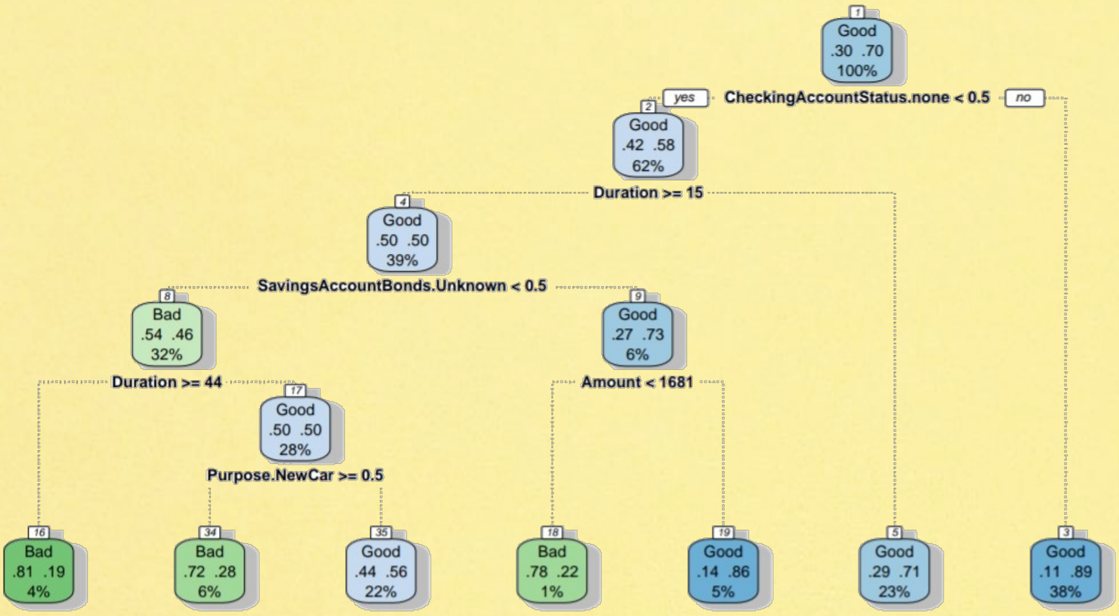
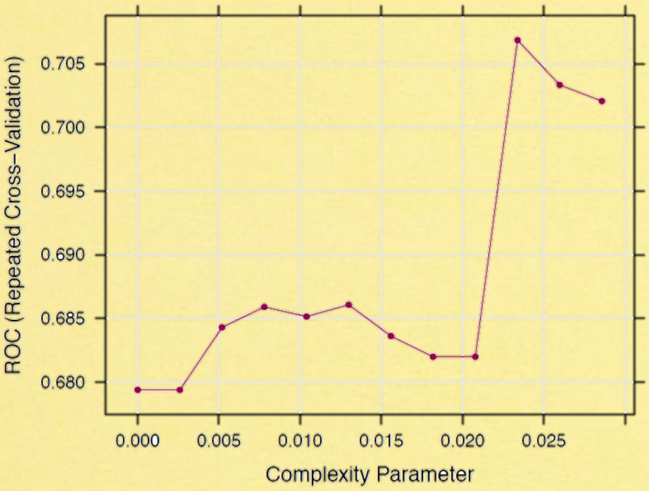
MACHINE LEARNING REPORT

Q1. Decision Tree:



ROC vs Complexity Parameter



Comments: By pruning the decision tree with a complexity parameter of 0.0233, the outcome of the pruned model looks more interpretable, less complex and easier to understand. With this, we are able to classify the data and predict defaults on consumer loans in the German market. Overall, the decision tree had a test error rate of approximately 26.3%.

Here, checking account status of the customer is the root node. From the decision tree, we can identify three possibilities of being classified as the ‘Bad’ class:

(a) If the customer does not have a checking account, credit duration is >= 15 months, amount of money available in savings/bonds is unknown and credit amount is less than 1681.

(b) If the customer does not have a checking account, credit duration is >= 15 months, amount of money available in savings/bonds is not unknown and credit duration is >= 44 months.

(c) If the customer does not have a checking account, credit duration is >= 15 months, amount of money available in savings/bonds is not unknown, credit duration is less than 44 months and purpose is a new car.

