



NUS
National University
of Singapore

NUS
BUSINESS
SCHOOL

DBA 5106 - Foundations in Business Analytics

Build Or Not?

Prediction Of Housing Prices & Profitability Of Home Additions



Group 23

Student Name	Student ID
ANKIT MALHOTRA	A0232322X
CHEN ZIYI	A0231946B
KYLE ASANO	A0231984X
PHILIPPINE GALLOT	A0231973B

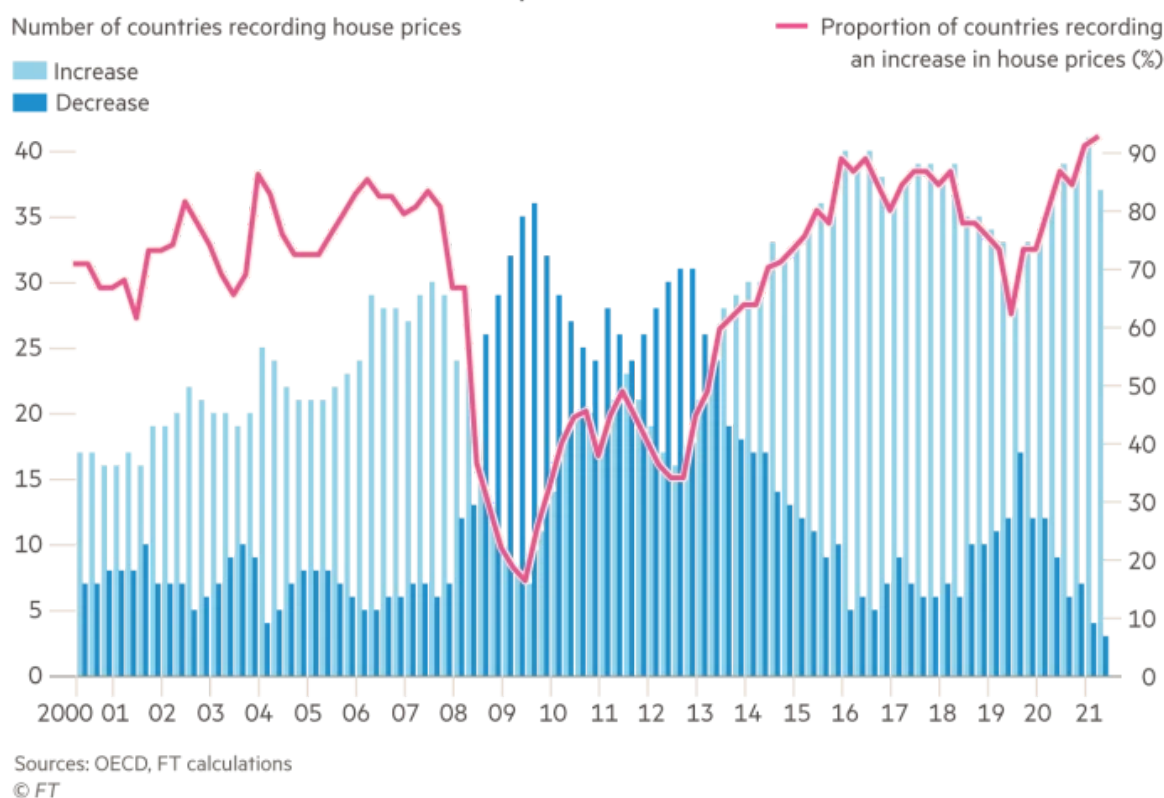
Table of Contents

Table of Contents	2
Introduction	3
Data Preparation	4
Data Collection	4
Dealing with missing values	5
Dealing with outliers	5
Data Exploration	6
Deep Dive on Analytical Models	7
Lasso Regression Model	7
Below is the plot showing the prediction error of our model:	9
Random Forest, Gradient Boosting and Bagging	9
EV Framework for home additions	11
Model Evaluation and Selection	12
Choice of the best model	12
Expected Value (EV) Results	12
Conclusion	15

I. Introduction

In the wake of the Covid-19 crisis and its tremendous socio-economic consequences, the housing market has been reshaped by a number of forces. On September 2nd, 2021, CNBC referred to the property market as being “*on steroids*”. Indeed, global house prices have never grown so fast over the last decade, as the figure below, extracted from the Financial Times, suggests. An economist at Goldman Sachs analyzed that housing markets in the US, Canada, the UK, and New Zealand “*are on fire*” as “*low interest rates and the shift to working from home are fuelling housing demand*”, and this trend seems to apply elsewhere as well.

Pandemic drives broadest house price boom for two decades



Yet, while prices are soaring, home buyers may want to rely on fundamental and objective criteria to evaluate whether the price of listing is a fair price or not.

This report tries to answer this question for residential properties based in Ames City (Iowa, USA). Our motivation for the problem derives from the perspective of an ordinary home buyer who faces a lot of dilemmas when choosing to purchase a flat taking into consideration several factors including pricing of the apartment, overall quality, year built/ last renovated, number of bedrooms and bathrooms, to name a few. More specifically, our goal is to predict the sale price of houses based on a large number of property characteristics. We can easily understand how

cumbersome it would be for a person to read the listings of all these houses and all these features, and to be able to infer precise and accurate criteria on features to predict a housing price, and this is precisely where the modelization through machine learning is relevant in this case.

