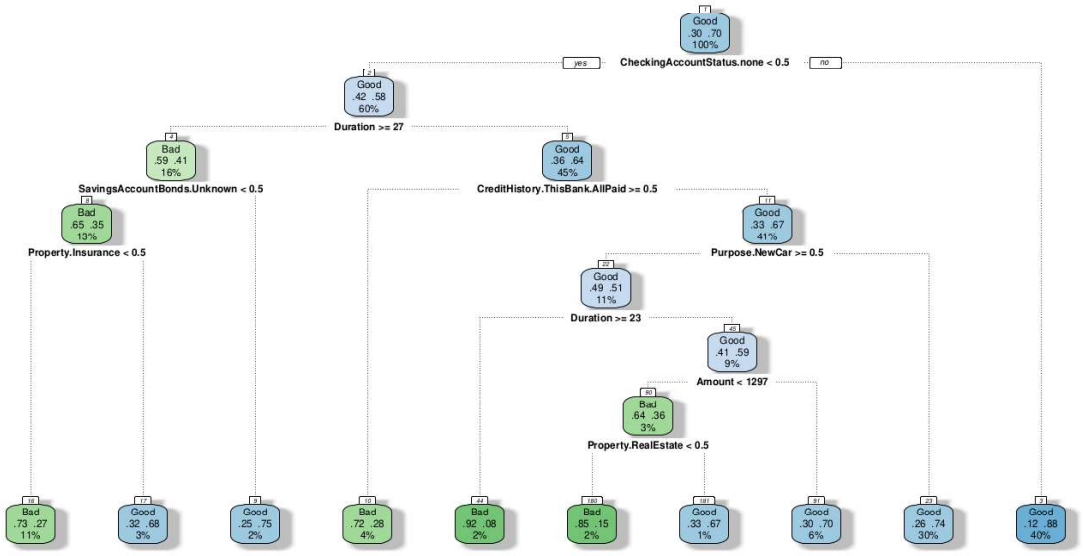
SMM636 Machine Learning – Individual coursework submission

Q1: (1) A decision tree was fit to the training data with the optimal tree size determined by 5-fold cross-validation. Following this, cost complexity pruning was performed to avoid overfitting on the training data. The tree shown below was obtained.



The optimal number of leaves is 10, obtained at an alpha value of 0.01190.

The checking account status is the most discriminative feature to determine if a person has good credit or not which is the root node. Overall, eight of the 59 features are used to determine the good/bad credibility class of the customer.

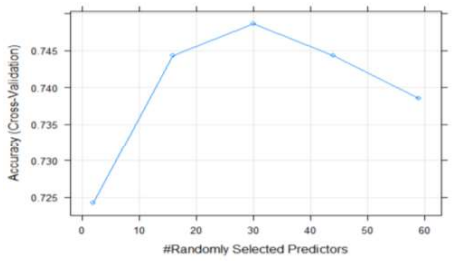
Out of the leaf nodes, 40% of the customers fall into the right-most node using just the CheckingAccountStatus.none, of which customers are labelled to be 'good' with 88% accuracy. On the other hand, the node with the highest purity has only 2% of the entire number of customers.

When the tree is used to predict on the test data dataset, the accuracy is 72% and the test error rate is 28%. The Confusion matrix is as follows.

		Predicted	
		Bad	Good
Actual	Bad	27	21
	Good	63	189

Q1: (2) A random forest with 1000 trees was fit to the training data with the number of features to create the splits tuned by 5-fold cross-validation. The optimal number of features was computed as 30. This can be seen in the plot below where the cv accuracy is plotted on the Y-axis vs the number of the predictors on the X-axis.

