MACHINE LEARNING FOR HOUSING ANALYSIS

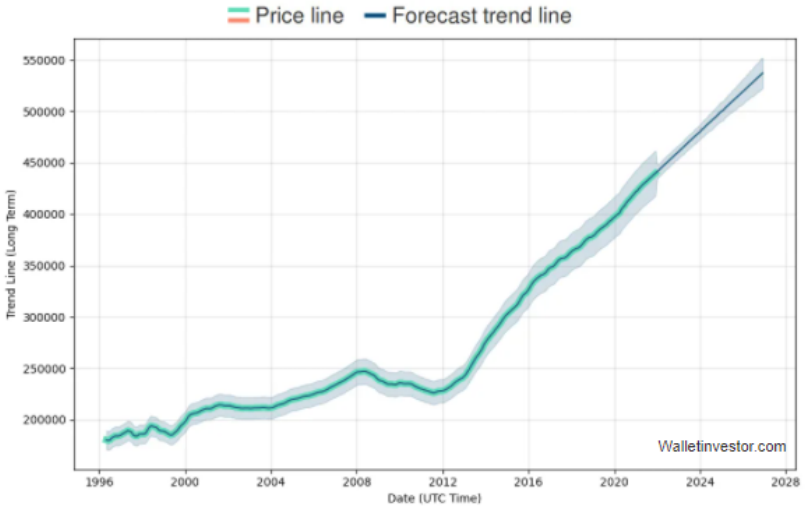
Agenda

- Introduction
  - What are the factors affecting housing price?
  - How do people forecast housing price?
  - What are limitations and challenges?
- Empirical cases
  - Case in the City of Austin
- Lab time

Austin City (Texas State) Forecast Chart, **Long-Term**

Predictions for Next Months and Years: 2021-2027

The long-term housing forecast is based on all the available median listing price recorded up to today.



https://walletinvestor.com/real-estate-forecast/tx/travis/austin-housing-market

Introduction

There are many factors that impact real estate prices, availability, and investment potential.

- Demographics
- Interest rates
- Macroeconomics factor
- Government policies
- Location
- Land prices
- etc.

KEY TAKEAWAYS

- There are a number of factors that impact real estate prices, availability, and investment potential.
- Demographics provide information on the age, income, and regional preferences of actual or potential buyers, what percentage of buyers are retirees, and what percentage might buy a vacation or second home.
- Interest rates impact the price and demand of real estate—lower rates bring in more buyers, reflecting the lower cost of getting a mortgage, but also expand the demand for real estate, which can then drive up prices.
- Real estate prices often follow the cycles of the economy, but investors can mitigate this risk by buying REITs or other diversified holdings that are either not tied to economic cycles or that can withstand downturns.
- Government policies and legislation, including tax incentives, deductions, and subsidies can boost or hinder demand for real estate.

https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp

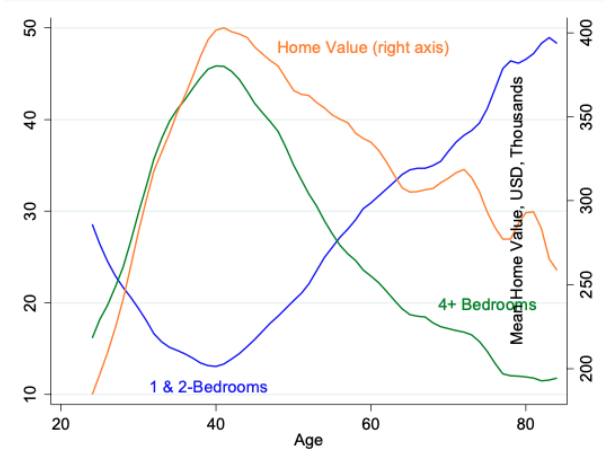
Introduction

Demographics

- Demographics are data describing the composition of the population, such as age, gender, income, migration patterns, and population growth.
- These statistics are an important factor, which affects the pricing method of real estate and the type of real estate demanded.

For example

- Baby boomers from 1945-1964
- Low demand in big houses when children moving out

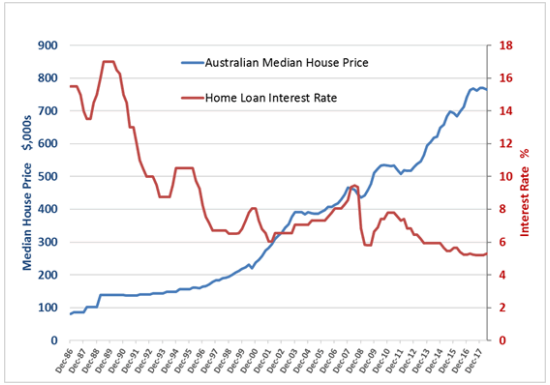


(Bolhuis and Cramer, 2020)