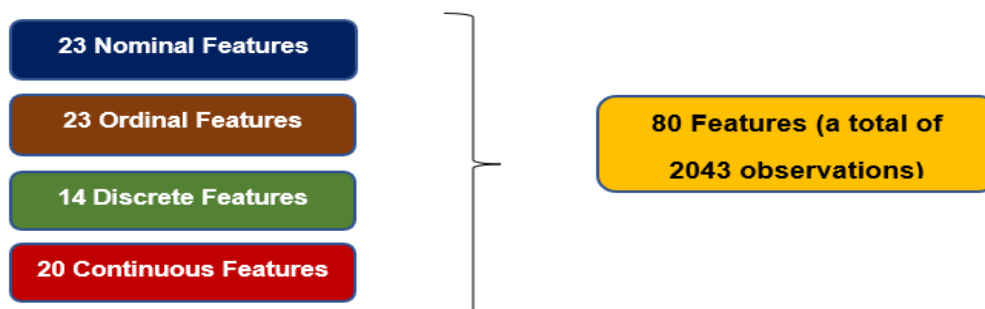
When taking the perspective of a home developer, the model could be extended to be part of the expected value (EV) framework, where a home developer could plan and decide home additions based on their profitability and how widely to target / build the addition. Similar models have been used in US “iBuyer” companies like Zillow and Opendoor, where a property is purchased from a seller, given an addition / upgrade, and then shortly thereafter brought back onto the market on their own platform (in US terminology, “home flipping”). The model in this project will select a range of home additions and apply the EV framework, as a starting point for planners; it will also highlight where underestimating uncertainty and over targeting can lead to losses. The latter is prescient, as iBuyer companies have faced a raft of issues in profitability (Opendoor has run net losses for the past 2 years) or exuberance (Zillow altered the weights of their models, leading to more purchases before cooling house prices; it has commenced a 25% retrenchment and incurred a \$420 M quarterly loss, see References); with correct tuning, it is hoped the EV framework can demonstrate the losses and upsides associated with “home flipping”.

In the forthcoming sections, we will walk through the data preparation, deep diving into analytical models used for prediction and model performance, and finally concluding with business applications, future recommendations, and eventual limitations.

II. Data Preparation

Data Collection

Our source of data was the Journal of the American Statistical Association. Below is an infographic that represents our dataset and breakup of its types in measurement scales:



Our dataset contained 2,043 observations and 80 features containing comprehensive information on largely each and every aspect considered such as overall condition of the house, overall material and finish of the house, original construction date, remodel date, type of

foundation, basement condition etc. when making a housing purchase decision. The dataset pertains to residential properties of the Ames city in Iowa state of US and describes a large number of features, from its interiors (number of rooms, frontage, area, heating quality, etc.) to its surroundings (name of the neighborhood, street, etc.). For our analysis, the target variable is the Sale Price, and the 79 remaining variables are the features we will use to predict it.

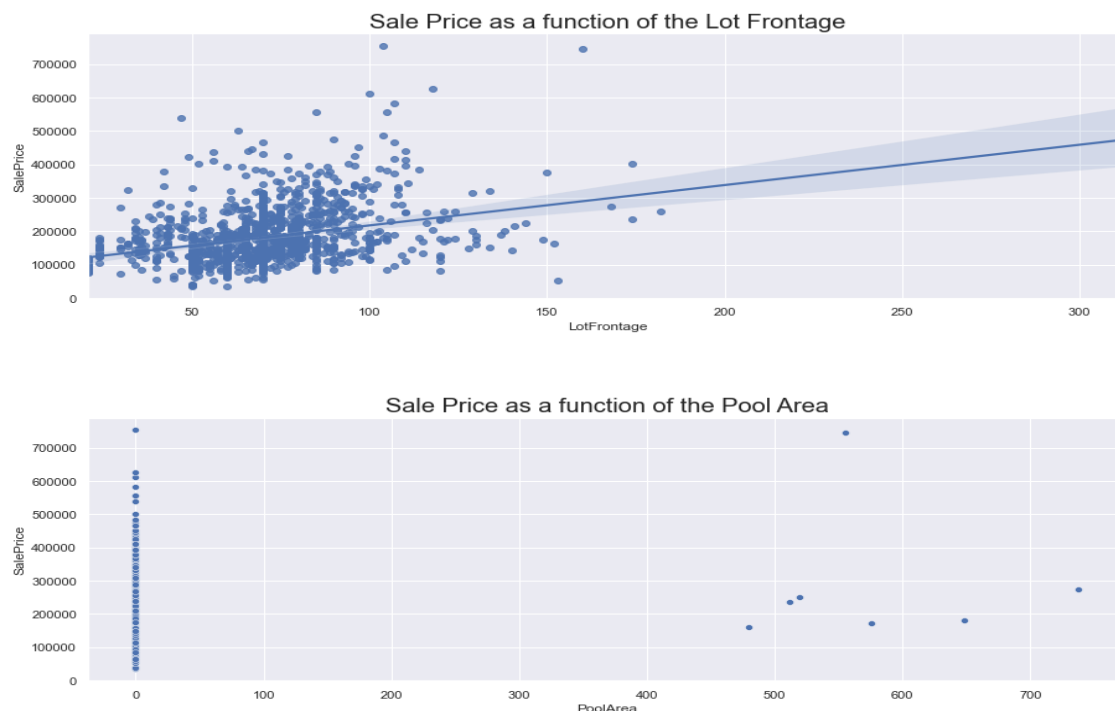
Dealing with missing values

We observed that our dataset had missing values for a few features. First, we removed the rows where the Sale Price is missing since they would have not helped us train or validate our models. We were then left with 1,460 observations. Then, we removed the features with over 50% of the missing data while for features with less than 50% of the data missing, we imputed the values with mean/ median, as appropriate.

