Using Machine Learning to Create High Performance Models for AVM without Linearity Constraints

Luke JORGENSEN and Joshua JORGENSEN, United States of America

Keywords: Large Spatial Data, Machine Learning, Mass Appraisal, Property Evaluation

SUMMARY

The purpose of this paper is to present the case for further exploring the use of machine learning models for AVM as an alternative to the broadly used linear regression models. In this paper a host of machine learning algorithms such as Random Forest and Light Gradient Boosting will be defined and implemented on a large set of single-family property (real estate) sale data. The predictions from the machine learning algorithms will then be compared to the linear regression model to assess the performance of the machine learning models by comparison. In addition, this paper will explore the application of Explainable Artificial Intelligence (XAI) algorithms to give insight to how the predictions are being made at a global and local level. Preliminary conclusions indicate machine learning algorithms can offer not only a high level of performance, but also explainability when compared to traditional statistical regression techniques.

Using Machine Learning to Create High Performance Models for AVM without Linearity Constraints

Luke JORGENSEN and Joshua JORGENSEN, United States of America

1. INTRODUCTION

In order to fund our governments (local, state, and federal) we need to have systems in place for taxation. One of the fundamental components of Florida's tax system on the local level is ad valorem property taxes. Therefore, it is crucially important that the value of all the properties in all jurisdictions are equitable ensuring everyone can pay their fair share of taxes. This job is constitutionally assigned to the county property appraisers (assessors) and their staff. Unfortunately, each county can have hundreds of thousands of individual parcels of property, where it becomes unrealistic to have individual valuation staff try to value every single parcel in a timely fashion. The value of a mass appraisal model is that a function can be used to value mass quantities of property instantaneously to high degrees of accuracy. Mass appraisal models allow property appraisers the ability to equitably value each property every year for taxation as well as quickly and fairly make adjustments to values. Linear regression has been the accepted and standard practice for mass appraisal models.

2. THE DATA

The data used in this study is a single-family housing market area in Lee County, Florida. Lee County contains in excess of 240,000 heterogeneous housing units from the ultra-exclusive Gasparilla Island to bedroom communities in Cape Coral and vast swaths of subdivisions in Estero and Bonita Springs. According to International Association of Assessing Officers (IAAO) and codified in Florida Statute, there are three commonly accepted valuation practices for modeling a mass market. They consist of the cost approach, the sales comparison approach, and the income approach. For purposes of this paper, the sales comparison approach is used as the basis for valuation. The sales comparison approach tries to value a market using the recorded sales within that market. To do so, ten years of sale data was pulled from the market area totaling 24,566 sales. For duplicate sales only the most recent sale was kept. Since the sales comparison approach is being used, the dependent variable (the value that the model is attempting to predict) is going to be sales price for the model.

3. LINEAR REGRESSION FOR MASS APPRAISAL

Linear regression is the standard method used for the creation of mass appraisal models. The mass adoption of linear regression is in part by its low technical barrier of entry and its simplistic explainability to an audience. Linear regression (also known as ordinary least squares) fits a line to linear combinations of the independent variables to predict the dependent variable. Linear regression has the form of $y = b_1 * x_1 + b_2 * x_2 + \dots$ where y is the dependent variable, b_n is the coefficient and x_n is an independent variable, where all combinations of x are linear. Linear regression has allowed for the adoption of mass appraisal on very large and diverse areas. By creating linear regression models, an appraiser can have an unbiased and accurate assessment for most of, if not the entire county by fitting a regression model to their data. Moreover, the coefficients of the model can explain how the model is making its prediction. For example, if the coefficient of area of the house is 10.02 we know that each additional square foot of the house increases the value by $10.02 indicating the value attributable to the size of an 1,800 square foot house is $18,036. While linear regression is a very powerful tool for mass appraisal it does have limitations. Some of the limitations are the constraints with independence of variables. While linear regression requires all predictors to be independent, many times predictors used in the prediction of a parcel's value are not and thus this assumption is broken and might lead to a poor performing model. Machine learning algorithms can naturally handle these limitations as well as offer high levels of explainability. To follow up, linear regression also assumes the relationship of variables to the outcome are linear, while again in predicting a parcel's value this may not be the case. The distance to open water or the size of a house are examples of this limitation. If a predictor is not linear the linear regression model will not truly capture the trend in the data. While techniques such as using splines, variable transformations, or logistic regression can help address non linearities, the explainability is severely impacted because the coefficient of the model can no longer be thought of as a single unit change. Since machine learning algorithms are utilizing non fixed rules to learn latent patterns in the data, machine learning models can naturally learn nonlinear relationships.

