Assignment #5: Automated Variable Selection, Multicollinearity, and Predictive Modeling (100 points)

Data: The data for this assignment is the Ames, Iowa housing data set. This data will be made available by your instructor.

Assignment Instructions:

For this assignment we will set up a predictive modeling framework, explore the use of automated variable selection techniques for model identification, assess multicollinearity, assess the predictive accuracy of our model using cross-validation, and compare and contrast the difference between a statistical model validation and an application (or business) model validation.

(1) Define the Sample Population

- Define the appropriate sample population for your statistical problem. Hint: We are building regression models for the response variable SalePrice. Are all properties the same? Would we want to include an apartment building in the same sample as a single family residence? Would we want to include a warehouse or a shopping center in the same sample as a single family residence? Would we want to include condominiums in the same sample as a single family residence?

- Define your sample using 'drop conditions'. Create a waterfall for the drop conditions and include it in your report so that it is clear to any reader what you are excluding from the data set when defining your sample population.

A description of your sample data should be included as its own section in your assignment report.

(2) Create a train/test split of the data for a basic cross-validation. We will use a standard SAS trick to keep 'two' data sets as one data set. We will do this by defining a new response variable train_response.

- We will split the sample into a 70/30 training/test split. We will 'train' each model by estimating the models on the 70% of the data identified as the training data set, and we will 'test' each model by examining the predictive accuracy on the 30% of the data.

```
data temp;
set clean_data_subset;
* generate a uniform(0,1) random variable with seed set to 123;
u = uniform(123);
if (u < 0.70) then train = 1;
else train = 0;

if (train=1) then train_response=SalePrice;
else train_response=.;
run;
```

We will use the response variable train_response when fitting our models. Show a table of observation counts for your train/test data partition somewhere in your assignment report.

(3) Model Identification by Automated Variable Selection and Predictive Accuracy

- Model Identification: Using the training response find the 'best' models using automated variable selection using the techniques: adjusted R-Squared, MaxR, Mallow's Cp, forward, backwards, and stepwise variable selection. Identify (list) each of these six models individually. Refer to them as Model_AdjR2, Model_MaxR, Model_MCp, Model_F, Model_B, and Model_S. Did the different variable selection procedures select the same model or different models? Report the final models and the summary tables from each variable selection technique. The summary tables and their discussion should be its own section in your assignment report. Do not include the goodness-of-fit plots anywhere in this report. Graphical goodness-of-fit is not part of this assignment.

- For each of these six models compute the adjusted R-Squared, AIC, BIC, mean squared error, and the mean absolute error for each of these models for the training sample, and the mean squared error and the mean absolute error for the test sample. Which model fits the best based on these criteria? Did the model that fit best in-sample predict the best out-of-sample? Discuss these quantitative measures of goodness-of-fit. This discussion should be its own section in your assignment report.

Note that SAS can be used to compute some of these measures using a separate PROC REG statement from the PROC REG statement used for variable selection. Other metrics will need to be computed using a SAS data step and a PROC MEANS statement.

- For each of these six models assess multicollinearity by computing the Variance Inflation Factors. Do any of these models have multicollinearity concerns? The discussion of multicollinearity should be its own section in your assignment report.

(4) Operational Validation

- We have validated these models in the statistical sense, but what about the business sense? Do MSE or MAE easily translate to the development of a business policy? Define the variable 'Prediction_Grade' (define the variable using format $7.). Let's consider the predicted value to be 'Grade 1' if it is within ten percent of the actual value, 'Grade 2' if it is within fifteen percent of the actual value, and 'Grade 3' otherwise. How accurate are the models under this definition of predictive accuracy? Use PROC FREQ to provide a table of the model's operational accuracy.

SAS Tip: There are two ways to do this: (1) use logical statements (if – else if – else), or (2) use a user defined SAS format created using PROC FORMAT. If you use logical statements, then before you write those statements (at the top of your ladder) put ***format Prediction_Grade $7.;***. This will make sure that SAS creates a character variable of length 7 so that you can fill it with the values 'Grade 1', 'Grade 2', and 'Grade 3'.

Assignment Document:

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information. The document should be submitted in pdf format. Name your file Assignment5_LastName.pdf.