Dealing with outliers

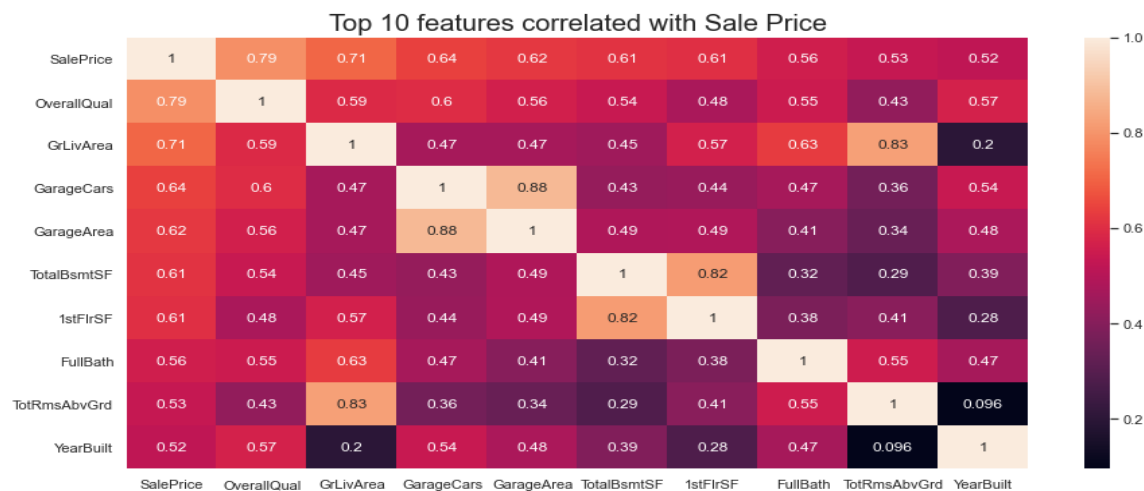
From an initial exploration of the dataset, we noticed the existence of a small number of outliers (see the graphs below). We usually want to get rid of outliers because they increase the error variance, can cause bias and influence our estimates.

A solution to get rid of these outliers is to perform a Robust Scaler standardization. This Scaler removes the median and scales the data according to the quantile range (which, by default, is the inter-quantile range). We performed this transformation after splitting the data into training and testing sets, and only using the data from the training set, as using any information coming from the test set would bring a potential bias in the evaluation of the performance.

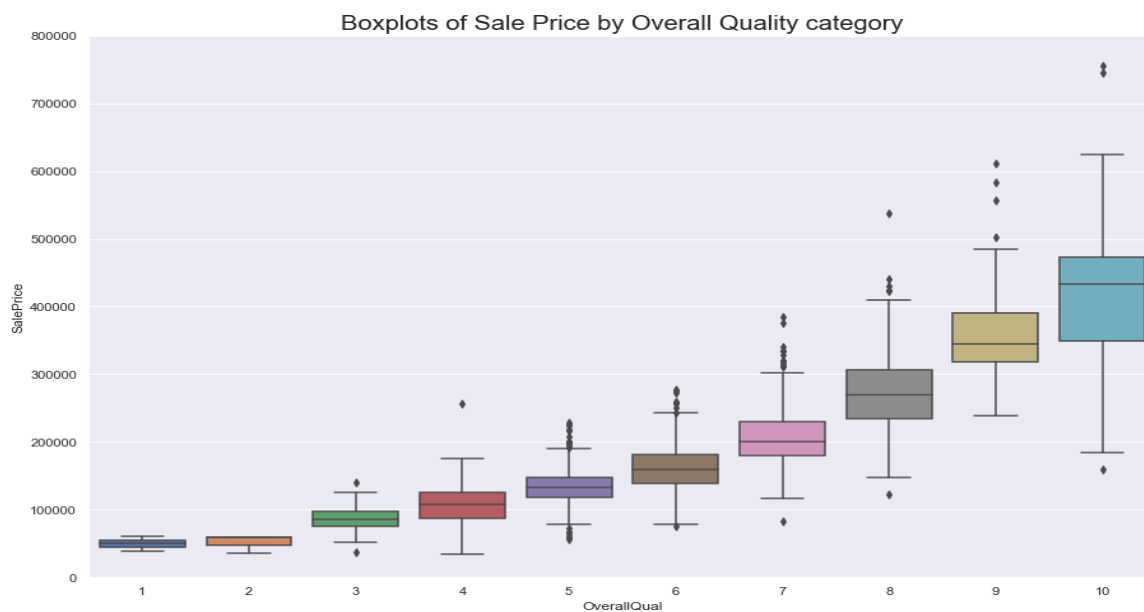


Data Exploration

Based on the initial exploration of the dataset, we observed that the sale price of a house has a high positive correlation with the overall quality of the house and with above ground living area square feet, which is quite intuitive.

