**Adult Census Income**

**Problem Statement**

The Adult Census Income data was extracted from the 1994 Census bureau database by Ronny Kohavi and BarryBecker (Data Mining and Visualization, Silicon Graphics). The data set contains the features about the people who earns more than $50K or less than $50K a year. Our task is to predict whether a person earns over $50K a year using various classification algorithms. Exploratory data analysis is done on the dataset to achieve insights and the pre-processing pipeline is done to get the data ready for the training.70% of the data is used for training purpose and 30% for the testing purposes. Four Classification models are trained and their performances are compared with various performance metrics like accuracy score, cross validation score and the Receiver operating characteristic curve.

**INTRODUCTION**

The Adult Census Income data was extracted from the 1994 Census bureau database. The purpose of this study is to get the insight about the income of the persons from 1994 in a particular reign and can infer about the income inequality if there's any from the data set and to successfully predict whether a person make $50K a year or not. This Adult Census data set contains the features about the people who earns more than $50K or less than $50K a year. Some features are age of the person, their nationality and education. In this project the Exploratory data analysis is done and various insights are concluded from Exploratory Data Analysis. The feature engineering is done to prepare the data for the Machine Learning models, then the Classification models are trained and the performances are compared with various performance metrics like accuracy score, cross validation score and the Receiver operating characteristic curve. The classification algorithms that are used are Logistic regression, Random forest Classifier, DecisionTreeClassifier, and the Support vector Classifier. The Best performing model is finalized for the predictions.

The Adult Census data set is of shape (32561,15) i.e. It has 15 attributes and 32,561 rows.

The dataset provides 14 input variables and 1 target variable that are a mixture of ordinal, categorical and numerical data types. Following are the variables is our dataset:

1. Age.
2. Work class.
3. Final Weight.
4. Education.
5. Education Number of Years.
6. Marital-status.
7. Occupation.
8. Relationship.
9. Race.
10. Sex.
11. Capital-gain.
12. Capital-loss.
13. Hours-per-week.
14. Native-country.

Target Variable:

1. Income

**Attributes:**

Attributes age, hours-per-week are already self-explanatory.

education-number describes number of years of education.

Capital-gain is income gain from investment sources other than salary/wages.

Capital loss is the income loss from investment sources other than salary/wages.

The education and education number are related as the education number is the numerical conversion of the education column which consist of categorical values.

The relationship features describe the relationship status about the person if they are Unmarried or married in various category.

The race feature is used to tell the race of the person like black, white, Asian etc.

The column Age has values between 16 and 100.

The fnlwgt (final weight) feature is the weight on the Current Population.

Income is the target Variable that is to be predicted that has two categorical values ">50K" and "<=50K".

**Exploratory Data Analysis**

Now, we will do exploratory data analysis to get the insight about the data and how target variable depends on various attributes.

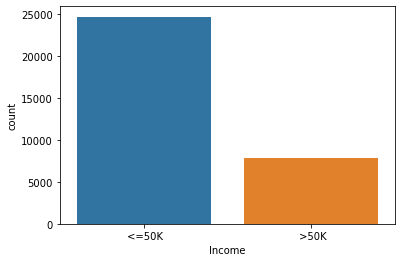
First, we are analyzing our target variable i.e. Income.

<=50K 24719

>50K 7841

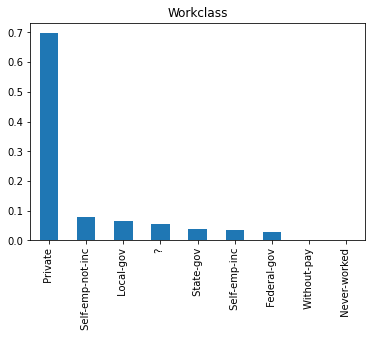
We get to know that in our dataset, the count of people whose income is less than 50K is higher than the people who are earning more than 50K.

We can visualize it in below mentioned figure.



* On X-axis our target variable i.e. Income is shown.
* On Y-axis No of counts is shown.

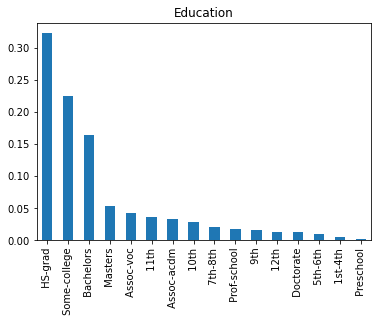
**UNIVARIATE ANALYSIS OF INDEPENDENT VARIABLES:**

1. **WORKCLASS**:

The figure.2 shows the bar graph of Attribute “Workclass”. It can be seen that around 70% of people are working in Private sector.