4. MACHINE LEARNING IN MASS APPRAISAL

Machine learning is a type of algorithm in artificial intelligence (AI) where the algorithm learns to recognize and subsequently make predictions from data. Machine learning has seen rapid growth over the past decade and much of this growth is due to the high level of performance machine learning has to offer. Machine learning algorithms such as the ensemble methods explored in this paper utilize non-fixed rules to learn latent patterns in the data. This allows machine learning models to overcome obstacles such as nonlinear data and non-independent variables in the model. Moreover, ensemble machine learning methods are a reasonable alternative to the singular prediction linear regression model, as ensemble methods are the combination of multiple models and help improve accuracy. All of these techniques allow for machine learning models to typically be much more accurate predictors than the standard practice of statistical regression. In this paper four machine learning models are created and tested: a Random Forest regressor, a Gradient Boosting regressor, Histogram-based Gradient Boosting regressor, and a Light Gradient Boosting machine. A Random Forest is a type of meta-estimator that utilizes decision trees to classify/predict and subset the data. Averaging is then used to create a more accurate model and to mitigate over-fitting (Scikit-learn, 2011). Gradient Boosting is used to build an additive estimation model, the model utilizes forward stage-wise structure that is suited for the performance of arbitrary differentiable loss functions. In every individual staged step, a regression tree is fitted to the negative gradient of the associated loss function (Scikit-learn, 2011). A Histogram-based Gradient Boosting algorithm is a more time-efficient version of the Gradient Boosting program with built-in support for missing values (Scikit-learn, 2011). Light Gradient Boosting machine is a gradient boosting framework with the advantages of being efficient with low memory consumption, while at the same time being able to handle very large sets of data (Ke, 2017). The data is subsequently split into training and testing data for all the models. This allowed for an unbiased comparison of the model performance by training the model on one set of data and then viewing its performance on a test set that the models have never seen before.

5. MODEL CONSTRUCTION

A model is considered to be a high-performing model if it has a high level of accuracy, utilizing as few independent variables/predictors as possible. Thus, to compare these models a limited number of predictors (no more than 20) were used in the model. The features used in the models include base living area, land square footage, home quality, home type, time of sale, as well as several other predictors. The models consist of both numerical (e.g., living area) and categorical predictors (e.g., having a pool). In order to optimize performance in all models tested, a transformer was utilized to standardize numerical features while one hot end encoding was utilized for categorical features. The Linear Regression, Random Forrest, Gradient Boosted and Histogram-based boosting models were trained from the Sklearn library (Scikit-learn, 2011) in python while the Light Gradient Boosting Machine was directly from the LGBM library (Ke, 2017).

6. MODEL PERFORMANCE METRICS

For a regressor model to be considered high performing it is desired to have an R squared value as close to 1 with as few independent variables as possible. The R squared value indicates on a scale between 0 to 1 what percentage of the additional error from the mean the model is able to explain. However, in the realm of property appraisal there exists additional means of determining if the model is a good indicator of a true housing market. For the values of a housing market to be considered fair, it is common practice to preserve vertical and horizontal equity. Here we rely upon the standards of equity proffered by the IAAO; the State of Florida has largely adopted the IAAO standards and guidelines (International Association of Assessing Officers, 2013). Vertical equity ensures consistency in appraisal levels by value ranges. The performance of the vertical equity is measured by the price related differential (PRD). Equivalently, horizontal equity can be defined as ensuring two or more property groups are appraised at the same percentage of market value when they have similar characteristics. In order to measure the horizontal equity, the coefficient of dispersion (COD) is utilized. In order to compare performance on both horizontal and vertical equity both the COD and PRD are used in the model comparison.