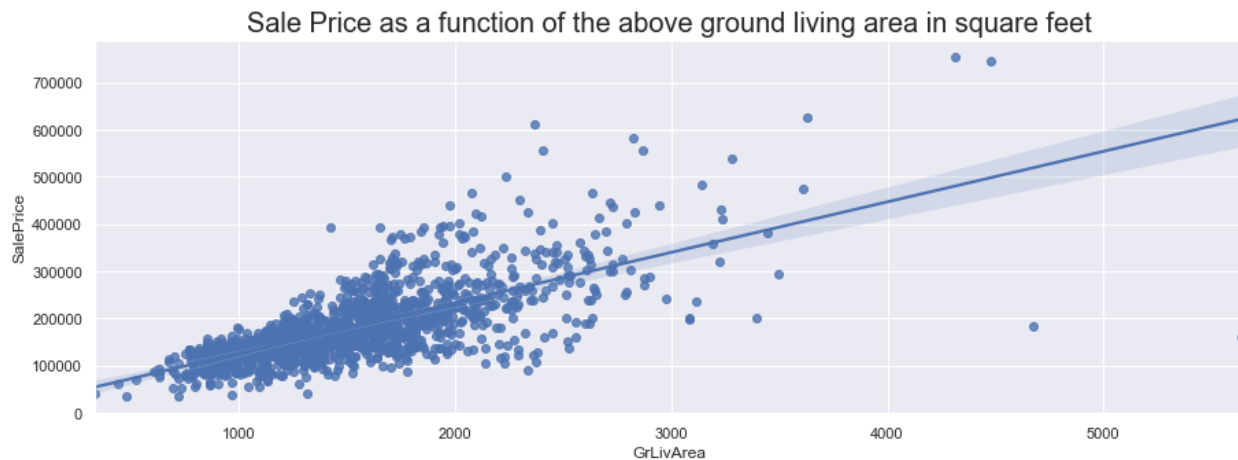


Let's further explore the relation between the Overall Quality ('OverallQual') and the Sale Price, as we can see that there is a 79% correlation between the two. The description given on the data describes the Overall Quality as the features which rates the overall material and finish of the house.



From the figure above, we can see that lower the overall quality, the more accurately the overall quality predicts the sale price of the unit. It seems like given the overall quality of a unit, we can predict quite accurately the sale price of the house.

Let's also explore the 2nd feature in the list of those most correlated to the sale price, which is 'GrLivArea', the above ground living area (in square feet). The graph below shows that there is a clear linear relation between the two.



These analyses enable us to understand the data more in depth and to have a better intuition on the important features. We can now move on to the machine learning part of the project, i.e. we are going to fit models on our data in order to do predictions.

III. Deep Dive on Analytical Models

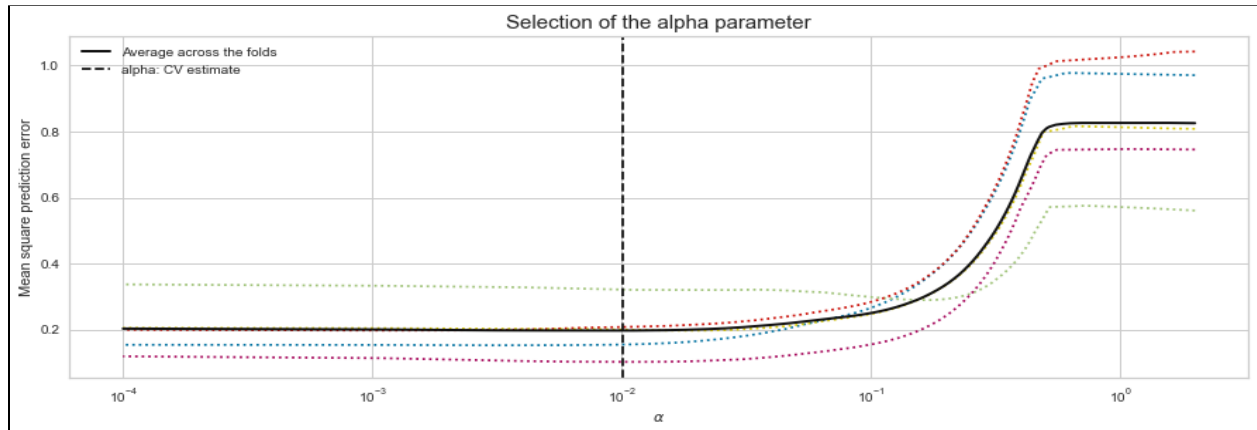
In this report, we chose to focus on 4 regression models: Lasso, Random Forest, Gradient Boosting and Bagging. We are also presenting an analysis based on the expected value framework, and a stochastic variant of it.

