***CENSUS INCOME PROJECT***

**ABSTRACT:**

In this Census Income Project, we are going to predict whether the people makes over $50K or below $50K with collection records of People residing in various federal districts / states of the USA. It can be used to know people’s standard of living in USA according to their age, race, native, education etc., And the detailed analysis, Machine Learning Model predictions done can be used for specific Business Requirements, Challenges and Improvements.

**MACHINE LEARNING AND DATA SCIENCE FOR BUSINESS:**

Machine learning is a branch of [artificial intelligence (AI)](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn from experience, make predictions and gradually improving its accuracy. It is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data science will increase, requires to assist in the identification of the most relevant business questions and subsequently the data to answer them. Following are the ways Data science can add value to Business :

* Empowering management and officers to make better decision
* Directing actions based on trends—which in turn help to define goals
* Challenging the staff to adopt best practices and focus on issues that matter
* Identifying opportunities
* Decision making with quantifiable, data-driven evidence
* Testing these decisions
* Identification and refining of target audiences

**BUSINESS OBJECTIVE:**

**“Prediction task is to determine whether a person makes over $50K a year”**

Data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). CPS (Current Population Survey) is done on monthly basis by Population division in USA collecting records of People residing in various federal districts / states.

A set of reasonably clean records was extracted using the following conditions:

* + (AAGE>16)
  + (AGI>100)
  + (AFNLWGT>1)
  + (HRSWK>0)

**Data overview:**

Description of fnlwgt (final weight)

Weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian non-institutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls. These are:

1. A single cell estimate of the population 16+ for each state.
2. Controls for Hispanic Origin by age and sex.
3. Controls by Race, age, and sex.

**Data Ratifying:**

All three sets of controls are used in weighting program & "rake" through them 6 times, so that by the end we come back to all the controls we used.

The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population.

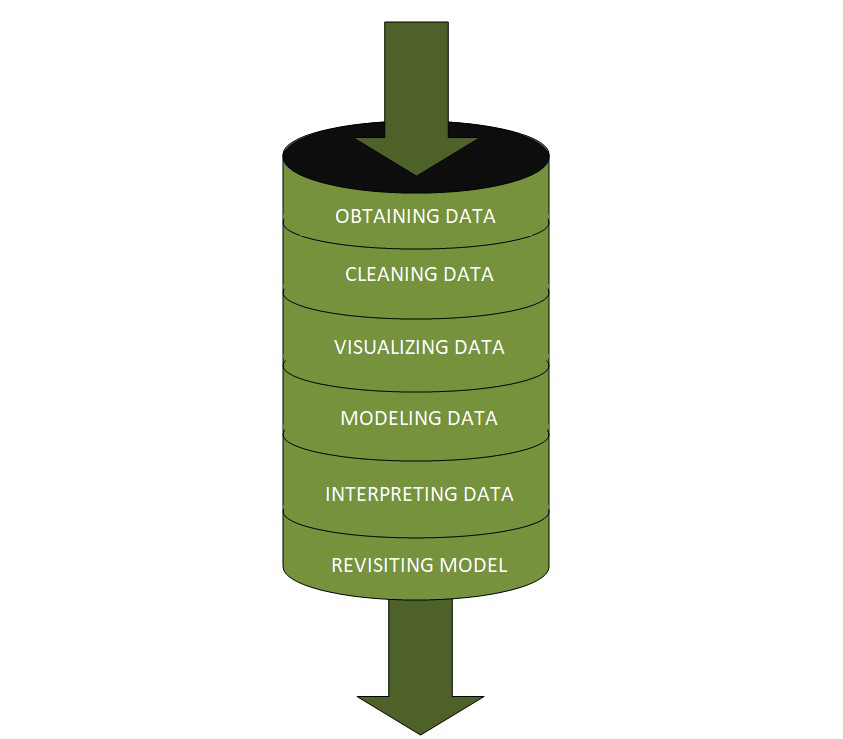
People with similar demographic characteristics should have similar weights.

Point to Remember:

CPS sample is a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.

**DATASCIENCE PIPELINE:**

The data science pipeline is a collection of connected tasks that aims at delivering an insightful data science product or service to the business organization. The responsibilities include collecting, cleaning, exploring, modeling, interpreting the data, and other processes of the launching of the product. This final product can be used for to achieve Business Goals.



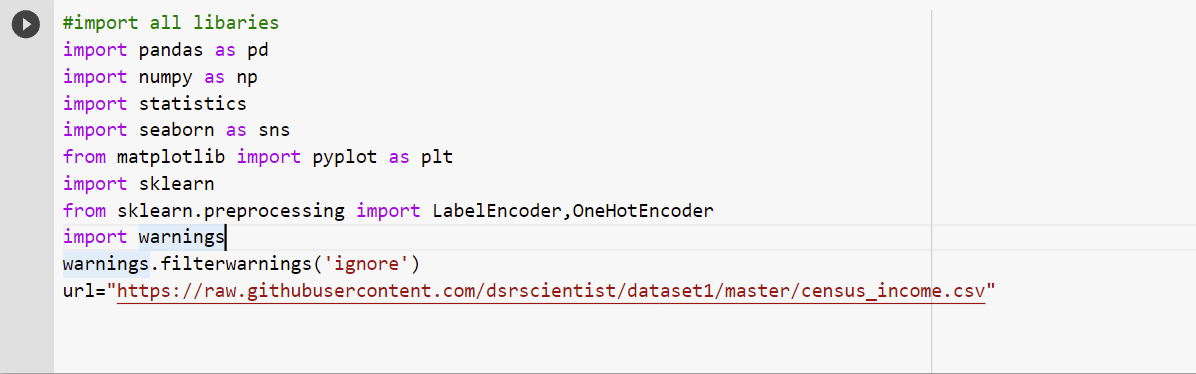
**Software Tools, Libraries and Packages Used:**

Software: Jupyter Notebook (Anaconda 3)

Language: Python

Libraries:

1. Pandas
2. Numpy
3. Matplotlib
4. Seaborn
5. Sklean
6. Scipy
7. Statsmodels
8. Pip-Package install Manager
9. Imblearn etc.,



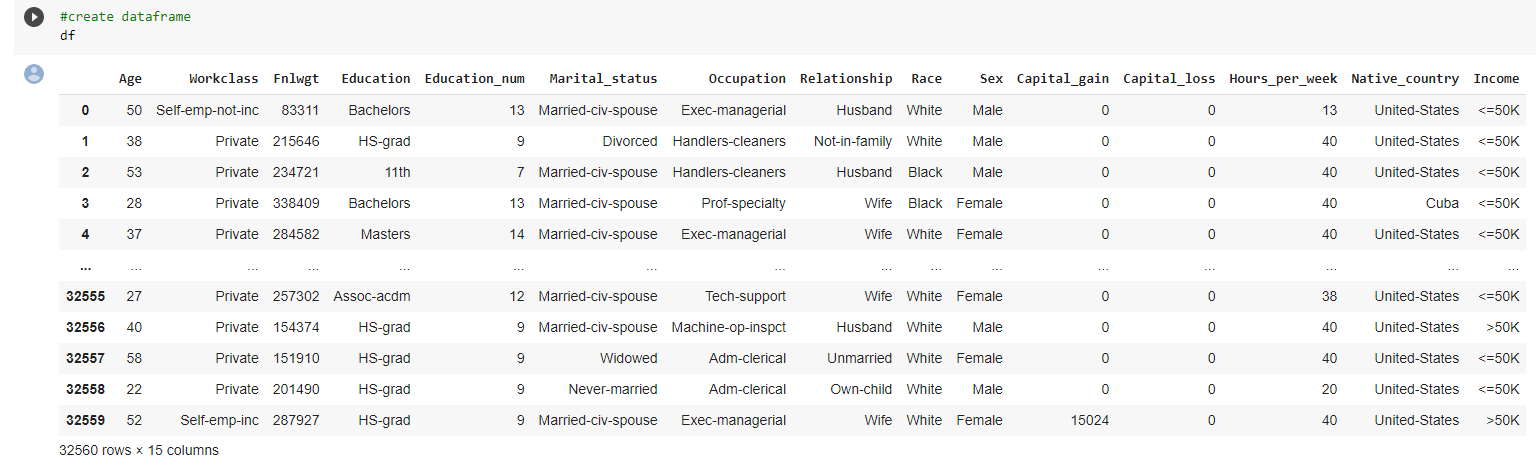
|  |  |  |
| --- | --- | --- |
| **Category** | **Tool** | **Function** |
| Data loading and analysis | Import pandas as pd | Pandas is a Python library that is used for faster data analysis, data cleaning and data pre-processing. Pandas is built on top of numpy. So, numpy gets some superpower with pandas. It offers data structures and operations for manipulating numerical tables and time series. |
| Import numpy as np | NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python |
| Data visualization | Import matplotlib.pyplot as plt | Matplotlib is a plotting library used for data visualization. |
| Import seaborn as sns | Seaborn is also a plotting library. It is more advanced than matplotlib but works with matplotlib |
| Scikit Learn Preprocessing Libraries | Sklearn.preprocessing | Package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators. In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate. |
| Sklearn.preprocessing import LabelEncoder | Label Encoding in Python can be implemented using the Sklearn Library. Sklearn furnishes a very effective method for encoding the categories of categorical features into numeric values. Label encoder encodes labels with credit between 0 and n-1 classes where n is the number of diverse labels. |
| Import statistics | Import statsmodels.api as sm | From scipy import stats This module provides functions for calculating mathematical statistics of numeric (Real-valued) data. This library provides a number of common functions and types useful in statistics. It focus on high performance, numerical robustness, and use of good algorithms |

And all other libraries required will be discussed in further proceedings below.

**DATA ACQUISITION:**

<https://raw.githubusercontent.com/dsrscientist/dataset1/master/census_income.csv>

**Data Type Classification:**



Age, Fnlwgt, Capital\_gain, Capital\_loss, Hours\_per\_week are all the columns which has integer data type and it is categorical ordinal in nature.

Education\_num is integer data type and it is categorical nominal in nature.

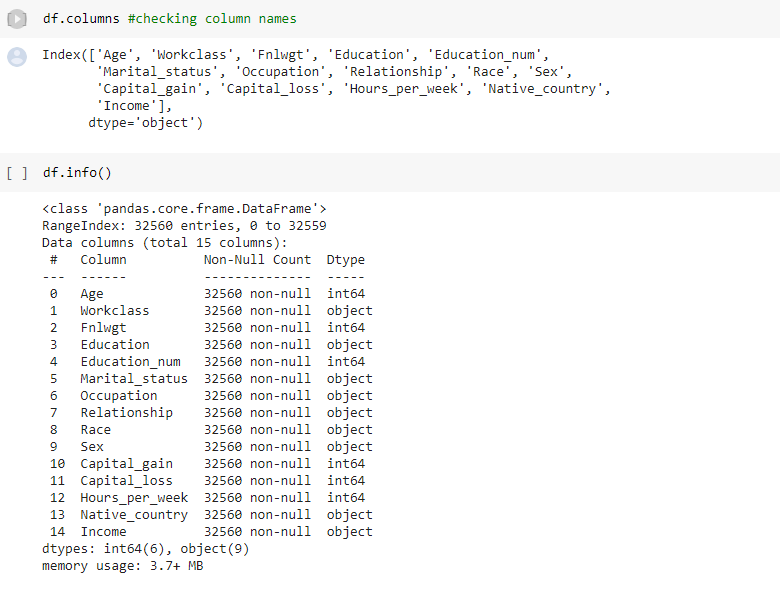
Workclass, Education, Marital\_status, Occupation, Relationship, Race, Native\_country are all the columns which has string (object)data type and it is categorical nominal in nature.

Methodology:

We are going to predict the target column Income which has string data type and it has value either >50k or <=50k. Its like yes or no answer (Only two categories).

It is Similar to a “Binary Classification problem”.

The data types and missing value counts of the features or columns are given below:



**Exploratory Data Analysis:**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions

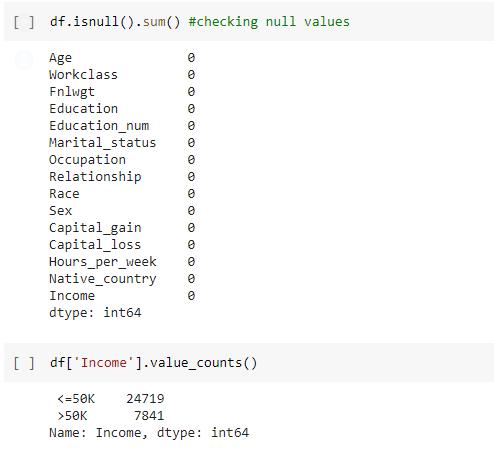
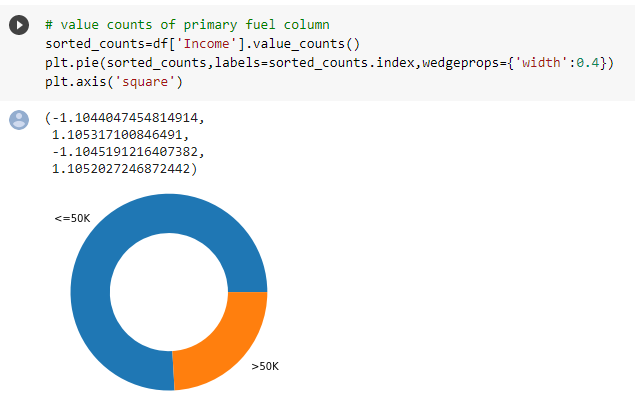
**TYPES OF EXPLORATORY DATA ANALYSIS:**

* Univariate Non-graphical
* Multivariate Non-graphical
* Univariate graphical
* Multivariate graphical

Multi-Dimensional Comparison Done across Each Variable:

Univariate Analysis (Graphical & Non-Graphical Analysis):

High Level Data analysis shows that people getting <=50000 are higher in counts than people earning more than >50000 salary

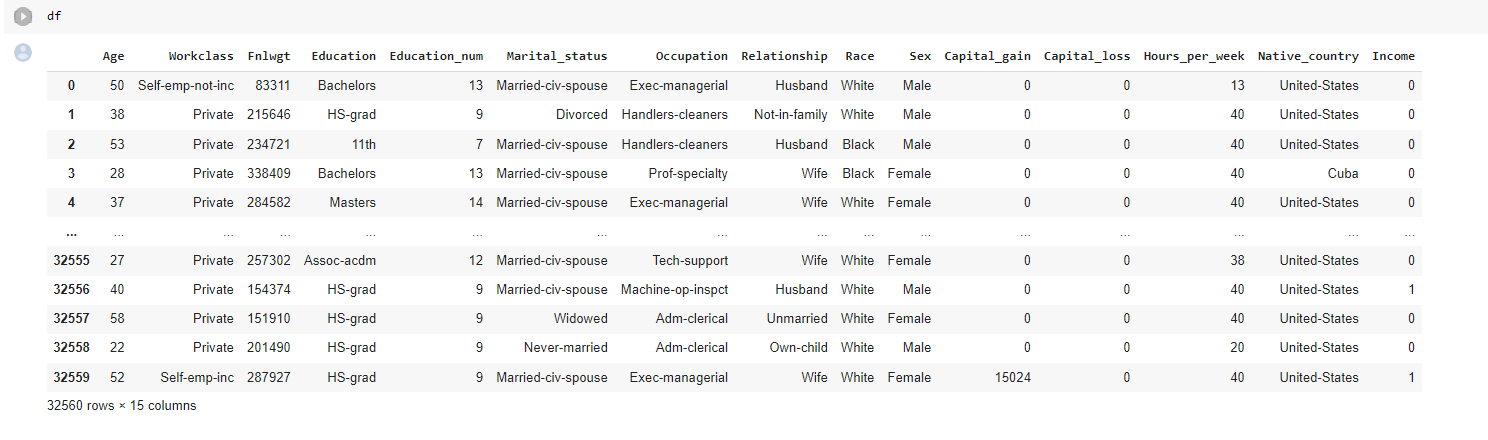
 

**Null Values & Standardization:**

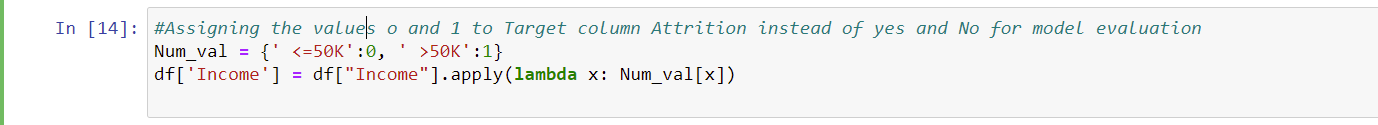
In order to compare / use the data for analysis, it is vital to keep the dataset uniform without missing values, Perform the (?) empty values with most frequent values.



