Take-Home Final Exam for ISyE 7406

Arthur R. Ward

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

ISyE 7406: Data Mining & Statistical Learning

Dr. Xiaoming Huo

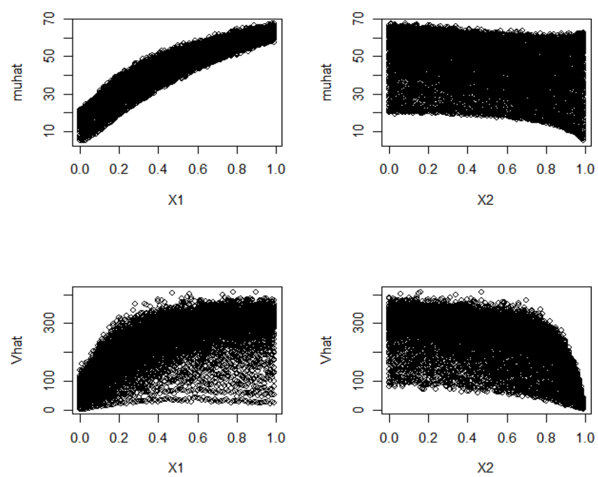
December 8th, 2024

Introduction

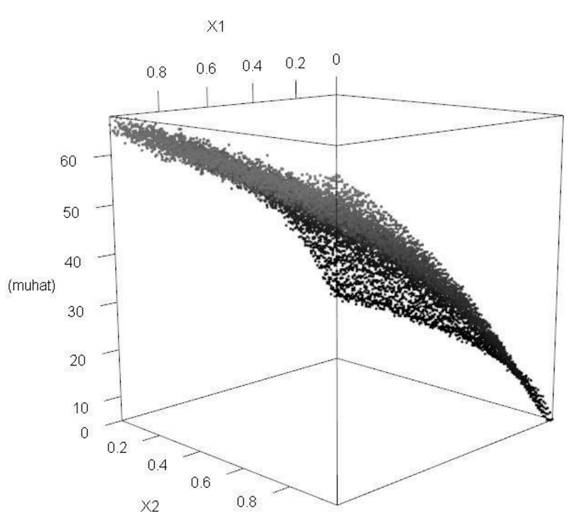
The objective of this analysis is to develop and compare predictive models for estimating both the mean and variance of the random variable Y= Y(X1,X2), given the independent predictors X1 and X2. These estimates will be evaluated using a separate testing dataset to assess the performance of the models in terms of their MSE.

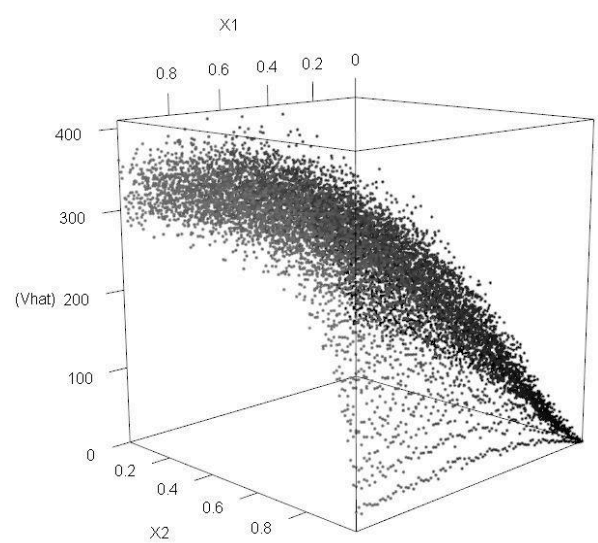
Data Exploration

The data was generated as follows: Uniform design points when $0 \le X1 \le 1$ and $0 \le X2 \le 1$ for $X1_i = .01 * i$ for $i = 1, 2, ..., 99$, and $X2_j = .01 * j$ for $j = 1, 2, ..., 99$, for a total of 100 * 100 = 10,000 combinations of $(X1_i, X2_j)$ and for each of the combinations 200 independent realizations of Y were generated, Therefore, the training data is composed of 10,000 combinations of X1 and X2 with 200 corresponding independent realizations of Y for that pair, resulting in a 10000 * 202 dataset including X1 and X2. The testing dataset was generated by taking 50 random design points from X1 and X2. Thus, there are 50 * 50 = 2500 combinations to evaluate and compare models.



A series of 2 dimensional plots show the relationship of variables X1, X2, to muhat, and vhat, see above. The relationships are generally not linear to weakly linear with exception of X1 and muhat, which shows a moderate positive linear relationship in that dimension.The plots below show 3-dimensional relationships X1 and X2 vs muhat and vhat. The moderate linear relationship of X1 and muhat results in a somewhat defined hyperplane when expanded into X2, suggesting that a linear model for muhat may have good results. The relationship of X1, X2, and vhat show a much more dispersed and somewhat concave relationship. While there is a vaguely defined hyperplane for vhat, random forest, and XGBoost, or splines will likely provide better estimates because of their more robust tolerances for nonlinear relationships.





