Problem Set 10

Yeganeh Karbalaei

April, 2025

1 Question 9

Below is a table showing the optimal values of the tuning parameters for each algorithm along with their out-of-sample performance:

Algorithm	**Accuracy**	**Optimal Parameters**
Logistic Regression	0.85	penalty = 0.00
Decision Tree	0.87	cost_complexity = 0.00, tree_depth = 15.00, min_n = 10.00
Neural Network	0.85	hidden_units = 9.00, penalty = 0.08
k-Nearest Neighbors	0.84	neighbors = 30.00
Support Vector Machine	0.86	cost = 1.00, rbf_sigma = 0.25

Table 1: Optimal tuning parameters and out-of-sample accuracy for each algorithm

By comparing the performance of the algorithms, we see the Decision Tree achieved the highest out-of-sample accuracy at 87%, after the Decision Tree, Support Vector Machine got 86% accuracy. The Logistic Regression and Neural Network both achieved 85% accuracy, while the k-Nearest Neighbors model had the lowest accuracy at 84%. It means it correctly predicted whether someone was a high earner or not for 84% of the individuals in the test dataset.

The Decision Tree model performed best with a maximum depth of 15 levels and a minimum of 10 observations required in a node before further splitting. It means the tree could have up to 15 sequential decision points from the root to the furthest leaf node. This relatively high depth allows the model to capture complex patterns and relationships in the dataset.

The Support Vector Machine performed well with cost parameter 1.00 and rbf_sigma of 0.25, indicating a balanced approach towards the classification margin and a mid kernel width. 0.25 for kernel width means SVM is considering data points that are relatively close to each other as similar.