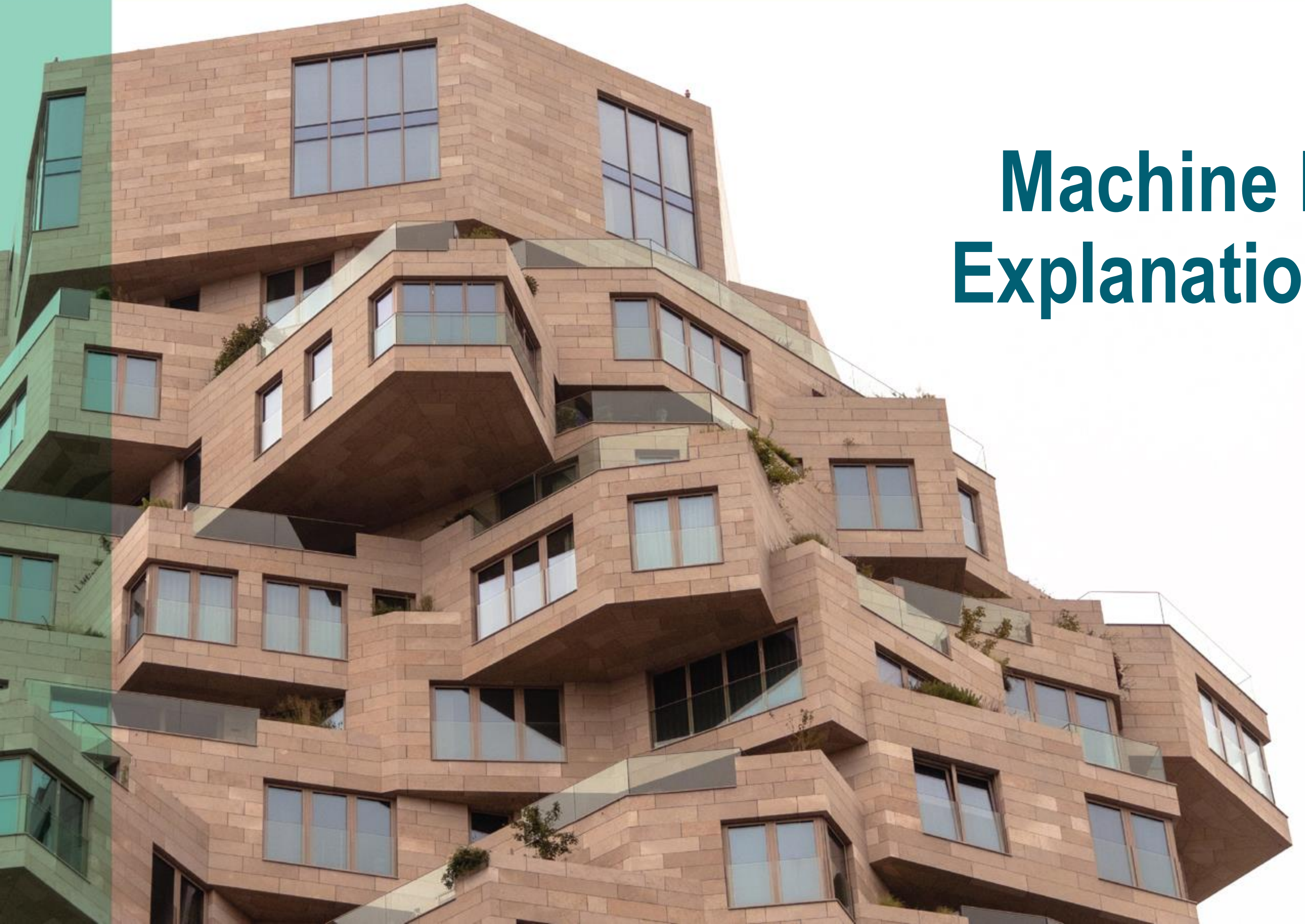
INTERNATIONAL RESEARCH SYMPOSIUM
Amsterdam, The Netherlands · December 4 - 5, 2024
ASSESSMENT INNOVATION & COLLABORATION WITH A FOCUS ON AI



Machine Learning and Model Explanation in Mass Appraisal

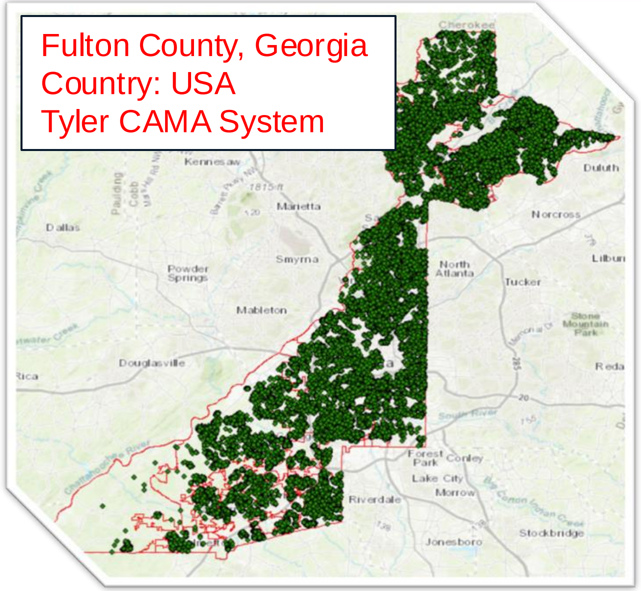
Joseph Wehrli (Joe)
Tyler Technologies

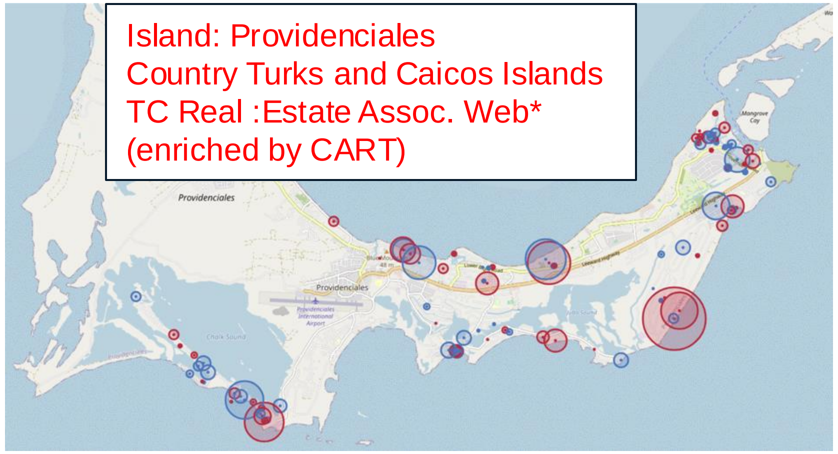
IAAO
IFI IPTI
International Property Tax Institute

Outline

Machine Learning and Model Explanation in Mass Appraisal

- ❑ AI/ML-based Direct Market Model
  - Gradient Boosting
  - General Regression Neural Network
- ❑ AL/ML Model Explanation
- ❑ AI/ML-based Comparable Sales Model
  - Similarity Model
  - Adjustment Model





Butler County, Ohio
Country: USA

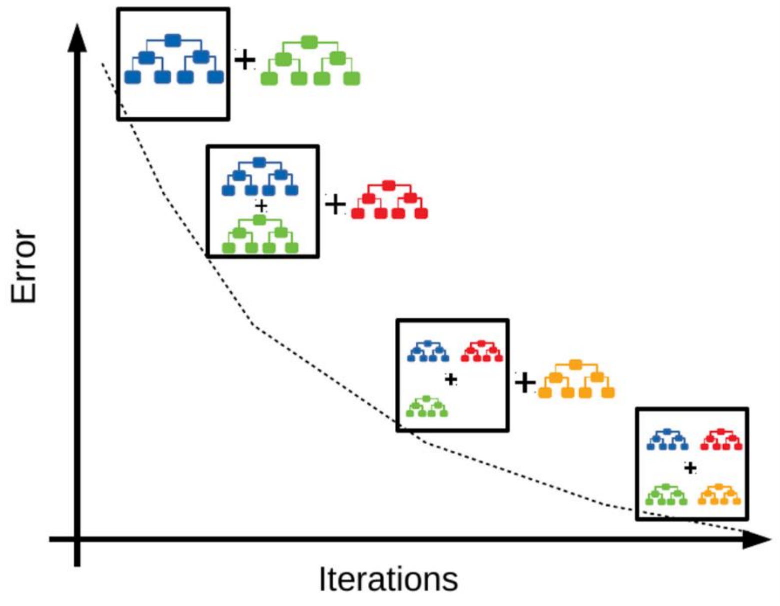
Model Calibration Differences with AI (ML)

- Avoidance of memorization, tending to the generalization capability of the ML model, i.e., how well does it do on examples it has seen vs examples it has not seen
- Usage of Hold-out datasets and more advanced K-Fold training and test set regimes
- Hyperparameter tuning
- Usage of variable importance and explanation techniques in lieu of parametric equation forms (rates) and fit parameters to review

AI-Based Direct Market Models

Gradient Boosting in Machine Learning

- Gradient Boosting Machines (GBM) is a powerful ensemble technique which combines the predictions of multiple weak learners sequentially to create a single more accurate strong learner.
- The weak learners are usually tree-based models.
- GBMs are among the current state-of-the-art ML techniques on tabular data in a variety of tasks such as prediction and regression.
- Can handle both numerical and categorical data, which eliminates the need for data conversion or transformation.



(Reference: https://medium.com/@hemashreekilari9/understanding-gradient-boosting-632939b98764 /)

Popular Gradient Boosting Tools

- Besides scikit-learn implementations, the three most famous boosting algorithm implementations that have provided various recipes for winning ML competitions are:
  - CatBoost
  - XGBoost
  - LightGBM

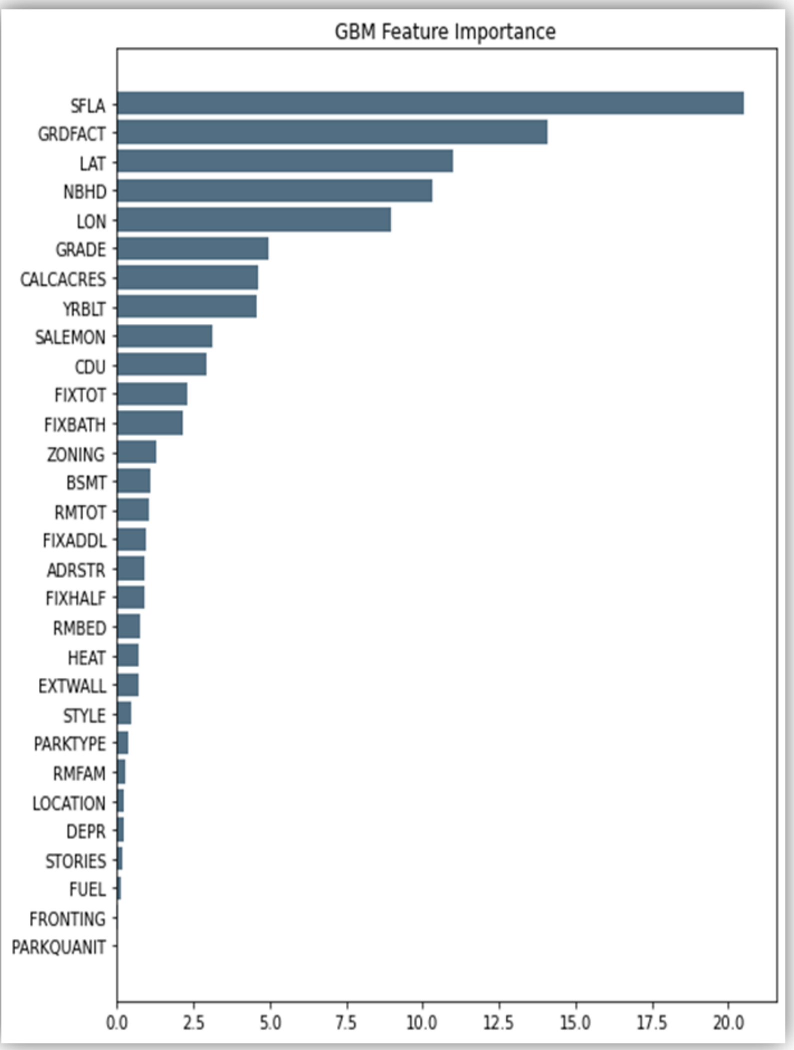
CatBoost (coined from “Category” and “Boosting”) is our choice of GBM engine

- Best supports Categorical and Text data

- Offers fastest prediction time and best performance
(based on internal benchmark comparison research)

Feature Importance in Gradient Boosting Models

- GBMs provide a score, called feature importance, that indicates how useful or valuable each feature was in the construction of the boosted decision trees.
- This importance is calculated explicitly for each attribute in the dataset, allowing attributes to be ranked and compared to each other.
- The more an attribute is used to make key decisions with decision trees, the higher its relative importance.



GBM Model Tuning

- Training CatBoostRegressor GBM model with randomly split sales dataset; 80% for training and 20% for test.
- GBM model hyper-parameters:
  - Tree-Specific Parameters: **max_depth, min_samples_leaf, max_features**, etc
  - Boosting Parameters**: learning_rate, n_estimators, subsample**, etc
  - Other Parameters: **loss, random_state**, etc
- Tuning method: grid search, random search, and Bayesian optimization
- Our Current Engine Choice: Optuna
  - Using a combination of Bayesian optimization and helping algorithms.
  - Efficiently search large spaces and prune unpromising trials for faster results.

GBM Regression Example Results

City: **Atlanta, GA USA (Fulton County)**
NBHD=14663

Sales data: 322 sales in 2017, 2018, 2019
Training data: 257, Test dataset: 65

RMSE - CatBoost (training): 24,470.47
RMSE - CatBoost (test): 39,259.87

Best hyperparameters:
'iterations': 1636,
'learning_rate': 0.0107,
'depth': 7,
'subsample': 0.5877,
'colsample_bylevel': 0.7625,
'min_data_in_leaf': 46,
'l2_leaf_reg': 2.0

GBM Error Metrics (Test Data Set):

RMSE	MAE	MAPE	Mean Sales Ratio	Median Sales Ratio	COD	COV
39,259	31,240	11.9	1.037	1.01	11.7	19.28

GBM Regression Example Results

Island: **Providenciales**
Country: **Turks and Caicos Islands**
Condo Listings
*TC Real Estate Association web data**

Sales data: 110 listings for 2023

Best hyperparameters:
'iterations': 1318,
'learning_rate': 0.0780,
'depth': 4,
'subsample':0.5819,
'colsample_bylevel':0.6597,
'min_data_in_leaf': 4

GBM Error Metrics (Sales Data):

RMSE	MAE	MAPE	Mean Sales Ratio	Median Sales Ratio	COD	COV
382,922	248,934	8.00	1.009	1.006	7.92	10.68

Observations that can be empirically tested

- Simple GBMs with <u>no</u> added feature engineering, and <u>no</u> “a priori” market area delineation often perform good enough for homogenous property stock if there is sufficient temporal and spatial information represented in the base attributes
- GBMs aggressively memorize and overfit
- GBMs are constant piecewise estimators; the valuation functions learned have jump discontinuities all over; ensembles dampens it (weighted sums) but do not remove it entirely
- Varying attributes causes no change in valuation *until at least one decision tree in the ensemble forest* has an evaluation that draws from a different leaf node

GRNN : General Regression Neural Network

- General Definition:
  - A General Regression Neural Network (GRNN) is a type of neural network used for regression tasks
  - GRNNs perform a type of non-parametric regression

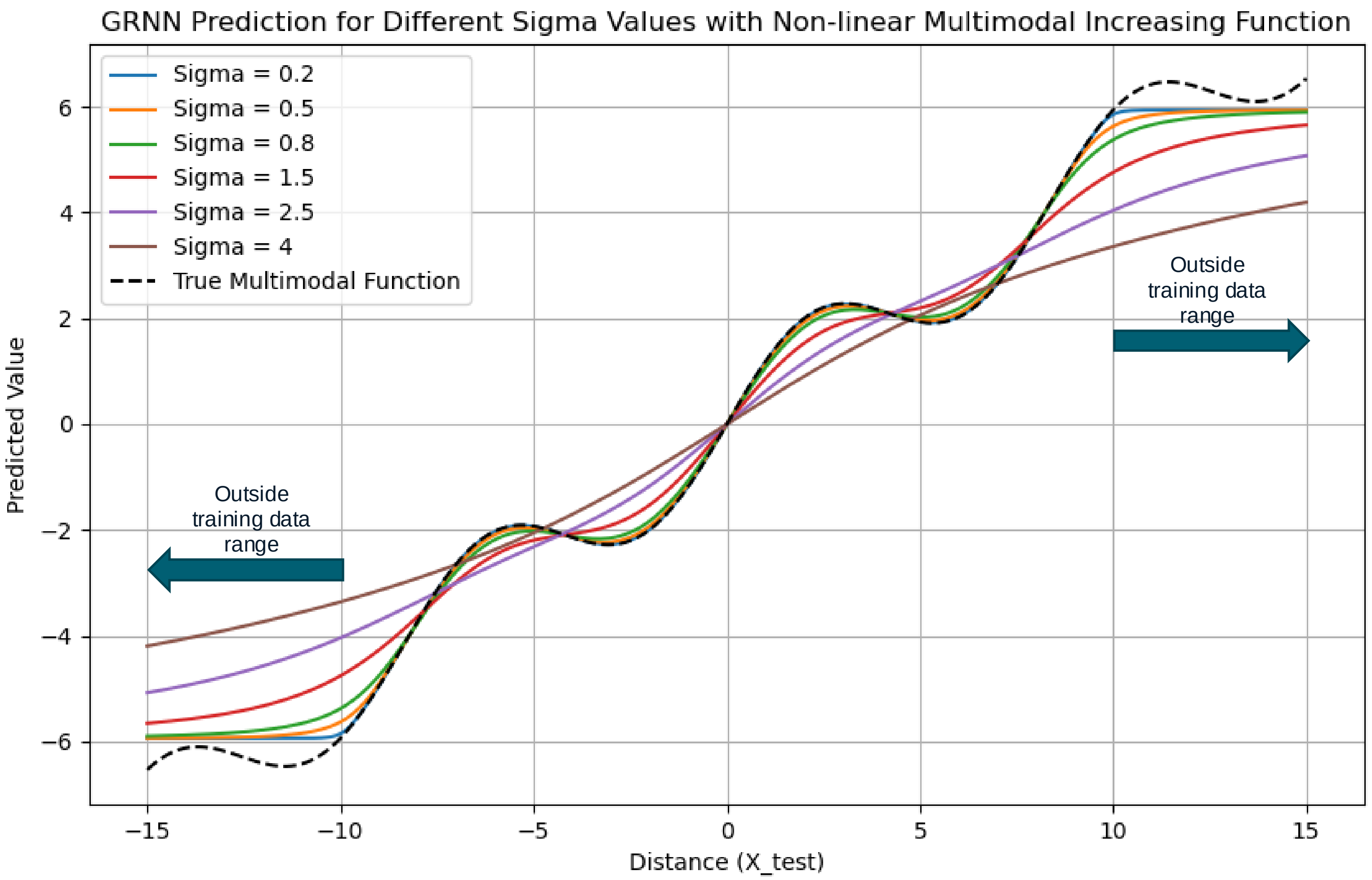
- Very Special Qualities:
  - Requires no iterative training process
  - It can model complex, nonlinear relationships
  - Particularly robust in handling noisy and sparse data
  - Standardized data preprocessing i.e., full automation

How does the GRNN work?

