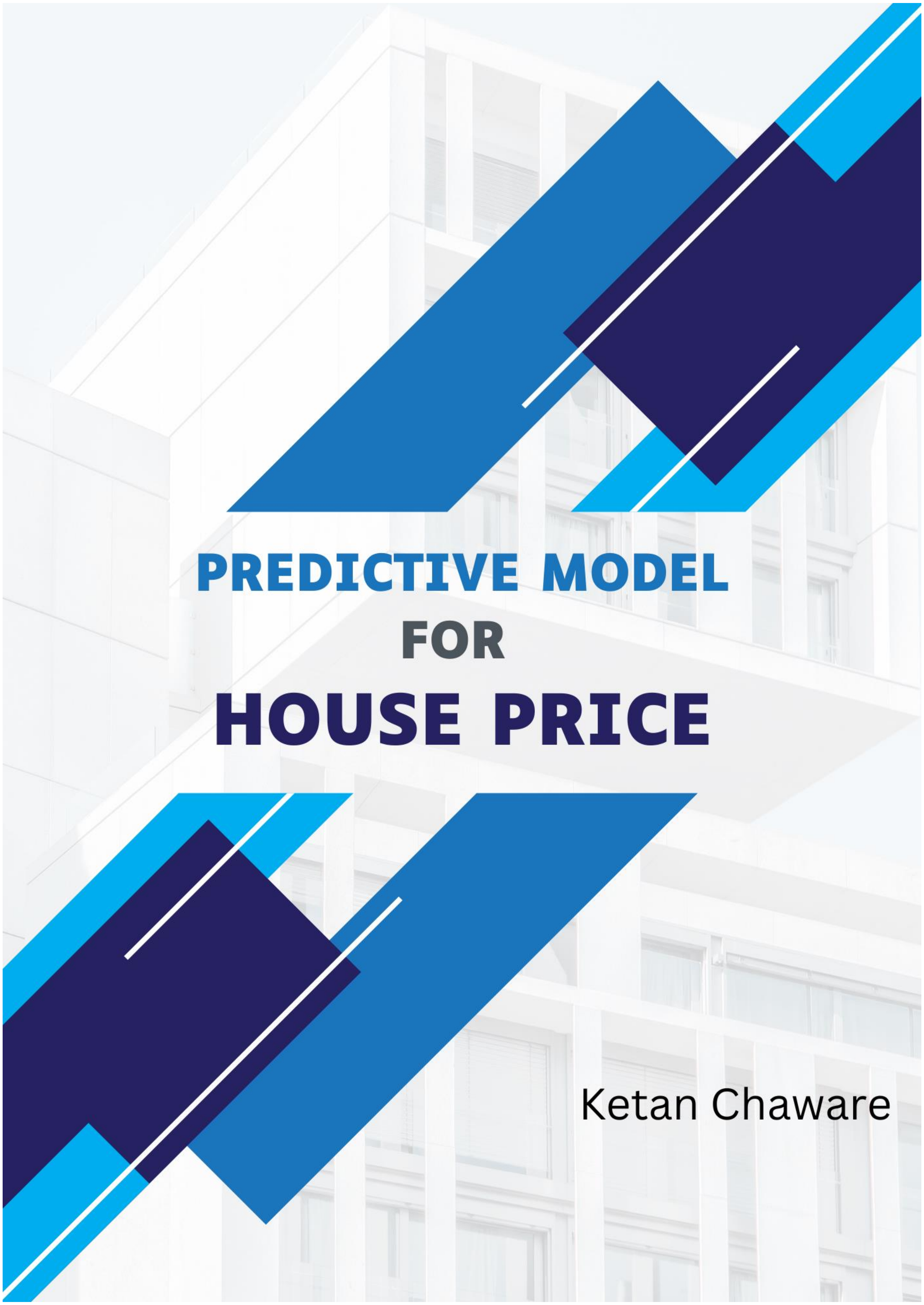


PREDICTIVE MODEL FOR HOUSE PRICE

Ketan Chaware

Table of Contents

House Price Dataset Analysis

Knowledge of Business

The real estate market involves buying, selling, and renting properties. Property prices are influenced by various factors such as location, size, condition, and market trends. Understanding these factors can help in making informed decisions regarding investments and sales.

Business Goal

The primary goal of this project is to predict house prices accurately based on various features. This can help stakeholders, including real estate agents, buyers, sellers, and investors, make data-driven decisions.

Model Objective

The objective is to develop a predictive model that estimates the price of houses based on features such as the number of bedrooms, bathrooms, square footage, location, and other relevant characteristics.

Problem Statement

How can we accurately predict the prices of houses based on their features?

