Census Income Prediction Using Machine Learning

Hi Everyone….I am Abhishek and Today I am gonna give a quick 6 steps guidance on how to predict census income using machine learning.

Introduction:

So now I am doing a complete analysis of census income project. Here I will perform all major tasks that are necessary to make a machine learning model, I am going to do complete analysis from EDA, visualization to model building and finding the key observations from the analysis, that will help us to predict the income.

So, following are the required steps needed for our project:

1.Problem definition

The problem statement is explained below:

Description of fnl wgt (final weight)

The 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:

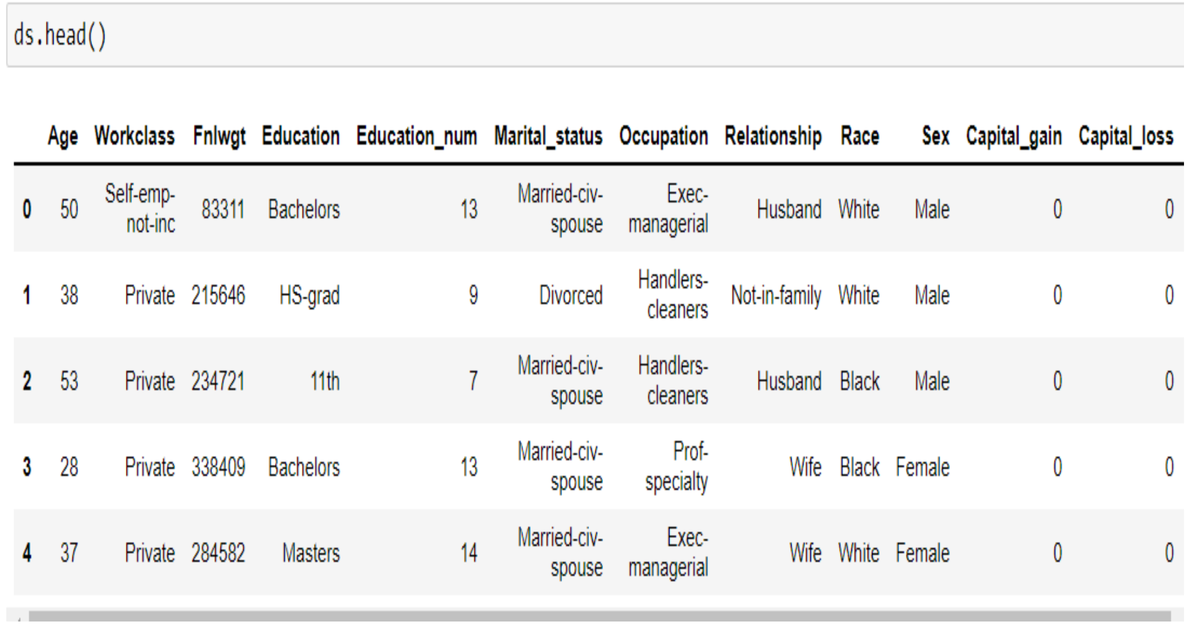
1. A single cell estimates of the population 16+ for each state.
2. Controls for Hispanic Origin by age and sex.
3. Controls by Race, age and sex.

We use all three sets of controls in our weighting program and "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. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.

Now we are clarified with the problem statement so let's proceed further.

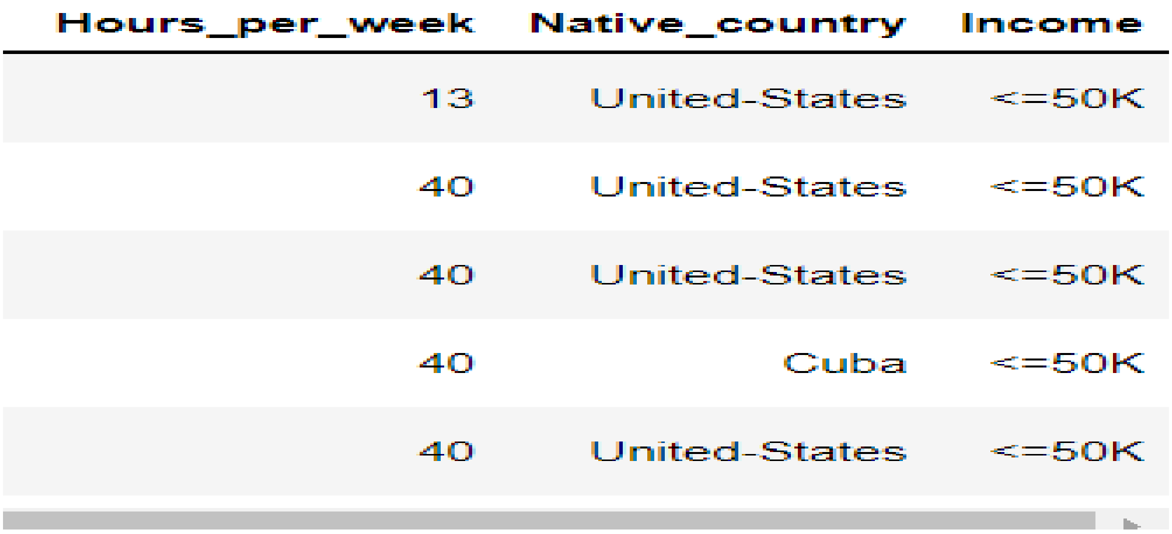
2. Data Analysis

Now we are just analyzing our dataset, to make the picture clear I am sharing a picture of the dataset containing first few rows, now let's see what insights should we gain from it. So here is the dataset as we can see, I have divided the table in two pictures as we have so many columns in our dataset.



```
ds.head()
```