Lasso Regression Model

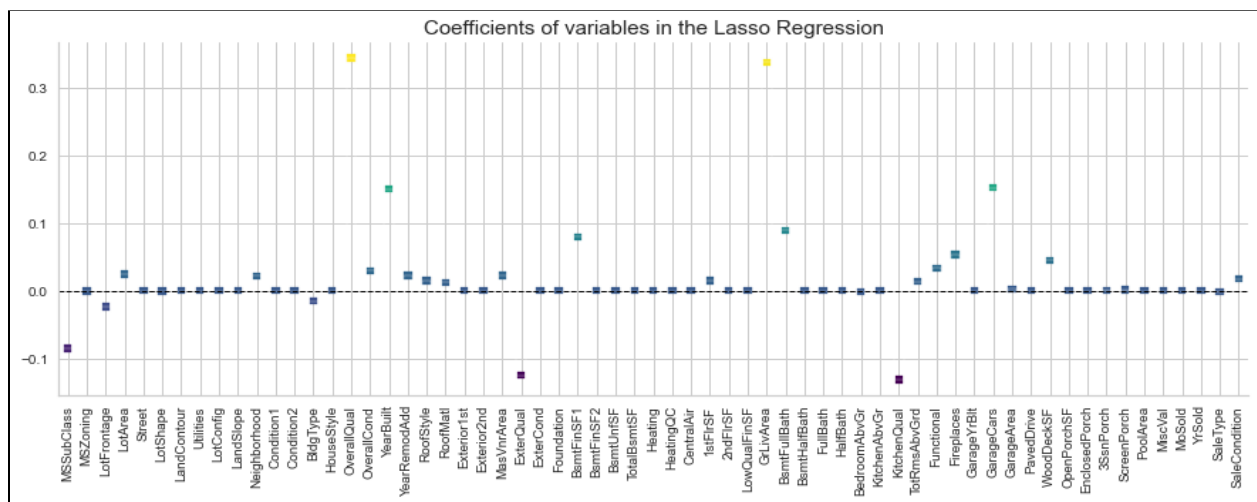
The dataset entails many features. It is reasonable to think that some of the features are not relevant or not significant and should not be included in our model. Therefore, it is appropriate to use Lasso Regression which can set the value of some coefficients to zero and exclude them from the regression.

First, we finalize the data preparation, which entails the encoding of 27 categorical features. Then, we split the data between training set (75% of the data) and test set (the remaining 25%). We can then standardize the data using the abovementioned Robust Scaler. Notably, we only use the data from the training set to perform this standardization, as using any information coming from the test set before or during training is a potential bias in the evaluation of the performance.

We perform the Lasso regression using *scikit-learn's* *LassoCV*, which not only fits a Lasso regression model on the data, but also performs cross-validation to determine the optimal alpha parameter, which is the regularization parameter. The model outputs an alpha of 0.0101, selected as the alpha giving the minimal mean square error on the average of the folds' mean square error on the figure below:



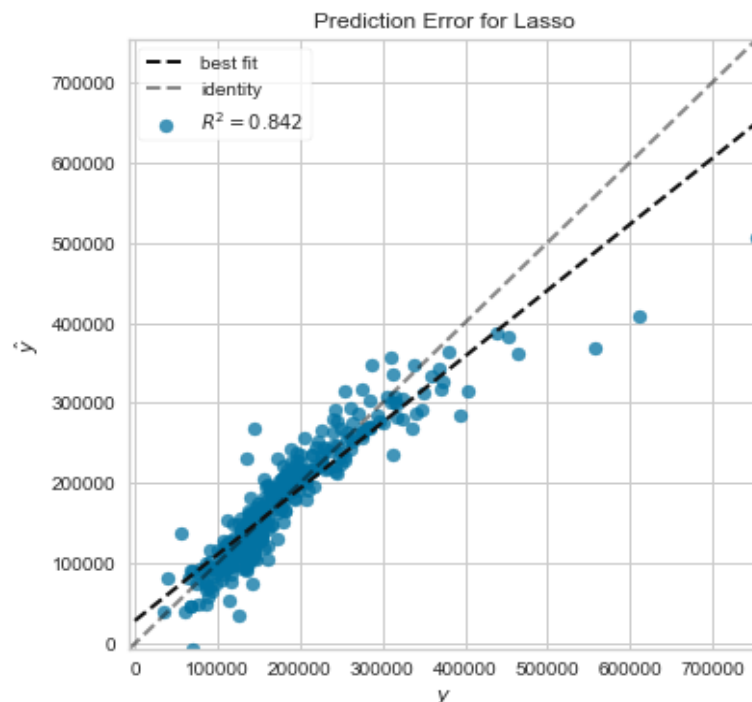
The model sets 27 coefficients to zero, as we can see in the figure below.



The model performs quite well to the extent that it does not overfit the data. Indeed, the out-of-sample Root MSE (mean square error) and MAPE (mean absolute percentage error) are equal or even below the in-sample metrics.

Metric	In-sample	Out-of-sample
Root MSE	196,555.59	196,090.42
MAPE	99.3%	99.3%

Below is the plot showing the prediction error of our model:



Random Forest, Gradient Boosting and Bagging

In order to reduce the variance and further control the risk of overfitting, we also explored a ensemble methods in advanced regression which includes Bagging, Random Forest(RF) and Gradient Boosting (GB) on the full training dataset before using just the important features to predict the sale price.