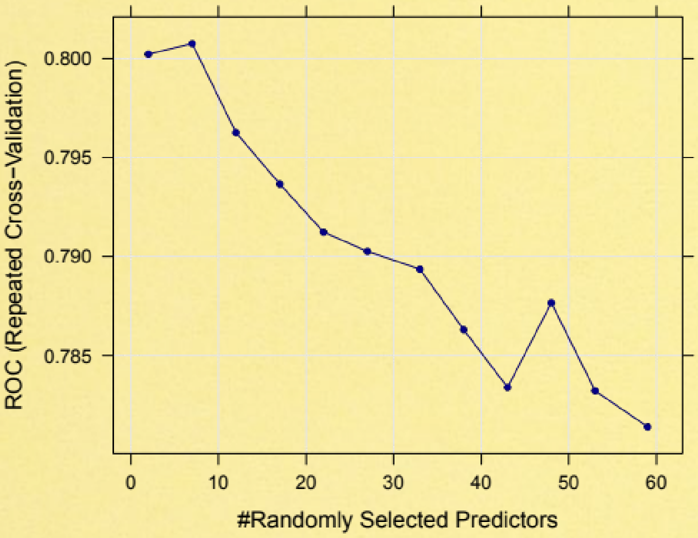
Q1. Random Forest:

Comments: Using a 5 fold cross validation, 'ntrees' (i.e. number of trees) of 1000 and 'mtry'(i.e. no. of variables tried at each split) of 7, the random forest arrived at an accuracy of approximately 77% (test error rate: 23%). The OOB error rate was similar to the test error rate at 23.14%. Hence, on average, the model misclassified 23.14% samples during training.

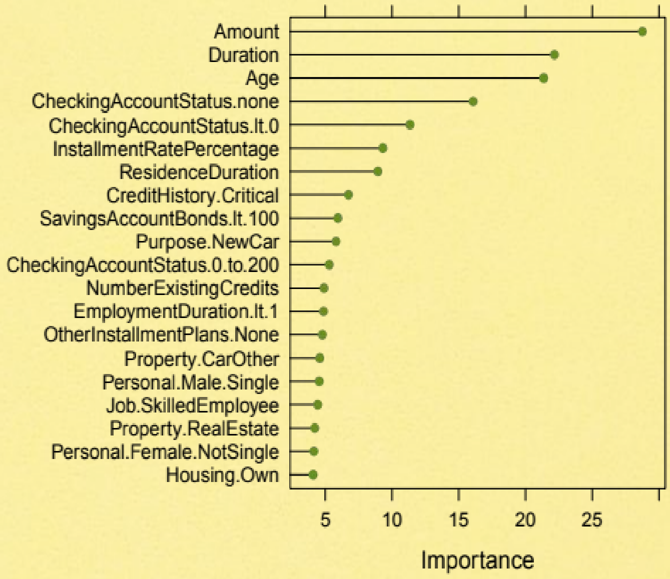
An overview of the confusion matrix:

- o. True Positives (446): Instances classified as 'Good' that the classifier predicts correctly
- o. True Negatives (92): Instances classified as 'Bad' that the classifier predicts correctly
- o. False Positives (118): Instances of class 'Good' that the classifier incorrectly predicted as 'Bad'
- o. False Negatives (44): Instances of class 'Bad' that the classifier incorrectly predicted as 'Good'

ROC vs Mtry



Variable Importance Plot



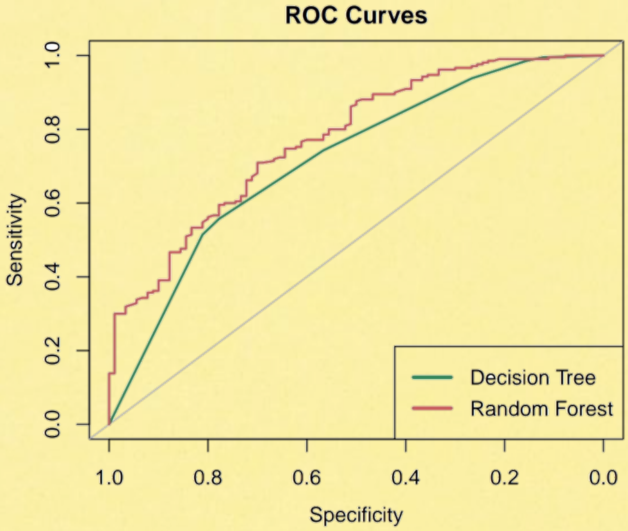
A **Variable Importance Plot** highlights the significance of each variable in predicting the target. According to the Variable Importance Plot, 'Credit Amount', 'Duration' and 'Age' are the three most important/influential variables in classifying the German credit data and predicting defaults on consumer loans in the German market. In fact, this makes sense intuitively. Analysing variables such as amount, duration and age can help evaluate credit worthiness of a customer:

Amount and Duration: The credit amount a customer is seeking and the time available to pay back the loan can often be predictive characteristics of a potential bad credit. Can the borrower afford to repay the loan? Can they keep up with monthly installments?

Age: Age often reveals important characteristics of a person. If a customer is too young, are they financially stable? Do they have appropriate financial knowledge? On the other end, if a customer is too old, what are their potential health concerns? what is their probability of potential unemployment? and what is their source of regular income?

Q1. ROC Curve

The ROC curve represents the trade-off between sensitivity and specificity. Generally, classifiers that give curves closer to the top-left corner indicate a better performance. Hence, as compared to a decision tree (represented in green), the random forest (represented in red), with a curve closer to the top-left corner, was more accurate in classifying the data and predicting defaults on consumer loans in the German market. This is also reiterated by the AUC scores (higher the AUC, better the model tends to perform):

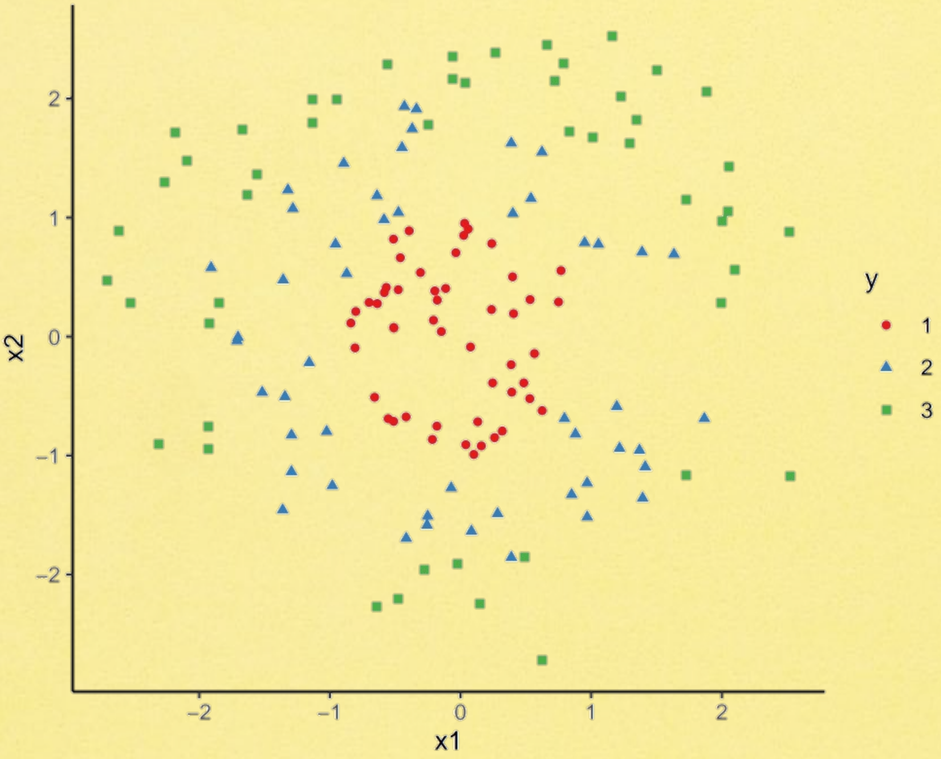


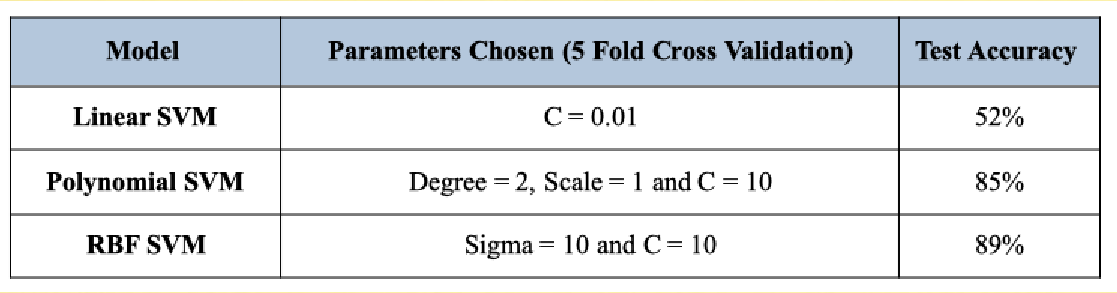
Decision Tree AUC Score: 0.72

Random Forest AUC Score: 0.77

Q2. Scatter Plot:

The following is the scatter plot of the linearly inseparable simulated data. Overall, there are three classes (50 observations each) arranged in concentric circles, differentiated by three distinct shapes & colours:

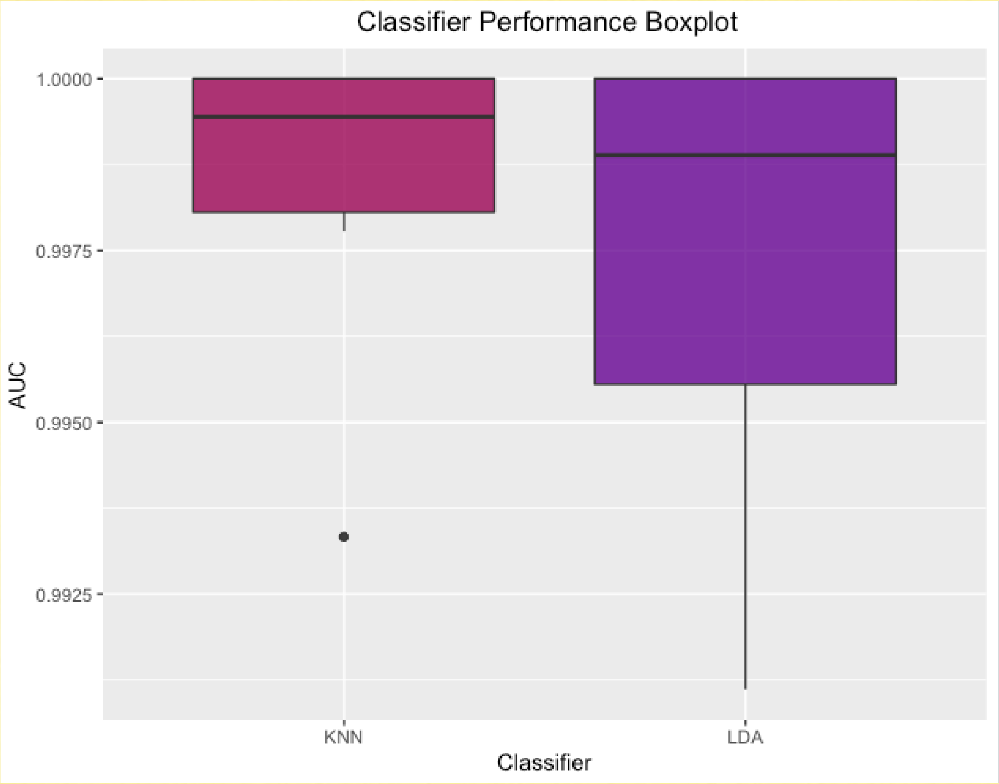




Model	Parameters Chosen (5 Fold Cross Validation)	Test Accuracy
Linear SVM	C = 0.01	52%
Polynomial SVM	Degree = 2, Scale = 1 and C = 10	85%
RBF SVM	Sigma = 10 and C = 10	89%