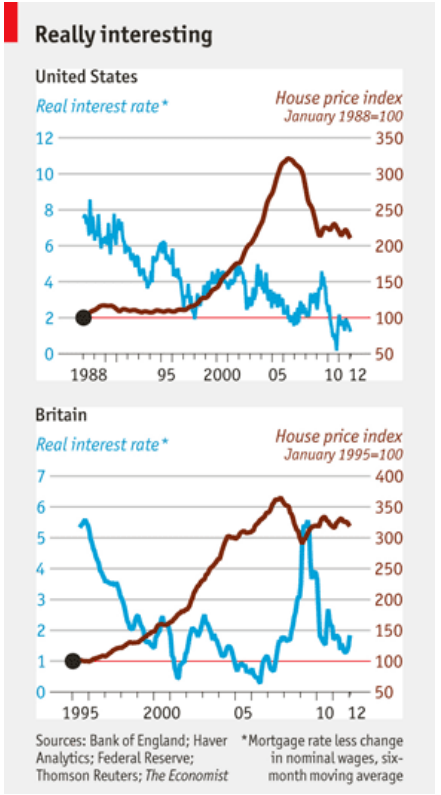
Introduction

Interest rates

- Interest rates also have a significant impact on the real estate market.
- The lower the interest rate, the lower the cost of a mortgage to buy a house, which will increase the demand for real estate and push up prices again.



(Buckingham, 2018)



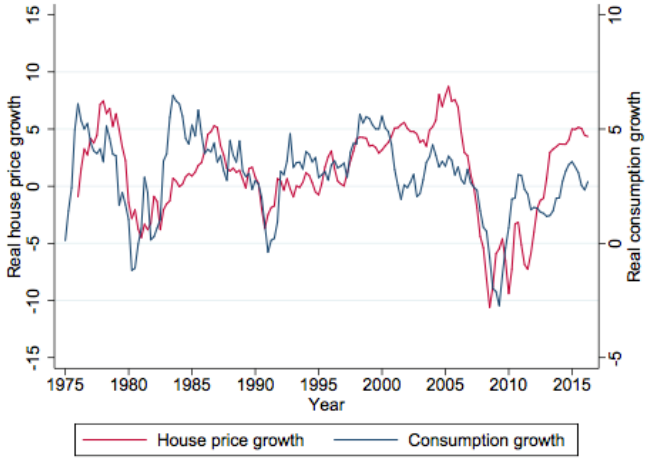
(Buttonwood, 2012)

Introduction

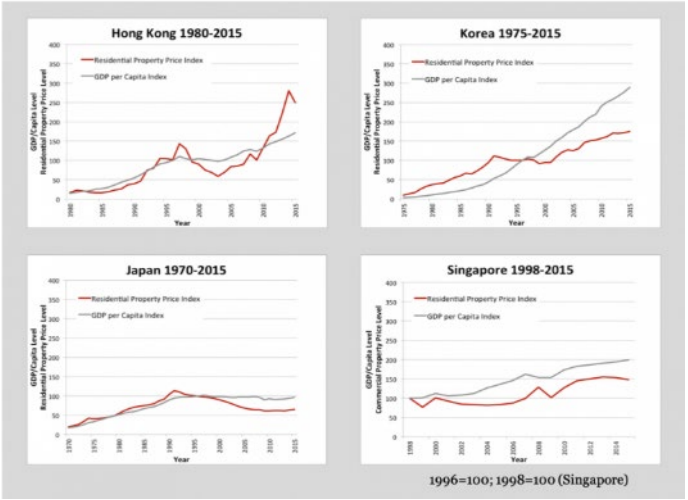
Macroeconomics factor

- Another key factor affecting the value of real estate is the economy. This is usually measured by economic indicators, such as GDP, employment data, manufacturing activity, commodity prices, etc.

(Cloyne et al., 2018)



(Asian Green Real Estate, n.d.)



Introduction

Government policies

- Legislations (property tax, deductions, credits, subsidies etc.) are factors that can have a sizable impact on property demand and prices.

For example

- In 2009, the U.S. government introduced a first-time homebuyer's tax credit

https://www.youtube.com/watch?v=HjjJqFBmVEk

Introduction

Location

- Location is an essential factor in predicting property values.
- Distance to downtown, grocery stores, hospitals, open spaces are common criteria in measuring the location factor.

https://www.pinterest.com/pin/483433341230595738/

Introduction

How do people forecast housing price?

Spatial economic theory tells several location factors should be the outcome of trade-offs between different levels of accessibility in location choices and building stock development decisions.

The value of land and buildings should reflect these trade-offs and the competition among agents to locate at points of greater or lesser transportation advantage (Alonso, 1964). Despite this theory, and despite decades of analysis and modeling activity, there generally lacks robust relationships between accessibility and land value that can be confidently used in policy analysis.

This is a particularly important gap for transit infrastructure investment decisions, whose benefit-cost evaluations often critically hinge on the land development and land value increases expected from such major investments.

