12/6/2021

House Sale Price Prediction – Snohomish County

Prepared by:

- Mangesh Jadhav
- mail2mangeshjadhav@gmail.com

House Sale Price Prediction- Snohomish County

Table of Contents

1) Data Set Overview (Property Sales in Snohomish County)

Residential property sales in the Snohomish County (Washington) are public and offer useful information of the real estate market. Assessors, gather this data from official documents and upload the consolidated dataset for evaluating annual taxes on residential properties.

A) Dataset Source

- This can be downloaded from Washington Snohomish County Assessors FTP Portal: ftp://ftp.snoco.org/assessor/property_sales
    - File Name: AllSales_2021_10_06.xlsx
- Has 5 years (2016-2020, and partial 2021) of property sales in Snohomish County, Washington.
- Records have key property attributes like:
    - Sale Price and Sale Date (recorded at Parcel_Id level and can be looked up)
    - Sale Qualification Code (Qualified, Forced Sale, Estate Sale etc.)
    - Plot Area (Sq. Ft.)
    - Built-up Area (Sq. Ft.)
    - Number of Bedrooms
    - Built Year of the House
    - Property Class (Single Family, Condominiums etc.)
    - House Type (1 Storey, 2 Storey, Split Entry etc.)
    - Grade (Quality of Building represented in numeric code 15-95)
    - School District Number

B) Data Dictionary

- Field Names
    - Property_Class_Codes

C) Assumptions & Data Limitations

To simplify data exploration and processing, this dataset was further limited to look into:

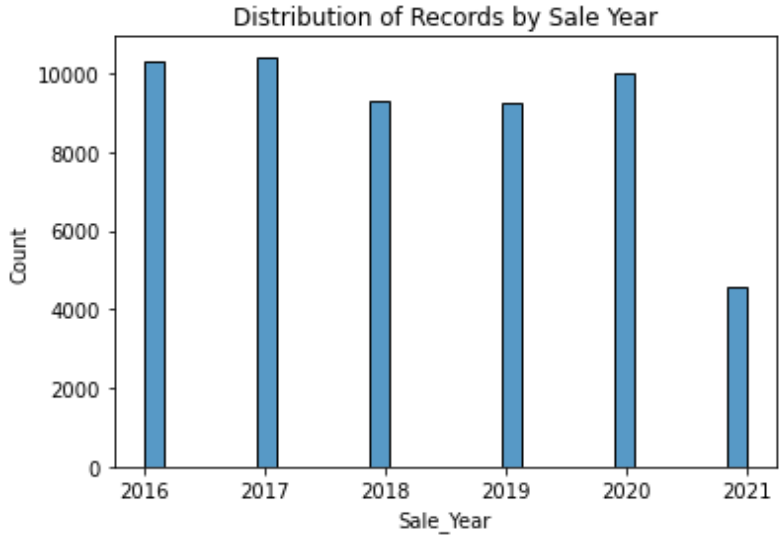
- 1.5k property sales after May 2021 were used for testing the model predictions
- Property Type = 111 i.e. Single-Family Residence – Detached (One Structure)
- Sales Qualification Code = Qualified
- Consider only recent sale of property (ignore previous sales in past 5 years; if any)
- Exclude property sales with outliers
    - NULLS or 0's as the area of Land_SqFt or House_SqFt
    - Price_Per_Built_SqFt (outside of 3 standard deviations from the mean)
    - Built_To_Land_Ratio (outside of 3 standard deviations from the mean)
    - Number of Bedrooms (0 or 6+) to avoid data skew
    - House Type in (Quad Level, 2+ Sty B, Multi-Level) due to lack of datapoints
    - School District Numbers in (63, 330, 306) due to lack of datapoints

2) Data Set Exploration

Below are the high-level observations on the dataset:

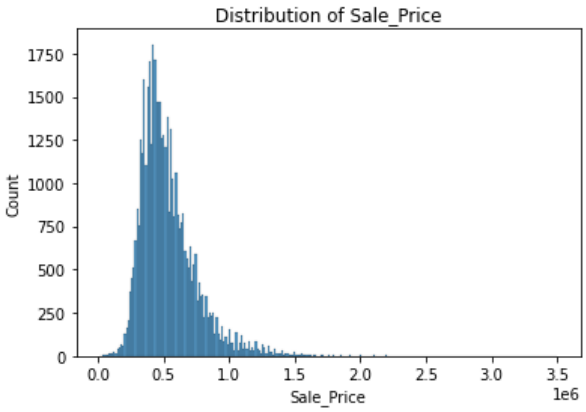
A) Number of Records (45k Sale Records)

