ISYE 7406 - Data Mining and Statistical Learning

Final Exam

Eric Wissner

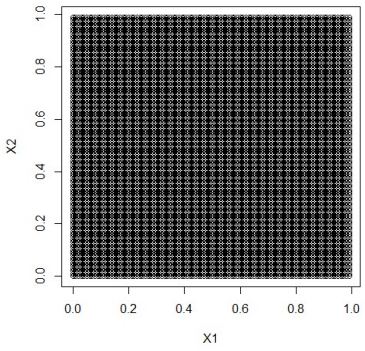
Introduction

For the final exam, students were asked to develop two analytic models based on a training data set of 10,000 observations. These 10,000 observations contained two predictor variables (X1 and X2) and 200 values generated from an unknown function of those variables. The Mean and Variance of those generated values for each X1/X2 combination ultimately represented the response variables that were used to train the models.

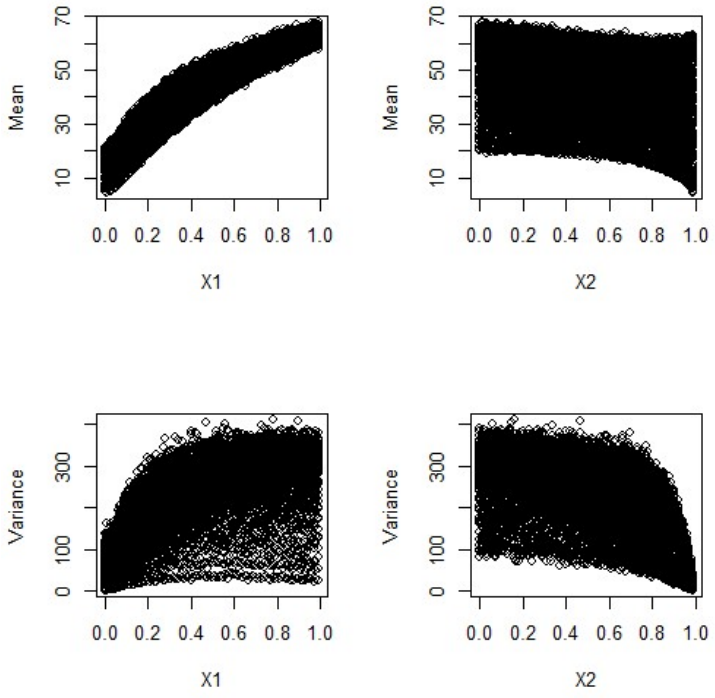
Students were then to use these models to predict the Mean and Variance for 2,500 X1 and X2 values from an unknown testing data set. Ranges of target mean squared error values for each set of predictions were provided.

Exploratory Data Analysis

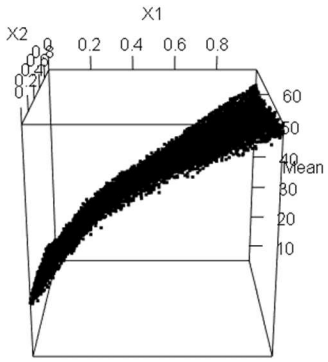
The directions for the exam stipulated that the training data set included each combination of X1 and X2, each uniformly distributed between 0 and 1 inclusively with intervals of 0.01. To that end, there was no need to test for multicollinearity or other relationships between the predictor variables. The image to the right illustrates the even coverage across these data points.

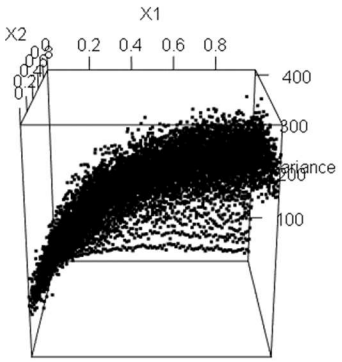


Relationships between the predictor variables and each of the response variables can, however, be explored. The following image shows those relationships isolated by each response and predictor variable. The curves appear to suggest some degree of nonlinearity in the relationships, and perhaps are polynomial in nature.

