PROJECT REPORT

(Project 2 – Classification Analysis)

(EAS 508 – Statistical Learning and Data Mining I)

Team Members:

1. Jahnavi Gangu (jgangu@buffalo.edu)
2. Sai Teja Ankuru (saitejaa@buffalo.edu)
3. Utkarsh Mathur (umathur@buffalo.edu)

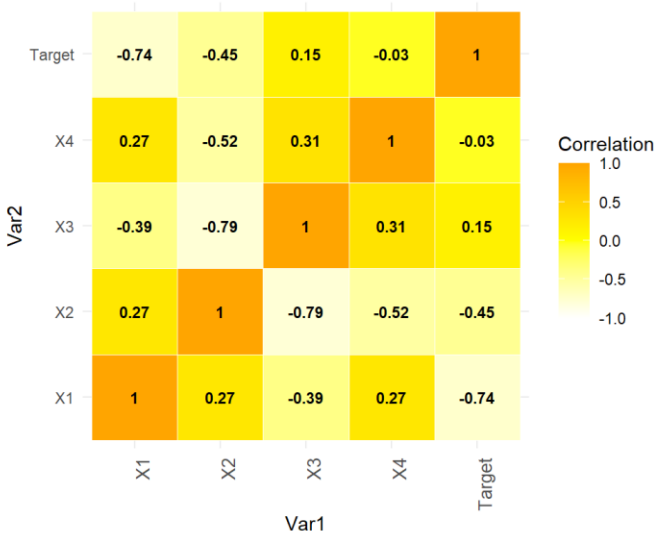
Instructor:

Dr. Leonid Khinkis

---

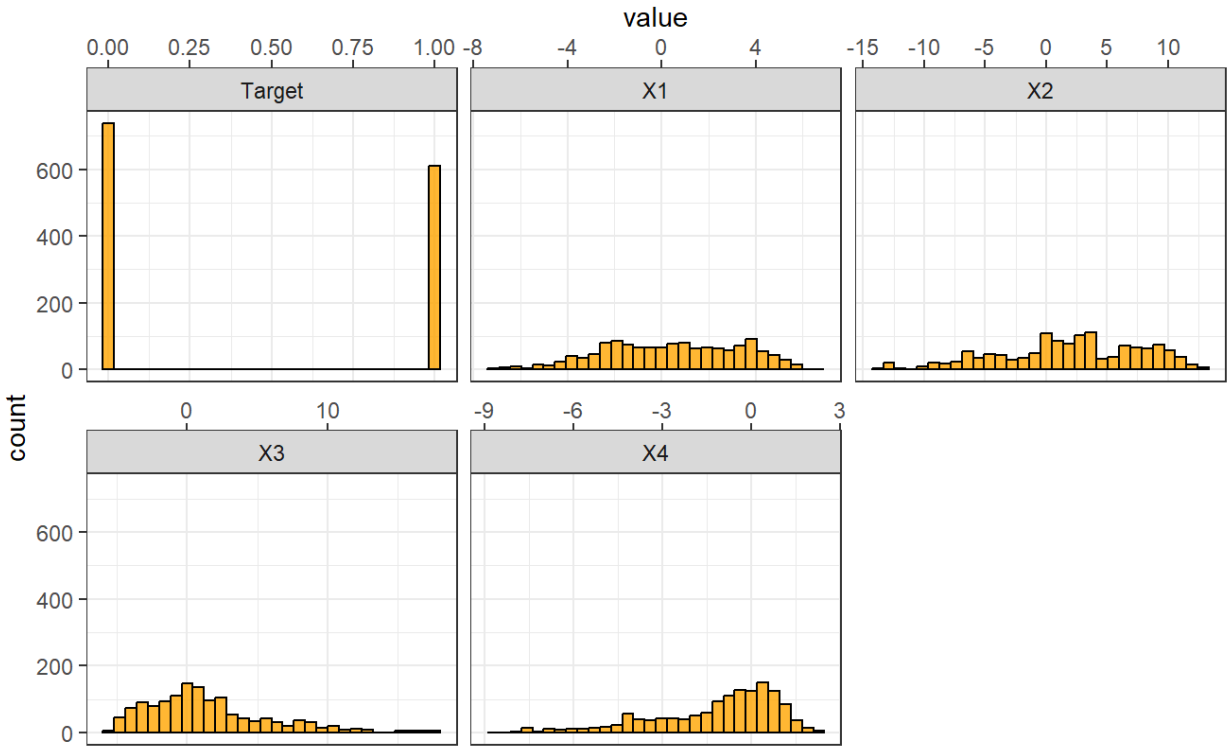
1. INTRODUCTION

- The aim of the project is to practice classification analysis on a data set of 1348 observations and 4 predictors. The dataset was divided into a training set of 1088 observations and test set of 260 test observations.
- The correlation matrix of the predictions is given in Figure 1. As we can observe there is a slight correlation between predictors X2 and X3 which can be neglected as there are only 4 predictors.



(Figure 1 – Correlation Matrix)

- The distribution of features are given in Figure 2. From these distributions we can observe that there is a straight-forward connection between the predictors and response.



(Figure 2 – The distribution of prediction features)

- We have used 7 models namely, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Ridge Classification, Random Forest Classification, Decision Tree Classification, and Support Vector Machines.

2. RESULTS & ANALYSIS

2.1. LOGISTIC REGRESSION

- The Logistic Regression model produces 99.23% accuracy.
- It can be seen from the model fitted model parameters that the p-value of X4 feature is very high which suggests that X4 explains limited variability of response variable.

2.2. LINEAR DISCRIMINANT ANALYSIS

- The accuracy of LDA is 99.23% which is the same as Logistic regression. This suggests that LDA has limited regularization effect on the test accuracy.

2.3. QUADRATIC DISCRIMINANT ANALYSIS

- The accuracy of QDA is 99.23% which is the same as Logistic regression. This suggests that QDA has limited regularization effect on the test accuracy.

2.4. RIDGE CLASSIFICATION

- The accuracy of Ridge classifier is 97.31% which is the same as Logistic regression. This suggests that Ridge classifier has negative regularization effect on the test accuracy.

2.5. RANDOM FOREST CLASSIFICATION

- The accuracy of Random Forest is 100% with 96.64% variance explained.
- This suggests that Random Forest does very well on the training data as well as test data but as the total number of observations is very less it also hints towards slight overfitting.

2.6. DECISION TREE CLASSIFICATION

- The accuracy of Decision Tree is 87.69% which is very low as compared to the peer models.
- This can be anticipated from the data distributions as there is no clear distintion between the two classes with respect to any one predictor variable.

2.7. SUPPORT VECTOR MACHINES

- The accuracy of Support Vector Machine is 100% with 403 support vectors (out of 1088 observations).
- This suggests that Support Vector Machine does very well on the training data as well as test data but as the total number of observations is very less it also hints towards slight overfitting.

MODEL NAME	TEST ACCURACY
Logistic Regression	99.23%
Linear Discriminant Analysis	99.23%
Quadratic Discriminant Analysis	99.23%
Ridge Classification	97.31%
Random Forest Classification	100%
Decision Tree Classification	87.69%
Support Vector Machines	100%

(Table 1 – Test Accuracy Results)

3. CONCLUSION

- We conclude that the Random Forest Classifier and Support Vector Machine produces the best test accuracy of 100%. Although it can also indicate overfitting on training data and highly correlated training and test datasets.
- The predictor X4 is highly uncorrelated to the response variable and that is why it has less influence on the prediction fit.
- The Decision Tree Classifier returns the worst results on test data which can be justified by lack of distinction between classes on the basis of anyone of the predictor variables.

4. REFERENCES

i. stats-learning-notes (tdg5.github.io)
ii. Notes for Elements of Statistical Learning — ESL Notes 0.1 documentation (esl-notes.readthedocs.io)
iii. Resources - Second Edition — An Introduction to Statistical Learning (statlearning.com)