The data pre-processing and train-test-split steps are the same as the above model. *scikit-learn's RandomForestRegressor* is used to perform random forest regression. We explored different numbers of predictors (m) to get the optimal m from lowest possible out-of-sample root MSE error. In order to get an overall summary of the impact of each predictor, a variable importance plot is obtained to get the important features. When m is set to the full set of 77 predictors, it becomes bagging. GB Regressor builds an additive model in a forward stage-wise fashion. It iteratively builds weak regressor and aggregates them into a strong regressor. Results are shown in Table1.

While bagging and random forest improves prediction accuracy, it forgoes the interpretability. However, we can still get an overall summary of the impact of each predictor. By obtaining a variable importance plot, we learned that the top 12 variables that had the importance above 0.1(Table 2).

Metric	In-sample performance			Out-of-sample performance		
	Bagging	Random Forest	Gradient Boosting	Bagging	Random Forest	Gradient Boosting

m	77	60	76	77	60	76
Root MSE	11514.64	11529.94	15015.43	33642.69	33046.87	31381.51
MAPE	0.040	0.040	0.071	0.010	0.10	0.095

Table1 Performance Comparison of different models with all features

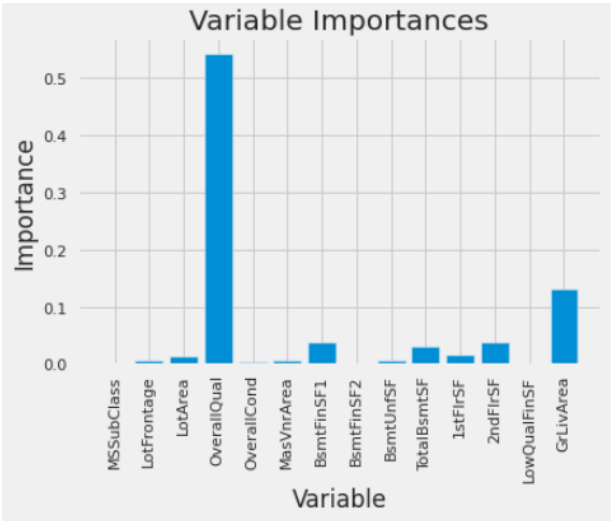
Variable	Importance	Variance Importance plot
OverallQual	0.55	
GrLivArea	0.14	
TotalBsmtSF & 2ndFlrSF	0.04	
BsmtFinSF1& 1stFlrSF	0.03	
LotFrontage & Fullpath & TotRmsAbvGrd & Fireplaces & GarageCars	0.01	

Table 2 Variable Importance from Random Forest

By performing RF Regressor and GB Regressor models using the top 12 important features, the model did not yield better performance. For RF, The best out-of-sample root MSE is 33035.75 and MAPE is 12.5% when $m = 11$. For GB, performance is the same under $m=1$ to 12 and the out of sample RMSE is 38222.62 and MAPE is 17.17%. Both models performed worse than the Gradient Boosting Model with all features.

Hence, with more available data and more number of predictors, gradient boosting performed the best with the smallest error by using weak regressors step by step and gradually improves from the information obtained from the previous regressor.

EV Framework for home additions

For determining if a home addition is profitable or not, the EV framework is applied to selected features in the dataset - ``2ndFlrSF``, ``OpenPorchSF``, and ``WoodDeckSF`` indicate the presence of a 2nd storey, an open porch, or a wood deck respectively. After conversion to binary variables (where a feature is assumed to present if its original scalar was greater than 0), this can be fed as the target to a classification model (``XGBClassifier``), fitted to the training data

and predicted against the test data, and then the class probabilities can be compared against the thresholds for generating confusion matrices. In turn, these confusion matrices would be multiplied by the 'cost priors' or rewards and penalties associated with each confusion matrix outcome; under naive assumptions, this could be the mean cost or reward, or a fixed scalar informed by domain experts.

In our solution, we explore a naive assumption (as shown below), and then a stochastic scenario with means generated by 2-stage least squares regression (2SLS). The former is simple to implement and easier to interpret, but is highly subject to confounding variables (e.g. the difference in means between home with/without an addition can be influenced by a confounding variable; homes with the addition can have greater variability). Ergo, the stochastic scenario returns the range of profits and losses for a given addition, allowing planners to choose between greatest reward or least variation.

True Positive Mean marginal <code>SalePrice</code> from home addition	False Positive Mean per-square-foot cost of building addition * Mean square feet of home addition ¹
False Negative No reward or cost	True Negative No reward or cost

¹ Average generated by US real-estate-tech companies HomeLight, Fixr, and Angi HomeAdvisor (See References)

The steps below are used for the stochastic scenario:

1. Generate a binary variable for the selected home addition
2. Generate a list of instrument variables (IV) that could proxy the home addition (endogenous variable) in relation to `SalePrice` (dependent variable)
3. Feed the IV, dependent variable, endogenous variable, and other remaining exogenous (independent variable) to 2SLS
4. Confirm that endogenous variable is valid through its p-value and T-statistic
5. Check that fitted 2SLS model passes the Woolridge and Hausman tests for exogeneity
6. Generate reward for true positive from 5 samples, from a distribution based on the 2SLS mean and standard deviation
7. Proceed with all other steps for EV for each sample true positive reward

Using causal inference for the true positive, the EV framework can compare marginal rewards tightly attributed to the home addition, rather than a simple difference in means (which also does not include variability). The only assumption unchanged between EV and stochastic EV is that the opportunity cost of not building a home addition, the false negative, need not be included since it has no cash flow impact.