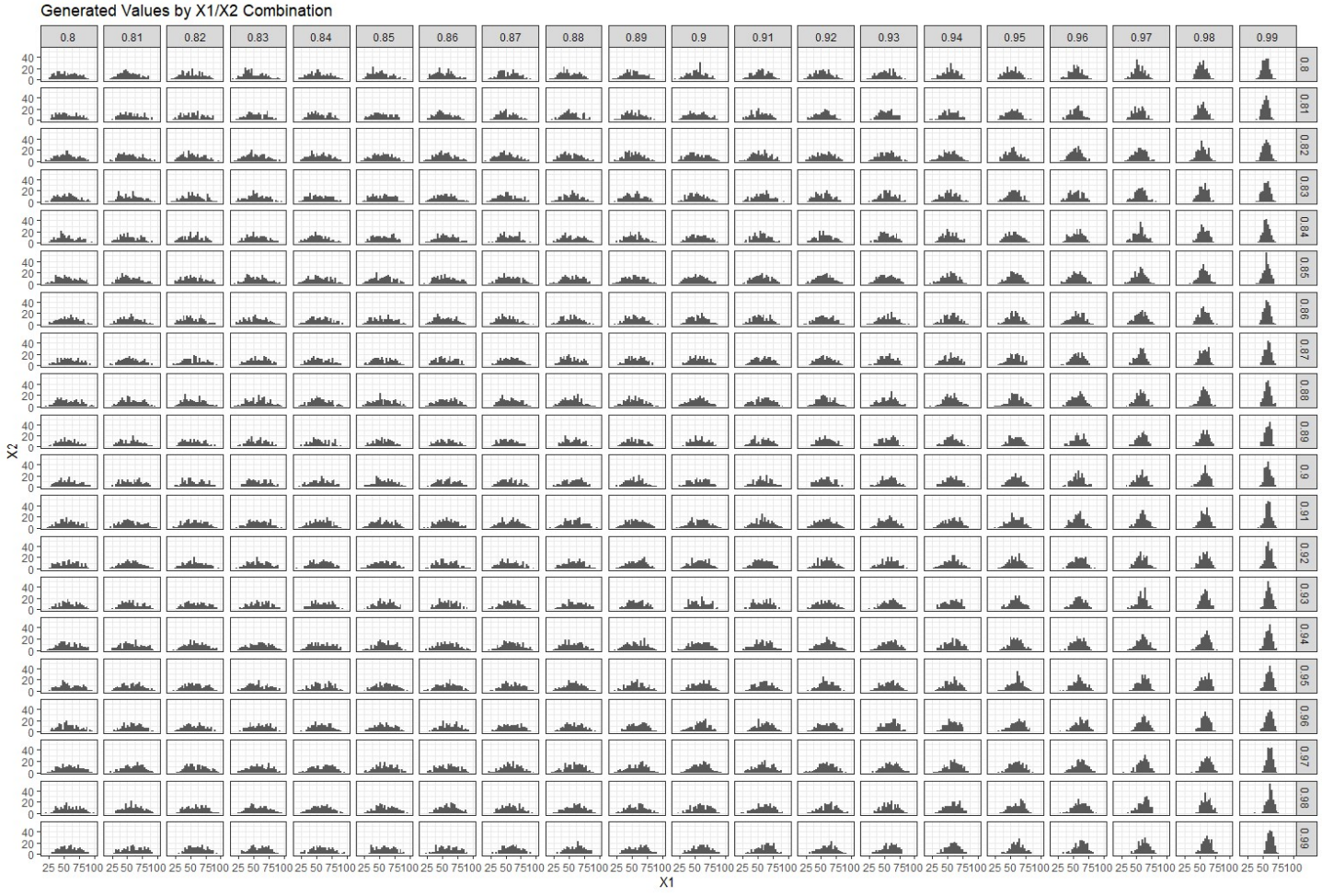


When considering both the X1 and X2 variables together, three-dimensional renderings available from the 'plot3d' function in the 'rgl' library can be helpful in getting a better sense of the shape of the data. These tools are interactive, allowing the analyst to rotate the view to fully see the 3D space. Static versions of these renderings are included below to illustrate the benefit (Mean chart is on the left; Variance is on the right).





Additional insights about the distribution of the 200 data points associated with each X1/X2 pair were possible using trellised histograms. Viewing the entire 100x100 matrix was not meaningful, so the chart was broken up into twenty-five 20x20 X1/X2 sections. An example is provided below. Note that, although details from any one specific cell are not clear, the overall visualization represents the changing shape of the distributions (which, in turn, affects both the resulting Mean and Variance response variables).



Methodology

Before building any models, a few preparatory steps were taken with the data. With the potential for an interaction to exist between the predictor variables, a new variable (X1X2) was created by multiplying X1 and X2. Additionally, the training data set was randomly divided into ten subsets for subsequent use in Monte Carlo cross validation (MCCV).

Using these MCCV loops, the following models were trained, tuned, and assessed:

Linear Regression. This was unlikely to be the final model based on the shape of the data as noted above. It did, however, provide a "worst case scenario" and an MSE for the other models to beat. Two versions were created: one with the original two predictor variables and one that included the X1X2 interaction variable.

K Nearest Neighbors (KNN). In this method, predictions are made based on the data points (quantity of "k") nearest to the data points for the record in question. Those neighbors effectively vote, and the average of those data points becomes the prediction for that record. The parameter of k was tuned by iterating over options between 3 and 151, inclusive, within each MCCV loop.

Generalized Additive Models with Local Smoothing. These models are advantageous because they do not assume a linear relationship between the predictor and response variables. Multiple versions of this model were evaluated: with smoothing on just the original variables, with smoothing also on the new X1X2 interaction variable, and with tensor product smoothing on the original X1 and X2 variables. Additionally, different values of "k" were assessed to best capture a suitable number of degrees of freedom while balancing computing performance.

Neural Network. A limited number of neural network models were trained and evaluated. Having already identified a model that returned an MSE within the target range, however, additional tuning and exploration was not conducted to increase their performance. The computation time for each variant, considering the MCCV loops, was deemed unnecessary.