Columns are now cleaned and it doesn’t have empty (?) values,



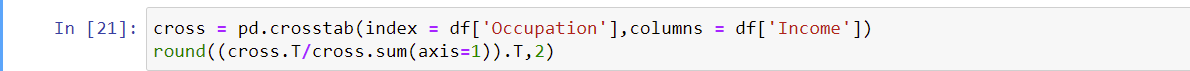
Since ML algorithms understands only numeric data. Replace the >50K with 1 and <=50K with 0 in target values.

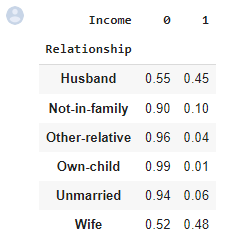
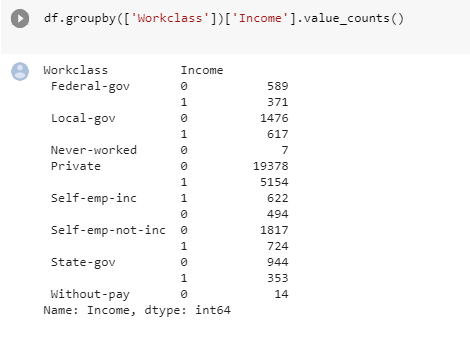


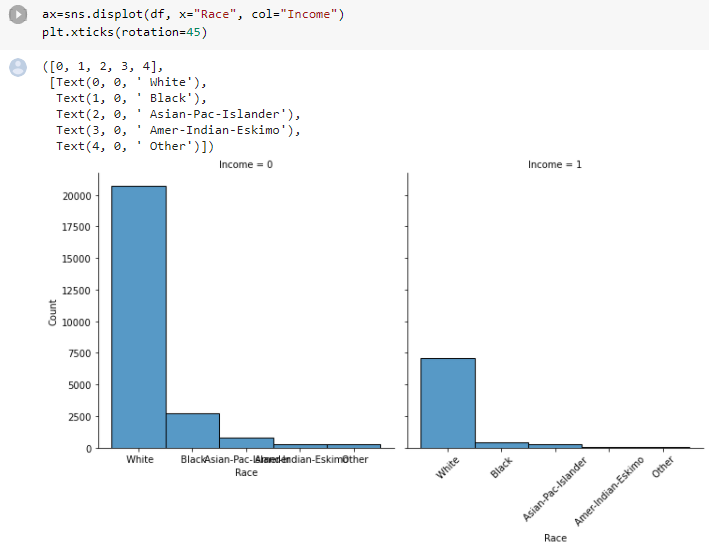
**Bi-variate Analysis:**

Assumption: Assign people earning more than 50k as “1” & Less than / equal to 50k earning as “0”

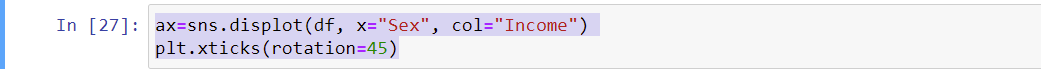
1. Occupation: People whose occupation is Exec-managerial has higher chance of getting income > 50k
2. Relationship: Husband and wife relation having higher chance of getting income >50k
3. Work class: Private working people are getting income>50k mostly

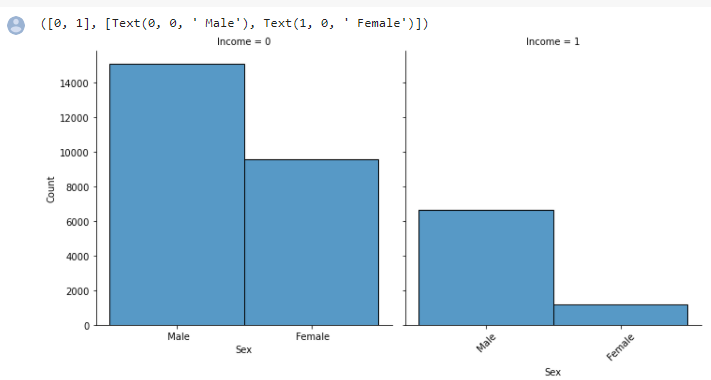




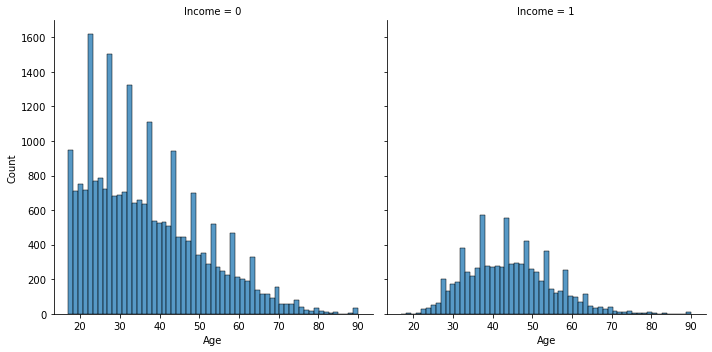
1. Comparison by Race: White people who earns more than 50k are higher in count than all other races

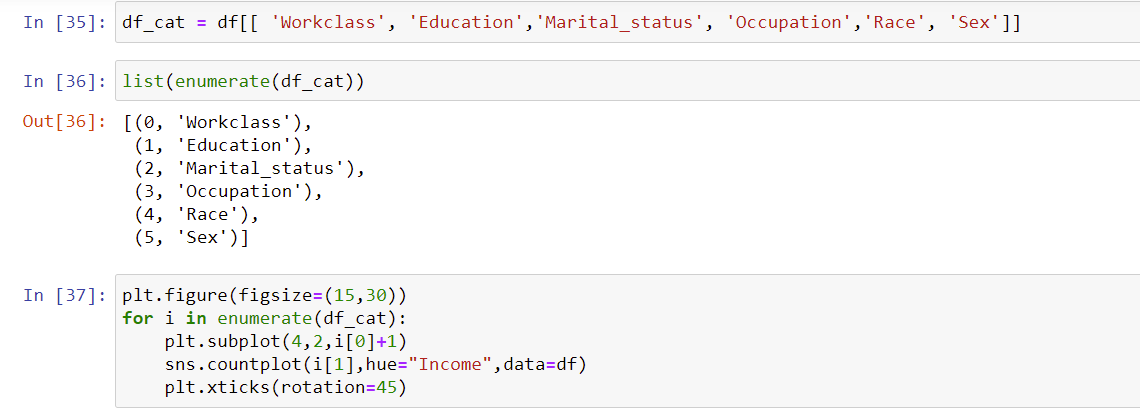


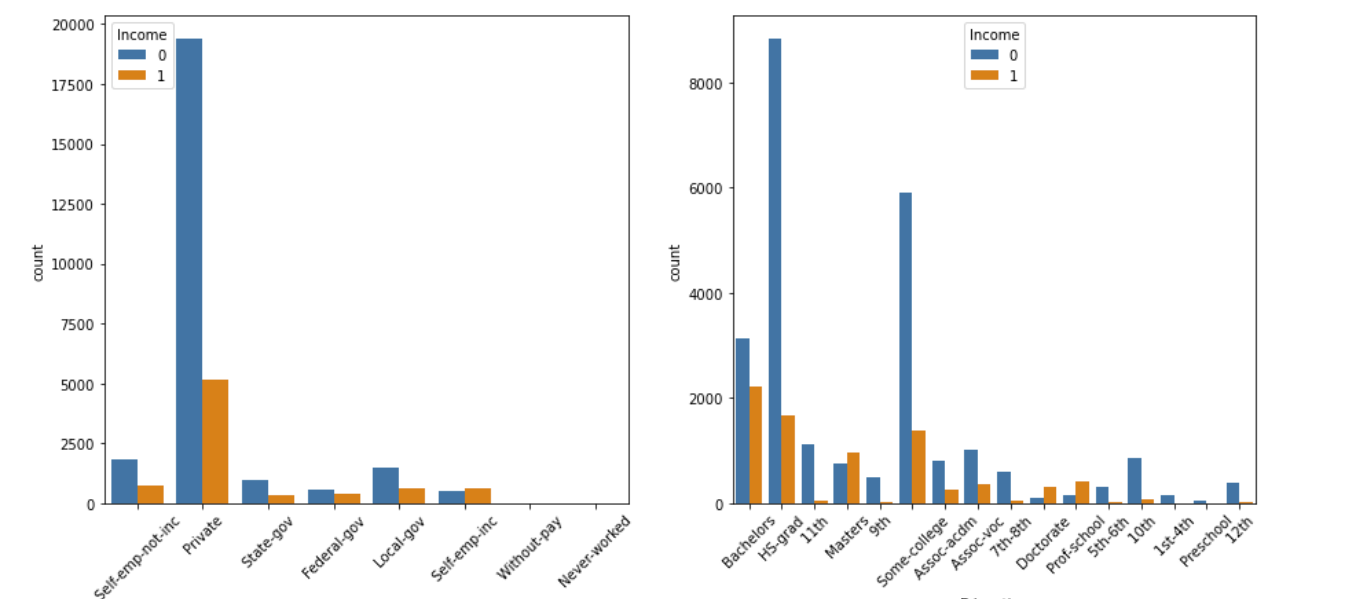


* Gender / Sex: Comparing in Salary less than or equal (or) higher than 50K. No of males count is higher in both categories.
* Age Comparison: Shows higher no of people across all ages earns <=50k only (shown below)

