NAME : TANYA KALE

PROJECT BATCH : DS2402

INTERNSHIP NO : 2402

Hello Everyone In this blog i am going to tell you how I’ve done an end to end project with details & proper description.

**What You Will Learn In This Blog**

1-Bussiness ProblemUnderstanding

2-Data Collection

3-Data Understanding

4-Data Preprocessing & Data Cleaning

5-Data Exploration

6-Feature Engineering

7-Predictive Modeling

8-Model Evaluation

9-Model Deployment

**Note-***This blog is going to be very easy you just need some patience and read it very carefully.*

**Lets Start**

First in every machine learning project we need to understand the problem statement if i talk about this project problem statement **[**This data was extracted from the [1994 Census bureau database](http://www.census.gov/en.html) by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). **[*The prediction task is to determine whether a person makes over $50K a year*.]**

**Importing Required Library-**

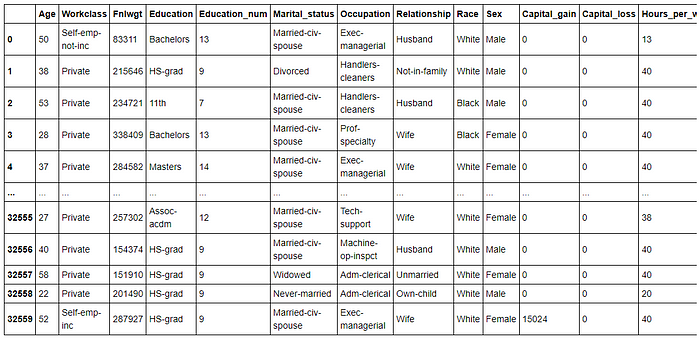


**Census Income Project-**

In this dataset we have 14 independent variable and 1 dependent variable.

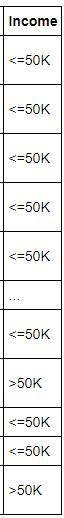
**Independent Variable-**

Age,Workclass,Fnlwgt,Education,Education\_num,Marital\_status,Occupation,Relationship,Race,Sex,Capital\_gain,Capital\_loss,Hours\_per\_week,Native\_country



These are the inputs or i can say dependent variables

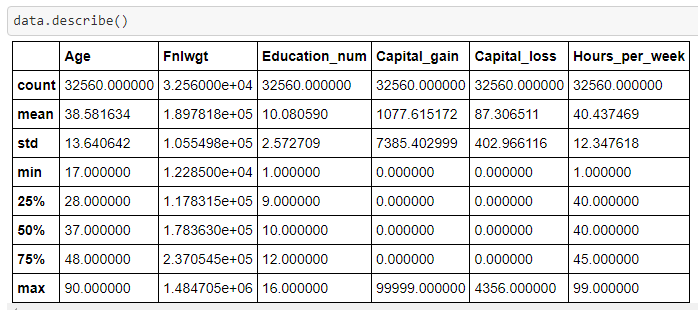
**Dependent Variable or Target Variable**



This is my dependent variable which shows if the person has more then 50k income or less then 50k income.

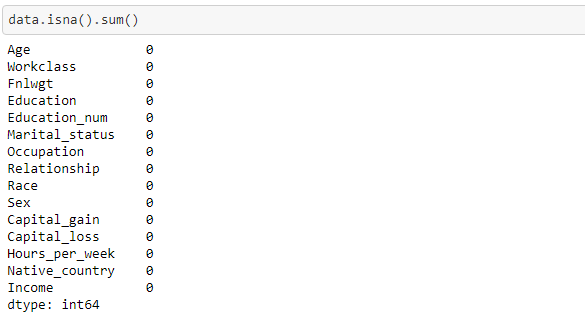
Before going to direct machine learning part we need to do some statistical analysis and EDA.

First i will use describe function that will show me the spread of data how much is my min,std,mean,max value. So i can understand my data and if there is some -ve value that shoul not be there. **example-***Like we have column education num, Hours per week so these value can’t be -ve so if there were some values that were -ve i might filter them.*



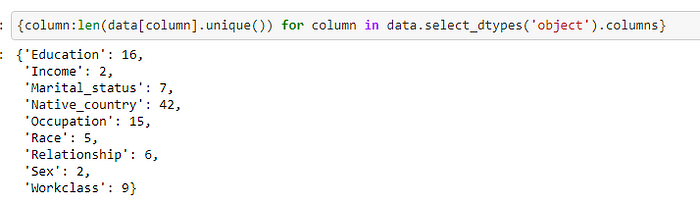
As we can see the min age of people in this dataset is 17 and the max-age is 90 this seems to be an outlier and the 75% of people age is 48 and the education\_num has a min value of 1 and max is 16 that seems to be quite normal like this we can analyze other columns.

