**Adult Data Analytics Report**

This dataset contains information about the annual incomes of people form 42 different countries and it was extracted from the census bureau database found at <http://www.census.gov/ftp/pub/DES/www/welcome.html> and was download from <https://archive.ics.uci.edu/ml/index.php>. The goal of doing this project is to determine whether a person makes over 50k a year or not. To analyze this data, python and a few different libraries were used, such as NumPy, pandas, seaborn etc.

The data has train and test sets. There are some question marks, “?”, in the datasets. sep=',\s', na\_values=["?"] was used to replace the cells with “?” when the data was read. Pandas\_profiling was used to get EDA. The train dataset has 15 variables, 24 duplicate rows and 32561 observations. For train data, workclass has 1836 missing values, occupation has 1843 missing values, and native-country has 583 missing values. The test dataset has 15 variables, 5 duplicate rows and 16282 observations. For test data, workclass has 964 missing values, occupation has 967 missing values, and native-country has 275 missing values. For both datasets, capital-gain and capital-loss have more than 91% zeros. The duplication and missing values were dropped.

A picture containing text, receipt

Description automatically generated

The above table shows the correlation matrix between numeric attributes and income. We can see that income has a weak uphill (positive) linear relationship

with age, education-num and hours-per-week. The following is the correlation plot.

A picture containing square

Description automatically generated

The following image shows one of the examples of EDA. We can see that for the same amount of education-num, male people make more than female people.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

The above plot indicates that for the same education-num, people older than 29 years old have higher income than younger people. That means that the median aged people earning more than $50k is significantly greater than the median aged people earning less than $50K. People have higher income when they get longer education. But income doesn’t necessary increase with education-num when education-num is above 15.

Chart, bar chart

Description automatically generated

The above plot shows the relationship between workclass and income for every age period. We can see that people who are self-employed with incorporated business had the highest income for every age period. People who worked in federal government had the second highest income when they are older than 29 years old.

To build the model, column “fnlwqt” was droppe as it was not important to predict salary. get\_dummies was used to convert all the object data types into numeric ones. Random Forest Classifier, Decision Tree Classifier and Gradient Boosting Classifier were used to build the models. Gradient Boosting Classifier had the highest accuracy, 0.86. Therefore, the model built with Gradient Boosting Classifier was used to predict the income.

Text

Description automatically generated

The following the histogram of final prediction. We can see that most people make less than 50k a year.

Chart, histogram

Description automatically generated