Project Report ON Census income

The task is to predict whether a person makes over $50K a year or not from Census data.

**1. Problem Definition:**

In this project we had to detect whether a person makes over $50K a year or not based on that person’s demographical, Education and Financial details. This is a binary classification problem. We have to use the dataset that was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker.In this Census dataset there are many features like Age, Workclass, Fnlwgt , Education and more.

**2.  Data Analysis:**

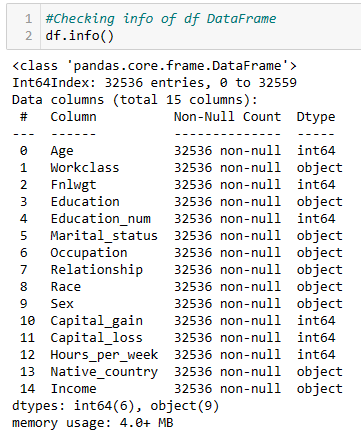
For Data Analysis we have to understand each column’s data and it’s characteristics. In Data analysis part EDA is a very important part and it’s help us to understand every feature’s characteristic and give us a better understanding of our data statistically and graphically. In Exploratory Data Analysis we performed lot of tasks like:

**i) Cleaning Data:**

We checked for duplicate rows but in my case my dataset doesn’t have any duplicate rows so I don’t have to delete any rows.

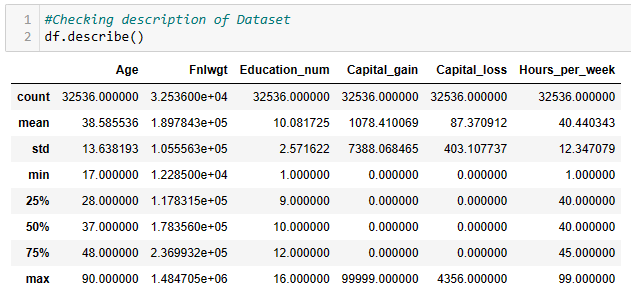
We have to check for null values if there is any, we have to fill those value or delete those according to our need.

Checked the unique values and looking for blank values also if there any we have fix those also.

****

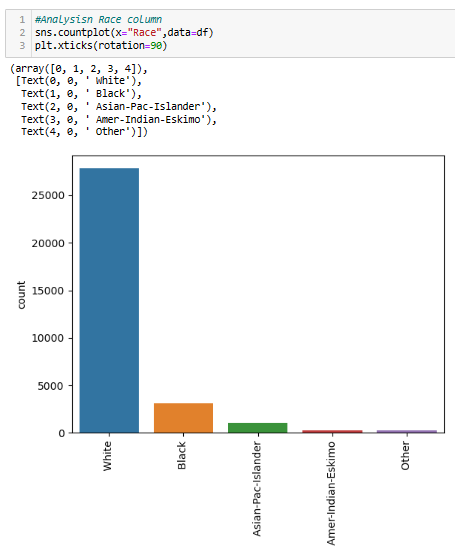
**ii) Data Exploration:**

For Data Exploration we can check the statistical summery(Mean,Median,Mode,Percentile) of all numerical columns and try to understand there characteristic. Also how those column’s data was distributed (Right skewed,Left Skewed,Normal).



**iii) Univariate, Bivariate and Multivariate Analysis:**

In this part we analyzed each column solely after that we analyzed that column with different columns data but max two columns at a time that called Bivariate analysis and we got clear understanding what relation both the columns have with each other and after that we go for Multivariate analysis where I analyzed three or more columns.



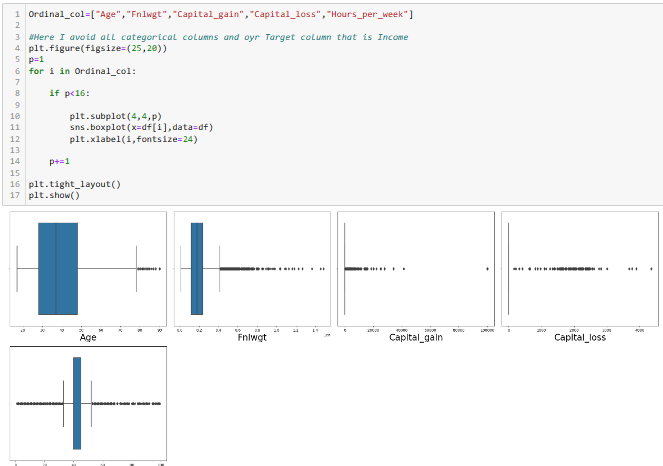


**iv) Visualization:**

In this part we had use different visualization techniques like Count plot, Scatter plot, Line plot, Bar and many other to see the data visually.