7. MODEL PERFORMANCE COMPARISON

Given that the market that is being modeled is an active and heterogeneous residential market area, IAAO states that the COD should be between 10 to 15. The PRD should be between

0.98 and 1.03 (International Association of Assessing Officers, 2013). Additionally, we would like the R squared of the model to be as close to 1 as possible indicating the model has accurately predicted all parcel values. Using the same exact data sets to train and test all the models with the same exact predictors, it was observed that the linear regression model had an R squared of 0.783, the Random Forest had an R squared of 0.879, the Gradient Boosting had an R squared of 0.918, Light Gradient Boosting had an R squared of 0.916, and Histogram-based Gradient Boosting had an R squared of 0.916 as well. Thus, it can be seen when comparing the R squared values the ensemble machine learning methods outperform the linear regression by a substantial margin as seen in Figure 1.

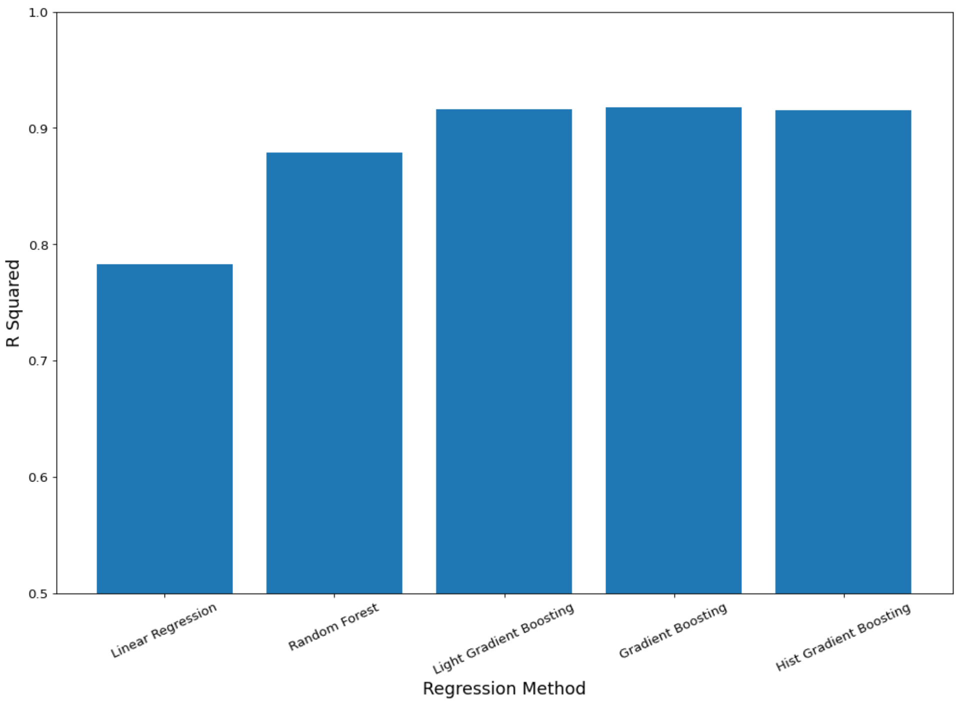

Figure 1

Now we need to see if the machine learning models maintain or surpass performance on vertical and horizontal equity as compared to linear regression models. For PRD the Linear model had a value of 0.983, the Random Forest had a value of 0.994, the Gradient Boosting had a value of 0.999, Light Gradient Boosting had a value of 1.00 and Histogram-based Gradient Boosting had a PRD of 0.997. Thus, it can be observed in Figure 2 that the machine learning models have an equivalent to superior performance to that of the linear regression model when being compared by their PRD.

