House Price Predictive Model

Kyle Kim

Abstract

House is an asset that accounts for the majority of assets for average Americans. However, unlike stock prices, which have numerous metrics for pricing, house prices lack metrics for correct valuations. To my knowledge, the only house price metrics I have ever heard is the house index which functions like Dow Jones Industrial Average, S&P500, and NASDAQ. While I have been studying the housing market, mainly for BRRR strategy, a lot of resources I found on the internet normally identified certain prices cheap by comparing to neighboring houses. Such an approach is reasonable, however, lacks on separating supply or demand shock that any types of assets experience. For example, at the time of writing this, Nikola is valued at the market cap of $4.15B with $0 Sales and $-673.3M Income. In the past year, its stock prices saw a wide swing between about $90 and $ 10 per share. Applying this story to the housing market, if each share is counted as one house, one could have bought a house for $70 and thought it cheap simply because other houses were sold at $90. This is a bit of an extreme story since the

volatility of house prices is not as severe as that of stock prices, but people have experienced many house bubbles across the world in the past 30 years. The motivation for this model is to create a predictive model that reasonably prices a house to at least tell a person if he or she is overpaying or underpaying for it. If the model conducted in this analysis performs well, it should be tested on data from prior years. By doing so,

Data

The data is obtained from Kaggle, a website that holds many datasets and data science competitions. The dataset has a total of 2919 observations and 81 variables. 38 of 81 variables are numeric, and all others are categorical variables. Based on the variable description provided by the content holder, all variables reflect houses like the material used to build and sizes.

There were numerous missing values. Instead of removing them, KNN imputation was used to fill in the missing variables.

EDA

Because this is a predictive model, EDA was skipped. The success is only measured with test RMSE. The drawback is that each variable should not be interpreted for inferences. In other words, a variable's impact does not reflect the true effect on the house price. Test RMSE is chosen because its unit is the same as the response variable.

Modelling

The modelling started with multiple linear regression using all of the variables then gradually improved through using various methods to lower the test RMSE.

Model 1 – Multiple Linear Regression

As mentioned, model 1 is multiple linear regression with all variables in the model. There is a total of 79 predictors, minus Id and House Price. The model is expected to perform poorly due to multicollinearity. The observations were split into ½ for test and training data.

Model 1	
Test RMSE	$ 76,992.89
$Adj.R^2$	0.9179

Firstly, $Adj.R^2$ of 0.9179 tells that the model explained about 92% of variances in the training dataset, which is quite high. However, on the test dataset, it performed poorly. The five number summary for house sale price is below.

5 Number Summary	

Minimum	$135,751
1st Quantile	$168,697
Median	$179,203
3rd Quantile	$186,804
Maximum	$281,644

Based on the median house price, the test RMSE accounts for approximately 43% which is extremely high. In other words, on average, the model can price as extreme as 43% cheaper or more expensive than the current price. Hence, model 1 did not perform well. One possible reason for poor performance on the test data is overfitting because $Adj.R^2$ is high on the training data. This means that the model failed to generalize the trend. Instead, it accounted smaller trends in the training data which could be just noises. So, when a new dataset is introduced, the test data, the performance is poor. Coefficients of variables, F-test statistics, and P-value for each variable were not considered because the data didn't follow regression assumptions.

Model 2 – Stepwise Variable Selection

Some common solutions to handle overfitting problems are increasing the size of data, dropping variables (variable selection), dimension reduction, and regularization. Since the model currently has at least 79 variables, variable selection is considered first. The reason for having at least 79 variables is the way model handles categorical variables. To account for categorical variables, regression models automatically create dummy variables that are consist of 1s and 0s for each condition met. For example, if a predictor called "type" exists with 3 levels, wood, steel, and concrete, the model will have 2 predictors for the 3 levels. Hence, just for the imaginary predictor "type" alone, the model will have two predictors like "iswood" and "issteel".

For variable selection, there are three common methods. The forward method only adds a coefficient if and only if it improves the model's F test statistics. The backward method first uses all of the predictors to build a model then removes one by one if the model with the removed variable sees improved F-test statistics. Finally, the stepwise method adds and removes simultaneously for each variable. So, it adds a predictor then checks to see if removing any other predictor can improve the model. In this analysis, the stepwise method is chosen because multicollinearity is likely to exist between predictors.

Model 2	
Test RMSE	$ 77,831.38
$Adj.R^2$	0.9186

The outcome of model 2 addresses the same problem as model 1. Through variable selection, the number of predictors in the model is 160, and that of model 1 was 247. Even 87 predictors were removed from model 1, the performance has gotten worse and is still indicating overfitting. This blog post writes that there should be 10-15 observations for each predictor in a regression model. The model is trained on 1460 observations, so it is definitely lower than the ratio the blog post suggests. With the minimum information about the author and the lack of references, it is not safe to follow the suggestion; however, with the presence of many categorical variables, overfitting is still likely to be the case.