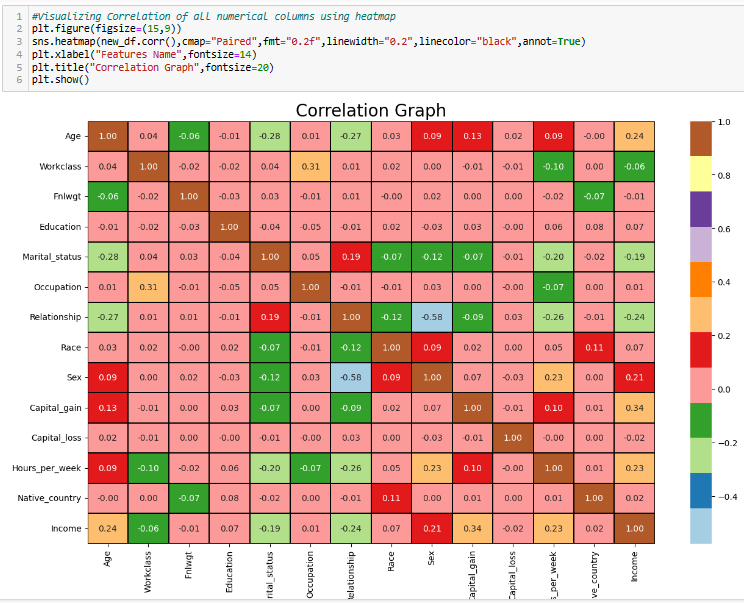
**v) Checking for outliers and remove the outliers:**

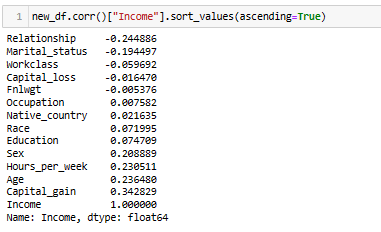
In this part we detect the outliers and after the detection we can remove them with different techniques like zscore or IQR.I can go with any technique where data loss percentage is less after removing the outliers.



**vi) Checking the correlation:**

In this step we can check all column’s correlation with each other where if the correlation coefficient value goes towards 0 then we can consider those two columns have weak correlation and if two column’s correlation coefficient value goes towards 1 then we can consider those columns have the strong correlation.





**vii) Documentation:**

Certainly, Documentation is an essential part for any project. From documentation we can gained knowledge, contains process that we followed, the decision we made for certain condition and also documentation can help us for our future reference. For example, in this project we can documented each of the stage like Project overview, Data description, Data distribution, Data cleaning, Data preprocessing, featured engineering, Data visualization, Encoding categorical columns, Checking multicollinearity, Splitting the data for train and test phase, Cross validation, Selecting the best model, Hyperparameter tuning, Checking my model works or not.

We can get many benefits from the documentation like:

1) We can preserve our knowledge.

2) Documentation can help others like if I documented my project then any one who can try to understand what techniques I applied they can easily know that.

3) We can share this documentation for a team and they can understand what they have to change or pick from this project.

4) From documentation we can improve constantly.

5) Anyone can audit the project and can understand easily because of documentation.

