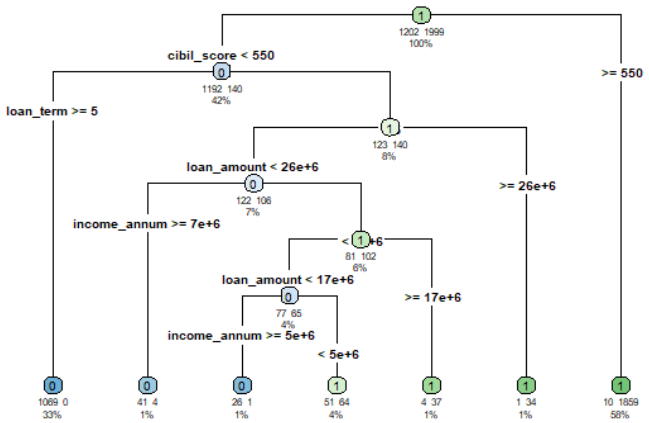
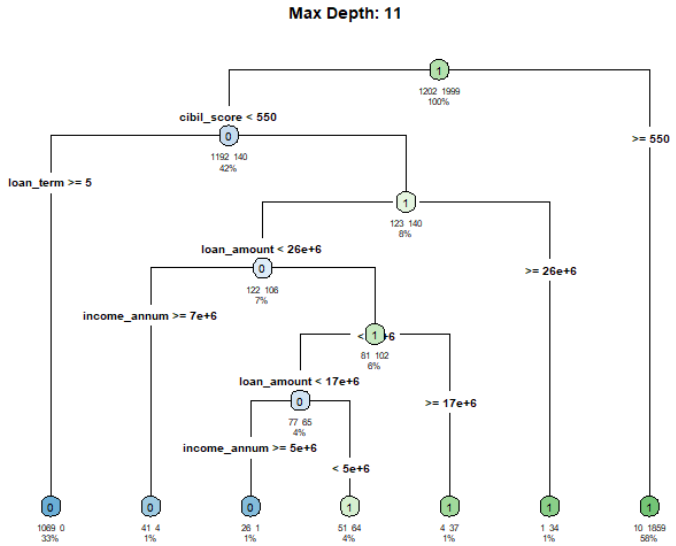
Roberto Jackson Baeza, 1861128
COSC 3337
Sept. 29th 2023

1A) 1. Using all attributes, build a Decision Tree model to predict whether a loan is approved: Train the Decision Tree model using the given maximum depths (3, 7, 11, 15). **8 points**

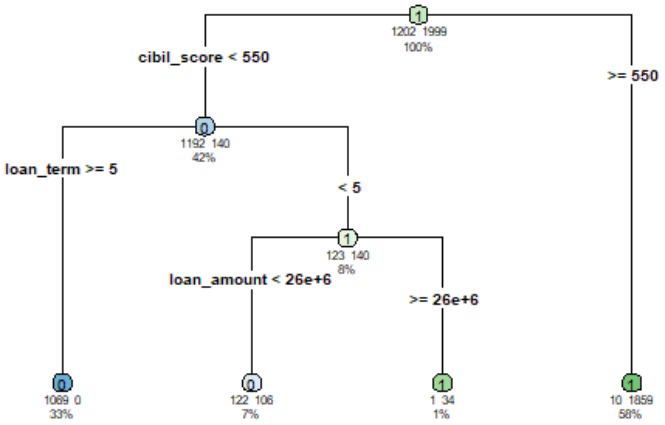
Max Depth: 15



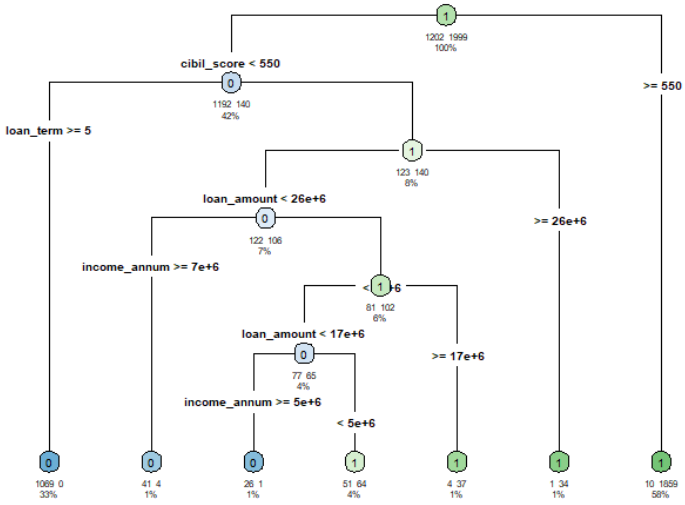
Max Depth: 11



Max Depth: 3



Max Depth: 7



Above are all of the decision trees I generated. Once I was over a max depth of 7 they stopped changing

1B) B. Perform 5-fold cross-validation for each of the 4 max depths and compute accuracy (mean of validation scores), precision and recall. Generate a table, as given below, for the obtained results. **5 points**

Max Depths	Accuracy	Precision	Recall
3	0.9599	0.9668	0.9330
7	0.9672	0.9396	0.9749
11	0.9672	0.9396	0.9749
15	0.9672	0.9396	0.9749

1C) Explain how the tree size/depth affects model performance in the context of overfitting/underfitting.