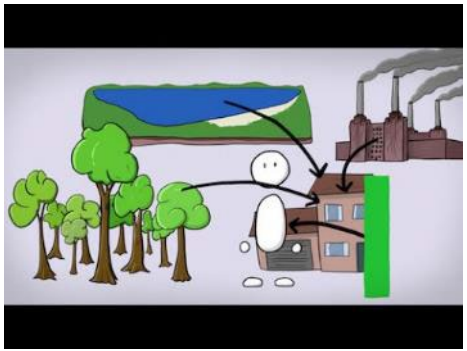
Introduction

How do people forecast housing price?

In the 1960s, economists developed the theory for determining property location in the urban land market (Alonso, 1964). The theory illustrates a model in which a household chooses to locate at a point where its bid–rent curve intersects with the actual one, in which the bid–rent curves have a declining gradient with the distance from the residential location to the central business district (CBD).

However, it might be necessary to consider the effect of other variables. The introduction of the hedonic pricing methodology by Rosen (1974) led to a way of attributing value to different properties' features.

A number of studies have observed the integration of physical, neighborhood and accessibility characteristics of the property in models trying to explain the differences in property values or house prices (Cervero and Duncan, 2002; Hess et al., 2005; Munoz-Raskin, 2007).



https://www.youtube.com/watch?v=LkXVCQam5kw

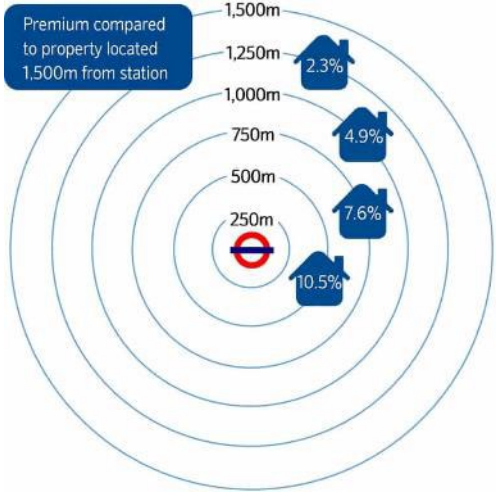
Introduction

How do people forecast housing price?

Most commonly, hedonic price models have used ordinary least squares (OLS) estimation (Haider and Miller, 2000; Tse and Chan, 2003), but more recently these models have been extended to incorporate spatial effects in multiple ways, feasible generalized least-square estimation (glm) and spatial econometric models (gwr).

There are several points of empirical evidence relating the changes in commercial and residential property market values and transport investment.

1. Brinkerhoff (2001) concludes that proximity to rail systems is valued by property owners, and there is little support that this proximity can decrease property values.
2. Martínez and Viegas (2009) mentioned that the impacts of accessibility are seen as positive, with some very large percentage increases.



https://www.theguardian.com/money/2014/aug/20/distance-from-station-value-of-house-nationwide-uk

Introduction

What are limitations and challenges?

- Marginal effects of these factors (Kestens et al., 2004)
- Nonlinearity (Yii et al., 2021)
- Multicollinearity (Raymond et al., 2015)
- Spatial autocorrelation (Wang el al., 2019)

How can we do?

- Wisely select dependent variables
- Test different models
- Considering spatial-related effects

Empirical studies

Research gaps:

- Traditional statical models and several machine learning approaches have been applied to predict residential property values, albeit most have focused one specific model without considering the economic/demographic conditions.

Data sources:

- Housing transaction data 2013-2019
- Sociodemographic data from census bureau
- Land use inventory from the City of Austin
- GTFS data from Capital Metropolitan Transportation Authority

Methods: SVM, xgbdt, RF, lasso, Ridege, ElasticNet

Outcome variables: single-family residential property value

Empirical studies

- Is there a significant spatial changes in “hotspots” of market value in Austin from 2015-2021?

Pilot test 1

Empirical studies

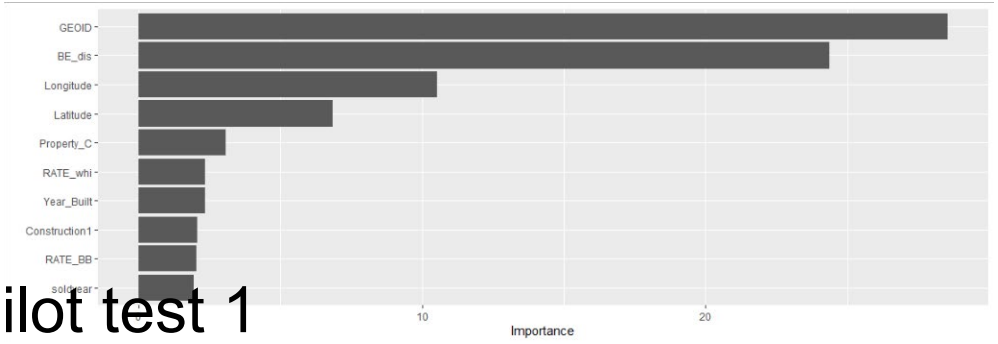
SALES model (obs=property location)