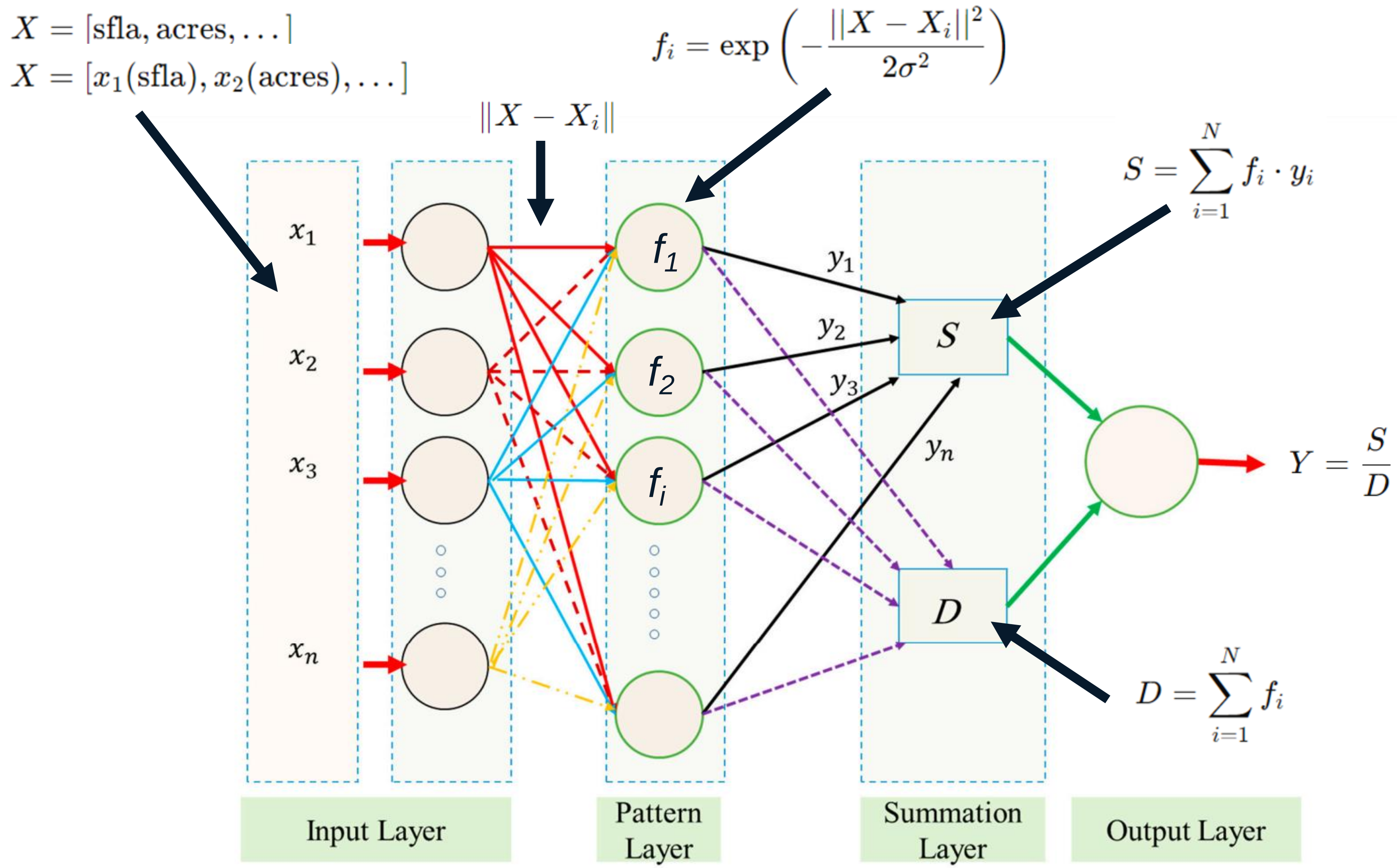
- **Smoothness Control**
  - GRNN uses a smoothing parameter, often denoted by σ (sigma), which controls the influence width of the training points (sale properties = training points).
- **Distance Measure**
  - GRNN computes the distance between the input and each training sample. This distance determines how much influence each training point has on the prediction for the new input.
- **Weighted Averaging**
  - The output of GRNN is essentially a weighted average of the training outputs, where the weights depend on the distances between the input and the training points, modified by the sigma parameter.
- **Non-linear mapping**
  - GRNN automatically performs nonlinear mapping of input features to the target outputs without explicitly requiring the user to specify the form of nonlinearity.

GRNN Hyperparameter - σ (sigma)

- The smoothing parameter σ (sigma) plays a crucial role in determining network performance.
- The farther a valuation point is from the training data, the less influence the training points will have due to the weight decay (based on the Gaussian distribution).
- Sigma controls the rate of this decay: a smaller sigma means the model is more sensitive to distance, and only nearby points significantly influence the prediction. A larger sigma means more distant points also contribute to the prediction.



GRNN Architecture



- Optionality (not needed in practice, academically interesting however)
- Select subsets or clusters of sales
- Continuous learning with decay
- Nuance in the distance function
- Nuance in activation function f
- Constraints in weights
- Usage of link functions for Y
- Introduce autoencoders of X
- Adapt with more advanced deep or generative network techniques

Interpretation of GRNN valuations

- GRNN calculates the distance between a subject and all sales. The model then uses a Gaussian (bell curve) weighting scheme to weigh the contribution of each training sale to the prediction.

- GRNN is interpretable since any subject (non-sold) valuation can be traced to and decomposed into the relative contributions from actual market prices (sales/sold parcels).

- The network nodes (sales) with the highest excitement, can be interpreted as “explanatory comparable sales” that are selected on the combined basis of both distance and predictive contribution

Explainable by highest weighted sales

Sale ID	Distance (d_i)	Sale Price (y_i)	Weight (f_i)	Weighted Contribution (f_i * y_i)
Sale A	0.3	295000	0.955997482	282019.2571
Sale G	0.4	298000	0.923116346	275088.6712
Sale B	0.5	300000	0.882496903	264749.0708
Sale J	0.6	307000	0.835270211	256427.9549
Sale F	0.7	305000	0.782704538	238724.8842
Sale C	0.8	320000	0.726149037	232367.6919
Sale H	0.9	312000	0.666976811	208096.765
Sale I	1	299000	0.60653066	181352.6673
Sale D	1.2	310000	0.486752256	150893.1993
Sale E	1.5	315000	0.324652467	102265.5272

Direct Market Model – GRNN and LAD LP Solver Regression

- Area: Butler County, Ohio
  - NBHD=R0215001 Sales: 94 2020, 2021, 2022

Method	R2	RMSE	MAE	Mean A/S	Median A/S	A/S COD	COV
GRNN	0.878	21138	15598	1.007	1.00	4.504	6.44
LAD	0.823	25444	17844	1.008	1.00	5.036	7.80

  - GRNN and LAD LP (Least Absolute Deviation Linear Program) Solver using same variables

    "CALCACRES","RMBED","FIXBATH","RMFAM","FINBSMTAREA","DEPR","SFLA","STORIES1","EXTWALL2","STYLE3","STYLE9","GRADEAm","GRADEBp","GRADEBm","GRADECp","CDUVG","CDUGD","RQOS"

When to consider GRNN over GBM

Case	GRNN Strength	GBM Limitation
Small Dataset or Sparse Data	Performs better with small datasets due to its non-parametric nature (RBF network)	Requires larger datasets to leverage its boosting process effectively
Noisy Data with a Smooth Underlying Function	Handles noisy data well, particularly when the underlying function is smooth	May overfit noise in the training data, even with regularization and early stopping techniques
Real-Time or Fast Training or Valuation Needs	Fast to train. Produces very fast predictions once trained, beneficial for real-time or interactive systems	Slower to train. Slower prediction times due to the complexity of the ensemble of trees
Lack of Extensive Feature Engineering	Works well with raw data processed in standardized ways	Works well with raw data processed in standardized ways. Can benefit from more attentive feature engineering
Smooth Function Approximation	Approximates smooth and continuous functions well by averaging outputs based on input vector similarity	May struggle with smoothness due to its piecewise constant models (decision trees)

Spatial analysis becomes critical with AI

- Example Area: Fulton County, GA
- Sales data:

  31,125 residential single-family sales

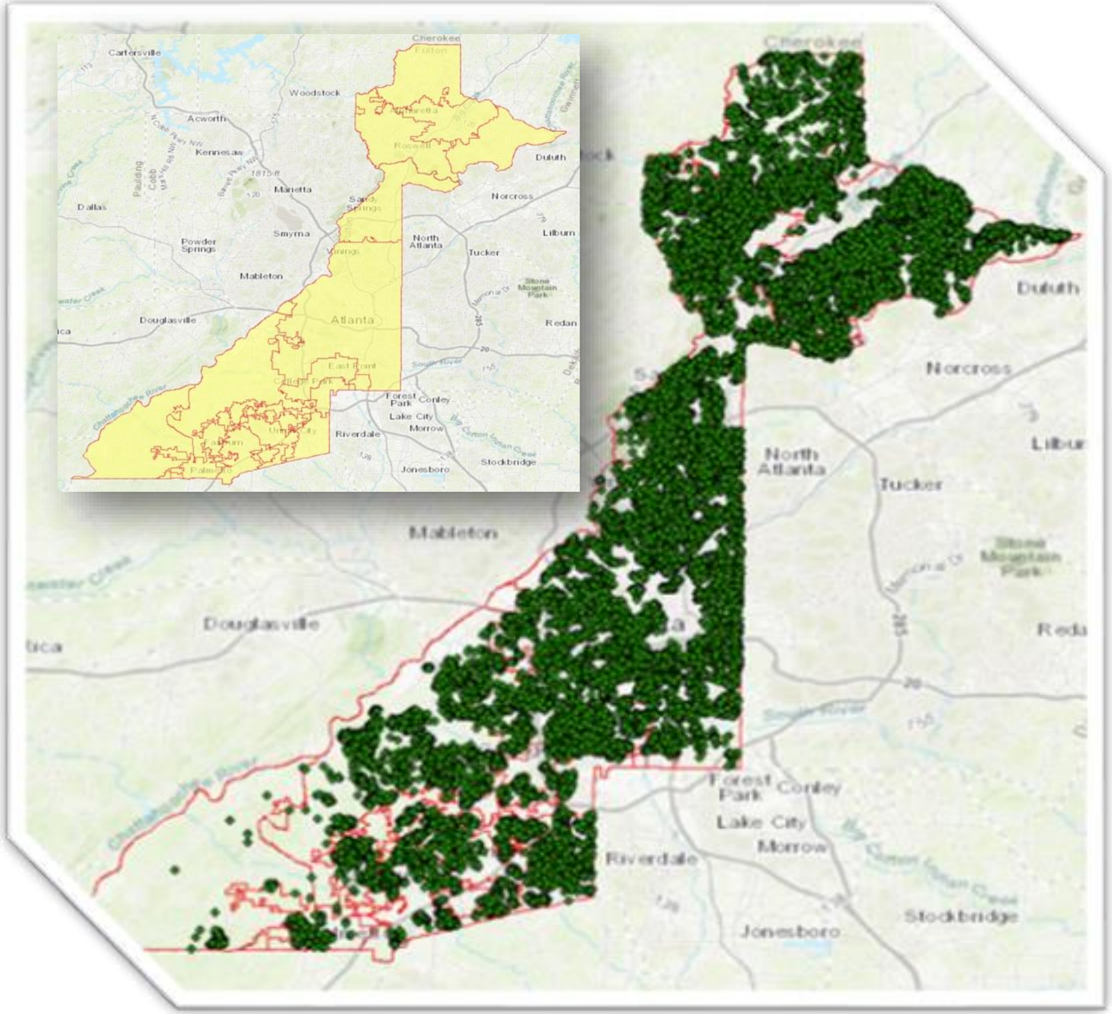
  from Jan 1, 2017 to Dec 31, 2019

- 17 Numerical Variables/Features

  CALCACRES, FRONTING, STORIES, YRBLT, EXTWALL, RMTOT, RMBED, RMFAM, FIXBATH, FIXHALF, FIXADDL, FIXTOT, BSMT, HEAT, FUEL, SFLA, GRDFACT, DEPR ,LAT, LON, SALEMON

- 13 Categorical Variables/Features

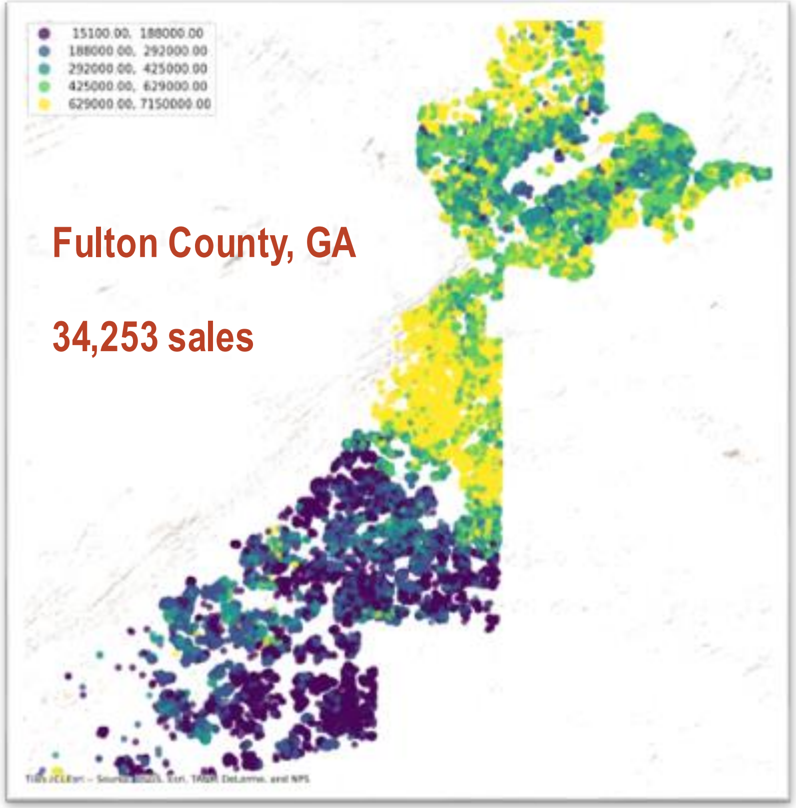
  NBHD, STYLE, ZONING, GRADE, CDU, LOCATION, ADRSTR, BSMT, HEAT, FUEL, FRONTING, EXTWALL, PARKTYPE