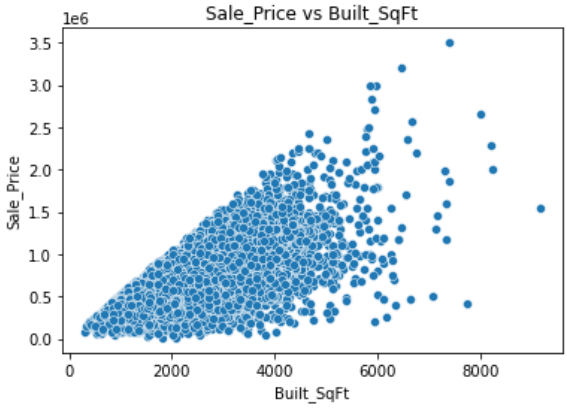
- Raw dataset had 104k sales records with 51 columns spanning across 5+ years.
- After de-duplicating and outlier removal the processed dataset had:
    A) 45k records and 10 features
        - Train Dataset: 43.6k property records (for the duration of 2016-2021 April)
        - Test Dataset: 1.5k property records (for the duration of May-2021 onwards)



B) Sale Price (Target Variable)

- Raw dataset had the Sale Price values ranging from $1 to $200 Million.
- Processed dataset had Sale Price values ranging from $10k to $3.5 Million.
    A) Normal distribution with most of the values between $250k to $900k
- Approximately Linear relationship between Sale Price vs Built Sq. Ft.



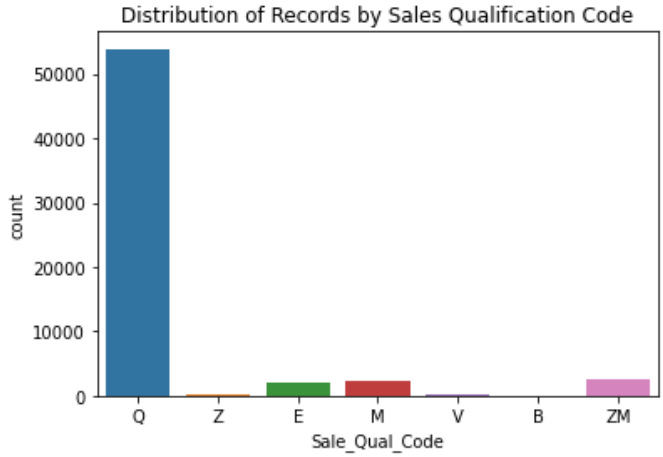


C) Property Class = Single Family Residence – Detached (One Structure)

- Most of the values in the dataset are for Prop_Class = 111 which is Single Family Residence – Detached (One Structure). Processed dataset was limited to Prop_Class = 111.

D) Sale Qualification Code = Qualified

- Qualification Code has values as Q, Z, B, E, V or M where Q = Qualified, Z = Qualified but includes other property, B = Qualified but improvement added after sale occurred, E = Estate Sale, V = Forced Sale, M = Miscellaneous). Processed dataset was limited to Sale Qualification Code = **Q**ualified.



E) Multiple records for Parcel_Ids (Use recent sales record)

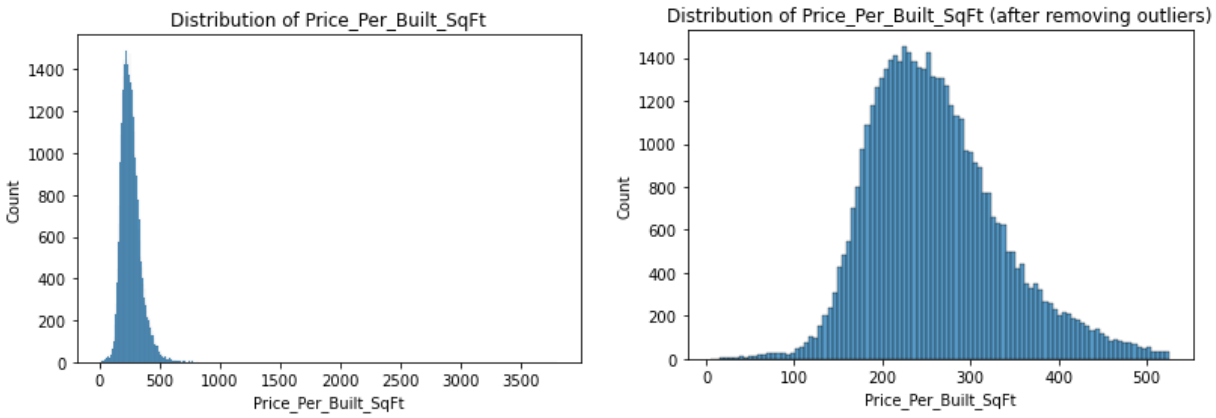
- Raw dataset had multiple sale records for same property. For simplicity I've limited it to recent sale transaction.

```
Count of Records: 53872
Count of Unique Values in Parcel_ID: 47073
```