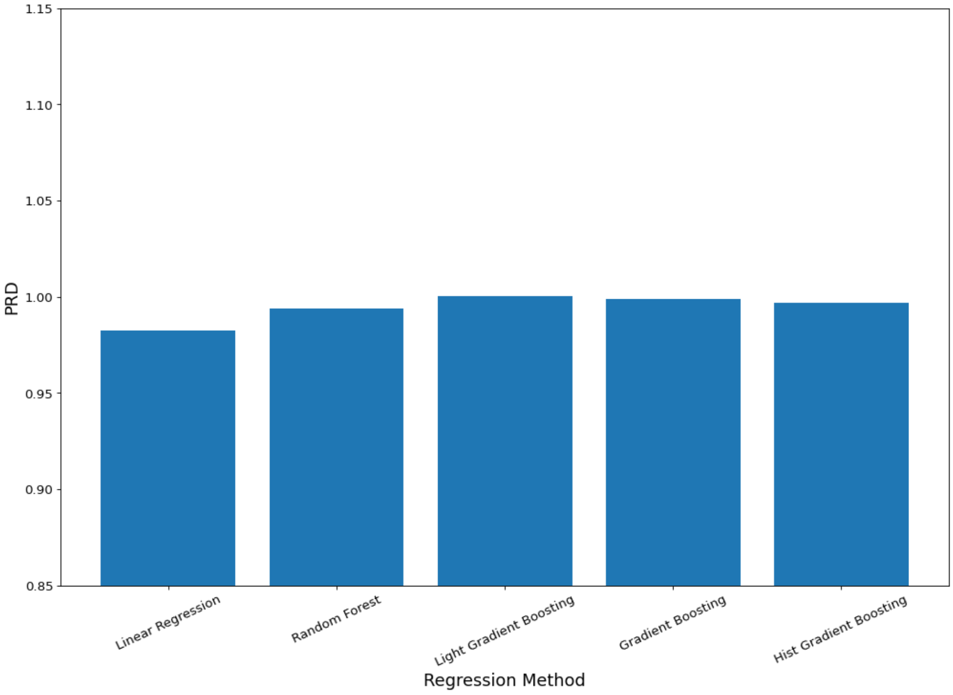

Figure 2

With a limited number of features being used on a large population area the linear model has a very poor performing COD of 41.62, although the machine learning models are performing within IAAO standards. Random Forest had a COD of 10.75, Gradient Boosting had a COD of 10.65, Light Gradient Boosting had a COD of 9.90 and Gradient Boosting had a COD of 10.49 as shown in Figure 3.

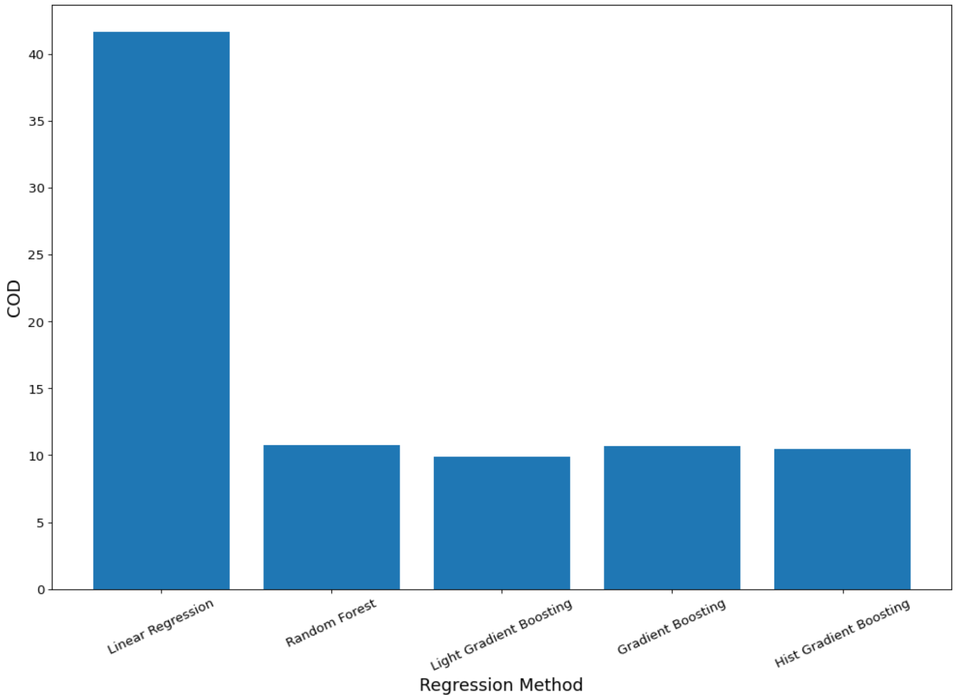

Figure 3

It is clear that the machine learning models are statistically more accurate than the standard linear model with their higher R squared values, lower COD values, and equivalent to superior PRD values. They preserve equity in accordance with the IAAO standards while the linear model fails to meet expectations and provide higher levels of accuracy.