Model Building and Testing

Linear Regression

A linear regression model was fitted primarily to serve as a baseline reference for more advanced models. Preliminary exploratory data analysis showed that linear regression might be somewhat suitable for explaining variation of muhat with regard to X1. Minor data cleaning in preparation for regression was used to help ensure optimal results. The data was standardized and principal component analysis was used.

Random Forest

Random forest models for muhat and Vhat were trained using tidydodels package. The original training dataset was separated into a derivative training and test set using a 75/25 split. Model hyperparameters mtry, min_n, and trees were first tuned using bootstrap resampling to help minimize the risk of overfitting.

XGBoost

Gradient boosting with XGBoost was tested with both cross validation and bootstrapping to tune a complex set of 6 hyperparameters. Tuning the number of trees was deemed unnecessary after preliminary runs showed marginal changes in RMSE with tree counts above 1000. The remaining parameters: tree_depth, min_n, loss_reduction, sample_size, mtry, and learn_rate were optimized using grid_space_filling to better cover the wide array of parameters possibilities and reduce processing time.

Generalized additive Model

Generalized additive models (GAM)s were built using polynomial regression engines. These models were trained with stepwise sampling and tested various degrees of freedom from 1-15 along with interaction terms for X1 and X2. Interestingly, GAM models without interaction terms performed worse than the tree based models but outperformed the baseline linear regression model.

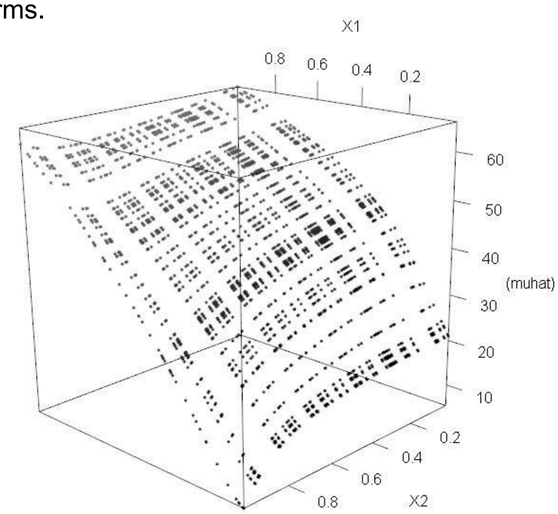
Test Results

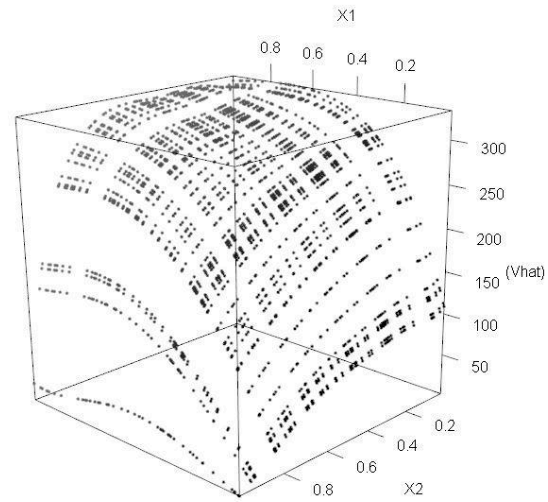
Each model was tested with 2500 data points split from the original data0 training set. The table below shows the final models with their associated MSE metrics plus R2.

Model	**Mu MSE**	R2	**Var MSE**	R2
Linear Regression	9.19028	0.9508	2160.109	0.7219
Random Forest	1.35398	0.9928	557.7987	0.9425
XGBoost	1.26903	0.9933	542.6240	0.9325
GAM w/ X1X2	1.22887	0.9935	531.8752	0.9338

Conclusion

As expected, the linear model performed somewhat well with regard to X1 and managed to explain a reasonable amount of variation in muhat. However, the linear model was unable to fit the much less linear X2 dimension and provided poor estimates for Vhat, earning the highest MSE score of any model run. Non-linear models performed much better overall with ensemble models Random Forest and XGBoost providing satisfactory results. The final Generalized Additive Models performed very well and produced the lowest MSE for both prediction cases, owing to the inclusion of the additional interaction terms.





The final predictions are plotted above. Both models seem highly reminiscent of the distributions shown in earlier data exploration. The visual interpretation along with associated MSE scores suggest that the models have adequately predicted the mean and variance of the random variable Y= Y(X1,X2).

Appendix

Software Used

Tuning for XGBoot, Random Forest, and GAMs was done with **tidymodels**, **usemodels**, and utilized **doParallel**. All work was carried out in **RStudio**.

Tidymodels: tidymodels
Usemodels: usemodels 0.0.1 - Tidyverse
doParallel: R doParallel: A Brain-Friendly Introduction to Parallelism in R | R-bloggers

Acknowledgment: Blog | Julia Silge was a very helpful resource for learning to use the above libraries.