Assignment 6

I chose to make binary predictions (over and under 50k) and the three models I chose are Logistic Regression, Decision Tree, and Random Forest. I chose logistic regression because I wanted to include a model that is more closely related to traditional statistics and less so to machine learning. I chose the decision tree because it is a fundamental model for machine learning beginning, and I wanted to compare a simple probability based model against regression. Lastly, I chose the random forest model because by design it has to work better than a single shot decision tree, I wanted to experience that myself.

The order of F1 scores from highest to lowest is as follows: Random Forest (0.8673), Decision Tree(0.8603), Logical Regression(0.8575). Given that it is the minority class, all models struggled more with the "over 50k group"compared to "below 50k" group. The results of logistic regression and decision tree are quite similar. The decision tree surpasses logistic regression the accuracy of "over 50k" group and also in general F1 value, but the difference is quite small and hidden when rounded decimals. Random forest on the other hand, surpasses both of these models in all aspects although. The F1 scores of the models respect Especially for the "over 50k" group, the precision, recall, and F1 scores of random forest is higher than both models. Random Forests overall F1 macro score is also slightly better than the other two models.

When we look at the false positives, logical regression had the highest false positives with 320 cases, followed by decision tree with 281, and random forest with 245. The decision tree manages to surpass regression in this aspect, but random forest bests both of the models. In terms of false negatives, logical regression seems to have the lowest value, followed by random forest and decision tree, but the overall accuracy of the models does not allow this to make a difference. Ultimately Random Forest has yielded the best results with considerable difference, followed by a decision tree as the second best, and logical regression as the last with small differences with the second.