	Parcel_Id	SD_Nbr	Prop_Class	Sale_Qual_Code	Sale_Date	Sale_Price	Land_SqFt	Built_SqFt	Grade	Grade_Desc	Yr_Blt	House_Type
0	795300003900	2	111	Q	2018-06-15	415000.0	6534.0	1578.0	45.0	Avg	1994.0	11.0
1	795300003900	2	111	Q	2020-05-18	499950.0	6534.0	1578.0	45.0	Avg	1994.0	11.0

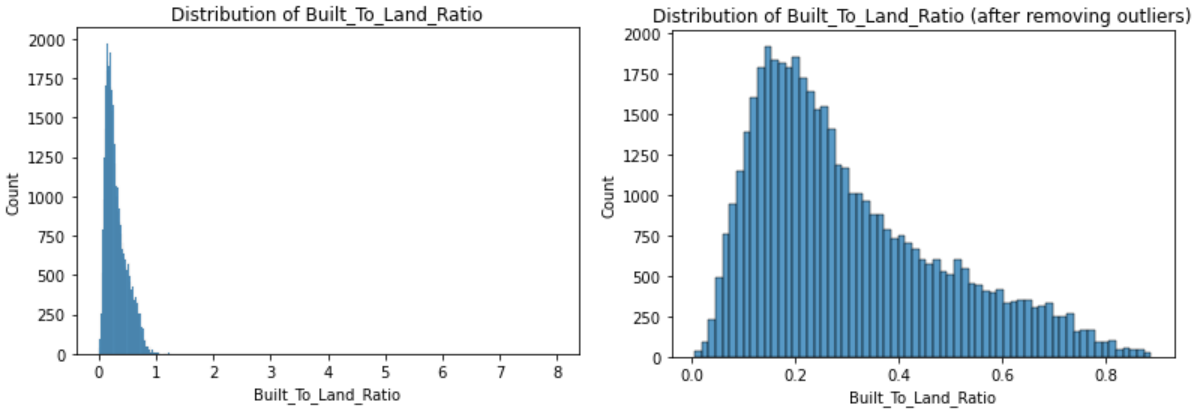
F) Price_Per_Built_SqFt (New Feature) to exclude outliers

- After processing the dataset Price_Per_Built_SqFt is normally distributed with most of the values in the range of $150/SqFt to $400/SqFt



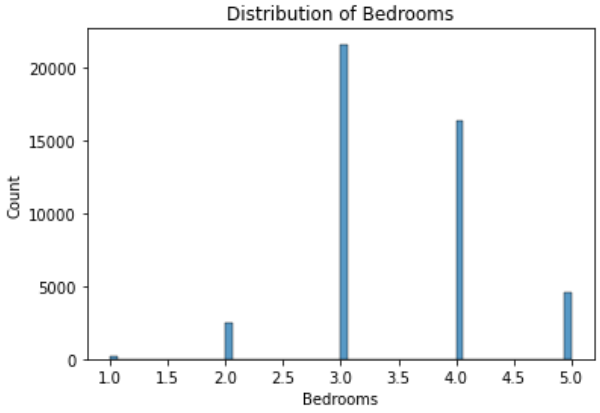
G) Built_To_Land_Ratio (New Feature) to exclude outliers

- After processing the dataset Built_To_Land_Ratio is distributed with most of the values in the range of 15% to 50%.



