Executive Summary for Housing Price Prediction

Daji Liang | 730043364 | dl579@exeter.ac.uk

1 Introduction

Housing is one of the basic needs for human beings, and when buying a house, the price is a crucial factor to consider[7]. [6] indicates that housing price indices determining the housing market trend benefits sellers and real estate professionals. At the same time, it also points out that it can help the government formulate real estate policies and prevent financial crises caused by the depreciation of the real estate market and house prices. This study analyses housing price data from Kaggle[5] and uses regression, a machine learning(ML) technique, to investigate *how price varies with size, the number of bedrooms and bathrooms, year built, and neighborhood.*

2 Insights

1. **Introduction to Dataset**
   After we load the whole dataset and use `head(), info()` function, we can get Figure 4a in Appendix B. There are total 6 columns:

   - 'SquareFeet': the size of each house in square feet and the type is int;
   - 'Bedrooms' & 'Bathrooms': the number of bedrooms and bathrooms in each house;
   - 'Neighborhood': the area where the house is and the type is object which needs to be encoded before applying to the ML models;
   - 'YearBuilt' & 'Price': which year it was built and the price of the house.

   From Figure 4b, there are 50000 rows and no 'None' value in the data.

2. **Distribution**
   After removing the rows of 22 negative prices and using `describe()` to get Figure 5a in Appendix B, we can use `histogram()` to plot Figure **1a, 1b** and Figure 5 in Appendix B:

   - 'SquareFeet': the distribution of size varies from 1000 to 2999, the mean and medium are similar and from the Figure **1a** we can see it is like uniform distribution;
   - 'Bedrooms' & 'Bathrooms' & 'YearBuilt': from Figure 5b-5d, the number of bedrooms and bathrooms are from 2 to 5 and from 1 to 3 separately while the year of built varies from 1950 to 2021;
   - 'Neighborhood': from Figure 5e-5f, the distribution grouped by bedrooms and bathrooms is similar to uniform distribution too;
   - 'Price': from Figure 1b, 5g, 5h, they look like normal distribution and the price rises slightly with the increasing number of bedrooms or bathrooms and the declining distance from the city centre. Moreover, the price is averagely distributed in each group.

3. **Correlation**
   Before we choose the ML technique, we have to know the correlation of columns. Using the `corr()`, we can get Figure **1f** showing us the correlation matrix. Hence, we can plot the correlation between Size and Price, which is the biggest number in the heatmap, and get Figure **1c** representing the size and price positively correlated. We can also see the correlation of other columns from Figure 6 in Appendix B.

3 Prediction

1. **Choice of ML Technique**
   It is widely known that supervised learning, such as regression and classification, and unsupervised learning, such as clustering and dimensionality reduction, are two major and crucial domains within the field of machine learning. One of the biggest differences between them is whether the dataset contains explicit labels or outcomes. In this dataset, supervised learning is our choice.
   Similarly, the main distinction between regression and classification lies in their respective objectives: regression predicts continuous numerical outcomes, such as predicting house prices, while classification predicts discrete categorical outcomes, such as spam and image classification. Hence, we choose regression as the ML technique to address the research question.

2. **Description of Regression Models**
   In this study, I choose 6 regression models and I will explain the reason, pros and cons:

   - Linear & Lasso & Ridge:
     Linear is easy to use for real estate modeling due to its less complex implementation[4] while [3] found that the margin of error is slightly higher than the average.
     Lasso is a regularization and variable selection method using L1 regularization. [12] shows that it performed better than Ridge on multicollinearity house price prediction model and it is useful for eliminating trivial genes.
     Ridge uses L2 regularization to avoid overfitting. It is good for group selection, which is the opposite of Lasso.
   - Random Forest(RF): It can be used for both regression and classification and to address price prediction, analysis and decision-making in [3, 8, 9]. It is effective with numerical features in tabular data[4].
   - Gradient Boosting(GB) & XGB: The applicability is the same as RF and [11] shows the high accuracy of GB in house price market. It is better than RF on objective function under the gradient out while XGB, imporved over GB, incorporates several enhancements for better performance.