There are no null values in this dataset



You can use this code data.isna().sum() for checking if there is some null values or not and if you just want to see total null values you can use data.isna().sum().sum()

**Now we have to analyze our object columns**

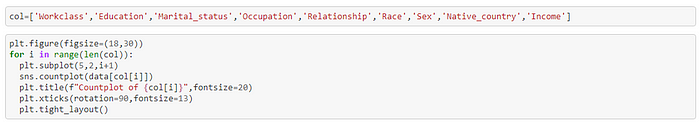


This code is showing how many unique values we have in each column this code really helps you to analyze like if you want to do EDA by plotting so if there is a more unique value you need to increase the size of the figure so you can clearly see each and every value. and for machine learning, you can see if there are fewer unique values like income, marital status, relationship, race, Workcalss we can use OrdinalEncoder if unique values are large i can use label encoder or hashtag encoder or Onehot Encoder.

**EDA-**

**Univariate Analysis**

For univariate analysis we can use countplot to analyze which value has the high count like we have education\_num so from this we can see which education\_num mostly people have. example-if 10 has the high count so we can say mostly people have a bachelor degree.

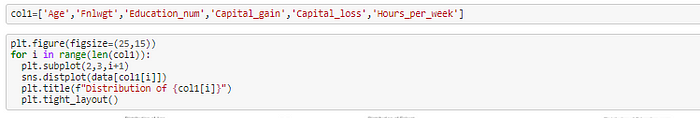


The variable col have all the obejct columns and in the second line of code i am using a for loop so it will draw a countplot for all object columns.

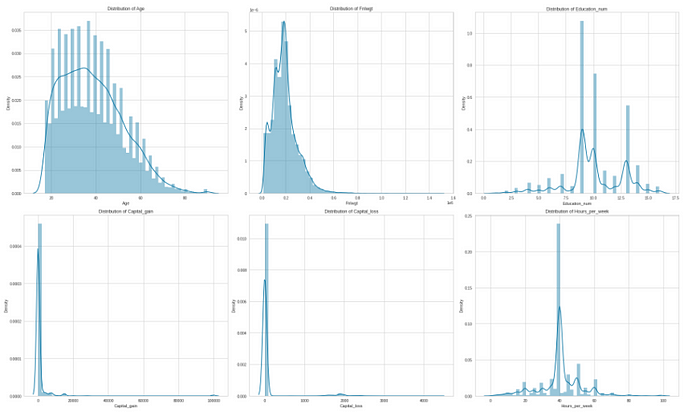
1-From workclass Private has the highest count and other
 
 Private 22696
 
 Self-emp-not-inc 2541
 
 Local-gov 2093
 
 ? 1836
 
 State-gov 1297
 
 Self-emp-inc 1116
 
 Federal-gov 960
 
 Without-pay 14
 
 Never-worked 7
 
 2-From education the count of every class is
 
 HS-grad 10501
 
 Some-college 7291
 
 Bachelors 5354
 
 Masters 1723
 
 Assoc-voc 1382
 
 11th 1175
 
 Assoc-acdm 1067
 
 10th 933
 
 7th-8th 646
 
 Prof-school 576
 
 9th 514
 
 12th 433
 
 Doctorate 413
 
 5th-6th 333
 
 1st-4

From upper graphs we can easily see which value have high count like in this dataset we have more peoples who have a salary of less then 50k and from education i can say there are more peoples who have did HS-Grad and very less people who have only did pre-school. like this we can do analysis

**Univariate Analysis To Check Distribution Of numerical columns-**



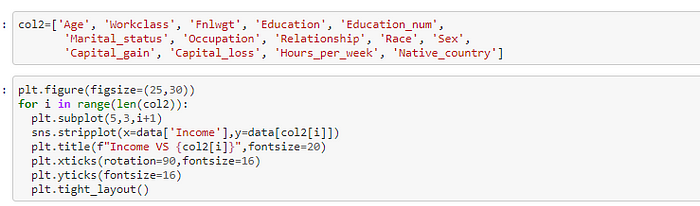
For plotting the distribution plot I am using distplot from seaborn and in the upper code I have a for loop that will draw one by one every numerical column.

