ISYE 7406: Final Report

Introduction:

This project focuses on predicting both the mean and variance of a random variable Y based on two independent variables X1 and X2. Being able to accurately predict the mean and variance is important in fields like finance, manufacturing, and healthcare, where understanding uncertainty and expected outcomes plays a major role. The goal is to build models that can estimate two functions: the mean $\mu(X1,X2)=E(Y)$ and the variance $V(X1,X2) = Var(Y)$, using machine learning methods. Because the relationship between X1, X2, and Y is likely nonlinear, several different models were explored, ranging from basic linear regression to more flexible approaches like Random Forests and Boosting.

The training dataset, provided in the file "7406train.csv", included 10,000 rows, each representing a unique combination of X1 and X2. Each row contained 200 independent realizations of Y for that specific combination. The first two columns were the input variables X1 and X2 and the next 200 columns contained the observed values of Y. For each (X1,X2) pair, the sample mean and the sample variance were calculated based on 200 realizations. This resulted in a clean modeling dataset with four columns X1, X2, μ, and V.

Methodology:

Before model building, Exploratory data analysis was conducted to better understand the relationship between the variables. Scatterplots showed that X1 had a strong positive relationship with both the mean and variance, while X2 showed a weaker and slightly negative relationship. This suggested that models selected need to be able to capture nonlinear effects, particularly for variance. Additionally, histograms of muhat and Vhat showed that both were right-skewed, with Vhat having a much wider spread. This indicated that predicting variance could be more challenging than predicting the mean. These initial observations helped guide the choice of models and tuning decisions in the next steps..

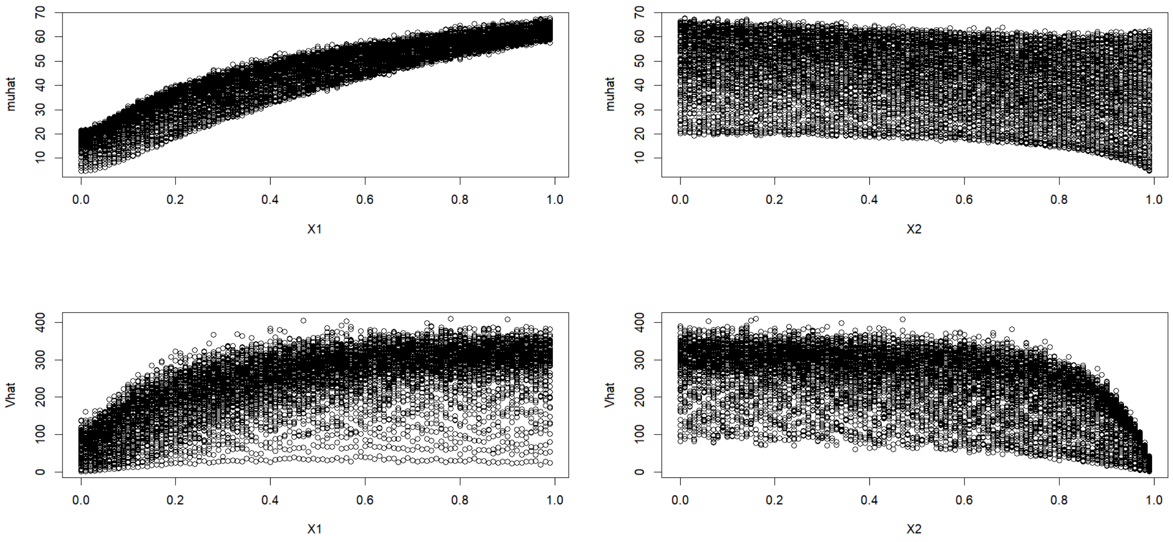


Figure 1: Scatterplots (X1 and X2 vs muhat and Vhat)