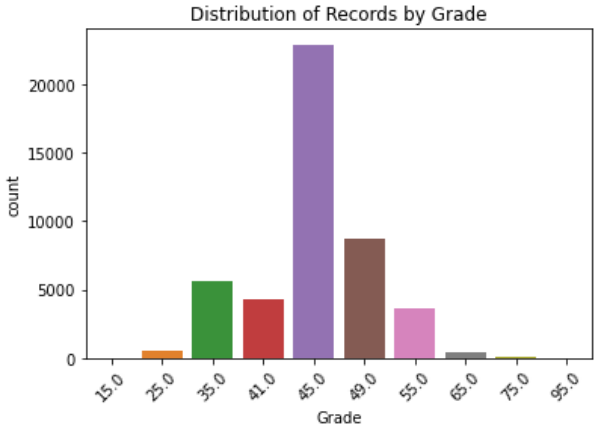
H) 3 or 4 Bedrooms are Popular in Single Family Houses

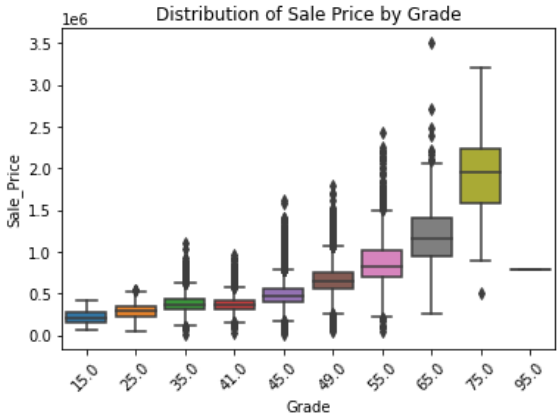
- 3- and 4-bedroom houses are popular among single family houses in Snohomish County.
- I've excluded houses with 0 or 6+ bedrooms to avoid data skew



I) Normally distributed Grade (or Quality of House)

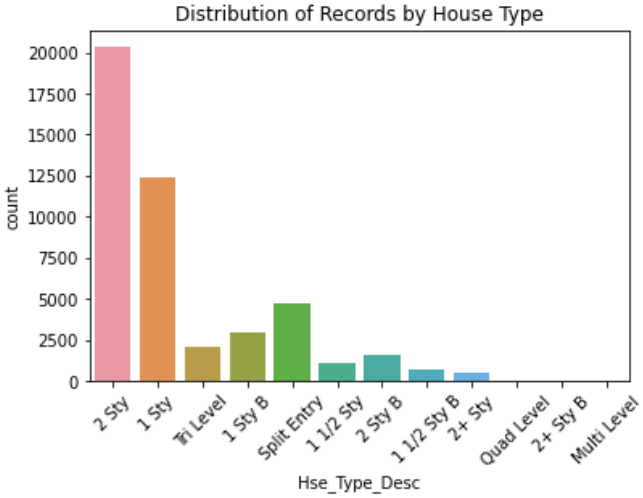
- Grade is a numeric measure for quality if the house with normally distributed values between 15 to 95.
- Sale price increases with the grade (or Quality) of the house.

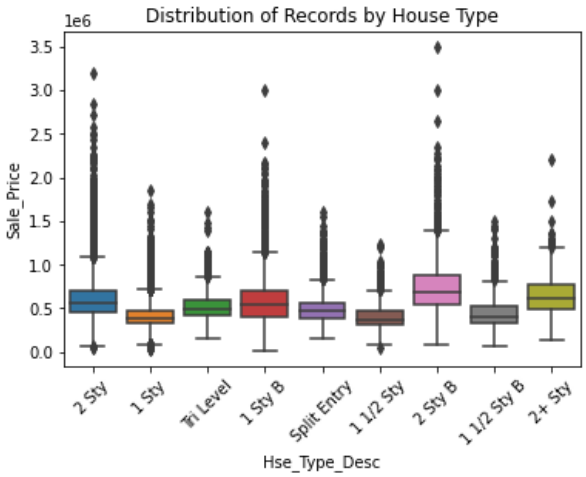




J) Popular House Types (2 Storey, 1 Storey)

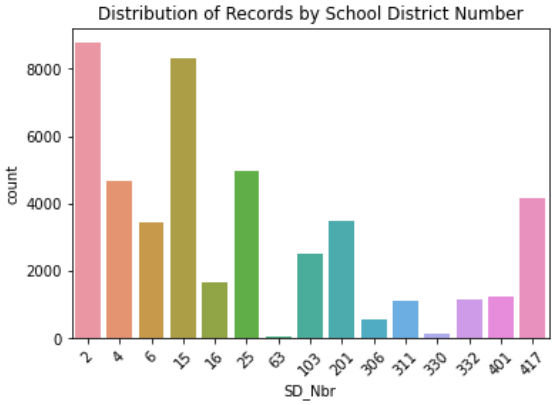
- Most popular house types (as seen in the sales records) are 2 Storey and 1 Storey houses.
- I've excluded 'Quad Level', '2+ Sty B', 'Multi Level' House Types given their low representation in data.
- Generally, 2 Storeyed houses have higher sale prices than 1 Storey or tri-Storeyed.