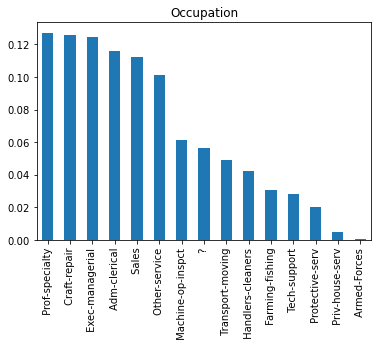
One more important point to note here is that we have some unkown values under “Workclass” showing as “?”. We will deal with this while preprocessing our data later.

Figure.2

 Figure.3

1. **EDUCATION:**

The figure.3 shows the bar graph of attribute “Education”. We can easily visualize from the figure that around 33% of person are having degree in HS-graduation.

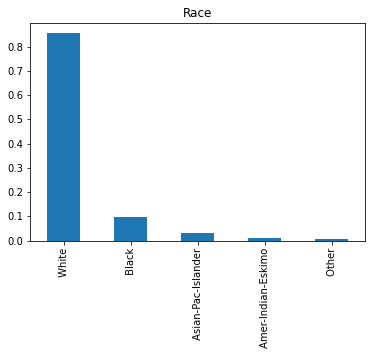
Figure.4

1. **OCCUPATION:**

The figure.4 shows that most of the people are working as Prof-speciality, carft-repair, Adm-clerical and in sales.

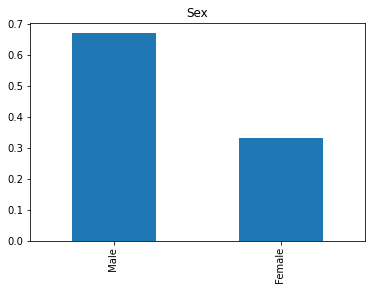
People least engaged in Armed forces.

One more important point to note here is that in “Occupation” also we have some unkown values showing as “?”. We will deal with this while preprocessing our data later.

Figure.5

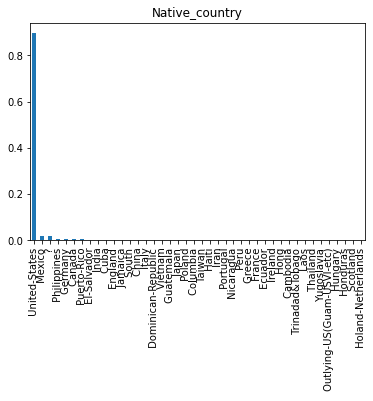
1. **Race:**

The figure.5 shows that around 85% of people in our dataset belonged to white race. Remaining 15% belonged to black, Asian, and other.

 Figure.6

1. **Sex:**

The figure.6 shows that around 70% are males in our dataset and remaining 30% are females.

Figure.7

1. **Native Country:**

This figure shows that 85% people are from United States.

### Analysis of Categorical Independent Variable with Target variable:

### Analysis of “Workclass” WRT “Income”:

### 

Figure.8

Above figure.8 shows that 70% people are working in private sector and most of them are earning less than 50K.

### Analysis of “Education” WRT “Income”:

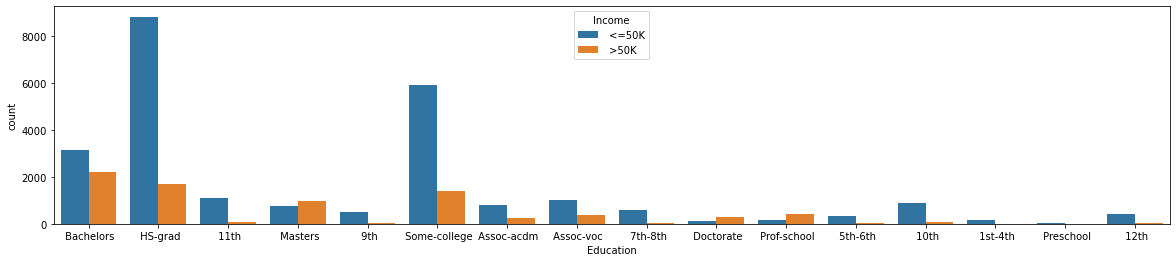


Figure.9

Above figure.9 shows that 33% of people are having HS-gard degree and most of the are earning less than 50K.