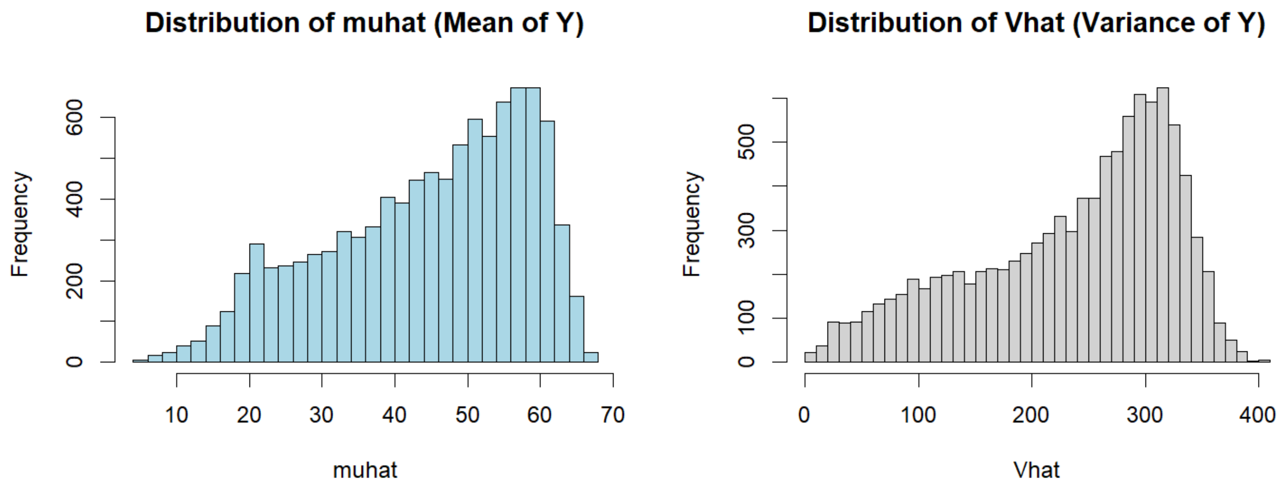


Figure 2: Distributions of Mean and Variance

The training data was randomly split into 80% for training and 20% for validation. This split made it possible to check how well each model performed before using it to predict the final testing data. Mean squared error (MSE) was used as the main way to evaluate models throughout the project. To estimate the mean and variance functions accurately, several modeling strategies were explored.

The first models were basic linear regression models to predict the mean and variance based on X1 and X2. Initially, only additive relationships were included without interactions. Later interaction terms between X1 and X2 were included to capture possible combined effects. Linear regression served as a simple baseline for comparing other, more complex models.

Random Forest models were then trained to predict the mean and variance separately. Random Forest is an ensemble method that combines multiple decision trees which helps capture nonlinear patterns and reduces overfitting. The models were trained with 500 trees (ntree = 500) and used a minimum of 10 observations per leaf (nodesize = 10). The number of variables considered at each split (mtry) was left at the default. Random Forest handled complex relationships between X1 and X2 without the need for manual feature engineering.

Boosting works by fitting models one after another, with each new model trying to fix the mistakes of the previous ones. The boosting models used 500 trees (n.trees = 500), a small learning rate (shrinkage = 0.01), and an interaction depth of 3. Five-fold cross-validation was included during training to help control overfitting. Boosting performed especially well for predicting the variance.

Generalized Additive Models (GAMs) were tested next. GAMs use smooth curves instead of straight lines, which allows more flexibility to capture nonlinear relationships without needing to specify exactly how the inputs and outputs are related. A Gaussian distribution with an identity link was used for the models.

Every model was evaluated based on its performance on the validation set. Separate models were trained for predicting the mean and the variance, and their respective MSE values were compared. Final models were then used to make predictions on the testing dataset.

Results:

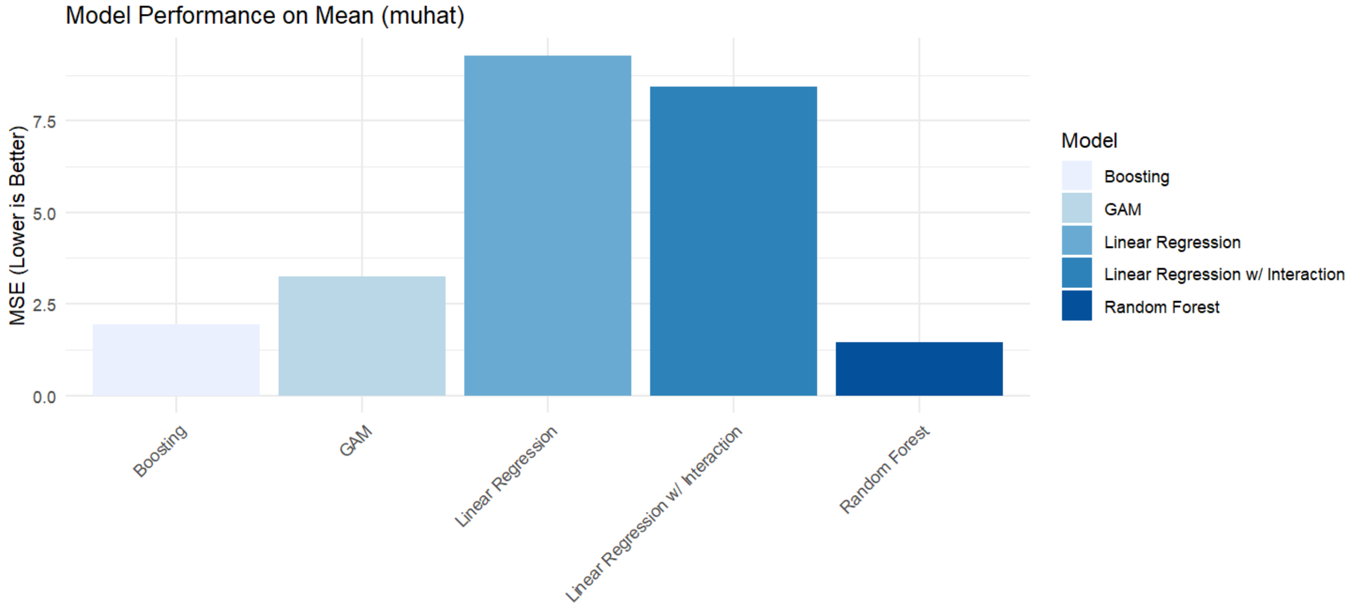
Model performance was evaluated by comparing the MSE of each model on the validation set. Results were organized separately for mean prediction (μ) and variance prediction (V).