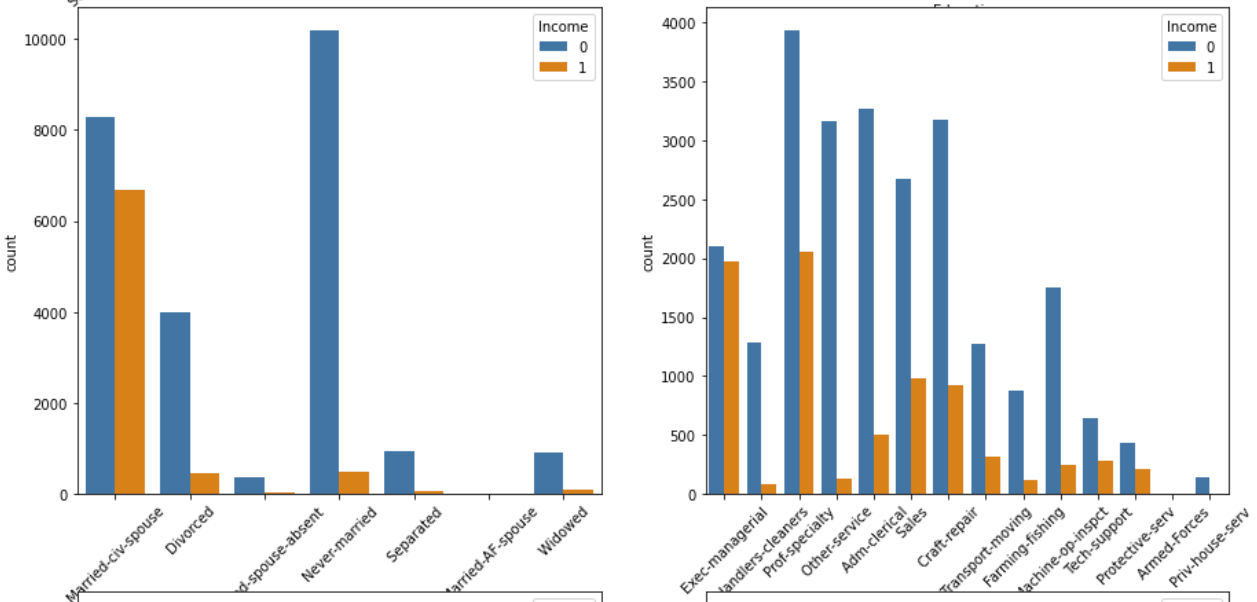




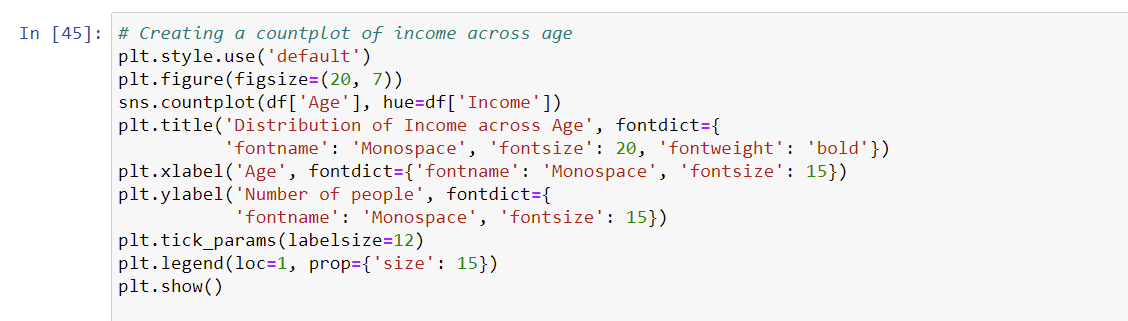


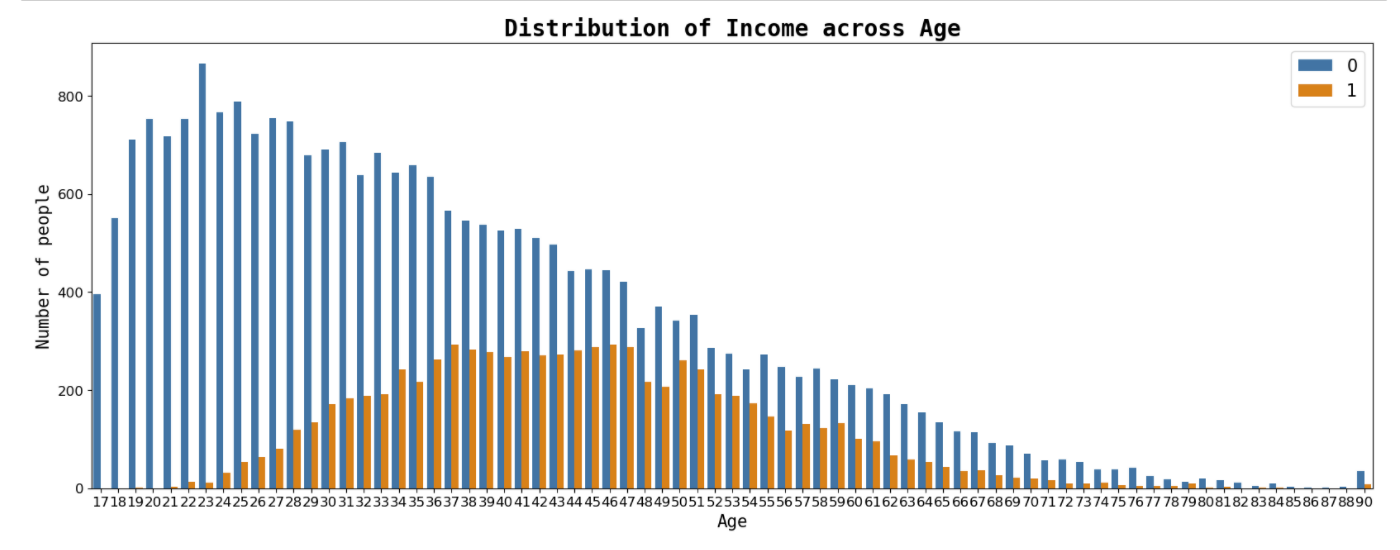


* Bachelor Degree and Private Employees earns more than >50K when compared in overall.



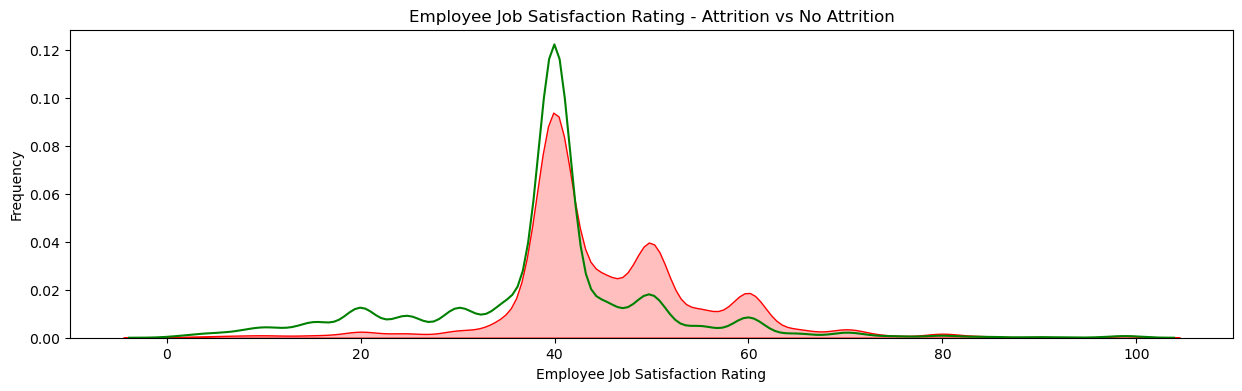
* Married people and prof people earns more than >50K





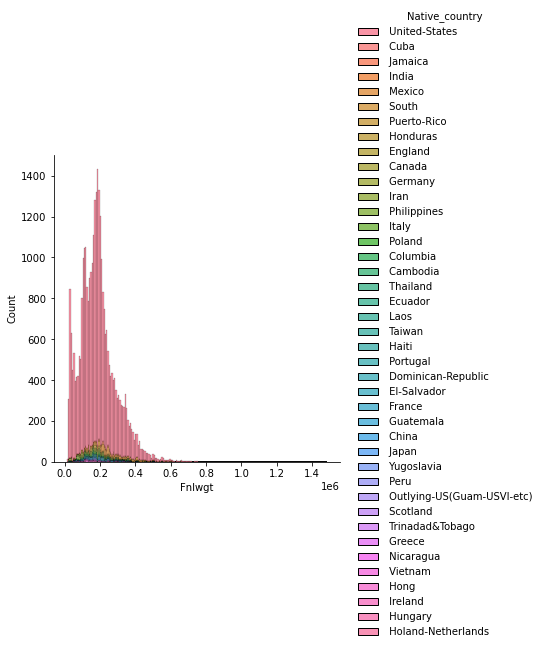
The people earns income >50K are lesser in counts across all ages.

* Earning <=50k: More no of people distribution between age range of 17 ~ 50
* Earning > 50k: Middle aged people with age range of 33 ~ 52



**The Green lines shows more no of people earns <=50K in overall.**





**Origin Country:**

Weighted tallies are higher in counts for People origin from US and Cuba, so we can conclude from above all origin people who earns >50k are lesser in numbers than people who earns <=50k

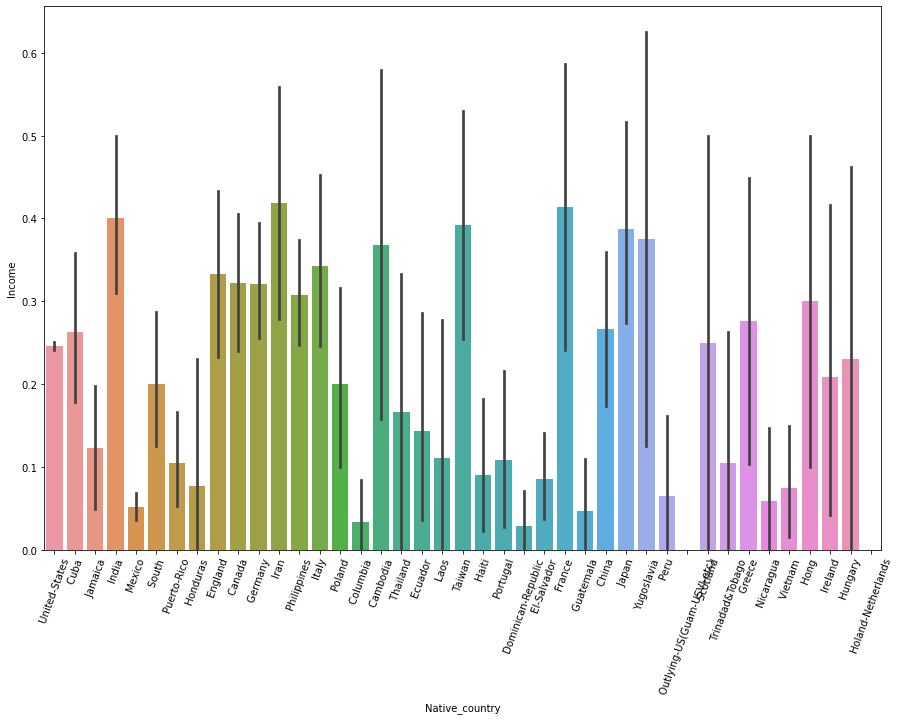
**Variable transformations:**

Pool of Variables are consolidated to compare across various indices for category “0” & “1”.

Analysis clearly depicts the data population over Category of “1” (Earnings over 50k)

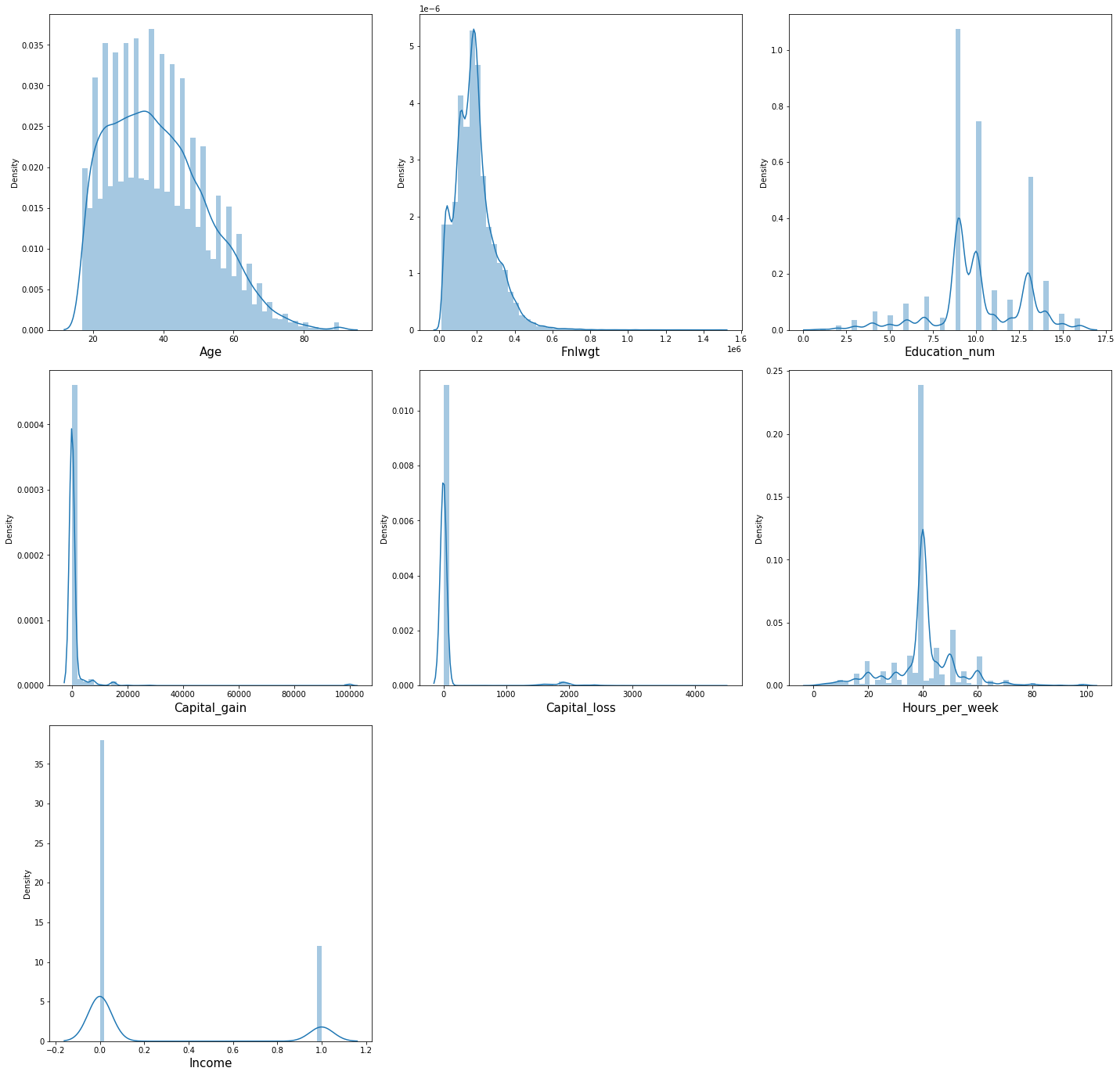
* Private employees earns >50k are more in counts
* Bachelors earns >50k are more in counts
* Married couples earns >50k are more in counts
* Prof-specialty earns >50k are more in counts
* White people earns >50k are more in counts
* Male people earns >50k are more in counts

People of Origin from Iran & India are among the top in people earning more than 50k



**Data Skewness:** All data columns have slight skewness in distributions





**Multivariate analysis**

Graph shows correlation between the two variables. that is,

“if either one decreases or increases what will be the impact on the target variable”.

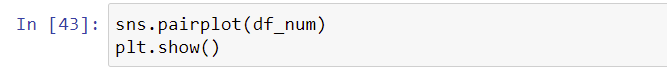
Outcome:

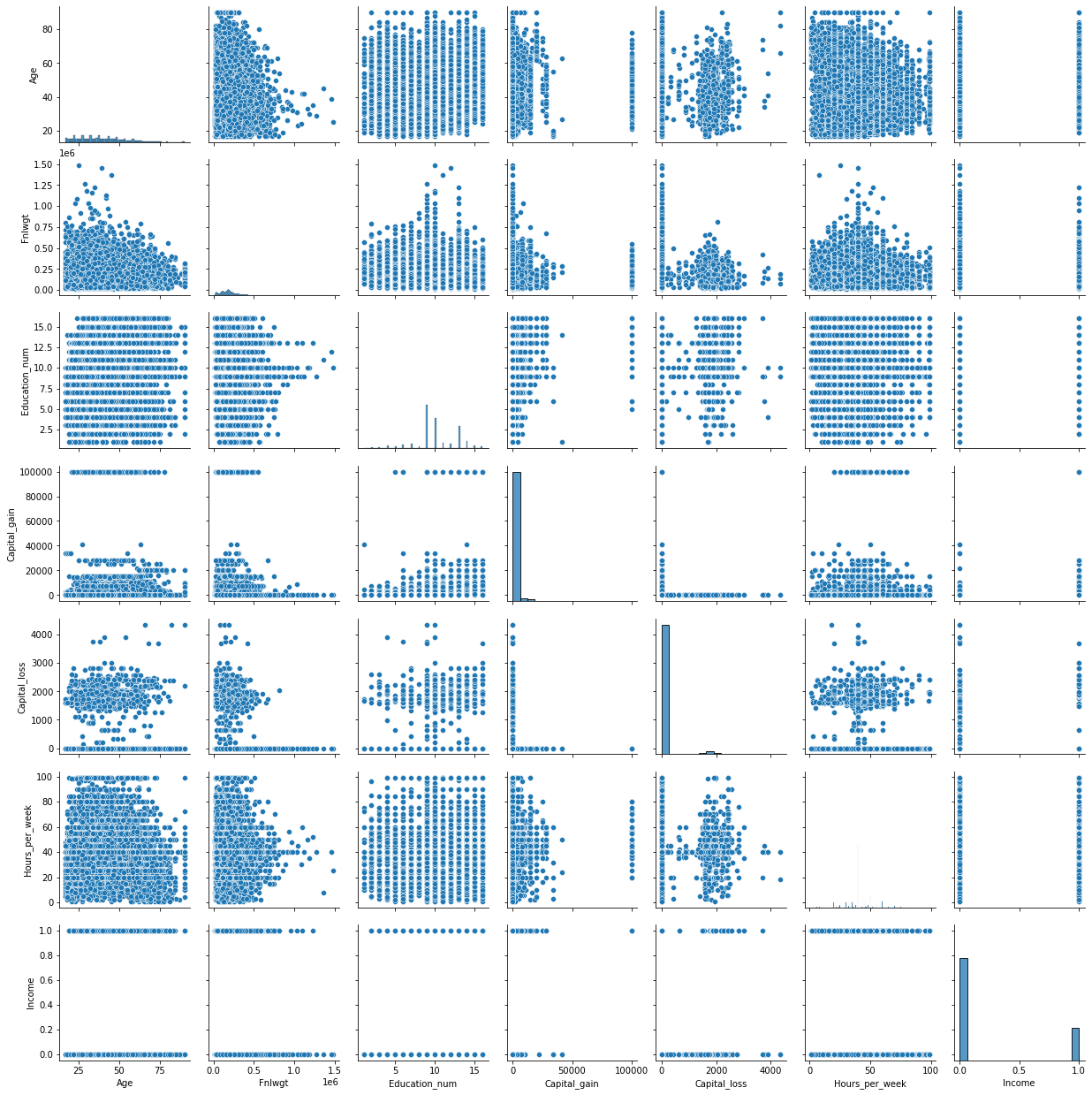
We can observe when income is lower, the capital gain is low.