### Analysis of “Occupation” WRT “Income”:

### 

Figure.10

Above figure.10, shows relation ship between Occupation and Income. People engaging in each occupation are mostly earning less than 50K. In every occupation type the group of people earning less than 50K is dominating.

### Analysis of “Race” WRT “Income”:

### 

Figure.11

Above figure.11, shows that 85% people belongs to White race and most of them are earning less than 50K.

### Analysis of “sex” WRT “Income”:

### Figure.12 shows that 70% are males and most of them are earning less than 50K.

Figure.12

### Analysis of “Native Country” WRT “Income”:

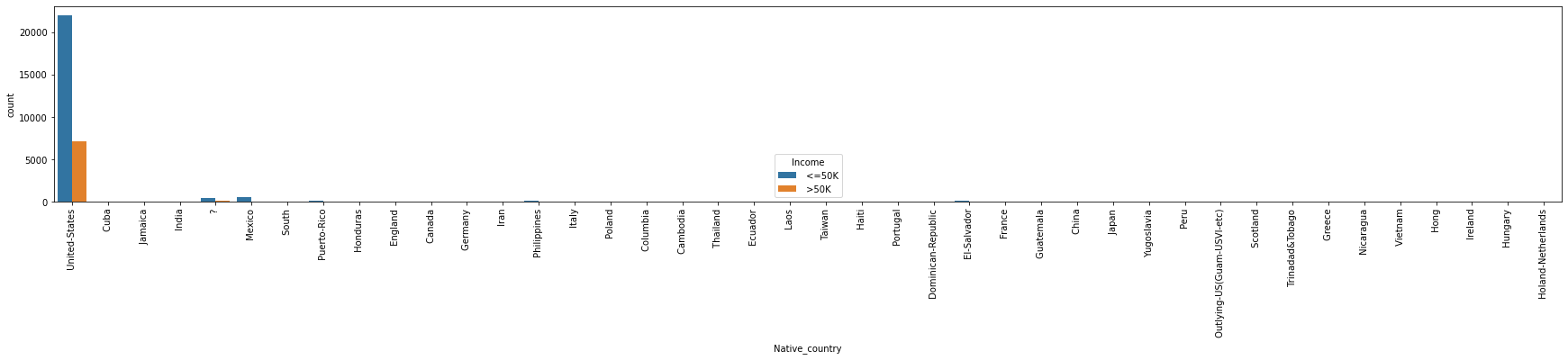
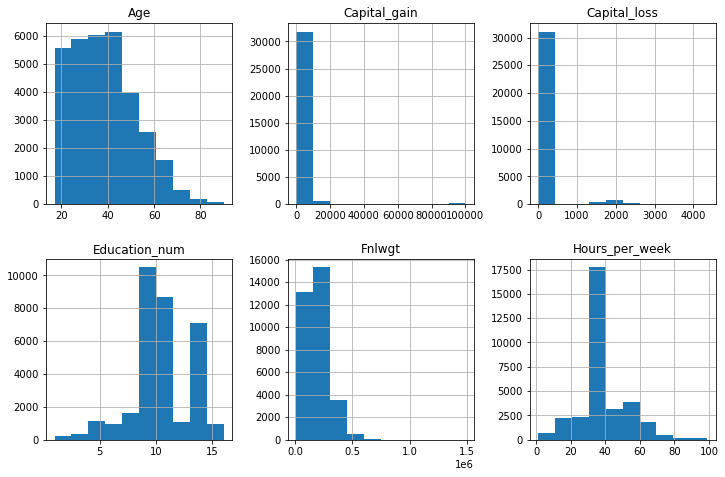


Figure.13

Figure.13 shows that 85% people belongs to United States and most of them are earning less than 50K.

**ANALYSIS OF CONTINOUS VARAIBLES:**

 Figure.14 shows the attributes which are continuous in nature. Age, capital\_gain, capital\_loss, education\_number, Fnwlgt, Hours\_per\_week are given in numeric form.

We can clearly analyse that age, capital gain, capital loss, fnlwgt is rightly skewed.

Hours per week & education\_num are normally distributed.

