**Analysis Question 2**

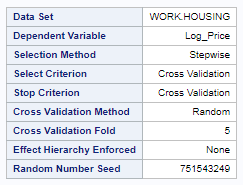
*Problem Statement:* What is the best model for predicting the sale prices of homes in Ames, Iowa based on the variables provided? Specifically, what models are generated by using Forward, Backward, and Stepwise selection methods on the provided training data set and what model produces the highest Kaggle score?

*Selection Models \*Results and parameters to be moved to appendix*

**Stepwise**

**See Appendix [] for sesults and parameter estimates.**

Selection options:

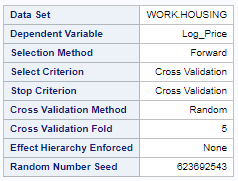


Results: *\*Results and parameter estimates to be moved to appendix – located in html files*

Model parameter estimates:

**Forward**

Selection options:

****

Results: *\*Results and parameter estimates to be moved to appendix – located in html files*

Model parameter estimates:

**Backward**

Selection options:

Results: *\*Results and parameter estimates to be moved to appendix – located in html files*

Model parameter estimates:

**Custom**

Selection options:

Results: *\*Results and parameter estimates to be moved to appendix if we don’t choose one of the above 3 models*

Model parameter estimates:

*Model Comparisons*

*\*Note: CV options were used with each selection method with PROC GLMSELECT, so the models are not exactly reproducible, even with the same code and dataset.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward | .92 | 21 | .139 |
| Backward | .91 | 21 | .134 |
| Stepwise | .92 | 20 | .269 |
| CUSTOM | .00 | 00 | .000 |

*Conclusions*

* The model that provided the best predicted sale prices (by kaggle score) for homes was based on the backward selection method using SAS options that select variables for the model based on cross validation. We decided that the best approach to forming our three initial predictive models was to examine the variables in detail and cull them down before running the models to minimize over-fitting as much as possible and produce the most meaningful models. This approach would give us the best we could achieve with each selection method so we would be left to choose from one of the three as our final custom model.
  + **\*We have the option to pick one of our 3 models or include a different one; if someone comes up with a better model before we’re done we can change this part of the conclusion**
* Easy to over-fit the training set with such a large number of variables, categorical variables with large number of levels are most influential in over-fitting issues (shown by high cv press). Taking the log of the Sale Price greatly reduces cv press.
* (\*this might be discussed in assumptions) high correlations between many of the variables; sale price mostly determined by a few characteristics - homes with certain key features will share many characteristics that result I similar trends that affect sale price.
* It is easy to see where adjusted R2 falls short versus trying to minimize multicollinearity, as running models in SAS with many variables (and all of the interactions between them) can achieve adjusted R2 on the training set of very near 100%, but the predicted values on the test set are less accurate than simpler models that attempt to minimize overfitting.