

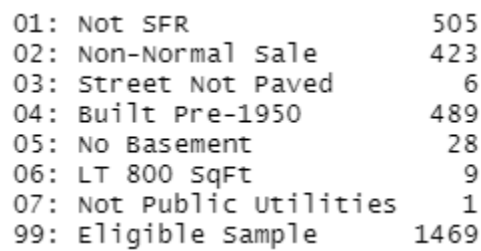
MSDS 410 Assignment 7

Sameer Khan

Section 1: Sample Definition and Data Split

1.1: Sample Definition

For this assignment, the goal is to build regression models for the home sale price. To generate the most accurate predictions we will need to create a subset of our data which accurately represents a sample of single-family homes. The Ames dataset is very large and contains observations which are out of the scope of our model, so some data points must be dropped based on a set of conditions. An appropriate sample population for this problem would be a set of single-family homes sold at a normal price. Other conditions can be created to decrease the chances of outlier homes being present in the data in terms of features and price. Homes that have unpaved streets, were built before 1950, do not have a basement, are small (<800 sqft), and don't use public utilities could be considered 'unusual' homes that might confuse our model. A waterfall for these drop conditions has been created which shows the number of observations that are removed from the Ames data frame during each condition check (Figure 1).



01: Not SFR	505
02: Non-Normal Sale	423
03: Street Not Paved	6
04: Built Pre-1950	489
05: No Basement	28
06: LT 800 SqFt	9
07: Not Public Utilities	1
99: Eligible Sample	1469

Figure 1: Number of data points removed for each waterfall condition

1.2: The Train/Test Split

Our data is split into a 70/30 train/test split. This means that 70% of the dataset will be considered 'in-sample' and used to train the models. The remaining 30% of the data will be considered as 'out-sample' data and used to validate/grade our models. The counts for each of the training/test datasets are shown below in Figure 2.

Data Split	Observation Count
Training Data	1037
Testing Data	432

Figure 2: Observation Counts for Train/Test sets

Section 2: Model Identification and In-Sample Model Fit

2.1: Forward Variable Selection

The forward variable selection model was found to have the formula:

SalePrice ~ TotalSqftCalc + KitchenQual + BsmtUnfSF + GarageYrBlt + QualityIndex + LotArea + BsmtQual + MasVnrArea + LandContour + TotalBsmtSF + BedroomAbvGr + GarageArea + BsmtExposure + ScreenPorch + Fireplaces + WoodDeckSF + OpenPorchSF

One variable, 'ExterQual' was removed due to having a VIF value larger than 20.

2.2: Backward Variable Selection

The backward variable selection model was found to have the formula:

SalePrice ~ LotArea + LandContour + MasVnrArea + BsmtQual + BsmtExposure + BsmtUnfSF + TotalBsmtSF + FirstFlrSF + SecondFlrSF + LowQualFinSF + BedroomAbvGr + KitchenQual + Fireplaces + FireplaceQu + GarageYrBlt + GarageArea + GarageQual + GarageCond + WoodDeckSF + ScreenPorch + QualityIndex

All predictor variables were kept because they had VIF values less than 20.

2.3: Stepwise Variable Selection

The stepwise variable selection model was found to have the formula:

SalePrice ~ TotalSqftCalc + KitchenQual + BsmtUnfSF + GarageYrBlt + QualityIndex + LotArea + BsmtQual + MasVnrArea + LandContour + TotalBsmtSF + BedroomAbvGr + GarageArea + BsmtExposure + ScreenPorch + Fireplaces + WoodDeckSF + OpenPorchSF

All predictor variables were kept because they had VIF values that were less than 10. We notice that this model is very similar to our forward variable selection model.

2.4: Model Comparison

The three models we created have metrics displayed below in Figure 3. They all have a similar adjusted R-square value of about 0.93, indicating that the chosen variables account for 93% of the variation in SalePrice. The AIC and BIC values are also very close for each model, especially backward and stepwise selection. It is impossible to say that one model is significantly better than another; backward selection model has a lower MSE and MAE than the other two models, but it also has higher AIC and BIC values.

Metric	Forward Selection	Backward Selection	Stepwise Selection	Junk Model
Adjusted R-Squared	0.9234	0.9247	0.9234	0.8431
AIC	9993.438	9994.645	9993.438	24259.25
BIC	10107.42	10145.26	10107.42	24293.86
MSE	542840498	522190659	542840498	14482682665
MAE	17069.17	16770.59	17069.17	89889.1

Figure 3: Metrics

Section 3: Predictive Accuracy

Based on the MSE and MAE criteria for the training data, we would expect the backward selection model to perform the best for prediction on the test data. For the test data the MSE and MAE data is shown below:

Metric	Forward Selection	Backward Selection	Stepwise Selection
MSE	1336144148	10350658360	1336144148
MAE	21954.04	71853.35	21954.04

Figure 4: MSE and MAE values

Based on these results for the test data, the forward selection and stepwise selection perform better than the backward selection model in terms of both MSE and MAE. This is different than our in-sample results. I personally prefer the MSE to MAE because it has a heavier penalty for larger errors between actual values and predicted values. Since the goal of our model building is to get as close to an accurate estimate as possible, MSE is the better metric to evaluate by. A model is overfitted when it has better predictive accuracy in-sample than it does out of sample.

Section 4: Operational Validation

Prediction grades for each model are shown below in Figure 5:

Model	Grade 1	Grade 2	Grade 3	Grade 4
Forward Selection	0.5949	0.1342	0.1689	0.1018
Backward Selection	0.3981	0.1111	0.0787	0.4120
Stepwise Selection	0.5949	0.1342	0.1689	0.1018
Junk Model	0.5902	0.2106	0.1459	0.053

Figure 5: Prediction Grades

These results confirm our test MSE/MAE values and verify that forward selection and stepwise selection models are more accurate than backward selection. Both of models fit underwriting quality with a Grade 1 (10%) accuracy of almost 60%. It is concerning for backward selection that 80% of its predictions are either very accurate (within 10%) or very inaccurate (>25%) compared to their actual values.