Model Development Process

I started my EDA process by checking the dimensions of available data, the data types, and the amount of missingness in the data. There were three variables with data missingness, work class, occupation category, and country, that I decided I would impute before moving forward with any modeling.

Next was moving on to explore the continuous variables by using histograms to visualize the distributions. Most values for both capital\_gains and capital\_losses were 0 and the lack of variability makes them unlikely to be good predictors, though I waited to see how they looked when broken down by the target variable. For each continuous variable I plotted boxplots by the target variable, which confirmed my earlier thoughts about capital\_gains and capital\_losses, and they did not vary across the target. The remaining three continuous variables did show differences between the two groups and so I kept them when developing my model.

Chart, waterfall chart

Description automatically generatedWith the categorical variables I found that the variable country had many records sharing the same value and that it had many categories, most of which would have to be small, so I dropped it from further exploration. For the remaining categorical variables, I used bar charts with each variable individually and then broken down by the target to explore whether there were any concerns due to small categories or similar frequencies in the two target groups. Two variables, workclass and race, did not show much difference between the two groups, but the rest of the variables did, including marital status, which I’ve included a chart of below:

Looking at the chart single, divorced, separated, and widowed people are more likely to make under 50k while the majority of those making over 50k are married. Also, this charge shows that there are a number of small categories here, which is why I chose to collapse both this variable and occupation category. Education was also a candidate for collapsing, but after examining the relationship between the predictor variables I chose to instead drop the variable and just keep education\_num as a predictor, since the two are related variables.

With the EDA completed it was time to move on to model development. This started with preparing the data including imputation of missing values, creating collapsed variables, converting the categorical variables to dummy variables, and splitting the data 70/20/10 between train, validation, and test. I chose to go with 6 different potential model algorithms, Logistic Regression, KNeighbors, LinearSVC, DecisionTreeClassifier, AdaBoost, and RandomForest. I fit all of them to the train data and tested them on the validation data to evaluate performance. Reviewing the results of the scoring on the training and validation sets, the KNeighbors, DecisionTreeClassifier and RandomForest all overfit the train set, with train scores up to 0.18 higher than the validation scores. Between the remaining three models both the AdaBoost and LogisticRegression outperformed the LinearSVC by about 0.08. In deciding between the two remaining models both performed similarly enough that I chose LogisticRegression as the final model since it is well understood by most audiences and it is easy to interpret the coefficients, which might provide additional value. The final step was to score the LogisticRegression model’s performance on the test data, which was in line with both train and validation, at 0.8295.