is one of economists' primary use cases for statistical models, and an extensive history of statisticians and economists have developed theoretical foundations for inference in econometrics. In the economics literature, a lack of inference tools for ML and AI models reveals itself in slow adoptions rates in the field (though quickly changing)[5]. In the ML and

[4] See chapter 2 of James et al. (2013) for a discussion of prediction and inference in ML and traditional models, and discussion of ML models' improved forecast accuracy versus traditional models.

[5] Although see Athey and Imbens (2019) and Coulombe (2021b) for examples of inference on some ML

AI literatures, the lack of inference tools reveals itself in the rapidly growing literature on explainability and interpretability.

Fortunately there is a promising path forward for AI/ML inference, driven by two observations. First, the marginal relationships between the target and model inputs that are captured by coefficients in a traditional linear regression model can be estimated in a more general way that applies to any model. The approach we focus on in this paper is described in Friedman (2001) as the partial dependency function or "partial dependency plot" (PDP).[6] As described in the following section, the slope of the PDP is exactly the coefficient in a traditional linear regression[7]. This is due to the fact that Friedman (ibid.) constructed the PDP to be a generalization of the *ceteris paribus* reasoning that is taught regarding regression coefficients in introductory statistics courses. We refer to the slope of the PDP as the Partial Marginal Effect (PME) for reasons outlined below. Second, as described in Efron and Hastie (2016), the bootstrap and related methods can provide a straightforward if computationally intense way to calculate variance of a wide range of functions of data. We employ the bootstrap to find the variance in the marginal relationships captured by the PME.

When we apply the PME to a traditional linear regression model and bootstrap to obtain the variance, we replicate the traditional point estimates and variances of the coefficients that one obtains in a standard regression table. When applied to an ML model, we obtain a generalization of the regression table, which allows us to conduct inference on the ML model analogous to inference on a traditional econometric model.

2.1 PME: An Intuitive Discussion

This section uses two analogies to provide an intuitive description of what the PME captures, before turning to the mathematical details. A key insight is that the PME helps an economist understand the properties of a fitted model itself.

For the first analogy, suppose that we have a fitted model. We can think of the PME as providing a summary statistic about the distribution of results for the following experiment:

Take an observation and plug it into the fitted model and get a prediction.

models and applications in economics.

[6] As we will describe in more detail later, there are a number of ways to generalize the marginal relationships that OLS coefficients embody. See Appendix A.

[7] see Appendix B for proofs

> Change nothing else about this observation except for a single characteristic. For example, change square footage for a house, but leave number of rooms, lot size, etc, unchanged. How much does the model output change? Do this with many observations to obtain a distribution of these effects. What is the average of this distribution of these experiments, over the domain of the variable in question?

In this sense, the PME is communicating something about what a fitted model would predict, if it were asked to predict an observation where only one characteristic was changed. We are learning something about the fitted model itself through this process: the distribution of these outcomes over the variable of interest.

Alternatively, we can think of the PME as analogous to conducting a type of field experiment applied to a model that makes predictions. For example, in Bertrand and Mullainathan (2004), the authors submitted a number of resumes to a hiring process and then change a single characteristic of the resumes (the name) to examine how changes in the outcomes (number of call-backs). The PME essentially implements this experiment on a fitted model.

If this sounds like the interpretation of coefficients in multiple regression, that's because it is. Friedman (2001) constructed PDPs to implement and generalize the *ceteris paribus* reasoning for interpreting multiple regression coefficients as taught in most introductory econometrics courses. The main difference is that Friedman (ibid.), and almost all subsequent ML and AI literature, discusses the PDP in terms of the level of the relationship, not the slope of the relationship, which is what traditional multiple regression coefficients capture[8]. Appendix B proves that the PME (the slope of the PDP) is equivalent to the traditional multiple regression coefficients.

2.2 PME: Mathematics and Multiple Regression Coefficients

Consider a typical description of multiple regression coefficients, drawn from Abdi (2004) in the Encyclopedia of Social Sciences Research Methods:

[8] The reason for this is likely that tree-based models (what Friedman (2001) invented the PDP to describe) do not have smooth slopes in their PDP, unlike other methods like support vector machines, deep neural nets, or kernel ridge regressions (SVMs/SVRs, DNNs, KRRs, respectively), and looking at the PDP level is natural for a tree. However even for tree-based methods, approximations of the slope can be examined and provide insights similar to those of smooth methods.

> "[A] regression coefficient… gives the amount by which the dependent variable (DV) increases when one independent variable (IV) is increased by one unit and all the other independent variables are held constant."

The partial dependency function in Friedman (2001) is defined so as to directly capture this reasoning for any fitted model, in levels of model outcome.

This is useful because it means a student who has internalized the intuition of multiple regression coefficients can employ the same reasoning (and the same caveats!) to understand the relationships in data represented by an ML or AI model.

This reasoning is implemented in the mathematics of Friedman's PDP applied to a fitted model $\hat{f}$. Write the fitted model as:

$$\hat{f}(x) = \hat{f}(x^{(k)}, x^{(\neg k)}) \tag{1}$$

where $\hat{f}$ is the fitted model, x is a vector of input variables, and $(x^{(k)}, x^{(\neg k)})$ simply separate out the single input variable $x^{(k)}$ from all other input variables, $x^{(\neg k)}$. For example, in the hedonic house pricing model to be described, $x^{(k)}$ might represent square footage; then $x^{(\neg k)}$ would represent all other variables which are not square footage, such as number of bedrooms, age, neighborhood, etc..

As in Friedman (ibid.), the PDP, denoted ν, of fitted model $\hat{f}$ for an input variable k at a value q is

$$\begin{aligned}\nu_k(q) &= E_{x^{(\neg k)}}[\hat{f}(q, x^{(\neg k)})|q] &(2)\\ &= \int_{x^{(\neg k)}} \hat{f}(q, x^{(\neg k)}) \mathbb{P}(x^{(\neg k)}) dx^{(\neg k)} &(3)\end{aligned}$$

where $\mathbb{P}(x^{(\neg k)})$ in equation 3 is the marginal distribution over the input data. Thus the PDP holds all other independent variables constant by employing the marginal distribution directly and holding it fixed[9]. This is a direct construction of the *ceteris paribus* logic employed in the traditional interpretation of regression coefficients.

[9] The alternative, using the conditional distribution $\mathbb{P}(x^{(\neg k)}|x^{(k)} = q)$, would no longer effectively hold other variables constant. In fact different ways of employing the conditional distribution instead of marginal distribution produces alternatives to PDP; see Section A for further discussion.

Using the fitted model and data, the PDP can be estimated via Monte Carlo as

$$\hat{\nu}_k(q) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(q, x_i^{(\neg k)}) \tag{4}$$

where N is the total number of observations and $\hat{nu}_k(q)$ is calculated for each q in some range of interest over $x^{(k)}$.

That is, for each $q \in \left[x_{low}^{(k)},\ \ x_{high}^{(k)}\right]$, the above mean is taken over the empirical marginal distribution of the fitted values $\hat{f}(q, x_n^{(\neg k)})$, where the $x^{(\neg k)}$ values in the data are literally held constant. The resulting PDP function is in levels of the predicted $\hat{y}$ variable; the slope of this function is the PME which can be obtained analytically or by first difference approximation. Appendix B proves that for a linear model, the PME is equivalent to the traditional multiple regressions coefficient.

This partial dependence function of course is a point estimate. If we want the variance in the estimate, then we can employ the non-parametric (pairs) bootstrap to the entire process as described in Algorithm 1.

Algorithm 1 Bootstrap PDP

```
Preallocate output matrix $Z$ with dimensions $(J \times B)$
Allocate vector $Q$ as a J-length vector of equally spaced values from $\left[x_{low}^{(k)},\ \ x_{high}^{(k)}\right]$
for b in 1 to B: do:
    $X_b, y_b$ =bootstrap($X$)
    Estimate $f_b$ s.t. $\hat{y}_b = f_b(X_b)$
    for j in [1,J]: do
        $q = Q_j$
        $Z_j^{(b)} = f_b(q, X_b^{(\neg k)})$
Return $Z$
```

This is computationally expensive but also an embarrassingly parallel problem.

2.3 Partial Marginal Effect Terminology

The PME is a version of *marginal effects* in economometrics; see eg. Cameron and Trivedi (2005), or Williams (2012) for extensive discussion. In other fields these have different names. For example Gelman, Hill, and Vehtari (2020) refers to the marginal effects as "average predicted comparison," and Hanmer and Ozan Kalkan (2013) uses the terminology of “observed-value approach.”

It is also important to note that the term *marginal* takes on different meanings depending on the context. In economics *marginal* often refers to taking a derivative of a function – hence *marginal utility* is the first derivative of the utility function.

In statistics and other related disciplines, *marginal* can be shorthand for integration over the marginal distribution. The PME is marginal in both senses: as the slope of the PDP it is marginal in the sense of being the first derivative, and the PDP itself is constructed by integrating over the marginal distribution of the data. Appendix A describes alternative approaches to constructing these effects. These alternatives differ from the PDP primarily by using conditional (instead of marginal) distributions.

The PME is also *partial* in the sense that it only describes the effect of a single right-hand side input variable at a time, independently of all others. Importantly, if one has done manual variable transformation on a dataset that mechanically induces perfect dependency between two input variables – for example, if one has added a squared term, or added an interaction – the PME will treat each as if they are independent of one another[10].

3 Shapley Values

While the PME provides useful insights into the marginal effect of a feature – the change in model prediction as $x_i^{(k)}$ increases or decreases – it does not provide as much insight about the more general questions such as whether a feature should be included in the model in the first place, or which features are most important to a model's prediction. We can answer these types of questions with Shapley values (Shapley, 1953; Štrumbelj and Kononenko, 2014). Further we can extend our understanding by combining Shapley values PDP and bootstrapping.