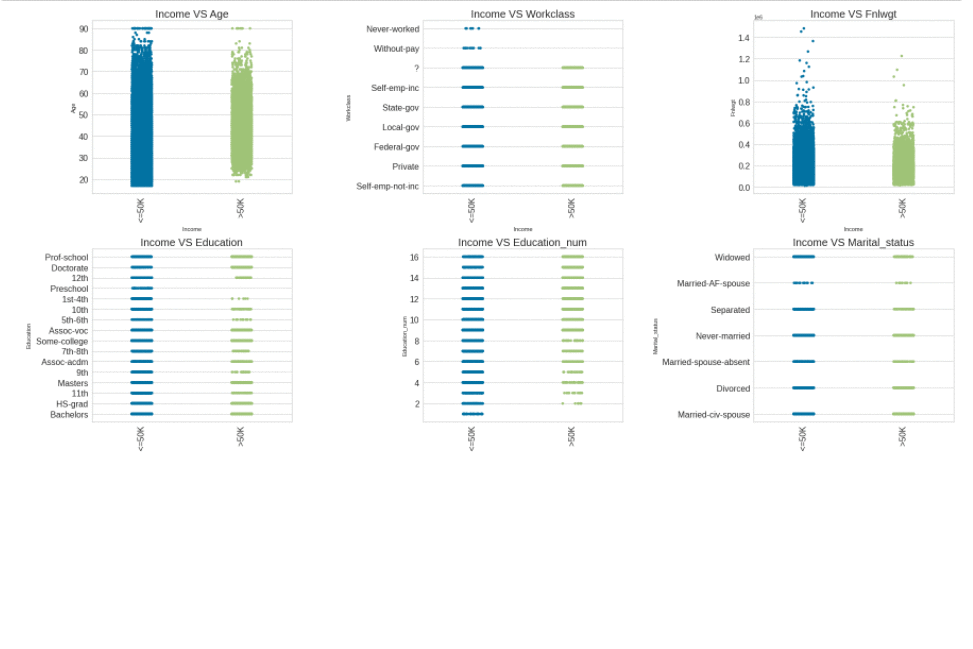


From upper code we get these all plots now we can see the distribution of all numerical features but there are some columns that are discrete like age and hour per week so we only have to check our continuous column because if my continuous column is not normally distributed i have to transform it into normal distribution.

**Bivariate Analysis-**

Bivariate analysis will help us to understand with respect to one feature how other features are performing in this step I am analyzing with respect to mt dependent feature how other are performing.



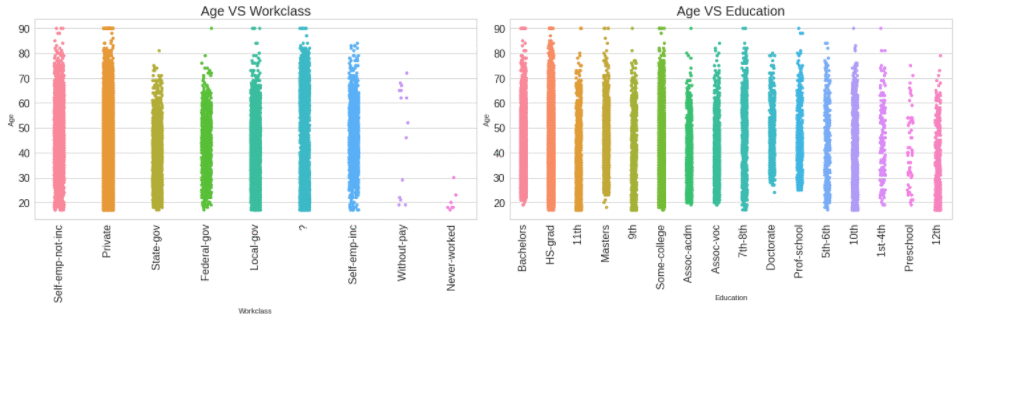


**Obseravtion- 1-**From age i can say from 25–65 people have high chances to get more then 50k, **2-**From workclass i can say except without pay and never worked other workclass have chances to earn more then 50k, **3-**From education i can say as the education increase there are more chances to earn more then 50k same with education\_num, **4-**From marital status i can say except married af spouse all have chances to earn more then 50k, **5-**From occupation i can say as the occupation increase there are more chances to earn more then 50k, **6-**from relationship i can say there is no such effect of the relationship on income same with SEX, 7-From capital gain i can say more the capital income more chances to earn more then 50k, **8-**From capital loss i can say less the capital loss more the income, **9-**for last two plot i plot them again to understand them, **10-**\*people earning >50k income work mean hours per week greater than tose earning, <50k while people from both the categories work from min to max hours per week\*, **11-***the plot shows people belonging to different contries have less chances of earning>50k which is wrong,this is because np.of individuals belonging from other countries other than U.S are very low but it is to be noticed that there are more people in the category <50k than >50k*