	Age	Workclass	Fnlwgt	Education	Education_num	Marital_status	Occupation	Relationship	Race	Sex	Capital_gain	Capital_loss
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0



Hours_per_week	Native_country	Income
13	United-States	<=50K
40	United-States	<=50K
40	United-States	<=50K
40	Cuba	<=50K
40	United-States	<=50K

- We have huge dataset as we can see that we have 32560 rows and 15 columns.
- All the column are named as 'Age', 'Work class', 'Fnlwgt', 'Education', 'Education_Num', 'Marital_status', 'Occupation', 'Relationship', 'Race', 'Sex', 'Capital_gain', 'Capital_loss', 'Hours_per_week', 'Native_country', 'Income'.
- We check the properties like shape, unique, dtypes etc.
- We have Income column as our target variable, it has two classes <=50k and >50k thus it is logistic regression problem.
- Now I will do the further analysis according to our problem type, which is logistic regression being a type of Supervised learning.
- So, let's do the further analysis.

3.EDA Concluding Remarks

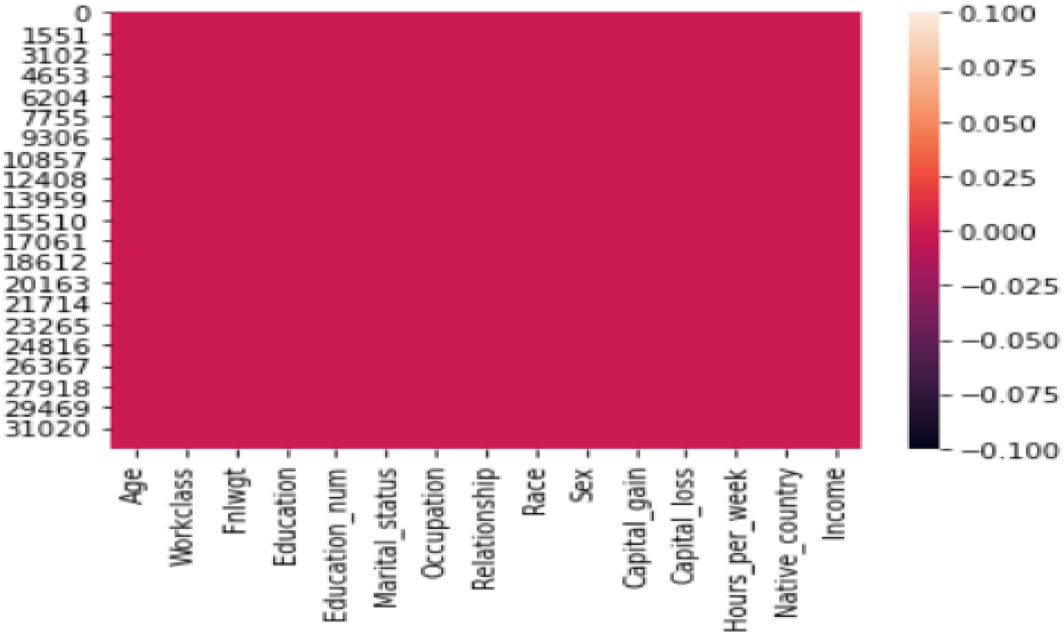
Now we will do the complete EDA of the dataset. By following these steps:

1) **Check missing values**
2) **Statistical Summary**
3) **Univariate Analysis**
4) **Bivariate Analysis**
5) **Multivariate Analysis**
6) **To check Skewness**

Let's discuss all the above steps in detail:

1) Checking null values

We can see that whether null values are present in the dataset or not, we can also use heat map like I have done.



From the heat map we can see that no null values are present in our dataset.

2) Statistical Summary

Statistical summary gives information about the mean, median, std, min, max etc.

key observations:

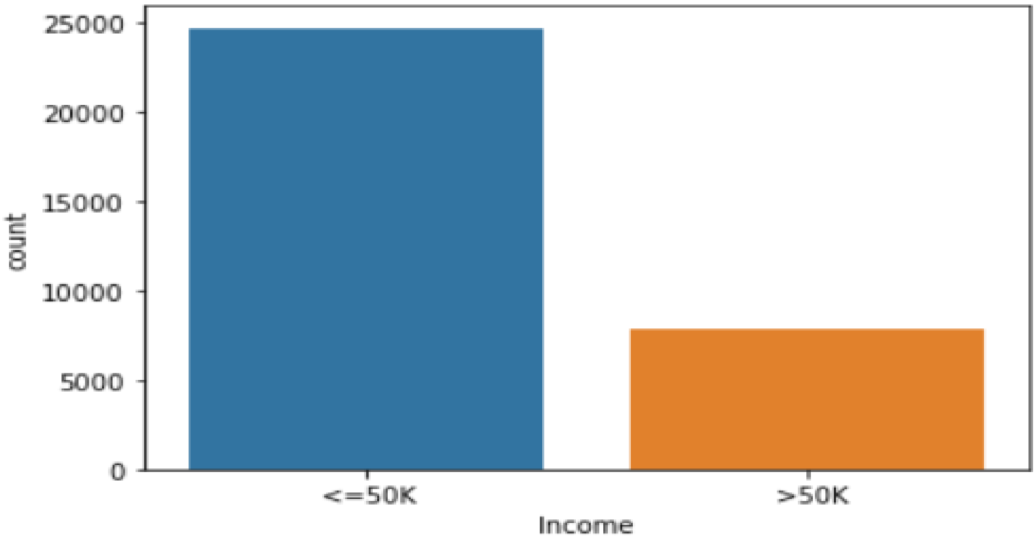
There is a small difference in mean and 50% percentile median.