IV. Model Evaluation and Selection

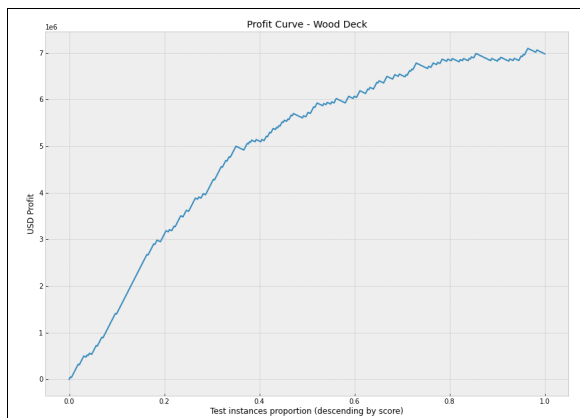
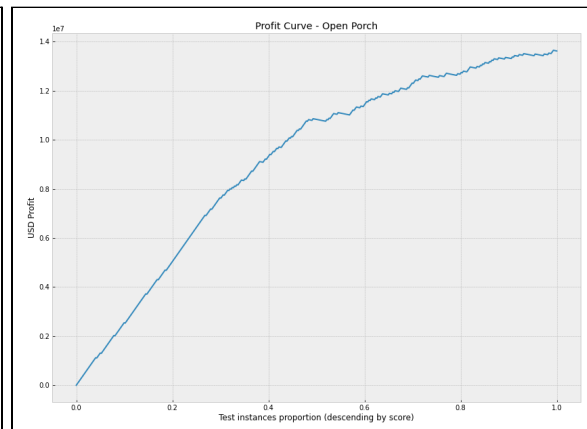
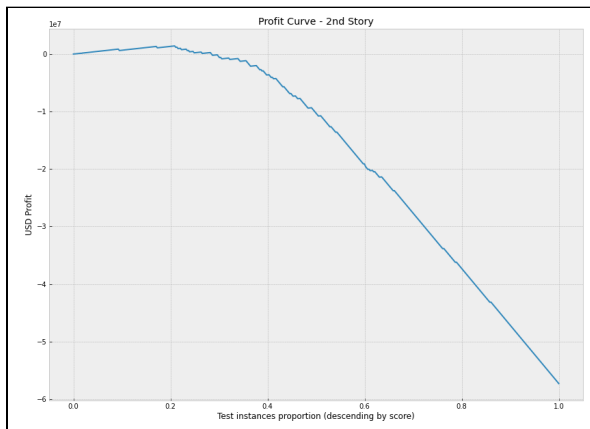
Choice of the best model

Our choice needs to be motivated by two factors. First, we need a good performing model given what is at stake: a model with poor performances could mean huge losses for the real estate agent or the home buyer, depending whether the price is overestimated or underestimated. Besides, we may want to choose a model easily interpretable as the stakeholders are likely to want to understand why a house was priced in a certain way. Therefore, gradient boosting is a strong greedy algorithm that iteratively learns from weak regression trees and aggregate to a strong regressor. With more data trained, it will become a stronger model in forward steps. With a low MAPE, it is suitable to predict future housing sale prices as well.

Although the Lasso Regression, as it is a linear model, is more easily interpretable, we decide to choose the Gradient Boosting model. Indeed, Lasso underperforms significantly. For Gradient Boosting, the out of sample lowest error metrics with a RMSE of 31381.51 and MAPE of 9.5%, vs. a RMSE of 196,090.42 and MAPE of 99.3% for the Lasso model. The latter offers performances which are too poor, so it would not be recommended to a potential client.

Expected Value (EV) Results

Native Assumptions - Difference in Means



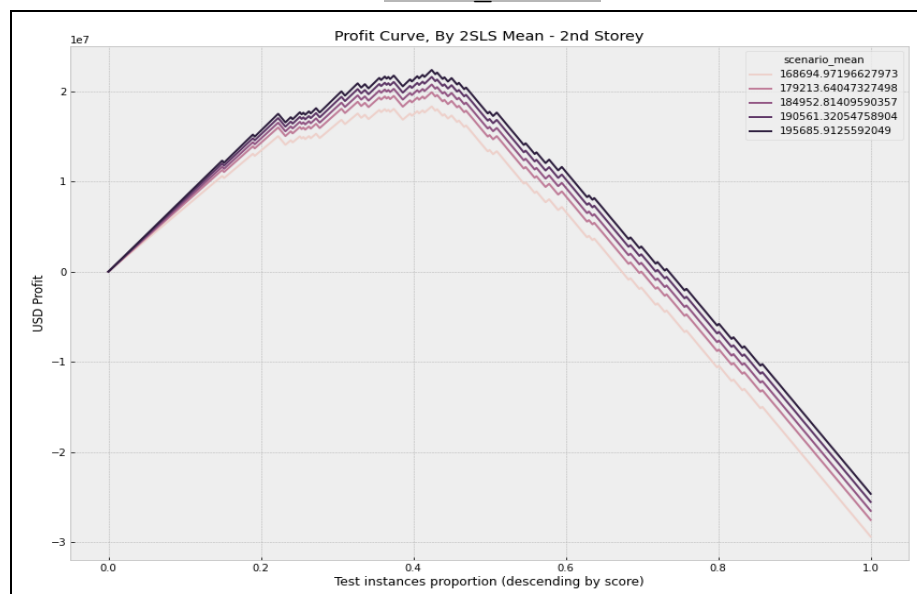
	TP Reward	FP Penalty	Max. Profit	Threshold
`class_porch`	\$65,930	83.9 sq.ft. * \$110	\$13,644,209	0.002
`class_wooddeck`	\$47,320	195.7 sq.ft. * \$60	\$7,092,433	0.013
`class_floor2`	\$22,067	800.7 sq.ft. * \$300	\$1,549,818	0.97

As shown above, the EV framework has an optimistic outlook for `class_porch`, with the highest maximum profit and most generous threshold; although it has higher per-square-foot costs, the mean size of such additions are smaller.

