Project 13: Income prediction

Name:	**Pritam Anurag**
Registration No./Roll No.:	19228
Institute/University Name:	IISER Bhopal
Program/Stream:	EES
Problem Release date:	February 02, 2022
Date of Submission: 24 / 04 / 2022	

1 Introduction

The dataset consists of information about different people working in the US. For each individual, we have information about their age, education level, marital status, occupation, relationship, race, gender, working hours per week, native country, capital gain/loss and the type of organisation that they are employed in. Using this information, we have to predict whether the income of the individual is above or below 50K per year. This is a classification problem where we are provided with a training dataset with class labels and our task is to predict the class labels of the test dataset using the algorithm giving the best accuracy.

Our training dataset consists of 43957 data points. On exploring the training data, we saw that 3 of the columns(workclass , occupation and native-country) have some missing values while the other columns don't contain any missing values. 24 percent of the data in the training dataset have income more than 50K while the rest 76 percent have an income of less than 50K.

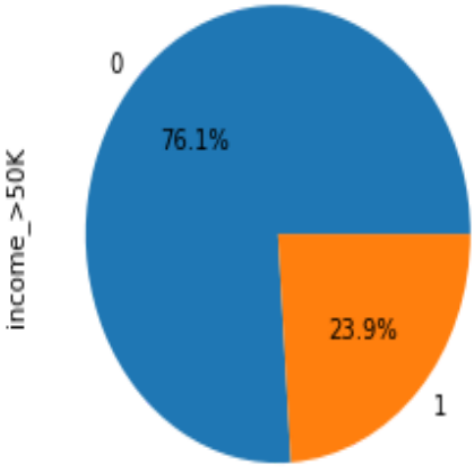

Figure 1: Data Distribution

Then we look at the data distribution of the different columns. The age of the individuals is distributed between 17 to 90 years with a mean of 38.6 years. The workclass of the individuals is distributed across 8 categories, with the highest number of data points belonging to the private sector. We observe that Never-Worked and Without-pay workclass are very less. So we merged the two workclass in another category of workclass and named it as others. The data contains null values represented as NaN. It is observed that wherever Workclass is NaN, occupation is also NaN There are

2498 missing values which we need to handle. We see that the education column and the educational-num columns are equivalent. More advanced education is given a higher number. As it is easier to work with numerical data, we ignore the education column and only consider the educational-num column. The marital status column has 7 categories. The highest number of people have marital status as Married-civ-spouse, whereas lowest values are Married-AF-spouse. The data has no missing values. There are missing values in occupation column, and have workclass also missing Number of missing values is equal to 2506, which are slightly higher than Workclass The distribution among top occupations are quite similar in numbers (5000) The people who have workclass as other have occupation value Null as an addition in the dataset There are 14 categories in Occupation column. Race column has 5 categories and no missing values. The highest number of people are white and the rest of the numbers are significantly less. So we merge the significantly lesser category Asian-Pac-Islander'and 'Amer-Indian-Eskimo to 'other' category.The ratio of male is to female in the training dataset is close to 2:1. Under the hours per week column, people with 40 hours per week have the highest count (20513). Majority of the people have hours per week with 20 - 60, rest of the values are quite less. The relationship column contains 6 categories with the highest number of datapoints belonging to the husband category. There are 122 categories belonging to the capital gain column with the maximum people with no capital gain. There are 40330 rows where capital gain = 0. There are 3627 rows where capital gain has a non-zero value. There are 97 categories belonging to the capital loss column with the maximum people with no capital loss. There are 41884 rows where capital gain = 0. There are 2073 rows where capital loss has a non-zero value. Under the native country category, we see that other than US, we have very less number of data points belonging to other countries. So we combine all the other countries into one category i.e others. Then we conducted bivariate and multivariate analysis. Using that we found out that, the majority of the people in the US are whites, other races are significantly low. Black still have a considerable number as compared to rest. We also see that males have a higher chance of having income more than 50K than women. There is also a somewhat positive correlation between education and the chances of earning more than 50K. Husband and wife have a high chance of getting income more than 50k Rest of the relations have very less chances to be more than 50k.