8. HYPERPARAMETER TUNING

It is possible to further tune the machine learning models by performing hyperparameter tuning. With machine learning models the hyperparameters are values that control how the machine learning models learn from the data they are training on. These models come with default hyperparameter values that typically work very well across a multitude of applications but may not be optimal to all applications. By adjusting these hyperparameters we can potentially further increase the performance of the model. Precautions have to be taken as there are typically many hyperparameters for a machine learning model and sometimes a single hyperparameter can have an infinite number of values. One solution to tune these models is by defining all possible values that should be searched and use Bayesian Statistics to test out combinations of hyperparameters and perform a series of trials looking for the optimal set of hyperparameters. To implement this in our research we used Optuna (Akiba, 2019). Optuna is a hyperparameter optimization framework. By implementing Optuna on the Light Gradient Boosting model with only 50 trials the R squared increased from 0.916 to 0.924, the PRD moved from 1.00 to 0.999, and the COD further dropped 10.49 from to 9.19.

Similarly, when Optuna hyperparameter tuning was used on the Gradient Boosting model the R squared increased from 0.918 to 0.919, the PRD increased from 0.999 to 1.00, and the COD decreased from 10.88 to 9.88. Implementing hyperparameter tuning on high performing models nets a marked increase in their performance.

9. MODEL EXPLANATION

While having a high performing model is very important, it is also crucial that the model can be explained to the public. Since values need to be easily defendable to the public or in legal challenges, linear regression has been considered ideal since the coefficients of the model help to easily illustrate the values impact. To account for nonlinearities in linear regression splines, variable transformations, or logistic regression is applied. Implementing these techniques on linear regression, sometimes results in explanation that can be quite convoluted since the coefficients no longer indicate a single unit change. Typically, since machine learning uses non-fixed rules in their algorithms, machine learning has been thought of as a black box. Thankfully the field of Explainable AI or XAI has been addressing this issue. One such method that is used in this research is SHAP. SHAP stands for SHapley Additive exPlanations and can create explanations by quantifying the contribution that each predictor brings to the prediction made by the model (Lundberg, 2017). This is done based on the idea that the outcome of each possible contribution of predictors should be considered to determine the importance of a single feature. One of the key features of SHAP is that SHAP provides a unified local and global explanation. Thus, SHAP for global explanations can show how a given predictor's value impacts the prediction over the entire data set for all value ranges of that predictor. In the below example, we can see how base (living) area is impacting the value of the parcel according to the model.

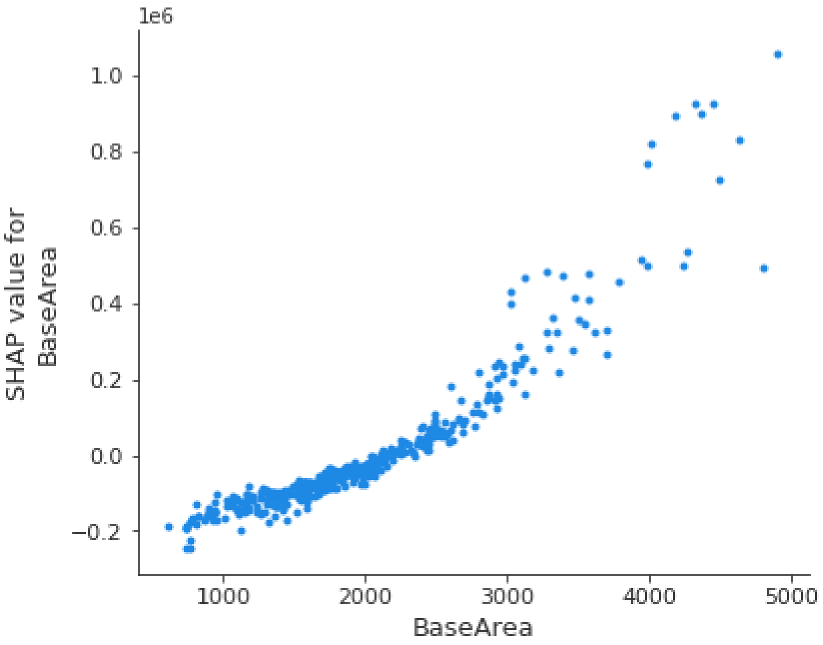

Figure 4

In Figure 4, as expected, we can see that as the house size increases there is a positive effect on the home's value. Moreover, rate curves could now be extracted from the explanation allowing appraisers to ensure these match industry standards. Lastly, one of the strongest points to observe is that this curve is not linear and illustrates how the model has learned from the data: it captures non linearities that naturally exist. Another example would be how a home's age impacts the value as shown in Figure 5.