**Bivariate Analysis Age Vs All The Features-**

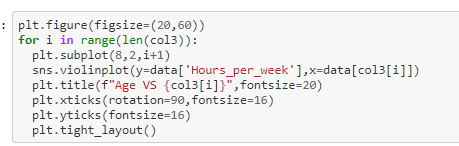


Here I am plotting stripplot and checking what age of people are more in which field and what age of people affect our dependent varaible

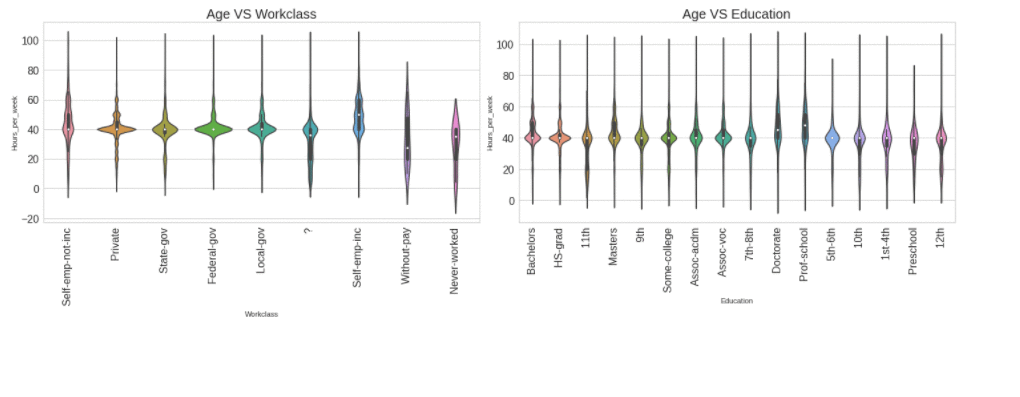


**Observation**- **1-**Individuals working in the gov. sector have almost age 70 to 80 with few outliers which must be the retirement age for them. **2-**there are no individual who do not work after age of 30. **3-**there are no individuals of age >70 belonging to the pre school education category while doctorates and professors appers from late 20’s as they have to study for more years to get to that level of education. **4-**same is the case with education num, as the education number increase age also in increased. **5-**there are no people after the age of 50 in the married to armed forced category with just a few outliers. **6-**Widowed category has seen increase as the age seems to increase there are very few widows at an early age. **7-**there are more no. of working men at higher age than women. **8-**there are very few people belonging from other countries with high age

**Bivariate Analysis Of Hours Per Week Vs all the Features-**

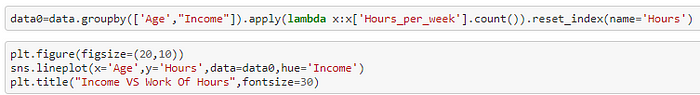


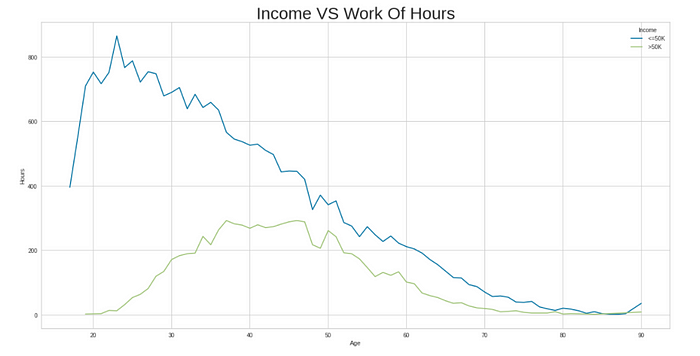
Here I am drawing violinplot to understand the relationship between hours per week vs all the features



**Obseravtion**- **1-**Govt Employees do not work more than 80 hours a week that also with rare cases, **2-**it is been seen that people with less education work more no. hours of the week which is logical, **3-**no armed force person works more than 60 hours a week while farmers and transport movers has working hours mean high than other occupation,**4-**More no. of individuals who have a relationship as own child have high density for working only 20 hours a week. **5-**female works for less no. of hours as compared to men’s, **6-**From relationship people who are husband work more as compare to others and who owned child they work less as compare to other,**7-**From Race white people work more as comapre to other, **8-**From education people who are prof. and are doing doctorate work more no.of hours