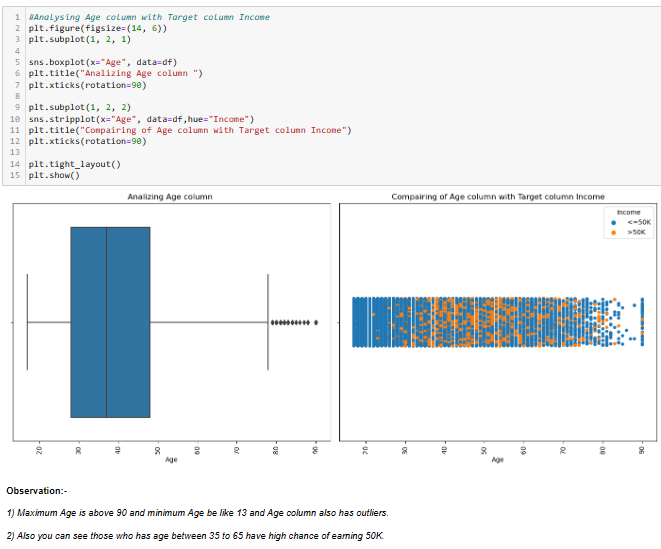
**3.  EDA Concluding Remarks:**

EDA revels the potential of a person that he can earn more than $50k or above in a year or not age wise, Education wise, Occupation wise, Country wise etc. Understanding these pattern helps us for feature selection. From my project we can see a various key factors for the prediction of whether a person makes over $50K a year or not. Those are:

**i) Income Distribution:**

Income distribution plays a vital role for the prediction that a person can make more than $50k or above in a year or not. We can inspect the Income distribution among the different age group, education and occupation.

Also, age group can be a crucial point for our prediction.



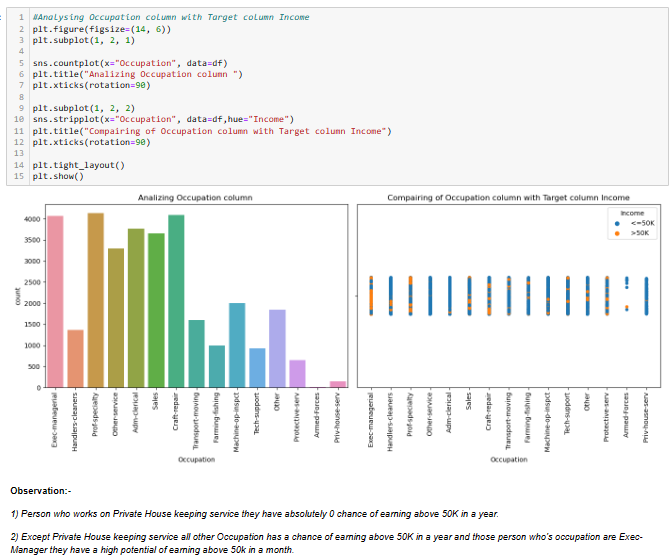
**ii) Education:**

Education also can play a vital role for the income. If the person have higher education he has a chance to get higher income.



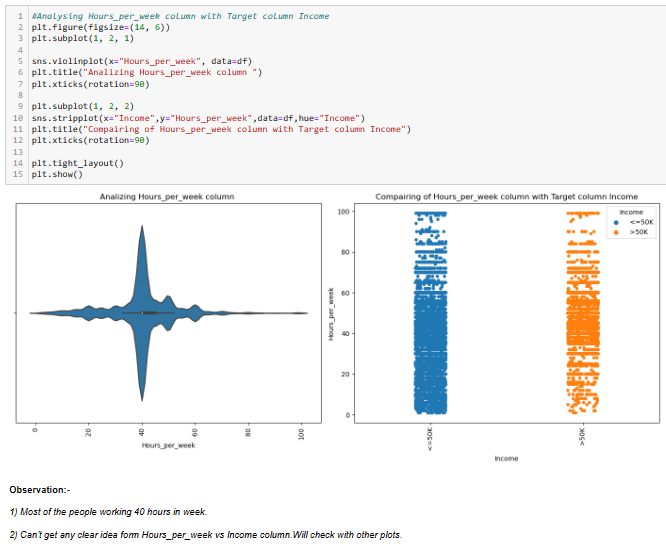
**iii) Occupation:**

Also, occupation is also a very good factor for prediction like the people who choose low skill jobs they have very low chance for making high volume of income.