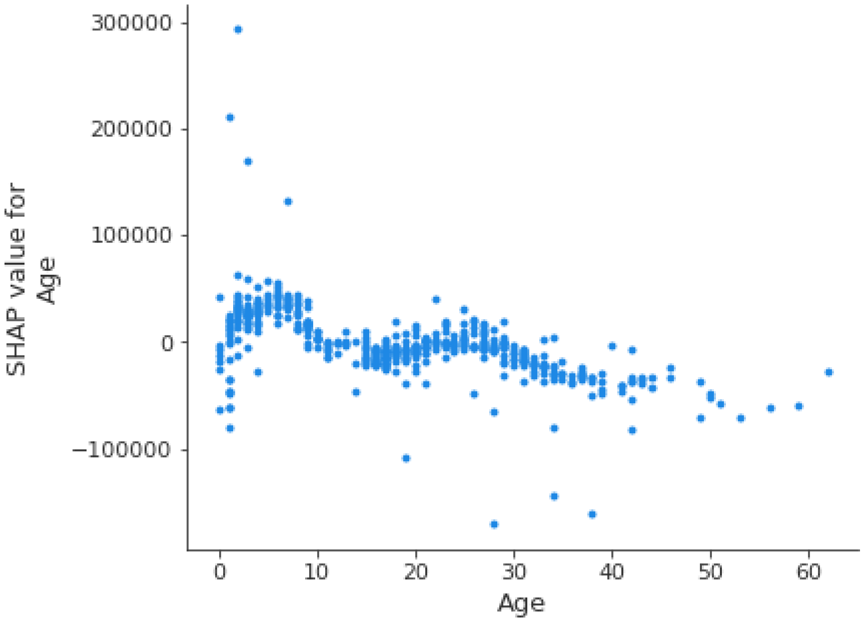

Figure 5

In Figure 5, the data shows as homes get older the value impact is negative. An interesting note: there is a slight increase in value for homes between 20-30 years old: this illustrates the model's ability to capture nonlinearities in the market (such as housing renovations). In addition, we have the ability to view interactions. Let's say we wanted to see the importance of base area’s relationship to age. Figure 6 illustrates the impact of base area on age.