2014-2016 predict 2017

Focusing on sold price per square foot

Average error: $24.31/sqft

Method: XGBDT



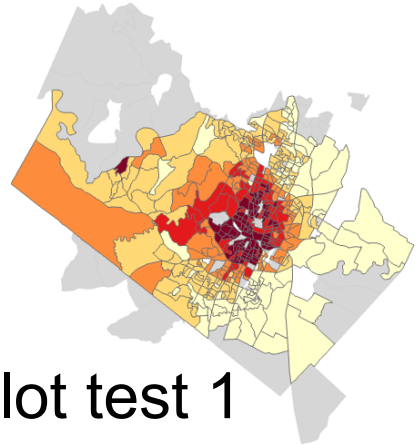
Pilot test 1

Dependent variable	
SPSqFt	Sold price per square foot
Timestamp	
soldyear	Which year sold (2014-2017)
Independent variable: location	
GEOID	GEOID (census block group)
Longitude	Longitude of housing
Latitude	Latitude of housing
Independent variable: housing status	
Year_Built	When the house built
Type	What type of sales (rescale, new, under construction, tear down, to be built)
X_Stories	Number of floors
CDOM	The gap between list and sold
Occupant_T	Occupation status (own, rent, vacant)
Property_C	Status of property (poor, fair, good, excellent)
Unit_Style	Remarks of property
Exterior_Features	Exterior features line 1
Exterior_Features2	Exterior features line 2
Exterior_Features3	Exterior features line 3
Exterior_Features4	Exterior features line 4
Construction1	Construction remark line 1
Construction2	Construction remark line 2
Foundation	Foundation type
Independent variable: sociodemographic factors (census block group)	
RATE_BB	Rates of children/teenagers within house
RATE_hhO	Rates of house occupied
RATE_hhinc	Rates of household income larger than 1000000
RATE_carO	Rates of car (>=1) occupied
RATE_whi	Rates of white only
DEN_pp	Population density
DEN_hh	Household density
Independent variable: surrounding built environment (0.6-mile)	
BE_pthubs	Number of public transit hubs
BE_ptroutes	Number of public transit routes
BE_permits	Number of new permits
BE_ptstops	Number of public transit stops
BE_dis	Distance from housing to state capital
Independent variable: regional status	
T_pp	Travis county population
T_GDP	Travis county GDP

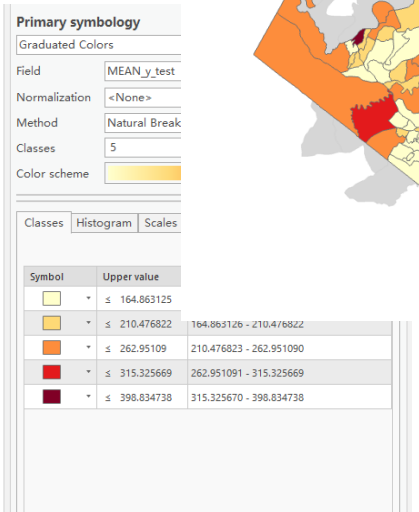
Empirical studies

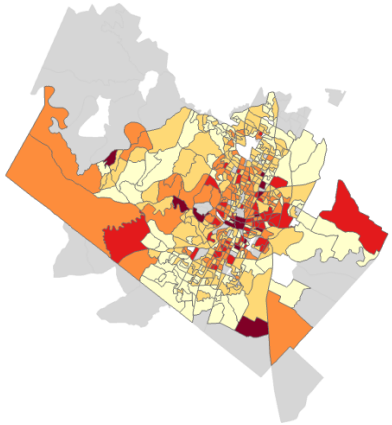
- SALES model (obs=property location)
- 2014-2016 predict 2017
- Focusing on sold price per square foot
- Average error: $24.31/sqft

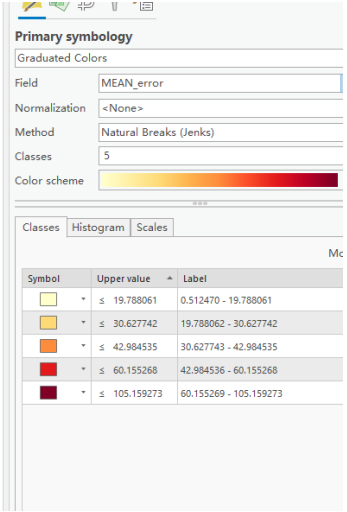
Error (abs(sold price-predicted value))
Aggregated to CBG for visualization



