Summary

Model 1: Uses Logistics Regression and variables identified from assignment 02 from correlation plot and mutual information

Model 2: Uses Logistics Regression and variables used in class

Model 3: Uses Naïve Bayes and variables identified from assignment 02 from correlation plot and mutual information

Model 4: Uses Naïve Bayes and variables used in class

Variables Description

Variables Identified from Assignment 02	Variables Used in Class
1. Channel Condition 2. Structural Evaluation 3. Age 4. Age Square	1. Average Daily Traffic 2. Percentage of Truck 3. Reconstruction Variable 4. Age 5. Age Square

- Uses Culvert Condition as the output: 0 for Satisfactory & 1 for Unsatisfactory

Metrics Table

Parameters	Model 1	Model 2	Model 3	Model 4
Accuracy	0.955	0.6310	0.899	0.631
Precision	0.917	0.646	0.886	0.680
Recall	1.0	0.585	0.916	0.503
F1 Score	0.957	0.614	0.901	0.578
AUC	0.9657	0.6707	0.8943	0.6349

Discussions & Model Comparison

- For logistic regression, model 1 performs better in all aspects than model 2. The features selected in model 1 was based on the correlation plot and thus have better accuracy. The channel condition and structural evaluation showed strong correlation with the output parameter which are used in model 1. The features selected in model 2 was based on the variables used in class for a different dataset. As it can be seen from the correlation plot, the reconstruction parameter shows weak correlation with the output variable. This has been used in model 2.
- Similarly, for Naïve Bayes algorithm model 3 which uses the identified variables from assignment 02 shows greater accuracy than model 4 which uses the variables used in class.
- Between model 1(logistic regression) & model 3(Naïve Bayes Algorithm), model 1 shows greater accuracy with almost a perfect precision in recall score. Thus, it can be concluded that model 1 is the best working models from all the models presented.

- Between model 2(Logistic Regression) & model 4(Naïve Bayes algorithm), even though both show the same accuracy the AUC value of model 2 is greater than model 4. Even though the precision score of model 4 is greater than model 2, it falls short of the recall score. Thus, while predicting the unsatisfactory culvert condition, model 2 will perform better than model 4.
- From the presented models above, the models with logistic regression show higher accuracy than the models with Naïve Bayes Algorithm. Naïve Bayes Algorithm is a general model where it expects the features to be independent and Logistic Regression is a discriminative model that performs better using collinearity. Due to a large dataset and use of regularization, Logistic regression performs better. Naive Bayes algorithm might compensate the loss while working on a small dataset.