Final Exam

Introduction

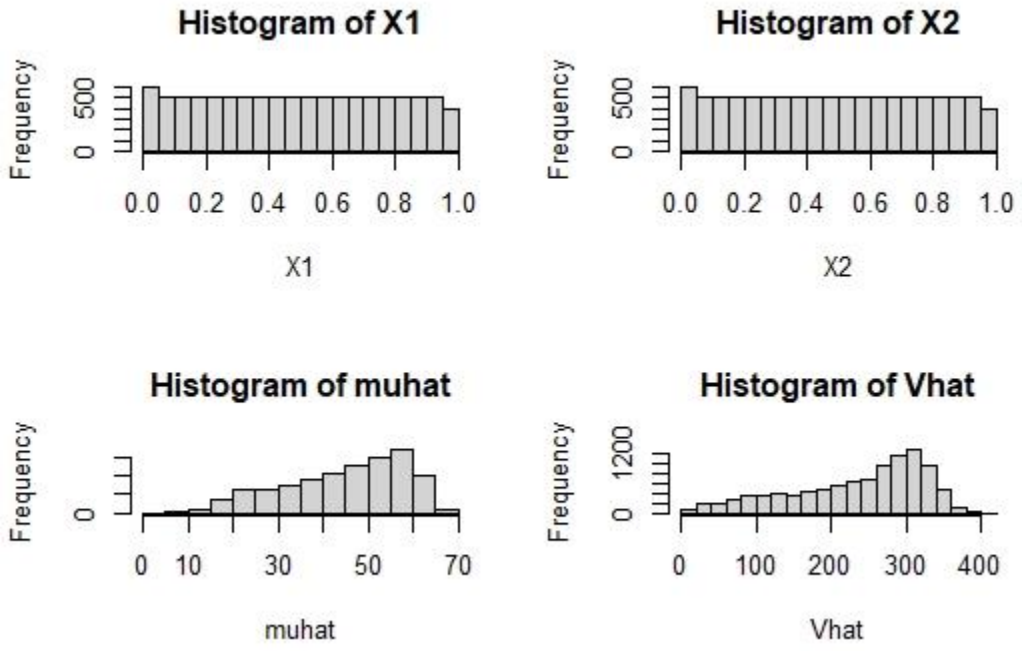
In probability and statistics, it is important to understand the mean and variance for any random variables. In this project, Y = Y (X1, X2) is a random variable whose distribution depends on two independent variables X1 and X2. The objective here is to use data mining/ machine learning methods to provide a useful surrogate model that allows us to conveniently predict or approximate the mean and variance of response variable Y = Y (X1, X2) as a function of X1 and X2. We are provided with 200 observed realizations of the Y values for some given pairs (X1, X2)'s that can be used to approximate the distributions of Y 's. Our task is to estimate or approximate the values of μ(X1, X2) = E(Y (X1, X2)) and V (X1, X2) = Var(Y (X1, X2)) for the given pairs (X1, X2) in the testing data set.

Note: This was an exam question in a Data mining course I took at GA Tech.