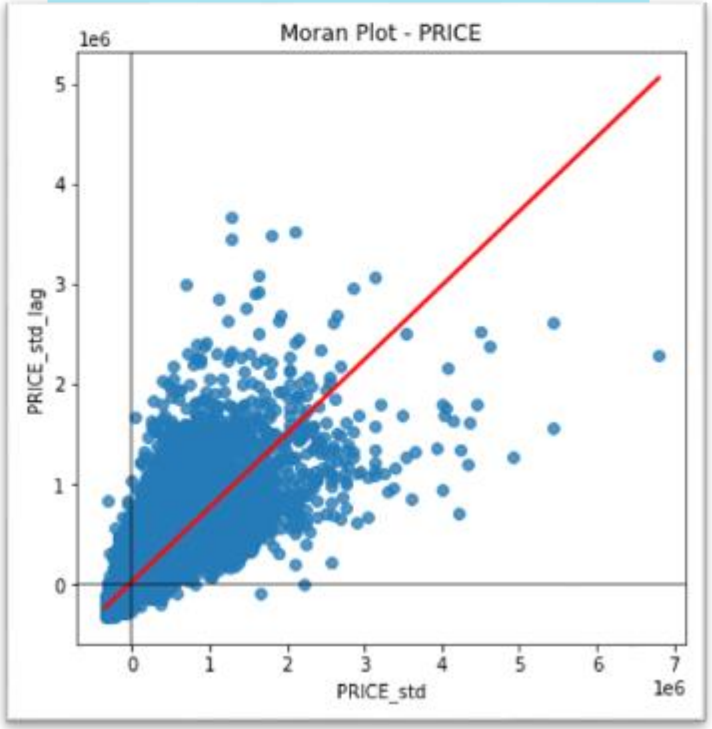
Spatial Autocorrelation: House Price

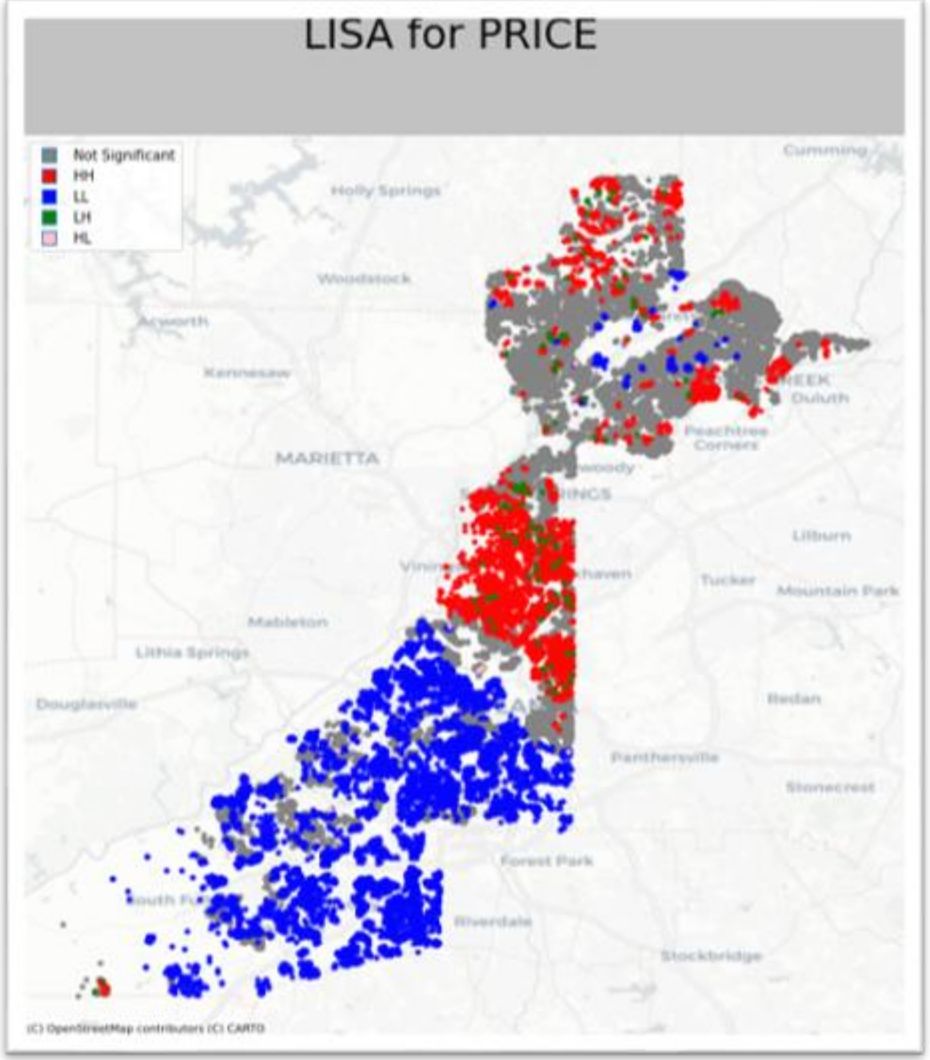
- **High positive spatial autocorrelation**
- **statistically significant**