When income is higher the capital loss is low

Hours per week work doesn’t show much impact on the target variable.

Income fnlwgt is high when income is low and fnlwght is low when income is high.





**Summary of Multivariate analysis:**

* Number of people who earns <=50k is high in counts the number of people who earns >50k.
* If either one decreases or increases what will be the impact on the target variable. i.e. We can observe when income is lower, the capital gain is low & vice-versa holds good.
* Middle aged people with age range of 33 ~52 has Earnings > 50k.

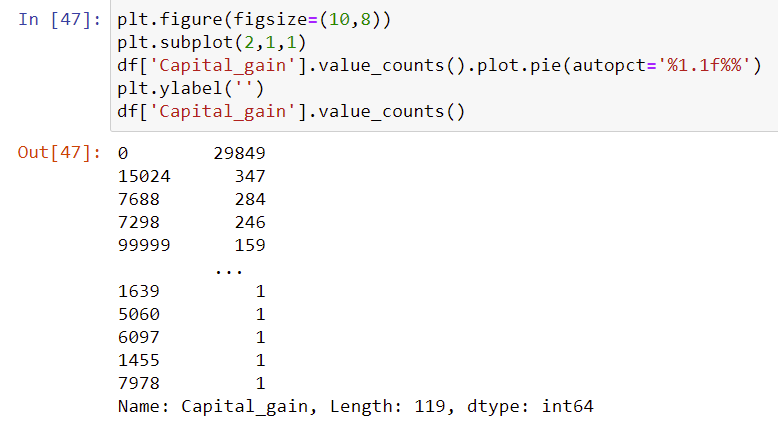
**DATA PRE-PROCESSING & FEATURE ENGINEERING:**

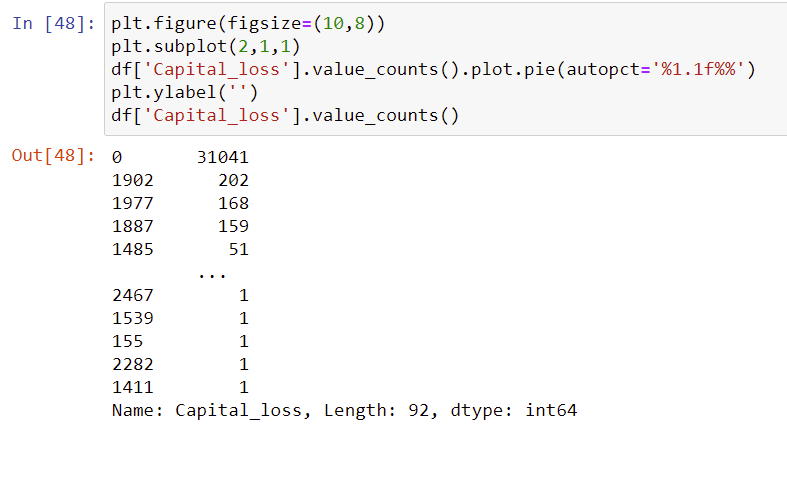
Preprocessing simply refers to perform series of operations to transform or change data. It is transformation applied to our data before feeding it to algorithm. When creating a machine learning project, and doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

****

Data pre-processing is a very vital input to machine learning models, It is to prepare the raw data & make it suitable for efficient machine learning model. These are the methods of data preprocessing and we are going to use the required ones in our project.

Since capital gain has 91.7% of all the values has 0. we can drop this column

****

Since capital losshas 95.3% of all the values has 0.we can drop this columntoo.****

**FEATURE ENGINEERING:**

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features. As you may know, a “feature” is any measurable input that can be used in a predictive model.

Feature engineering**, in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches.** It can produce new features for both supervised and unsupervised learning, with the goal of **simplifying and speeding up data transformations** while also **enhancing model accuracy.**

**Feature Engineering Techniques for Machine Learning**

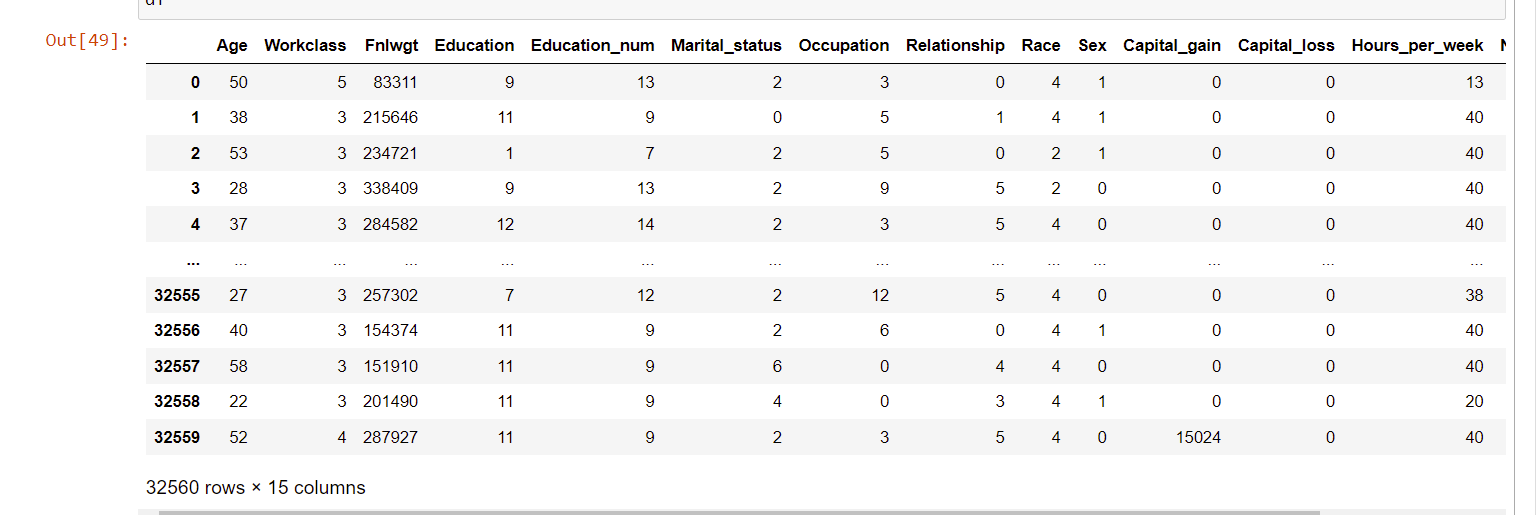
* **Imputation**
* **Handling Outliers**
* **Log Transform**
* **One-hot encoding/Label Encoding**
* **Scaling**

**Data Transformation:**

**Label Encoding:**

**As we mentioned above in library installation,** Label Encoder is used to encode labels by assigning them numbers. It is used to encode single or multiple columns. Thus, if the feature is color with values such as [‘white’, ‘red’, ‘black’, ‘blue’]., using Label Encoder may encode color string label as [0, 1, 2, 3]

****

****

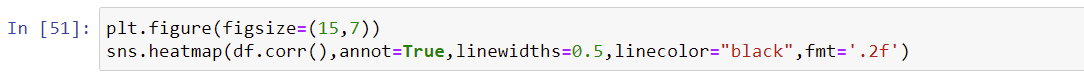
**Correlation with Heatmap:**

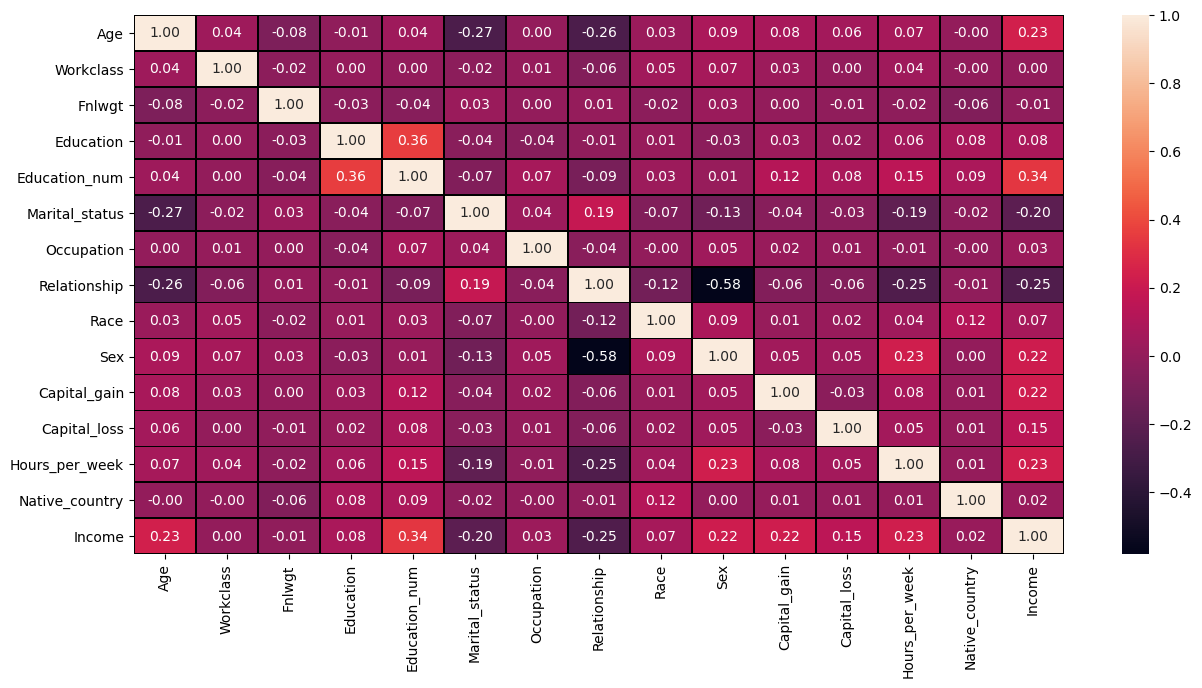
The correlation coefficient is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect [negative correlation](https://www.investopedia.com/terms/n/negative-correlation.asp), while a correlation of 1.0 shows a perfect [positive correlation](https://www.investopedia.com/terms/p/positive-correlation.asp). A correlation of 0.0 shows no linear relationship between the movement of the two variables.Correlation statistics can be used in finance and investing. Pearson correlation is the one most commonly used in statistics. This measures the strength and direction of a linear relationship between two variables.

It can also be defined as the measure of dependence between two different variables. If there are multiple variables and the goal is to find correlation between all of these variables and store them using appropriate data structure, the **matrix data structure**is used. Such matrix is called as **correlation matrix.**

Correlation heatmap is graphical representation of **correlation matrix**representing correlation between different variables.

**For to do feature selection and make feature ready for the model building.we check correlation of variables using heatmap.And describe method for the census data set.**

****



* Education num is the most positively correlated with the target column Income.
* Relationship and marital status is the most negatively correlated with the target Income column.
* Work class has less correlation with the target Income column
* Sex,Hours\_per\_week,Age columns also has good correlation with the target Income column

We can drop the columns which has zero correlation with the target Income column.

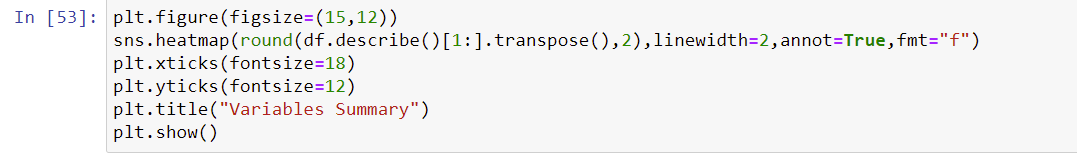
We can check the positive and negative correlation impacts as,

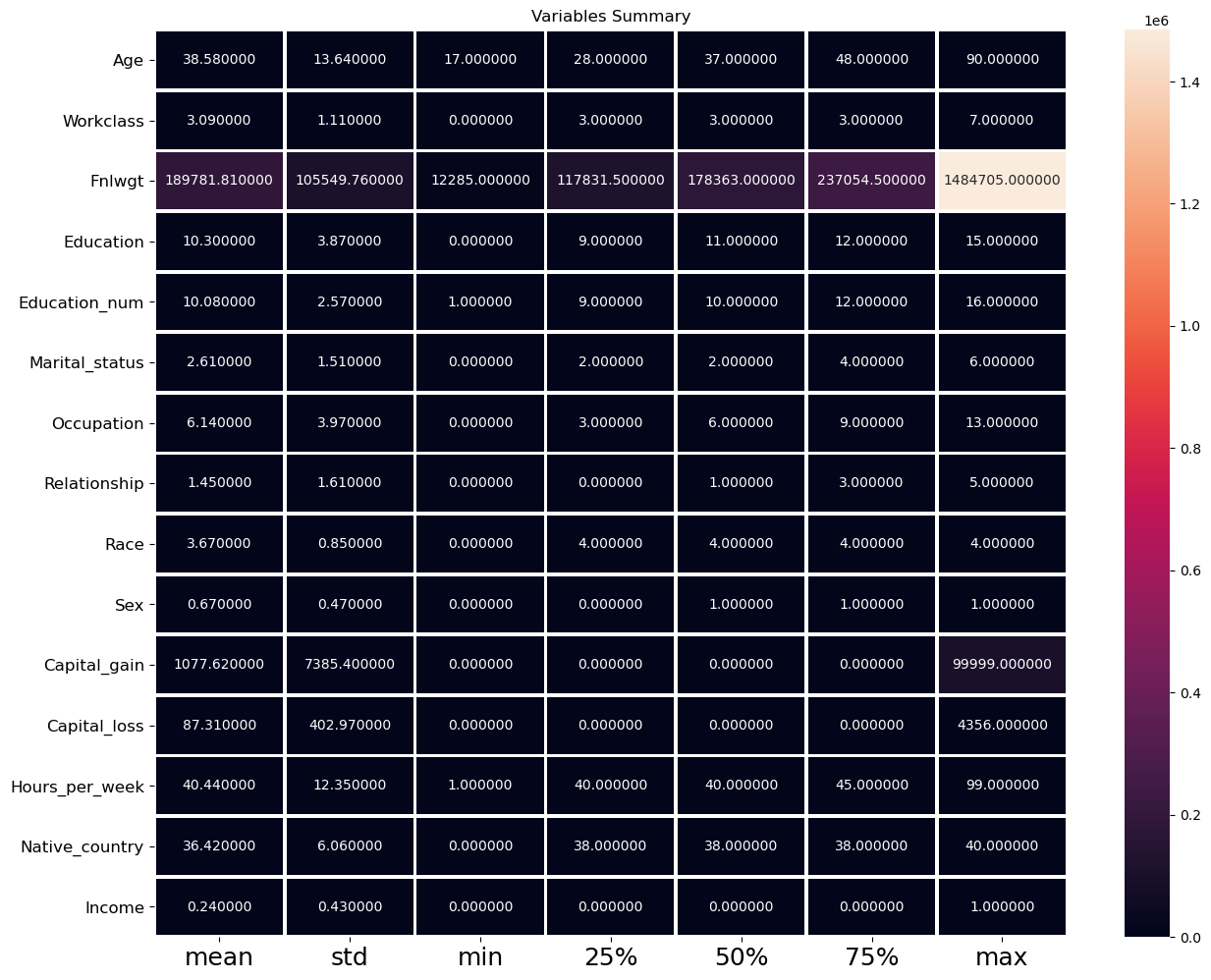
For e.g:

* if education increases income increases.
* if relationships increase income decreases.