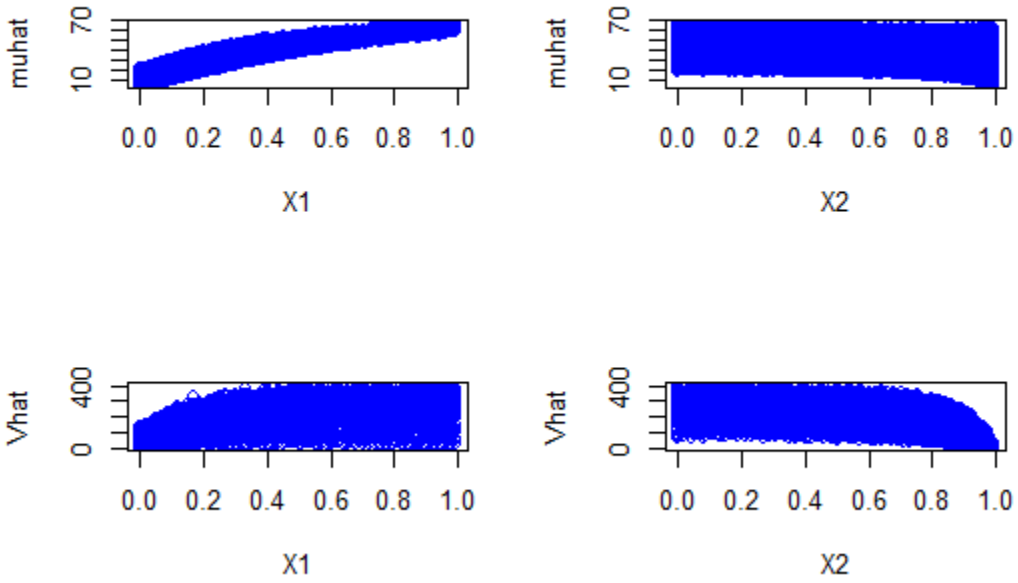
Exploratory data analysis

Training dataset contains 10000 values for X1,X2 between 0 and 1 uniformly distributed and 200 realizations of Y's for each datapoint. The first transformation performed on the input dataset was to replace the 200 realizations of Y's with a mean and variance for each row. These two variables will be our response variables.

As mentioned above, the independent variables in the training dataset are uniformly distributed from 0 to 1. The response variables have a skewness to the right. muhat is in the range of 0 to 70 and Vhat is in the range of 0 to 400.



We will now plot the dependent variables against the independent variables to understand the pattern of relationships

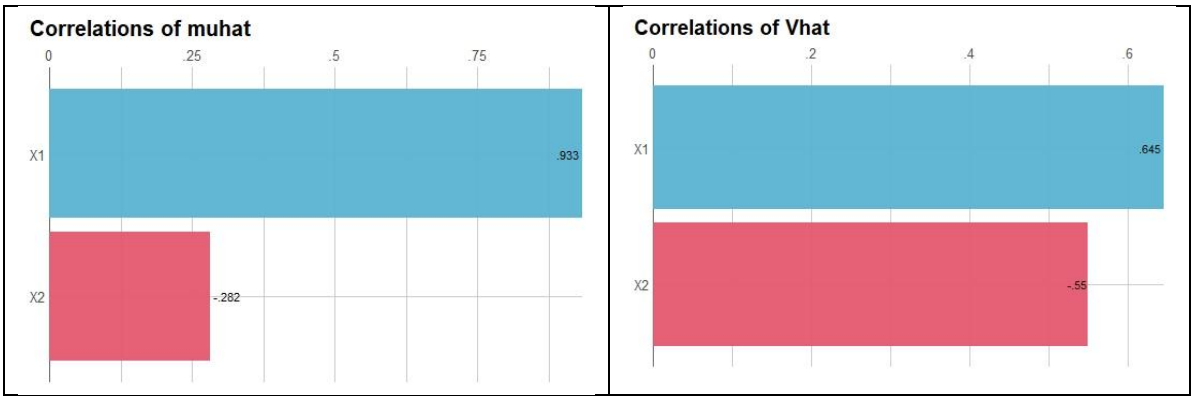


It looks like there is a strong positive relationship between the response variables and and X1 and a small negative correlation between response variables and X2. Also, the relationship does not look like a simple linear relationship. We will experiment with linear, polynomial and non-linear regression models to estimate the response variables.

Next, I plotted the correlation coefficient between response variables and independent variables and it confirmed the above findings as follows

Muhat has a strong positive correlation with X2 and a relatively weaker negative correlation with X2

Vhat has good positive relationship with X1 and a fair amount of negative correlation with X2.