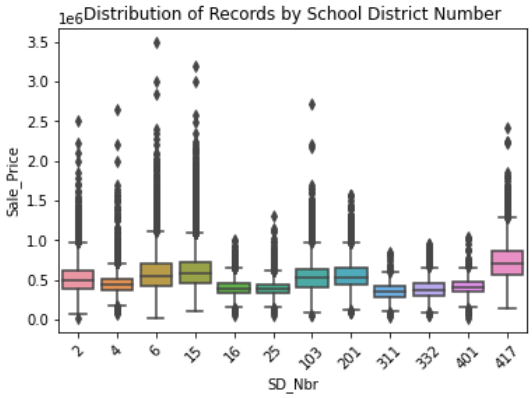




K) School District #417 has the Higher Sale Prices

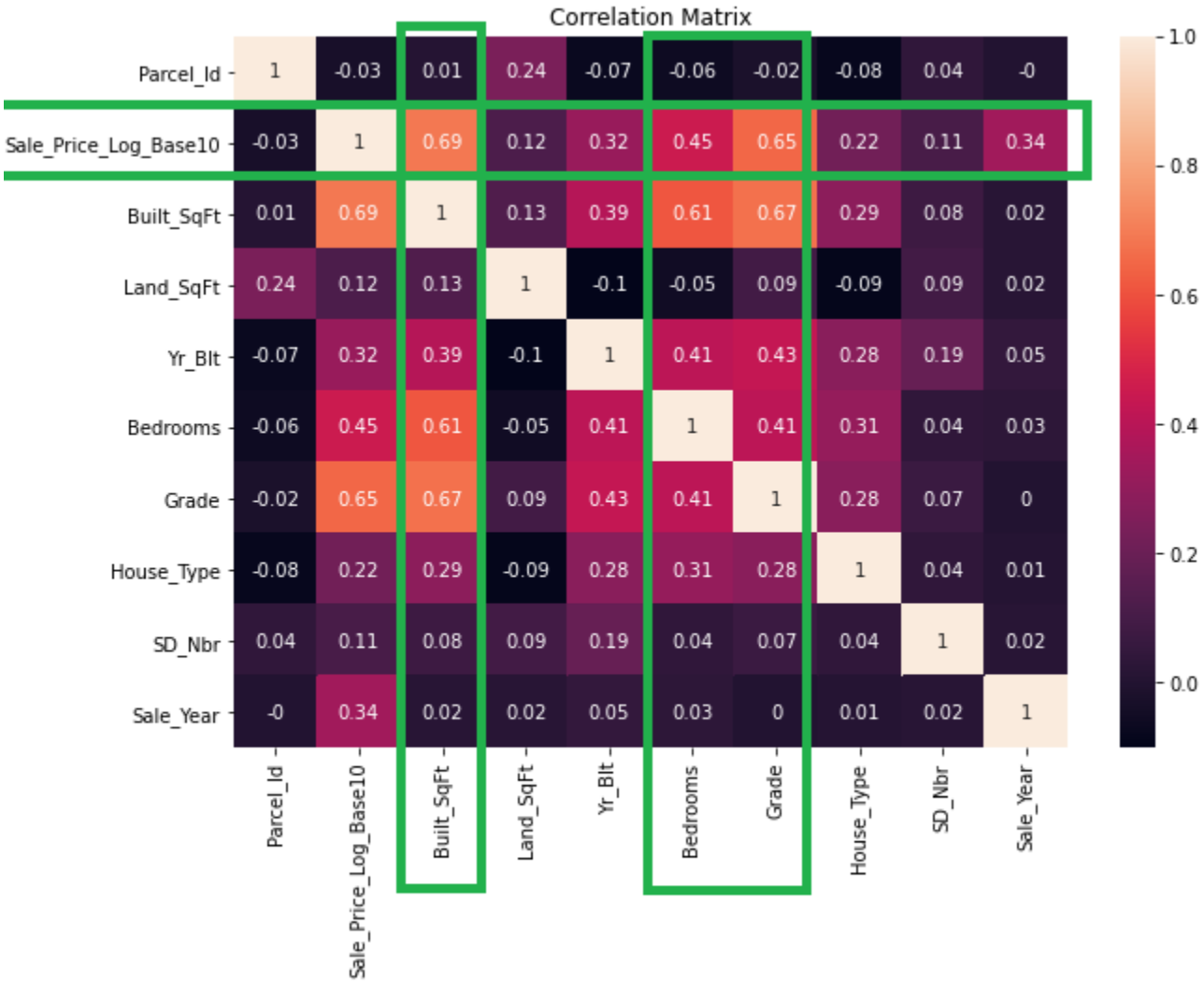
- School District #417 has the highest median sale price (usually higher than 75th percentile of other school districts)
- I've excluded 63, 330, 306 School Districts given their low representation