Moran I: 0.72
p-value for Moran's I : 0.001
Not significant data: 19283 (57.8)%

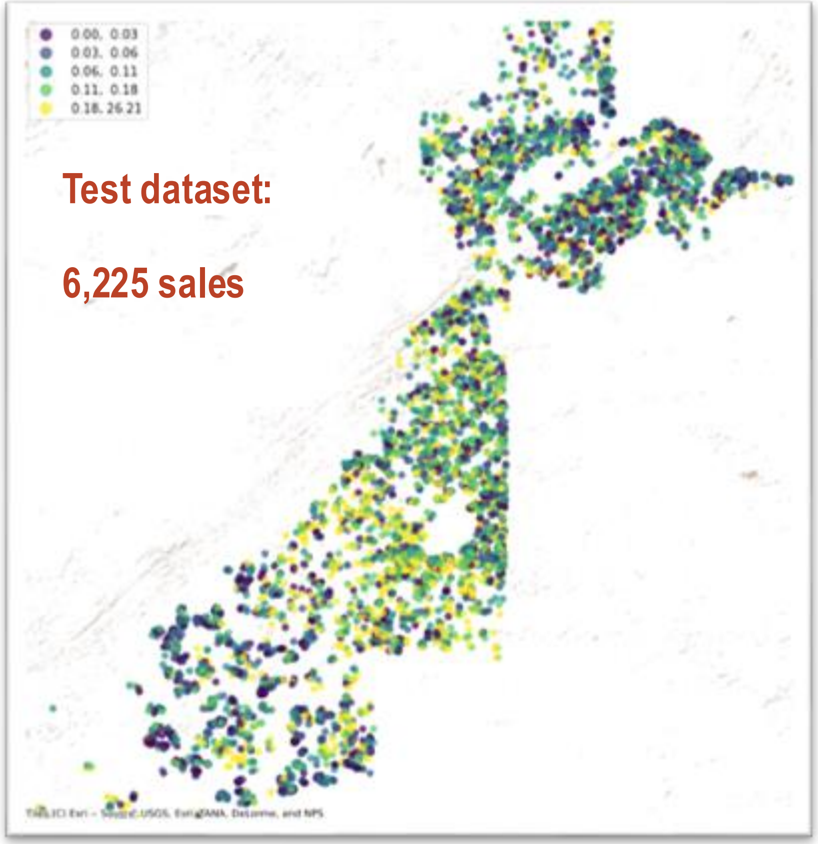


Choropleth Map

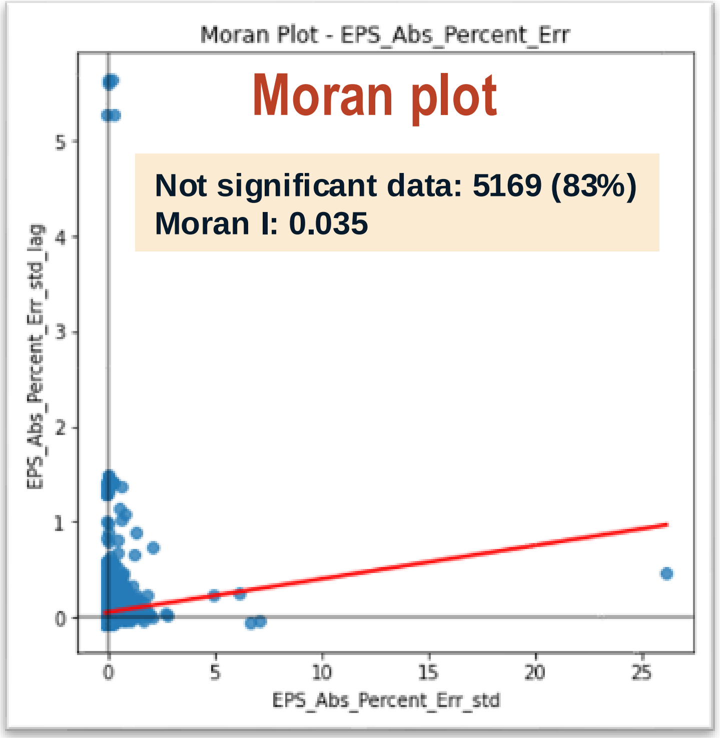


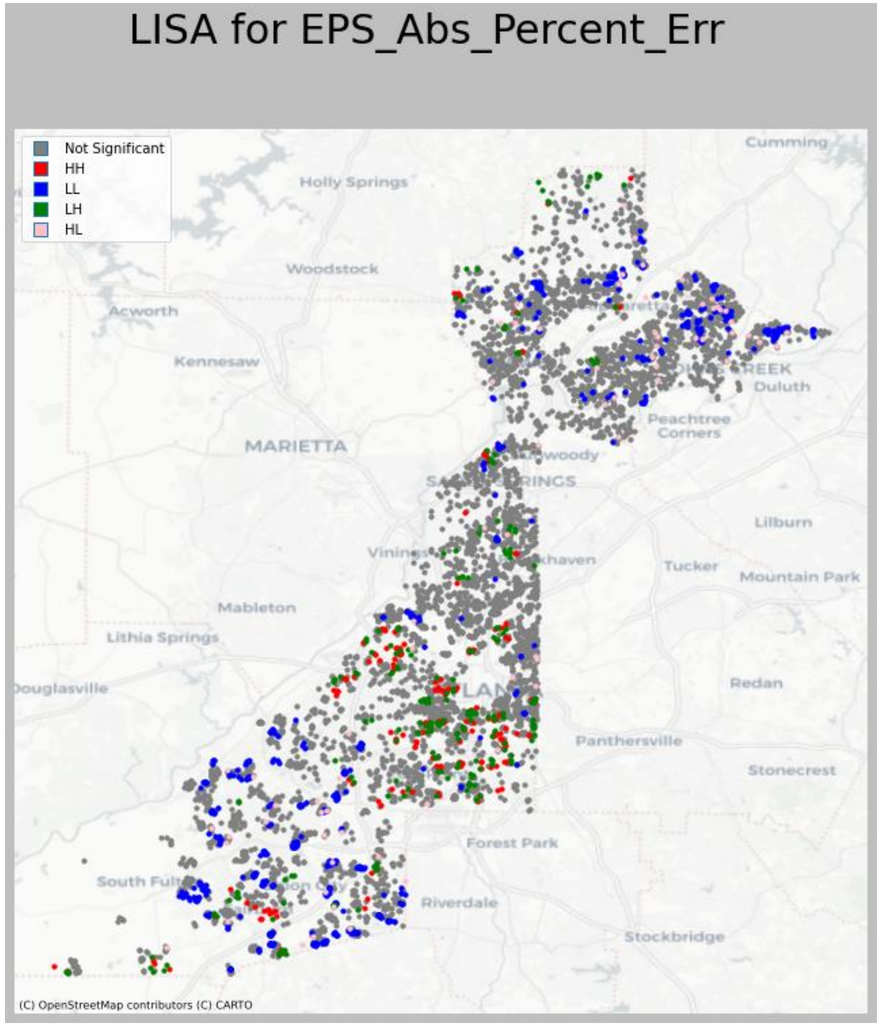


Spatial Autocorrelation: GBM Prediction

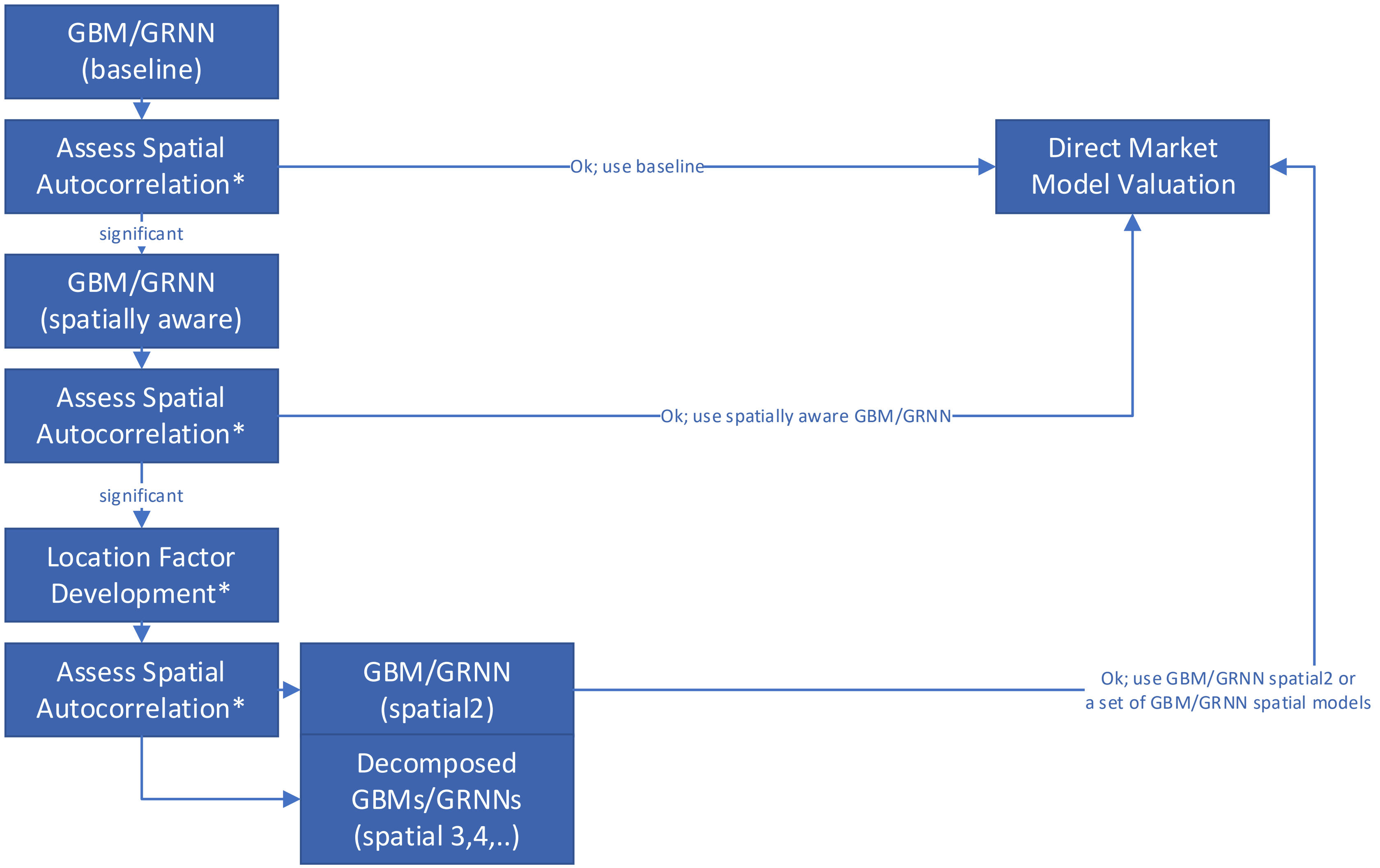


Choropleth Map



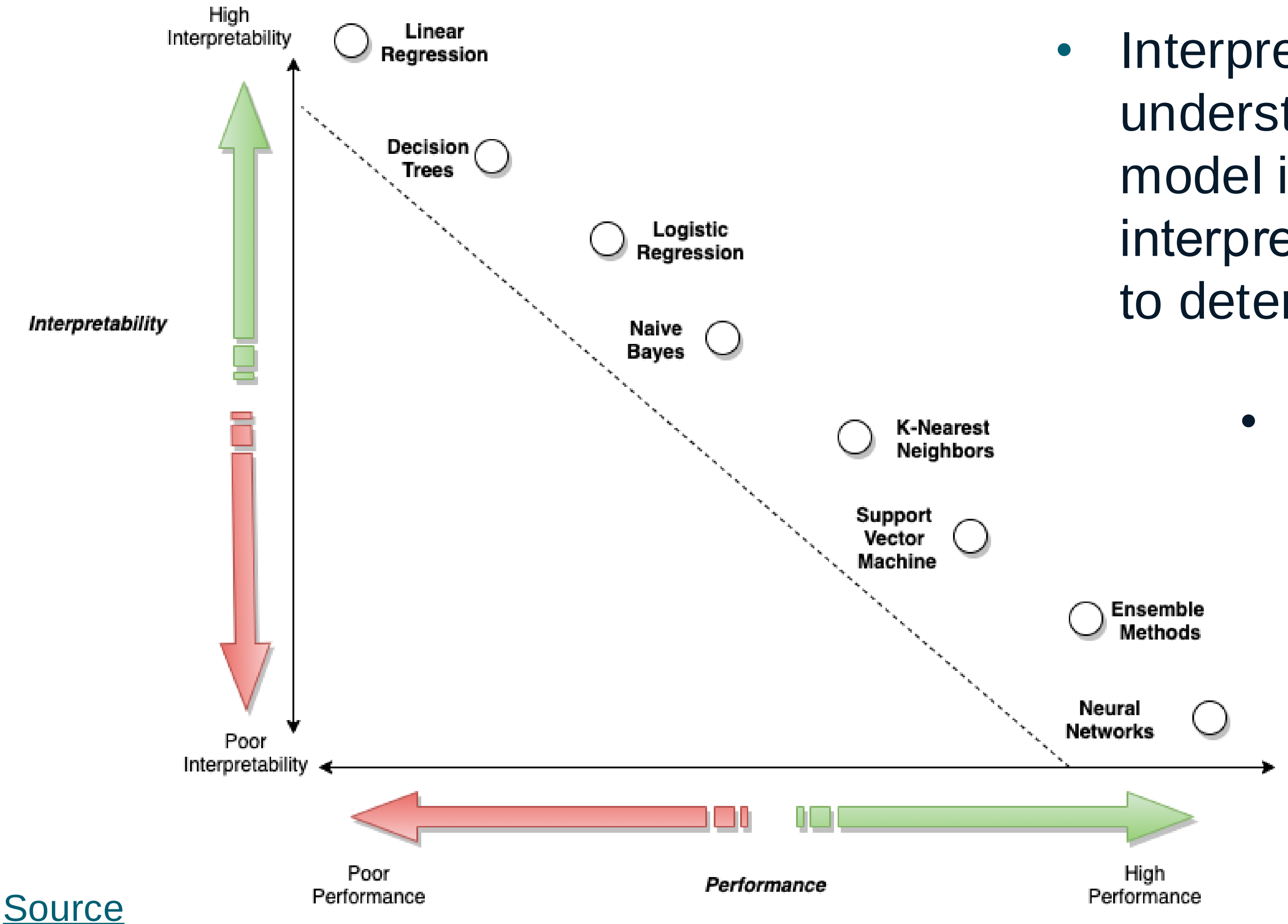


ML Workflow -- Direct Market Model



AI Model Explanation and Trust Tools

Interpretability Versus Explainability

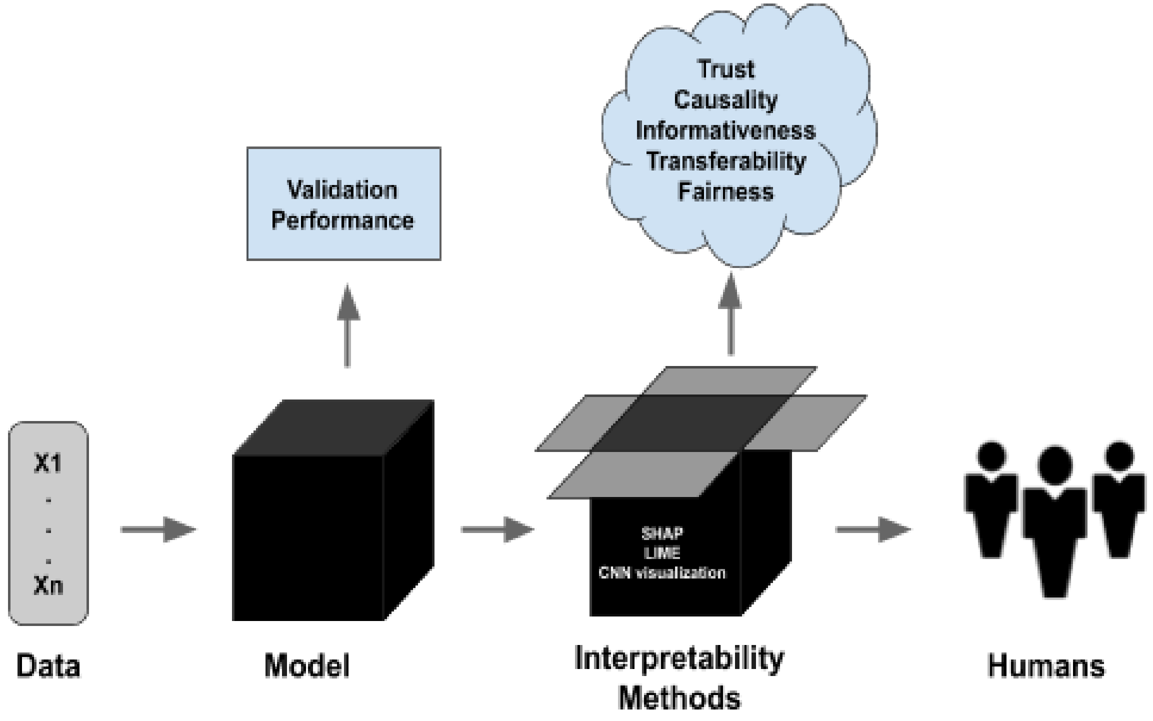


Source

- Interpretability: high model transparency, understand exactly why and how the model is generating predictions by interpreting model's weights and features to determine the given output.
  - Explainability: the ability to explain the decision-making process of an ML/AI model.
    - For complex models with poor interpretability, model explanation methods/tools are needed to interpret them.

Machine Learning Model Explanation

- Most ML/AI models are ‘black-box’ models. Their internal workings are not easily understandable.
- Model explanation is a critical necessity during calibration of AI/ML models for Mass Appraisal and Valuation.
- Model explanation enables users to understand why the model produced a particular result and valuation, so it is explained clearly and understandably.
- Enhance trust, transparency and fairness.



(Source: https://blog.ml.cmu.edu/2020/08/31/6-interpretability/)

Model Explanation Methods

❑ Model-specific / Model-agnostic

- Model-specific: specific to certain models, have interpretable inner mechanics like coefficients and weights.
- Model-agnostic: can be applied to any ML models after the model has been trained. Don't have access to model internals and work by analyzing feature input and output pairs.

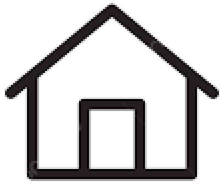
❑ Global / Local scope

- Global: describe the average behavior of a machine learning model and provide an overall explanation of the model's behavior.
- Local: explain individual prediction, capturing the reasons behind only the specified prediction.

Scope	Model-agnostic
Global	Partial Dependence Plot (PDP) Feature Importance Global Surrogate etc.
Local	Local Surrogate (LIME) SHAP Individual Conditional Expectation (ICE) etc.

Shapley Values

Based on coalitional game theory, each feature value of the property can be thought as a “player” in a game where the prediction is the payout.



AREA = 3000 BEDROOM = 4 STORY = 2 CONDITION = POOR

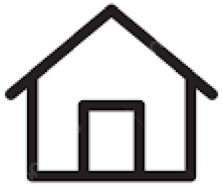
Marginal Contribution of **BEDROOM=4** in **One coalition** { AREA = 3000 STORY = 2 }

- For one coalition, compute the predicted price with and without the **BEDROOM=4** and take the difference to get the marginal contribution.
- Replace the feature values of features that are not in a coalition with random feature values from the dataset.
- The Shapley value is the (weighted) average of marginal contributions for all the possible coalitions.

SHAP (SHapley Additive exPlanations)

- SHAP is a method to explain individual predictions based on shapely values of each feature.

The difference of $50,000 explained:



Prediction:	$500,00
Average Prediction:	$450,000
Difference:	$50,000

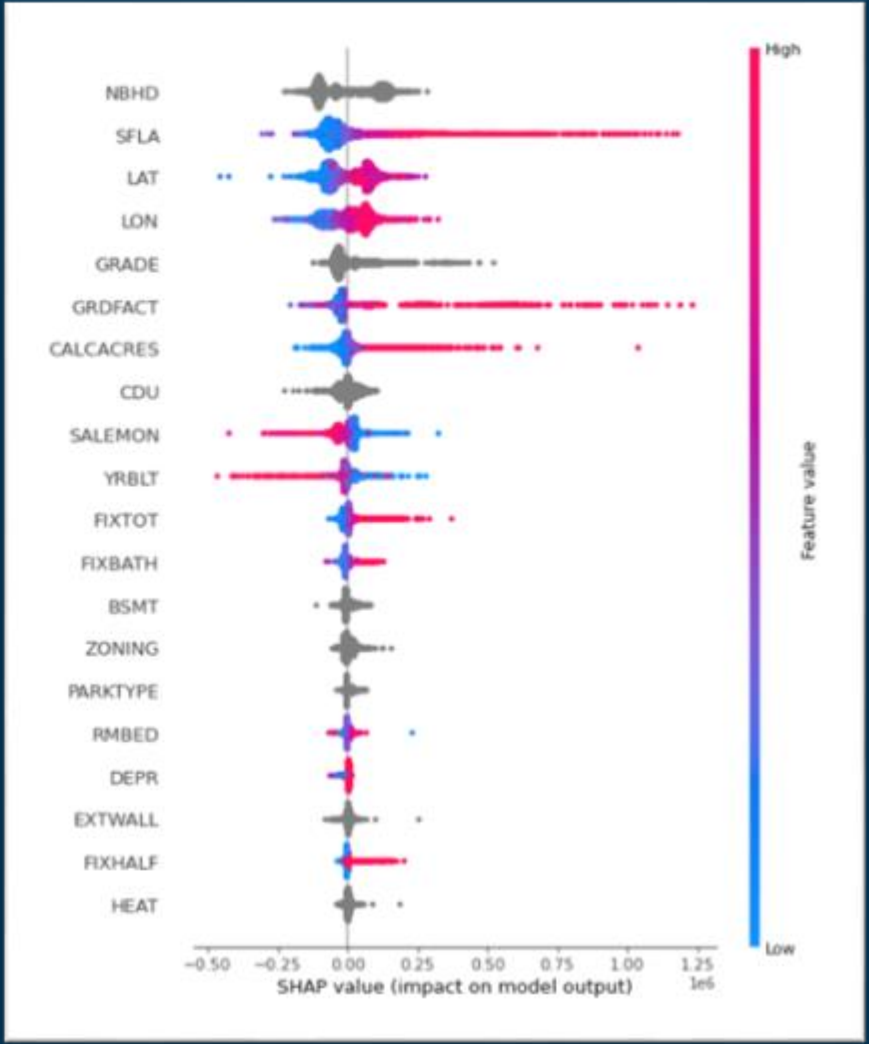
Feature	Feature Value	Marginal Contribution (Shapley Values)
AREA	3000	50,000
STORY	2	10,000
BEDROOM	4	10,000
CONDITION	Poor	-20,000
	Sum:	50,000

Marginal contribution of each feature

SHAP: Global Explanation

Summary Plot

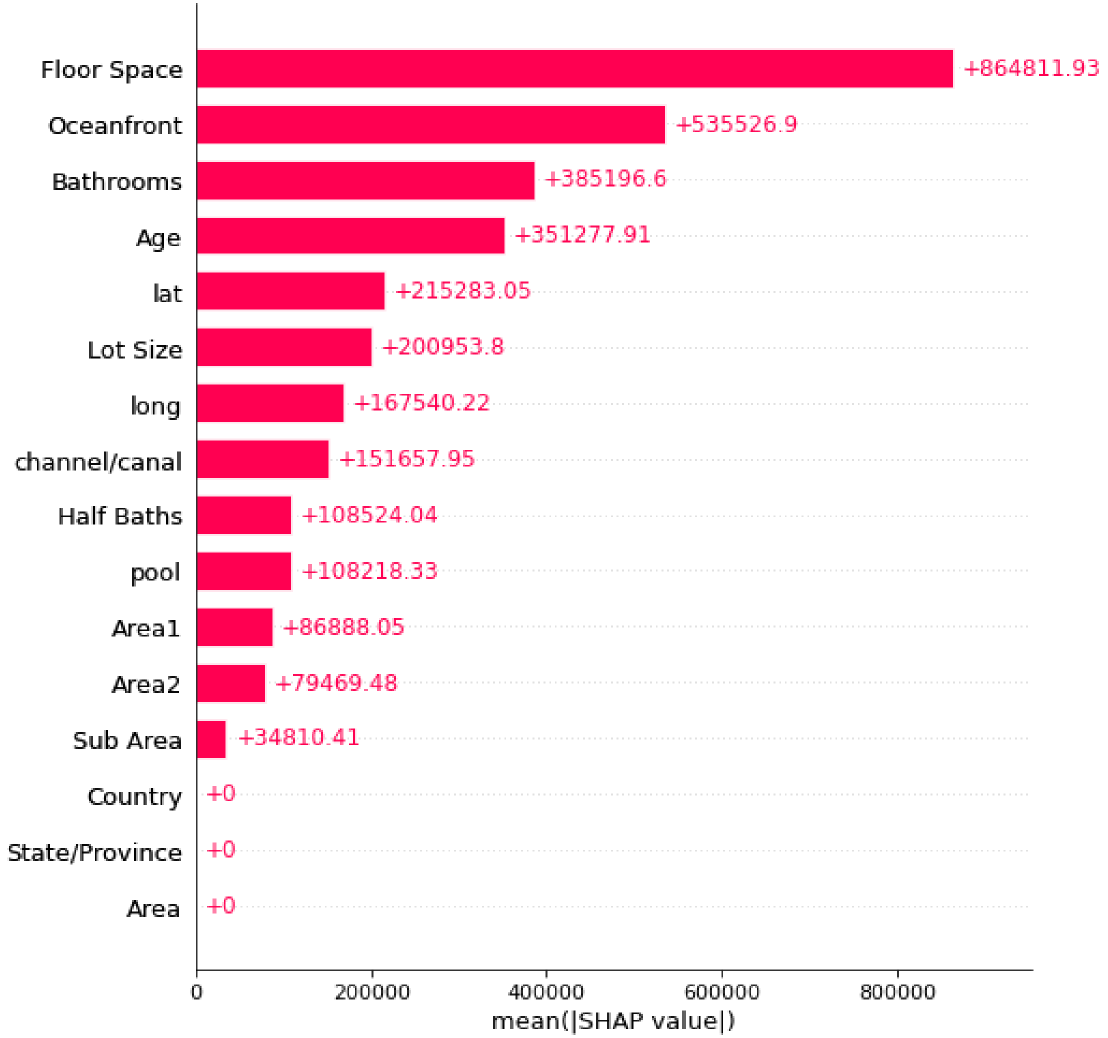
- Each point on the summary plot is a Shapley value for a feature and an instance.
- Show the relationship between the value of a feature and the impact on the prediction.



Area: Fulton County, GA Country: USA

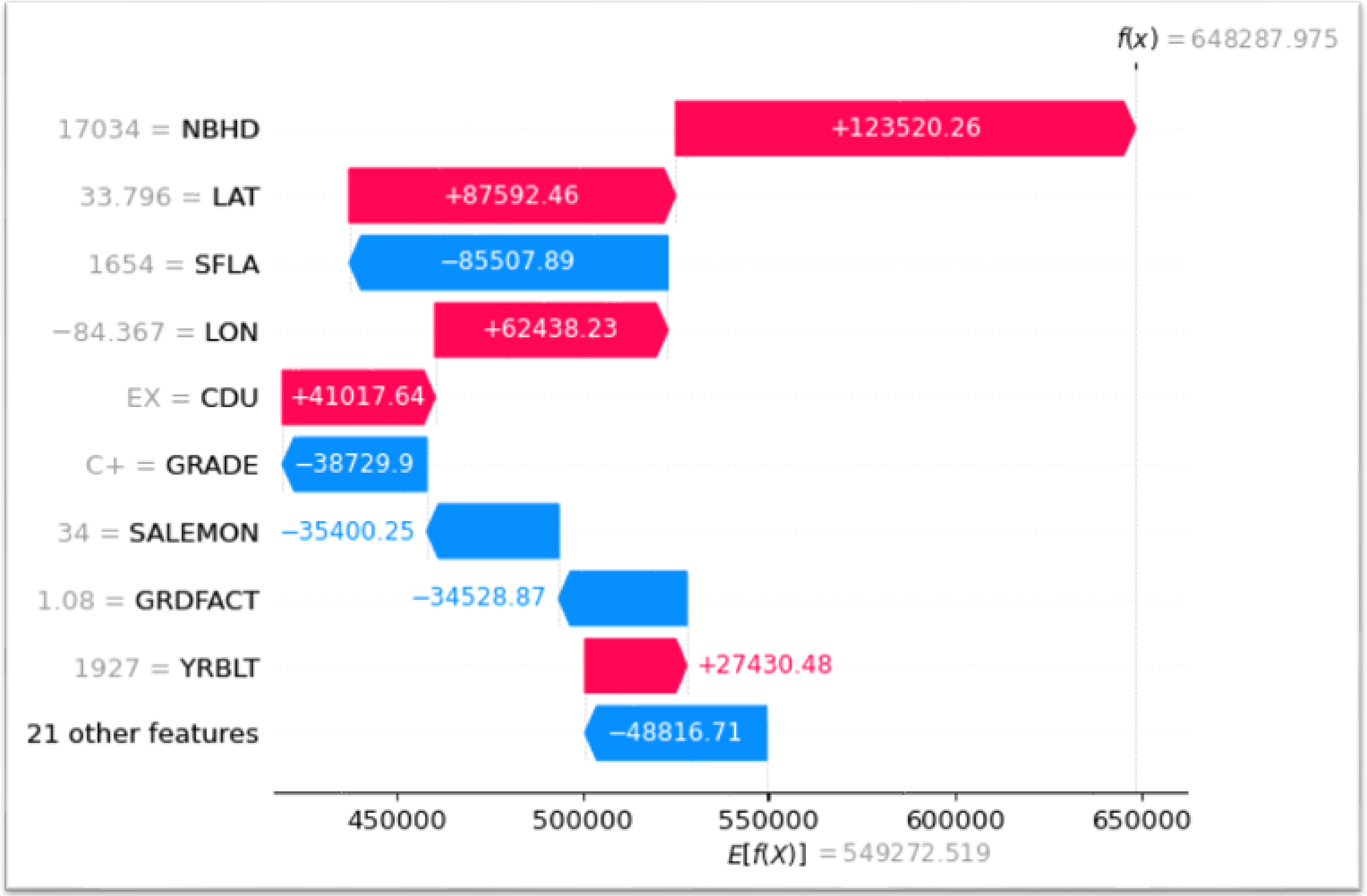
SHAP: Global Explanation

Shapely values used as feature importance



Island: Providenciales
Country: Turks and Caicos Islands

SHAP: Local Explanation



Area: Fulton County, GA   Country: USA

Waterfall Plot

- Visualize the contribution (shapely values) of each feature on the property's prediction.