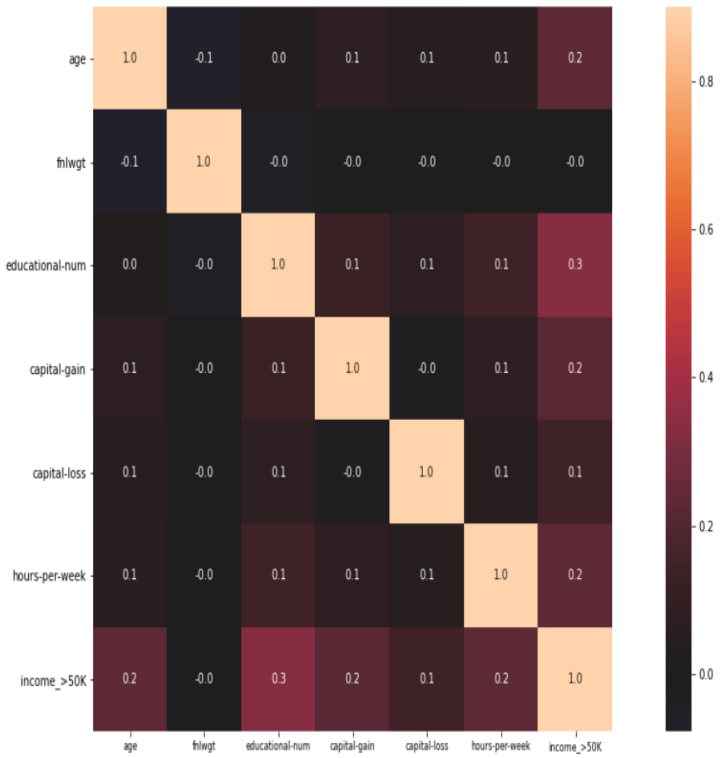

Figure 2: Overview of Corpus

Income has highest correlation with Education-num (30 percent) Income has correlation with

capital gain, hours per week and Age (20percent, 20 percent and 20 percent respectively)

The null values of workclass are all changed to private as this had the highest number of values. Then the data was merged in a meaningful manner. Then we check for outliers. We see that, by dropping outliers, we will be dropping around 15 percent of the rows which have income greater than 50k, hence we choose not to drop the outliers data.

2 Methods

First we perform one hot encoding on the data to convert the categorical data into numerical data. Then on the encoded data, we run the following models:
Decision tree, Guassian Naive Bayes, Logistic Regression, KNN, Random Forest, Support Veector Classifier, Gradient Boosting, AdaBoost.
After running these models on the unscaled encoded data, we scale the data and run the above models on the scaled data. After this, feature selection was done on the scaled data. Then the best parameters were chosen and then again the models were run using them. After that we do hyperparameter tuning on the feature selected data. And then, as our dataset was skewed, we resampled the data. And then we ran the base models on the resampled data. And after that we did hyperparameter tuning on the resampled dataset, and in this case, the model which provided us with the best accuracy was chosen as the model to predict the class labels of the testing dataset.

After this, before the class labels of the test data was predicted, it was also scaled ,it was encoded and feature selection was also conducted on the model.

3 Evaluation Criteria and End Result

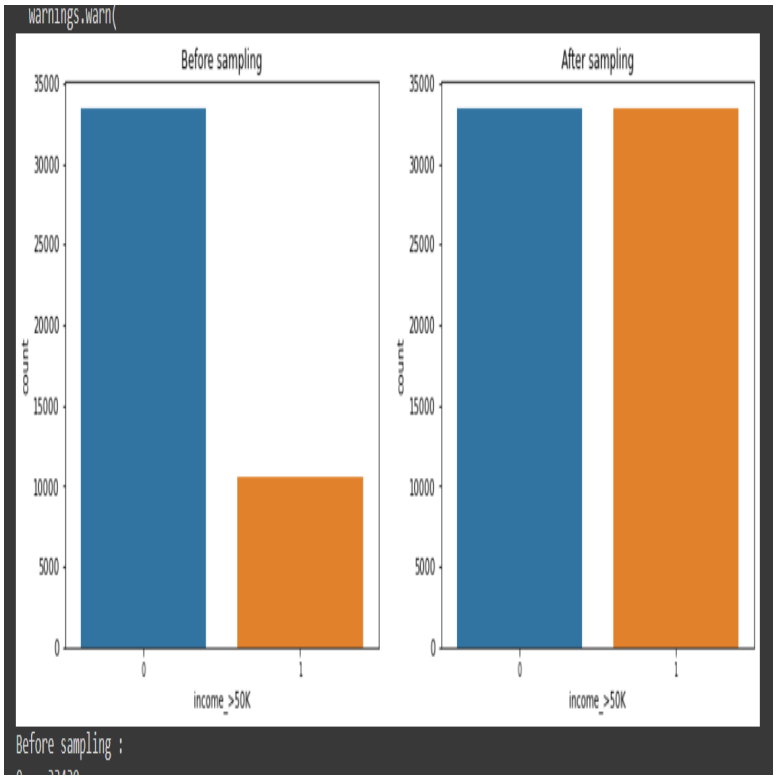

Figure 3: Overview of Corpus

The above figure shows the classlabel distribution after resampling. Resampling of data leads to increase in accuracy of the prediction.

In the base model, the random forest had the best accuracy and f-measure. And this continued to be the best model after all the steps. One interesting thing that we observed was that, after

resampling, the macro averaged f-measure increased, while the accuracy decreased slightly.In the base model the accuracy was coming 92 percent. In the end, after doing all the things with the data, and hyperparameter tuning we see that random forest algorithm had an accuracy of 94 percent, which was higher than all the other algorithms used. Hence this was used to find the class labels of the test dataset.