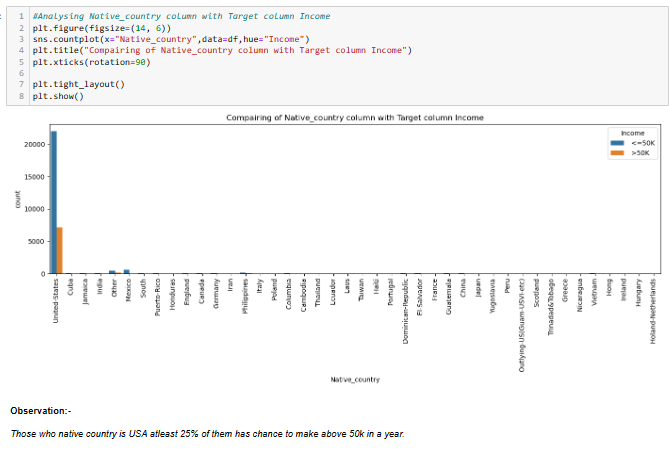
**iv) Work hours:**

Working hours are also matters for the income. In maximum cases if a person work more time then he gets more income but there is some issue also because like if a person do low skill jobs and do extra time also he can’t make that amount of money like a person who is in a high skill job.



**v) Country:**

Country is another crucial feature for the income because if you see most of the person who making $50k or above in a year most of them are from USA.

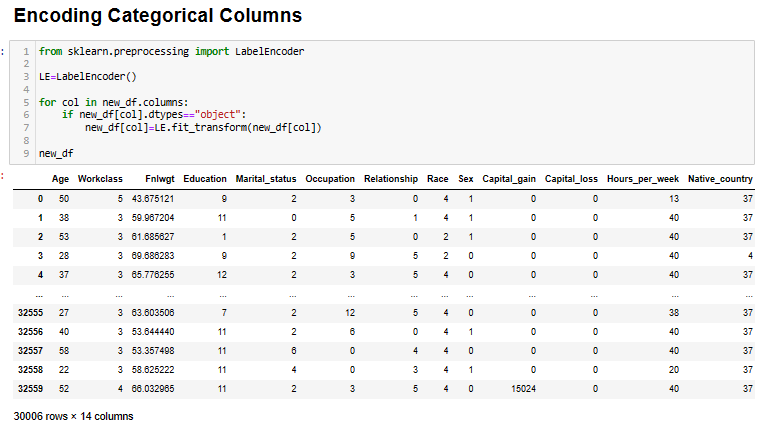


**4.  Pre-processing Pipeline:**

Data pre processing is also a crucial step for our machine learning models. It ensures that our dataset would be suitable for our machine learning. In data pre processing we had use different methods that would be described below:

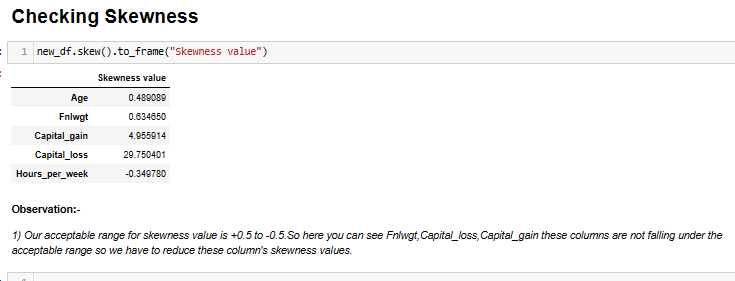
**i) Encoding:**

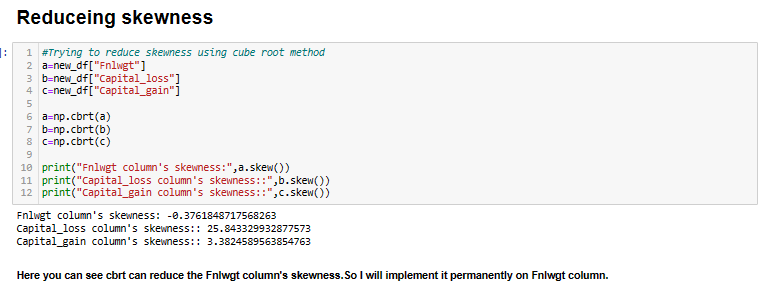
In this step we converted any column’s categorical values with numbers. There are multiple encoding techniques like OneHotEncoder,LabelEncoder,OrdinalEncoder. In this dataset I use LabelEncoder according the need of my dataset.