Methodologies and model Training

I split the training data provided to us into a training and a validation dataset for cross validation and model evaluation. 75% of the input data is used in training dataset and the remaining used for validation dataset using random split. A 5 or 10-fold cross validation was performed on each of the models for hyperparameter tuning and to select the optimal model. Data was normalized for all models except those based on decision tree and linear regression. A box-cox transformation was tried for regression model, but that did not improve the model performance. Models Following regression models were fit on the training dataset to estimate expected values of the response variables muhat and Vhat.

1. Linear Regression - Linear regression is a linear approach for modelling the relationship between a scalar response variable and one or more independent variables.

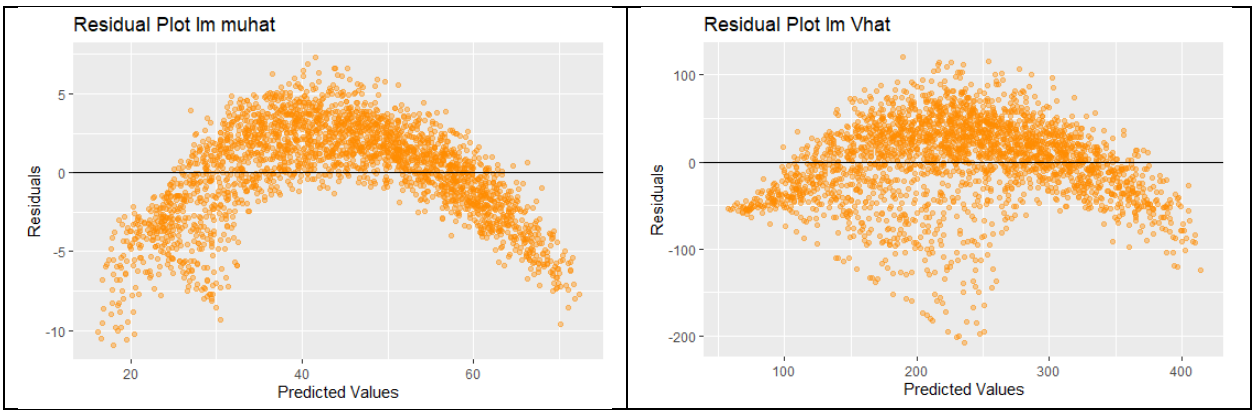
The exploratory data analysis suggested that a basic linear regression model may not fit the data well. However, I used this model to get an understanding of the effect of each explanatory variable on the response variables and response variable and to obtain an initial benchmark for model accuracy.

The linear equation estimated by lm is

muhat = 29.15621 + 44.11818 * X1 - 13.29390 * X2

Vhat = 219.850 + 198.070 * X1 - 166.262 * X2

This shows that X1 has a positive effective and X2 has a negative effect on the response variables. The adjusted R-squared is quite high at 0.9498. However, MSE on validation data for muhat and Vhat were the highest for this model as the relationship is not linear, 9.42 and 2235.03 respectively. The residual plots show a pattern that confirms that the residuals are not random. The linear regression model is not a good choice for our modeling task. Therefore, I tried various methods that can model non-linear relationships.



2. Polynomial Regression - Polynomial regression is a form of regression analysis in which the relationship between independent variables Xi and the dependent variable Y are modelled as an nth degree polynomial of Xi. Even though polynomial regression fits a non-linear relationship between Xi

and Y, it is considered a special case of linear regression as the regression function is linear in terms of coefficients.

