#### CENSUS INCOME PROJECT

In this project we will be predicting whether a person can makes over $50k a year.The problem statement is as below.

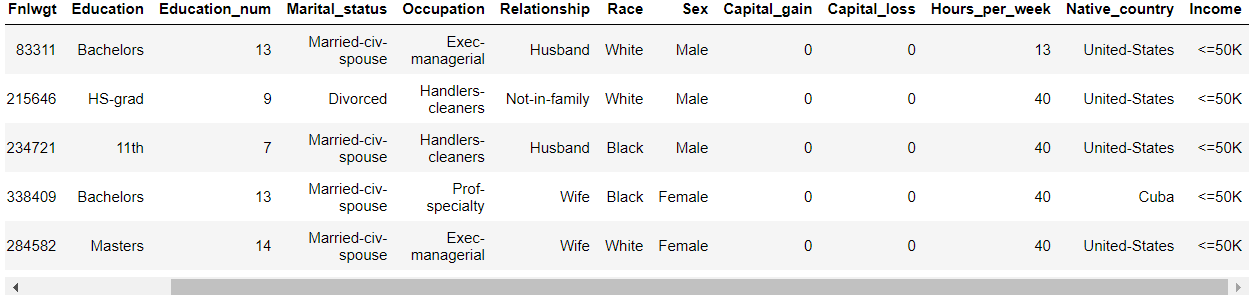
#### Problem statement:

This data was extracted from the 1994 Census bureau database by Ronny kohavi and Barry Becker (Data mining and Visualization, Silicon Graphics).A set of reasonably clean records was extracted using the following conditions:((AGE>16)&&(AGI>100)&&(AFNLWGT>1)&&(HRSWK>0)).The prediction task is to determine whether a person makes over $50k a year.

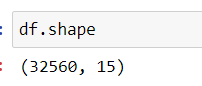
#### Data Analysis & EDA:

In this Data Analysis part we will import our data set which is a csv file as shown in the below output and this step starts with importing necessary libraries.

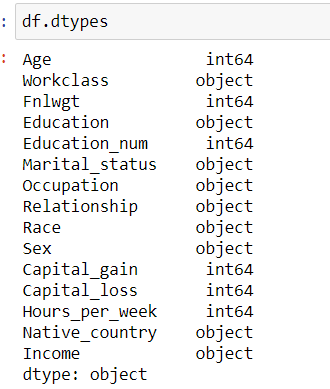
After importing necessary libraries and our data set we will be able to see that our Target variable which is **INCOME**  is numerical and we need to predict whether a person can make a income of greater than 50k.As the Target variable is numeric it is a regression problem.To find all the columns in the data set we use the df.head() function which gives all the variable in our data set.



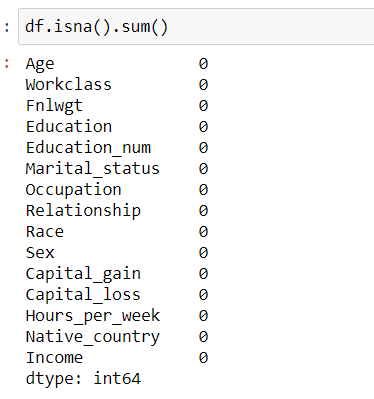
When we use the .head function it will return the First five rows our data set as shown above.The next step is to find out the shape of our data set using .shape() function.So when we use .shape function on our data set the output will be as shown below.



So our data set has 32560 rows and 15 columns including our target variable.Now let us check the data type of each column by using df.dtypes function which returns the below output.



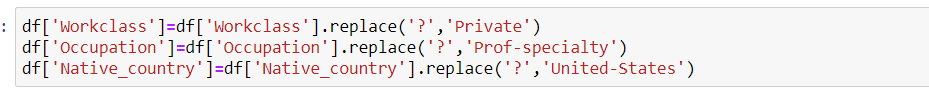
As shown in the above output the our target variable is in Categorical format which has numerical values.Now let us check the null values in our data set.



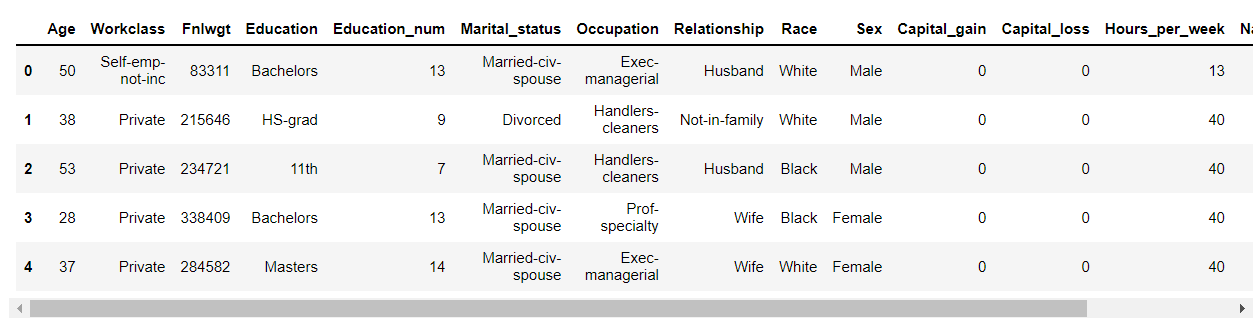
The above output shows there are no null values but we need to check whether we have any special characters.For this we use value\_Counts() which returns the count of each variable in the column as shown in the below.