The Shapley value $\psi_k f(x_i^{(k)}) \equiv \psi_k f(x_i^{(k)}, x_i^{(\neg k)})$ is the marginal contribution of a variable, k, to a model's prediction, $f(x_i)$, for a particular observation, i averaged over all possible combinations with its covariates $(x_i^{(\neg k)})$. Write the demeaned PDP of $f(x^{(l)} = q)$ as $\tilde{\nu}(x^{(l)} = q) = E_{x^{(\neg l)}}[f(q, x^{(\neg l)}) - E_x[f(x)]]$, then we can write the Shapley value of $x_i^{(k)}$ as

[10] In such a case, the PME must be adjusted to properly recover the complete marginal effect. An example of how to do this is provided in an online appendix.

$$\psi_k f(x_i) = \frac{1}{K} \sum_{s \subseteq x_i^{(\neg k)}} \binom{K-1}{|s|}^{-1} \big(\tilde{\nu}(x_i^{(k)} \cup s) - \tilde{\nu}(s)\big). \qquad (5)$$

Where s is a subset of covariates at values observed in $x_i^{\neg k}$ and K is the total number of variables. The Shapley decomposition of a model's output, $\Psi(f(x_i)) = \{\psi_1 f(x_i), \ldots \psi_K f(x_i)\}$, is a linear decomposition that can be described as an additive feature attribution (Lundberg and Lee, 2017) of f i.e. $\sum_k \psi_k f(x_i^{(k)}) = f(x_i) - E[f(x)]$. This property of the Shapley decomposition makes interpretation straightforward; $\psi_k f(x_i^{(k)})$ tells us the gain[11] in model output on observation i from including variable k *at its value for observation* i, marginalized over its inclusion with all possible combinations of covariates as observed for observation i.

To better understand the intuition and interpretation of the Shapley value, consider a linear model, $f(x) = xB$. OLS estimates of B provide a summary interpretation of the effect of x on $f(x)$. That is, $\beta_k = \frac{\partial f(x)}{\partial x^{(k)}}$ tells us about the overall effect of $x^{(k)}$ on $f(x)$ regardless of its observed value. By contrast, $\psi_k f(x_i^{(k)})$ incorporates the value of x_i directly. In the linear model case[12], $\psi_k f(x_i^{(k)}) = (x_i^{(k)} - E[x^{(k)}])\beta_k$.

Non-parametric models and machine learning models are more complex than linear models and generally explore various nonlinearities in the data. In these circumstances, the effect of $x_i^{(k)}$ on $f(x_i^{(k)})$ is not necessarily independent of $x_i^{(\neg k)}$. By producing the average contribution of including $x_i^{(k)}$ over all possible combinations of observed covariates in $x_i^{(\neg k)}$, Shapley values resolve the attribution of the effect of $x_i^{(k)}$ on $f(x_i^{(k)})$ in a way that is 'fair'[13]. As a consequence however, $\psi_k f(x_i^{(k)})$ is dependent on the specific covariate profile $x_i^{(\neg k)}$.

We can gain a more general understanding of the effect of including $x^{(k)}$ in the model by marginalizing out the covariate profile $(x^{(\neg k)})$, i.e. by calculating $E_{x^{(\neg k)}}[\psi_k f(x_i^{(k)})|x_i^{(k)}]$. We can estimate this quantity as $SPDP(x) = \nu(\psi_k f(x))$ and we can produce bootstrap estimates of $SPDP(x)$ by applying the following algorithm:

Note that the quantities contained in $\tilde{Z}$ are just a normalized version of the quantities in

[11] This gain is expressed relative to $E[f(x)]$ allowing for comparison between different models.

[12] see Appendix B for discussion

[13] The properties of fairness are discussed in Young (1985). Crucially, for a decomposition to fairly describe feature attribution, it must be accurate (i.e. the sum of feature attributions must sum to the model output) and it must exhibit coalitional monotonicity whereby if $f(x_i^{(k)}) - E[f(x)] \geq g(x_i^{(k)}) - E[f(x)]$ then $\psi_k f(x_i^{(k)}) \geq \psi_k g(x_i^{(k)})$. See discussion in Lundberg and Lee (2017) for application of fairness to feature attributions specifically and in which the authors establish that the only linear decomposition approaches to feature attribution that can be described as fair are those that are derived from Shapley values.

Algorithm 2 Estimate SPDP and SFIPDP

```
Preallocate output matrix $Z$ with dimensions $(J \times B)$
Preallocate output matrix $\tilde{Z}$ with dimensions $(J \times B)$
Allocate vector $Q$ as a J-length vector of equally spaced values from $\left[x_{low}^{(k)},\ x_{high}^{(k)}\right]$
for for b in 1 to B do
    $x_b, y_b =$bootstrap$(x)$
    Estimate $f_b$ s.t. $\hat{y_b} = f_b(x_b)$
    for for j in 1 to J do
        $q = Q_j$
        $Z_j^{(b)} = \frac{1}{N}\sum_i \psi_k f(q, x_{bi}^{(\neg k)})$
        $\tilde{Z}_j^{(b)} = \frac{1}{N}\sum_i \frac{|\psi_k f(q, x_{bi}^{(\neg k)})|}{\sum_{h\in\{k,\neg k\}} |\psi_h f(x_{bi}^{(h)})|}$
Return $Z$ as SPDP, and $\tilde{Z}$ as SFIPDP
```

Z, which gives us the Shapley feature importance partial dependency plot (SFIPDP). The SFIPDP values can be plotted against quantiles of k to give a sense of how much influence k would have at a specific value and how likely it would be to actually observe such a value.

4 Applied Exercise: Ames Housing Data

To concretely illustrate the PDP, PME, SPDP, and SFIPDP methods discussed in sections 2 and 3 we apply them to a hedonic house pricing model run on tax assessor data. The data, described in De Cock (2011), is comprised of houses sold from 2006 through 2010 in Ames, Iowa. The data contains approximately 3000 observations with models spanning roughly 90 different variables, including square footage, number of beds and baths, number of fireplaces, and neighborhood and amenities information.

Our hedonic model measures the response of the log of house prices to the most-often included dependent variables across the two meta-studies on hedonic house pricing, Sirmans, Macpherson, and Zietz (2005) and Zietz, Zietz, and Sirmans (2008).

The estimated model, for both OLS and the ML models, is presented in Table 1. In this discussion, we focus on the results from five models. The simplest model is a linear model estimated via OLS. Two of the models are ensemble models based on decision-trees: random forest (RF) and gradient boosting machine (GBM). These ensemble models are notable in that they essentially learn complex piecewise functions to fit the data. By contrast, the other two models we discuss – support vector machine (SVM) and deep neural networks

Table 1: Model Specification

Target	Input Features	Additional Controls
Log(sale price)	Square Footage	Neighborhood
	Age	Sale Condition
	Lot Area	Central Air
	Garage Area	Condition 1
	Bathrooms	
	Bedrooms	
	Bathrooms:Bedrooms	
	Fireplaces	
	Time Trend	

(DNN) – learn 'smooth' (i.e. continuous) functions that fit the data.

Generally, each of the five models performed well. In-sample and out of sample R^2 scores are presented in Table 2. In sample, the linear model exhibits inferior fit to the other models while the random forest model exhibits near perfect fit[14]. The difference between the linear model and the ML models is less striking with regard to out of sample fit (R^2 measured via 10-fold cross-validation), where the OLS model performs better than the DNN and only slightly worse than other models. In this particular problem, the ML models may not dominate the OLS model in terms of out-of-sample prediction, but as shown below they contribute additional insights regarding non-linearities in the marginal relationships via the PME.

Finally, note that as is often the case, the random forest model obtains very high in-sample fit, while remaining competitive out-of-sample as well. See Coulombe (2021a) for discussion of this property. An implication is that random forests may be particularly useful for inference if we can describe the marginal relationships well.

Table 2: OLS and ML model performance

	OLS	GBR	RF	SVR	DNN
In-sample R^2	0.847	0.897	0.979	0.914	0.887
Out-of-sample R^2	0.837	0.842	0.841	0.839	0.808

[14] This result is consistent with Coulombe (2021a).

4.1 PDP of Linear Model and OLS Estimates

Turning to the OLS model results, we can interpret the effect of each of the variables in Table 1 by examining the OLS coefficient estimates. These are presented in Table 3. The PME-based coefficient estimates are presented in the second column of Table 3 along with bootstrap estimates of the standard error. The PME-based coefficient estimates are equal to the OLS based estimates (to machine precision) and the corresponding estimates of standard errors are only slightly different.

The OLS estimates show effects that are generally consistent with findings documented in the meta analyses in Sirmans, Macpherson, and Zietz (2005) and Zietz, Zietz, and Sirmans (2008). That is, it finds square footage, lot size, number of bathrooms, garage area and number of fireplaces as statistically significant with a positive effect on house price while Age is statistically significant with a negative effect. The time trend is not estimated to be statistically significant, which fits with the mixed findings regarding the time trend in Sirmans, Macpherson, and Zietz (2005). The total effect of bedrooms is estimated as negative as long as the home has at least 1 bathroom, but it is not necessarily always statistically significant. This is likewise consistent with the mixed findings in Sirmans, Macpherson, and Zietz (ibid.). See Appendix C for additional discussion.

Table 3: OLS and PME-derived model estimates

	OLS	PME
Square Footage	0.000235*	0.000235*
	(9.45e-06)	(9.28e-06)
Age	-0.00242*	-0.00242*
	(0.000313)	(0.000308)
Lot Area	3.95e-06*	3.95e-06*
	(1.19e-06)	(1.26e-06)
Bedrooms	0.0142	0.0142
	(0.0108)	(0.00976)
Bathrooms	0.0914*	0.0913*
	(0.0184)	(0.017)
Bed x Bath	-0.015*	-0.015*
	(0.0051)	(0.00448)
Time Trend	-5.56e-05	-5.56e-05
	(0.000203)	(0.000191)
Garage Area	0.000201*	0.000201*
	(2.8e-05)	(2.69e-05)
Fireplaces	0.061*	0.061*
	(0.006)	(0.00642)
Adj. R^2	0.845	0.845
N	2874	2874

Robust standard errors presented in parentheses for OLS model. Bootstrap estimated standard errors presented for Δ PME model.

4.2 ML Models and PMEs

The nonlinear nature of the ML models makes it difficult to present the marginal effects of the ML models in tabular form. Though less concise than Table 3, examining the PME reveals interesting nonlinear relationships learned by the ML models. These relationships often vary considerably and cannot be easily summarized with a single number or concisely described using a table[15].

