Capstone Final Report

Adult Census Income

By Cem Karaman

The purpose of this project is not only to predict adult income level, if it is greater or less than $50,000 annually, but also to compare variables in which of them is easier to predict, given the same dataset.

The dataset used in this project is available at Kaggle.com and it is one of the widely used datasets. The data is clean with relatively few missing data points. Even after eliminating some of the missing values, we are still left with plenty of data to make meaningful comparisons.

We uploaded the data file to Google Drive and used Google Colab to run our code. Google provides easy access to your Drive within Colab which makes working with files a breeze. We started with basic exploratory data analysis (EDA); data types, counts, missing values, histograms and more.

The majority of the participants are high school graduates with college graduates coming next. At least two thirds of the participants are male, which makes this dataset a male dominated group. Very most of the participants are in a relationship. Again, most of the participants are white, which very much depends on where this survey was made. The data is US census data of a certain region but unfortunately this information is not available to us.

One of the interesting aspects of the data was the availability of survey participants’ country of origin. A fact derived from data was US born citizens’ higher level of income, compared to non-US born citizens. Although there was a difference in income levels, this difference is not significant.

Another analysis necessary for any dataset is to look for correlations. We used a heatmap to visualize correlations between variables. Although this was color coded, a need for a numerical analysis was understood. We ranked the relationships from highest to lowest and we did not observe any correlations that would make us worry about collinearity problems. No need for principal component analysis.

After EDA and cleaning the data, we started preparing our data for supervised machine learning techniques. The first step was to encode data entries to numerical values which is required by sci-kit learn modules. We used LabelEncoder() for this task.

Throughout the project we will use three different machine learning techniques, decision tree, random forest and XGBoost. For each of these techniques, we chose four different target variables: income, education level, marital status and relationship status. For each of the three techniques, we tried to predict four different target variables which makes twelve models in total.

After assigning the target variable, the rest of the variables became explanatory variables or independent variables. To repeat, the two aspects of our project are:

1. We compared different machine learning techniques to choose the best one of them in predicting our target variables,
2. We compared four different target variables in their predictability under each of the learning techniques.

Before going for a grid search of hyperparameters, we tried a decision tree model with default parameter values. This model’s prediction accuracy became the benchmark to beat for the rest of the models. For all of our models, we split the data into training and testing sets. For grid search we used the GridSearchCV() module of Python. We tried the most commonly used hyperparameters in grid search. The grid search parameters produced significantly better results than models with default parameters.

Unsurprisingly, every consecutive model improved the model before it, with XGBoost producing the best results. Besides this, predicting marriage was easier than prediction income, education level or relationship status. In the order of easiness, marriage was first, income second, relationship third and education level with a big margin below fourth. Moreover, this ranking did not change across different supervised learning techniques.

I learned a lot from this project. I had the opportunity to apply many Python, EDA and machine learning techniques with this project. Many times you think you know the subject, until it comes the time to apply what you supposedly know. I sharpened my skills in grid search, matplotlib and seaborn graphs. I am looking forward to working on other projects to improve my skills and make a meaningful contribution to the understanding of data.