Importance of the Problem

Accurate house price predictions can provide significant advantages in the real estate market, including:

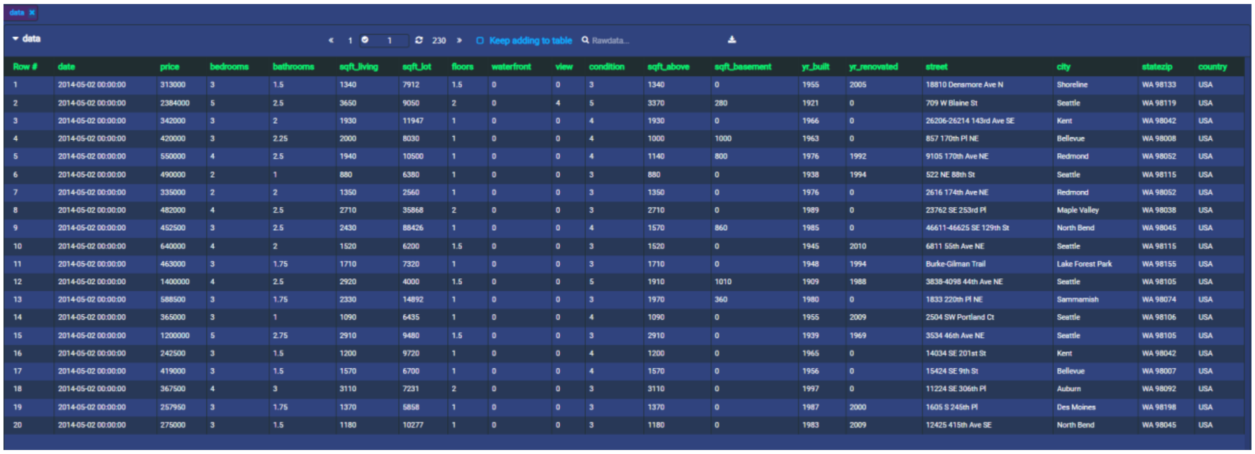
- Better investment decisions
- Fair pricing for buyers and sellers
- Enhanced market efficiency

Historical Insight

Real estate prices have historically been influenced by economic conditions, interest rates, and demographic trends. Analyzing past data can help identify patterns and factors that drive property values.

Data Source and Type

The data is sourced from the Titanic dataset available on Kaggle. This regression dataset includes variables such as date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, Waterfront, View, Condition, sqft_above, yr_built, yr_renovated,street, city, country, Statezip.