Consider the PME of the SVM for square footage in Figure 1. Inspecting the PME graphically tells us that the premium on each additional square foot of a home grows rapidly until a home reaches about 2000 feet in size, at which point the SVM estimates a 2.5% increase in home price per each additional 100 square feet. But beyond 2000 square feet, there are diminishing returns to this premium and, per the SVM, the premium effectively drops to zero as square footage exceeds 5000 feet.

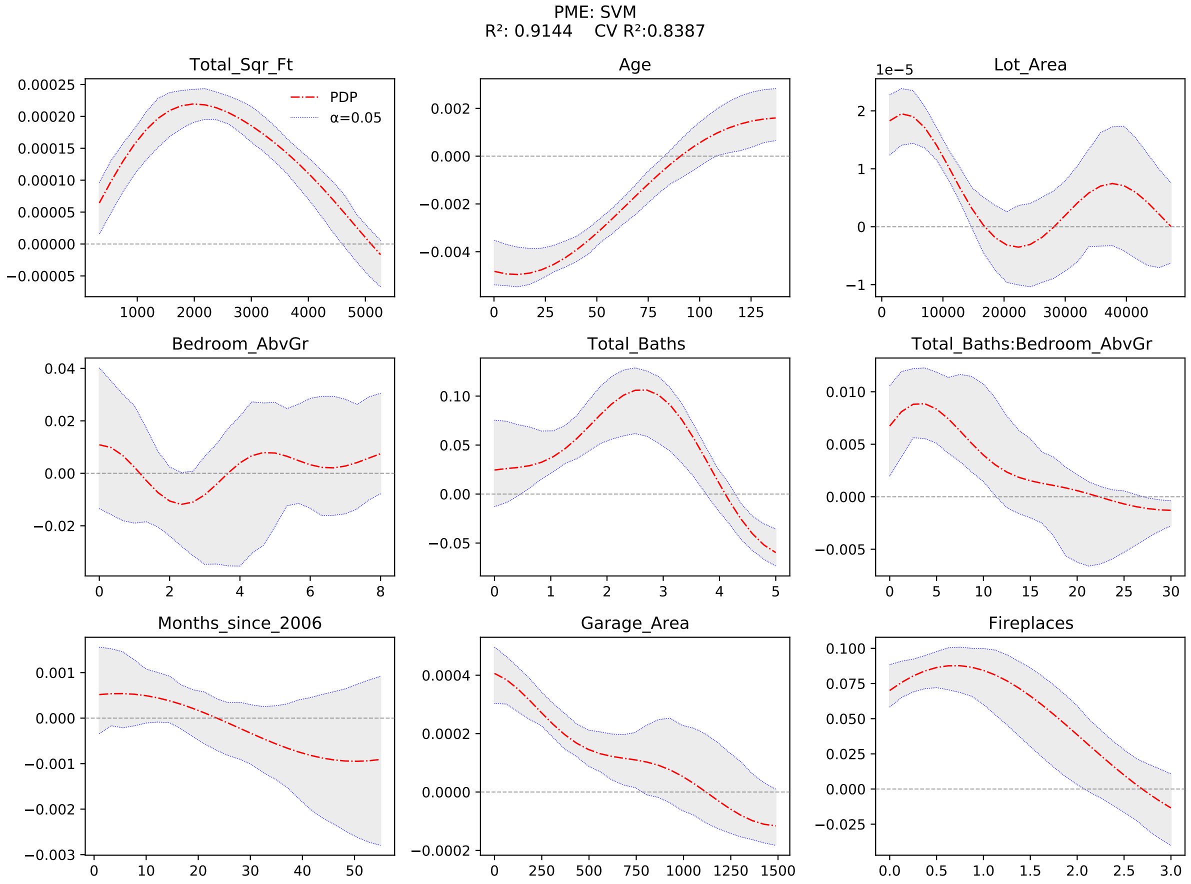
The SVM's PME in Figure 1 also reveal an interesting, nonlinear effect of age on home price. For a new house, increasing its age by one year corresponds with a reduction in predicted price of about 40 basis points. This is true for houses that are 0-25 years old, and then the marginal effect steadily moves towards zero as the age of a house increases. For a house around 80 years old there is approximately no change in predicted price due to a change in age, and past around 110 years a 1-year increase in age is actually predicted increase the value of the house by around 10-20 bps.

Lot area depicts an intuitive pattern: the PME increases between 0 and 5000 square feet before falling to zero near 15000 square feet (about 1/3 of an acre). Beyond 15000 square feet, the bootstrap standard error of the PME render it statistically indistinguishable from zero. The shape of the lot area PME suggests that for small, likely urban lots where space is scarce, lot area commands the largest premium. In more suburban areas, space is less scarce and lot sizes are larger but command less of a premium per square foot. Lot sizes above 1/3 of an may be in more remote portions of the metro and thus command even less of a premium per square foot.

As in the OLS model, the PME for bedrooms in location (2,1) in Figure 1 is largely indistinguisable from zero over essentially the entire range, and the same is true for the time trend in location (3,1). Baths and the beds*baths interaction both are significant for

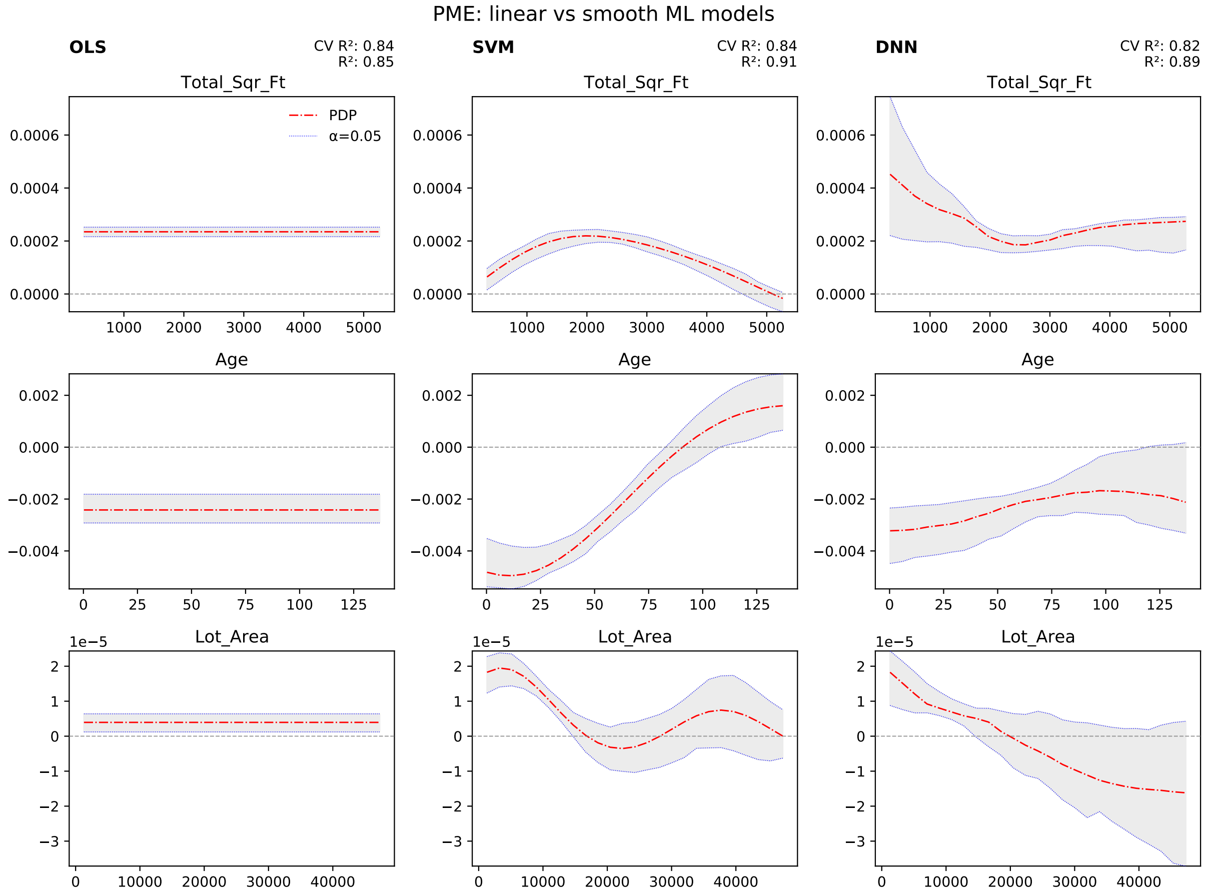
[15] It might be possible to accomplish this for a pre-specified nonlinearity but the advantage of many ML models is precisely that they will find nonlinear relationships without ex-ante specification.

Figure 1: SVM Partial Marginal Effects



much but not all of their range, and garage area and fireplaces both depict a similar trend – the marginal effect is significant and positive for low values, with the PME dropping until it is indistinguishable from zero by the end of the range.

Figure 2: PME for smooth models (SVM, DNN)



PMEs can be used to compare across multiple models. Figure 2 compares the OLS model with two ML models that have smooth marginals: the support vector machine (SVM) and a deep neural network (DNN). These figures show distinct non-linear effects. Notably, we see diminishing effects for a number of variables as they approach values far outside of their mean.

Square footage in the first row displays the most stylistic difference across models, although it is positive and significant over nearly the entire range in all models.[16] For age,

[16] Many machine learning methods are universal function approximators, and discover non-linearities and interactions endogenously. If multiple ML methods do not converge on the same approximate form for a given estimation task this may imply that there is not yet enough data.

both ML models show that younger houses experience a greater negative marginal price imapct versus older houses. [17]. For lot area, both ML models display a similar path for the marginal effect: houses with small lots expect the greatest price increase due to a lot size increase, and this effect steadily decreases until about 15,000 square feet, or approximately 1/3 acre.

The PMEs are most useful for models with smooth marginal functions, such as the SVM and DNN depicted already. For models with non-smooth marginals, such as tree-based models, the PDP in levels[18] are likely a better tool for model interpretation.