Results

Models were created first for the estimated Mean. Each of the models were built and tuned as noted in the table below. The average mean square error for each model over the ten Monte Carlo cross-validation loops is also listed.

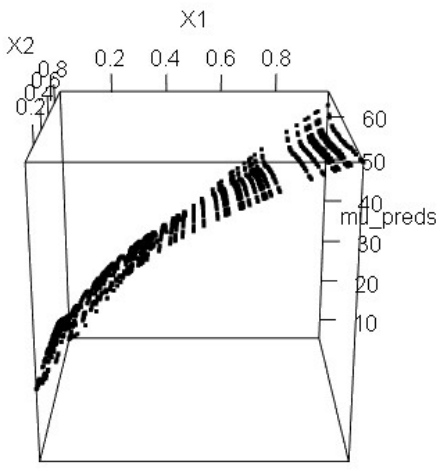
Model	Notes	MCCV MSE
Linear Regression	Original variables	9.203908
Linear Regression	Including X1X2 interaction variable	8.401605
K Nearest Neighbors	Best "k" from 3 to 151 selected and used within each MCCV loop	1.242015
General Additive Model	Local smoothers for original 2 variables (k=5)	3.202775
General Additive Model	Local smoothers for original variables (k=10); tensor product smoother for original variables (k=15)	1.199151
General Additive Model	Local smoothers for original variables (k=10) and X1X2 interaction variable (k=20); tensor product smoother for original variables (k=25)	1.194896
General Additive Model	Local smoothers for original variables and X1X2 interaction variable; tensor product smoother for original variables (k=15)	1.191786
General Additive Model	Local smoothers for original variables and X1X2 interaction variable; tensor product smoother for original variables (k=20)	1.186734
General Additive Model	Local smoothers for original variables and X1X2 interaction variable; tensor product smoother for original variables (k=25)	1.186816
Neural Network	SoftPlus activiation function (like ReLu); threshold=0.04; hidden=5	1.471997
Neural Network	SoftPlus activiation function (like ReLu); threshold=0.04; hidden=(4,2)	1.391394
Neural Network	SoftPlus activiation function (like ReLu); threshold=0.04; hidden=(4,4)	1.326498
Neural Network	SoftPlus activiation function (like ReLu); threshold=0.04; hidden=(3,3,3)	1.313974

The General Additive Model using local smoothers and tensor product smoothing (with k=20) was the best model (MSE = 1.186734) and was used to predict the estimated Mean values for the test data provided.

The final model:

```
m3c_ss <- mgcv::gam(muhat ~ s(X1, k=20) + s(X2, k=20) + s(X1X2, k=20) + te(X1,X2, k=20), data=data0)
```

The three-dimensional shape of the predicted values appears to resemble that of the training data provided.



Models were then trained for the Variance response variable. Similar tactics were employed to tune parameters, including tweaking the k-values in the General Additive Model. The table below shows the models, associated notes, and the MCCV MSE.

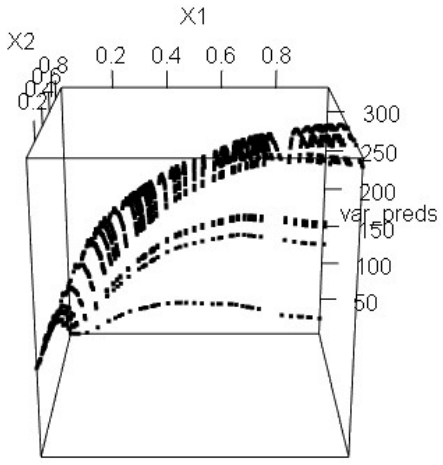
Model	Notes	MCCV MSE
Linear Regression	Original variables	2172.0956
Linear Regression	Including X1X2 interaction variable	2172.5596
K Nearest Neighbors	Best "k" from 3 to 151 selected and used within each MCCV loop	561.2063
General Additive Model	Local smoothers for original 2 variables (k=5)	756.2603
General Additive Model	Local smoothers for original variables; tensor product smoother for original variables (k=10)	533.6394
General Additive Model	Local smoothers for original variables and X1X2 interaction variable; tensor product smoother for original variables (k=10)	531.0233
General Additive Model	Local smoothers for original variables (k=10) and X1X2 interaction variable (k=20); tensor product smoother for original variables (k=25)	530.4121

Once again, the General Additive Model that utilized both the interaction variable and tensor product smoothing (with k=10, 20, and 25, respectively) was the optimal model with an MSE of 530.4121.

The final model:

```
v_ss <- mgcv::gam(Vhat ~ s(X1, k=10) + s(X2, k=10) + s(X1X2, k=20) + te(X1,X2, k=25), data=data0)
```

As with the estimated Mean values, the estimated Variance values can be plotted in a 3D space for comparison against the Variance values from the original training data set.



Conclusion

With Cross-Validation MSE scores within the top end of the targeted range of scores, my expectation is that these models will have performed well on the 2,500 records of unseen test data. Per the Final Exam directions, those predictions have been saved to a file named “1.Wissner.Eric.csv”.