SHAP: Local Linear Explanation

Prediction: $416,621.994 Average Prediction: $429,403.613

Contribution of SFLA(SFLA =2160): $41489.46

Contribution of NBHD: $-12,507.60

Prediction = Average Prediction + SFLA + NBHD + ……

416621.994 = 429403.613 + 41489.46 + -12507.60 + ……

$$\text{Rate (Coeff.) of SFLA} = \frac{41489.46}{2160} = 19.2$$

416621.994 = 429403.613 + 19.2 x 2160 + -1 x 12507.60 + ……

$$y = \text{Intercept} + B_1 * X_1 + B_2 * X_2 + \ldots\ldots$$

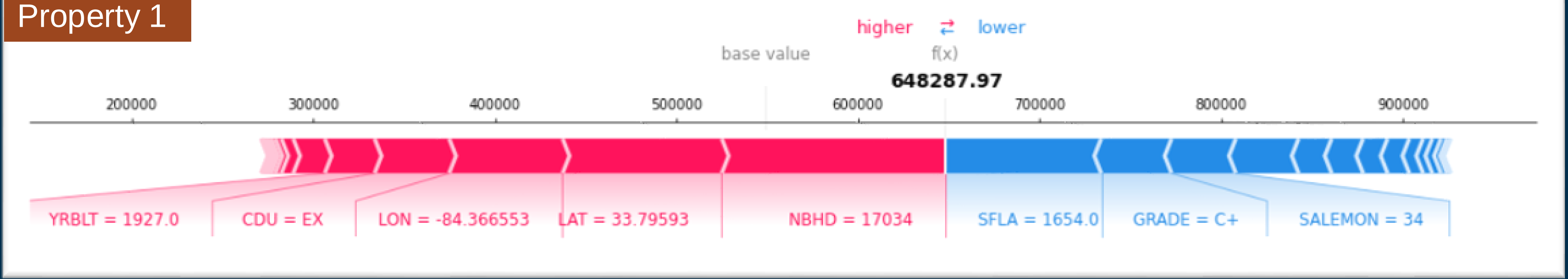
Area: Fulton County, GA Country: USA

SHAP: Local Explanation

Area: Fulton County, GA   Country: USA

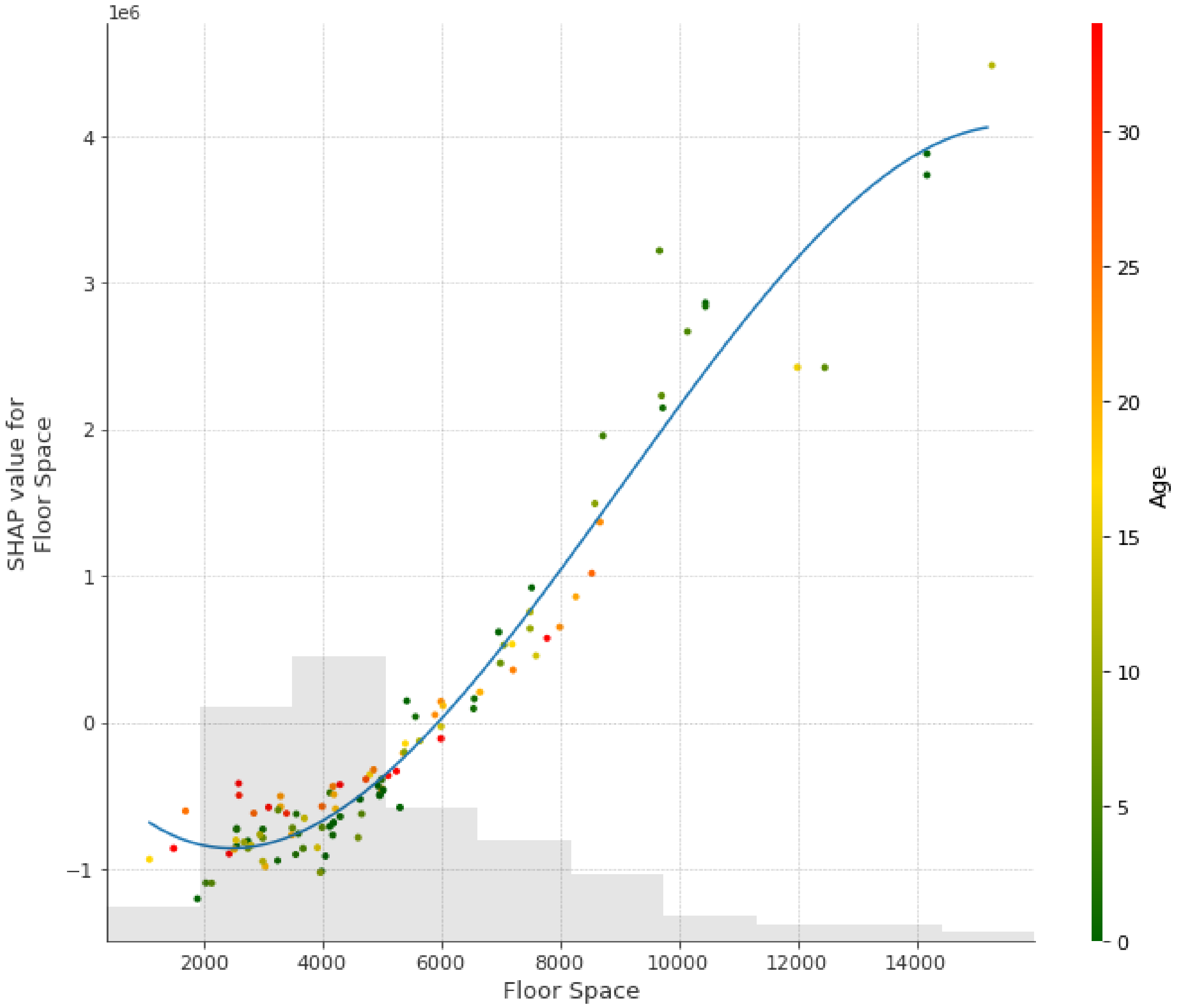
Force Plot

Property 1



Property 2

SHAP Dependency Plot: Floor Area



Island: Providenciales
Country: Turks and Caicos Islands

Observations that can be empirically tested

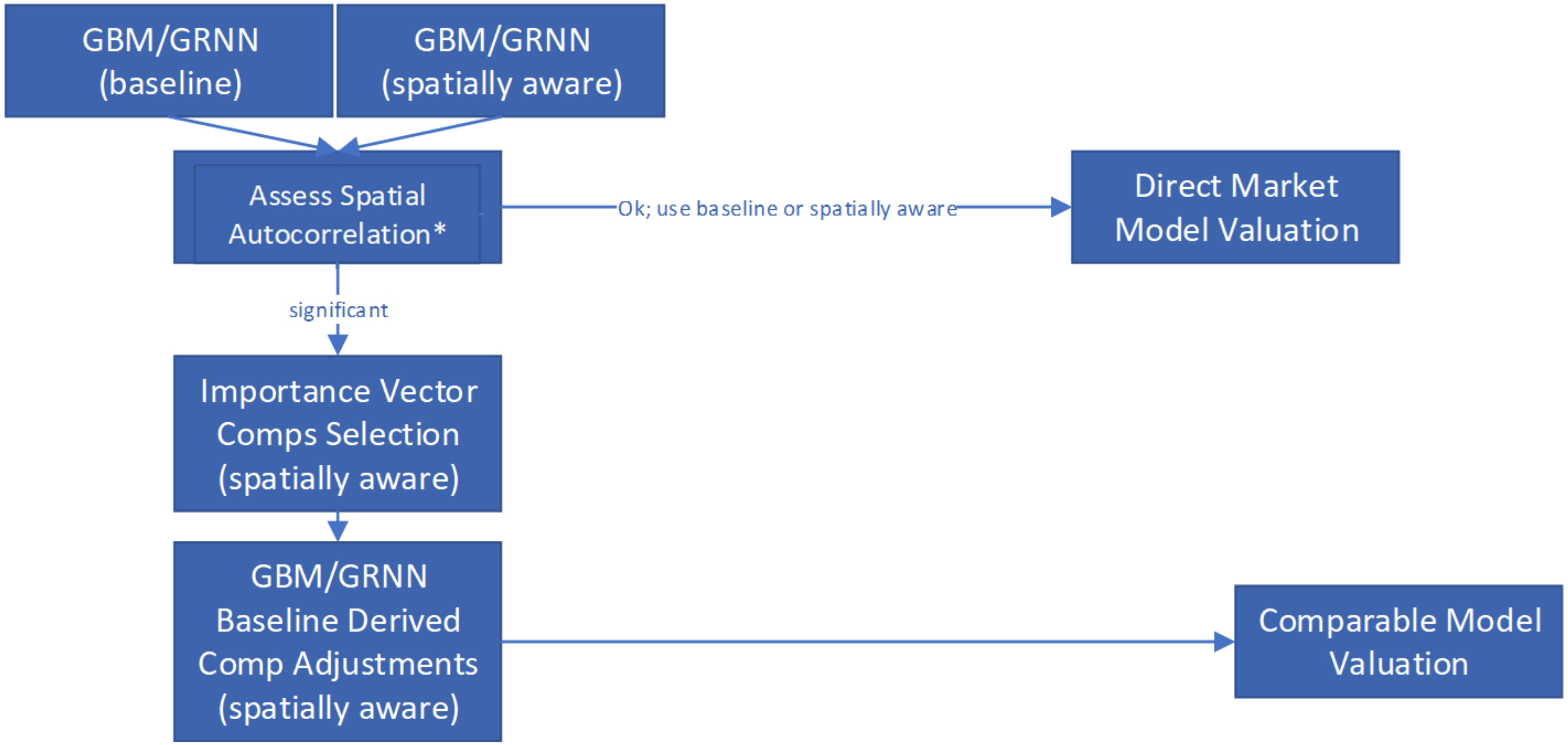
- Global importance from SHAP can be used as a “weights” in any attribute weight distance function (established in next section*)

- Averaging groups of individual SHAP values provide marginal values for the groups (and thus approximate “linear” coefficients)

- Similar properties should have similarly proportioned SHAP values – provides a similarity metric for comparable selection and inequity determination

AI-based Comparable Sales Models

ML Workflow – Comparable Market Model



NON-AI Comparable Sales Model

1
- Filter valid sales, remove outliers
- Segment market
- Time adjust sales

2
- Define adjustment function i.e. regression equation
- Define similarity function
- Choose Weights

3
- Calculate similarity
- Select most similar comps per subject
- Apply constraints / penalties

4
- Adjust the selected comps for subject
- Combine adjusted comps
- Estimate market value

Weights used in Traditional Comparable Approach

- MRA coefficients (betas)
- Empirical weights (manually tweaked subjective values)
- The value of weights depends on the scale of the input features
- Empirically picked weights needs a large amount of time, model specific, subjective

Variable Name	Weights-Variable	Weights-Constan	Subj Data	Comp Data	ABS Diff	[W*(diff)]^2
LANDVAL	0.0018		226,000	244000	18000	1050
SFLA	0.075		5,300	5000	300	506
AGEMAX	5		20	20	0	0
FIXTOT	10		22	22	0	0
GFACT	200		1.85	2.5	0.65	16900
SALEMON	5		0	24	24	14400
TOTGAR	0.09			0	0	0
FINBSMTOT	0.07				0	0
XCOORD	0.015		1250	3250	2000	900
YCOORD	0.015		1250	3250	2000	900
STORIES		50	1	1	0	0
NBHD		150	410	400	1	22500
NGROUP		100	1	1	0	0
STYLE		50	1	1	0	0
						57156
		Sum of Squares		57156.01		
		Distance Points		239		

AI/ML Importance Based Comps Approach - Select

- Similarity Measure
- Suppose there are N candidate properties and K attributes/features used for comps selection, the Euclidean distance between the i^{th} candidate property and the subject property:

$$D_i = \sqrt{\sum_{j=1}^{K} \frac{W_j}{\sum W_j} \begin{Bmatrix} \left(\frac{X_{ij}-X_{sj}}{s_j}\right)^2 & X_{ij}\ is\ numerical \\ \begin{matrix} 1 & X_{ij} == X_{sj} \\ 0 & X_{ij} <> X_{sj} \end{matrix} & X_{ij}\ is\ categorical \end{Bmatrix}}$$

i=1,2.3 …. N

j=1,2.3 …. K

D_i : Weighted Standardized Euclidean Distance

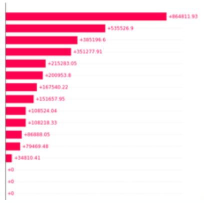
W_j : j^{th} attribute feature importance weight

X_{ij} : the value of the j^{th} attribute of the i^{th} property

X_{sj} : the value of the j^{th} attribute of the subject property

s_j : standard deviation of j^{th} attribute

- **Feature importance values of each attribute are used as the weights: W_j <=**



AI/ML Importance Based Comps Approach - Value

- Estimate subject market value

Based on the previous Similarity Measure, select top ***n*** candidate properties as comparable sales to estimate a subject market value

$$ESP = GBM_{subj} + \underbrace{w_1\, Resid_1 + w_2\, Resid_2 + \ldots + w_n\, Resid_n}_{\text{Weighted GBM adjustments}}$$

n >= 3 and n <= 5 in practice

ESP : Estimated Subject Price

GBM_{subj} : GBM model prediction for the subject

$$Resid_n = SP_n - GBM_n$$

SP_n : Sale price of the nth comparable property

GBM_n : GBM predicted price for the nth comparable property

w_n : Inverse distance weight

**Any estimator can be substituted for GBM in formula

[Reference: Improving Mass Appraisal Valuation Models Using Spatio-Temporal Methods by Richard A. Borst, PhD]

AI Based Comparable Sales Selection Test

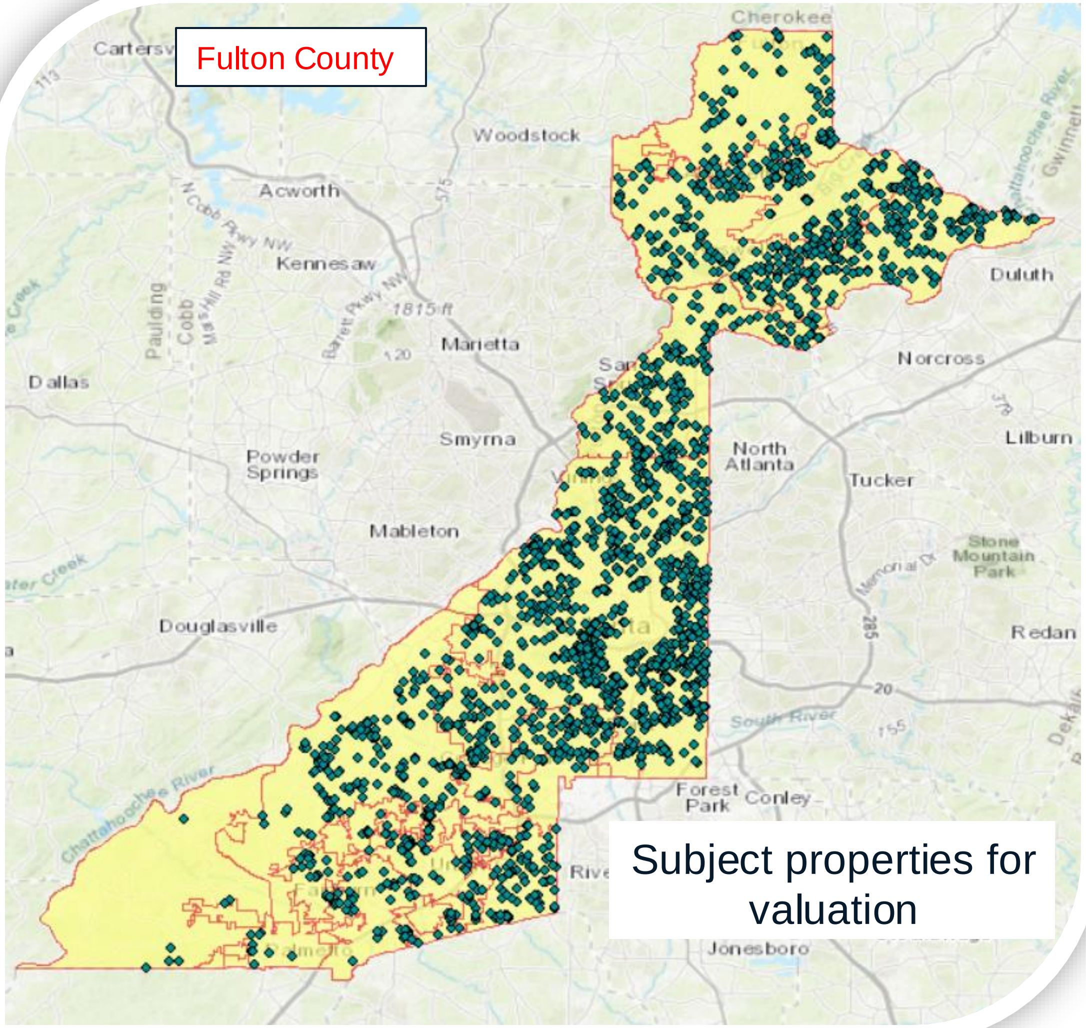
- Valuation Date: Jan 1, 2020
- Issue:

No sales really sold on Jan 1, 2020, therefore no true sales prices for calculating valuation accuracy.

- Solution:

Randomly selected 2,272 residential single-family sales in Dec 2019, Jan 2020 and Feb 2020, use their sale prices as the true prices (approximately).