**Describe() with heatmap:**

Pandas describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values. Let us plot these values with graphical representation of heatmap.





In Capital gain / loss there is so much of gap between the 75% and the max.

There are outliers and skewness because it has many number of missing values.

Other columns are ok.

Mean and std dev is also much close to 0

**Handling Outliers:**

The most important phase in Feature Engineering is handling outliers because it ensures that our model is trained on accurate data which leads to accurate models. An outlier may occur due to the variability in the data. It may indicate an experimental error or heavy skewness in the data(heavy-tailed distribution). We have three measures of central tendency namely Mean, Median, and Mode. They help us describe the data.

Below are some of the techniques of detecting outliers

* Boxplots
* Z-score

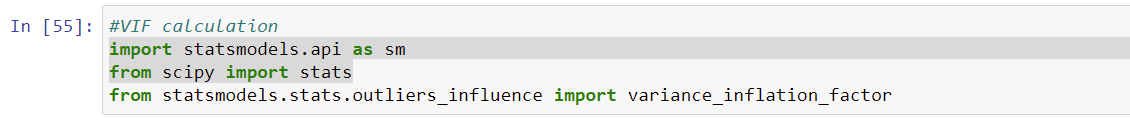
### **Detecting outliers using the Boxplots:**

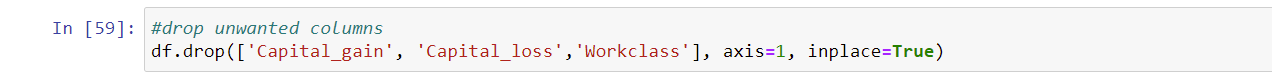


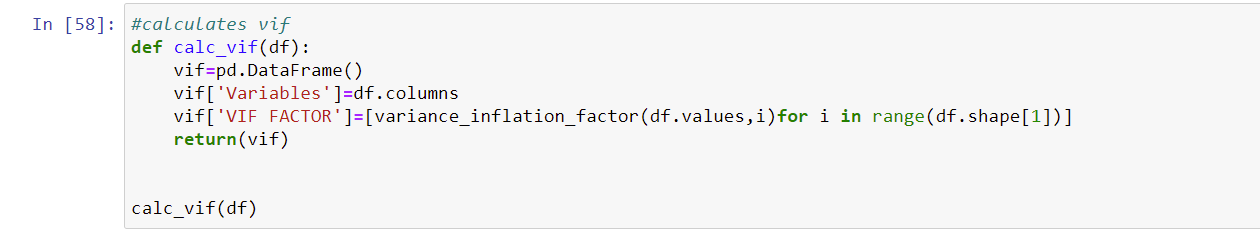
There are outliers present in Age, Fnlwgt and Hours\_per\_week columns. Thus these columns have skewness too.

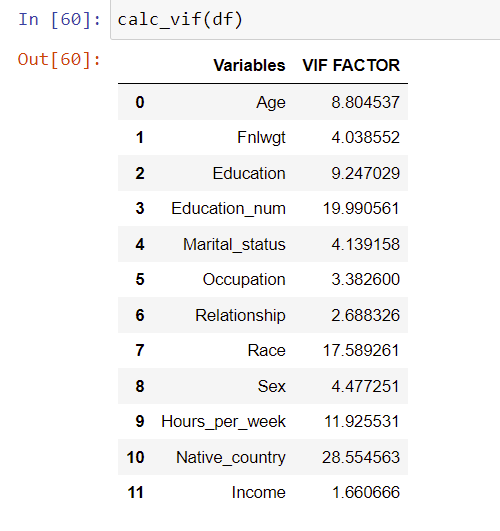
## Variance Inflation Factor (VIF)

Variance Inflation Factors (VIFs) measure the correlation among independent variables in least squares regression models. Statisticians refer to this type of correlation as multicollinearity. Excessive multicollinearity can cause problems for regression models. The stats models package has VIF library, Let us import the package.









The VIF values seems OK after dropping capital\_loss,capital\_gain,workclass columns.

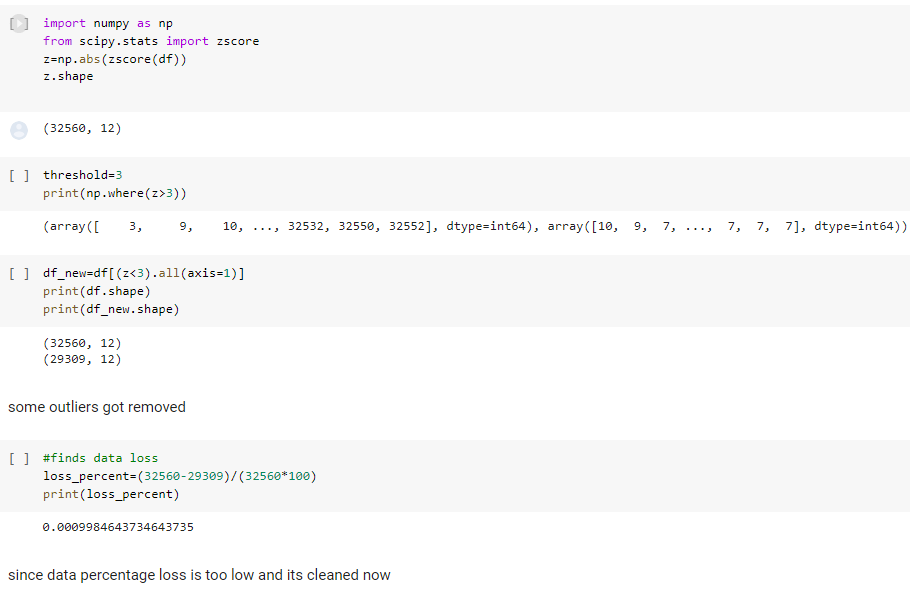
Because the capital loss and capital gain has high number of null values. The workclass column has zero correlation with the target column.

**Checking Z-score to remove outliers:**

SciPy is a scientific computation library that uses [NumPy](https://www.w3schools.com/python/numpy/default.asp) underneath. SciPy stands for Scientific Python. It provides more utility functions for optimization, stats and signal processing.  provides convenient and fast N-dimensional array manipulation. This Scipy package has Z-Score library. So let us import this.

Z-score is a numerical measurement used in statistics of a value's relationship to the mean (average) of a group of values, measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score.

**Criteria:**any data point whose Z-score falls out of 3rd standard deviation is an outlier.



Noticed that Data percentage loss is too low & its cleaned.

**Correlation model:**

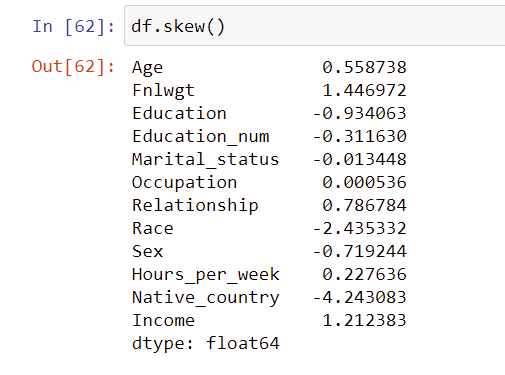
Graph depicts clearly the positive and negative correlation of each variables with target column, justifies the outcome outlined in Multivariate analysis, that higher the education higher the gain & vice-versa

****

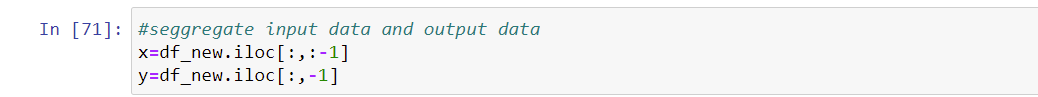
**SKEWNESS REMOVAL AND SCALING:**

Key step prior to initiating Machine learning models, optimizing, scaling the data to provide it as a input to start the modelling.

**Check Skewness:**

****

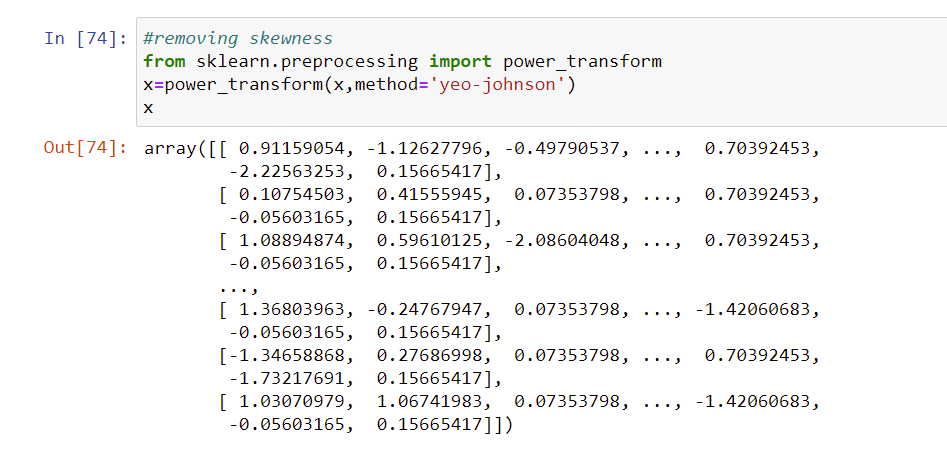
The skewness values should be in the range of -0.5 to 0.5.Age,Fnlwgt and Hours\_per\_week numeric columns has slight skewness.

****

Segregate input and output data for to remove skewness and then to scale the input data for effective model performance.

**SKEWNESS REMOVAL-(POWER-TRANSFORM):**

A power transform will make the probability distribution of a variable more Gaussian. This is often described as removing a skew in the distribution, although more generally is described as stabilizing the variance of the distribution. The log transform is a specific example of a family of transformations known as power transforms. The power\_transform library present in the Sklearn. Pre-processing package.



Data becomes normally distributed with less skewness now.

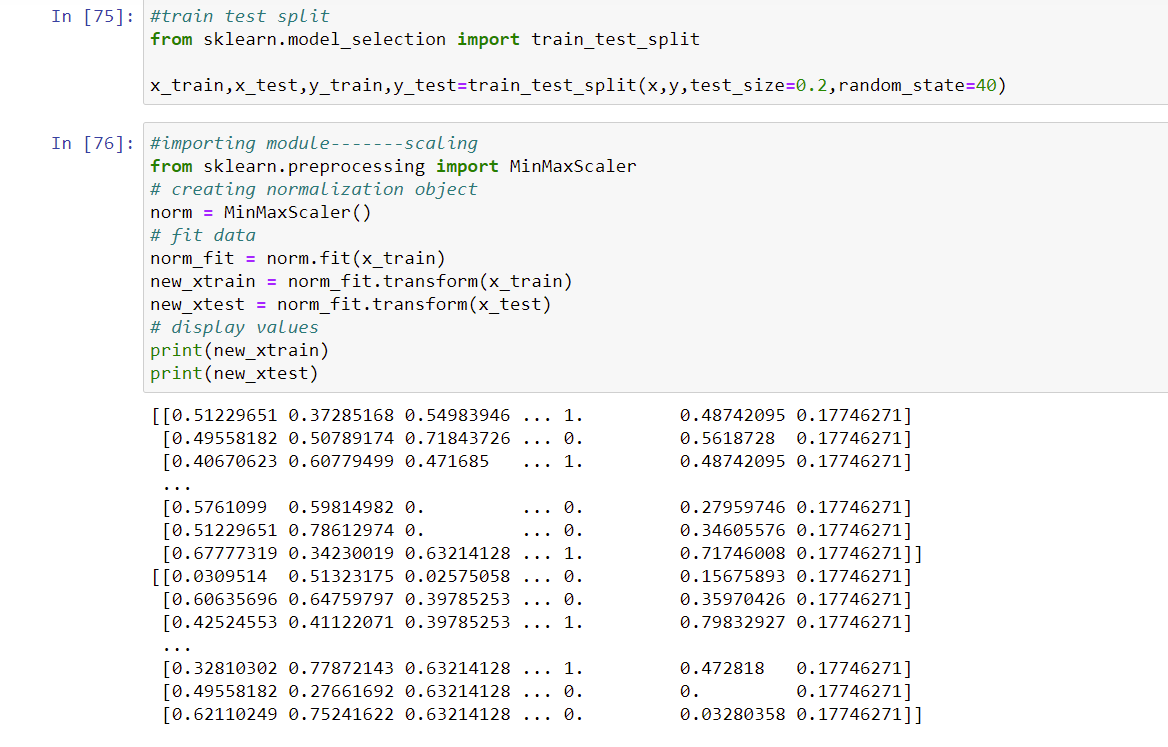
**MINMAX SCALER:**

MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range. It scales the values to a specific value range without changing the shape of the original distribution.

Before scaling we have to train test split the data.since we have to do skewness removal and scaling only on input data.