By contrast, the EV framework has the most pessimistic outlook for `class_floor2` (the 2nd storey addition), with the lowest maximum profits, and indicating only a small proportion of customers should be targeted. This is largely driven by the high cost of false positives, along with the low reward of true positives - of all the additions, `class_floor2` has the lowest difference in means.

At a high level, the EV framework's output (under naive assumptions) does fit some preconceptions - "building a 2nd storey is a costly affair; porches are simple to build and aesthetically pleasing". However, the naive assumption uses a difference in mean `SalePrice` that may include confounding variables, along with being a rudimentary attribution technique. The EV framework could be fed using a similar technique, using the coefficients of linear regression (with known collinear variables removed), but this still relies on manual curation of variables. Therefore, the EV framework could be fed sample means from the stochastic scenarios from 2SLS.

Stochastic EV With Means From 2SLS - `class_floor2`



Scenario Means	\$168,695	\$179,214	\$184,953	\$190,561	\$195,686
Maximum Profit	\$18,338,160	\$19,915,960	\$20,776,830	\$21,618,110	\$22,386,800
Maximum Loss	-\$29,401,330	-\$27,550,040	-\$26,539,950	-\$25,552,850	-\$24,650,920

Under stochastic scenarios with means drawn from 2SLS coefficients, the EV framework generates a much larger range of profits for `class_floor2`. In particular, the higher maximum profit is driven by the high marginal price of 2nd storey derived from 2SLS - in a sample of 5 means, the lowest reward is still 6x higher than the reward under naive assumptions. To initially test robustness of this coefficient, the Wooldridge regression test and Hausman test of exogeneity were applied to the 2SLS results, with both indicating that the formula for `class_floor2` and `SalePrice` was valid (p-value 0.000 with statistic 325.0579 and 316.6404 respectively).

The difference between 2SLS's coefficient for `class_floor2` and naive assumptions may be attributable to errors in the instrumental variables used for the 2SLS (`OverallQual` and `LotArea`), or (if done correctly) may be attributable to differences in the distribution of 1 storey or 2 storey homes. As a simple heuristic, the standard deviation of houses with 2nd storey is higher than that of 1 storey (84599.6, 73921.1 respectively), in addition to higher means. Nonetheless, with further tuning of instrumental variables and the 2SLS formula, the EV framework under stochastic scenarios could be used by residential planners for seeing a wider range of profits, and therefore a more comprehensive picture when planning home additions.

V. Conclusion

After exploring several models on predicting the sale price, we have chosen gradient boosting(GB) regression as the best model to be used for future prediction of saleprice. While GB may have lower interpretability than generalized linear regression model, it has much better performance in predicting more accurate sale price which is paramount for our client. Inaccurate prediction would be very costly for the home developers. With more available data, GB as the greedy stepwise algorithm that learns from weak regression tree and aggregate to a stronger model.

If deployed to iBuyer companies or residential development planners, the EV framework (and it's stochastic variation) could also provide a range of both profits for home additions, and how to deploy them. The framework could identify how much of residential district should have a certain addition, as well as the expected losses if a policy must be globally implemented (e.g. "due to legislation, all homes must be only 1 storey - what are the possible losses?").

As further improvements, the EV framework could also introduce stochasticity to the construction costs of additions, to reflect greater downsides when faced with supplier uncertainty. We would expect this to reduce the maximum profit displayed, but at the same time

give decision makers the flexibility of deciding between additions that have high reward, or the narrowest (and most certain) range of outcomes. This is directly applicable to the recent issues faced by iBuyer companies, where improvements to their bottom line can be made by selectively rolling out home additions - rather than adjusting the parameters and weights, which can lead to overoptimistic scenarios and losses.

VI. References

- <https://ireus.nus.edu.sg/wp-content/uploads/2020/10/iBuyers-Liquidity-in-Real-Estate-Markets-by-Tomasz-Piskorski-.pdf>
- <http://jse.amstat.org/v19n3/decock.pdf>
- <https://www.nerdwallet.com/article/mortgages/understanding-ibuyers>
- <https://content.knightfrank.com/research/84/documents/en/global-house-price-index-q1-2021-8146.pdf>
- <https://www.homelight.com/blog/cost-to-add-a-second-story/>
- <https://www.fixr.com/costs/porch-addition>
- <https://www.homeadvisor.com/cost/outdoor-living/build-a-porch/>
- <https://www.homeadvisor.com/cost/outdoor-living/build-a-deck/>
- <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
- <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html>

-----End of report-----