3. **Results**
   In Figure 1d, Ridge regression performs the best on MAE, RMSE and R2 score, so I use ridge to get the results: Figure **1e** and 7. Regarding the results of other models, they are in Figure 8. In conclusion, the price is mainly proportional to the size and rises slightly with the increasing number of bedrooms or bathrooms and the declining distance from the city centre.

4 Conclusion

Through this study, I have learnt that there are a lot of considerations like reasonableness to be taken into account in the processing of real datasets, for example, I need to remove the negative price and convert the object type to integer in this dataset. Moreover, when choosing the models to predict house prices, the balance between interpretability, complexity and performance as well as the trade-off between training time and accuracy have to be considered. In addition, a high correlation does not imply a cause-and-effect relationship between features, especially in small samples. Correlation between two features may be due to mutual causation or independent influence from another factor. For example, the example of Amazon's facial recognition technology has ever led to a 5% error rate in ACLU's testing[2, 10].

5 limitations

The dataset I chose does not include all the important factors in it which makes it not suitable for a real world prediction. [1, 7] show that gross domestic product, population, real property gains tax, construction costs, etc., are the factors affecting the house price, so we have to consider so many other factors when selecting a good dataset. Regarding the models, we can choose ANN, SVM, FLS, HPM, DT, KNN, PLS, NB, MRA, SA, etc. to fit different real datasets and problems. Moreover, we can also apply a log transformation and raise the values of 'Size' to the power of 0.8 and calculate the unit price per square feet to research (see Appendix A).

References

[1] Dennis R Capozza, Patric H Hendershott, Charlotte Mack, and Christopher J Mayer. Determinants of real house price dynamics. Working Paper 9262, National Bureau of Economic Research, October 2002.

[2] Drew Harwell. Oregon became a testing ground for Amazon's facial-recognition policing. But what if Rekognition gets it wrong? *Washington Post*, May 2019.

[3] Hakan Kuşan, Osman Aytekin, and İlker Özdemir. The use of fuzzy logic in predicting house selling price. *Expert systems with Applications*, 37(3):1808–1813, 2010.

[4] Thuraiya Mohd, Nur Syafiqah Jamil, Noraini Johari, Lizawati Abdullah, and Suraya Masrom. An Overview of Real Estate Modelling Techniques for House Price Prediction. In Naginder Kaur and Mahyudin Ahmad, editors, *Charting a Sustainable Future of ASEAN in Business and Social Sciences*, pages 321–338, Singapore, 2020. Springer Singapore.

[5] Muhammad Imran. Housing Price Prediction Data, 2023.

[6] Byeonghwa Park and Jae Kwon Bae. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, 42(6):2928–2934, 2015.

[7] Tze San Ong. Factors affecting the price of housing in malaysia. *J. Emerg. Issues Econ. Financ. Bank*, 1:414–429, 2013.

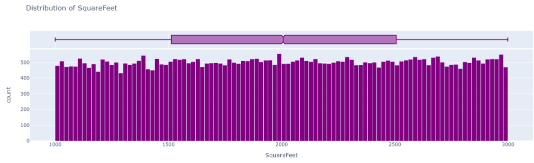
[8] Neelam Shinde and Kiran Gawande. Valuation of house prices using predictive techniques. *Journal of Advances in Electronics Computer Science*, 5(6):34–40, 2018.

[9] Michiel van Wezel, Martijn Kagie, and Rob Potharst. Boosting the accuracy of hedonic pricing models. Technical report, 2005.