Model	MSE (μ)	MSE (V)
Linear Regression	9.288	2260.947
Linear Regression w/ Interactions	8.433	-
Random Forest	1.456	557.994
Boosting	1.934	548.796
GAM	3.259	690.583

Table 1: Summary of validation MSE for mean and variance prediction

For predicting the mean of Y, the Random Forest achieved the lowest MSE, followed closely by the Boosting model. Linear regression models, even those with interaction terms, had much higher errors compared to the ensemble methods. The GAM model performed better than basic linear regression, but did not outperform the tree-based methods. Random Forest was especially strong at capturing complex nonlinear patterns between X1 and X2.

For predicting the variance of Y, Boosting performed slightly better than Random Forest, achieving the lowest MSE. Both tree-based methods significantly outperformed the linear models and GAMs. Variance prediction was more challenging overall, with higher error values compared to mean prediction. Boosting models helped capture complex variance patterns more effectively.



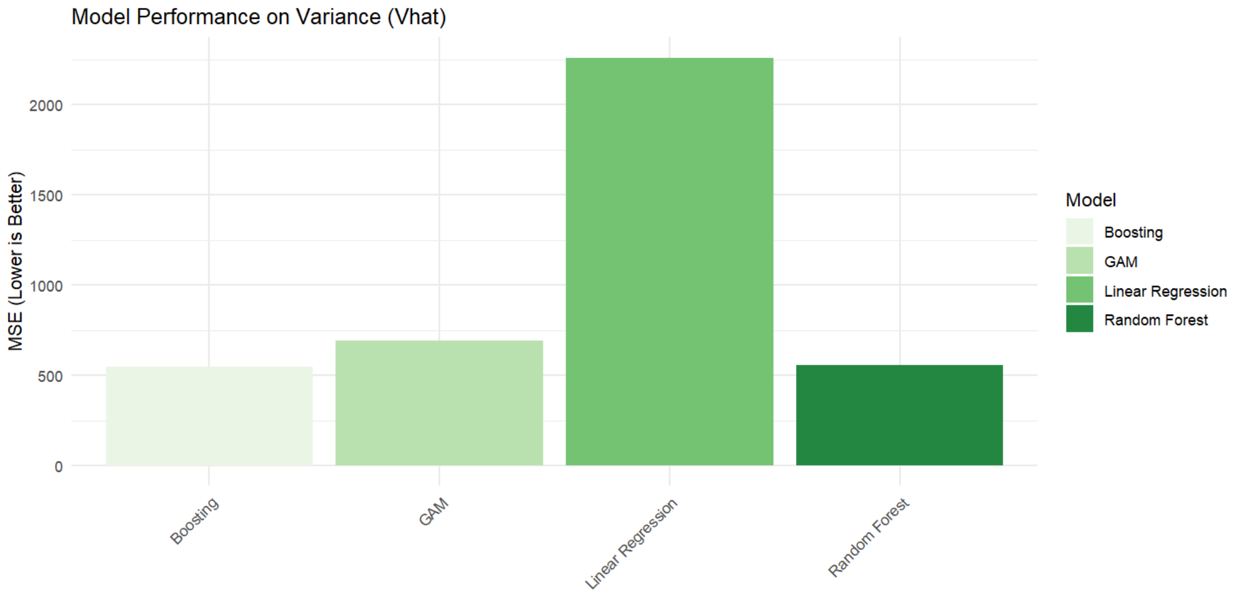


Figure 3: Model Performance

Conclusion:

Several modeling approaches were explored, including linear regression, linear regression with interaction terms, random forest, boosting, and generalized additive models. Based on the validation results, tree-based models performed best. Random Forest had the lowest error for predicting the mean, while Boosting had the lowest MSE for predicting the variance. Linear models, even with interaction terms, had larger errors and struggles to capture complex relationships. GAM models performed moderately, better than basic linear regression but did not outperform ensemble models.

Exploratory data analysis showed that X1 had a strong influence on the mean of Y, while the variance of Y was more random and harder to predict. Histograms showing the distributions of the outcomes looked fairly symmetric for the mean, but the variance was skewed to the right.

The results showed that flexible models, like Random Forest and Boosting, are important when working with nonlinear data. These methods gave much better predictions than simple linear models. Future work could involve more careful tuning of model settings, trying higher-order interaction terms, or testing more advanced methods like neural networks to improve accuracy even further.