Modeling Assignment3

Jordan Zhang

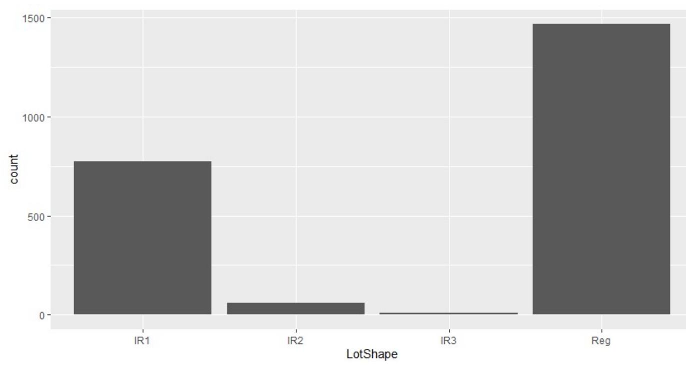
I picked three categorical variables as candidates of predictive variables: LotShape, Neighborhood and BldgType. These categorical variables have logical relationship to Saleprice, and the data for them are complete, without null entries.

There are Four levels of Lotshape: IR1, IR2, IR3 and Reg. And here is a table summarizing the Saleprice statistics for each level of Lotshape:

	Min	Median	Mean	Max
IR1	52000	185900	200318	470000
IR2	109000	207000	208716	402000
IR3	73000	201570	203928	375000
Reg	35000	141250	155797	468000

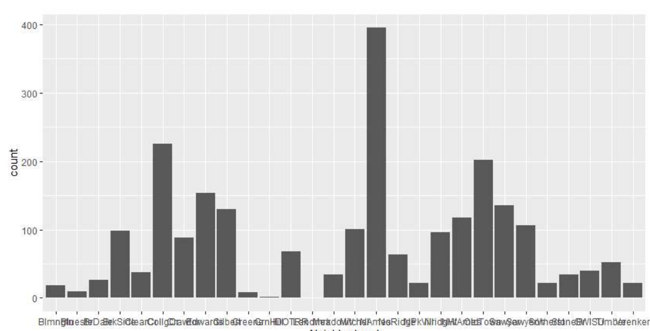
While there is no significant difference between IR1,2,3, the mean and median values of Reg is lower than that of IR groups.

Here is a bar plot displaying number of houses in each Lotshape:



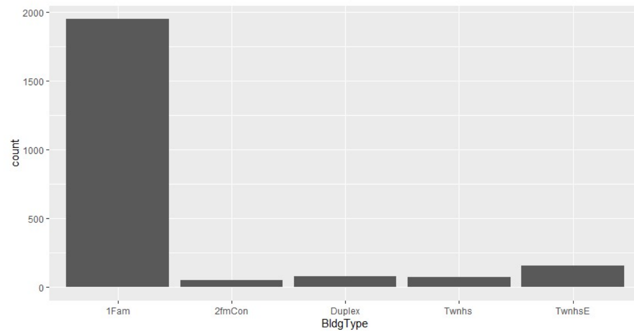
Most of the houses are in Reg and IR1 categories.

Neighborhood is logically related to the Saleprice. However, the bar plot of number of houses in each neighborhood showed a difficulty in using this variable for prediction:



The are about 30 different neighborhoods, with uneven distributions. This will make the model highly complicated, especially when interactions are taken into consideration.

BldgType also affects the Saleprice logically- the price of townhouse is usually different from single family houses. Here is the bar chart showing number of houses in each category:



Most houses are 1Fam houses. And here is a table summarizing the Saleprice statistics for each level of BldgType:

	Min.	Median	Mean	Max.
1Fam	35000	159925	175780	470000
2fmCon	55000	124500	126127	228950
Duplex	61500	136953	141191	269500
Twnhs	73000	119500	128834	230000
TwnhsE	75500	173000	181189	375000

The different Building Types did not show significant difference in Mean Sale Prices.

Comparing these three and other categorical variables in the dataset, I believe Lotshape is comparatively most predictive of Saleprice. Four dummy variables are created for IR1, IR2, IR3 and Reg. Lot_Reg will be used as basis case in the modeling process.

Here is the table of variables I included in the cleaned dataset for further variable selections:

'OverallCond'	'OverallQual'	'GrLivArea'	'FullBath'	'HalfBath'	'HouseAge'
'LotFrontage'	'LotArea'	'BsmtUnfSF'	'TotalSqftCalc'	'BedroomAbvGr'	'TotRmsAbvGrd'
'OpenPorchSF'	'LotIR1'	'LotIR2'	'LotIR3'	'GarageArea'	'WoodDeckSF'

Rows with NA entries for any of these variables are omitted and there are 1877 rows left. This will be final our population of study. The cleaned dataset was separated into training set and testing test, using the random number generator. The split is 70/30 train/test. And the actual training set contains 1313 observations, and the testing set contains 564 observations (30.05% of total data).