**ii) Check skewness and try to reduce it:**

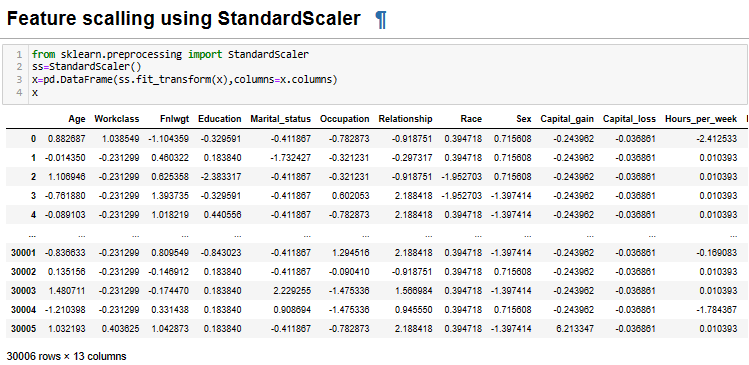
We know our acceptable skewness range is +0.5 to -0.5. So, if we found any column’s skewness above or below of this range then we have to try some transformation techniques to reduces this skewness. We can use different transformation techniques like cbrt, sqrt, log, log1p, PowerTransform, boxcox. But we have to take care of certain things like boxcox can not work on negative skewness column. And if I apply all the transformation one by one for a column and after used all the transformation on that column and still it’s skewness did not come under our acceptable skewness then we have to consider that the earlier skewness of that column’s skewness is it’s default skewness and we have to accept that.





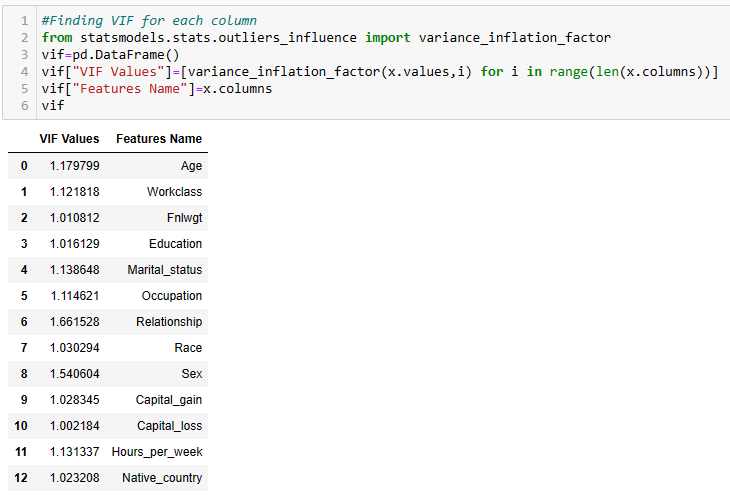
**iii) Feature scaling:**

Feature scaling is a preprocessing technique that standardize the independent variable. The main goal of the Feature scaling is it brings all features values in a similar scale. And by using Feature scaling method my model can train easily. For an example of Feature scaling, we can understand this with an example like in weight column if there are two values like one is in grams and one is in kilo gram then machine could get confused so in this picture Feature scaling can play a vital role that standardize both value in scaling and machine learning model can understand this very easily.