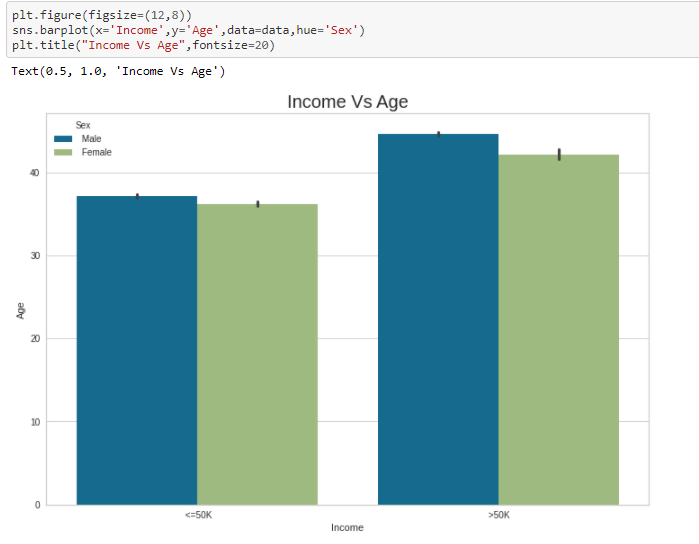
Income Vs Work Of Hours-





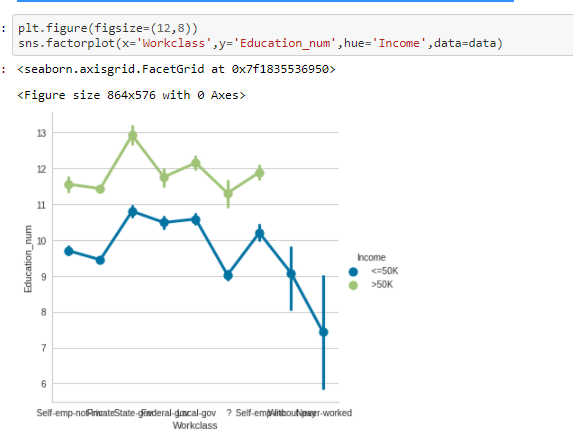
**Obseravtion-**People who get salary <50k they work more and people who get salary >50k they work less as compared to other.

**Income Vs Age**-*Checking As the Age increase of people will there salary also increasing or not.*



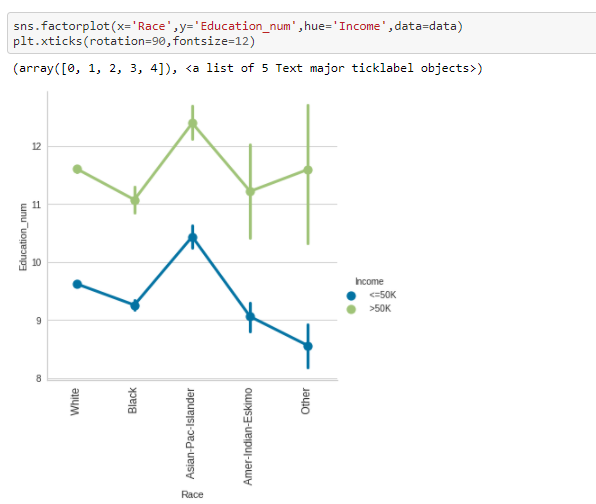
**Obseravtion-**  
As the age increase, people paid more but males paid more always as compared to females

WorkClass Vs Education Num-*This helps us to understand as the education increased will the workclass and income also increasing.*



**Obseravtion-**Some people belonging to a particular workclass might have less education and some workclass might require more education level but no matter whether workcalss, people in the same workclass, if they have higher education level they earn more. it is also to be noticed that there is no person from without pay and never worked workclass category who earn more than 50k which is logical.

Race Vs Education Num*-This will help us to understand which race people are more educated and which are earning more.*

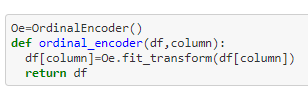


**Obseravtion-** Asian pacific race have comparatively more education than the fellows who earn same as much as they do,belonding to other races.Indians and some other races earns >50k with lowest education level

**Machine Learning-**

In Machine Learning We Can’t Give Raw Data To Machine Learning Models So We Have To Do Various Type Transformation And Have To Follow Many Steps Have So I Will Go Step-By-Step.