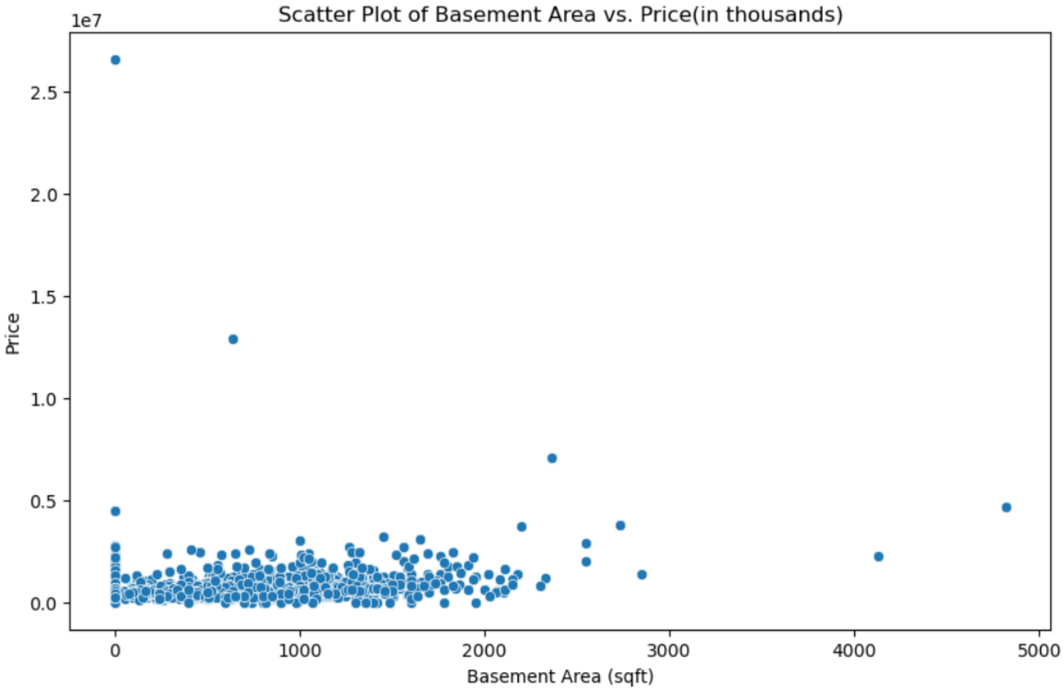
Row #	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	street	city	statezip	country
1	2014-05-02 00:00:00	313000	3	1.5	1340	7912	1.5	0	0	3	1340	0	1955	2005	18810 Densmore Ave N	Shoreline	WA 98133	USA
2	2014-05-02 00:00:00	2384000	5	2.5	3650	9050	2	0	4	5	3370	280	1921	0	709 W Blaine St	Seattle	WA 98119	USA
3	2014-05-02 00:00:00	342000	3	2	1930	11947	1	0	0	4	1930	0	1966	0	26206-26214 143rd Ave SE	Kent	WA 98042	USA
4	2014-05-02 00:00:00	420000	3	2.25	2000	8030	1	0	0	4	1000	1000	1963	0	857 170th Pl NE	Bellevue	WA 98008	USA
5	2014-05-02 00:00:00	550000	4	2.5	1940	10500	1	0	0	4	1140	800	1976	1992	9105 170th Ave NE	Redmond	WA 98052	USA
6	2014-05-02 00:00:00	490000	2	1	880	6380	1	0	0	3	880	0	1938	1994	522 NE 88th St	Seattle	WA 98115	USA
7	2014-05-02 00:00:00	335000	2	2	1350	2560	1	0	0	3	1350	0	1976	0	2616 174th Ave NE	Redmond	WA 98052	USA
8	2014-05-02 00:00:00	482000	4	2.5	2710	35868	2	0	0	3	2710	0	1989	0	23762 SE 253rd Pl	Maple Valley	WA 98038	USA
9	2014-05-02 00:00:00	452500	3	2.5	2430	88426	1	0	0	4	1570	860	1985	0	46611-46625 SE 129th St	North Bend	WA 98045	USA
10	2014-05-02 00:00:00	640000	4	2	1520	6200	1.5	0	0	3	1520	0	1945	2010	6811 55th Ave NE	Seattle	WA 98115	USA
11	2014-05-02 00:00:00	463000	3	1.75	1710	7320	1	0	0	3	1710	0	1948	1994	Burke-Gilman Trail	Lake Forest Park	WA 98155	USA
12	2014-05-02 00:00:00	1400000	4	2.5	2920	4000	1.5	0	0	5	1910	1010	1909	1988	3838-4098 44th Ave NE	Seattle	WA 98105	USA
13	2014-05-02 00:00:00	588500	3	1.75	2330	14892	1	0	0	3	1970	360	1980	0	1833 220th Pl NE	Sammamish	WA 98074	USA
14	2014-05-02 00:00:00	365000	3	1	1090	6435	1	0	0	4	1090	0	1955	2009	2504 SW Portland Ct	Seattle	WA 98106	USA
15	2014-05-02 00:00:00	1200000	5	2.75	2910	9480	1.5	0	0	3	2910	0	1939	1969	3534 46th Ave NE	Seattle	WA 98105	USA
16	2014-05-02 00:00:00	242500	3	1.5	1200	9720	1	0	0	4	1200	0	1965	0	14034 SE 201st St	Kent	WA 98042	USA
17	2014-05-02 00:00:00	419000	3	1.5	1570	6700	1	0	0	4	1570	0	1956	0	15424 SE 9th St	Bellevue	WA 98007	USA
18	2014-05-02 00:00:00	367500	4	3	3110	7231	2	0	0	3	3110	0	1997	0	11224 SE 306th Pl	Auburn	WA 98092	USA
19	2014-05-02 00:00:00	257950	3	1.75	1370	5858	1	0	0	3	1370	0	1987	2000	1605 S 245th Pl	Des Moines	WA 98198	USA
20	2014-05-02 00:00:00	275000	3	1.5	1180	10277	1	0	0	3	1180	0	1983	2009	12425 415th Ave SE	North Bend	WA 98045	USA

Methodology: Supervised Learning – Classification

Exploratory Data Analysis (EDA)

EDA involves analyzing the dataset to understand its structure, detect anomalies, and identify relationships between variables. This includes:

- Summary statistics
- Distribution plots
- Correlation analysis



The above scatter chart compares price with basement area.