Figure.14

**DATA PREPROCESSING**

**Data preprocessing** is very essential step in any **data** mining process. It directly impacts the predictions of the model. If data is unclean, have missing vales, missing attributes or contains outliers, if skewness is present, then all these factors degrade the quality of our results and our predictions will be biased.

First, we will check for missing or NaN values through heat-map:

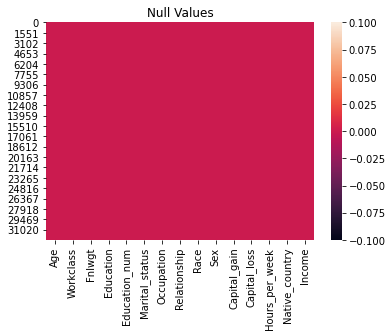
 We can clearly visualize that no NaN value is present in our dataset.

Figure.15

Now, we will handle unknown vales which is showing as “?” in attributes Workclass, Native country & occupation. First, we replaced “?” by NaN values and then fill these NaN values with Mode. Workclass, Native country & occupation all three attributes are present in object form or categorical data. That’s why we replaced NaN vales with mode. The **mode** is the value that appears most frequently in that particular attribute. We already analysed that mode of “Occupation”, “Native country”, “Workclass” is Prof\_speciality, Unites States, Private respectively.

## **Drop irrelevant columns:**

#### fnlwgt is not related to our target variable income. That’s why we will remove this column before building the model.

#### The column education is just a string representation of the column education-num. So, we decided to drop the education column.

**LABEL ENCODING:**

Label Encoding means to convert all the categorical values into numerical values so as to convert it into the machine-readable form.

First, we have imported the label encoder and then a looped Label Encoder is used to convert all the labels in every column to the numeric form for the training and testing purpose. The attributes that has categorical values are education, occupation, marital status, race, sex, native country, relationship and income. The target variables i.e. income had only two categories >50K & <50K. That’s why our target variable has only two values 1 and 0.

**Correlation matrix heatmap:**

Checking correlations is very important to analyse data. A heatmap has been plotted to check the correlation between the attributes, if there is positive or negative relationship. This is one of the methods to decide which attributes affect the target variable the most.

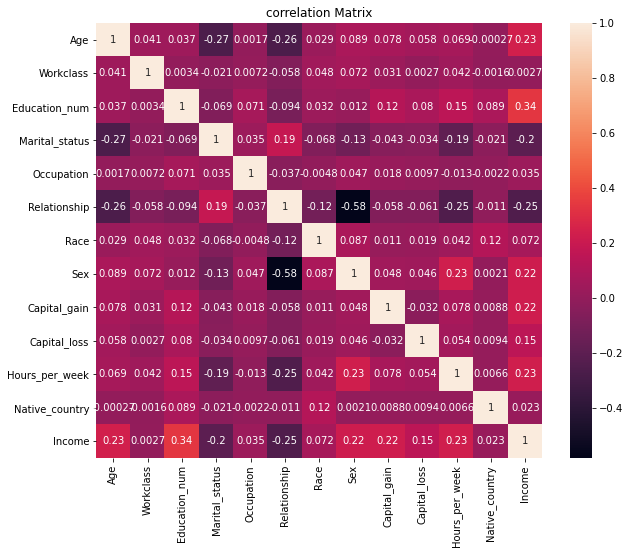


Figure.16

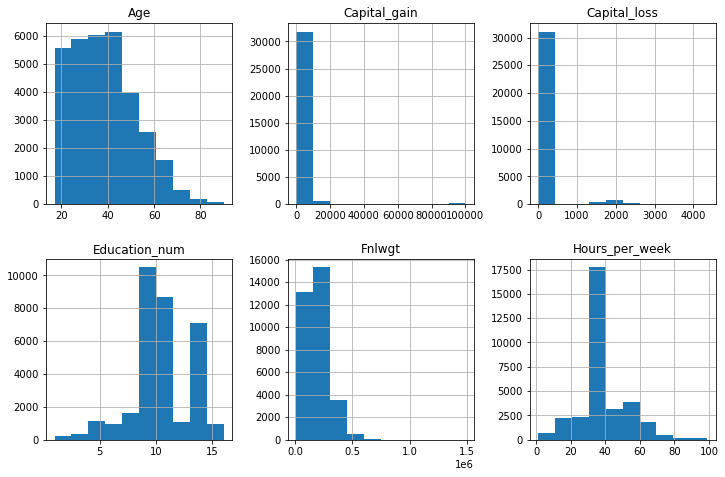
The figure.16 is the heatmap of the correlation matrix of the attributes. It can be inferred that Income per week has a fair positive correlation with the education number and age. The capital gain and the capital loss have a negative relationship.