- Set SALEMON of all subjects to 0



AI Based Comparable Model Example

- Market Price Estimation - Weighted GBM Adjustment - Valuation Date: Jan 1, 2020

No. of Subjects: 2,273 (Since no sales really sold on Jan 1, 2020, randomly selected 2,272 residential single-family sales in Dec 2019, Jan 2020 and Feb 2020, and used their sale prices as the ground truth proxy prices).

Comps from	AVG Price	Median Price	R2	RMSE	MAE	MAPE	RRSE	RAE	COV
GBM	445,701	330,000	0.91	113,334	42,104	10.28	0.29	0.16	25.62
Permutation			0.91	113,092	41,734	10.30	0.29	0.16	25.57
SHAP			0.91	112,798	41,909	10.38	0.29	0.16	25.51
Trad. CAMA			0.85	148,156	66,467	15.36	0.38	0.26	33.49

Comps from	AVG Price	Median Price	Median Sales Ratio	Mean Sales Ratio	COD	PRD
GBM	445,701	330,000	0.998	0.977	10.288	1.005
Permutation			0.998	0.978	10.315	1.004
SHAP			0.997	0.977	10.398	1.005
Trad. CAMA			0.953	0.945	15.179	1.034

Fulton County, Georgia USA

GA Individual Subject and Comps Comparison

Subject Location

Subject sold on
1/27/2020
Price: **$713,000**

Comps selected using SHAP feature importance

Comps selected in CAMA/MRA

AI/ML Estimated Price: $760,164.47

Residual: - $47,164.46

MRA Estimated Price: $649,870.00

Residual: $63,130.00

Area: Fulton County, GA
Country: USA

GA Individual Subject and Comps Comparison

	SUBJECT	COMP1	COMP2	COMP3	COMP4	COMP5	CAMA_COMP1	CAMA_COMP2	CAMA_COMP3	CAMA_COMP4	CAMA_COMP5
PARID	14 004300060160	14 004300060160	14 004400071034	14 004400070697	14 004300020982	14 004400040864	14 004300030155	14 004400070697	14 004300030650	14 004300060277	14 004400040864
PRICE	713000	669000	625000	585000	515000	500000	590000	585000	550000	570000	500000
SALEDT	1/27/2020	1/12/2018	11/15/2019	11/22/2019	7/12/2019	10/19/2018	7/15/2019	11/22/2019	6/6/2019	8/8/2019	10/19/2018
ADRSTR	CHEROKEE	CHEROKEE	MILLEDGE	GRANT	PAVILION	GRANT	BASS	GRANT	GRANT PARK	ORMOND	GRANT
NBHD	14269	14269	14269	14269	14269	14269	14269	14269	14269	14269	14269
STYLE	1	1	1	1	1	1	8	1	8	8	1
ZONING	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5	R5
GRADE	A-	A-	B+	C+	B+	B+	B+	C+	B+	B+	B+
CDU	EX	EX	VG	EX	VG	GD	EX	EX	VG	VG	GD
LOCATION	6	6	6	6	6	6	6	6	6	6	6
BSMT	3	3	2	3	2	3	2	3	2	2	3
HEAT	4	4	4	4	4	4	4	4	4	4	4
FUEL	1	1	1	1	1	1	1	1	1	1	1
FRONTING	9	9	9	9	9	9	9	9	9	9	9
EXTWALL	6	6	1	1	1	1	1	1	1	1	1
PARKTYPE	3	3	3	3	3	3	3	3	2	3	3
SFLA	1768	1768	1479	1794	1900	1747	1828	1794	1786	1894	1747
GRDFACT	1.45	1.45	1.35	1.08	1.35	1.35	1.35	1.08	1.35	1.35	1.35
CALCACRES	0.3122	0.3122	0.1221	0.1722	0.1339	0.1879	0.1531	0.1722	0.1066	0.1825	0.1879
STORIES	1	1	1	1	1	1	2	1	2	2	1
YRBLT	1920	1920	1920	1920	1920	1920	1997	1920	2003	1998	1920
RMTOT	7	7	6	6	5	7	4	6	8	8	7
RMBED	3	3	3	3	3	3	3	3	3	3	3
RMFAM	0	0	0	0	0	0	1	0	1	1	0
FIXBATH	2	2	2	2	2	2	2	2	2	2	2
FIXHALF	1	1	0	0	0	1	1	0	1	1	1
FIXADDL	2	2	2	4	2	3	4	4	5	4	3
FIXTOT	10	10	8	10	8	11	12	10	13	12	11
DEPR	100	100	98	100	90	96	100	100	98	98	96
LAT	33.732122	33.732122	33.739125	33.739325	33.735825	33.742222	33.735001	33.739325	33.732873	33.731723	33.742222
LON	-84.37406	-84.37406	-84.374322	-84.376634	-84.374146	-84.376834	-84.377473	-84.376634	-84.3773	-84.375305	-84.376834
SALEMON	0	23	1	1	5	14	5	1	6	4	14
PARKQUANIT	2	2	2	2	2	2	2	2	2	2	2

Fulton County Georgia USA

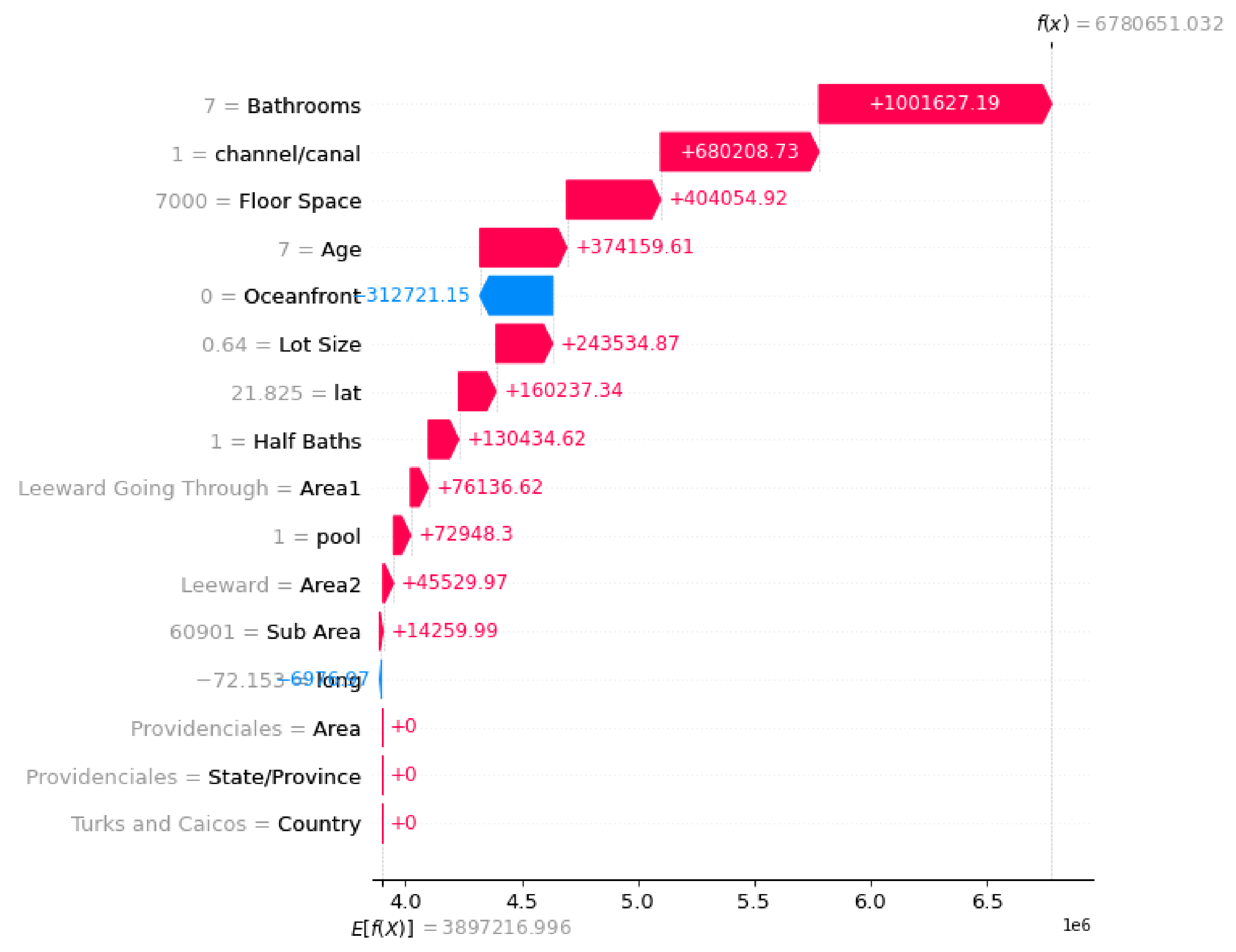
TC Example #62

MLS: 2300737
PRICE: $6900000
GBM Prediction: $6780651
COMPS Estimated Price: $6736257
COMPS Residual: $163742

	SUBJECT	COMP1	COMP2	COMP3
MLS#	2300737	2300337	2300750	2300732
PRICE	6900000	5900000	13500000	7995000
Area	Providenciales	Providenciales	Providenciales	Providenciales
Area1	Leeward Going Through	Leeward Going Through	Leeward Going Through	Leeward Going Through
Area2	Leeward	Leeward	Leeward	Leeward
State/Province	Providenciales	Providenciales	Providenciales	Providenciales
Country	Turks and Caicos	Turks and Caicos	Turks and Caicos	Turks and Caicos
Sub Area	60901	60902	60903	60902
Bathrooms	7	6	6	5
Half Baths	1	0	1	1
Floor Space	7000	6000	9678	10150
Lot Size	0.64	1.02	0.7	0.84
Age	7	14	5	5
lat	21.82469576	21.82084348	21.81484993	21.81938766
long	-72.15332717	-72.15065955	-72.14379561	-72.15332766
channel/canal	1	1	1	1
pool	1	1	1	1
Oceanfront	0	0	0	0

Island: Providenciales
Country: Turks and Caicos Islands

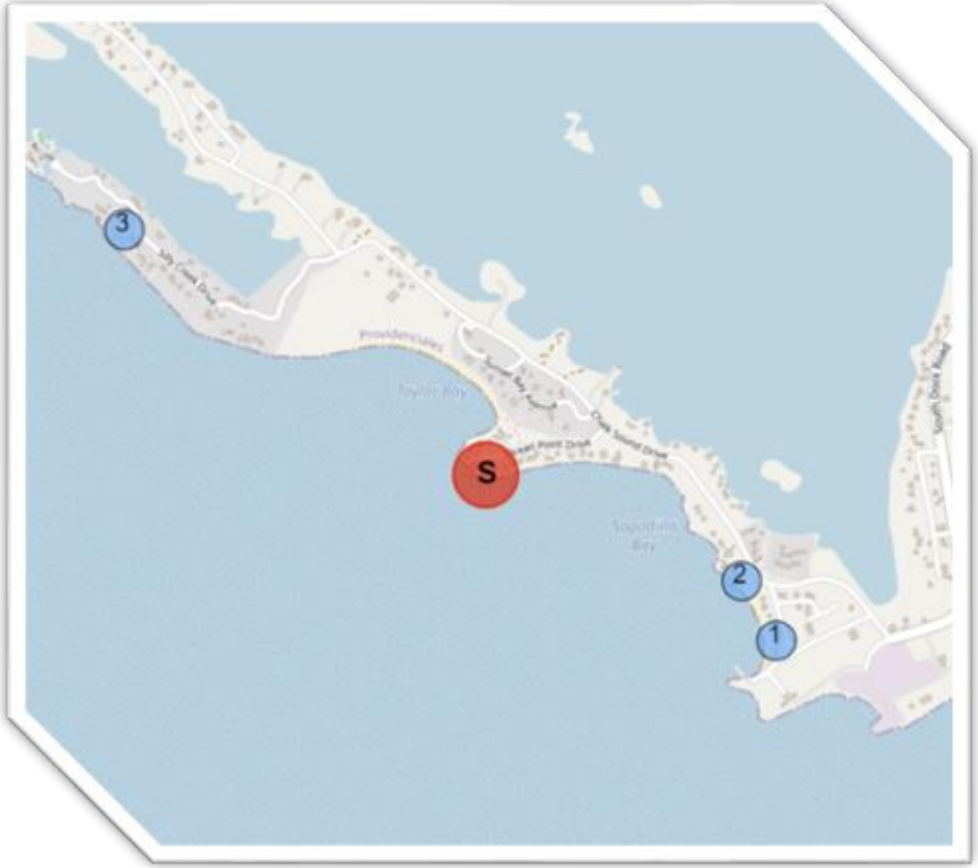
TC Example #62 (Waterfall)



TC Example #101

MLS: 2300188
PRICE: $5750000
GBM Prediction: $5789056
COMPS Estimated Price: $5665319
COMPS Residual: $84680

	SUBJECT	COMP1	COMP2	COMP3
MLS	2300188	2300731	2100323	2300728
PRICE	5750000	6950000	5999999	5850000
Area	Providenciales	Providenciales	Providenciales	Providenciales
Area1	Chalk Sound	Chalk Sound	Chalk Sound	Chalk Sound
Area2	Sapodilla Bay	Sapodilla Bay	Sapodilla Bay	Chalk Sound
State/Province	Providenciales	Providenciales	Providenciales	Providenciales
Country	Turks and Caicos	Turks and Caicos	Turks and Caicos	Turks and Caicos
Sub Area	60612	60612	60612	60400
Bathrooms	6	5	7	5
Half Baths	0	1	1	1
Floor Space	7200	7498	8000	7060
Lot Size	0.69	0.51	0.79	0.88
Age	17	12	23	8
lat	21.74695565	21.74182254	21.7436802	21.75447766
long	-72.29327271	-72.28360085	-72.28477619	-72.30527695
channel/canal	0	0	0	0
pool	0	1	1	1
Oceanfront	1	1	1	1

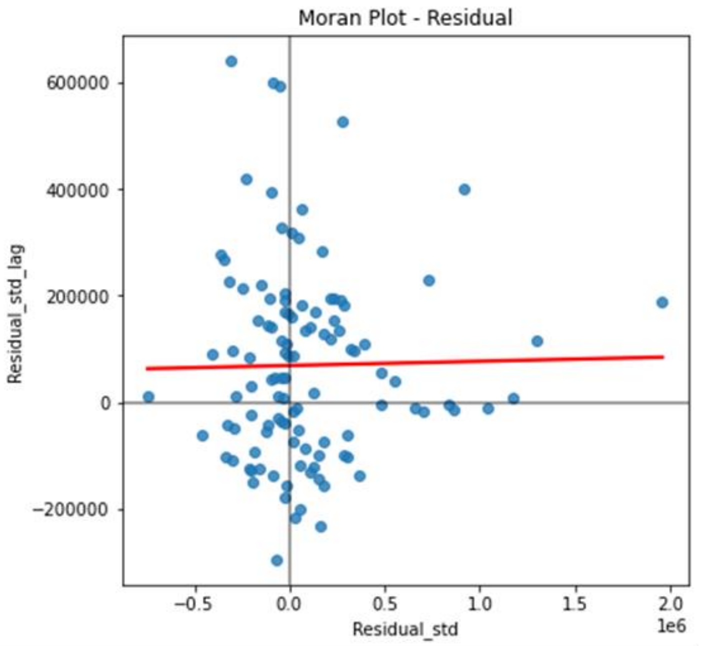


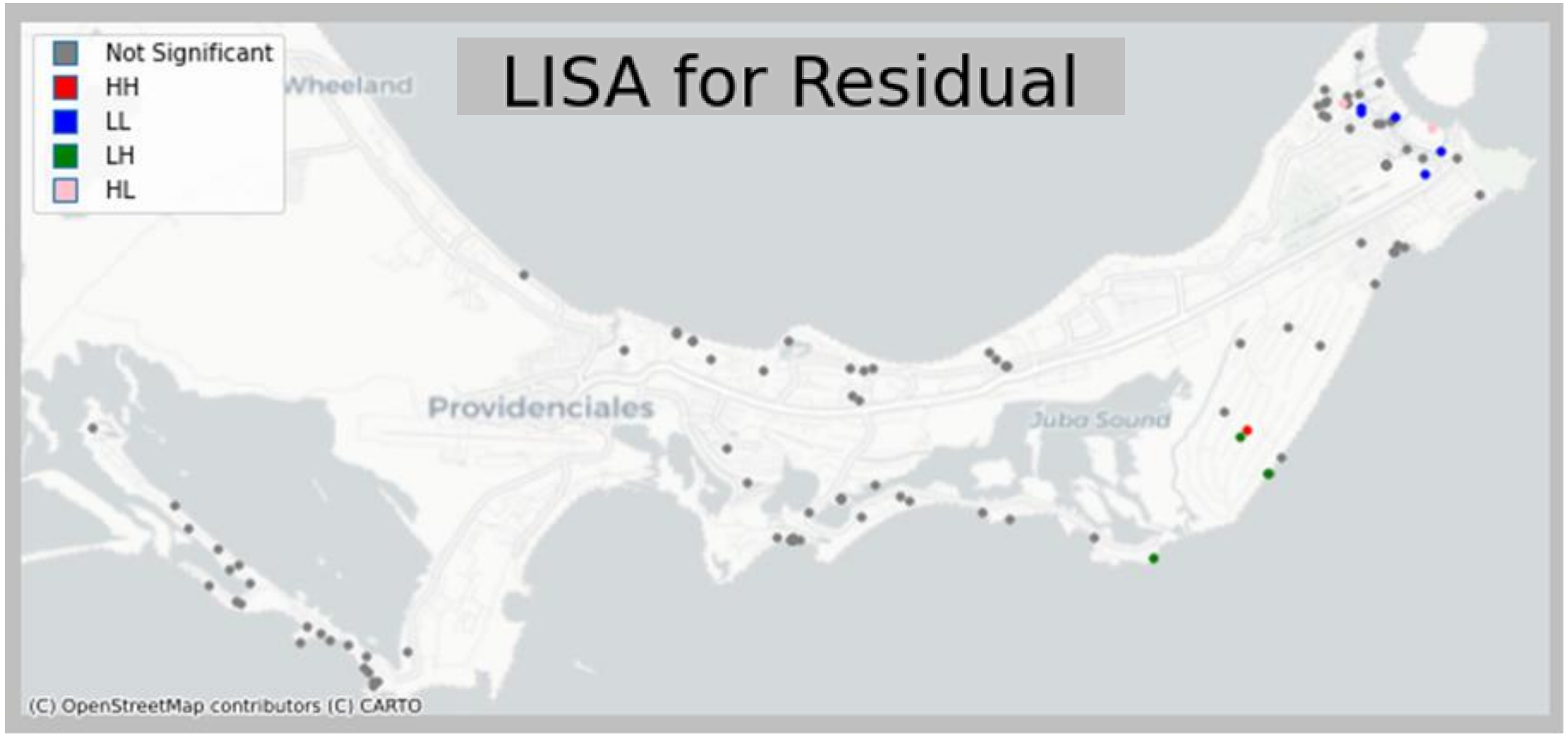
Island: Providenciales
Country: Turks and Caicos Islands