- Underfitting can lead to various problems as the model does not take into account deeper patterns in the data so the Accuracy, Precision and Recall will all be low as it does not create a good fit or model.
- Overfitting can lead to various problems as the model listens to excessive noise in the data which misleads the model to think that random variation is a pattern. Precision and Recall will be very high on training data but significantly lower on the test data as it does not create a good fit for the model.
- Tree size/depth plays a big part in Underfitting and Overfitting. A small size constraint will give a tree to underfitting as it does not have the space to differentiate patterns in the data. Large size trees will give themself to overfitting as the model will overcorrect to fill the large tree.
- Overall for our model because we used five fold cross validation and chose the best models generated from that process our four models are all pretty good. Also for the three models with a tree depth greater than 3 they all optimized for the same model. So our model does not suffer from over or underfitting

1D) Explain the meaning of the difference in accuracy, precision and recall scores in relation to the task; only if there is a significant difference

- Overall for our data there is a marginal difference between the model for depth of 3 and model for 7, 11, and 15, all our values are very high and indicate that we have a strong predictive model.

- Accuracy: Calculates the overall proportion of correct predictions given by the model. For our models the Max Depth: 3 Model is only marginally worse, at 0.9599 than the models for Max Depth 7, 11, and 15, at 0.9672. Overall all the models are accurate predictors of Loan_Status.
- Precision: Calculates the proportional of all correct true predictions compared to correct true and false true predictions. Here we can see that our Max Depth: 3 Model is better, at 0.9668, than the other three models, at 0.9396. This suggests that the smaller model produced less false positives when compared to the total number of predictions than the other three models. Overall all the models are precise predictors of Loan_Status
- Recall: Calculates the proportion of predicted true outcomes compared to the total number of true outcomes. In this category the Max Depth: 3 Model, at 0.933, is somewhat lower than the other three models, at 0.9749. This means that the Max Depth 3 Model produced more false negatives than the other three models. Overall all the models have a very high recall.

2A) Using all attributes, build an SVM Model to predict whether a loan is approved: Train the SVM model using the given kernel functions (linear, polynomial, sigmoid, sigmoid with different s value).

2B) Perform 5-fold cross-validation for each of the 4 kernel functions and compute accuracy (mean of validation scores), precision and recall. Generate a table, as given below, for the obtained results

Kernel Function	Accuracy	Precision	Recall
Linear	0.9239	0.9475	0.9292
Polynomial	0.6477	0.6529	0.9252
Sigmoid	0.6222	0.6222	1
Sigmoid Coeff	0.6222	0.6222	1

2C) Discuss the impact of different kernels on model performance

- Overall we can see that between the three measures that the Linear kernel is the most accurate and precise. Overall the Linear kernel does very well in all measures. The other three Kernels do moderate in Accuracy and Precision and very well in Recall.
- For Accuracy all of the kernels do at least moderate in the category with all being above 0.6222. Sigmoid and the Sigmoid Coef are tied for the lowest at 0.6222. Polynomial is marginally higher at 0.6477, While the Linear kernel is significantly higher than the others at 0.9239, this means that overall the Linear Kernel produced significantly less false positives and negatives
- For Precision All of the kernels are at least moderate with the lowest being Sigmoid and Sigmoid Coeff tied for 0.6222. Polynomial kernel is marginally higher at 0.6529, Linear

is significantly higher at 0.9475. This shows that the linear kernel produced significantly less false positives than the other models.
- For Recall Sigmoid and Sigmoid Coeff have the highest at 1, while the Linear and Polynomial are both above 0.92 they are also very high. Each of the four kernels produced very few False negatives

2D) Explain the meaning of the difference in accuracy, precision and recall scores in relation to the task.

- Here Accuracy shows the proportion of points that were sorted by the model into the correct classification. In the context of the task that means the model would correctly give a loan status prediction based on the other data points.
- Precision calculates the proportion of points that were correctly predicted to be positive of all the points predicted positive. In the context of the task that means the proportion of points correctly given a 1 loan_status to all the points given a 1 loan_status
- Recall calculus the proportion of points that were correctly predicted to be positive to the correct positives and the false negatives. So in the context of the task this would determine the proportion of points correctly predicted to have a loan_status of 1, to the total number of points that do have a loan_status of 1.

3) Interpret the tables you generated in questions 1B and 2B; compare the performance of the Decision Tree and SVM models. Which model performs better? Why do you think that is the case? What would you recommend to further improve each model's performance?

- I think overall the Decision trees did a better job at predicting overall. Taking an average of the three measures we did, all Decision trees are greater than all the SVMs, and the worst SVM is significantly worse than the worst decision tree.
- The Max depth of 7 decision tree is the best fit for predicting the data overall. I think this based on an average of all three data points it is the largest.
- I think more control over the creation of the models could help improve the model and better understand them. For me all of the Depths greater than 7 created the exact same decision tree which is frustrating. Also for the sigmoid even with a coefficient of 5 it produced no meaningful change in the model overall.