Figures 3 and 5 depict the PMEs and PDPs, respectively, for OLS versus two tree-based models: a gradient-boosted machine (GBM) and a random forest (RF). These tree methods do not have inherently smooth marginal effects under the basic tree structure. Accordingly, the PME plot is quite volatile; non-zero effects are only observed in the differences between points that straddle 'jumps' in the piecewise model function. The PME can be made more legible by evaluating it on fewer points over the range of the input variable, but this comes with the risk of imprecision in interpretation. For tree based models it may be preferable to interpret the PDP directly, although this is necessarily more qualitative. Future work will explore trees with natural non-zero slopes and explore additional solutions for tree-based models.

The broad patterns seen in the SVM and DNN models in Figure 4 are reflected in the GBM and RF in Figure 5. Increasing square footage is associated with increased price, with tight 5% confidence bounds over much of the range. For homes aged 0 to about 35 years old, any increase in age is associated with a steep drop in price, but after that, an increase in age is only associated with a modest drop in price, often indistiguishable from zero. For houses with a lot size of about 0 to about 12,000 square feet (a little over 1/4 acre), an increase in lot size corresponds with an increase in predicted price; otherwise lot size increases don't appear to increase expected price (seen most clearly in the RF).

[17] This nonlinearity in the effect of age on house price is well known; see Goodman and Thibodeau (1995). This discussion highlights (1) that the ML models account for the nonlinearity without prior specification and (2) that PMEs allow us to depict the nonlinear relationship that the ML models have learned.

[18] That is, the effect function in log(price).

Figure 3: PME for non-smooth models (GBM, RF)

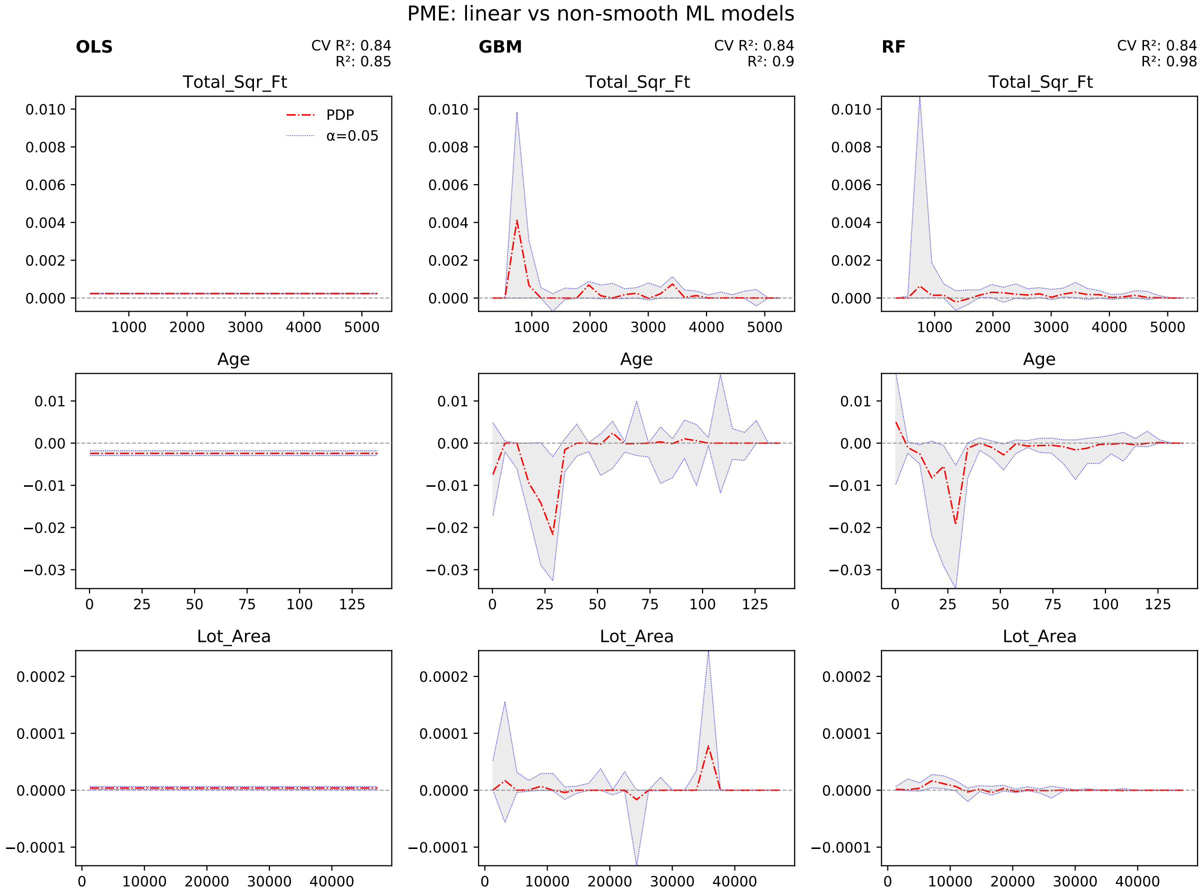


Figure 4: PDP of linear model and smooth ML models

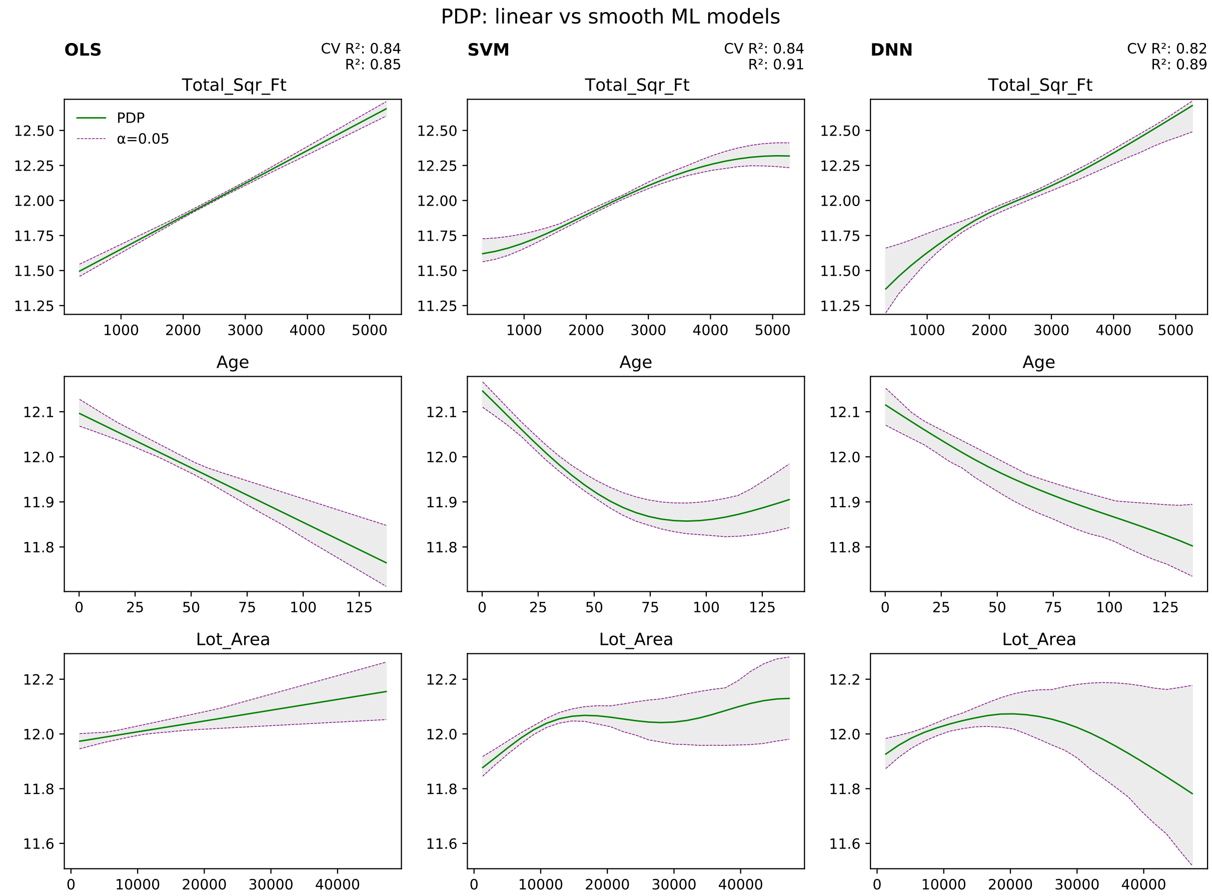
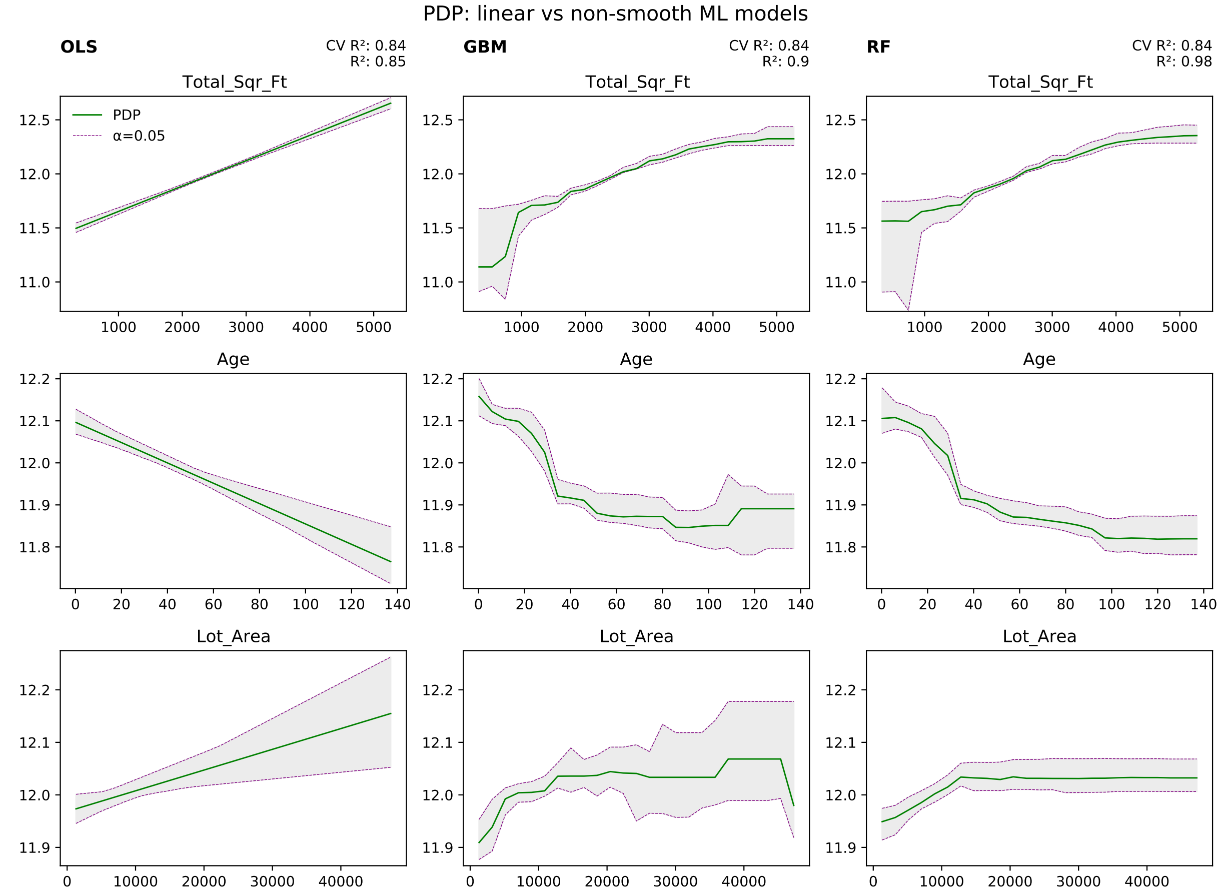