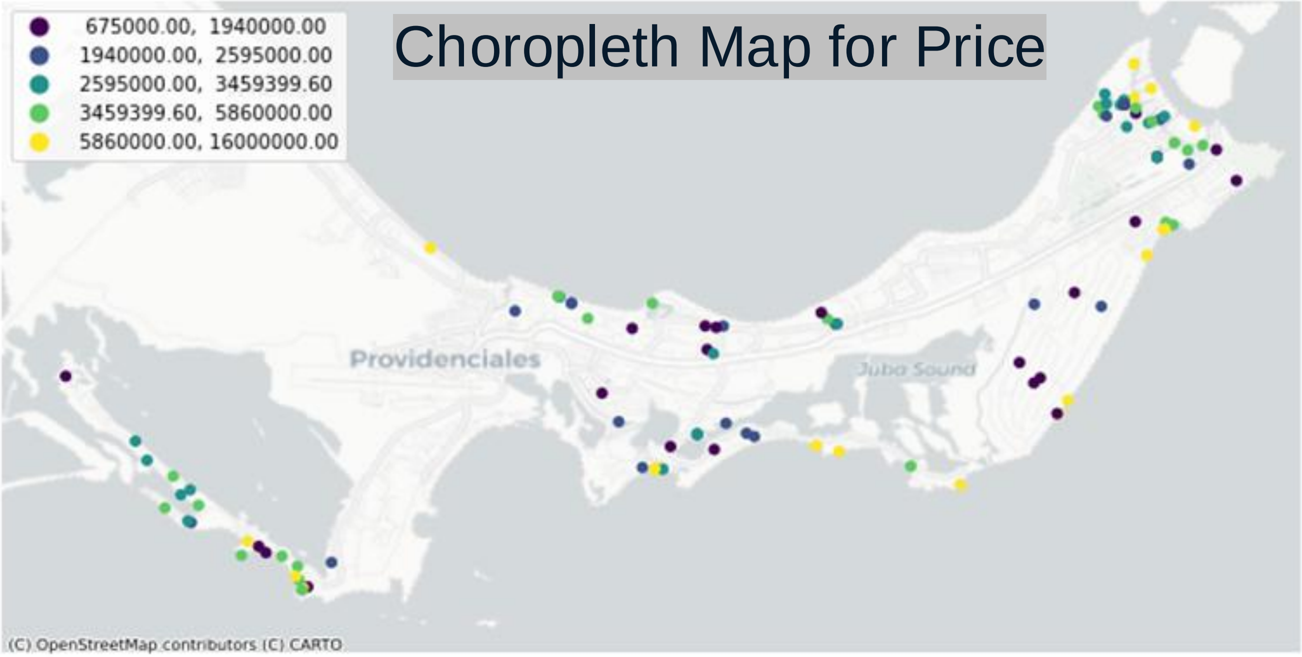
Direct Market Model Spatial Residual

- **very weak spatial autocorrelation**
- **not statistically significant**

Moran I: 0.008
p-value for Moran's I : 0.346
Not significant: 97/110 = 88.2%



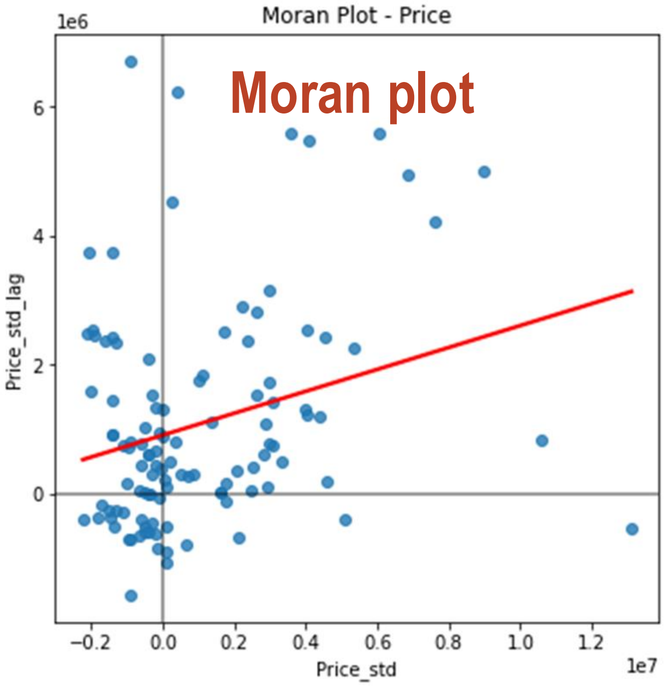




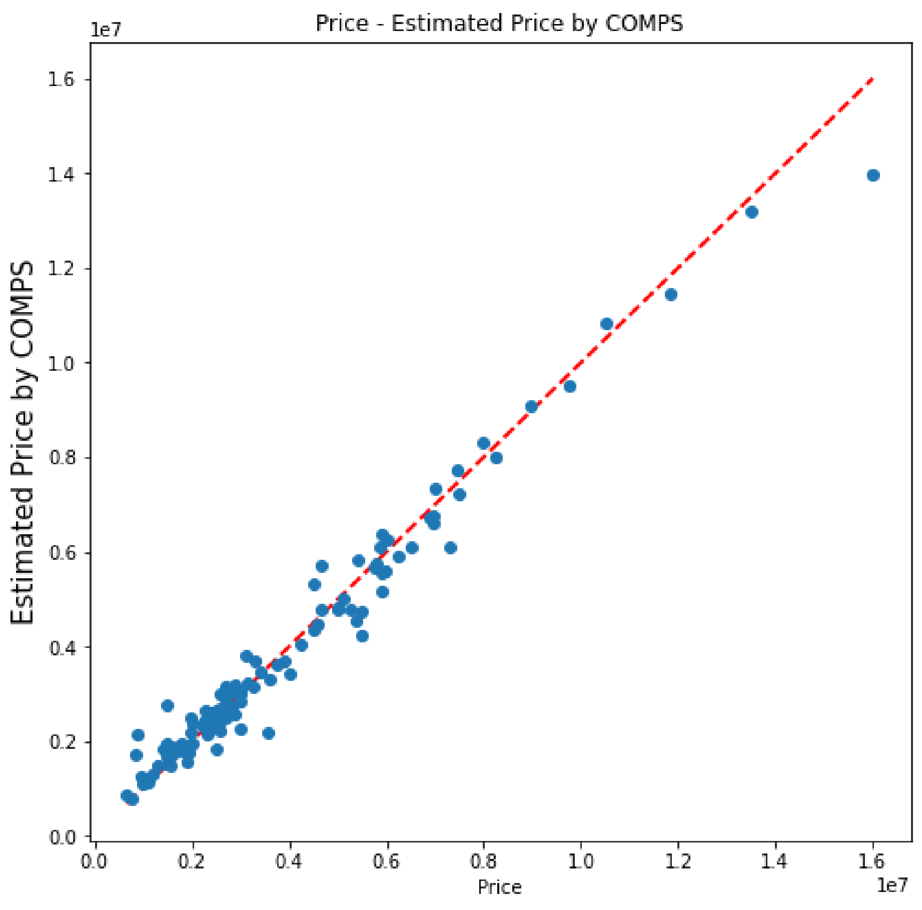
Price

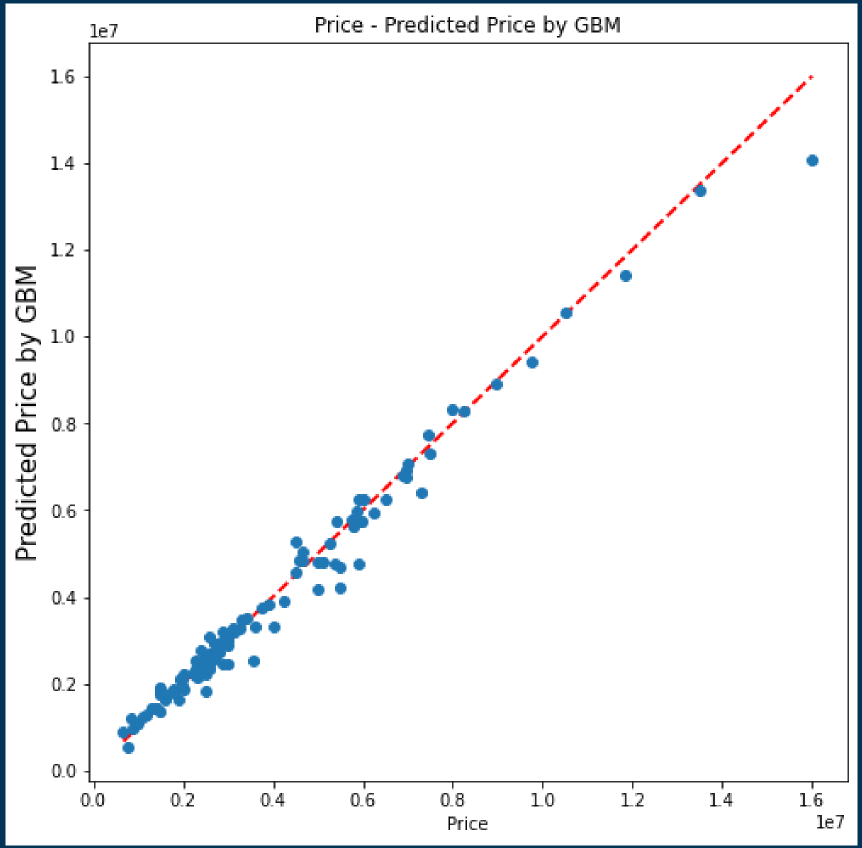
- weak positive spatial autocorrelation
- statistically significant

Moran I: 0.169
p-value for Moran's I : 0.006
Not significant: 93/110 = 84.5%



Prediction vs Price (Comp VS Direct)





Island: Providenciales
Country: Turks and Caicos Islands

Direct vs Comps Results

Comparable Sales Model / COMPs Error Metrics:

RMSE	MAE	MAPE	Mean Sales Ratio	Median Sales Ratio	COD	COV	PRD
466,886	325,762	11.87	1.046	1.003	11.84	13.02	1.04

Direct Market Model / GBM Error Metrics:

RMSE	MAE	MAPE	Mean Sales Ratio	Median Sales Ratio	COD	COV	PRD
382,922	248,934	8.00	1.009	1.006	7.92	10.68	1.02

Island: **Providenciales**
Country: **Turks and Caicos Islands**
Condo Listings
*TC Real Estate Association web data**

Sales data: 110 listings for 2023

Key Takeaways

1. AI/ML techniques are usable by today's practitioner; they can be for market data understanding by individual/fee assessors as well as with the mass appraisal approach.
2. Usage of AI/ML simplifies Direct Market Model and Comparable Sales Models.
3. Usage of SHAP to explain what ML models have learned about the market.
4. Can be implemented with Open-Source Python libs (can be integrated w/Esri).
5. Turnkey Incorporation into Commercial CAMA Systems (Tyler Technologies) – Greatly simplified & repeatable process

Joe.Wehrli@tylertech.com

linkedin.com/in/wehrli/

***

CO-AUTHOR in R&D and technical approaches:

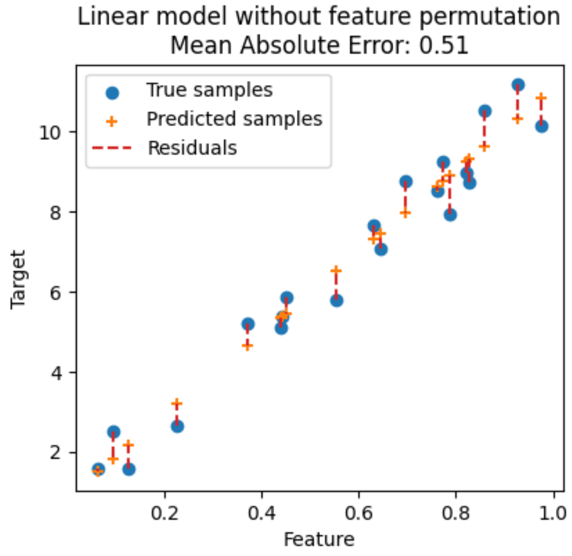
Langyue Wang (Larry)
larry.wang@tylertech.com

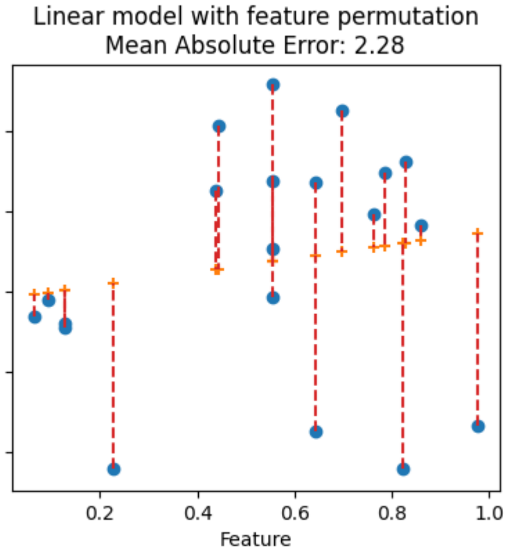
INTERNATIONAL RESEARCH SYMPOSIUM

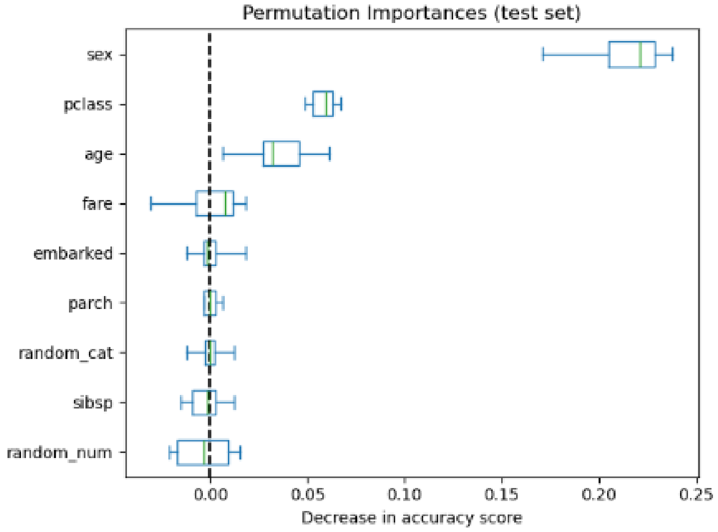
Amsterdam, The Netherlands · December 4 - 5, 2024

Permutation Feature Importance

- Permutation feature importance measures the degradation of the model's score after randomly shuffling the values of a single feature
- A feature is "important" if shuffling its values increases the model error, because the model relied on the feature for the prediction
- Permutation feature importance is **model-agnostic.**







(Ref: https://scikit-learn.org/stable/modules/permutation_importance.html)