Model 3 – Lasso Regression

The variable selection wasn't successful in model 2. While assuming that deeper variable selection is needed, another way of solving overfitting is regularization. Regularization is achieved by decreasing coefficients of predictors to reduce their impact on the model and therefore solves multicollinearity. So, lasso regression reduces variances by decreasing flexibility and at the cost of increased bias. Lasso regression is chosen as a method of regularization because it also removes certain predictors from the model.

For lasso regression, lambda, a discounting factor that reduces coefficients, is chosen by K=5 validation method. The training set is randomly split into 5 equal sizes then trained on 4 chunks and tested on the remaining 1 chunk. Hence, it tests 5 times. The best lambda value is 259.5566.

Model 3	
Test RMSE	$ 74,205.81
Dev Ratio	0.9207111

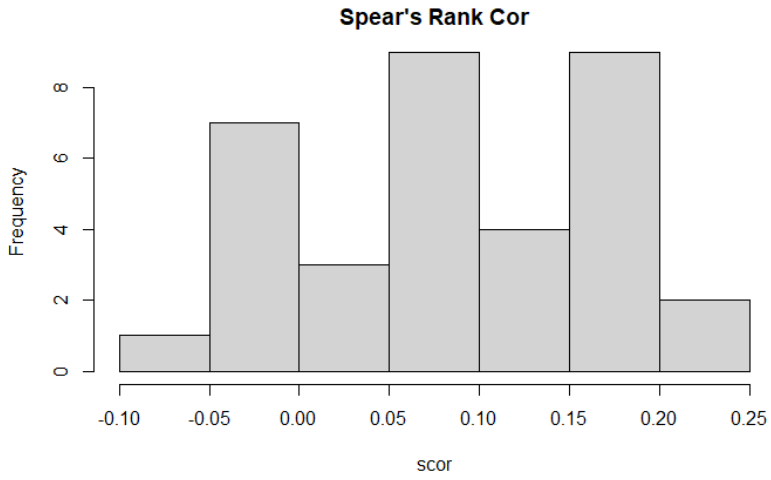
Note: according to the package document, dev ratio has the same equation as R^2.

Both model 2 and model 3 select variables to reduce multicollinearity. Additionally, model 3 uses a regularization method that further reduces the variance. However, the test RMSE has barely improved. Variances of explanation measures and the performance on the test dataset repeatedly show the issue with overfitting even after using methods that tackle overfitting problems.

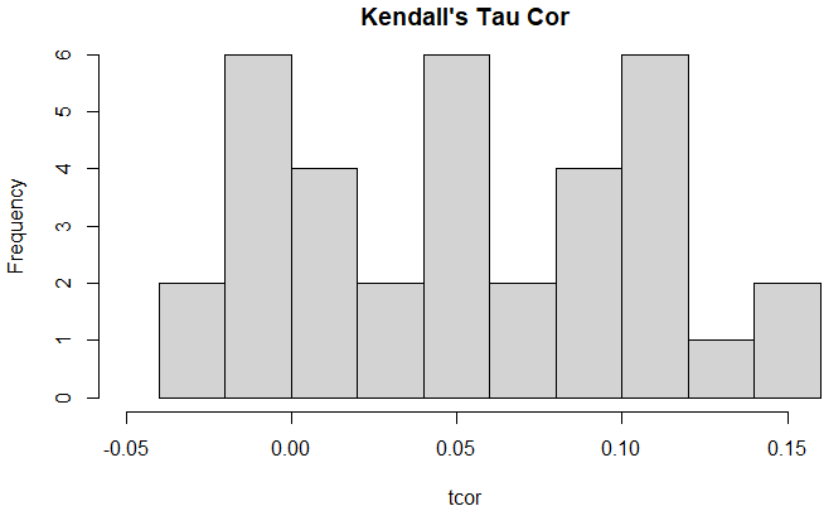
One assumption for the cause of the problem is nonlinearity. The equation for MSE (RMSE is just square root of MSE) is: $MSE = Variance + Bias^2$. Both model 2 and model 3 focused on the variance portion of the test RMSE. Since model 3, a method that reduces variance, has barely improved, it's reasonable to suspect the bias portion of equation. Bias is like a structural problem that the model is missing structural problems like interactions or higher order terms. Before moving on to the next model, variables are reviewed to identify possibility of interactions and higher order terms.

Nonlinearity

To measure nonlinearity, Spear's Rank correlation and Kendall's Tau correlation are checked. The form measures the correlation of ranked data, and the latter measures concordance, the direction of data as one variable exists. These two methods only consider continuous predictors.



The above histogram is showing the distribution of Spear's Rank correlation, and none of them shows significant nonlinear relationship with the response variable.



The histogram above shows Kendall's Tau correlations between the response variable and explanatory variables which aligns with the outcome of spear's rank correlation. In conclusion, these correlations are not showing any evidence of nonlinearity between the response variable and explanatory variables.