Figure 5: PDP of linear model and non-smooth ML models



4.3 Feature importance and marginal effect of inclusion via SPDP and SFIPDP

Recall that SPDP and SFIPDP are complimentary measures to PME. In general, the PME values can be interpreted as coefficient estimates in the spirit of statistical regression pedagogy. That is, we can think of the PME as telling us about the effect of a variable as it increases or decreases and taking as a given that the variable will be included in the model. The SPDP is subtly different in that it tells us about the expected effect of including a variable in a model at a given value. The SFIPDP is more distinct from the PME and can be thought of as a measure of feature importance. As such, the SFIPDP enables us to compare input features to one another in terms of their overall impact on model output.

Figures 6 and 7 plot SPDP and SFIPDP results for Total Square Feet, Year Built, and Lot Area as constructed for OLS, SVR, and GBM modelling approaches. As discussed in Section 3, the SPDP for the linear model is linear with a slope equal to the estimated OLS coefficients.

For the SVM model, the nonlinearities are less pronounced for some variables (square footage and age), but are more readily apparent for others (lot area). It is also notable that for age, the SPDP of the SVM model is nearly flat and close to zero. This suggests that the inclusion of age has a relatively limited impact on the model output. The corresponding panel from Figure 7 supports this, showing that, on average, the inclusion of age in the SVM model accounts for between 5% and 10% of the model output[19].

For the GBM model, the SPDPs follow the same general direction of the linear model, but are nonlinear, reflecting the nonlinearities captured by these models. Unlike the SVM model, the SPDP of age for the GBM model in figure 6 is quite sizable. Here, the the inclusion of the age of a home variable will, on average, reduce the estimated price by about 20% for houses more than 80 years old. For newer houses, the inclusion of the age variable increases the estimated home price. For the newest homes, including the age variable causes the model to attach a 15% premium to the estimated home price. This premium declines rapidly with the age of the home until age 30, whereupon the inclusion of age begins to act as a drag on the estimated home price. The drag from including age grows more slowly with age after age 35. In terms of relative importance, the GBM model appears to place more

[19] It is additionally notable that both the SPDP and SFIPDP for the SVM exhibit rather wide bootstrap confidence intervals (for age and other variables, when compared to OLS and GBM models). This suggests sensitivity to the composition of the training data that might otherwise be corrected through more careful hyperparameter tuning.

weight on the age variable than the SVM model. This is highlighted in Figure 7 where we can see the SFIPDP of the age variable for the GBM model accounts for as much as 50% of the model output[20]

[20] There is a notable decline in the SFIPDP as the age variable approaches the average age of homes in the dataset. As discussed in Section 3, this is expected behavior for the SFIPDP.

Figure 6: SPDP for Linear, SVM and GBM models for selected variables

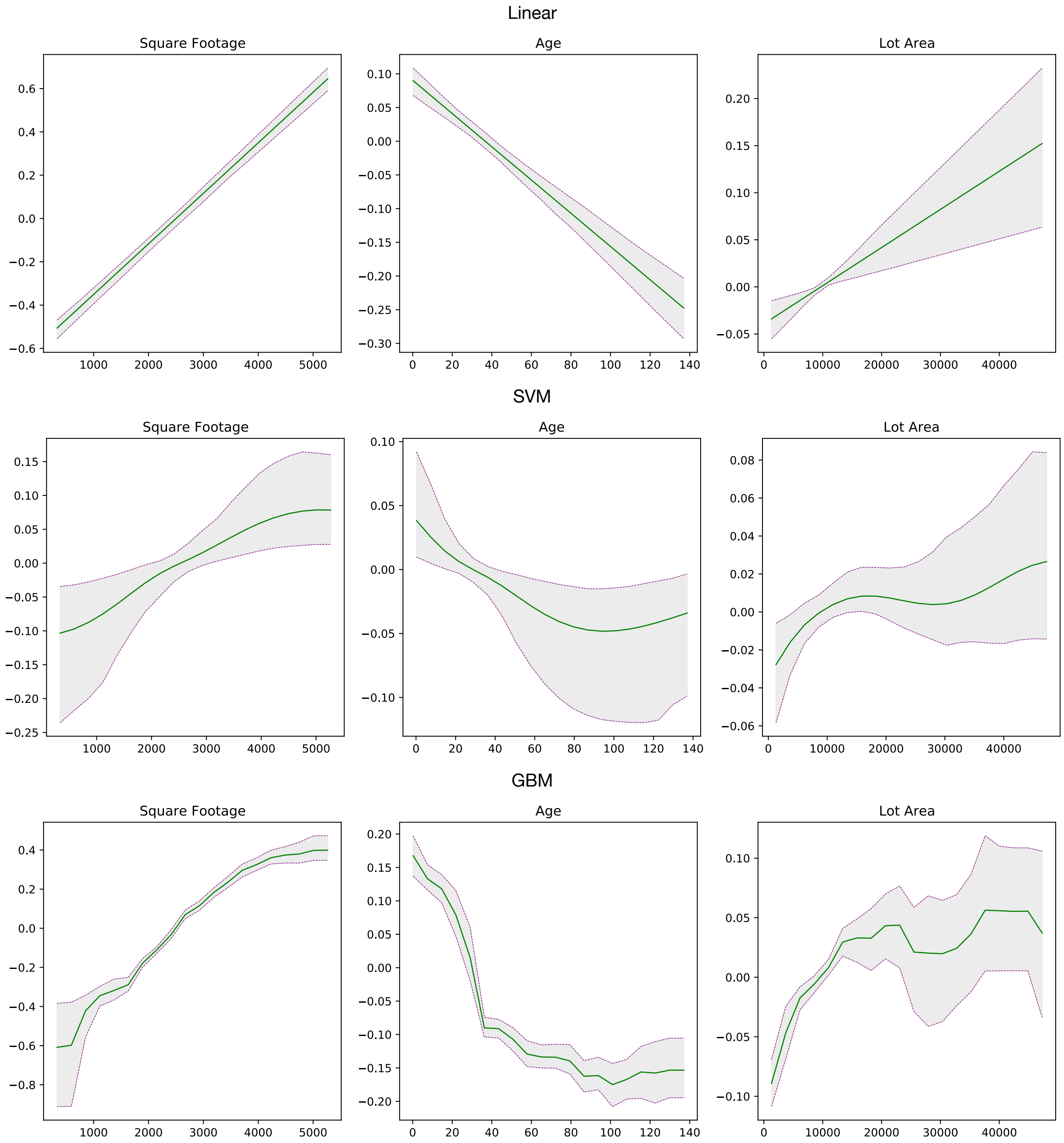
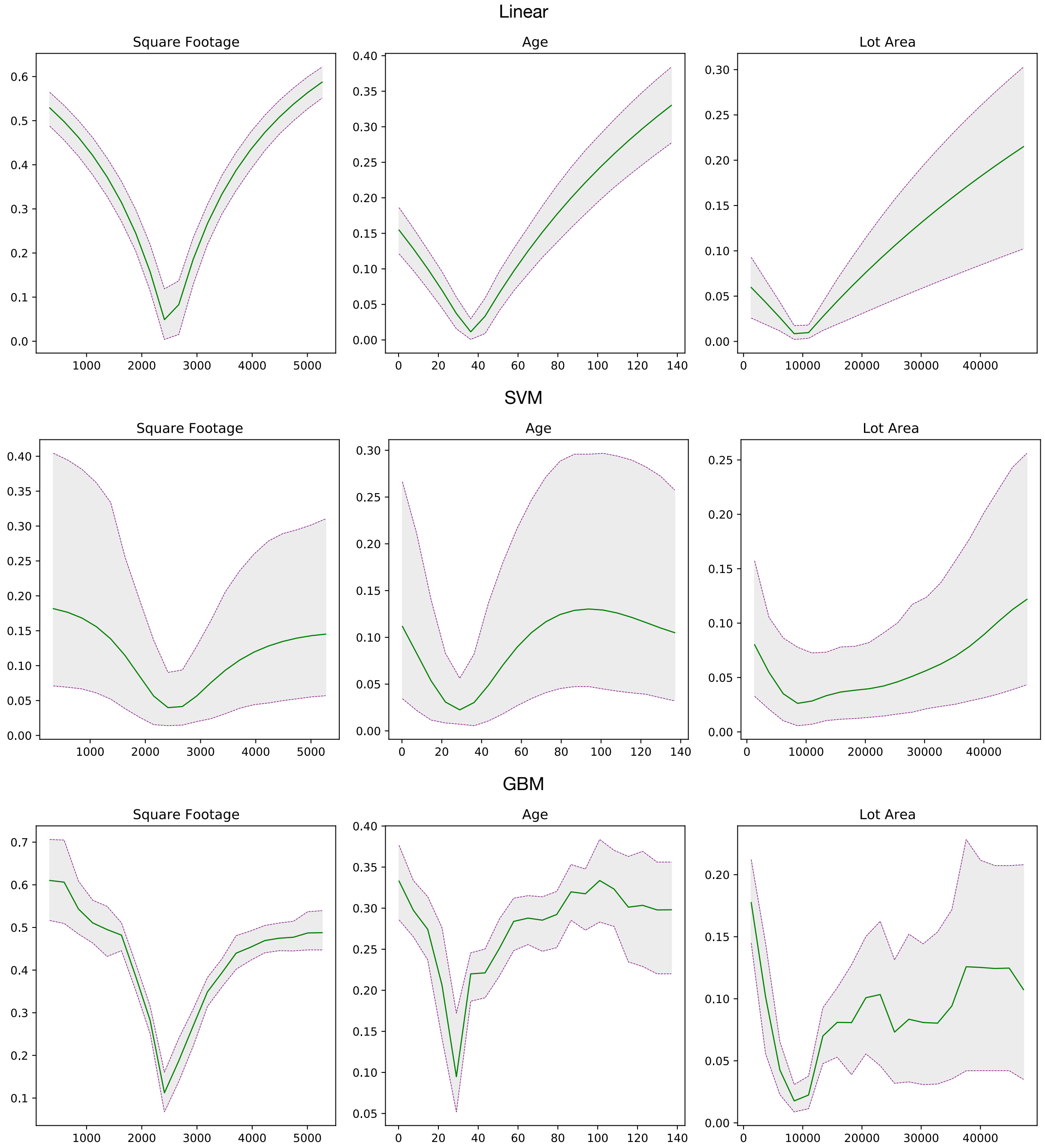