There is a huge difference in 75% percentile and max column in many columns like Age, Fnlwgt, education_num, Capital_gain, Capital_loss, Hours_per_week.

From the above observations we can say that extreme outliers are present in our dataset.

3) Univariate Analysis

Since we are working on a classification problem, we plotted count plot to see the proportion of each type of class.



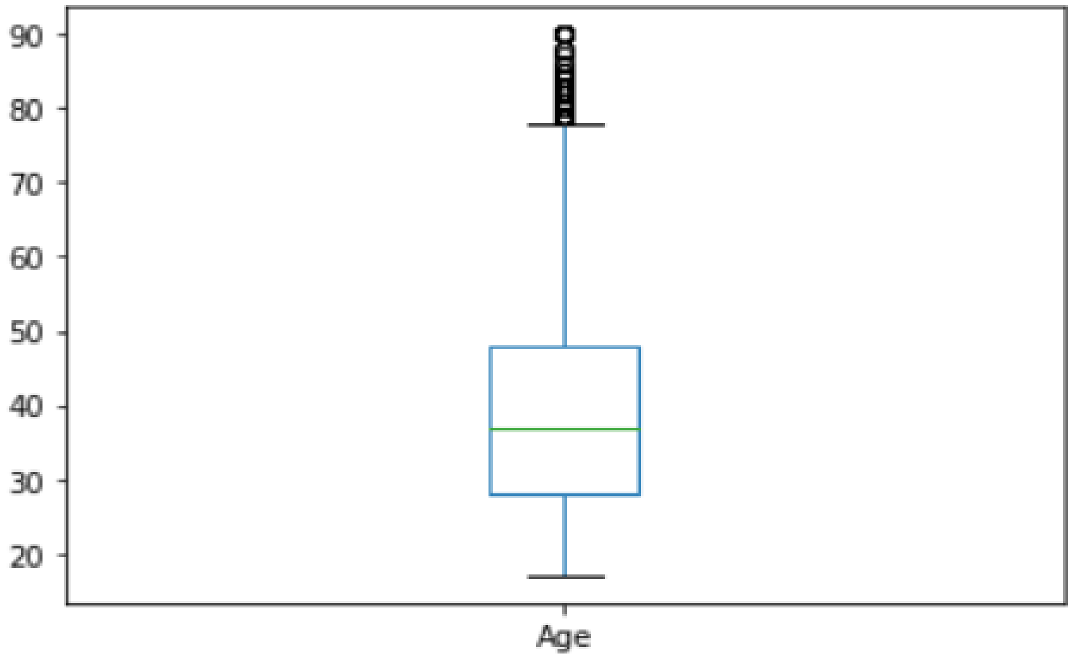
Here we can see that class imbalance is present in our dataset.

Another example of univariate analysis

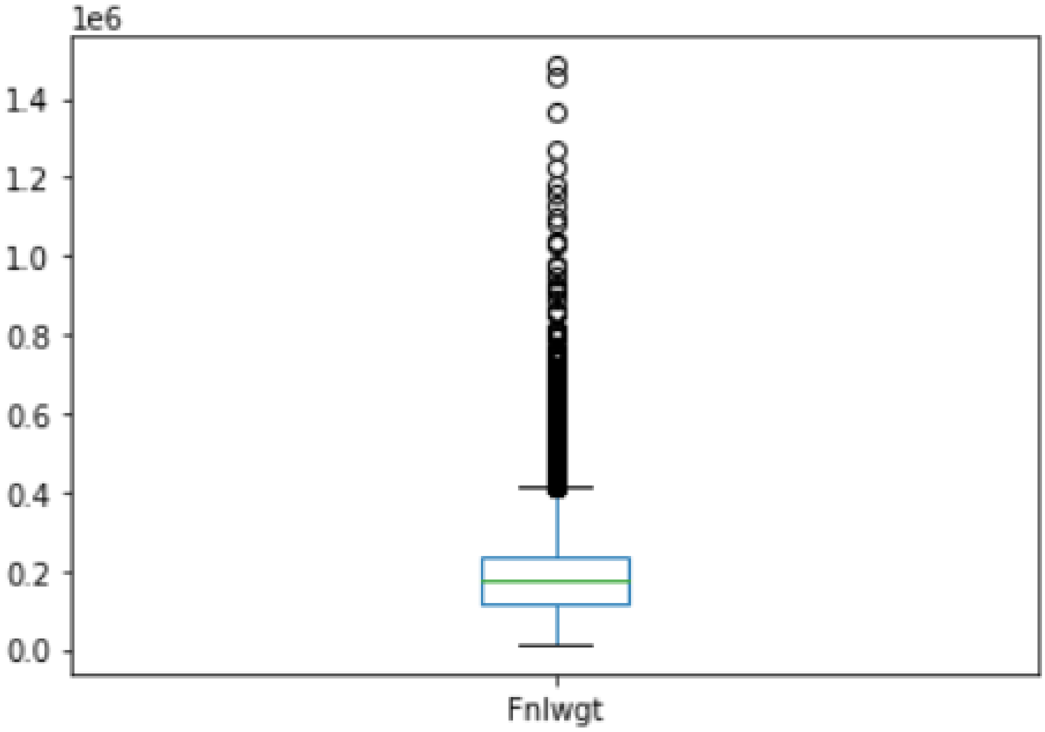
Univariate Analysis

In univariate analysis I have plotted box plots, from these plots we can see the mean, median, max, min and we can also see whether outliers are present or not, as we can see in below picture, we have outliers in many columns like Age, Fnlwgt,