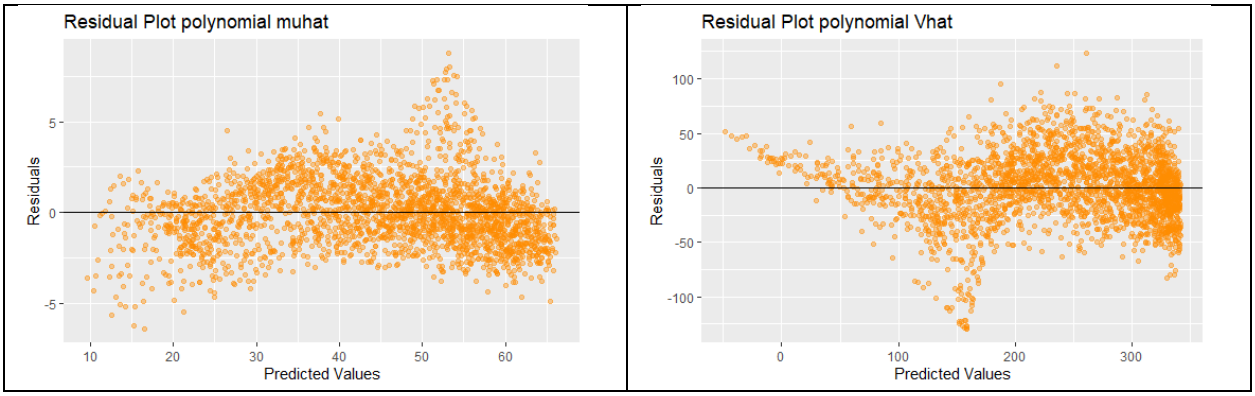
I tried Polynomial regression with various polynomial functions to fit the curvilinear relationship we saw in the plots above in EDA. Out of the few nonlinear relationships I tried, the following model gave the best test results:

Muhat = -6.7075 + 74.46753 * X1 - 3.63085 * X2 - 30.65249 * X1^2 - 9.80826 * X2^2

Vhat = 111.784 + 548.040 * X1 + 147.494 * X2 - 353.661* X1^2 - 317.829 * X2^2

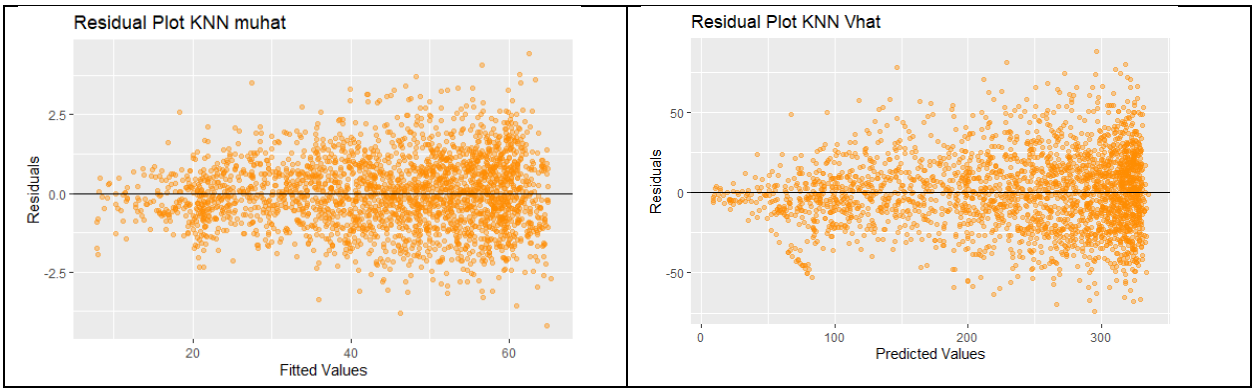
This model suggests that all the variables are signification with a very low p-value

The MSE on validation dataset for muhat and Vhat are 3.68 and 1006.81. These MSE's are much better than the simple linear model. The residual plots look better for muhat., but there still is some pattern.



3. KNN - The k-nearest neighbors algorithm, also known as KNN, is a non-parametric, supervised learning classifier, which uses proximity to make predictions about an individual data point. While it is typically used as a classification algorithm, it can also be used for regression.

KNN performed very well on our dataset. It gave much better MSE's and the residual plot look better too. The best value for k was reported as 23 for muhat and Vhat. The residual plots look better in terms of randomness suggesting KNN does a better job in modeling our data. It shows some level of heteroscedasticity. I tried the model after box-cox transformation but the model performance did not improve with it. Therefore, I continued experimenting with other models.