floor vs price of house

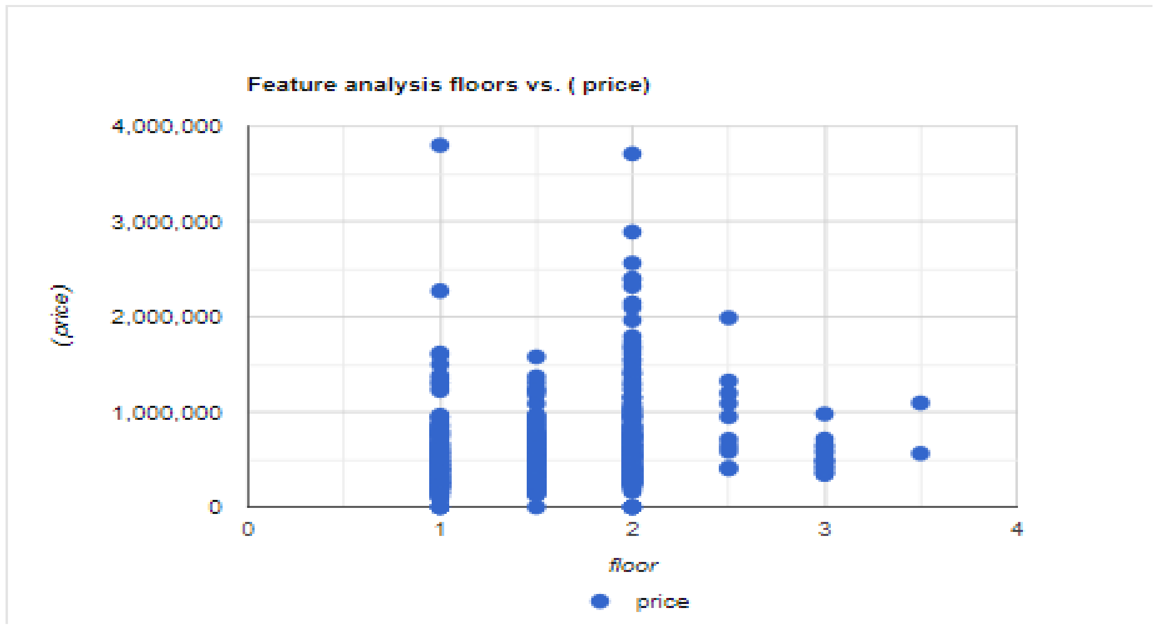


Chart comparing price with floors.

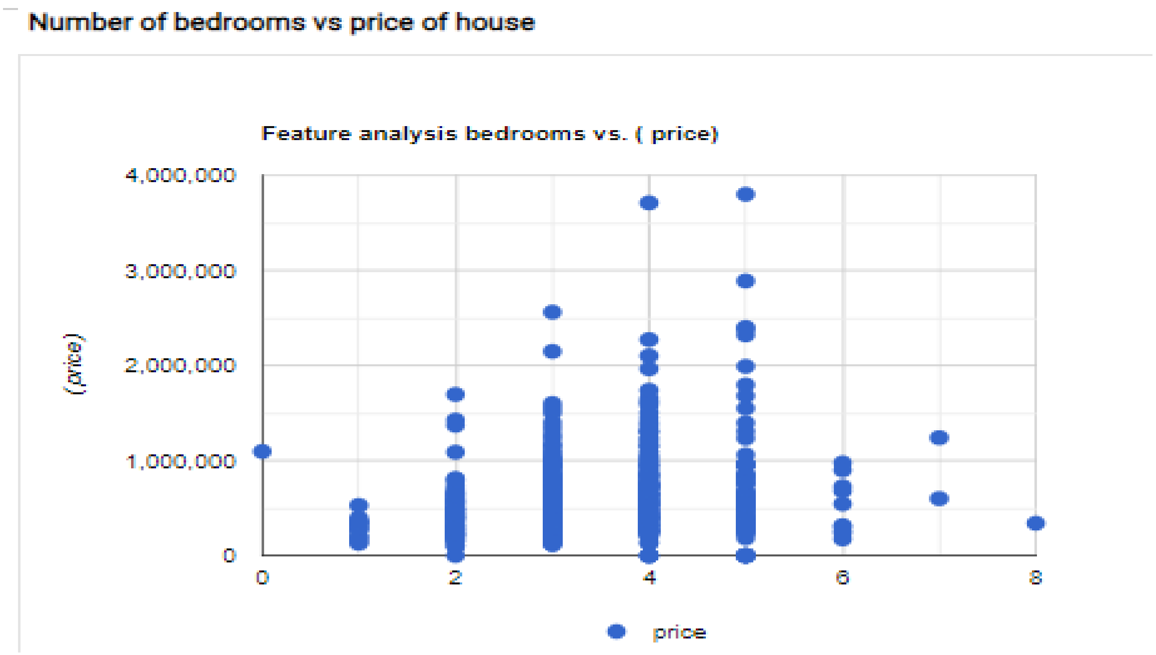
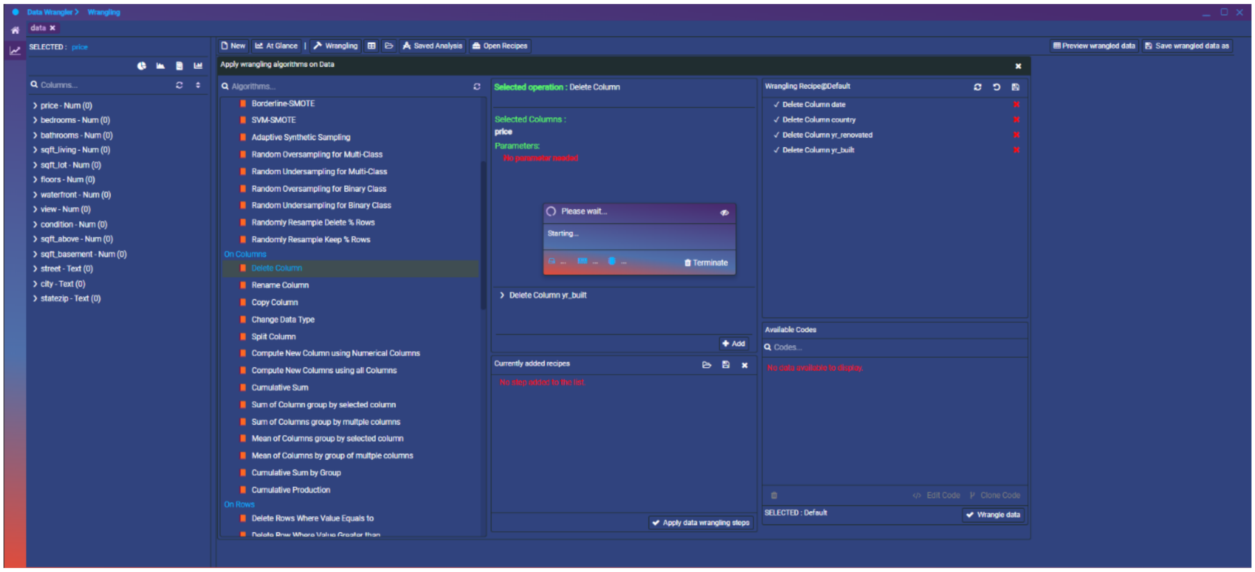