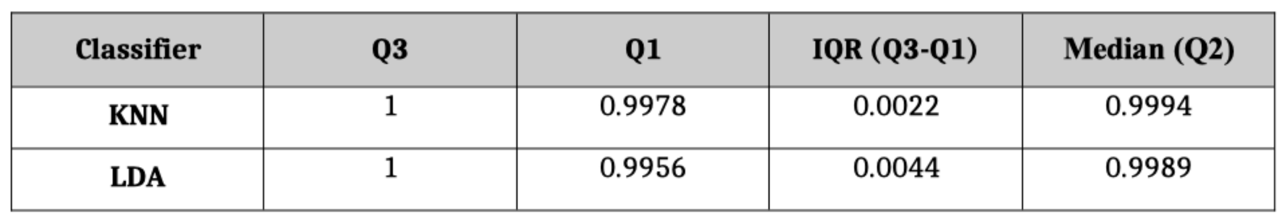
Comments: Firstly, it is important to interpret why the Linear SVM performed so poorly. Often Linear Kernel performs effectively when the data points are linearly separable (i.e. the decision boundary is a straight line). Clearly, our simulated datapoints are arranged in concentric circles and hence it does not make sense to use a linear classifier (therefore, resulting in a low accuracy of 52%). While both Polynomial kernel and RBF kernels have performed well, the latter has slightly outperformed the former. As compared to polynomial kernels, RBF is much more efficient. It combines multiple polynomial kernels of varying degrees multiple times in order to map the data into a high dimensional space, making it easier to separate using a hyperplane. In addition, RBF kernels work quite well with noisy or overlapping datasets. Hence, RBF is quite powerful in the above scenario (and therefore, the high accuracy).

Q3. Boxplot:



Comments:

(i) Comparison of Location: The median AUC score of KNN classifier (0.9994) is slightly greater than the median AUC score of the LDA classifier (0.9989). In addition, the range of AUC scores of both classifiers overlap.



Classifier	Q3	Q1	IQR (Q3-Q1)	Median (Q2)
KNN	1	0.9978	0.0022	0.9994
LDA	1	0.9956	0.0044	0.9989

(ii) Comparison of Skewness: Since the medians of both the boxes are closer to Q3 than Q1 (Q3-Median < Median-Q1) the distributions are negatively skewed or left skewed.

(iii) Comparison of Dispersion: By analysing the interquartile ranges and the whiskers, it can be interpreted that there is more variability detected in LDA AUC data, whereas KNN AUC values are tightly grouped.

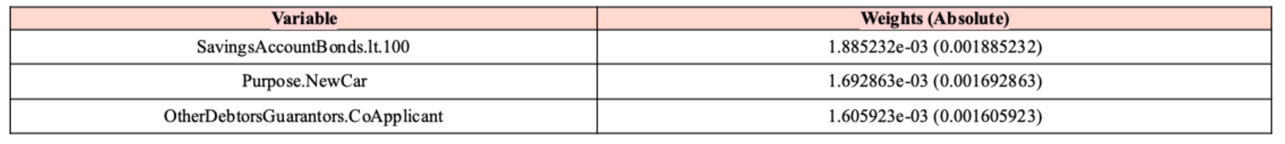
General Conclusion: Overall, it is safe to say that both KNN and LDA have performed exceptionally well (with AUC scores above 0.99) in classifying the thyroid data and in predicting whether a patients suffers from hyperthyroidism or not. (Note: We can observe how KNN with cross validation is very powerful and generates complex decision boundaries, hence resulting in the high AUC's).

Q4. FDA:

FDA Formula: $\mathbf{w} = S_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \in \mathbb{R}^{p\times 1}$

Threshold Formula: $c = \mathbf{w}^T \cdot \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$

Interpreting Weights: The absolute value of the weights indicates the relative importance of the features in classification of data. Hence, based on the weights obtained, the three most important variables are:



Variable	Weights (Absolute)
SavingsAccountBonds.lt.100	1.885232e-03 (0.001885232)
Purpose.NewCar	1.692863e-03 (0.001692863)
OtherDebtorsGuarantors.CoApplicant	1.605923e-03 (0.001605923)

Interpreting Signs: Interpreting the sign of the weights can reveal critical information about the direction of relation between the predictor and response. A positive weight for variables (i.e. age, checking account status and resident duration) indicates that as the value of these variables increases, the likelihood of belonging to the class good also increases. Hence, higher age, improved checking account status and longer duration can be predictive characteristics of good credit. Similarly, a negative weight for variables (i.e. Job.UnemployedUnskilled and CreditHistory.delay) indicates that as the value of these variables increases, the probability of belonging to the class bad increases (hence, decreasing the probability of belonging to class good).