Analysis of the dataset was able to produce a model (KNN@10) that correctly predicts if a new instance is over\_50k by using 8 of the original 13 variables performing with a precision of .852 on test data. Education\_num, capital\_gain, and marital\_status were by far the most influential variables in the analysis. In the attached visualization, you can see three charts demonstrating the relationships between these three variables and the target class. Capital gains greater than five thousand account for large amounts of the over\_50k population. As an individual’s education level increases, the divergence between over 50K and under 50k increases. The attached chart also compares distributions of education levels across marital statuses for the target class. At the bottom, the performance of the KNN classifier is examined against capital gains and education level. Most errors were made in the vector space where capital gains was zero and education level was high (> Associates degree).

As part of my methodology, I extracted the data from the original database, assessing each variable’s summary statistics to ensure it did not require any cleaning or correction. In preparing the data for model building, I replaced all categorical variables with numeric dummies and replaced missing values with dummies (there were no nulls). I created the training (.75) and test sets (.25) from the cleaned data. Validation/model evaluation was accomplished through 10-fold cross-validation to maximize the amount of data available for training. I then calculated information gain for all variables with respect to the target class within the training set to assist in feature pruning during model building.

Model selection was completed over three rounds of evaluation. In the first round, four classifiers were evaluated for baseline results against all variables: RandomForest, LogisticRegression, KNN instance, and DecisionTree. The second round eliminated relationship, education\_level as features due to conflicts or duplication with other variables and eliminated country, race, and workclass as features due to low information gain. The removal of these features reduced the noise in the dataset and improved performance for three out of four models at p<.05. KNN was selected as the final model. While the Decision Tree classifier saw a higher average precision, I did not want tune this model any further as I was concerned I would overfit by increasing depth (I had set max depth to 10). The KNN model was then tuned for optimal value at k from 5-30. The optimized model was KNN@10 with an average precision of

.848 on the training set. When assessed against the test set, the model correctly predicted 0.852 of instances, indicating that the model could be generalized to assess new instances.