Chart showing bedrooms and price comparison.

Data Engineering and Wrangling

Data wrangling involves cleaning and transforming the data to make it suitable for analysis. This includes:

- Handling missing values
- Encoding categorical variables
- Normalizing or standardizing numerical features


Data Wrangler > Wrangling
data
New
At Glance
Wrangling
Saved Analysis
Open Recipes
Preview wrangled data
Save wrangled data as
Apply wrangling algorithms on Data
Columns...
Algorithms...
Borderline-SMOTE
SVM-SMOTE
Adaptive Synthetic Sampling
Random Oversampling for Multi-Class
Random Undersampling for Multi-Class
Random Oversampling for Binary Class
Random Undersampling for Binary Class
Randomly Resample Delete % Rows
Randomly Resample Keep % Rows
On Columns
Delete Column
Rename Column
Copy Column
Change Data Type
Split Column
Compute New Column using Numerical Columns
Compute New Columns using all Columns
Cumulative Sum
Sum of Column group by selected column
Sum of Columns group by multple columns
Mean of Columns group by selected column
Mean of Columns by group of multple columns
Cumulative Sum by Group
Cumulative Production
On Rows
Delete Rows Where Value Equals to
Delete Row Where Value Greater than
Terminate
Add
Currently added recipes
Apply data wrangling steps
Wrangling Recipe@Default
Available Codes
Codes...
Edit Code
Clone Code
Wrangle data

SELECTED : price

- price - Num (0)
- bedrooms - Num (0)
- bathrooms - Num (0)
- sqft_living - Num (0)
- sqft_lot - Num (0)
- floors - Num (0)
- waterfront - Num (0)
- view - Num (0)
- condition - Num (0)
- sqft_above - Num (0)
- sqft_basement - Num (0)
- street - Text (0)
- city - Text (0)
- statezip - Text (0)

Selected operation : Delete Column

Selected Columns :
price
Parameters:
No parameter needed

Please wait...
Starting...

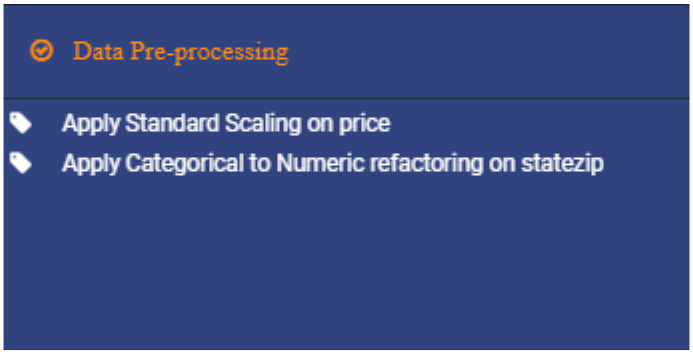
Delete Column yr_built

No step added to the list.

✓ Delete Column date
✓ Delete Column country
✓ Delete Column yr_renovated
✓ Delete Column yr_built

No data available to display.