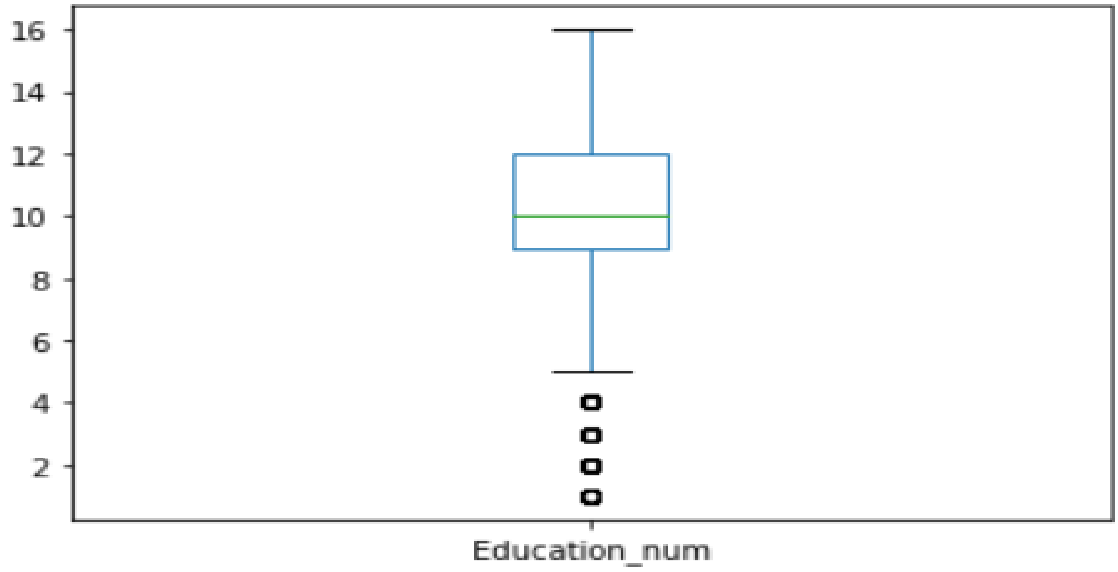
Education_num, etc. are shown in the below plot.



Here we can see that many outliers are present, which shows the ages of the peoples which is more than average.



Here in this column, we can see that extreme outliers are present.

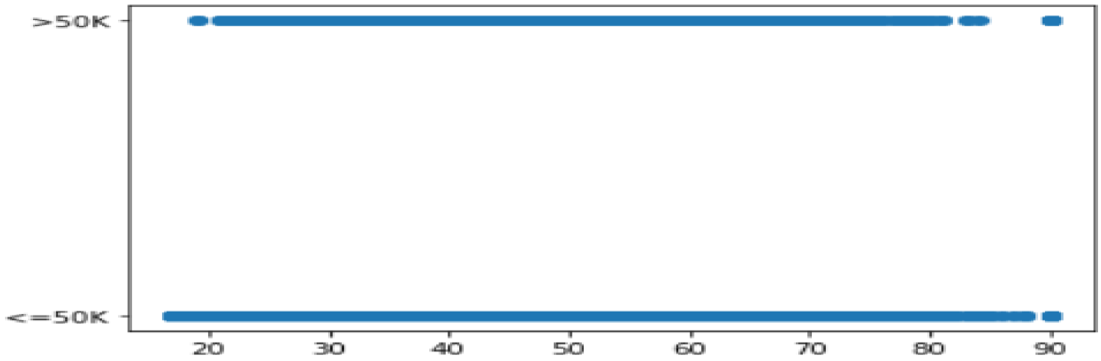


Here also we can see that outliers are present.

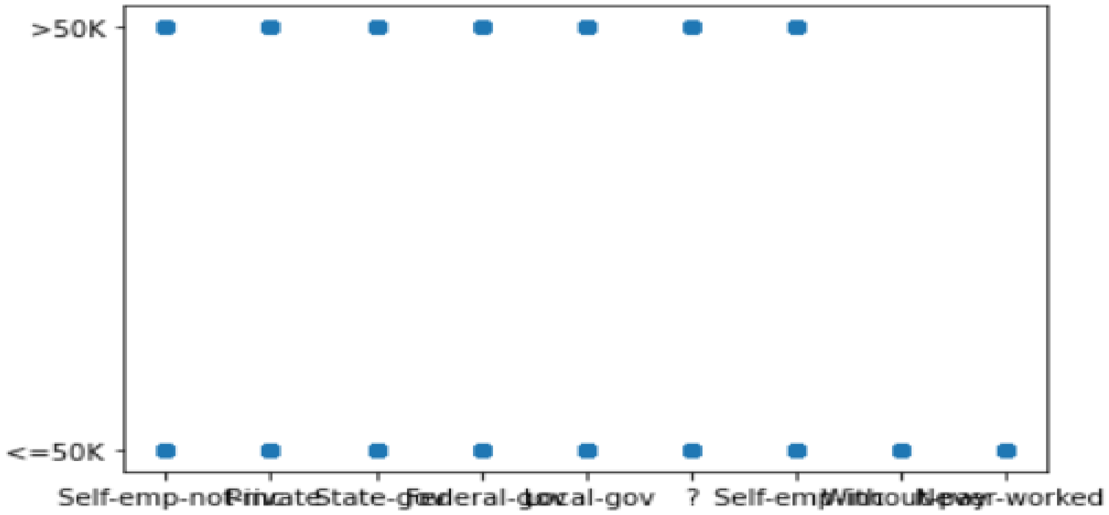
4) Bivariate Analysis

Now in Bivariate analysis I have used Scatter plot to see the relation of each column with the Income column.