Pilot test 1



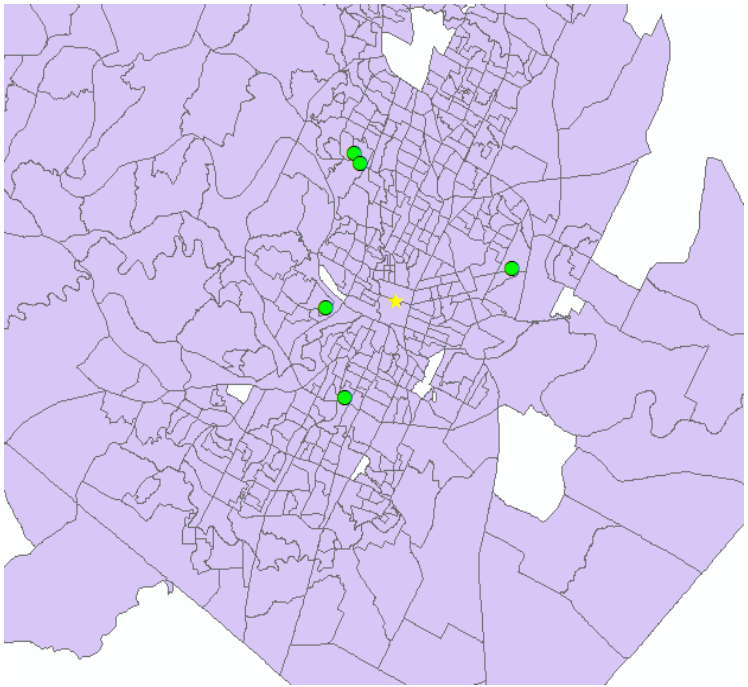




Empirical studies

Limitations in predicting housing prices in the future

- Test: using the list price right now to test.
- 09/10/2021 capture 5 property on Zillow (green)
- Average error of SALES model (using 2014-2017): $111.59/sqft

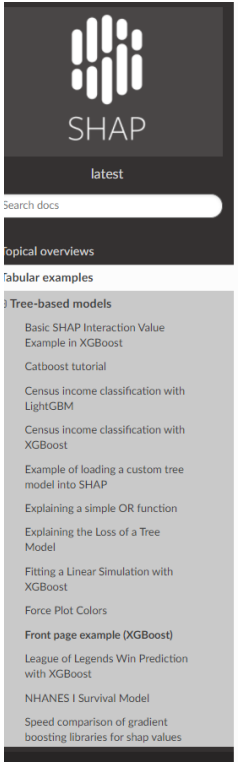


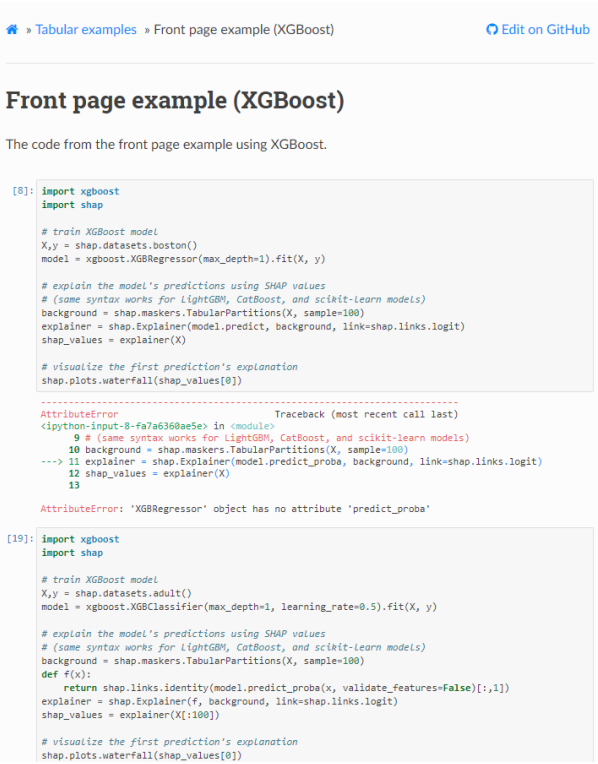
Pilot test 1

Empirical studies

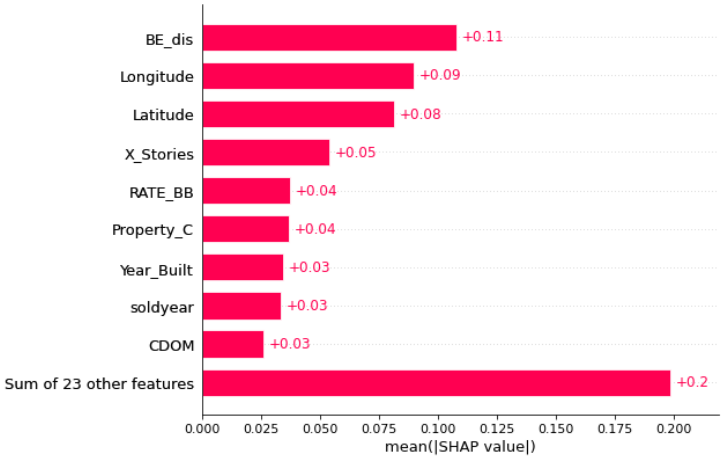
- (Idea 1) Explore how model performance changes across groups of property values
- (Idea 2) Explore how model performance changes across different spatial units