SELECTED : Default



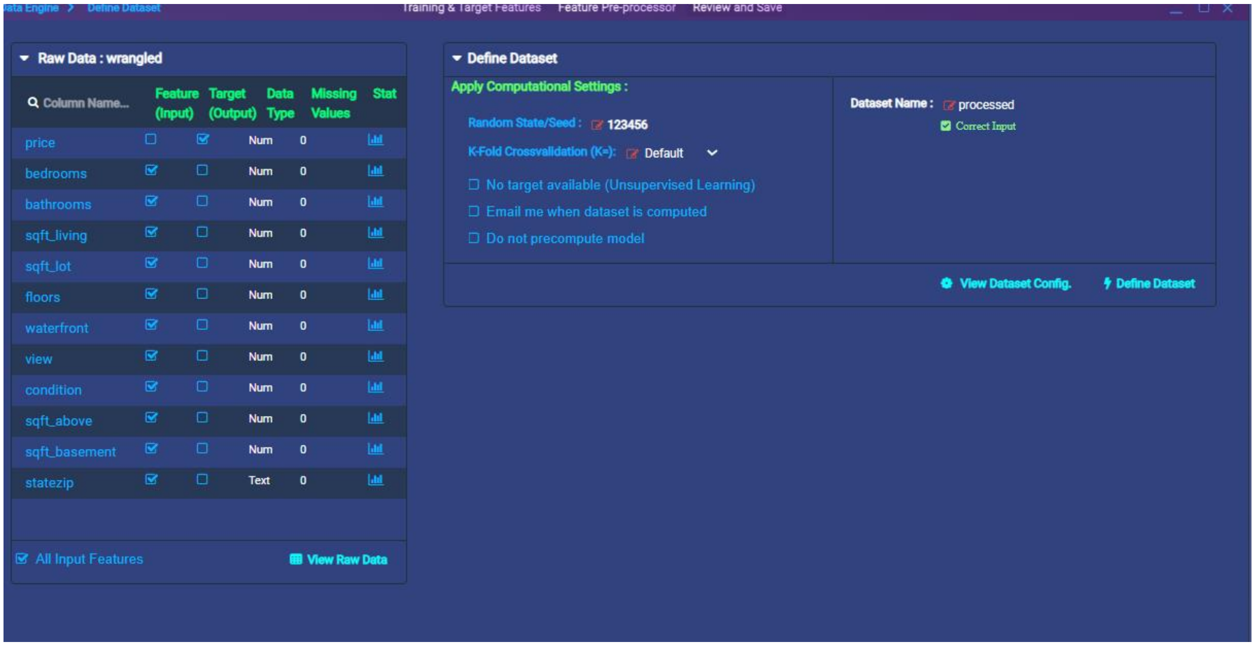
Data Pre-processing

- Apply Standard Scaling on price
- Apply Categorical to Numeric refactoring on statezip

Data Preparation for Modeling

Preparing the data involves splitting it into training and testing sets. Modelling involves selecting appropriate algorithms and training them on the data. Common algorithms for regression tasks include:

- Linear Regression
- Decision Trees
- Random Forests
- Gradient Boosting Machines



Raw Data : wrangled

Column Name...	Feature (Input)	Target (Output)	Data Type	Missing Values	Stat
price	□	☑	Num	0	
bedrooms	☑	□	Num	0	
bathrooms	☑	□	Num	0	
sqft_living	☑	□	Num	0	
sqft_lot	☑	□	Num	0	
floors	☑	□	Num	0	
waterfront	☑	□	Num	0	
view	☑	□	Num	0	
condition	☑	□	Num	0	
sqft_above	☑	□	Num	0	
sqft_basement	☑	□	Num	0	
statezip	☑	□	Text	0	

☑ All Input Features

Define Dataset

Apply Computational Settings :

Random State/Seed : 123456

K-Fold Crossvalidation (K=): Default

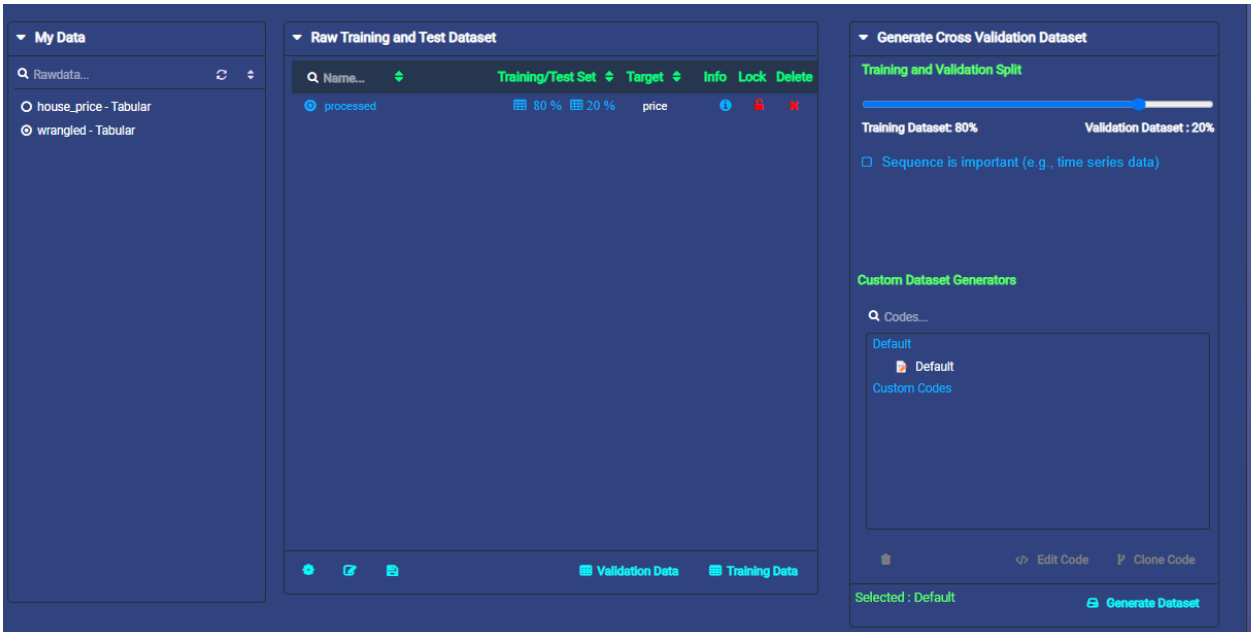
□ No target available (Unsupervised Learning)