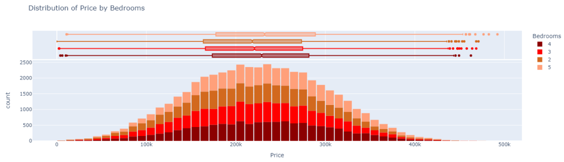
[10] Karen Weise and Natasha Singer. Amazon Pauses Police Use of Its Facial Recognition Software. *The New York Times*, June 2020.

[11] Nannan Zhang, Lifeng Wu, Jing Yang, and Yong Guan. Naive bayes bearing fault diagnosis based on enhanced independence of data. *Sensors*, 18(2):463, 2018.

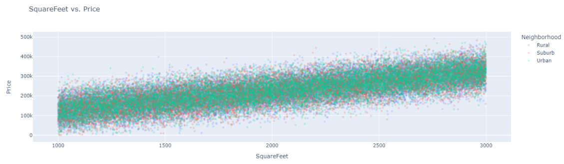
[12] Julia Zmölnig, Melanie N Tomintz, and Stewart A Fotheringham. A spatial analysis of house prices in the kingdom of fife, scotland. *GI_Forum*, pages 125–134, 2014.



(a) Distribution of Size



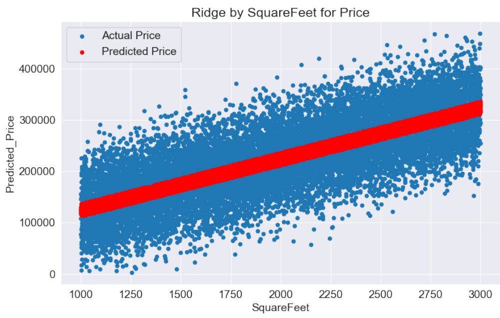
(b) Distribution of Price grouped by Bedrooms



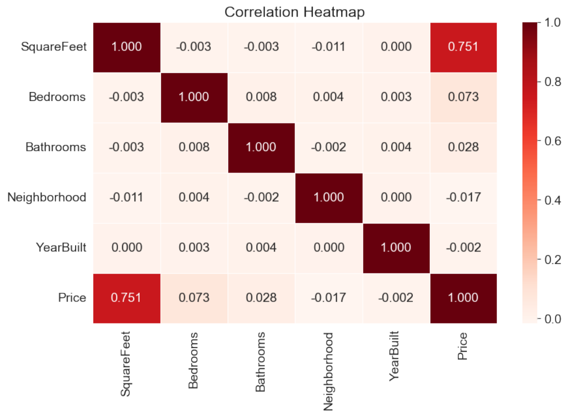
(c) Correlation between Size and Price grouped by Neighborhood

	Model Name	Mean Absolute Error MAE	Mean Absolute Percentage Error MAPE(%)	Root Mean Squared Error RMSE	R2 score
0	Ridge	39952.016876	23.698109	49963.589764	0.563508
1	Lasso	39952.075005	23.696517	49963.822024	0.563504
2	LinearRegression	39952.077259	23.696484	49963.828746	0.563504
3	GradientBoostingRegressor	40012.812908	23.760827	50025.399423	0.562427
4	XGBRegressor	40762.265292	24.182035	51090.677486	0.543593
5	RandomForestRegressor	42266.237859	24.831472	52881.862804	0.511030

(d) Table: Results of different Regression models



(e) Comparison between Predicted Price and Actual Price with Size based on Ridge model



(f) Correlation Heatmap

Figure 1: **Distribution, Correlation and Results**

A Appendix: Additional Analysis Details

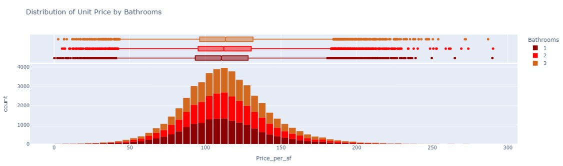
	Model Name	Mean Absolute Error MAE	Mean Absolute Percentage Error MAPE(%)	Root Mean Squared Error RMSE	R2 score
0	Ridge	39951.500703	23.689901	49962.793513	0.563522
1	Lasso	39952.475267	23.700204	49964.317939	0.563495
2	LinearRegression	39956.392148	23.718787	49969.807470	0.563399
3	GradientBoostingRegressor	40012.812908	23.760827	50025.399423	0.562427
4	XGBRegressor	40762.265292	24.182035	51090.677486	0.543593
5	RandomForestRegressor	42295.521628	24.838698	52904.343882	0.510614