Pilot test 2



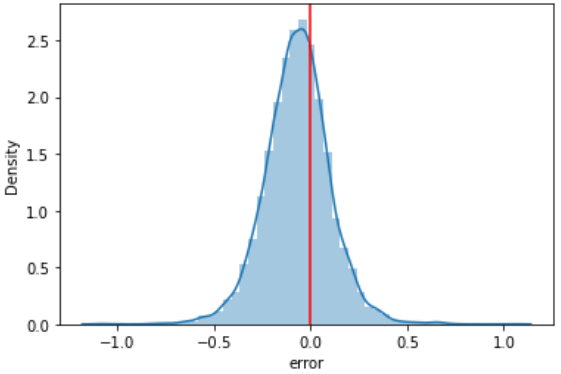


SHAP to visualize the results (https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Front%20page%20example%20(XGBoost).html)

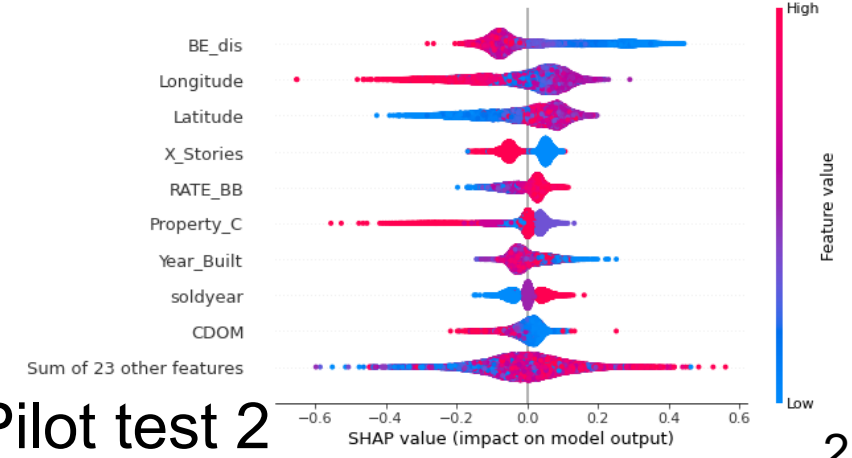


Distance to DT
Longitude
Latitude
Stories
Rates of BB in HH
Property construction
Year_built
Sold year
CDOM

1



Outcome variable = log($/SQFT)
Abs(average error) = 0.13

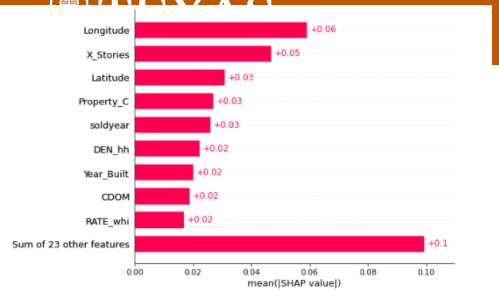


Distance to DT -
Longitude +
Latitude +
Stories
Rates of BB in HH +
Property construction -
Year_built -
Sold year
CDOM +

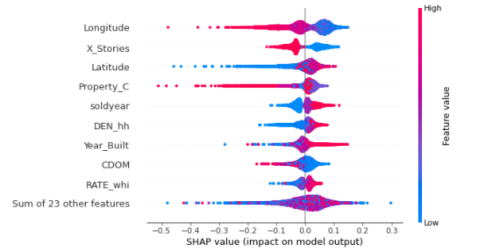
FULL model (obs=property location, n= 8527)

Pilot test 2

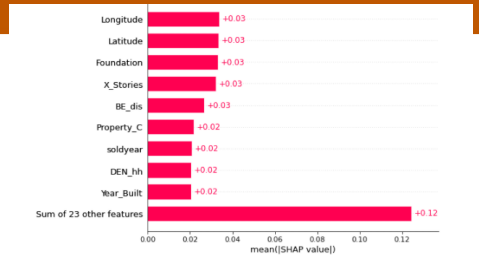
2



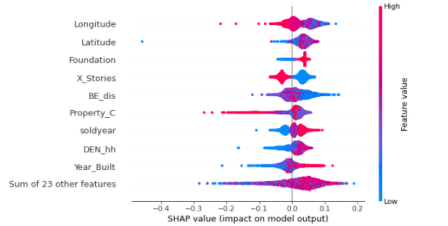

```
shap.plots.beeswarm(shap_values1)
```



1

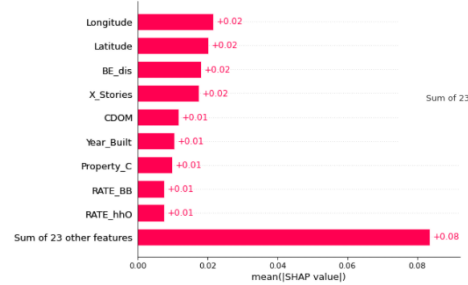



```
5]: shap.plots.beeswarm(shap_values3)
```