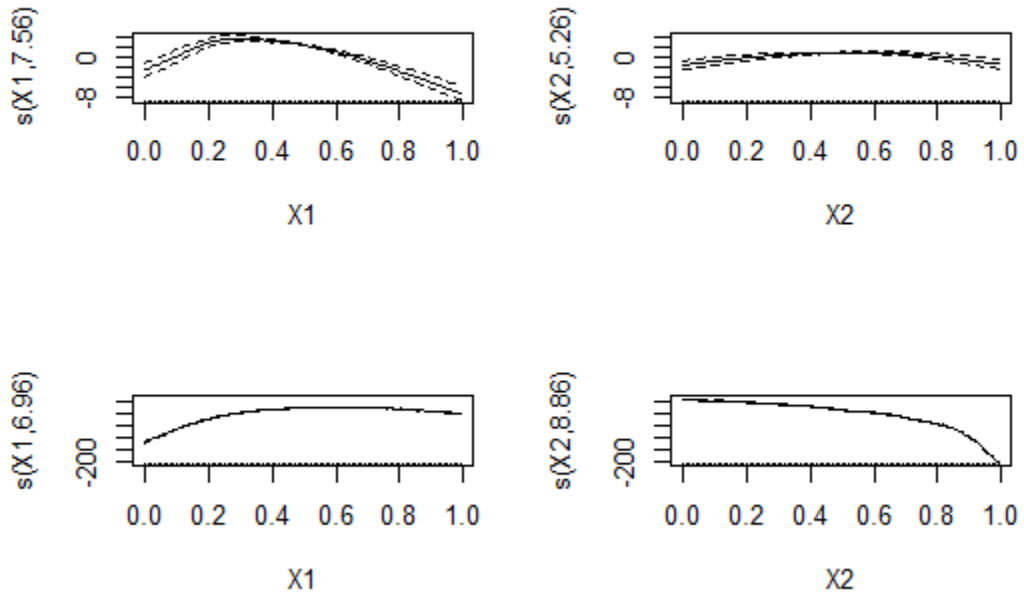
4. LOESS - Loess regression is a nonparametric model that uses local weighted regression to fit a smooth curve through points in a scatter plot. Loess curves can reveal trends and cycles in data that might be difficult to model with a parametric curve. The degree of smoothing can be controlled using the hyperparameter “span”.

LOESS model was fit to the training data. Multiple spans in the range of 0.15 to .85 were tried with a degree of 2 and the best span was found to be 0.15. This may overfit the data as the curve generated by a lower span is bumpier, but the testing error was found to be the lowest among all the models tried.

5. Generalized Additive Models - GAM is an additive modeling technique where the impact of the predictive variables is captured through smooth functions which—depending on the underlying patterns in the data—can be nonlinear. GAM provides a regularized and interpretable solution and can capture common nonlinear patterns that a classic linear model would miss.

GAM model was fit on the training data with automatic selection of smoothing parameter using REML (Restricted Maximum Likelihood). The following plot from gam shows that the model captured the non-linear relationship between independent and dependent variables quite well.