**CHECKING OUTLOIERS:**

An outlier means an observation that falls outside the overall pattern or we can say an abnormal distance from other values in a random sample from a population.

We have outliers present in various attributes named age, capital gain, capital loss & education number. From scipy.stats we imported Z-score and drop all the rows in which threshold value is greater than 3. But by dropping these rows we lost around 14% of our data and the results came from our predictions will be biased. Therefore, we will not drop these outlier values as these values are important for our predictions

**TREATING SKEWNESS:**



If the skewness is between -0.5 and 0.5 then the data is fairly symmetrical and represent normal distribution. If the skewness is between 0.5 and 1 or -1 and -0.5 then the data is moderately skewed. If the skewness is less than -1 or greater than 1then the data is highly skewed.

As we earlier analysed that skewness is present in age, capital gain, capital loss. So we will treat this with power transformation

Figure.14

**SPLITTING DATASET**

The shape of the dataset after dropping of the irrelevant columns is (32560, 12). We split the dataset where 70% is used for training the model and 30% for testing the model. Hence out of 32,560 data entries, 22,792 are used for training and 9,768 are used for testing the model.

**FINDING BEST RANDOM STATE**

Our model best performs at random state 22 and we are achieving 0.831 accuracy score.

Four Classification Algorithms are used.

1. Logistic Regression
2. RandomForestClassifer
3. DecisionTreeClassifier
4. SupportVectorClassifier

**Logistic Regression**

We are achieving 82% accuracy with Logistic Regression.

**DecisionTreeClassifier**

We are achieving 81% accuracy with Decision Tree Classifier.

**SupportVectorClassifier**

We are achieving 84% accuracy with SVC

**RandomForestClassifer**

We are achieving 85% accuracy with RandomForestClassifier.

**Cross Validation Score:**

Imported cross validation score to check the over-fitting and under fitting in our predictions.

**from** **sklearn.model\_selection** **import** cross\_val\_score

#### We proceed with RandomForestClassifier as it is giving highest accuracy\_score and there is minimum difference between accuracy\_score & cross\_validation\_score

We get best accuracy\_score from RandomForestClassifier.

**TUNNING WITH BEST PARAMETERS:**

Imported GridSearchCV from sklearn.model\_selection and find out the best parameters of RandomForestClassifier which performed best on our model.

Following are the best parameters for our model:

{'criterion': 'entropy', 'max\_features': 'log2'}

**ROC (Receiver Operating Characteristic) Curve**

Receiver Operating Characteristic curve or ROC curve represents a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the TPR against the FPR at various threshold settings. The true-positive rate (TPR) is also known as recall, sensitivity, or probability of detection in machine learning. The false-positive rate (FPR) is also known as the probability of false alarm or fall-out.

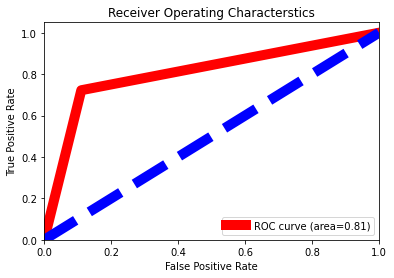


Figure.17:

The area under the curve is 81%.

**SAVE THE MODEL**

We finally save our best model by importing pickle. The use of pickle is widespread as they allow us to easily transfer data from one server or system to another and then store it in a file or database.

**CONCLUSION**

The main aim for this project is to successfully predict the income of the person whether a particular person is earning less than or more than $50K from dataset using various classification algorithms and compare their performances. The Exploratory Data Analysis is done on the data set to get insight about the data and get the correlations of the attributes present. The preprocessing pipeline is done and four Classification Algorithms are trained. The performances of the models are compared on the basis of accuracy score, cross validation score and ROC curve. The RandomForestClassifier comes out to be the best performing algorithm above all other models with an accuracy of 85.29% and over all generalizing well.

**DOWNLOAD JUPTYER NOTEBOOK**

Click on the below link to find my juypter notebook in GitHub:

<https://github.com/riturani2403/Data-Trained-practice-projects/blob/main/Adult_Census_Income.ipynb>