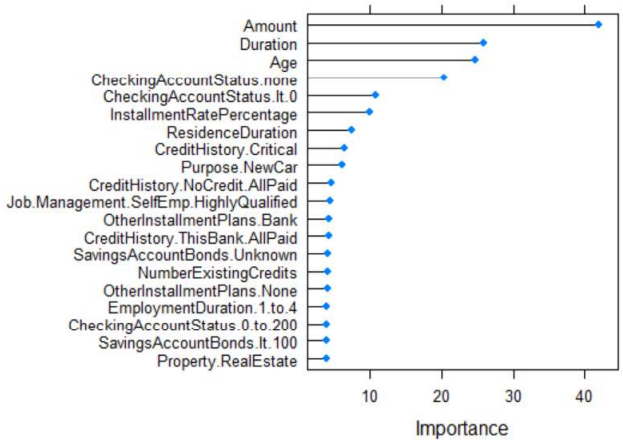
Soumya Ogoti



The test error rate for this model on the test data is 22% which is an improvement over the decision tree classifier. The Confusion matrix is as follows.

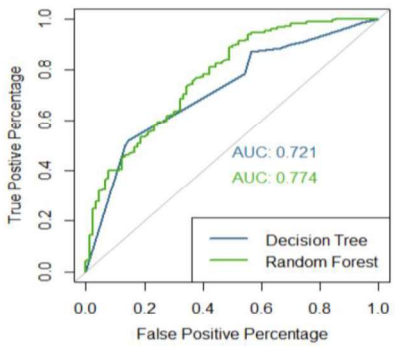
		Predicted	
		Bad	Good
Actual	Bad	40	17
	Good	50	193

The variable Importance plot for the top 20 most important variables in this model.



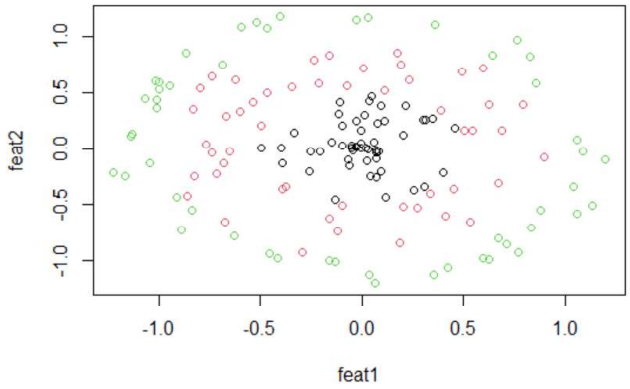
The most important feature for classification is the Amount (42.02) followed by duration (25.85) and age (24.81). The least important features are Purpose. Retraining (0.19) and Purpose.Other (0.28). The higher values of importance indicate that the Gini index by the highest amount if the feature is included.

Q1: (3) The ROC curves for the decision tree (blue) and the random forest (green) models are shown below.

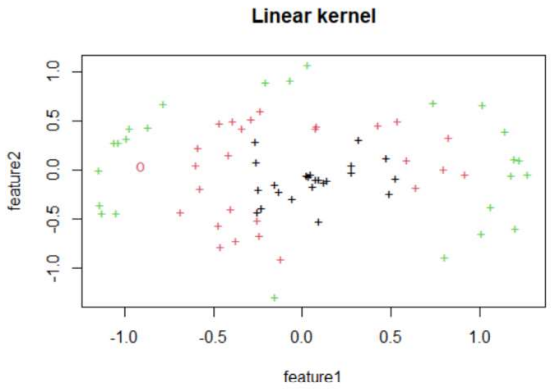


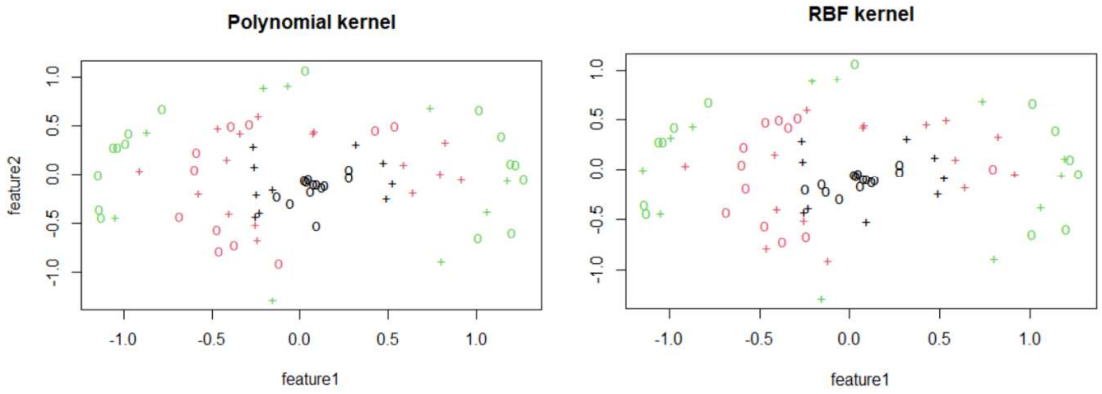
This plot is computed by interpreting the outputs of the two models as probability distributions and computing the tpr, and fpr for various thresholds. The area under the curve (AUC) is higher for the random forest classifier which indicates that it is a better model than the decision tree. This is because the predictive effort of multiple decision trees is taken into consideration in the random forest classifier.