L) Correlation Analysis for Sale Prices

- As per the correlation analysis the variable highly correlated with Sale Price are:
  - A) **Bult_SqFt** (Corr Coeff = 0.69)
  - B) **Grade** (Corr Coeff = 0.65)
  - C) **Bedrooms** (Corr Coeff = 0.45)



3) Define problem type (Regression: Predict Sales Price of Houses)

The **objective** is to predict the continuous output variable i.e. **Sale Price of Single-Family Houses in Snohomish County, WA** using the categorical and numerical features available in property sale dataset.

4) Model development (using AutoGluon & Ensemble Learning)

- For simplifying and automating the feature processing and model training, I've used **AutoGluon**. Using this feature, I've set the **training time limit to 3 minutes**.
- For evaluation, I've converted the **Sale Price to log 10 values**. This will make sure model training treats the prediction errors for high priced houses similar to low priced houses.
- For regression task, the evaluation function is set to **Mean Squared Error**.
- I've configured test dataset to **predict sale prices** in house **sale dates >= May-2021**.
- During 3 minutes AutoGluon evaluated 11 models and chose **WeightedEnsemble_L2** as the best performing one based on the score_val. If we need to reduce the prediction time, then **XGBoost** would give similar accuracy in lower time.

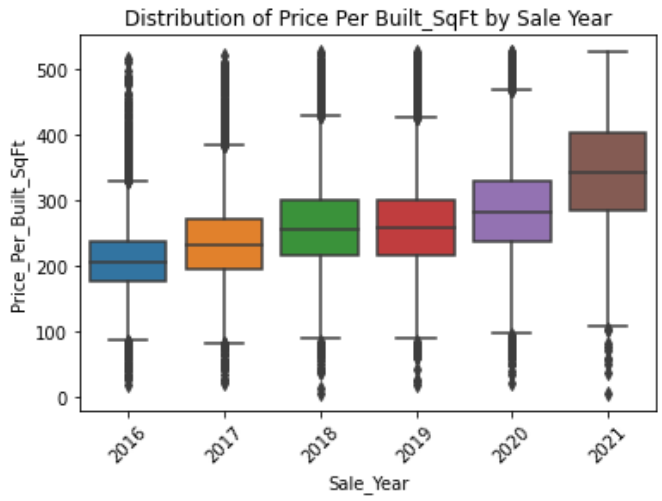
	model	score_val	pred_time_val	fit_time	pred_time_val_marginal	fit_time_marginal	stack_level	can_infer	fit_order
0	WeightedEnsemble_L2	-0.003938	0.998672	22.892883	0.000582	0.426133	2	True	12
1	XGBoost	-0.004108	0.020131	1.955198	0.020131	1.955198	1	True	9
2	LightGBMLarge	-0.004230	0.039961	1.286745	0.039961	1.286745	1	True	11
3	CatBoost	-0.004254	0.003600	5.703687	0.003600	5.703687	1	True	6
4	LightGBM	-0.004282	0.109182	2.364211	0.109182	2.364211	1	True	4
5	RandomForestMSE	-0.004431	0.202992	8.011673	0.202992	8.011673	1	True	5
6	LightGBMXT	-0.004517	0.463026	9.464273	0.463026	9.464273	1	True	3
7	ExtraTreesMSE	-0.004959	0.102788	2.771535	0.102788	2.771535	1	True	7
8	NeuralNetMXNet	-0.005009	0.175867	92.618857	0.175867	92.618857	1	True	10
9	NeuralNetFastAI	-0.005563	0.053319	51.663509	0.053319	51.663509	1	True	8
10	KNeighborsDist	-0.011356	0.103406	0.128989	0.103406	0.128989	1	True	2
11	KNeighborsUnif	-0.011467	0.105746	0.135082	0.105746	0.135082	1	True	1