□ Email me when dataset is computed

□ Do not precompute model

Dataset Name : processed

☑ Correct Input



My Data

○ house_price - Tabular

● wrangled - Tabular

Raw Training and Test Dataset

Name...	Training/Test Set	Target	Info	Lock	Delete
processed	80 % 20 %	price			✗

Generate Cross Validation Dataset

Training and Validation Split

Training Dataset: 80%

Validation Dataset : 20%

□ Sequence is important (e.g., time series data)

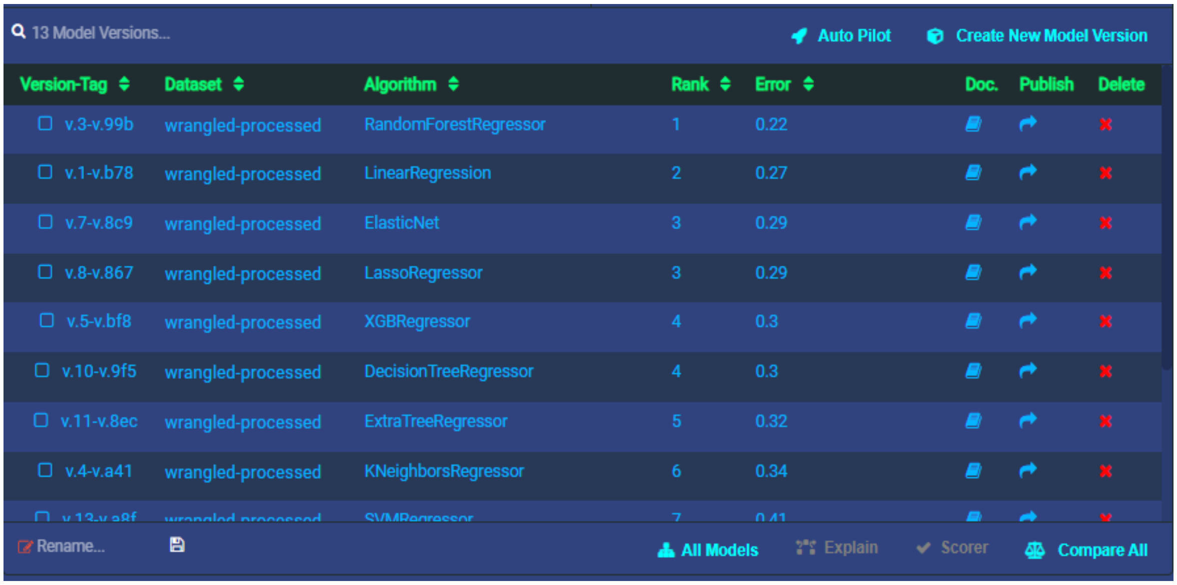
Custom Dataset Generators

Default

Default

Custom Codes

Selected : Default

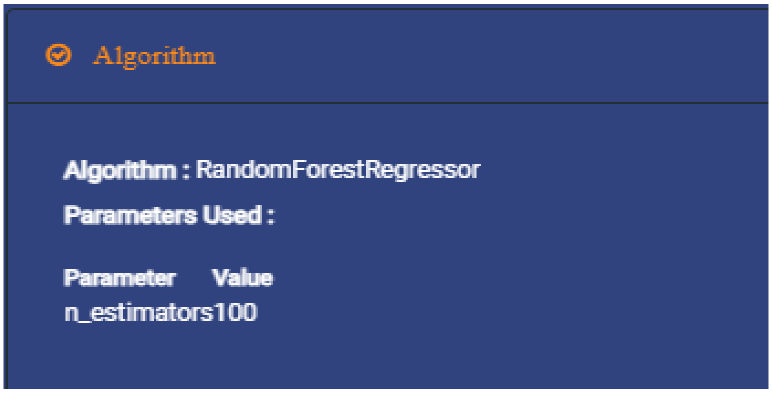


Version-Tag	Dataset	Algorithm	Rank	Error	Doc.	Publish	Delete
□ v.3-v.99b	wrangled-processed	RandomForestRegressor	1	0.22	📘	↪	✖
□ v.1-v.b78	wrangled-processed	LinearRegression	2	0.27	📘	↪	✖
□ v.7-v.8c9	wrangled-processed	ElasticNet	3	0.29	📘	↪	✖
□ v.8-v.867	wrangled-processed	LassoRegressor	3	0.29	📘	↪	✖
□ v.5-v.bf8	wrangled-processed	XGBRegressor	4	0.3	📘	↪	✖
□ v.10-v.9f5	wrangled-processed	DecisionTreeRegressor	4	0.3	📘	↪	✖
□ v.11-v.8ec	wrangled-processed	ExtraTreeRegressor	5	0.32	📘	↪	✖
□ v.4-v.a41	wrangled-processed	KNeighborsRegressor	6	0.34	📘	↪	✖
□ v.13-v.a8f	wrangled-processed	SVMRegressor	7	0.41	📘	↪	✖

Modeling Process

Algorithms Used

- RandomForestClassifier
- LinearRegression
- ElasticNet