(a) Log transformation of Size

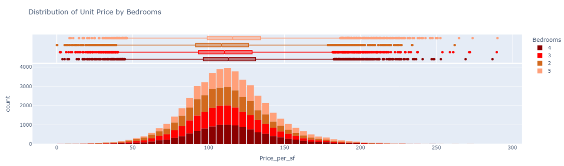
	Model Name	Mean Absolute Error MAE	Mean Absolute Percentage Error MAPE(%)	Root Mean Squared Error RMSE	R2 score
0	Ridge	39954.453643	23.713544	49967.111863	0.563446
1	Lasso	39956.191980	23.718603	49969.564841	0.563403
2	LinearRegression	39956.259301	23.718806	49969.655376	0.563402
3	GradientBoostingRegressor	40013.268391	23.760953	50025.758985	0.562421
4	XGBRegressor	40762.265292	24.182035	51090.677486	0.543593
5	RandomForestRegressor	42276.404059	24.846065	52865.545037	0.511331

(b) Power of 0.9 transformation of Size

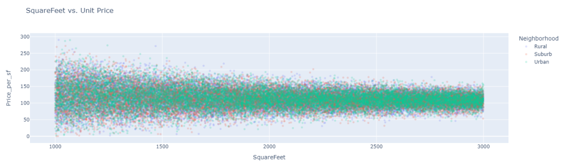
Figure 2: **Results for different transformation of Size**



(a) Distribution of Unit Price grouped by Bathrooms



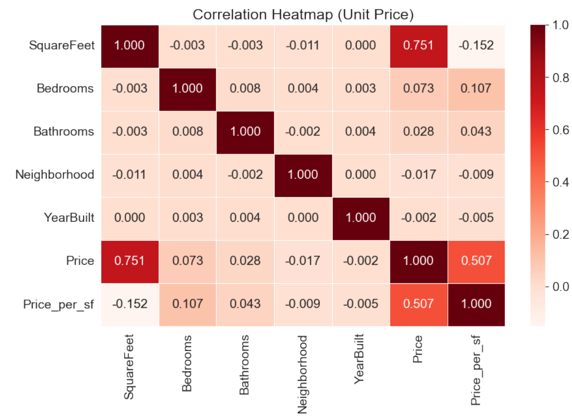
(b) Distribution of Unit Price grouped by Bedrooms



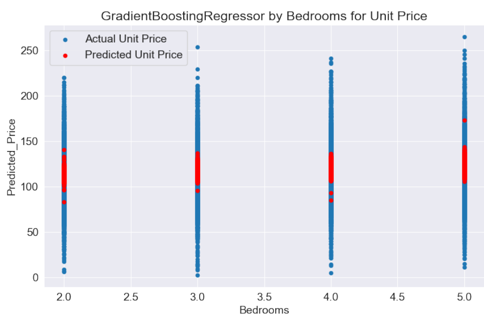
(c) Correlation between Size and Unit Price grouped by Neighborhood

	Model Name	Mean Absolute Error MAE	Mean Absolute Percentage Error MAPE(%)	Root Mean Squared Error RMSE	R2 score
0	Ridge	39952.016876	23.698109	49963.589764	0.563508
1	Lasso	39952.075005	23.696517	49963.822024	0.563504
2	LinearRegression	39952.077259	23.696484	49963.828746	0.563504
3	GradientBoostingRegressor	40012.812908	23.760827	50025.399423	0.562427
4	XGBRegressor	40762.265292	24.182035	51090.677486	0.543593
5	RandomForestRegressor	42306.322881	24.840810	52917.111197	0.510377