**TRAIN TEST SPLIT:**

The scikit-learn Python machine learning library provides an implementation of the train-test split evaluation procedure via the train\_test\_split() function. The function takes a loaded dataset as input and returns the dataset split into two subsets.train\_test\_split() will split arrays data into random subsets. The ideal split is said to be 80:20 for training and testing.

****

Note the data values got scaled in the range of 0 to 1 after minmax scaling.

**RESAMPLING TECHNIQUE:**

 Resampling is a technique of repeatedly drawing samples from available training data and refitting our model of interest to each of these samples to get additional information about our model.

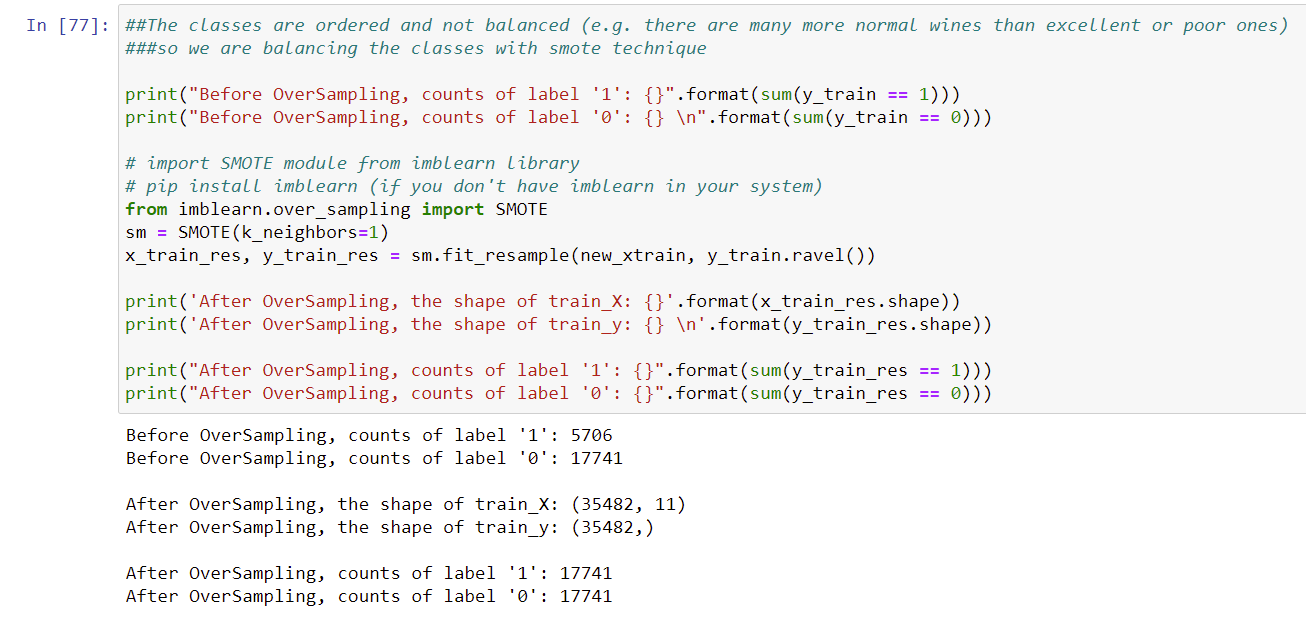
**Imbalanced Class Distribution**, generally happens when observations in one of the classes are much higher or lower than any other classes. In our analysis we found that the people earns <=50K are higher in numbers than the people earns >50K in the target variable. So thus two classes are imbalanced here.

Problem with Imbalanced dataset: Algorithms may get biased towards the majority class and thus tend to predict output as the majority class. Minority class observations look like noise to the model and are ignored by the model. Imbalanced dataset gives misleading accuracy score.

So we do SMOTE which is a resampling technique on the imbalanced our dataset.

**SMOTE:**

It is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.To do this let us now import SMOTE library from imblearn package.

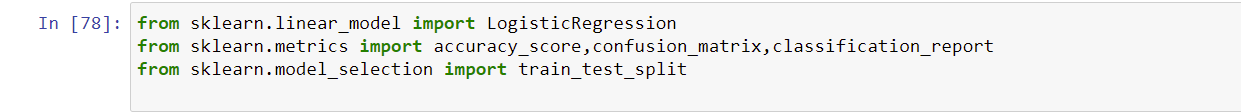


The classes got balanced now due to synthetic smote resampling.

**MODEL BUILDING AND PREDICTING:**

Various Models of Machine learning tools will be used in next steps to do various iteration & select the optimal model for problem solving and predictions.

Let us import the necessary evaluation metrics library and other libraries from sklearn.

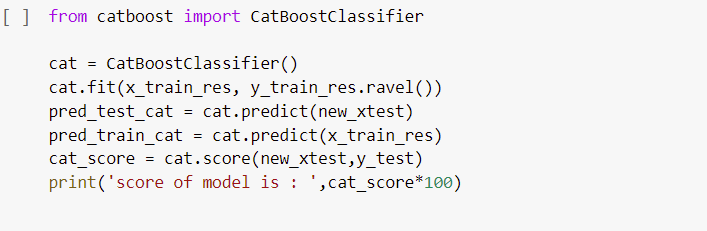


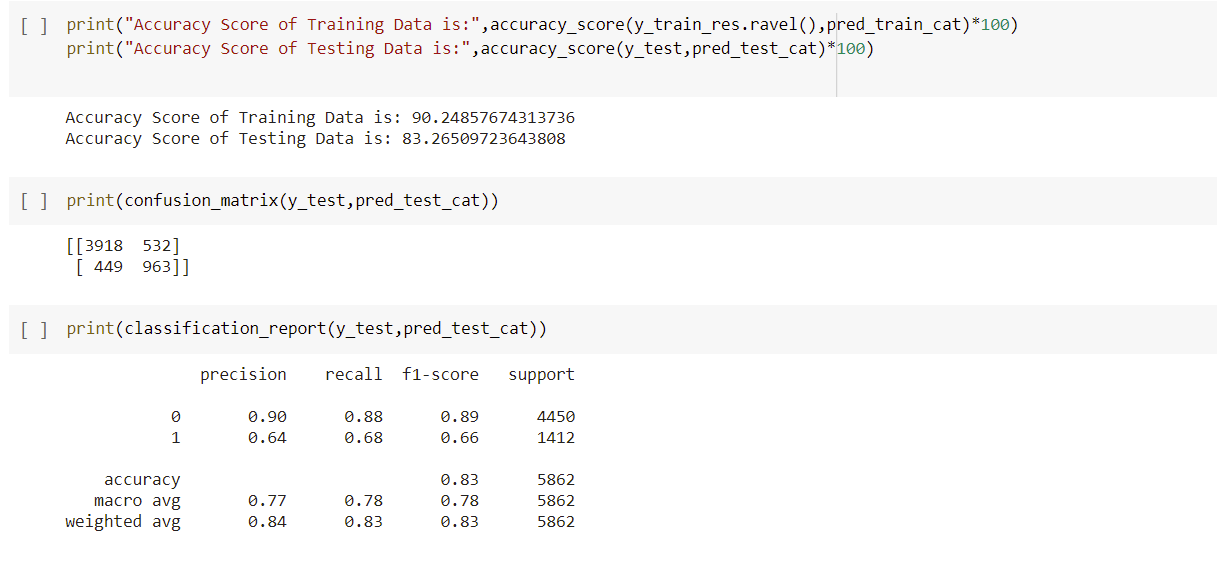
**1. CAT BOOST CLASSIFIER:**

**Gradient Boosting** is an ensemble machine learning algorithm and typically used for solving classification and regression problems.

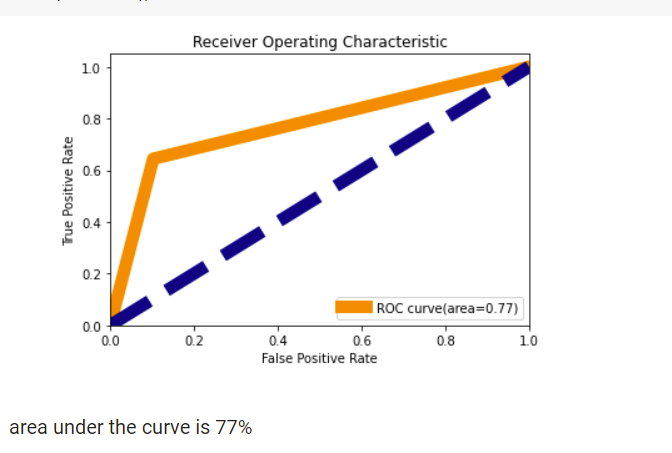
CatBoost implements oblivious decision trees (binary tree). It handles categorical features effectively by ordered target statistics. It has effective usage with default parameters thereby reducing the time needed for parameter tuning.  Unlike other gradient boosting algorithms (require numeric data), CatBoost **automatically handles categorical features**. One of the most common techniques for handling categorical data is one-hot encoding, but it becomes infeasible with many features. To tackle this, features are grouped in categories by target statistics (estimate target value for each category). Target statistics can be calculated in different ways: Greedy, Hold out, Leave one out and Ordered. CatBoost uses **Ordered** **target statistics.**

****

****

****

****

****

Cross-validation analysis & AUC/ROC(Area Under Curve – Receiver Operating Characteristics Curve)analysis is done to understand the performance of the model & outcome shows the Area under the curve is 70%.

The above Evaluation metrics confusion matrix, classification report and AUC\_ROC curve are used to measure the quality and accuracy of the model prediction.

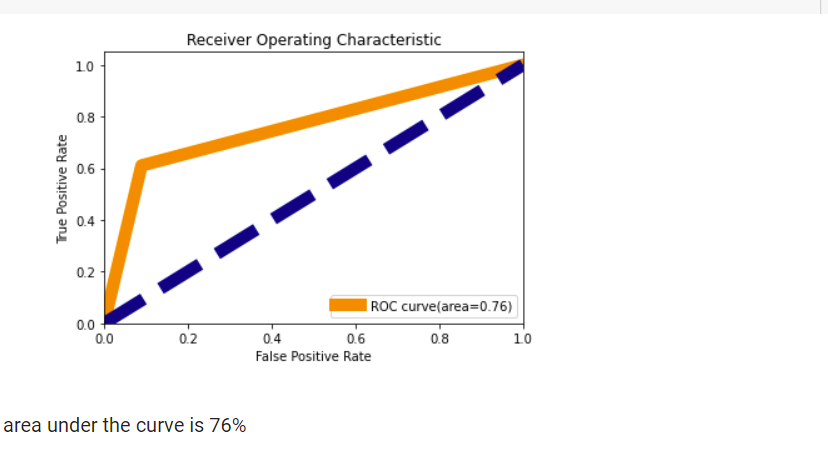
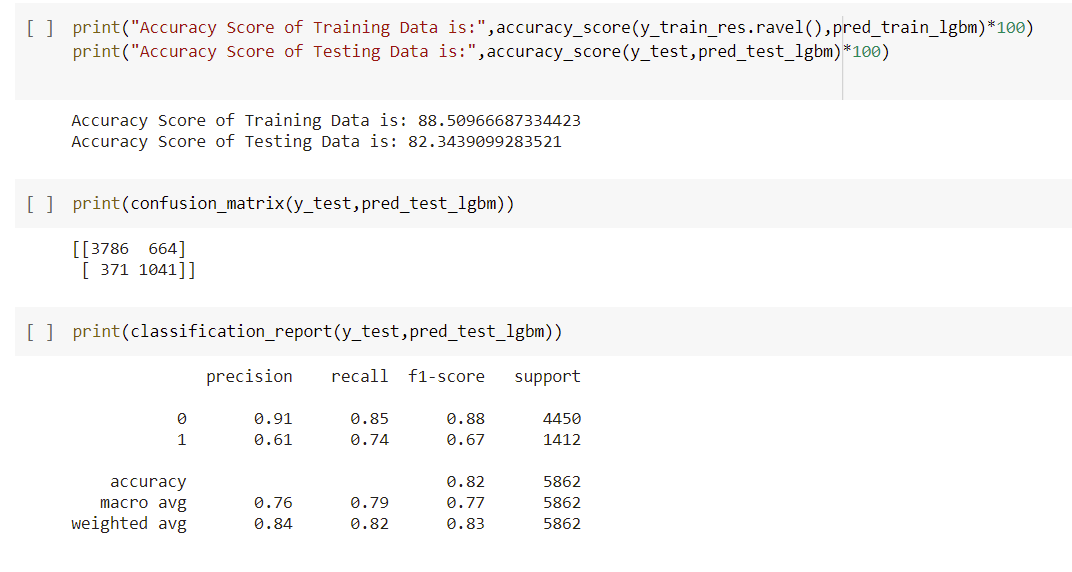
**CROSS\_VALIDATION:**

Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

**2. LIGHT GRADIENT BOOSTING CLASSIFIER**

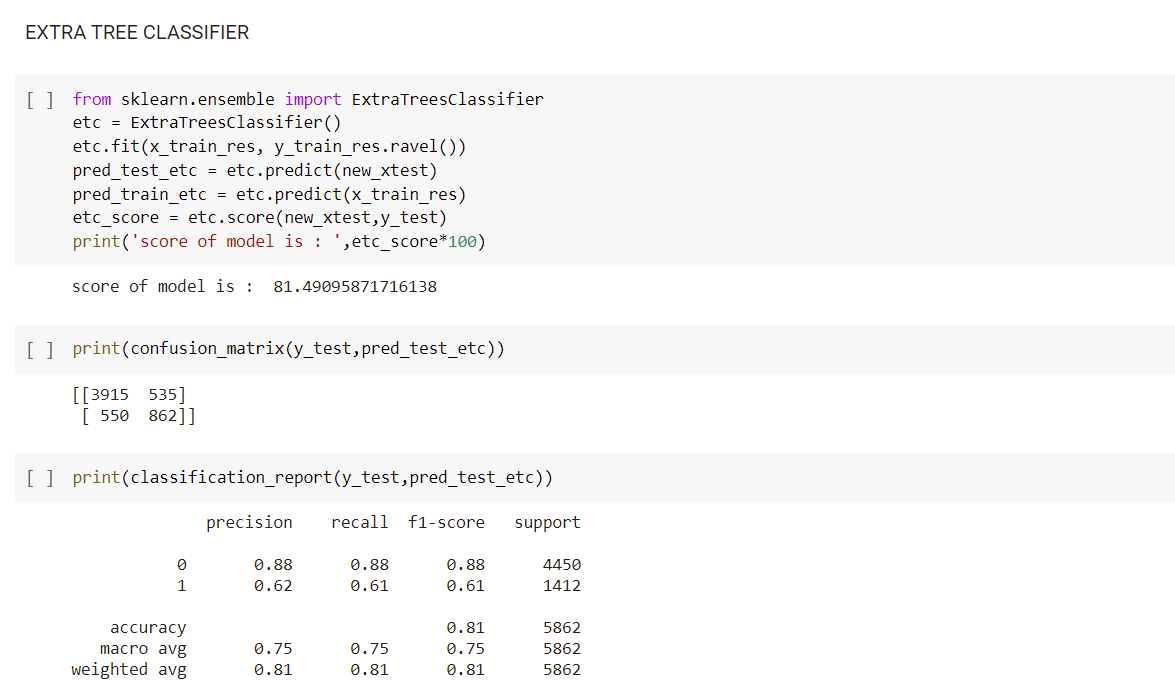
**LightGBM** is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.