The first row is for muhat and the second row for Vhat

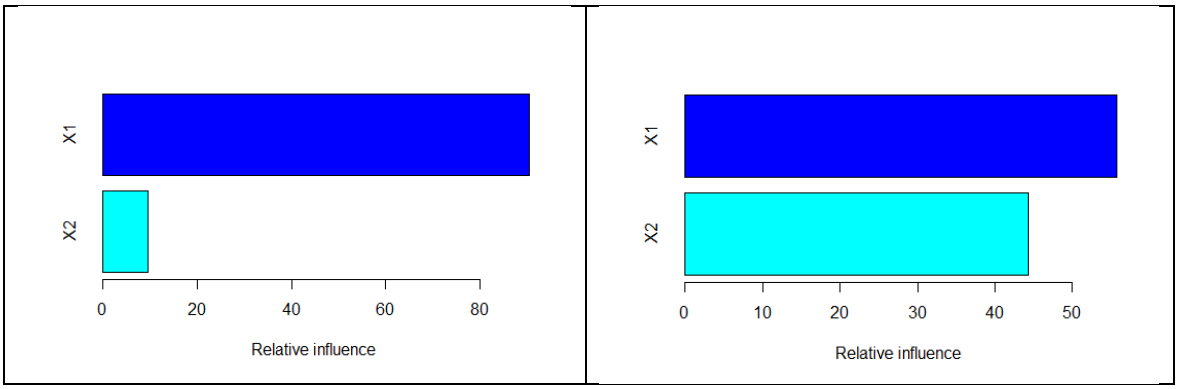


Random Forest - Random Forest is an ensemble of decision trees. These decision trees are constructed in a certain random way with a random sample of rows. At each node, a different sample of features is selected for splitting. Each of the decision trees make individual prediction. These predictions are then averaged to produce the final result.

Random forest was fit with ntrees=500 and the number of variables tried at each split is 1. It gave good MSEs.

6. Generalized Boosted Regression Models (GBM) - These models are a combination of two techniques: decision tree algorithms and boosting methods. Generalized Boosting Models repeatedly fit many decision trees to improve the accuracy of the model. For each new tree in the model, a random subset of all the data is selected using the boosting method. For each new tree in the model the input data are weighted in such a way that data that was poorly modelled by previous trees has a higher probability of being selected in the new tree. This means that after the first tree is fitted the model will consider the error in the prediction of that tree to fit the next tree, and so on. By considering the fit of previous trees that are built, the model continuously tries to improve its accuracy.

GBM are non-parametric and can thus handle skewed and multi-modal data. Gbm was one of the best models I tried. Based on gbm, the relative influence of X1 and X2 or muhat and Vhat are shown below.



7. Neural Network - Neural networks(NN) are a subset of machine learning algorithms inspired by the human brain, mimicking the way that biological neurons signal to one another.

Neural networks are comprised of node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. In my experiment, I have used single layer NN with one hidden layer and a deep NN with multiple hidden layers.

Single layer performed best with 7 nodes in the hidden layer and multiple layer Neural Network was tried with 2 hidden layers of size 4 and 3 respectively. Both the models performed relatively well.

Results

From all the different models tried, I found LOESS to be the best performing model with lease MSE for muhat and Vhat. Many other models such as KNN, Random Forest and GMG gave low MSE's that are close to that of LOESS. However, LOESS is computationally less costly compared to these models. Tree based models Random Forest and GBM works very well when the dataset contains many features. In our case there are only 2 features X1 and X2.

Here is a table of all the models and the respective MSE's:

MSE		
Model	**muhat**	**Vhat**
Linear Regression	9.42	2235.03
Polynomial regression aX1+bX2+cX^2+dX2^2	3.68	1006.82
KNN muhat - best value of k=27 Vhat - best value of k=25	1.3	553.67
LOESS muhat: best span = 0.15 Vhat best span = 0.15	1.27	531.38
Generalized Additive Models REML for optimal smoothing	3.3	706.7
Random Forest No.of variables tried at each split = 1 ntrees = 500	1.46	561.74
Generalized Boosted Regression Models(GBM) muhat best cv iteration 4986 Vhat best cv iteration 4846 interaction depth = 2 shrinkage=0.01	1.28	532.78
Single layer Neural network muhat with size=7 decay=0 Vhat with size=7 decay=0	1.93	555.5
Deep Neural Network hidden = (4,3)	1.92	548.86

Conclusion

From the analysis of input data and various model outputs, we have found that the response variables muhat and Vhat have a non-linear relationship with the predictor variables X1 and X2. Also, X1 has comparatively higher impact that X2 on the response variables, especially muhat.

The exact equation of the relationship between input variables and response variables were not found by the experiment but resulting MSEs and model outputs suggests that we have made a good prediction with models such as LOESS, KNN and the ensemble models.

The final prediction on the test data was made using LOESS and the results are uploaded in the attached csv file.

Thank you for reading!