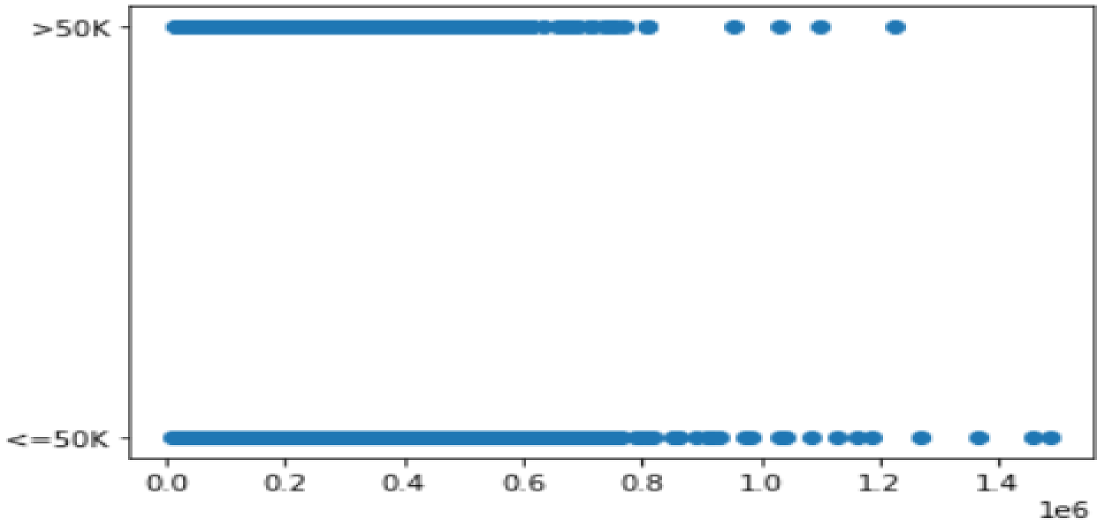
We can see the scatter plot of few columns like age, work class, fnlwgt etc. In the below image.



Here we can see the income of the peoples according to their ages in both classes also some outliers are present.



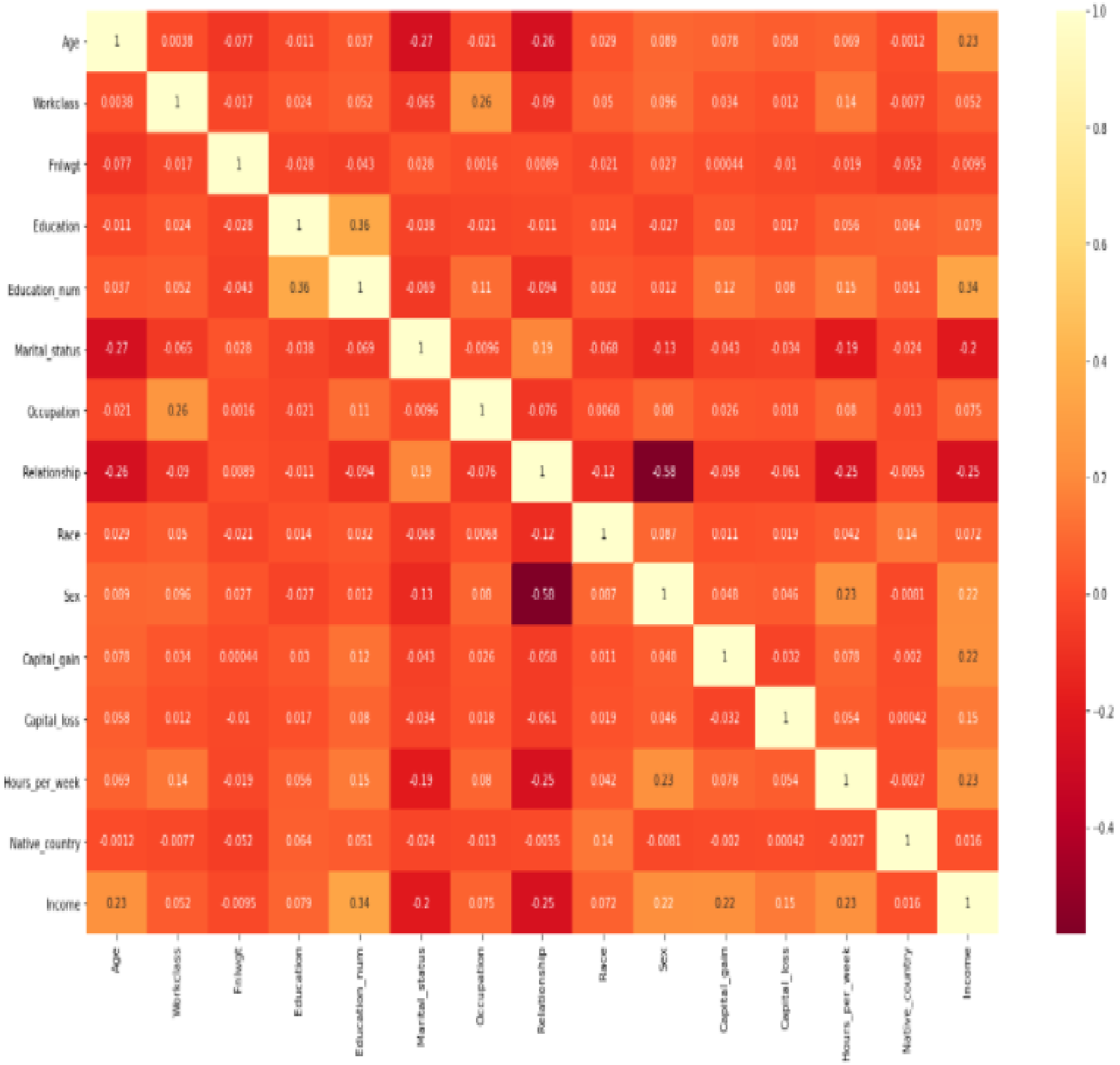
Here in the work class column we can see the work class of each class peoples.