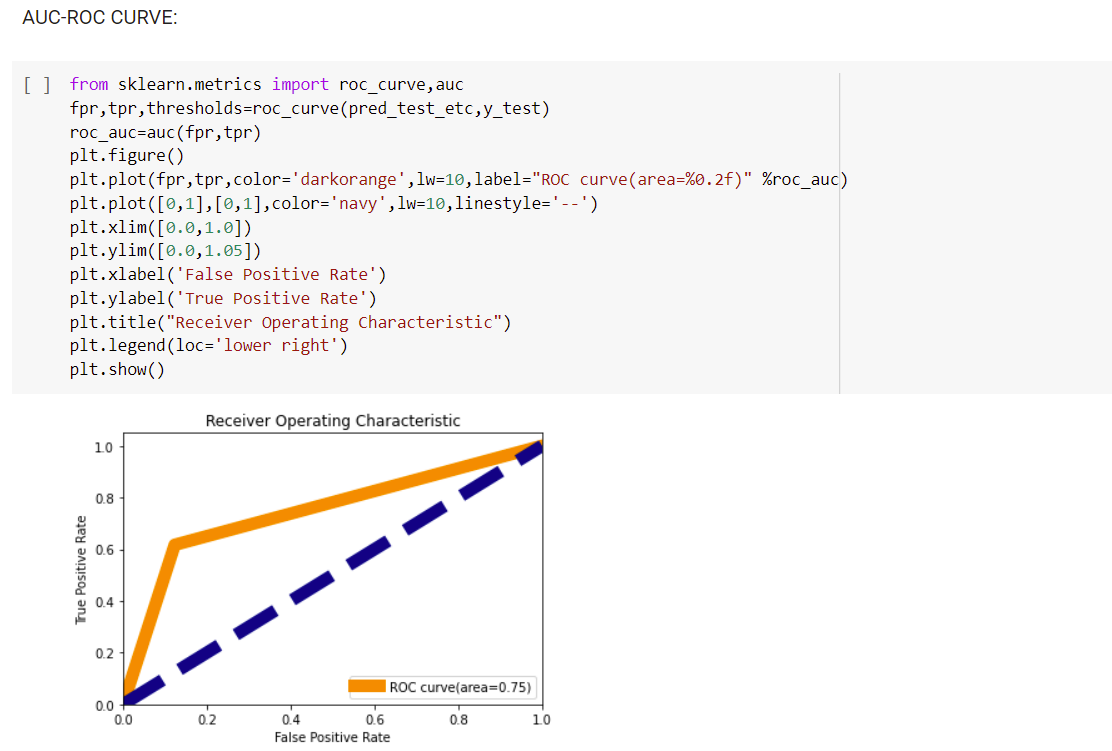
****

****

**3. EXTRA TREE CLASSIFIER**

**Extremely Randomized Trees Classifier (Extra Trees Classifier)** is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it’s classification result.

****

****

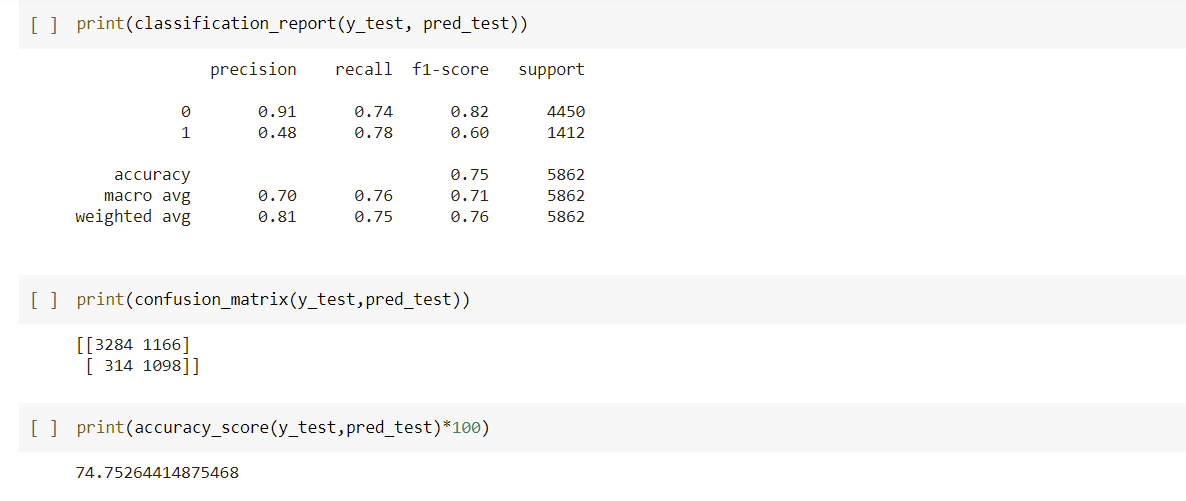
****

**4. HYPER PARAMETER TUNING :**

Grid search is used as an approach to hyper-parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid. GridSearchCV helps us combine an estimator with a grid search preamble to tune hyper-parameters.

Logistics Regression: Statistical analysis method / approach to classify the input data based on prior historical data. It is the most common tool used when we are dealing with Binary classification problems (i.e. 0 or 1 / Yes or no classification values). Best Parameters are selected and passed to GridSearchCV for optimization.