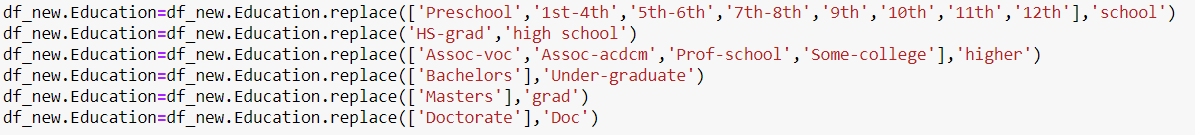
Similarly we will find the special characters in other columns and try to replace them with most frequently occurring values in those columns as shown below.

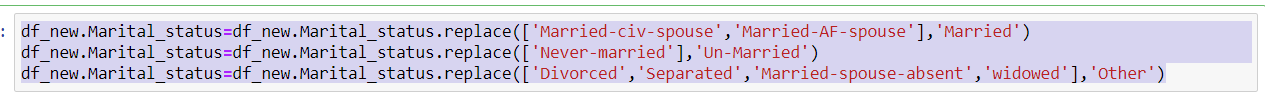


So after replacing the special characters the below is the output.

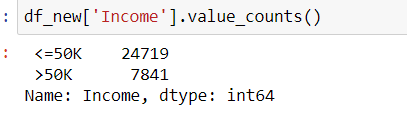


Now we need to split the multiple values in each column in to less content as shown below

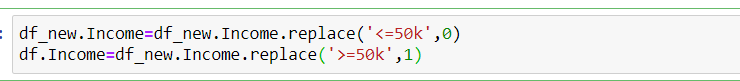


As shown in the above output we split-ed the classes in the columns having multiple classes in to less values.Similarly applying the same technique for remaining columns 

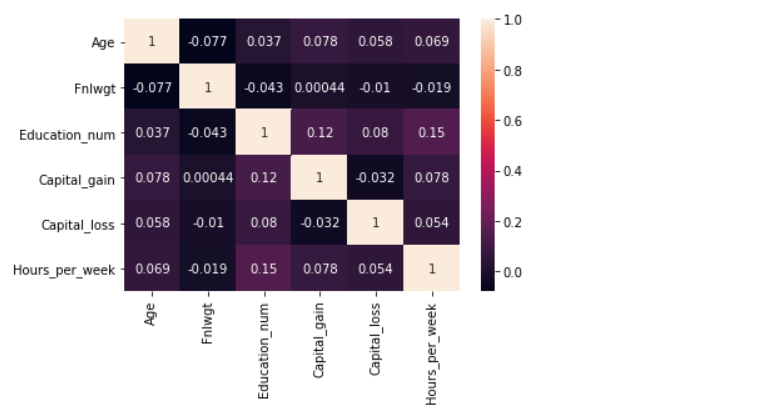
So after converting each variable in the each column in to lesser we will check the counts of variables in our target variable.



So our target variable has class imbalance issue as it has 24719 values of <=50k and 7841 values of >50k.we need to resolve the class imbalance issue. As it has multiple values we will convert them in to Zeros and One’s as shown below.

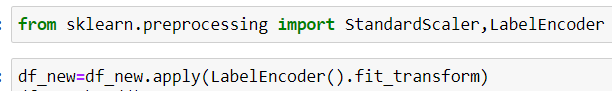


After replacing the Target variable in to 0’s and 1’s we will find the correlation between each variable by using a heatmap as shown below.

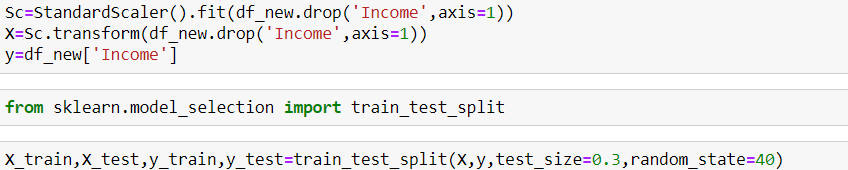
The above Output shows the co-relation between all the numerical values .From the all the Above EDA steps we tried to find out the the relation between all the variables and remove the un necessary columns in our data set and tried to identify the type of problem in our data set.