Here in this column, we can see the relation of final weight with the Income, we can also see the outliers.

5) Correlation Matrix

In multivariate analysis we can check the correlation matrix of the dataset, which clearly shows us the relation of each column with other column, here is the correlation matrix.



Observations-

Light shades are highly correlated

So, we see that Education_num column is highly positively correlated with correlation value 0.34.

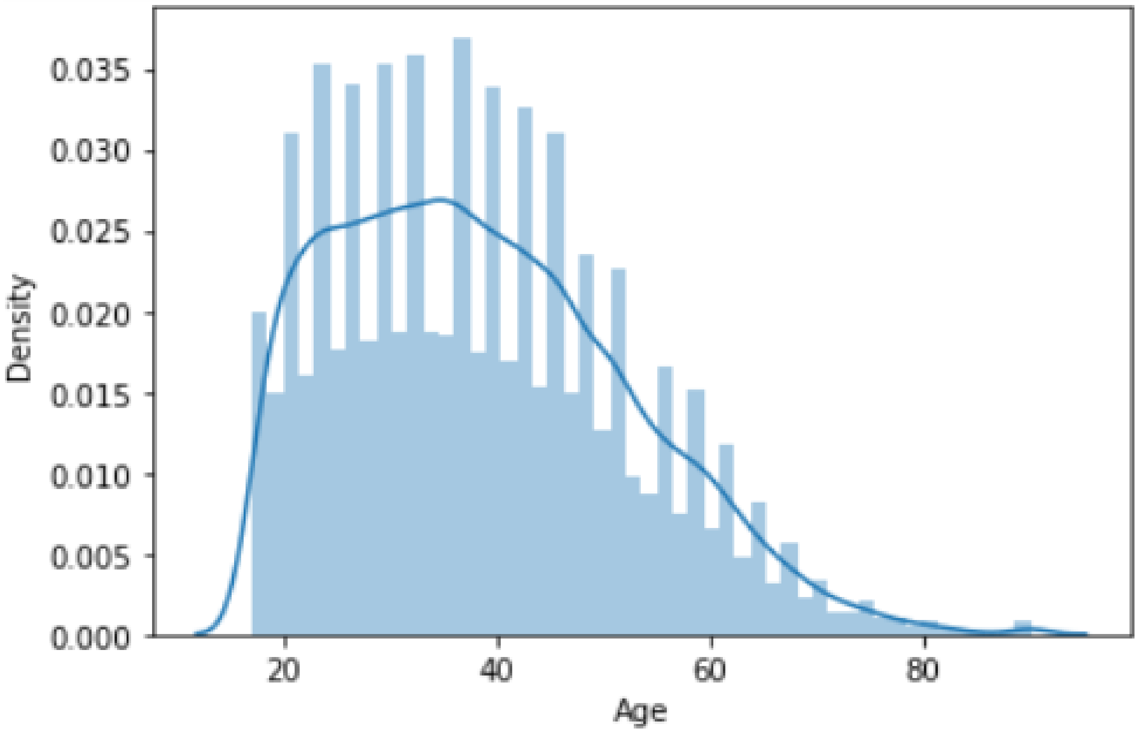
We can see that age, sex, capital_gain and hours per week column are also positively correlated with income column.

We can see that marital_status and relationship columns are negatively correlated with income column with correlation value -0.2 and -0.25.