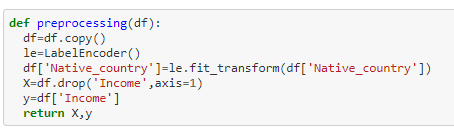
**1-Feature Transformation-***Feature transformation is a very important step, our machine learning model don’t understand the words it only understands numerical values so we have many techniques for doing feature transformation so we use each technique as per the situation like in this i am using OrdinalEncoder And LabelEncoder.*



https://miro.medium.com/v2/resize:fit:700/1*No_wG7sP9hM1BoIif5Xw9g.png

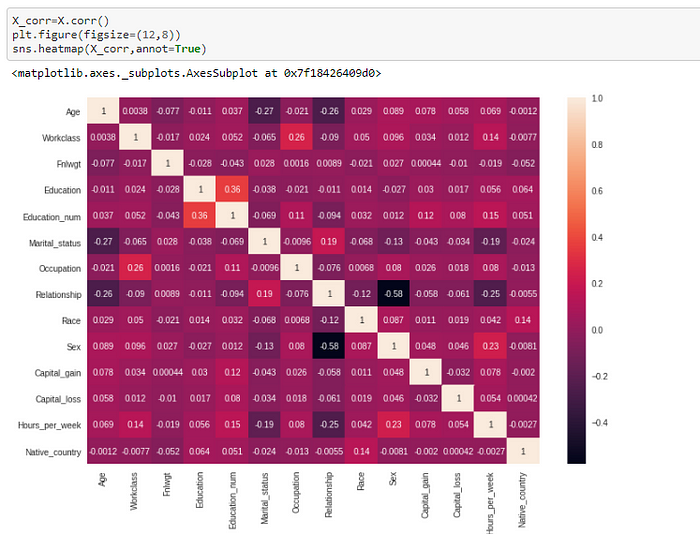
I am Converting all these columns as per order because all of these columns have values in order and have less unique values.

**2-Preprocessing-***In this preprocessing i am converting one categorical feature with the help of LabelEncoder because it has many unique values and also splitting my data into two part one is X which have all the independent variable and other is Y which has the dependent variable*



Preprocessign function which will return me X and y value

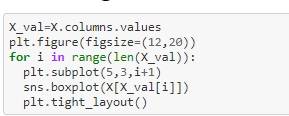
**3-Checking Multicollinearity [if any remove it]**-*Multicollinearity is a big problem in machine learning project where two features have high correlation which will affect our model result so in way to handle it we have many methods like we can use PCA and if i have two features which are highly correlated i can drop one of them.*



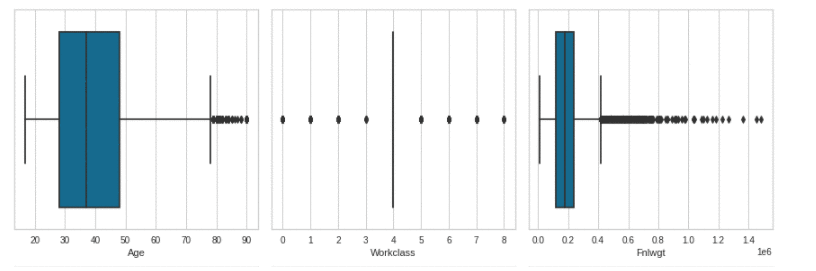
This is one of the way to check multicollinearity this method is called heatmap which shows the correlation between features correlation have the value of +1 to -1 where +1 shoes positive correlation and -1 shows negative correlation and 0 shows no correlation. So from this graph i can say *Correlation is seemed to be good like education having good +ve correlation and relationship and sex having -ve correlation and rest of the columns are also having correlation but at low level.*

**4-Checking Outliers [If Any Remove Them]-**

Outliers are extreme values that fall a long way outside of the other observations. The process of identifying outliers has many names in data mining and machine learning. There Are many ways to remove outliers and to detect it but the best way is BOX-PLOT which shows if there are any outliers or not.

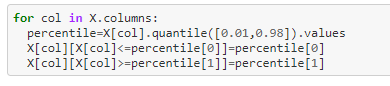


This code help me to plot boxplot for every feature that i have



There are mostly columns that are object and some columns that have discrete values like age and hour per week but there are some continuous values and yes we have some outliers in that features.

Removing Outliers-



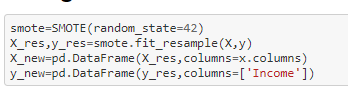
removing outlier using percentile method

There are many ways to remove outliers like Z-Score, IQR-Method but i am using percentile method which replace the outliers with percentile i am using this method because in data science data is very important and i don’t want to loose the data.

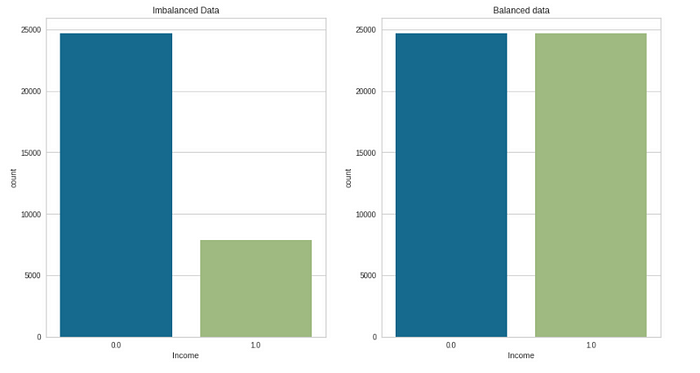
**5-Balancing Dataset-**

In classification machine learning problems we generally have a imbalanced dataset problem where we have 75% or 80% records of only one class and 20% records of second class so in this if we give this imbalanced data to our machine learning model it might give you good accuracy but when it come to performance metrics it will not give you good result because most of the time it learned patterns that related to the first class.