Q2: (1) A three-class dataset was simulated with 50 observations in each class with two features for every data sample. Data was generated in three concentric circles fashion to ensure that the classes are not linearly separable. The scatter plot is shown below.



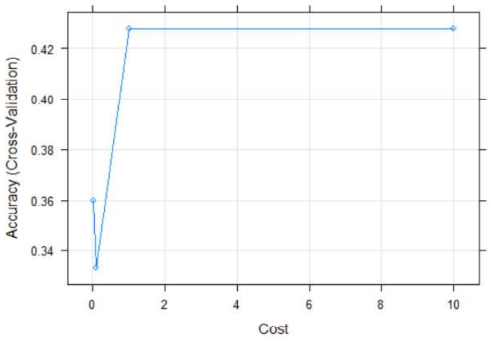
Q2: (2) The dataset was split into train-test (50-50%). Support vector machines with three different kernels – linear, polynomial and an RBF kernel were used for classification. The parameters for each model were tuned by 5-fold cross-validation. For linear kernel, no parameter except cost (an SVM parameter common to all three) was tuned. For the polynomial kernel, the degree of the polynomial and scale was tuned and for the RBF kernel, the sigma was tuned.



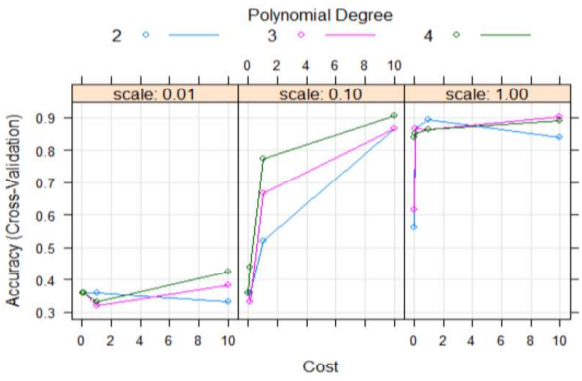


In the plots shown above, each of the data points is shown with an ‘o’ sign and support vectors are shown with the ‘+’ sign.

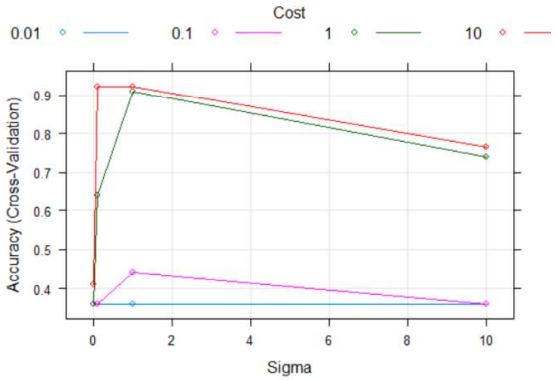
The linear kernel is unable to separate the data well and has an accuracy of only 46.66%, most of the points are also support vectors. This is expected as the data is in the form of concentric circles on which linear classifiers will perform poorly. The figure on the right shows the accuracy for different cost values.



The polynomial kernel performs significantly better and has an accuracy of 86.66%, the points along the boundary of the classes are support vectors as expected. The final degree of the polynomial kernel that classified the data well was 4. The figure on the right shows the accuracy for different degree and cost values.



The RBF kernel further improves the performance with an accuracy of 88%. This was for a sigma of 1. This boost in performance is because the non-linear boundaries in the dataset can be captured well by the RBF kernel.