The models were first developed using SalePrice, but the residual plot indicates significant bias and heteroscedasticity. So, the response variable is now logSalePrice.

18 predictive variables, including three dummy variables for Lotshape. Among them, variable HouseAge is calculated as Yearsold- YearBuilt, and TotalSqftCalc is calculated as BsmtFinSF1, BsmtFinSF2 and 'GrLivArea'. The highlighted variables are discrete (or dummy variables) and the rest variables are continuous.

The StepAIC function was then used for variable selection. The upper model is the Full Model containing all 17 predictor variables in the variable pool, and the lower model as the Intercept Model. The model containing single variable 'TotalSqftCalc' is used to initialize the stepwise model selection.

Here is a summary table of variables and coefficients for each variable and the p-value for significance test:

Selected Model	Estimate	Pvalue
(Intercept)	10.50112337	2E-16
TotalSqftCalc	0.000219172	2E-16
OverallQual	0.087189812	2E-16
LotArea	9.2559E-06	2E-16
HouseAge	-0.003031364	2E-16
OverallCond	0.058095943	2E-16
BsmtUnfSF	0.000128113	2E-16
GarageArea	0.000178497	2E-16
GrLivArea	6.67539E-05	6.02E-05
BedroomAbvGr	-0.015705566	0.0012
LotFrontage	0.000484738	0.00319
LotIR1	0.018036051	0.01405
LotIR3	-0.134496255	0.01324
HalfBath	0.011694972	0.11645

Three different variable selection methods yield the same 13 variables. (Essentially 12 variables considering LotIR1 and LotIR3 are both dummy variables of LotShape variable.) Five variables were dropped in this model.

Here is the coefficient of the junk model containing five highly correlated variables:

Junk model	Estimate	Pr(>|t|)
(Intercept)	9.808980057	2E-16
OverallQual	0.268064246	2E-16
OverallCond	0.135592609	4.48E-15
QualityIndex	-0.022289973	4.73E-13
GrLivArea	0.000121489	2E-16
TotalSqftCalc	0.000196204	2E-16

The VIF is calculated to explore the correlation between variables for both models:

For the selected model (3 methods yield same model):

Selected Model	VIF
GrLivArea	5.603232
TotalSqftCalc	5.224658
OverallQual	2.711916
HouseAge	2.096582
BsmtUnfSF	2.058228
GarageArea	1.712287
BedroomAbvGr	1.702906
LotArea	1.695351
LotFrontage	1.599639
OverallCond	1.32617
LotIR1	1.191151
OpenPorchSF	1.165626
LotIR3	1.027455

The largest VIF is 5.6, well below the 10 threshold. No variable needs to be removed in this model due to collinearity.

In comparison, here is the VIF table for the junk model:

Junk Model	VIF
QualityIndex	35.6318
OverallQual	21.94843
OverallCond	17.54195
GrLivArea	2.623704
TotalSqftCalc	2.595796

The three highly related variables: QualityIndex, OverallQual, OverallCond have very high VIF. This is because the variable Quality Index is essentially the product of the other two variables. This variable shows strong collinearity with the other two.

The Adjusted R square, AIC, BIC, mean squared error, and the mean absolute error for both models were calculated using the training data:

	Selected Model	Junk Model
Adjusted R2	0.9159	0.8233
AIC	-2131	-1249
BIC	-2053	-1213
Mean Sq Err	0.01128665	0.02245331
Mean Abs Err	0.08116837	0.1141099

The selected model has better predictive accuracy in all five metrics, comparing to the junk model: Higher Adjusted R2, lower AIC, BIC, MSE and MAE.

But generally speaking, when we compare multiple models, different metrics can have different rankings. And an analyst will need to make decision on which metrics to use for choosing a model.

Predictive Accuracy

Here is the table comparing both models' predictive accuracy on the test data set. The response variable is ln(Saleprice).

	Selected Model	Junk Model
Mean Sq Err	0.01264608	0.02249148
Mean Abs Err	0.08365691	0.1145258

The selected model is still more accurate than the junk model.

Both MSE and MAE are good metrics for predictive accuracy. In this case when many of the error terms are less than 1, the mean absolute error term will be larger than the mean square error term and thus relatively more obviously showing difference between models. I would prefer to use MAE here.

The selected model has slightly lower MAE for in-sample, comparing to the test set. The difference is acceptable. In general, if a model has much better predictive accuracy in-sample then it does out-of-sample, the model is likely over-fitting the training dataset.

Operational Validation

The predicted value is considered to be 'Grade 1' if it is within ten percent of the actual value, 'Grade 2' if it is within ten to fifteen percent of the actual value, Grade 3 if it is within fifteen to twenty-five percent of the actual value, and 'Grade 4' otherwise. Note the natural log transformed model will shrink the error rate, so the accuracy rate is calculated using the transformed back SalePrice values. Predicted Saleprice= exp(predicted log-SalePrice).