Figure 7: SFIPDP for Linear, SVM and GBM models for selected variables



5 Changes in preferences over time in Boise Idaho

This project was started just prior to the COVID19 pandemic. In the time that has followed, changes in work patterns, namely the rising prevelance of remote work has led to a migration of people away from high-cost of living areas and into smaller cities and towns. Of particular note San Francisco has seen a decline in population by as much as 6% from 2020 to 2021[21]. One of the more popular destinations for those leaving the Bay area has been Boise Idaho. How has the influx of migration from wealthy, high cost of living areas into these areas influenced house prices? To answer this, we extend the DNN and GBM models developed for the previous exercise to predict to house sales in the Boise metro area. By comparing models fit on 2019 home sales data with models fit on 2021 home sales data, we can see how home preferences have changed in response to the COVID-driven housing boom.

The data we use for this analysis comes from the Multiple Listing Service (MLS) dataset from the CoreLogic Real Estate database. We filter the data to include only home purchases in Ada and Canyon counties[22] in 2019 and 2021. The variables Generally, we include the same variables discussed in the previous section[23].

We focus here on two types of models: GBM and DNN. For each model type, we produce separate model fits on sample data from 2019 and again on sample data from 2021. This yields four total models: DNN-2019, DNN-2021, GBM-2019, and GBM-2021. We also fit linear regression models (linear-2019, linear-2021) to use as a performance baseline. Performance metrics are provided in Table 4. They show that the nonlinear, ML models better predict home pricing for both the 2019 and 2021 data, suggesting that the ML methods capture important nonlinearities in the data.

Generating the PME and SPDP from the fitted models allows us to understand how models changed between 2019 and 2021. The DNN models tend to produce smooth PME and SPDP plots. They are less smooth for GBM and somewhat more difficult to interpret[24]. For most variables, the PME and SPDP are similar for 2019 and 2021 models. There are, however, two notable exceptions that we highlight here: year built and lot size.

[21] Toukabri, Amel and Crystal Delbe. 2022. "New Data Reveal Most Populous Cities Experienced Some of the Largest Decreases." *America Counts: Stories Behind the Numbers. May 26,2022. Census.*

[22] This captures, primarily, the Boise-Nampa metro, though it also includes rural/exurban areas surrounding the metro.

[23] The primary differences are that we replace the neighborhood and "condition 1" – a measure of proximity to various neighborhood ammenities – with a school district fixed effect.

[24] As discussed in Section 4.2, this is due to the fact that GBM models are based on decision trees.

Table 4: Linear and ML Model Performance on Boise Idaho Data

	2019			2021		
	Linear	DNN	GBM	Linear	DNN	GBM
RMSE	7.25	6.79	5.96	11.5	10.8	10.3
MAE	4.29	3.83	3.31	7.04	6.68	5.78
R^2	0.747	0.751	0.828	0.718	0.786	0.792

RMSE and MAE shown in $10,000's. All metrics measure out of sample performance.

Figure 8 (left) compares the PME for GBM-2019 and GBM-2021 for lot size. The PMEs diverge at various points, but most dramatically at values below 0.25 acres (close to the size of a typical suburban lot), where the 2021 effects are positive and the 2019 effects are 0 or negative. Thus, compared to 2019, the GBM-2021 model suggests increased competition in suburban and urban parts of the metro and increased willingness of buyers to pay a premium for larger lots in desirable urban and suburban areas near urban centers.

The SFIPDP provides some additional insight (Figure 8, right). Broadly, GBM-2021 places a considerably higher importance on lot sizes when compared to GBM2019 and the difference in importance grows[25]with distance from the mean lot size (0.25 acres). This suggests that, at least from the perspective of the GBM models, lot size rose (broadly) in priority for home pricing.

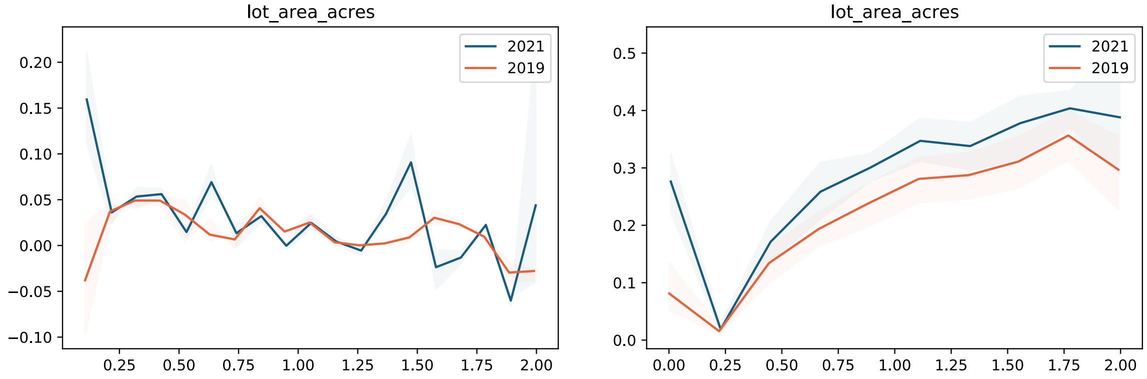
The effect of a home's year of construction (i.e. age) also seemed to change notably between 2019 and 2021. For older homes, such as those built between 1880 and 1975, the effect of a home's age was similar between 2019 and 2021. However, for more recent homes, our models suggest that newer homes commanded a lower premium than they would have in 2019. Figure 9 shows the PME of the DNN model for the year built variable. Note that for much of the figure, the PME for the 2021 model and 2019 model overlap and are linear, suggesting an effect on home price that decreases quadratically in age[26]. For homes built after about 1980, the PME for the 2019 model and 2021 model diverge. The PME for the 2019 model continues along its previous trajectory, suggesting that newer homes command a premium that increases quadratically with year of construction through 2019. In contrast,the PME for the 2021 model flattens at around 0.025, indicating a diminished

[25] This also notably diverges from the SFIPDP for the DNN models. The SFIPDP for GBM-2019, DNN-2019, and DNN-2021 are all fairly similar.

[26] The PME for the GBM model is provided in the appendix but bears similar (though more noisy) interpretation

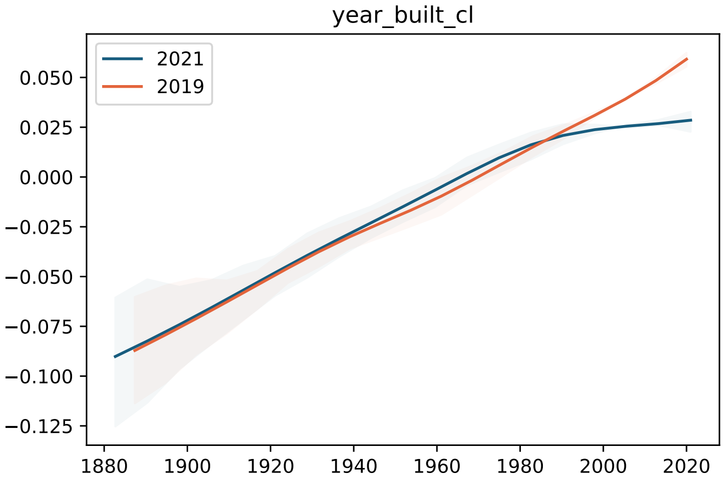
premium on newer homes as long as they were built relatively recently (within the last 40 years). It is not yet obvious why the premium on newer homes would taper in 2021. One potential explaination might be that more homebuyers in 2021 intended on renovating their newly-purchased homes. If this were the case, it might limit the premium paid for newer home features (since they would presumably be removed in the remodel).

Figure 8: The effect of lot size grew in 2021, especially for suburban sized lots



Left: PME Right: SFIPDP

Figure 9: The importance of year of development changed in 2021



PME

6 Conclusion

This paper has explored the use of model-agnostic tools to aid in the interpretation of machine learning models analogous to our interpretation of standard econometric models. It identifies the partial dependence function (PDP) of Friedman (2001) as directly implementing the paribus reasoning of multiple regression coefficients in a model agnostic way. It extends this framework by demonstrating that the slope of the PDP, the partial marginal effect (PME), reproduces the coefficients of a traditional linear regression. Bootstrapping the PME allows for the construction of a 'visual regression table' analogous to a standard regression table, displaying both point estimates and the variance in the point estimates of the PME. This approach allows traditional econometric inference to be conducted with non-linear ML models.