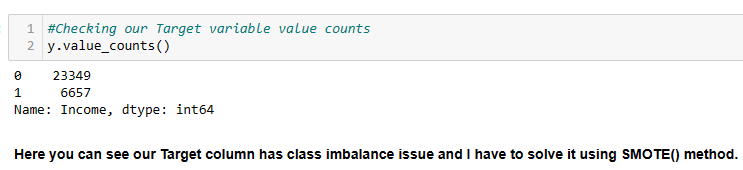
**iii) Checking for multicollinearity:**

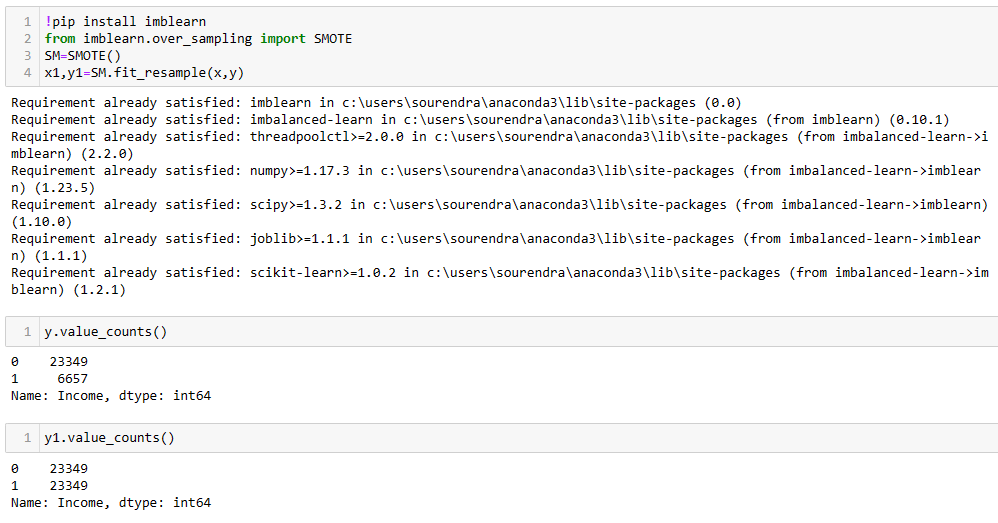
Multicollinearity collinearity occurs when two or more features highly correlated with each other. VIF is a common metrics that we use to check the multicollinearity. Below 10 is our acceptable VIF value range. Suppose we get some columns VIF value is above 10 then from those columns we have to delete that column which has the highest VIF value and after that we have to check the VIF value again for rest of the column. This process we have to continue until all the column’s VIF value range is in our acceptable range. This method is not applicable when we have less features. If you have less features and still you have some features that have VIF value over our acceptable range then we have to ignore those VIF values. In my case my all features have the VIF value below 10. So, I didn’t have to delete any columns. Also addressing multicollinearity can improve our model.



**iv) Checking imbalance issue for Target variable:**

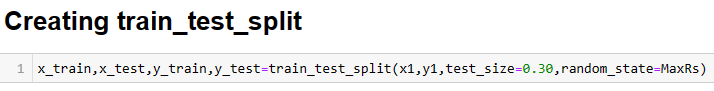
Imbalance class means one of the class significantly outnumbers the other class so in this case there is a high possibility that my model can generate a biasness towards the majority class. And this biasness can create poor performance while prediction. Imbalance dataset can create huge problem for our model and it can not predict accurately for that reason. By preventing the imbalance, we can train our model fairly and getting the prediction accurately. So that is why solving imbalance ness is very important. We can solve this imbalance issue by using SMOTE() method.