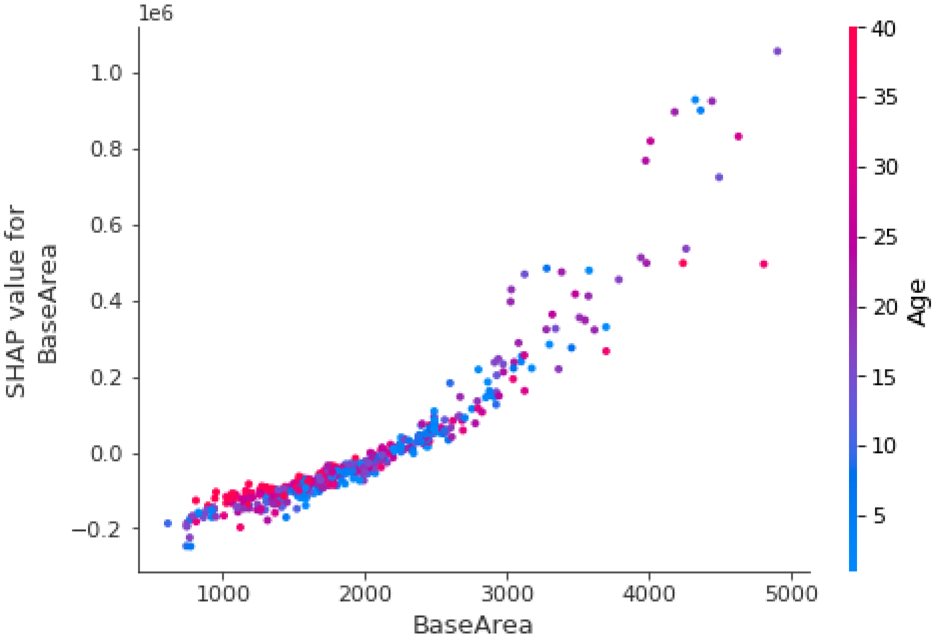

Figure 6

In Figure 6 the x axis shows the base area of the home, while the y axis shows the SHAP value or the impact the base area is having on the predictions. In addition, the points are now color coded to blue for a low home age and red being a high home age. We can clearly see now not only how base area is impacting a home's value, but also older homes tend to be smaller.

Local predictions in SHAP can give a detailed breakdown of how an individual home's value was predicted by the model. This will allow the appraiser to quickly and easily visualize how adjustments were made to come up with a house's value for example as shown in Figure 7.

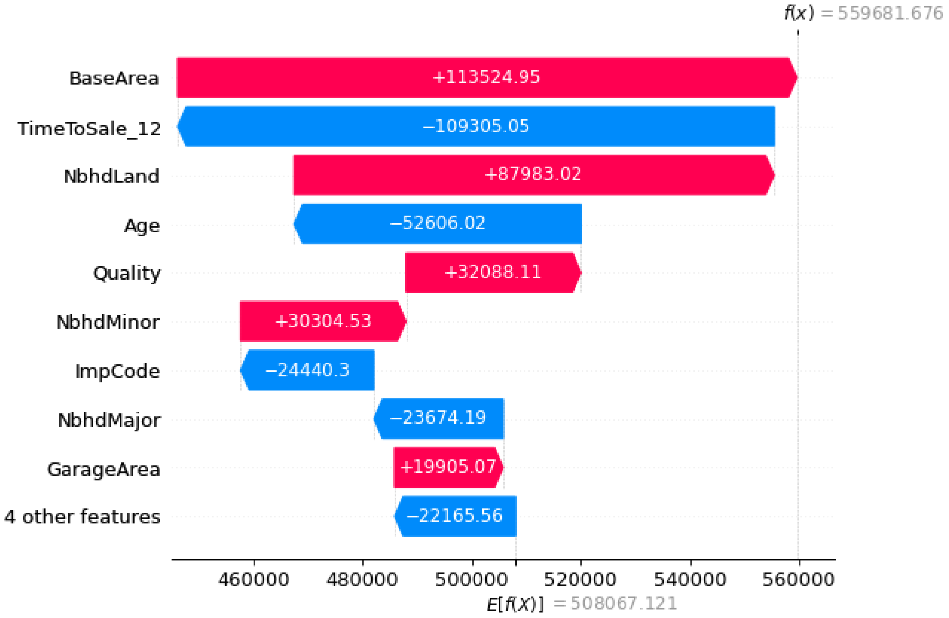

Figure 7

For Figure 7, the average predicted value in the comparable area is $508,067.12 and the predicted sale amount is $559,681.68. The values from the model are driven by the larger base area, adding $113,524.95 to the value. Making an adjustment for the time of the sale had a very negative impact of $109,305.05. The Nbhdland or property frontage of being on a canal had a significant positive impact of $87,983.02. The age reduced the value slightly, as expected, and the superior quality pushed the value upwards. By way of comparative figures, we've added a force plot of the same data in Figure 8.

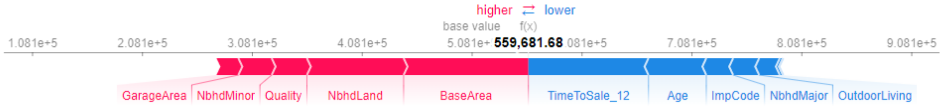

Figure 8

Figure 9 demonstrates the dynamic nature of the models. This allows us to adjust overall importance based on location and attributes accounting for changes naturally.