Here is a summary of prediction grades for the selected model and junk model, for the training set:

Train	Grade1	Grade2	Grade3	Grade4
Selected	69.69%	16.07%	10.89%	3.35%
Junk	22.99%	9.94%	16.16%	50.91%

Almost 70% of predictions are within ten percent of the actual value for the selected model, and the selected model outperforms the junk model significantly.

And the same analysis was done using the out of sample test dataset:

Test	Grade1	Grade2	Grade3	Grade4
Selected M	67.73%	16.84%	10.46%	4.96%
Junk Model	20.92%	9.75%	15.96%	53.37%

The prediction accuracy is close to that of the training set, with slight decrease in Grade 1 and increase in Grade 4. The selected model would qualify as having underwriting quality considering most of the predictions have Grade1 accuracy.

Final Model Selection

Here is the anova table of the AIC generated model:

	Df	Sum Sq
OverallQual	1	121.242
TotalSqftCalc	1	25.667
LotArea	1	4.352
HouseAge	1	3.305
OverallCond	1	2.752
BsmtUnfSF	1	4.061
GarageArea	1	1.169
GrLivArea	1	0.196
BedroomAbvGr	1	0.125
LotFrontage	1	0.104
LotIR1	1	0.08
LotIR3	1	0.071
HalfBath	1	0.028
Residuals	1299	14.819

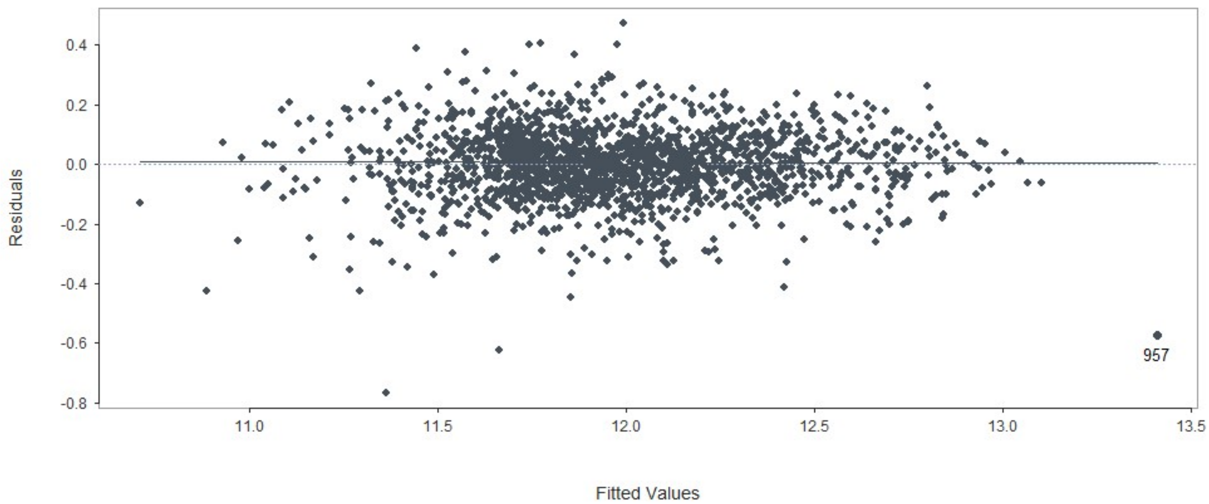
I decided to remove the three variables with least SumSq – contributing least to explaining the variance of the logSaleprice. I will also remove variable 'BedroomAbvGr' in the final model because it has negative coefficient that could not be intuitively explained – a likely indication of multicollinearity.

The final model contains 9 variables:

Variable	coefficient
(Intercept)	10.4853532
OverallQual	0.08841107
TotalSqftCalc	0.00021415
LotArea	4.1403E-06
HouseAge	-0.0033327
OverallCond	0.05874972
BsmtUnfSF	0.00011708
GarageArea	0.00018276
GrLivArea	7.4978E-05
LotFrontage	0.00089876

The R squared value is 0.9067. None of these variables is categorical/ dummy-coded. The coefficients are small in value because the response variable was log transformed. All variables have a logical correlation to the saleprice. For example, only HouseAge has a negative coefficient – the older a house is the lower the saleprice tend to be. Other variables are all positively correlated to the saleprice.

The residual vs predicted value plot is here:



The distribution of residuals is mostly centered around 0 and show no particular pattern. The final model containing 10 variables shows satisfying goodness of fit.

The past seven weeks of study has helped me tremendously in understanding the EDA, variable selection, modeling diagnostics and most importantly interpreting modeling results. The major challenge for me during first two weeks was the overwhelming number of variables. The combination of manual and automatic variable selection made it possible to find a most appropriate model for our purpose.

I believe that for data that has a business context, for example house sale price, it is better to have a simpler, explainable model that is accurate enough, rather than a max fit but complicated model. After all, the model is supposed to be helpful for making business decisions and thus needs to be interpretable.