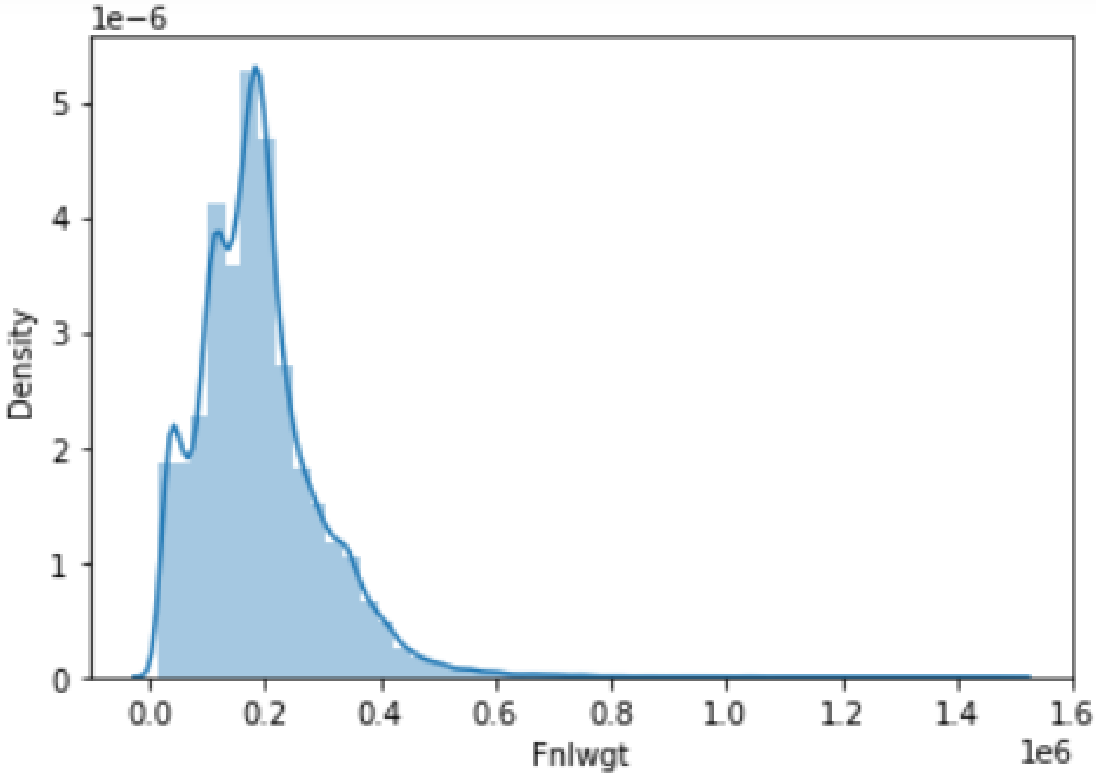
Some other columns are also slightly positively correlated with the income column.

6) Checking Skewness

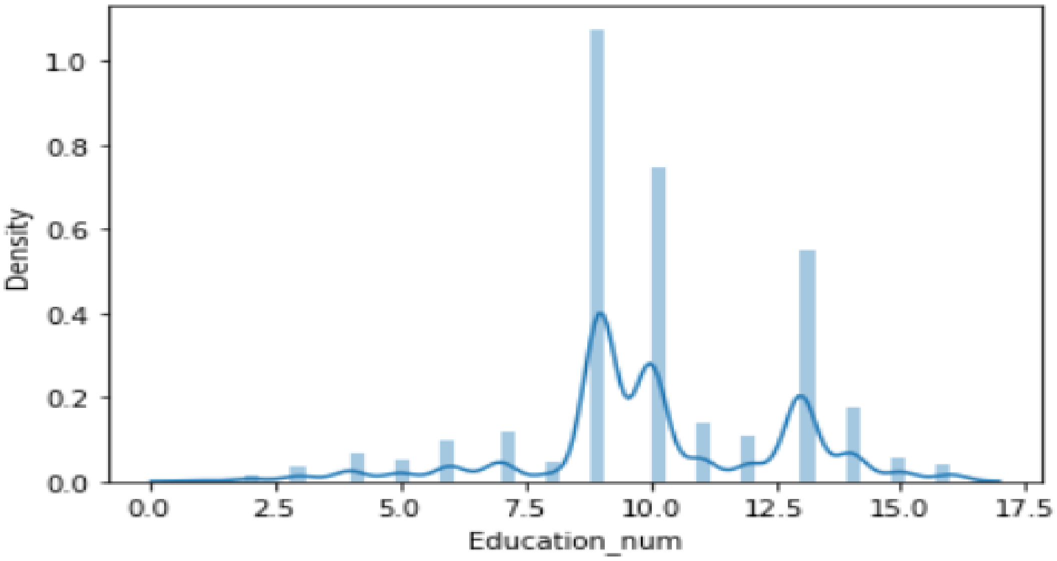
Now I have checked the skewness of each column by plotting density plot, to check whether the curve is normally distributed or not, how much skewness is present? We have skewness in many columns let’s see in age, Fnlwgt and education_num column.



Here we can see the skewness in the curve it's not normally distributed.



Here also we can see the skewness in the column Fnlwgt.



Here in this column also we can see the skewness. We have skewness in many other columns also.

4.Pre-processing pipeline

We have many steps included in pre-processing like Data cleaning, Data reduction, Data integration etc. Let's discuss each of them in detail the steps we have done in our project according to our requirement.

1) We have drooped the rows which are negatively correlated, here we have dropped the marital_status and Relationship column from the dataset due to its highly negative correlation with the income column. This column impacts our data negatively, thus we dropped that column.

2) Now we are removing the outliers present in our data, we have two methods to remove outliers one is zscore and other is by using IQR method here I have used zscore method to remove all the outliers present in our dataset, we have many outliers in my dataset due to this many rows are deleted from the dataset as outliers.

3) Here we have seen above that no null values are present in our dataset, so here is no need to handle missing values, if there were any missing data then we have to treat it with suitable method, but here is no need.

4) Next is that we check for whether any column is present in string format or not, if any column is present in string format, then we have to change it in integer format by applying Encoding technique, we have two methods in encoding one is One hot encoding and other is Label Encoding. Here in our dataset, we have many columns like work class, education etc. are in string format so we encode them by using label encoding technique. After using encoding we have all the columns in integer format.