- LassoRegressor



Algorithm : RandomForestRegressor

Parameters Used :

Parameter **Value**
n_estimators100

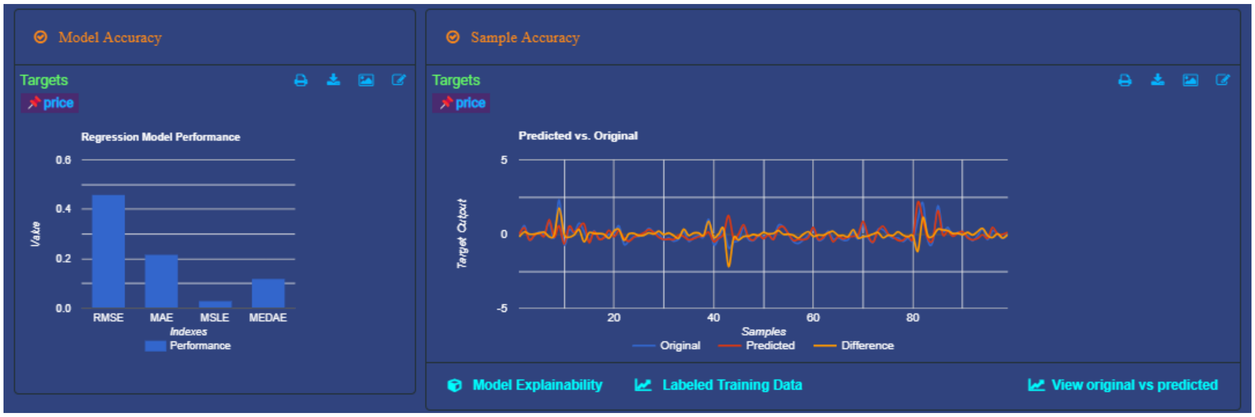
Parameters and Values

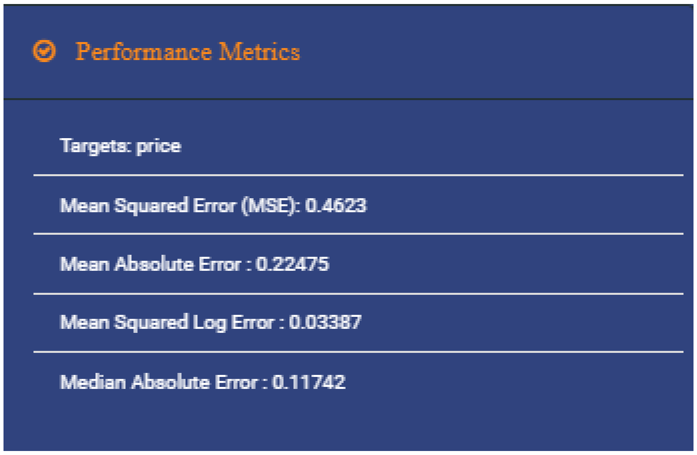
- **RandomForestRegressor:** n_estimators100

Results

Algorithm	Error
RandomForestRegressor	0.22
LinearRegression	0.27
ElasticNet	0.29
LassoRegressor	0.29

We chose RandomForestRegressor as it showed the least error of 0.22





Performance Metrics

Targets: price

Mean Squared Error (MSE): 0.4623

Mean Absolute Error : 0.22475

Mean Squared Log Error : 0.03387

Median Absolute Error : 0.11742

Conclusions

Improvements to Make in the Future

- ➢ Collecting more data to improve model accuracy.
- ➢ Incorporating additional features such as economic indicators and neighborhood characteristics.
- ➢ Exploring advanced modelling techniques and ensemble methods.

Solutions to the Client

Providing the client with:

- An interactive dashboard to visualize house price predictions.
- Detailed reports on factors influencing house prices.
- Recommendations for pricing strategies based on model insights.

Impact of Project

The project can significantly impact the real estate market by:

- Enhancing pricing strategies.
- Improving investment decisions.
- Increasing market transparency and efficiency.