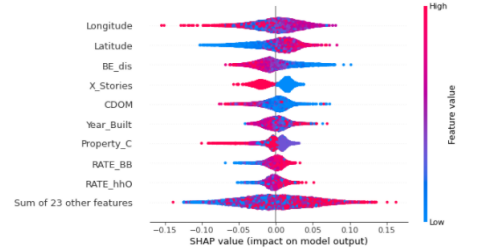


3

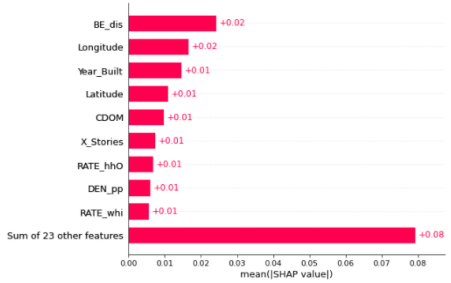
2



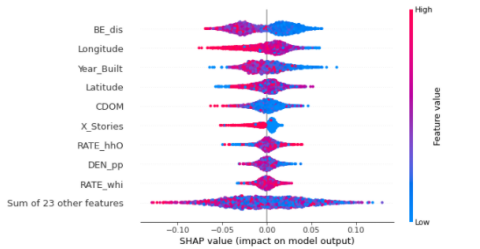

```
]: shap.plots.beeswarm(shap_values2)
```



4

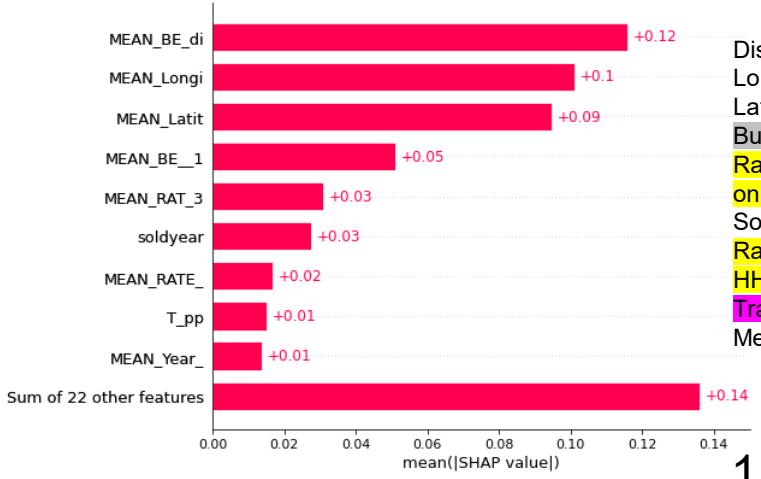



```
69]: shap.plots.beeswarm(shap_values4)
```



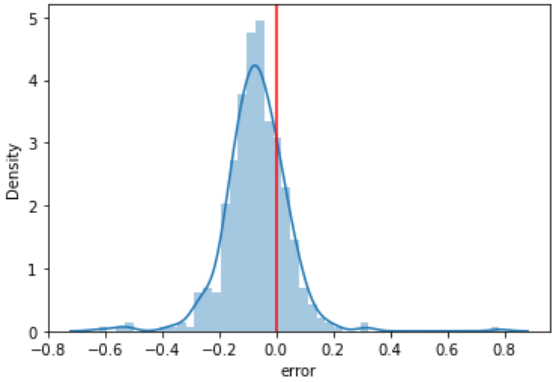
SUB model (obs=property location)

Pilot test 2



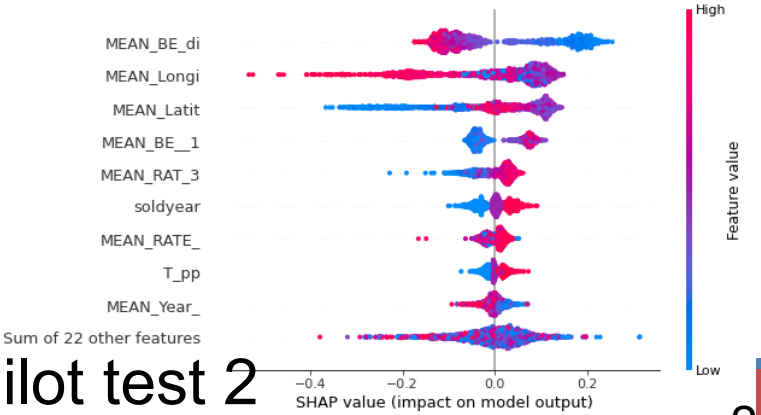
1

Distance to DT
Longitude
Latitude
Bus routes
Rate of white only
Sold year
Rates of BB in HH
Travis population
Mean year built



FULL model (obs= CBG)