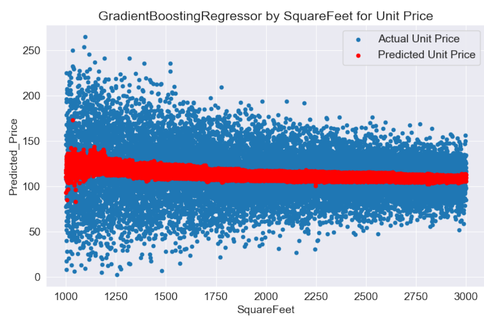
(d) Table: Results of different Regression models for Unit Price



(e) Correlation Heatmap for Unit Price



(f) Comparison between Predicted Price and Actual Price with Bedrooms based on GradientBoostingRegressor model



(g) Comparison between Predicted Price and Actual Price with Size based on GradientBoostingRegressor model

Figure 3: **Distribution, Correlation and Results for Unit Price**

B Appendix: Supplementary Graphs

	SquareFeet	Bedrooms	Bathrooms	Neighborhood	YearBuilt	Price
0	2126	4	1	Rural	1969	215355.283618
1	2459	3	2	Rural	1980	195014.221626
2	1860	2	1	Suburb	1970	306891.012076
3	2294	2	1	Urban	1996	206786.787153
4	2130	5	2	Suburb	2001	272436.239065

(a) First five rows of the data

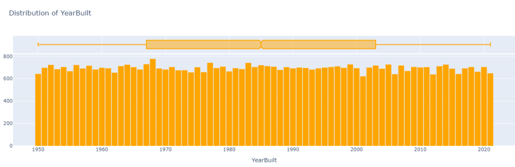
```
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   SquareFeet    50000 non-null  int64
 1   Bedrooms      50000 non-null  int64
 2   Bathrooms     50000 non-null  int64
 3   Neighborhood  50000 non-null  object
 4   YearBuilt     50000 non-null  int64
 5   Price         50000 non-null  float64
```

(b) Detailed Summary of DataFrame

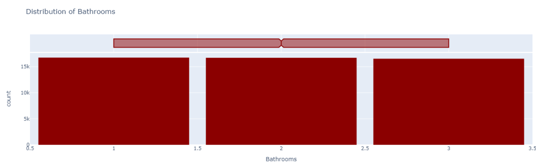
Figure 4: **The basic information of DataFrame**

	SquareFeet	Bedrooms	Bathrooms	YearBuilt	Price
count	49978.000000	49978.000000	49978.000000	49978.000000	49978.000000
mean	2006.752551	3.498659	1.995458	1985.404338	224931.667960
std	575.350298	1.116325	0.815859	20.718407	75995.682992
min	1000.000000	2.000000	1.000000	1950.000000	154.779120
25%	1514.000000	3.000000	1.000000	1967.000000	170007.487130
50%	2008.000000	3.000000	2.000000	1985.000000	225100.123857
75%	2506.000000	4.000000	3.000000	2003.000000	279395.826288
max	2999.000000	5.000000	3.000000	2021.000000	492195.259972