Furthermore, this paper demonstrates that SPDP likewise replicates the OLS coefficient estimates in the linear case and that, for the PDP in particular, the nonparametric bootstrap produces variances in those point estimates comparable to OLS standard errors. By understanding these tools in the linear context, we expect that researchers will feel more comfortable using them to interpreting ML models in the future. The hedonic house-pricing exercise demonstrates such an application and shows that ML models can both replicate the results of meta-analysis of the literature, add new insights about non-linear behavior of variables in the house-pricing models and help us understand changes in preferences over time.

Despite the promise of PMEs, Shapley Values and the SPDP, their use in interpreting 'black-box' models is not without caveats. Perhaps most crucially, these techniques require nuanced interpretation where interactions between variables are concerned; the ceterus paribus assumption cannot be forgotten (just as with traditional regression coefficients). Indeed a crucial avenue for further development will be the extension of these techniques to more fully illustrate interactions captured by a model.

References

Abdi, Hervé (2004). "Partial regression coefficients". *Encyclopedia of social sciences research methods*, pp. 1–4.

Apley, Daniel W and Jingyu Zhu (2020). “Visualizing the effects of predictor variables in black box supervised learning models”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.4, pp. 1059–1086.

Athey, Susan and Guido W Imbens (2019). “Machine learning methods that economists should know about”. *Annual Review of Economics* 11, pp. 685–725.

Bertrand, Marianne and Sendhil Mullainathan (2004). “Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination”. *American economic review* 94.4, pp. 991–1013.

Brainard, Lael (Jan. 2021). *Supporting Responsible Use of AI and Equitable Outcomes in Financial Services*. Speech at the AI Academic Symposium hosted by the Board of Governors of the Federal Reserve System, Washington, DC (Virtual Event), January 12, 2021. Board of Governors of the Federal Reserve System (US).

Breiman, Leo (2001). “Statistical modeling: The two cultures”. *Statistical science* 16.3, pp. 199–231.

Cameron, A Colin and Pravin K Trivedi (2005). *Microeconometrics: methods and applications*. Cambridge university press.

Čeh, Marjan et al. (2018). “Estimating the performance of random forest versus multiple regression for predicting prices of the apartments”. *ISPRS international journal of geo-information* 7.5, p. 168.

Coulombe, Philippe G. (2021a). “To Bag is to Prune”.

Coulombe, Philippe Goulet (2021b). *The Macroeconomy as a Random Forest*. arXiv: `2006.12724 [econ.EM]`.

De Cock, Dean (2011). “Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project”. *Journal of Statistics Education* 19.3.

Efron, Bradley and Trevor Hastie (2016). *Computer age statistical inference*. Vol. 5. Cambridge University Press.

Friedman, Jerome H (2001). “Greedy function approximation: a gradient boosting machine”. *Annals of statistics*, pp. 1189–1232.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020). *Regression and other stories*. Cambridge University Press.

Glaeser, Edward L and Todd Sinai (2013). *Housing and the financial crisis*. University of Chicago Press.

Goldstein, Alex et al. (2015). “Peeking inside the black box”. *Journal of Computational and Graphical Statistics* 24.1, pp. 44–65.

Goodman, Allen C and Thomas G Thibodeau (1995). "Age-related heteroskedasticity in hedonic house price equations". *Journal of Housing Research*, pp. 25–42.

Hanmer, Michael J and Kerem Ozan Kalkan (2013). "Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models". *American Journal of Political Science* 57.1, pp. 263–277.

James, Gareth et al. (2013). *An introduction to statistical learning.* Springer Science & Business Media.

Joseph, Andreas (2019). "Parametric inference with universal function approximators".

Leamer, Edward E (2015). "Housing really is the business cycle: what survives the lessons of 2008–09?" *Journal of Money, Credit and Banking* 47.S1, pp. 43–50.

Leamer, Edward E et al. (2007). "Housing is the business cycle". In: *Proceedings-Economic Policy Symposium-Jackson Hole.* Federal Reserve Bank of Kansas City, pp. 149–233.

Limsombunchai, Visit (2004). "House price prediction: hedonic price model vs. artificial neural network". In: *New Zealand agricultural and resource economics society conference*, pp. 25–26.

Lundberg, Scott and Su-In Lee (2017). "A unified approach to interpreting model predictions". *CoRR* abs/1705.07874. arXiv: `1705.07874`. URL: `http://arxiv.org/abs/1705.07874`.

McCluskey, William J et al. (2013). "Prediction accuracy in mass appraisal: a comparison of modern approaches". *Journal of Property Research* 30.4, pp. 239–265.

McMillen, Daniel P and Christian L Redfearn (2010). "Estimation and hypothesis testing for nonparametric hedonic house price functions". *Journal of Regional Science* 50.3, pp. 712–733.

Molnar, Christoph (2021). *Interpretable machine learning.* mimeo.

Piazzesi, Monika and Martin Schneider (2016). "Housing and macroeconomics". *Handbook of macroeconomics* 2, pp. 1547–1640.

Semenova, Lesia, Cynthia Rudin, and Ronald Parr (2019). "A study in Rashomon curves and volumes". *arXiv preprint 1908.01755.*

Shapley, L. S. (1953). "Stochastic Games". *Proceedings of the National Academy of Sciences* 39.10, pp. 1095–1100. ISSN: 0027-8424. DOI: `10.1073/pnas.39.10.1095`. eprint: `https://www.pnas.org/content/39/10/1095.full.pdf`. URL: `https://www.pnas.org/content/39/10/1095`.

Sirmans, Stacy, David Macpherson, and Emily Zietz (2005). "The composition of hedonic pricing models". *Journal of real estate literature* 13.1, pp. 1–44.

Štrumbelj, Erik and Igor Kononenko (2014). "Explaining prediction models and individual predictions with feature contributions". *Knowledge and information systems* 41.3, pp. 647–665.

Williams, Richard (2012). "Using the margins command to estimate and interpret adjusted predictions and marginal effects". *The Stata Journal* 12.2, pp. 308–331.

Young, H Peyton (1985). "Monotonic solutions of cooperative games". *International Journal of Game Theory* 14.2, pp. 65–72.

Zietz, Joachim, Emily Norman Zietz, and G Stacy Sirmans (2008). "Determinants of house prices: a quantile regression approach". *The Journal of Real Estate Finance and Economics* 37.4, pp. 317–333.

A Additional Generalizations of Marginal Relationships

The structure of the PDP and PME calculations, as well as the ALE calculattions discussed in the online appendix, suggest that there are in fact a number of estimators of the slopes of the marginal relationship of interest that may have different advantages either theoretically or in practice. We suggest a few here but otherwise simply note these for future work.

For example, one potential extension of ALE, not pursued in our current paper, is to apply the "derivate first, then integrate" reasoning of ALE to the conditional PDP described earlier. This would combine the "full information" benefit of PDP with the "local information" benefit of ALE, and may produce a smoother, more intuitive descriptive statistic. We suggest "conditional local effects" (CLE) as a descriptive name. Under such a setup the ICE lines and $\hat{y}$ points would still be useful and appropriate to display to potentially further aiding interpretation.

A simpler extension would be to use the "derivate first, then integrate" reasoning of ALE with the marginal distribution of PDP, resulting in something like a "partial local effects" (PLE) model which retains the ceteris paribus reasoning of PDP as well as some of the controls for correlations between variables that Apley and Zhu (2020) note is an advantage of the "derivate first" approach of ALE.

In fact, we can describe several different estimators for the marginal effect as a profile of three distinct choices:

1. Order of operation of calculating expected slope: (1) integrate then derivate: $\frac{\partial \mathbb{E}[\hat{f}(.)]}{\partial x_i^{(k)}}$, vs (2) derivate then integrate: $\mathbb{E}\left[\frac{\partial \hat{f}(.)}{\partial x_i^{(k)}}\right]$
2. Using the (1) marginal distribution vs (2) conditional distribution for the expectation (using $\frac{1}{N}$ vs conditional kernel-density-based weights for the expectation over ICE lines vs slopes), and
3. Using (1) global or (2) local data to form the expectation (eg. kernel density vs local sample).

PDP is {1, 1, 1} while ALE is {2, 2, 2}. The cPDP suggested above is {1, 2, 1}, m-plots as described in Apley and Zhu (ibid.) (not discussed here) are {1, 2, 2}, the "conditional local effects" (CLE) suggested above would be {2, 2, 1}, and the "partial local effects"

(PLE) would be {2, 1, 1}.[27] Other permutations are possible but may not yield much practical advantage. There are likely tradeoffs between how quickly these converge under the bootstrap, how variable they are, and how efficiently these can be calculated taking advantage of computational tricks (for example, using kernel density estimates to accelerate calculation).

Future research should examine the speed, variance, and convergence properties of the different estimators of marginal slope described above. In particular, it is important to understand how quickly bootstrapping converges for the different methods listed here. The method that converges with the nicest properties (for example, in the higher-order moments of the bootstrapped distributions) will allow for the most efficient bootstrapping estimate of of confidence intervals around the marginal slope relationships.

B Claims about PME, SPDP in linear models

Denote the PDP of a function f as $\nu_k(f(x^{(k)} = q))$ or $\nu_k(q)$ for short. And let $\nu_k(q) = E_{x^{(\neg k)}}[f(q, x^{(\neg k)})|q] = \int_{x^{(\neg k)}} f(q, x^{(\neg k)})\mathbb{P}(x^{(\neg k)})dx^{(\neg k)}$.

Claim 1: $\nu_k(q) = q\beta_k + E(x^{(\neg k)})\beta_{\neg k}$ if f is linear and $\partial \nu_k/\partial q = \beta_k$.

Proof: If f is linear, it can be written as $XB = \sum_j x_j\beta_j$. Accordingly, we can write $f(q)$ as $q\beta_k + x^{(\neg k)}B_{\neg k}$ and

$$\begin{aligned}\nu_k(q) &= E[q\beta_k + x^{(\neg k)}B_{\neg k}|q] \\ &= q\beta_k + E[x^{(\neg k)}]B_{\neg k}\end{aligned}$$

From this it follows that $\partial \nu_k/\partial q = \beta_k$ and thus the slope of the PDP, the PME, is equal to β ■

Write the de-meaned PDF as $\tilde{\nu}_k = E_{x^{(\neg k)}}[f(q, x^{(\neg k)}) - E_x[f(x)]|q]$

Claim 2: If f is linear, then $\partial \tilde{\nu}_k(q)/\partial q = \beta_k$

[27] Note that if this approach does not employ the quantiles-based slope approximation of ALE, it will simply replicate the PME as the derivation and integration will be interchangable in order.