Outcome variable = log($/SQFT)
Average error = 0.08

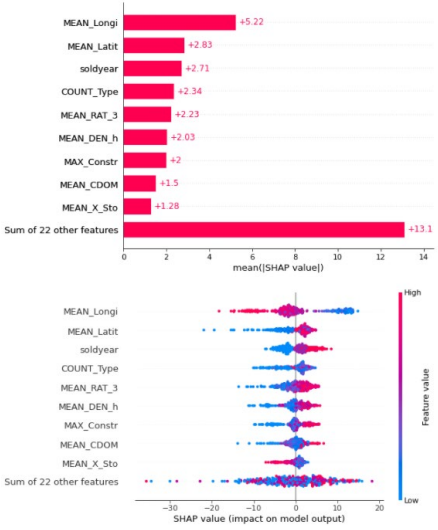


2

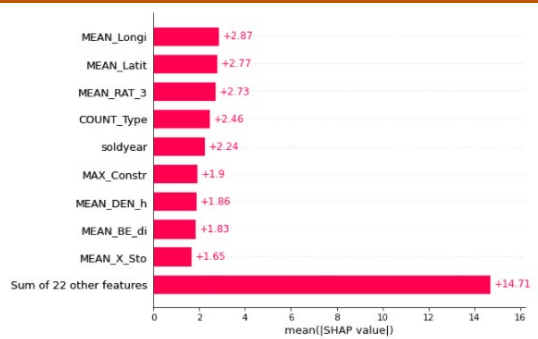
Distance to DT -
Longitude +
Latitude +
Bus routes
Rate of white only +
Sold year
Rates of BB in HH +
Travis population
Mean year built

Pilot test 2

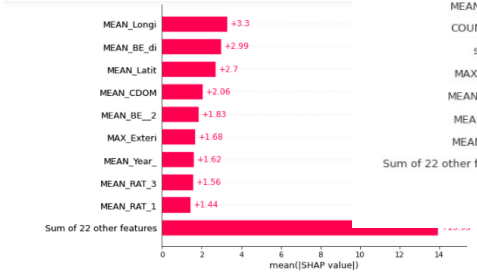
SUB model
(obs=CBG)

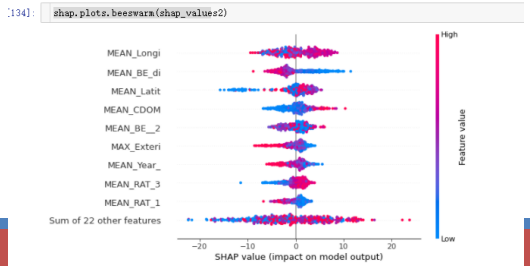


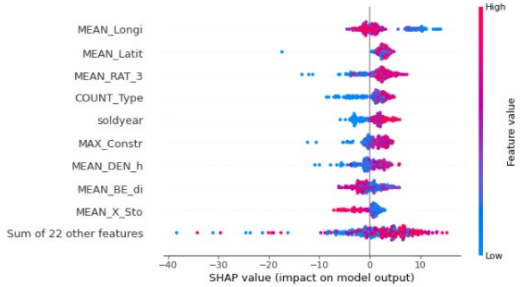
1



2

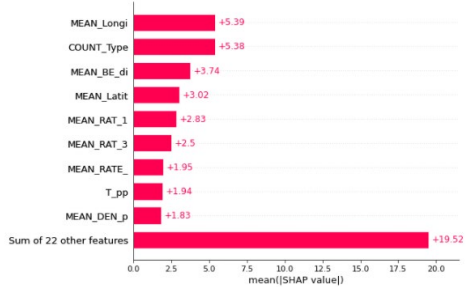


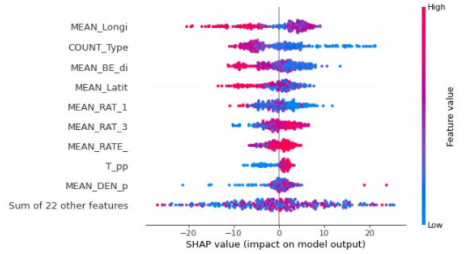




3

4





Pilot test 2

Empirical studies

- Using 2014-2016 SALES data predict 2017 single-family property value through GBDT

- (Idea 1) Explore how model performance changes across groups of property values
  - The GBDT model performs best in predicting single-family properties with low values. When the value of property increases, the performance of the model goes worse.

- (Idea 2) Explore how model performance changes across different spatial units
  - The GBDT model performs better in predicting spatial-aggregated average property value than individual properties.
  - Focusing on sociodemographic factors and BE factors of full models, rates of BB in households is the only one listed as top ten features (positive), While bus routes (-+), rates of white only (+), and rates of BB in households (+) are significant if aggregated to CBG.

- (Idea 1 x Idea 2)
  - (Choosing individual housing) Besides rates of BB in households, HH density is important in low value group; rates of white only is important in low-middle value group; rates of white only and HH density are significant in upper-middle value group; rates of white only and rates of HH ownership are significant in high value group.
  - (Choosing CBG) Besides rates of BB and white only, density and permits issued in low value group and ; HH ownership and public transit stops in low-middle; permits issued and HH den in upper-middle; bus routes, HH ownership, and HH den is high value group.

Pilot test 3

Empirical studies

Empirical studies

Single-family property values prediction (SVM, xgbdt, RF, lasso, ridege, elasticnet):

- RQ1. What are the variables having relatively high importance on property value predictions?
- RQ2. Does the importance change across neighborhoods?

https://www.wraltechwire.com/2021/11/02/zillow-unable-to-predict-housing-prices-to-wind-down-offers-program-everywhere/

LAB TIME

HW: using the technologies learned today to test and find the best models in the prediction