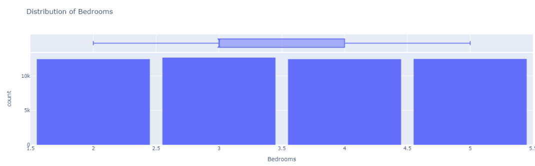
(a) Summary distribution of the data



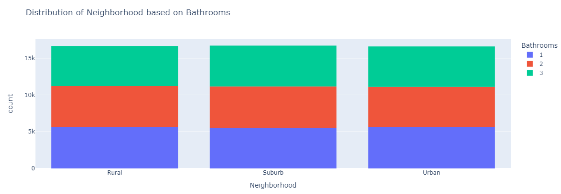
(b) Distribution of YearBuilt



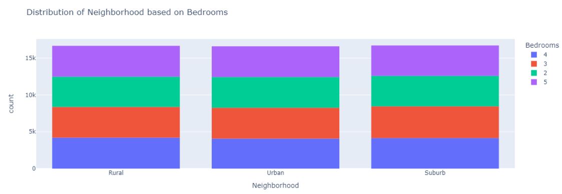
(c) Distribution of Bathrooms



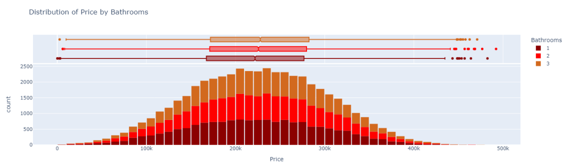
(d) Distribution of Bedrooms



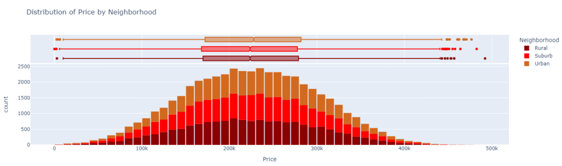
(e) Distribution of Neighborhood grouped by Bathrooms



(f) Distribution of Neighborhood grouped by Bedrooms

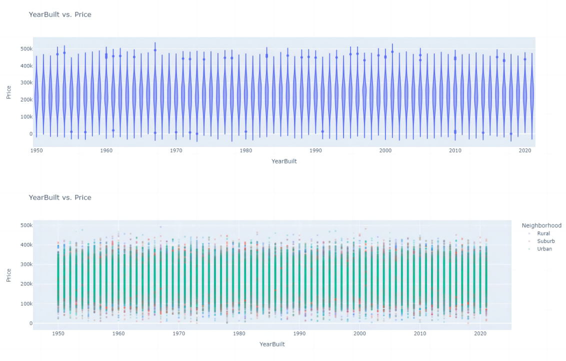


(g) Distribution of Price grouped by Bathrooms

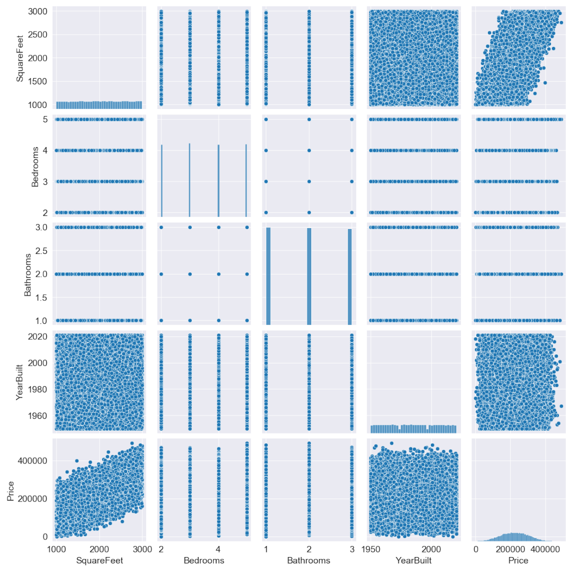


(h) Distribution of Price grouped by Neighborhood

Figure 5: **Distribution of Columns**



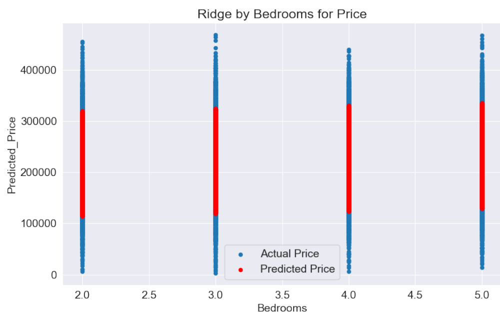
(a) Correlation between YearBuilt and Price