Q3: (1) The newthyroid.txt data was split into 70% training and 30% test sets. This was repeated ten times with random splits. For each split, a kNN classifier and an LDA classifier were fit.

For kNN, a 5-fold CV was used to choose k from a list of values (3,5,7,9,11,13,15) with AUC as the metric for evaluation. The best k for each split is shown below.

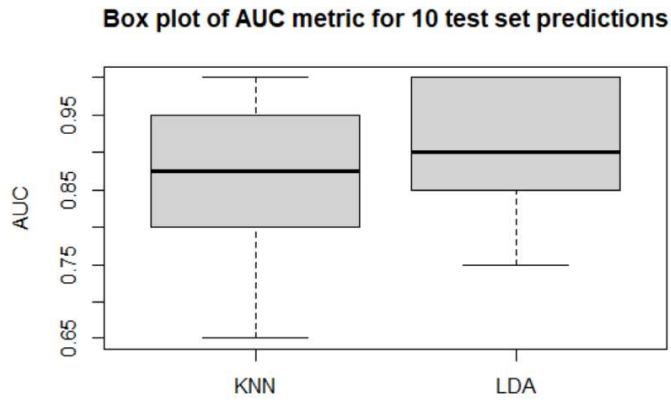
15	15	15	15	11	13	15	15	15	15

There are no parameters to tune for LDA.

The AUC scores for the models on the test dataset for the ten random splits are as follows.

Model	AUC scores									
kNN	0.65	0.9	1.0	0.95	0.9	0.75	0.85	0.95	0.8	0.85
LDA	0.75	0.9	1.0	1.0	0.85	0.9	0.85	1.0	0.8	0.9

Q3: (2) The box plots for the AUC scores are shown below.



Q3: (3) Both the methods, kNN and LDA perform well on the newthyroid.txt dataset. AUC is a good metric to compare the two methods over Accuracy because the dataset is unbalanced and has fewer hyperthyroidism samples in comparison to normal samples. With the AUC metric, the LDA classifier performs better than the kNN classifier with a higher mean AUC (0.89 vs 0.86). The variance of both methods is similar, 0.007 for LDA and 0.011 for kNN.

For this data, both methods predict very few false negatives (on average: 3) which is the desired outcome here as positive cases should not be missed.

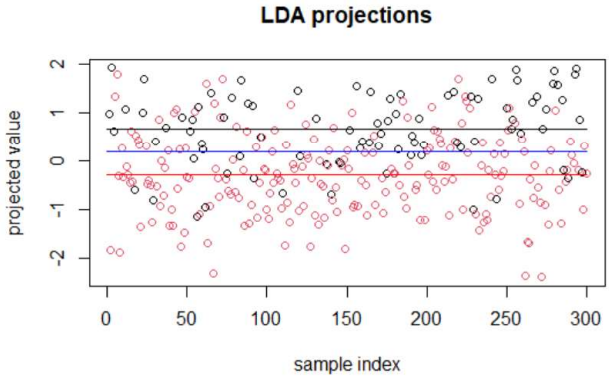
Q4: The myFDA function computes linear discriminant vector w according to Fisher's LDA formulation as

$$w \propto S_w^{-1}(\mu_1 - \mu_2)$$

where μ_1 and μ_2 are the sample means of the two classes respectively and $S_w = S_1 + S_2$ is the within-class scatter. S_1 and S_2 are the sample covariances of the two classes.

For the German credit data, PCA is applied to remove the features that are highly correlated and reduce the dimensions to those that explain 90% of the variance.

The myFDA function is then applied to the training split (70%) to obtain the w. As we performed PCA, w here corresponds to the PCs. The magnitude and sign of each component of w indicate the degree to which the corresponding feature contributes to the separation between the classes. Here, components of w having positive values indicate that the increase in the corresponding feature value will push the sample towards one class. Similarly, for negative values, it will push towards the other class. PC2 has the maximum positive component of the weight (0.3052834) and PC30 has the maximum negative component of the weight (-0.1671047). An accuracy of 72.66% was achieved.



The above plot shows the projections of the test data using the computed w. The colours (red, black) indicate the true classes and the corresponding horizontal lines are the class specific means. The blue line is the decision threshold.