Proof: If f is linear, then we can write

$$\begin{aligned}\tilde{\nu}_k(q) &= q\beta_k + E_{x^{(\neg k)}}[x^{(\neg k)}B_{\neg k}] - E_{x^{(k)}}[x^{(k)}]\beta_k - E_{x^{(\neg k)}}[x^{(\neg k)}]B_{\neg k} \\ &= (q - E_{x^{(k)}}[x^{(k)}])\beta_k.\end{aligned}$$

From this it follows $\partial\tilde{\nu}_k(q)/\partial q = \beta_k$ ■

Claim 3: If f is linear, then $\psi_k f(q) = (q - E[x_k])\beta_k$

Proof: Recall that we write the shapley value as

$$\psi_k f(q) = \frac{1}{K}\sum_{s\subset S(x)\backslash k}\binom{K}{|s|}^{-1}\tilde{\nu}_k(q\cup s) - \tilde{\nu}_k(s)$$

From the proof of Claim 2, it follows that if f is linear,

$$\begin{aligned}\tilde{\nu}_k(q\cup s) - \tilde{\nu}_k(s) &= (q - E[x^{(k)}])\beta_k - (s - E[x_s])\beta_s - (s - E[x_s])\beta_s \\ &= (q - E_{x^{(k)}}[x^{(k)}])\beta_k \ \forall s \subset S(x)\end{aligned}$$

Further, denote $S(x,i)$ as the set of all subsets of covariates in x of size i, then we can write

$$\begin{aligned}\psi_k f(q) = \frac{1}{K}\sum_{i}^{K}(E_{s\subset S(x,i)\backslash k}[q] - E[x_k])\beta_k \\ = (q - E[x_k])\beta_k ■\end{aligned}$$

C Stylized Facts

This appendix presents an excerpt of Exhibit 1 from Zietz, Zietz, and Sirmans (2008), and extends it to illustrate the stylized facts discussed in Section 4.1.

Column (1) is the name of the variable that is closest to the similar variable in the De Cock (2011) data. Columns (2)-(5) are drawn from Exhibit 1 in Zietz, Zietz, and

Sirmans (2008), and are, respectively, (2) the number of times the variables appear across the hedonic house pricing examples considered by Zietz, Zietz, and Sirmans (ibid.), then (3) the number of times the variable has a positive significant effect, (4) the number of times it has a negative significant effect, and (5) the number of times it is not significant in the studies considered.

Then columns (6)-(8) are constructed from columns (2)-(5) to indentify variables for which there are interesting stylized facts. Column (6) is simply the fraction represented by columns (3) divided by (4). When these values are close to zero or very large, then variables have a consistently positve or netagive effect across studies. When they are close to 1, then the relevant variable shows up has having opposite effects with equal frequencies, and we consider values close to the range 0.5-2 to be interesting in this regard: Bedrooms, Distance, and Time Trend.

Column (7) is the ratio: column (5) / (column (3) + column (4)), the ration of number of times not significant to significant. As with column (6), we look for variables with values close to the 0.5-2 range and find: Bedrooms, Distance, and the Time Trend.

Column (8), the final column, combines all three counts of significance for a unified measure of disagreement in results. It is a ratio of: the lesser of (column (3), column(4)) / (the greater of column (3) or (4) + column (5)). Thus it is (non-signicant results + the lesser count of negative or positive signicant results) / (the greater count of negative or positive results). The goal is to identify whether there are times in which one direction of results is close to equal to the number of times that results are not significant, or in the opposite direction. As before the interesting variables are close to 1. We identify the following variables that are close to the 0.5-2 range: Bedrooms, Time Trend, Distance, and there are a few additional values as well: total # Rooms and Fireplaces.

D Boise Analysis, additional discussion

In some ways the differing effects of age between the 2019 and 2021 appears more dramatic when considering the GBM model. Though more noisy than the DNN model, looking at the demeaned PME in figure 10 shows that the 2019 and 2021 models follow one another reasonably closely until about 2010, when they diverge substantially. This divergence in the effect of age occurs substantially later than it does for the DNN, but it also is more dramatic. For the 2019 model, the premium on newer homes continues to increase quadratically for

Table 5: Exhibit 1 Reproduced from Zietz, Zietz, and Sirmans (ibid.)

Variable	#Obs	#Pos	#Neg	#NotSig	Pos/Neg	NotSig/Sig	DisagreeRatio*
Age	78	7	63	8	0.11	0.11	0.24
Square Feet	69	62	4	3	15.50	0.05	0.11
Garage Spaces	61	48	0	13	0.00	0.27	0.27
Fireplace	57	43	3	11	14.33	0.24	0.33
Lot Size	52	45	0	7	0.00	0.16	0.16
Bathrooms	40	34	1	5	34.00	0.14	0.18
Bedrooms	40	21	9	10	2.33	0.33	0.90
Full Baths	37	31	1	5	31.00	0.16	0.19
AirCondition	37	34	1	2	34.00	0.06	0.09
Distance	15	5	5	5	1.00	0.50	2.00
# Rooms	14	10	1	3	10.00	0.27	0.40
Time Trend	13	2	3	8	0.67	1.60	3.33

*The ratio of (non-significant results + the lesser count of negative or positive significant results) / (the greater count of negative or positive results). This combines the three counts of significance into a single measure of ”disagreement in results.”

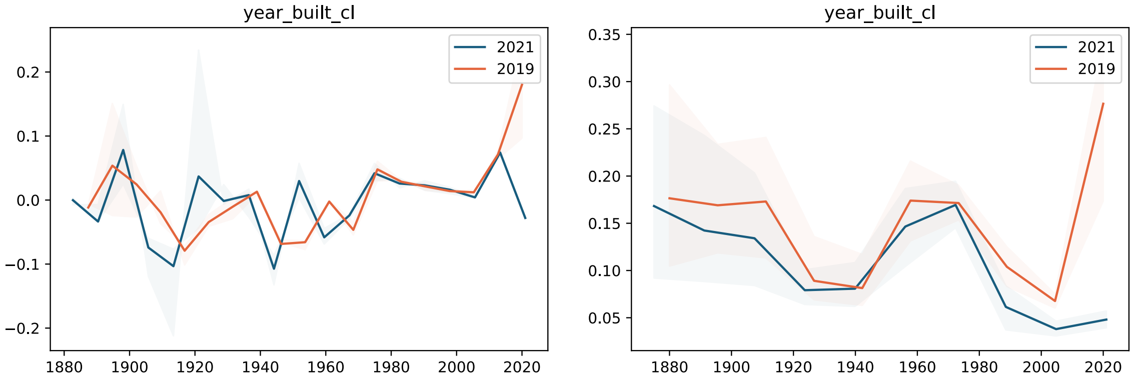
homes built after 2010. However for the 2021 model, newer homes bring zero premium or even demand a discount if they were built after 2010. According to the GBM, in 2019, a home built in 2019 commanded a 20% premium over a similar home built in 2010 while in 2021, a home built in 2019 would be priced similarly to (or even slightly lower than) a home built in 2010.

Looking at the SFIPDP provides a similar intuition. Here, we can see that for the 2021 model, a home's year of construction declines in importance around 5% for homes built after 2000 while the importance of a home's year of construction grows substantially over the same period in the 2019 model, reaching as high as 25% for homes built in 2019. This means that in 2019, newer homes had an increasing effect on the price of a home while in 2021, the age of a home did not matter as much, as long as it was built somewhat recently.

Figure 11 (left) shows the comparison of the PME for DNN-2019 and DNN-2021 for lot size. While the curves overlap at the upper range, they begin to separate for lower lot sizes and are meaningfully different in the region around .25 acres (around the size of a typical suburban size lot). Compared to 2019, this suggests increased competition in suburban and urban parts of the metro and an increased willingness of buyers to pay a premium for larger lots in desirable urban and suburban areas near urban centers.

Unlike the GBM, the SFIPDP values do not meaningfully shift between 2019 and 2021.

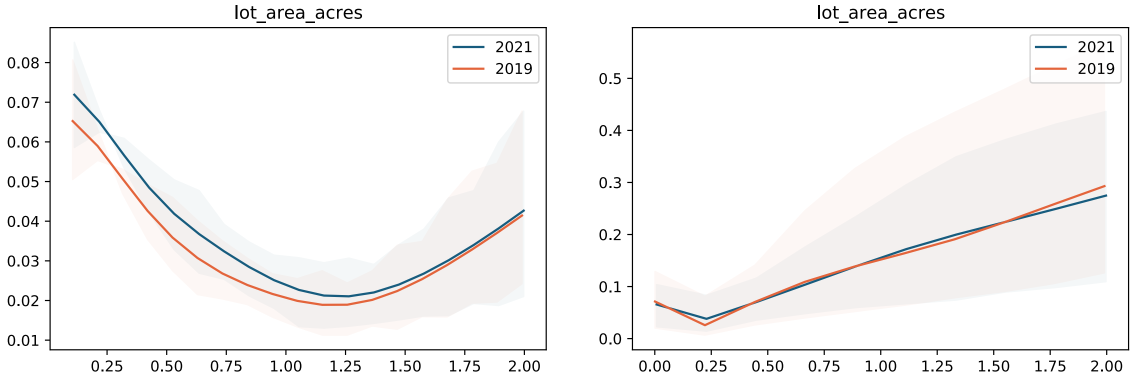
Figure 10: Year Built: PME and SFIPDP of GBM models



Left: demeaned PME Right: SFIPDP

This suggests that the increased priority in lot-size that is notable in the GBM model is otherwise captured by other variables in the DNN model. In some ways, this is unsuprising because the GBM model is forced to more aggressively choose among variables whereas the DNN models can more easily distribute effects across many features. Thus, the large shift in importance of lot size between GBM-2019 and GBM-2021 is distributed as small (insignificant) shifts in the SFIPDP of correlated features in the SFIPDP of the DNN models (e.g. bedrooms/bathrooms, school district, etc).

Figure 11: Lot Area: PME and SFIPDP of DNN models



Left: demeaned PME Right: SFIPDP