Model 4 – Random Forest

Random Forest is a type of bootstrapped tree. At each split of a tree, it takes a random sample of predictors, then takes the average of all trees for generalizations. While previous methods show the importance of variables by the magnitude of coefficients, random forest shows with the percentage increase in training MSE when excluding a predictor. This is the reason why model 4 is discussed before looking at interactions. Due to the large dimension in the data, instead of looking at every possible pair or 3081 pairs (79 Choose 2, excluding the response variable and "Id" column), a few important variables selected by Random Forest will be reviewed.

Model 4	
Test RMSE	$ 66, 853.58
Mean R^2	0.8763151

Test RMSE is the best so far, and, along with its mean R^2, it seems that the overfitting issue might be little relaxed because the mean R^2 decreased by about 5%. Here, it's called Mean R^2 because a total of 500 trees was built, and each tree has their own R^2, so the average was taken. However, as a model, test RMSE of about $ 66,000 is still quiet high given the five number summary of test sales prices earlier. One possible reason why random forest regression outperformed previous models is that it only considers 79/3 or 26 predictors, excluding any cases with dummy variables, at a time which means that it performed more rigorous variable selection than any other method. The caveat is that random forest uses random selection, hence, variable selection is not meaningful.

Interaction

Below is the predictors in the order of importance calculated by the model 4.

GrLivArea	Neighborhood	OverallQual	TotalBsmtSF	X1stFlrSF	GarageArea	BsmtFinSF1	X2ndFlrSF	BsmtFinType1
38.2026615	29.0701129	22.1890876	19.5837563	16.2178099	15.6586675	15.0070560	14.5816064	10.8630041
YearBuilt	LotArea	ExterQual	KitchenQual	MSSubClass	GarageType	CentralAir	Fireplaces	BsmtQual
10.7832892	10.4019978	10.3177695	10.0304863	9.5735393	9.1286471	8.8676665	8.3725898	8.2065351
TotRmsAbvGrd	MSZoning	GarageFinish	Exterior1st	YearRemodAdd	FullBath	LotFrontage	BldgType	GarageYrBlt
8.0089471	7.8781276	7.8101570	7.4816633	7.2649015	7.0205340	6.9427070	6.8795675	6.5395563
MasVnrArea	HeatingQC	WoodDeckSF	Exterior2nd	HouseStyle	BsmtUnfSF	BsmtFullBath	BedroomAbvGr	OpenPorchSF
6.0173472	6.0172993	5.9170377	5.7712270	5.5886391	5.5383703	5.3389301	5.2494763	4.9288481
HalfBath	KitchenAbvGr	Fence	BsmtExposure	MasVnrType	Foundation	OverallCond	BsmtFinSF2	Alley
4.8959761	4.7299341	4.3524701	4.2946219	4.1465747	3.7410808	3.3946161	3.1233260	2.9007849
PavedDrive	SaleCondition	BsmtCond	PoolQC	YrSold	LotShape	FireplaceQu	LandContour	LandSlope
2.4595304	2.3028326	2.2890933	2.2459364	2.0521340	2.0194147	1.9099956	1.7767116	1.7628335
EnclosedPorch	RoofStyle	SaleType	BsmtHalfBath	ExterCond	Condition1	ScreenPorch	MiscFeature	LowQualFinSF
1.7263592	1.5723224	1.4297111	1.3760844	1.2595922	1.2285783	1.0187297	0.8338167	0.6957466
BsmtFinType2	Utilities	Functional	GarageQual	Street	LotConfig	Condition2	X3SsnPorch	MoSold
0.1257992	0.0000000	-0.1690244	-0.6211626	-0.9068782	-0.9255500	-1.1056079	-1.4311031	-1.4514698
PoolArea	MiscVal	Heating	GarageCond	Electrical	RoofMatl			
-1.6686577	-1.7280614	-1.8571717	-1.9907972	-2.0195567	-2.0632108			

The number under the name of predictor indicates the amount of increase in training MSE when each predictor is excluded. Negative value means the training MSE decrease when a predictor is removed.

Interactions were verified with "lattice" package in R. This package outputs plots to inspect for interactions. No interactions were observed between the first 10 important predictors.

Conclusion

Several regression methods were tested for the best possible predictive model. Random Forest regression performed the best; however, it did not construct a reliable model.

Limitations & Future Work

Lack of Data

House price is determined not just data related to the house. It is also determined by other factors like school quality, crime rates, and location. All of these external data were not included in the data. The absence of these data could be the biggest reason why test RMSE barely improved despite many techniques used. Future work could be collecting more variables and selecting only a few important variables from the original dataset then combining them. During this process, PCA, ICA, and autoencoder can be used to verify the important variables or create more condensed variables.

Missing Value Imputation

KNN imputation was used with all of variables in the data. Using all variables cause the curse of dimensionality which basically states that data points are similar in direction wise as dimension increases.