So to deal with this problem we use sampling techniques like upsampling and downsampling. In this problem, i have used Upsampling, Even in upsampling we have many ways but in this problem i have used SMOTE,  
SMOTE is **an oversampling technique that generates synthetic samples from the minority class**. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.



Balancing dataset with SMOTE sampling

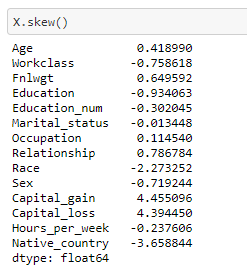


From this graph we can see the difference between imbalanced and balanced dataset

**6-Checking Skewness[if any remove it]-**

Skewness is the degree of distortion from a normal distribution for machine learning model.

Reducing skewness: A transformation **may be used to reduce A distribution that is symmetric or nearly so is often easier to handle and interpret than a skewed distribution**.



The value for skewness is 0.5 to -0.5 if we have a value greater then 0.5 or less then -0.5 it means our data is skewed and we can only reduce skewness of continuous columns as we can see our data also have some skewness.

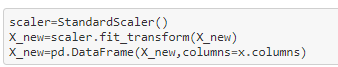
Removing Skewness-

https://miro.medium.com/v2/resize:fit:371/1*YP5XfY4dhxQiaapecgM_4Q.png

Removing skewness with the help of power transformer i am using method=’yeo-johnson’ because it will also deal with the -vs skewed data

**7-Feature Scaling-**

Feature scaling is a very important step in machine learning in datasets we usually have some extremely high values and some min values so feature scaling is used to convert them between 0–1 range so our model can easily interpret values but if we are using tree base model we don’t need to scale our values this is one of the advantages of tree base models.



Here i am using StandardScaler to scale values, we can only use standardscaler when my data is normally distributed and my data was skewed that’s why i had used power transformer to remove skewness so my data was converted into normal distribution so i am using StandardScaler.

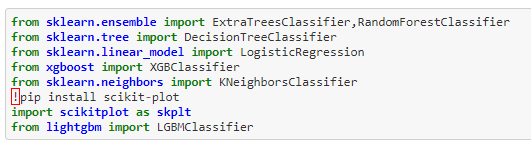
**8-Splitting Data into train and test-**

https://miro.medium.com/v2/resize:fit:651/1*0aUMpjQFEkoBtjPAD1GMgw.png

I am using train test split to split my data into train and test and i am using 30% of data for testing and rest of all data for training.

**9-Model Building-**

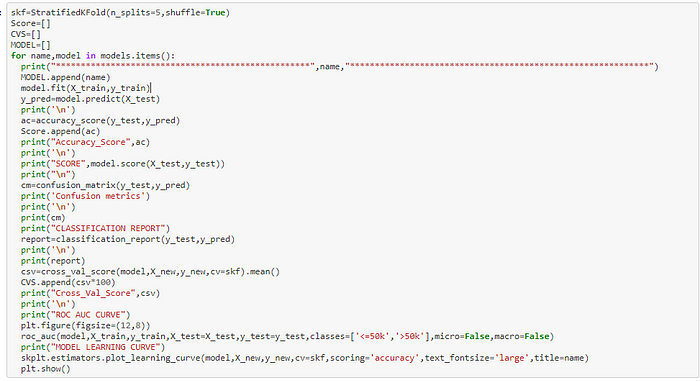
Importing library for machine learning model building