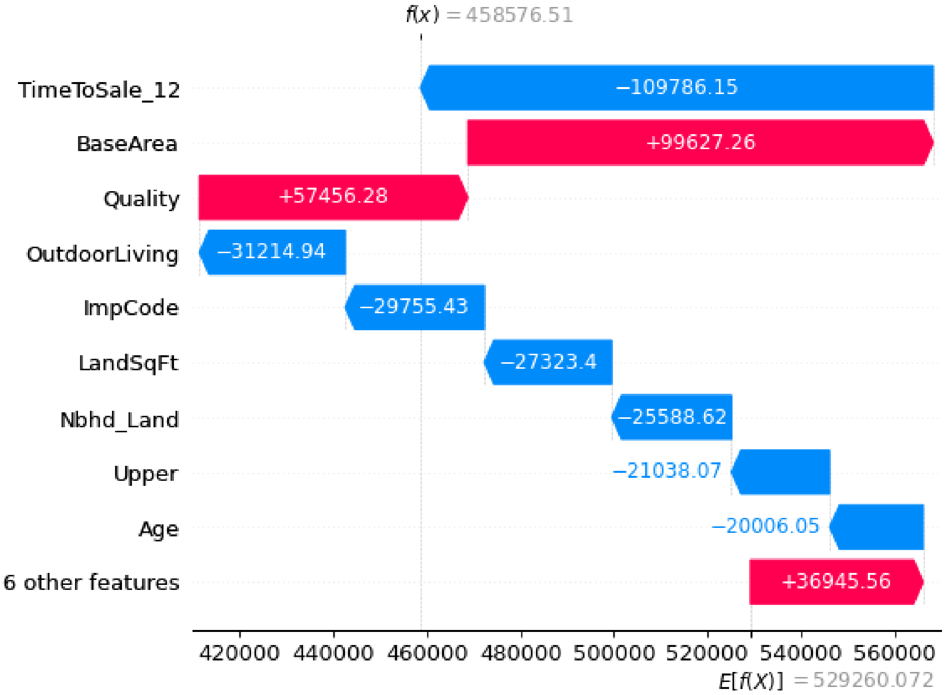

Figure 9

10. CONCLUSION

Employing machine learning models addresses many of the limitations of linear regression models. Machine learning in the AVM process significantly changes the efficacy of the models by accurately capturing trends, patterns, and interactions in the data. Machine learning models have not only outperformed the linear regression models, but they also maintained appraisal vertical and horizontal equity where the linear model failed to do so. Techniques such as SHAP were applied to the model in order to provide a high level of “explainability” to the outcomes that capture nonlinearities and other nuances. The benefits derived from SHAP are significant: providing a higher level of explainability to machine learning models, global and local explanations, clear explanations for the public and simple extraction of rate curves etc. The results of this research show that, given a set of significant data observations, machine learning algorithms outperform linear regression and are a viable and potentially superior option for providing mass appraisal values to large and diverse areas.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining.* pp. 2623-2631.

International Association of Assessing Officers (2013). Standard on Ratio Studies: A criterion on measuring fairness, quality, equity, and accuracy. *International Association of Assessing Officers.*

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research.* pp. 2825-2830.

BIOGRAPHICAL NOTES

Luke Jorgensen: Bachelor's in Applied Mathematics in Physics with a Physics minor from Stetson University and a Master’s in Applied Mathematics from Rensselaer Polytechnic Institute. Experience with nonlinear dynamics, wave theory, and differential equations. Certified Florida Evaluator designation from the Florida DOR. Currently employed at the Lee County Property Appraiser as a Data Analyst specializing in statistics, modeling and mathematics.

Joshua Jorgensen: Bachelor’s in Mathematics with a minor in Statistics from Florida Gulf Coast University. 11 years of experience in data science specializing in Machine Learning, Big Data, Data Mining, and Statistical Analysis. Currently employed by CGI as a Data Scientist.

CONTACTS

Data Scientist: Joshua Jorgensen
Organization: N/A
2480 Thompson St
Fort Myers
USA
Tel. +239-989-5348
Email: joshua.j.jorgensen@gmail.com
Web site: NA

Business Data Analyst: Luke Jorgensen
Lee County Property Appraiser
2480 Thompson St
Fort Myers
USA
Tel. +239-533-6100
Email: ljorgensen@leepa.org
Web site: NA