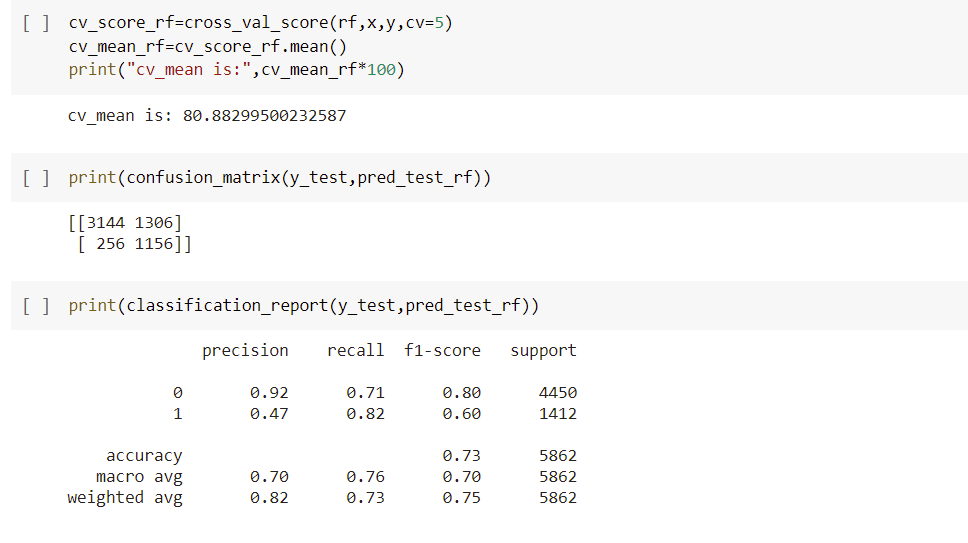




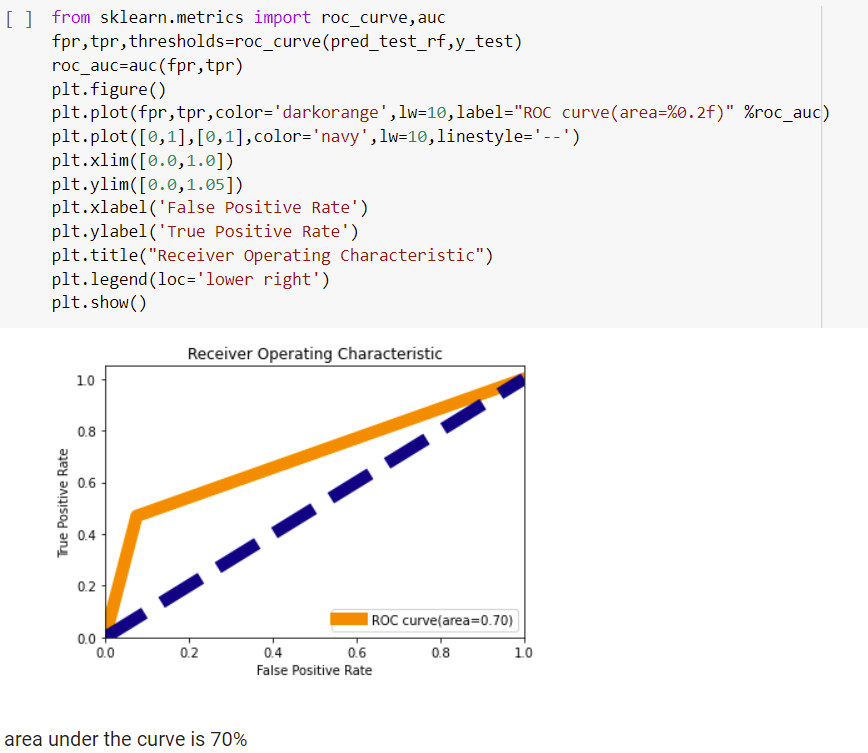


**5. RANDOM FOREST CLASSIFIER**



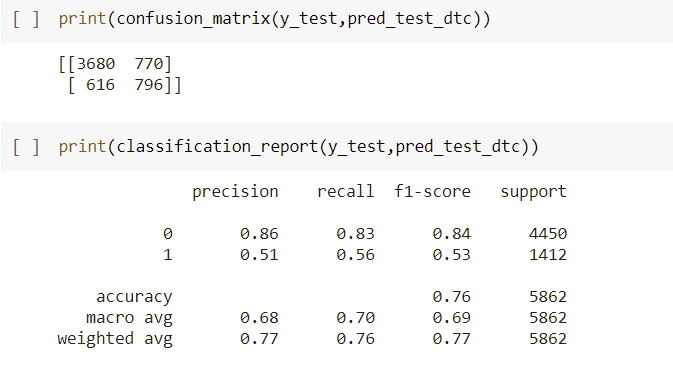


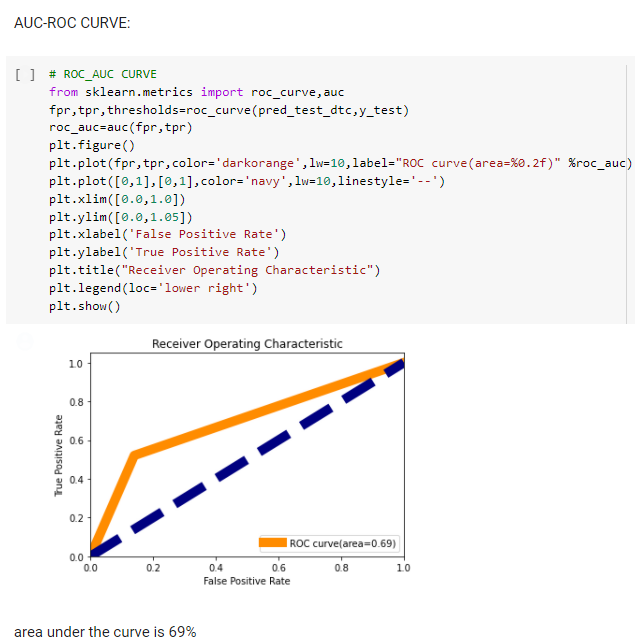
**AUC-ROC curve:**



**6. DECISION TREE CLASSIFIER**

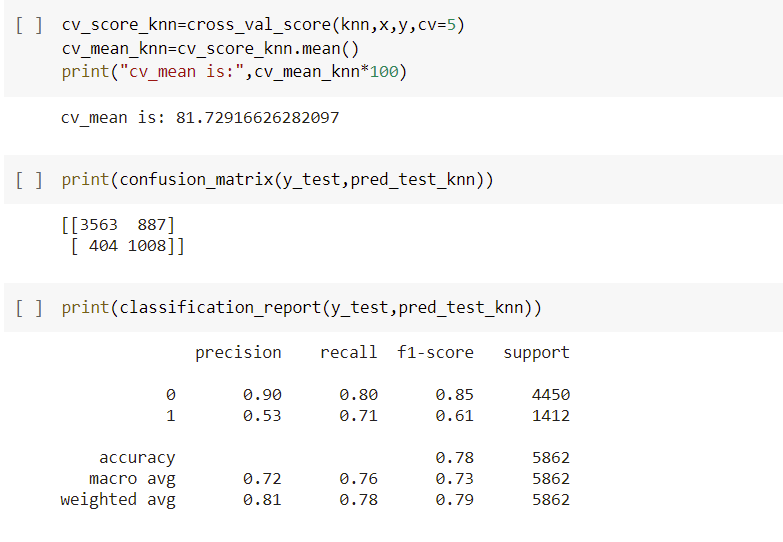
****

****



**7. KNEIGHBORS CLASSIFIER**

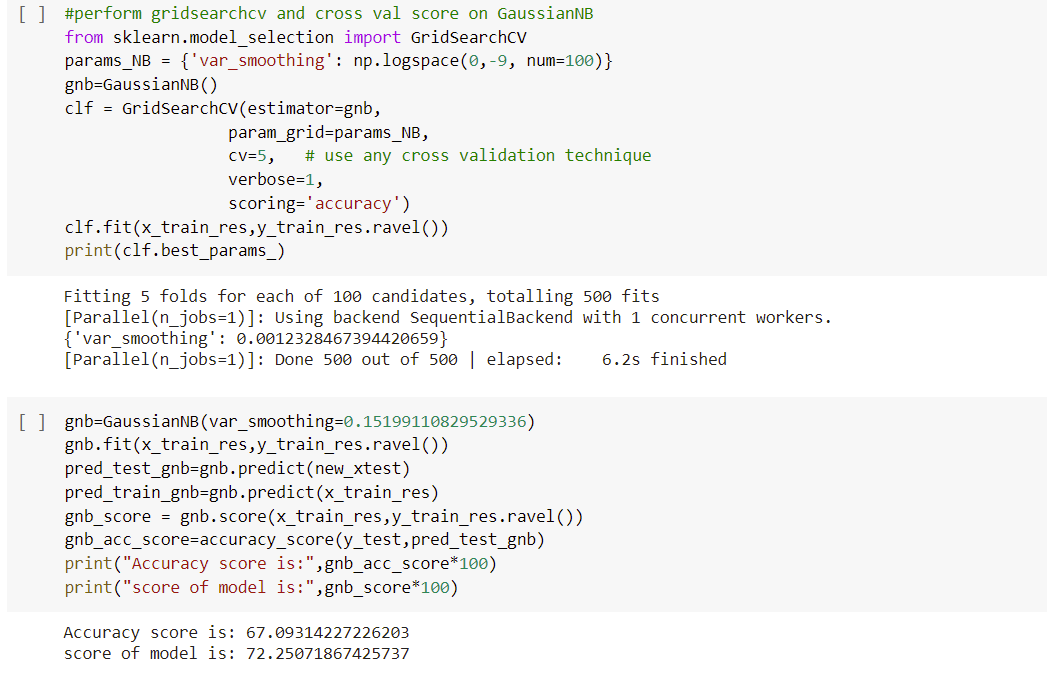
****

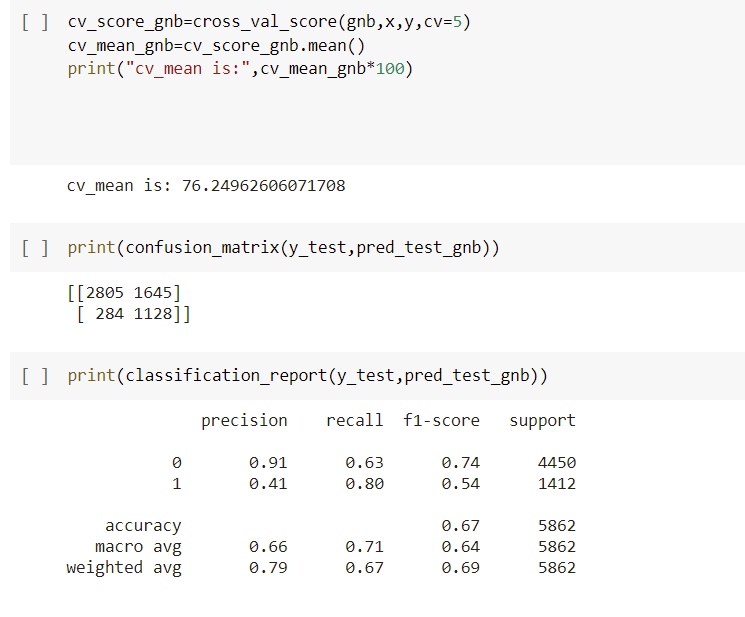
****



**8. GAUSSIANNB**

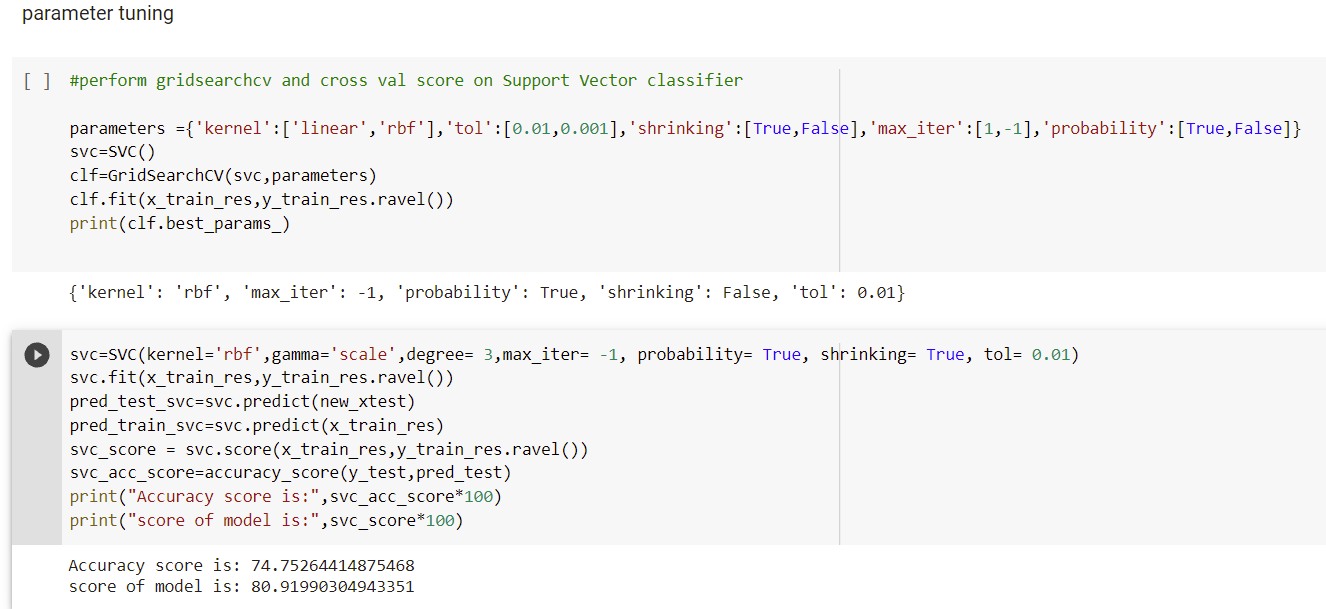
Parameter tuning:

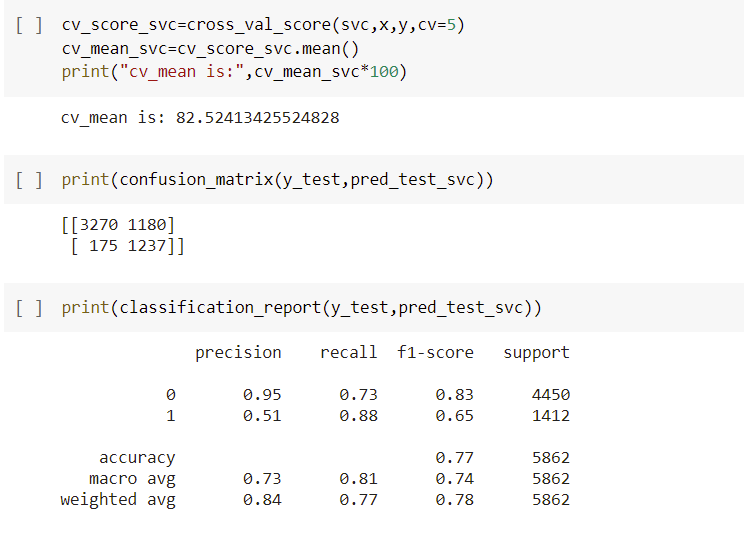
****

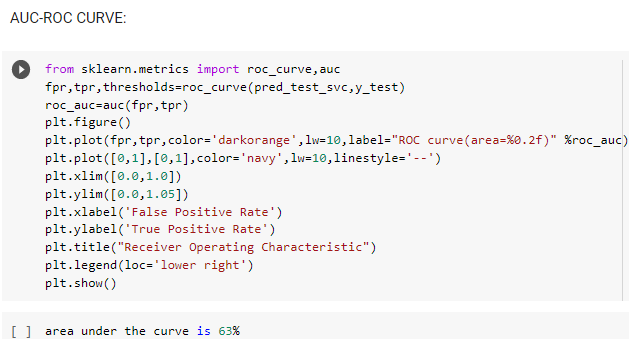
****



**9. SUPPROT VECTOR CLASSIFIER**

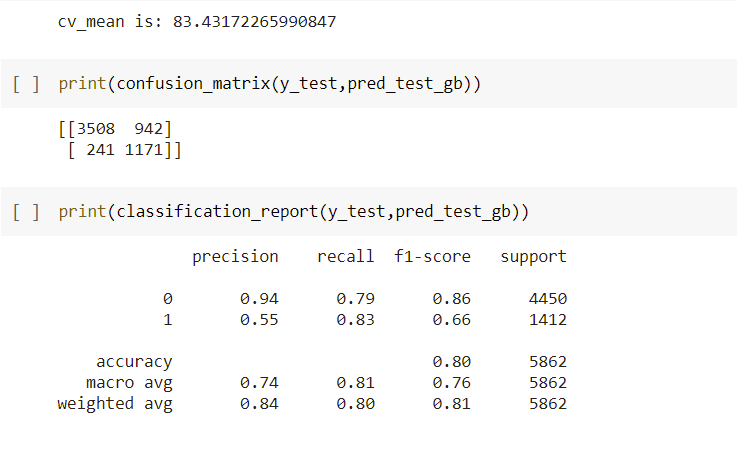
****

****



**10. GRADIENT BOOSTING CLASSIFIER**

****

****



**11. ADA BOOST CLASSIFIER**

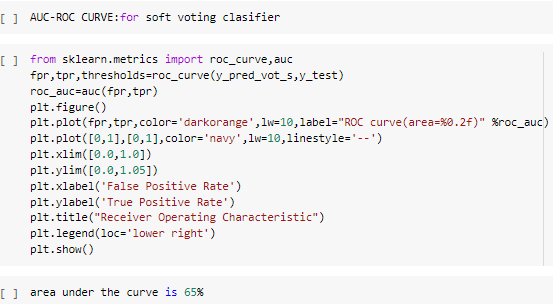


**12. XGBOOST**



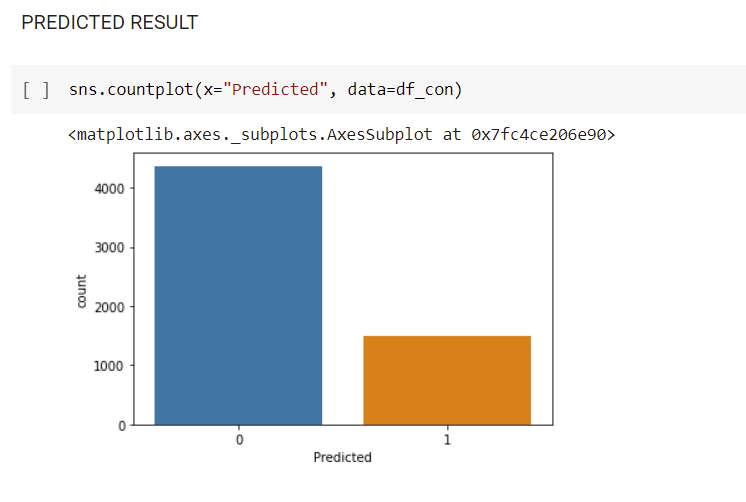
**13. VOTING CLASSIFIER**

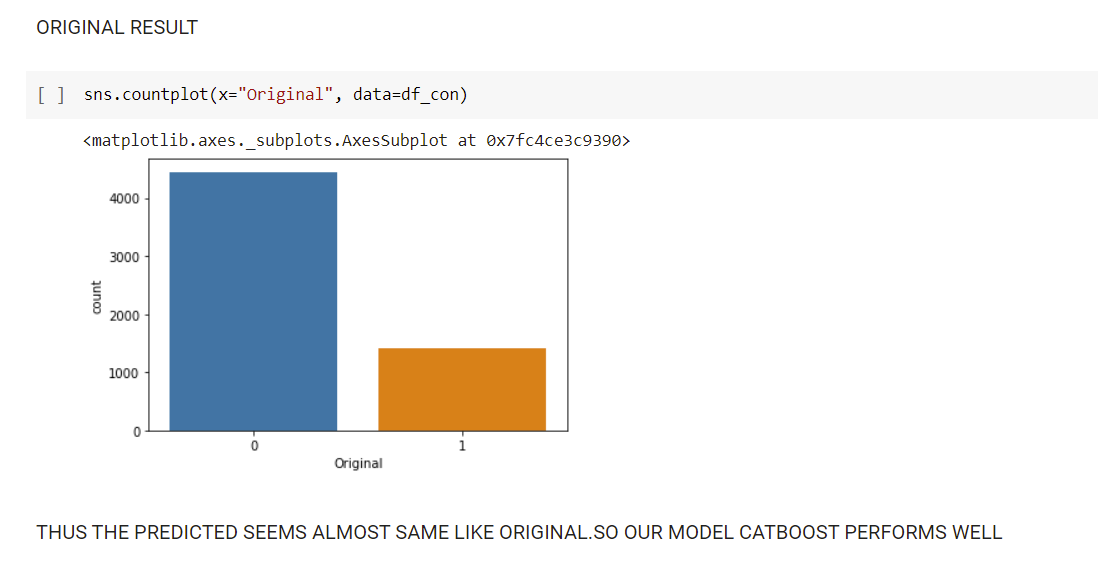


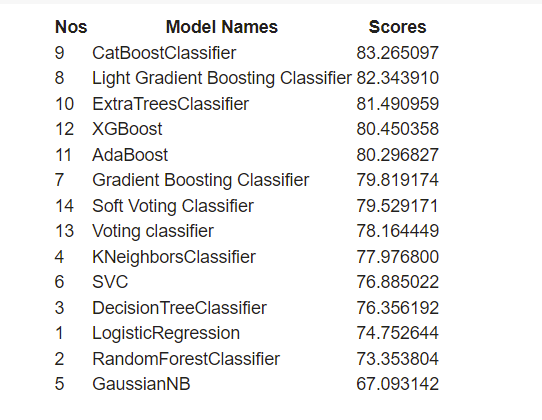


**CONCLUSION:**







**CONCLUSION:**

Key Findings and Conclusions of the Study

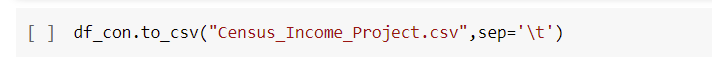
From the above models, CAT BOOST CLASSIFIER works well and gives 83.26% accuracy. CatBoost gives higher accuracy than others since it is based on gradient boosted decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees. The number of trees is controlled by the starting parameters.

So we save this model for prediction

From the analysis we can conclude that more number of people earns only Income of <=50000 irrespective of origin, sex, age, race etc., Only few people earns Income of >50000.

And the Income of people mostly based on their education, working hours, age, marital\_status and relationship.

SAVING THE OUTPUT TO CSV FILE:



Learning Outcomes:

CatBoost implements oblivious decision trees (binary tree in which same features are used to make left and right split for each level of the tree) thereby restricting the features split per level to one, which help in decreasing prediction time. One of the differences between CatBoost and other gradient boosting libraries is its advanced processing of the categorical features. But, if you have a classification problem with Heterogeneous data then also using CatBoost will be the safest thing to do because CatBoost is able to outperform the majority of the Boosting algorithms in the first run.

Recommendation:

Our model performs well. Morever, If less missing values would have present in dataset, then accuracy score would have been increased more. we found out many key findings. when capital loss and fnlwgt is low income of the US people will increase. And also the income increases when people did bachelor degrees(good education),worked in private, if people are from US native and also married etc., These are factors if concentrated it will in-turn improve the countries Economy and Standard of Living.

**Submitted by**

**Srividya**

**Acknowledgement:**

I would like to express my deepest gratitude to all the Mentors of DataTrained Academy, who gave this opportunity to do the project on Census Income Analysis and Prediction & also helped me to gain in depth knowledge of Machine Learning and DataScience to derive insights for organizational goals to meet business needs

For Implementation Code check my <https://github.com/srividya89/Evaluation-phase-week-two/blob/main/Census_Income_Project_(2).ipynb>