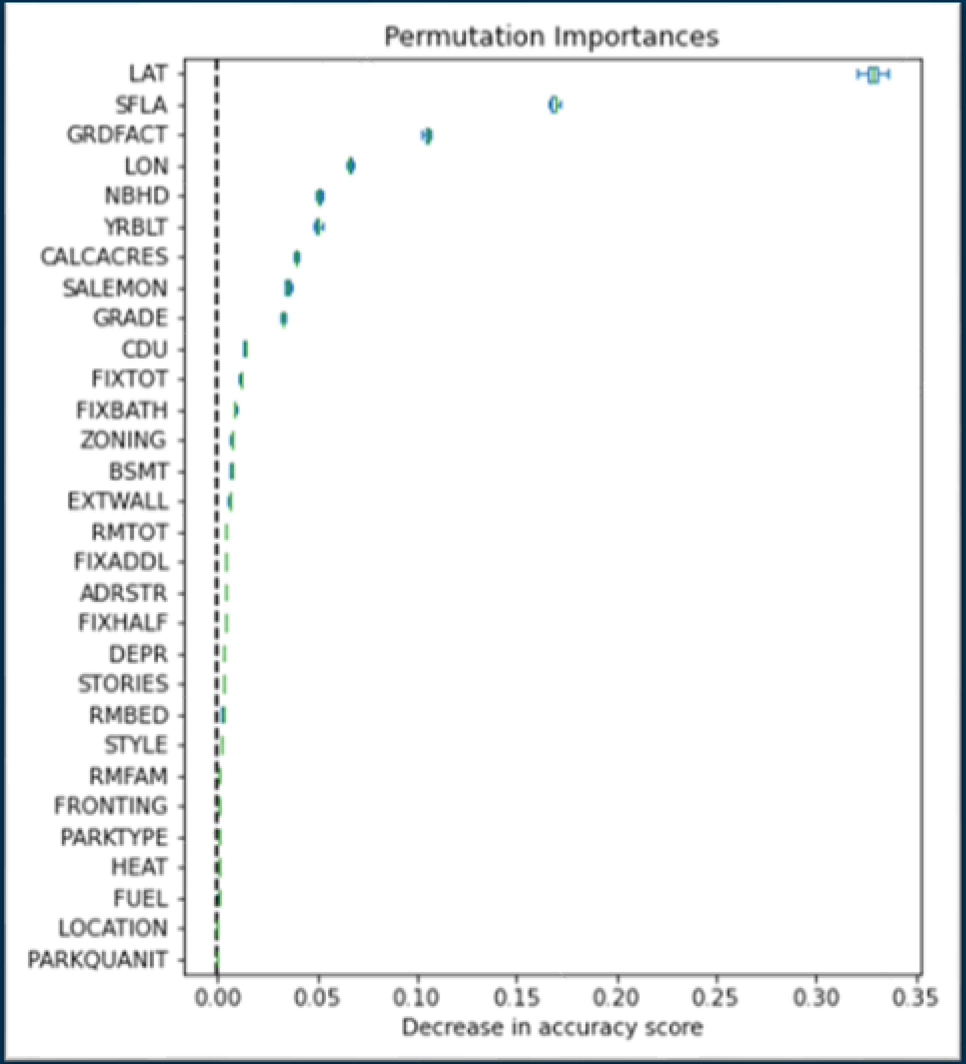
Permutation Feature Importance Example

- Use sklearn.inspection.permutation_importance function
- Use the same previous GBM model as the estimator
- Parameters:

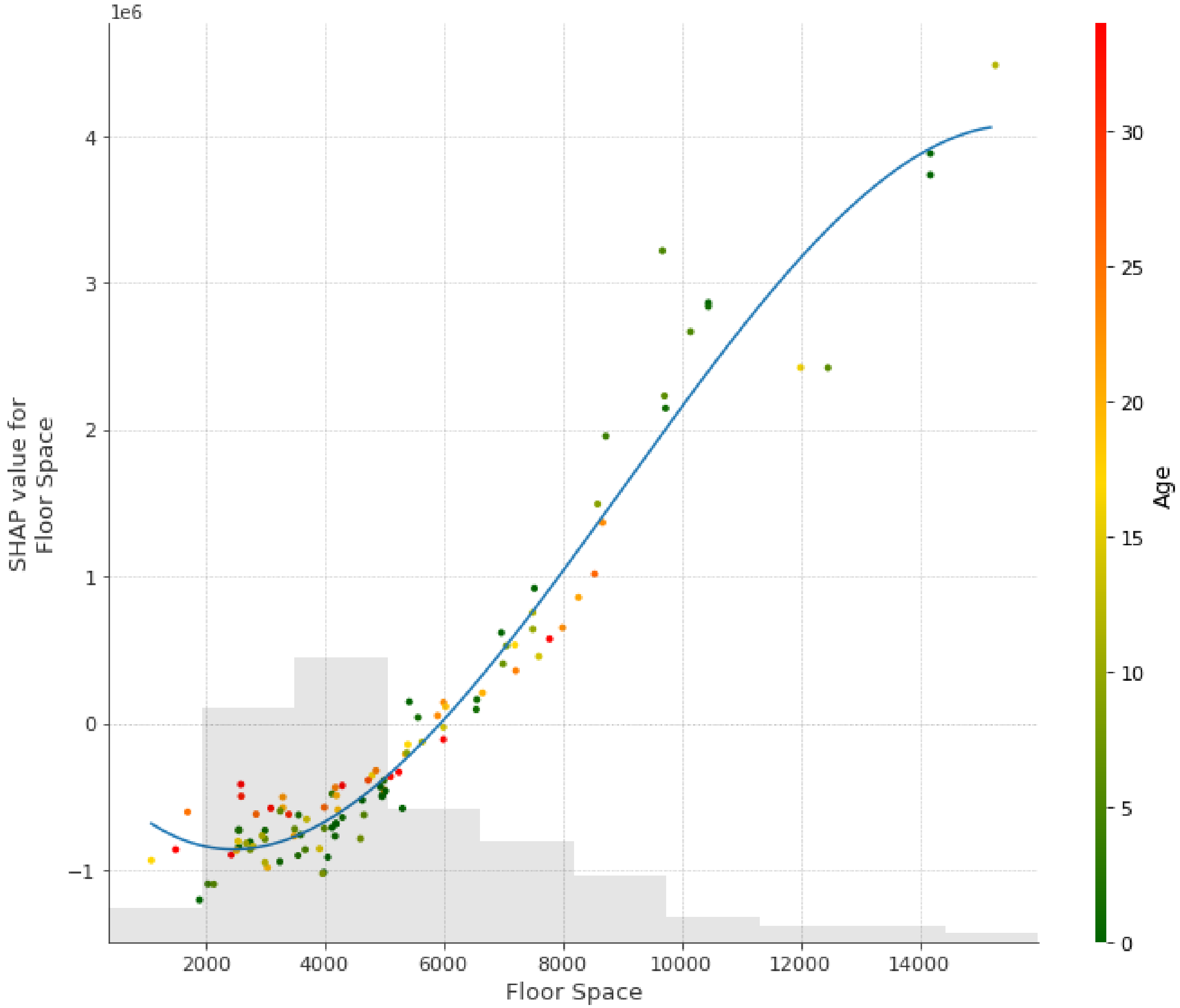
  n_repeats: 10

  (number of time to permute a feature)

  Others: default

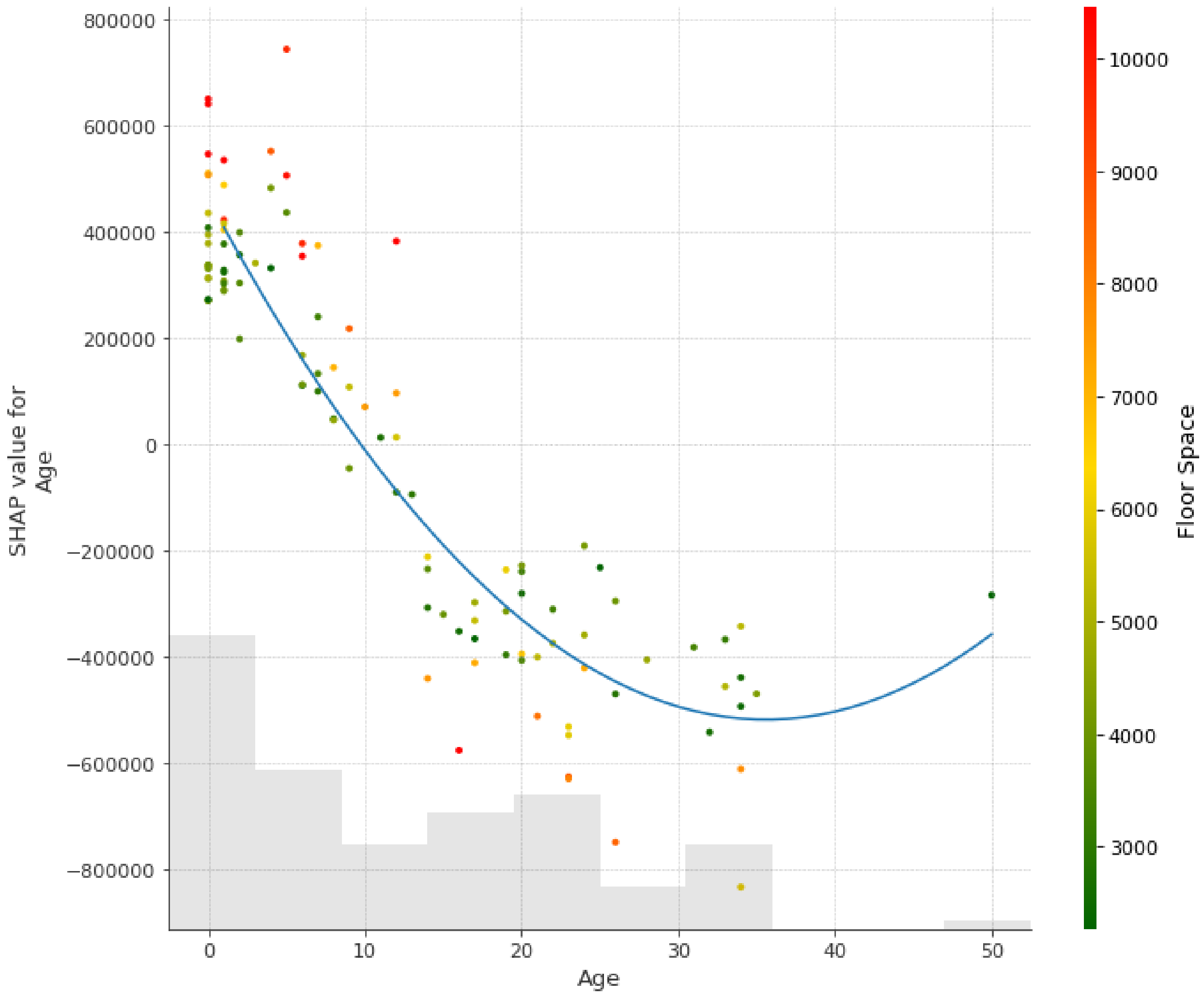


Floor Area



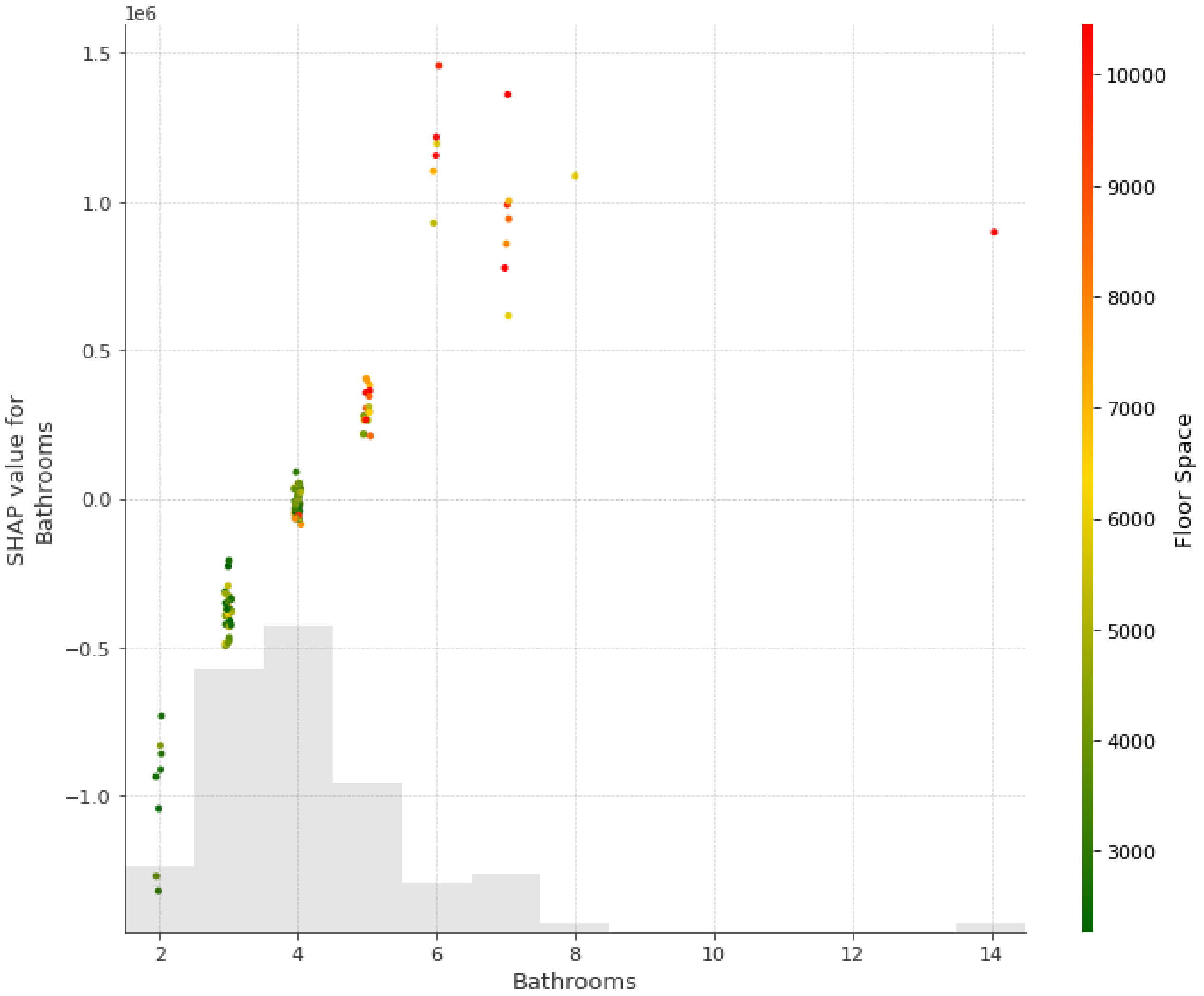
Island: Providenciales
Country: Turks and Caicos Islands

Age



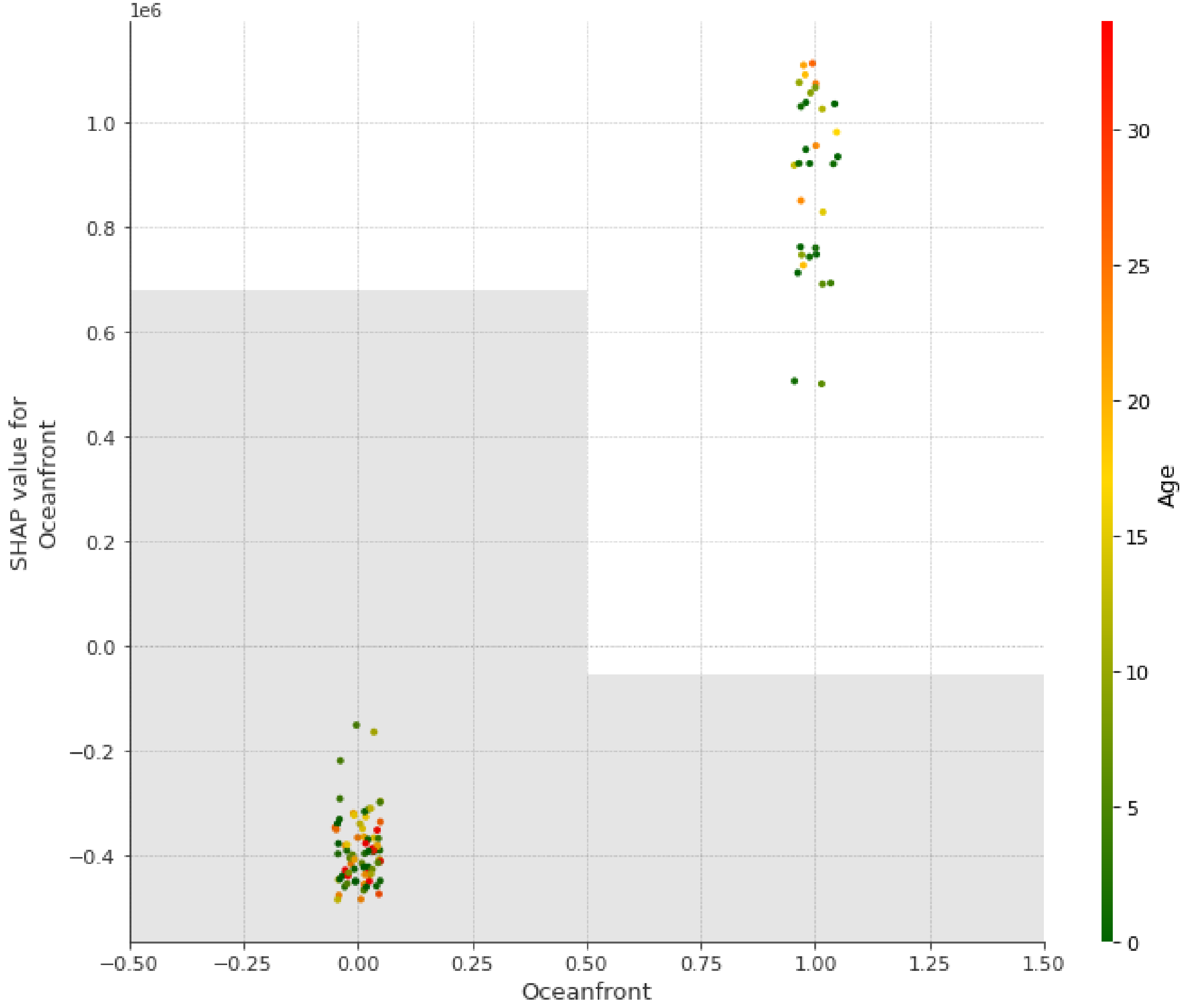
Island: Providenciales
Country: Turks and Caicos Islands

Baths



Island: Providenciales
Country: Turks and Caicos Islands

OceanFront



Island: Providenciales
Country: Turks and Caicos Islands