5) So now we are removing skewness from our dataset, as we have seen that skewness is present in all the columns except the sales column, to remove the skewness we have separated the target variable and the independent variable from the dataset. As we know that skewness between –0.5 to 0.5 is acceptable but more than it is not, so skewness except this range should be treated by using suitable method.

Here I am using power transformer method to remove the skewness from the columns. Till now we have treated the skewness in our column, now skewness is completely removed from our dataset.

6) Next point which comes is the feature engineering this is used when we have special characters in our columns, etc. But here we have no need to apply feature engineering because no such characters are present in our dataset.

7) The other thing is Handling the class imbalance problem; we have two methods to handle class imbalance first is oversampling and other one is undersampling we can do oversampling using SMOTE and undersampling by using NearMiss.

In our dataset I have used oversampling method and balanced the class.

8) The next thing which I can see is the **standardization** technique, we use this technique to scale our data. We have two methods to scale our data first one is standard scaler and the second one is min max scaler. We use these techniques only when there is huge difference between the ranges of any 2 columns, that’s why we use scaling.

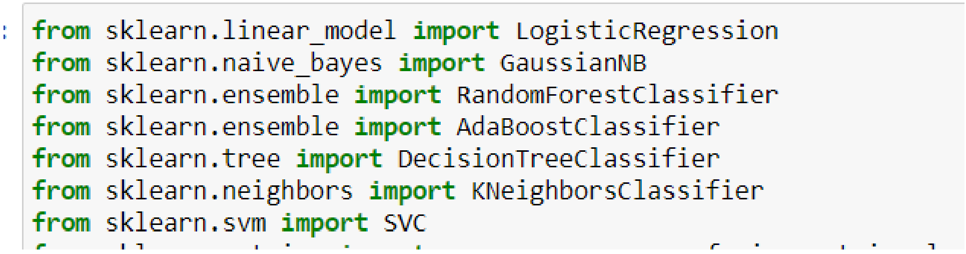
Standard scaler is used when data is normally distributed, its changes the mean=0, std=1 and the value ranges between –3 to +3.

Min –Max scaler is used when data is not normally distributed this method is also called normalization, it changes the data with mean=0, std=1 but range is 0 to 1. In our dataset we have seen there is no huge difference in ranges of the columns, thus here we have no need to apply standardization technique on our dataset.

9) The last thing which I can see is the PCA technique, this technique is used only when we have, we large numbers of columns and it's difficult to manage them all, but here in our dataset we have only 5 columns thus here we have no need to use this technique.

5.Building Machine Learning Models

Now we build a machine learning model, we will use multiple algorithms, as we know we are working on a classification problem so here we can only use classification models like logistic Regression, SVC, Decision Tree Classifier, K-Neighbors Classifier and evaluation matrix like Random Forest Classifier, Ada Boost Classifier and Gaussian NB. To use all these, we have to import each model from scikit learn as follows:



```
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
```

Firstly, we have to create train test split and thus we use train data for training our model and test data for testing our model performance. Here I have splited the data in 70% as train and 30% as test data. I have finded the best random state using the logistic regression model which is 42, I used this random state to train all the models. Now we use train data for training our model and test data for testing our model performance.

The accuracy score of the models shows the performance of the model if accuracy is more model is performing well, but if the accuracy is very low means model is not performing well.

Here we have seen that we are getting maximum accuracy 86% with Random Forest Classifier.

The next picture which comes around us is the **cross-validation technique,** as we know the score is also due to over fitting, thus we use cross validation method to come over it.

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to **validate** the model. ... The basic form of **cross-validation** is k-fold **cross-validation**.

Here in our model, when we checked the cross validation, I use cv=5 folds, the best cross-validation score is coming out to be for Decision Tree Classifier.

The picture which comes around us is the **cross-validation technique,** as we know the score is also due to over fitting, thus we use cross validation method to come over it.

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to **validate** the model. ... The basic form of **cross-validation** is k-fold **cross-validation**.

We are getting least difference in corss_val and accuracy is in Decision Tree Classifier.

Now we have used hyper parameter tunning to find the best parameter for our model by using GridsearchCV.

I have applied gridsearchcv on all the models and finded the best parameter for all the models and used these parameters in our model, now I am getting the best accuracy with decision tree classifier with(parameter=entropy).

Now we have concluded that decision tree regressor is the best model for our dataset.

So last step is to save our model to use it in future for predictions, we have two techniques for saving the model. First is using joblib and the second is by using pickle.

I have saved my model DTC using joblib, so that I can use it in future and predict the Income using this model.

The last thing we have to do is concluding remarks:

6) Concluding Remarks

Let 'see the conclusion in below points:

- The main goal of our project is to solve the problem and predict the census income.
- For this we used machine learning skills and solved the problem.
- We have done the complete analysis of the data using EDA, univariate, bivariate, multivariate, checking correlation, checking skewness, checking for outliers, checking for missing values by doing all this analysis we have collected the information about the data, whether it is skewed, having missing values or not etc.
- Next, I have done the pre-processing of the data and solved all the issues that we finded during EDA like Outliers, skewness, class imbalance etc.
- The last step is the model building I used classification algorithm and different evaluation matrix to prepare the models and finded decision tree classifier as the best model.
- At last, we can make predictions for census income using our model.

Author- Abhishek Behera