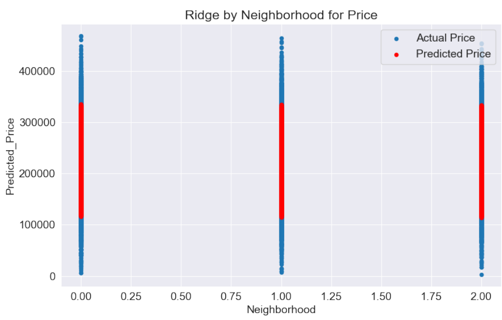
(b) Correlation on Pairplot

Figure 6: **Correlation of Data**

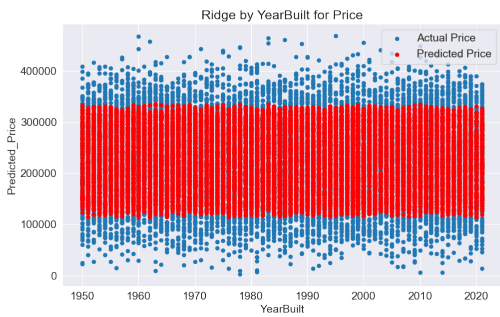
(a) Comparison between Bathrooms and Price



(b) Comparison between Bedrooms and Price



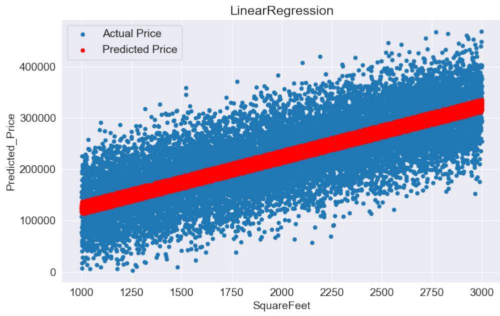
(c) Correlation between YearBuilt and Price



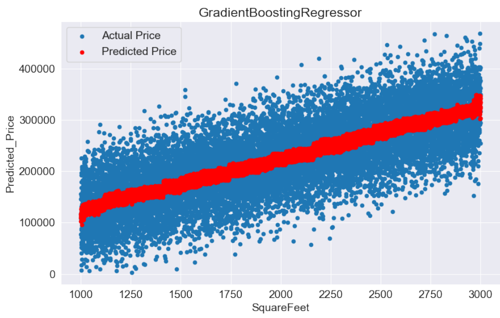
(d) Comparison between YearBuilt and Price

Figure 7: **Comparison between Predicted Price and Actual Price with other factors based on Ridge model**

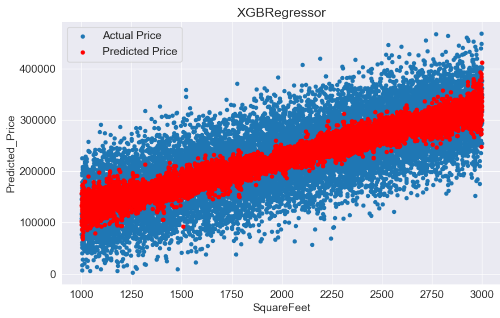
(a) Comparison based on Lasso model



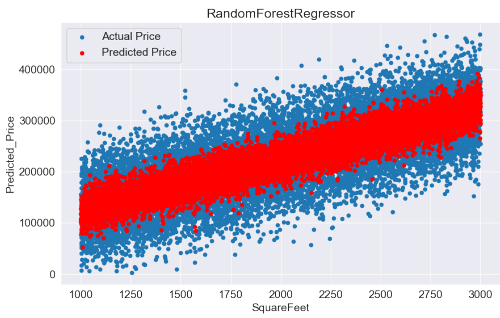
(b) Comparison based on LinearRegression model



(c) Comparison based on GradientBoostingRegressor model



(d) Comparison based on XGBRegressor model



(e) Comparison based on RandomForestRegressor model

Figure 8: **Comparison between Predicted Price and Actual Price with Size based on other models**