5) Analysis and Conclusions

- Converting the Log Values back to normal Sale Prices, and the % prediction error of a **known datapoint was ~8%** while the overall **RMSE for Log Values was 0.08**.
- Below are the key features ranked by their importance (with significant P-Values):
  A) **Built Area** of the House in Sq. Ft.
  B) **Grade** or Quality of the House
  C) **School District**
- In the correlation analysis **Bedrooms** had higher correlation coefficient of 0.45 with Sale Price. However, Bedrooms also had a high correlation of 0.61 with Built Area. And can cause high-collinearity. Thus, the importance of Bedrooms is lower.

	importance	stddev	p_value	n	p99_high	p99_low
Built_SqFt	0.011812	0.000415	0.000206	3	0.014191	0.009433
Grade	0.006643	0.000257	0.000249	3	0.008115	0.005172
SD_Nbr	0.005767	0.000235	0.000276	3	0.007114	0.004421
Parcel_Id	0.001955	0.000245	0.002595	3	0.003359	0.000552
Yr_Blt	0.001699	0.000287	0.004695	3	0.003344	0.000053
Land_SqFt	0.000712	0.000107	0.003744	3	0.001327	0.000097
Bedrooms	0.000631	0.000123	0.006181	3	0.001333	-0.000072
House_Type	0.000046	0.000084	0.220845	3	0.000526	-0.000434
Sale_Year	0.000000	0.000000	0.500000	3	0.000000	0.000000

- Surprisingly, the **Sale Year** was NOT significant or important in prediction. This seems contrary to the +20% YoY market trend in Snohomish County. I think this is probably due to the fact that historical sale price values are not adjusted for the appreciation. To verify this logic, if we plot Price/Built Sq. Ft across years, then we can see that overall real estate prices have gone up with each year.



6) Executive summary

Residential property sales in the Snohomish County (Washington) are public and offer useful information of the housing features via FTP Portal. The **objective** of this exercise is to predict the continuous output variable i.e. **Sale Price of Single-Family Houses in Snohomish County, WA** using the categorical and numerical features available in property sale dataset. The raw dataset had 104k sales records with 51 columns spanning across 5+ years. After de-duplicating and outlier removal the processed dataset had: 45k records and 10 features. 1.5k property sale records (with sales in May-2021 onwards) were used in **testing** the accuracy of the predictions.

For simplifying and **automating the feature processing and model training**, I've used **AutoGluon**. Using this feature, I've set the training time limit to 3 minutes. For evaluation, I've converted the **Sale Price to log 10 values**. This will make sure model training treats the prediction errors for high priced houses similar to low priced houses. For regression task, the evaluation function is set to **Mean Squared Error**. During 3 minutes AutoGluon evaluated 11 models and chose **WeightedEnsemble_L2** as the best performing one based on the score_val. If we need to reduce the prediction time, then **XGBoost** would give similar accuracy in lower time.

Converting the Log Values back to normal Sale Prices, and the % prediction error of a **known datapoint was ~8%** while the overall **RMSE for Log Values was 0.08**. Below are the key features ranked by their importance (with significant P-Values):

A) **Built Area** of the House in Sq. Ft.
B) **Grade** or Quality of the House
C) **School District**

In the correlation analysis **Bedrooms** had higher correlation coefficient of 0.45 with Sale Price. However, Bedrooms also had a high correlation of 0.61 with Built Area. And can cause high-collinearity. Thus, the importance of Bedrooms is lower.

Surprisingly, the **Sale Year** was NOT significant or important in prediction. This seems contrary to the +20% YoY market trend in Snohomish County. I think this is probably due to the fact that historical sale price values are not adjusted for the appreciation. To verify this logic, if we plot Price/Built Sq. Ft across years, then we can see that overall real estate prices have gone up with each year.

7) Github Code Repository

A) Github Repository:

- https://github.com/jmmangesh/House_Price_Prediction

B) Detailed code (Jupyter notebook if using python)

- https://github.com/jmmangesh/House_Price_Prediction/blob/main/Model_Snohomish_House_Price_Prediction.ipynb

C) Executive summary

- https://github.com/jmmangesh/House_Price_Prediction/blob/main/README.md