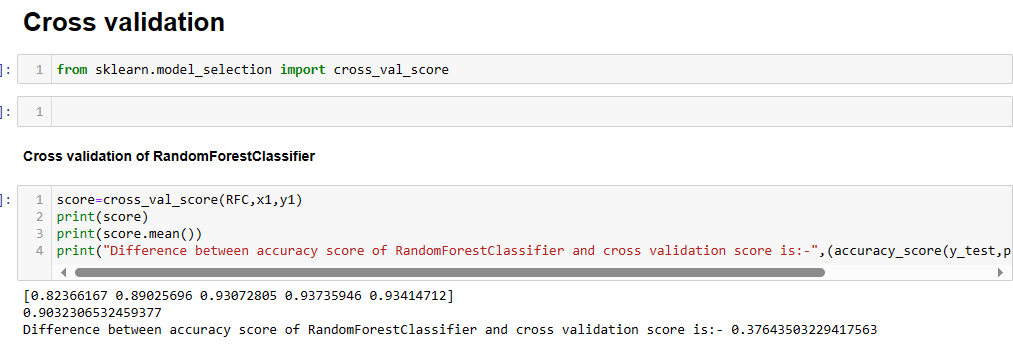


**5) Building the models:**

As we know from our problem statement this is a binary classification problem because here, I have to predict that a person can capable of making $50k or above in a year or not. So now we have to create several classification models like logistic regression, ExtraTreesClassifier, AdaBoostClassifier, GradientBoostingClassifier, DecisionTreeClassifier, KNeighborsClassifier, SVC and ensemble methods like RandomForestClassifier, BaggingClassifier. After creating model, I have to evaluating some matrics to judge the model like Accuracy, F1 score, Recall, Percision. While creating the model I have to set a portion of data for training like for me I set it 30%.



After checking the all model’s accuracy score I can not decide the best model because those accuracy are not accurate and I have to cross check it with cross\_val\_score. After that I have to see the difference between each model’s accuracy with it’s cross\_val\_score difference. From those models which have less difference between it’s cross\_val\_score with it’s accuracy I have to choose one of them.

****

**6) Hyperparameter tuning:**

So, after choosing the best model we will go for hyperparameter tuning. Hyperparameter tuning is the process where we can find the best set of parameters to improve the model performance. The goal of hyper parameter to find different combination of parameters value and find the best set of parameter’s value so that the model’s prediction accuracy would get increased. This process is very much required because hyperparameter can do significant impact on model’s ability to learn from the data and make the accurate prediction.

There are lot of methods of doing hyperparameter tuning but commonly we use only two of them i) GridSearchCV and ii) RandomizedSearchCV.

**i) GridSearchCV:**

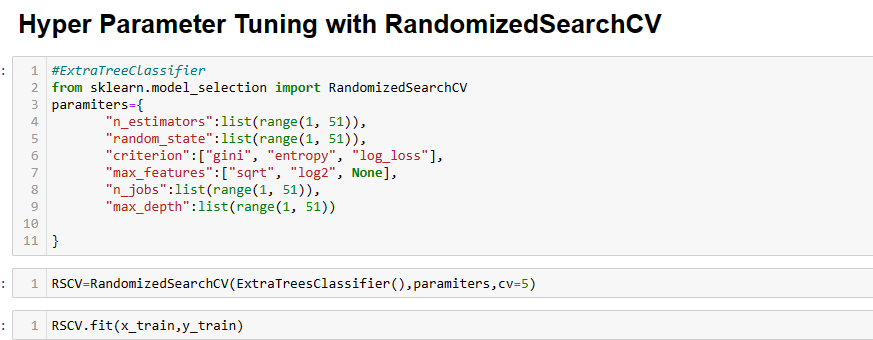
Grid Search searches all parameters thoroughly and evaluate the model performance using the cross validation for each combination and select the best parameters values combination with the best performance.

****

**ii) RandomizedSearchCV:**

Randomized Search involves randomly sampling parameters that we can use for our model. But it does not cover all best combinations of parameters value like GridSearchCV. But it takes lesser time than GridSearchCV.

So, in conclusion we can say Hyperparameter tuning is a crucial step for model development step. It helps to avoid the underfitting and over fitting and ensure that model will do well for unseen data.



**7. Concluding Remarks:**

The final phase is to check the model result. I can see my model can do good predictive performance accurately and can predict whether a person makes over $50K a year or not based on the person’s data. Analyzing give an instance that which variable contributes significantly towards our prediction.

So, in this model we use machine learning algorithm that can predict the income potential of predicting the persons capability that whether a person makes over $50K a year or not based on the Census data. This model can used by Data scientist, economists, Load distribution companies, Policy makers, Investment companies, Policy makers to understand their customers or client’s income distribution. But there are some limitations like if the if the world financial condition changes, then this model can’t predict accurately and there you can see some biasness.

If you need you can see this model’s code and documentation for this project on git hub. Here is the link:-

<https://github.com/Soumendra007/Evaluation-Project/blob/main/Third%20Phase/Census%20Income.ipynb>

Thank you for reading the whole project. Hope you enjoyed the detailed explanation for all the methods. That’s it from my side. Thank you once again.

**The End**