For this problem statement, i have used these model

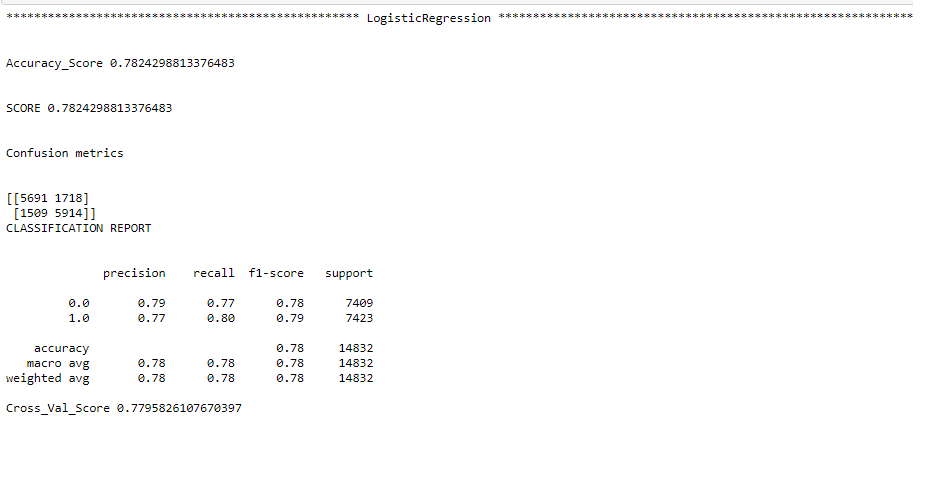


Mostly i have used tree-based model because it gives better result.



I have written this code where i am using for loop which will return me all the models one by one with all the performance metrics and roc-auc curve as well as model learning curve so i can check my model is overfitted or not. and i am using StratifiedKFold so it will select equal fold for each class.

Every Model Performance Result-

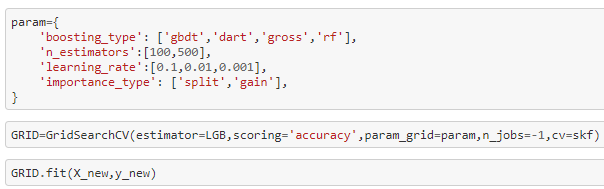


Model Performance Result

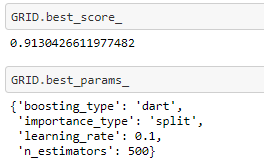
From every model that i have trained LightGBM is giving me good accuracy and performance metrics so i will do hyperparameter tuning of that only.

**10-Hyperparameter Tuning**

I will Do Hyperparameter tuning of my model so i can set some better result



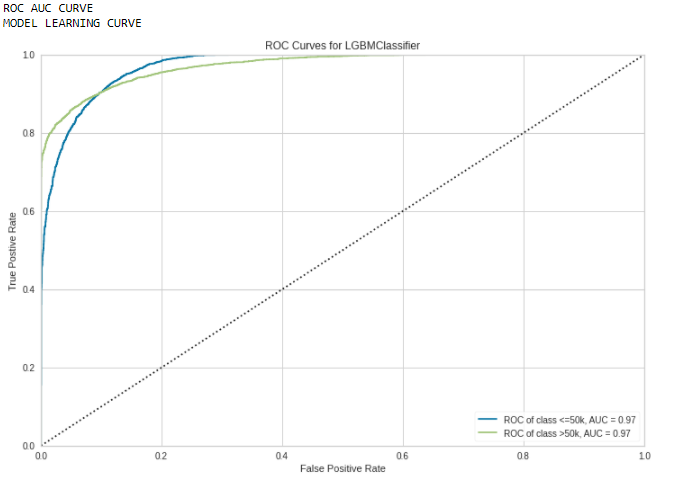
Here i am using some parameters so i can set some better result



The result i have increased the accuracy of my model and now i have new better parameters like boosting type is ‘dart’ and n\_estimator is 500 now.

***Final Model Performance Result and Roc-Auc Curve and Model Learning Curve***



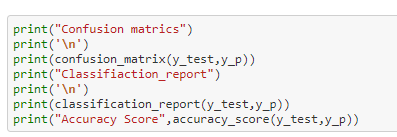


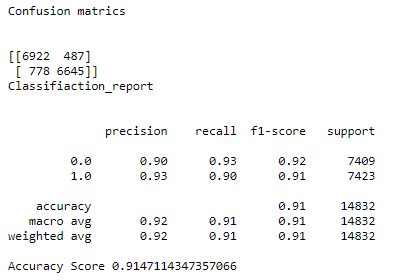
Final Model RUC-ROC Curve



Final Model Learnign Curve

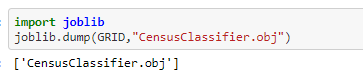
***Final Model Metrics-***





**Final Model Score And Performance Metrics**

***Saving The Model***



Saving the model for future use by using joblib it will save my model in obj format

**Model Deployment-**

**Conclusion**-*In this blog we have learned how to make an end to end machine learning project.*

for more code details and for model deployment code please go through this link-https://github.com/tanyakale993/Evaluation-Projects/blob/67f195416e86519188564e52ce626d734cf6c16e/CENSUS\_INCOME\_PROJECT.ipynb

Blog made by – Tanya kale