#### Preprocessing pipeline:

In this step we will try to preprocess our data before building a machine learning model.The steps involved are Encoding techniques and Feature selection and creating Train test split.Before preprocessing the data we need to convert all the categorical columns in to numerical as machine learning models do not understand the Categorical values.For our data set let us consider Using label encoding to convert all the Categorical columns in to numerical as shown below.

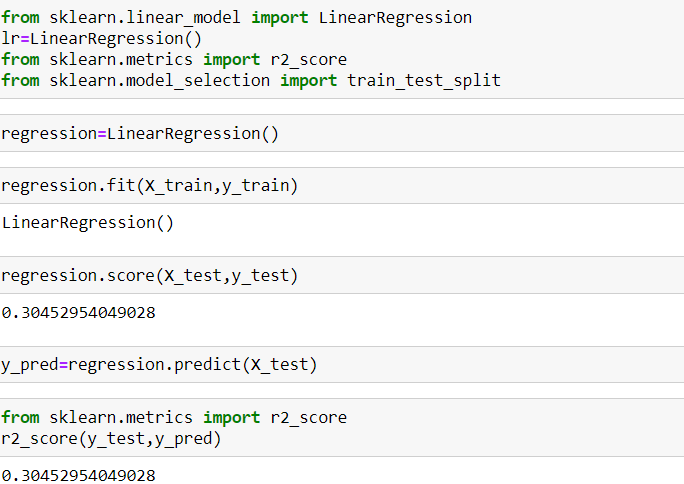


After applying the Encoding technique we will use Standard scaler to scale our data as shown below and from this we will create X and Y variables and train test split.



#### Building machine learning models:

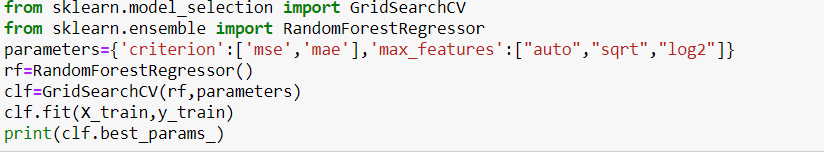
In this step after the pre processing we will try to build a model based on R2 score and Cross validation score of each regression algorithms.The model having less variance between the r2 score and Cross validation score is considered as best model.The R2 score and Cross validation Scores are calculated as shown below.



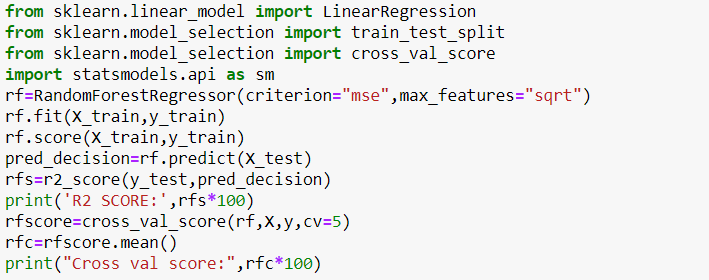
After calculating the R2 score we will apply regularization techniques like Lasso CV and Ridge CV to find the best alpha value as shown below.



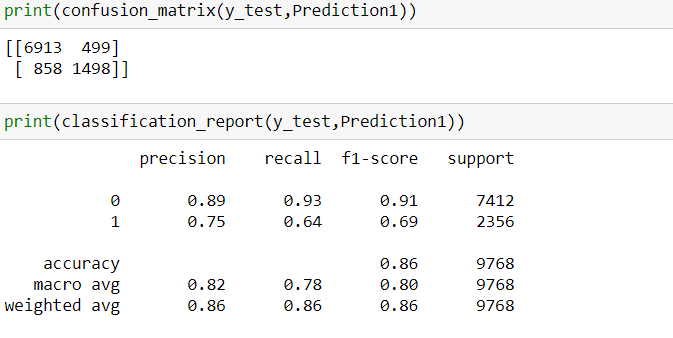


After applying the lasso CV and Ridge CV we will use ensemble techniques to find the best parameters in our data set using different Regression Algorithms.For time being let us use Random forest regressor to find the Best parameters as shown below.

After finding the best parameters we will use them to find the R2 score and Cross val score of our data set as shown below.



From this our model gives R2 score as 46% and Cross val score of 45% which has less variance so we consider it as best model and we will print a confusion matrix and Classification report for the best model in order to predict our target variable which is INCOME which is shown below.



From the classification report we are able to see that model gives 86% accuracy.After printing the classification report we need to calculate the precision and recall for both the classes which are >=50k and <50k which is shown below.

As the above output shows our model gives 89% precision when calculating the <=50k value and 75% precision when calculating the other class.

#### Conclusion:

Therefore from all the above explanation we are able to predict the Income variable by using random forest Regressor for time being.Similarly we can use different algorithms for calculating the precision and recall of the model.