**Default of credit card clients**



***Table content:***

* Introduction
* Problem Definition
* Data Analysis
* EDA Concluding Remarks
* Building Machine Learning Models
* Concluding Remarks

**Introduction:**

A credit card is a thin rectangular piece of plastic or metal issued by a bank or financial services company, that allows cardholders to borrow funds with which to pay for goods and services with merchants that accept cards for payment.

The credit card industry is big business, and it is dominated by a handful of companies.

# According to Credit card market share statistics There were some 41 billion U.S. general purpose credit card transactions in 2018 (based on cards issued by the four major networks – Visa, Mastercard, American Express, and Discover), accounting for about $3.8 trillion in dollar volume.2General purpose credit card payments value had grown from $3 trillion in 2016 to $3.32 trillion in 2017.

And the total number of credit card transactions in the U.S. was40.8 billion in 2017, up from 37.3 billion in 2016, according to the 2013 Federal Reserve Payments Study.

**Dataset**

**The dataset for this project consists of default payments from a major credit card company in Taiwan and is available at** [**UCI Machine Learning Repository**](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)**. The data set consists of 30,000 instances and 24 attributes consisting of gender, education profile, marital status, age, history of statement balance, payment status and binary status of default ( 1 or 0).**

**Details of the Features:-**

# Features:

**ID: ID of each client**

**LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit**

**SEX: Gender (1=male, 2=female)**

**EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)**

**MARRIAGE: Marital status (1=married, 2=single, 3=others)**

**AGE: Age in years**

**PAY\_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months,8=payment delay for eight months, 9=payment delay for nine months and above)**

**PAY\_2: Repayment status in August, 2005 (scale same as above)**

**PAY\_3: Repayment status in July, 2005 (scale same as above)**

**PAY\_4: Repayment status in June, 2005 (scale same as above)**

**PAY\_5: Repayment status in May, 2005 (scale same as above)**

**PAY\_6: Repayment status in April, 2005 (scale same as above)**

**BILL\_AMT1: Amount of bill statement in September, 2005 (NT dollar)**

**BILL\_AMT2: Amount of bill statement in August, 2005 (NT dollar)**

**BILL\_AMT3: Amount of bill statement in July, 2005 (NT dollar)**

**BILL\_AMT4: Amount of bill statement in June, 2005 (NT dollar)**

**BILL\_AMT5: Amount of bill statement in May, 2005 (NT dollar)**

**BILL\_AMT6: Amount of bill statement in April, 2005 (NT dollar)**

**PAY\_AMT1: Amount of previous payment in September, 2005 (NT dollar)**

**PAY\_AMT2: Amount of previous payment in August, 2005 (NT dollar)**

**PAY\_AMT3: Amount of previous payment in July, 2005 (NT dollar)**

**PAY\_AMT4: Amount of previous payment in June, 2005 (NT dollar)**

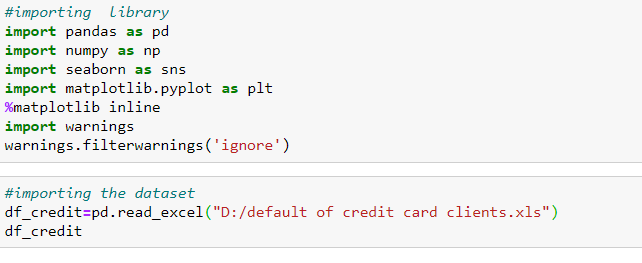
**PAY\_AMT5: Amount of previous payment in May, 2005 (NT dollar)**

**PAY\_AMT6: Amount of previous payment in April, 2005 (NT dollar)**

**default.payment.next.month: Default payment (1=yes, 0=no)**

***Data Analysis***

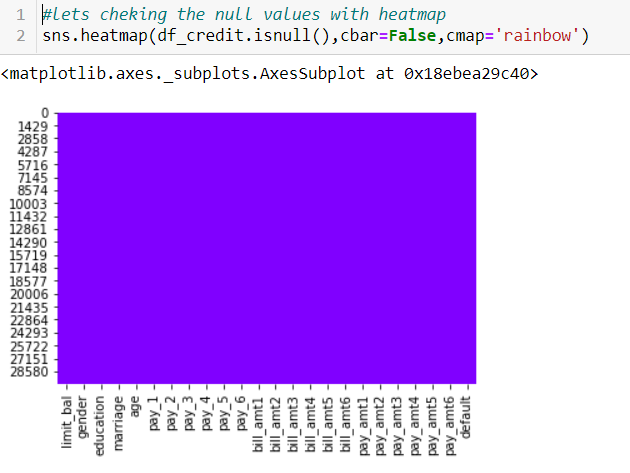
**First of all, I read data from xls file and created a data frame to store the data and created df\_credit table to store the data.For this, I wrote the following code:**

****

**Next, I Dropped the ID columns because That one is use less for us.**

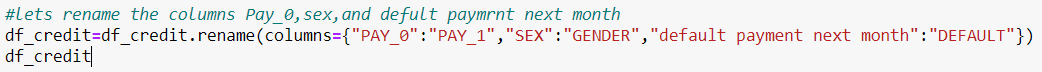
****

**Next ,I checked the null values is there or not,**

****

**From the above I got the conclusion that I don’t have any kind of NaN values in This Data**

**Next,I Rename the columns :**

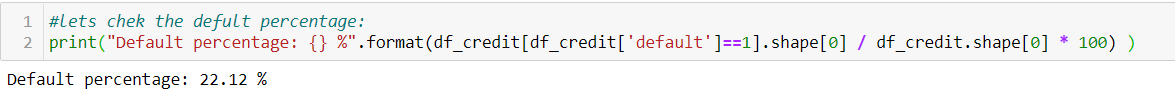
****

After that I make all columns name to lowercase for Handling the data easily .

Next, I did the value counts of each columns:-

****

Next, checked the Default %:-

****

From the above result we got to know that there is default Percentage is 22.12%.

# Next , I did some Statistics Summaray:

# 

# I got the info from the above chart:-

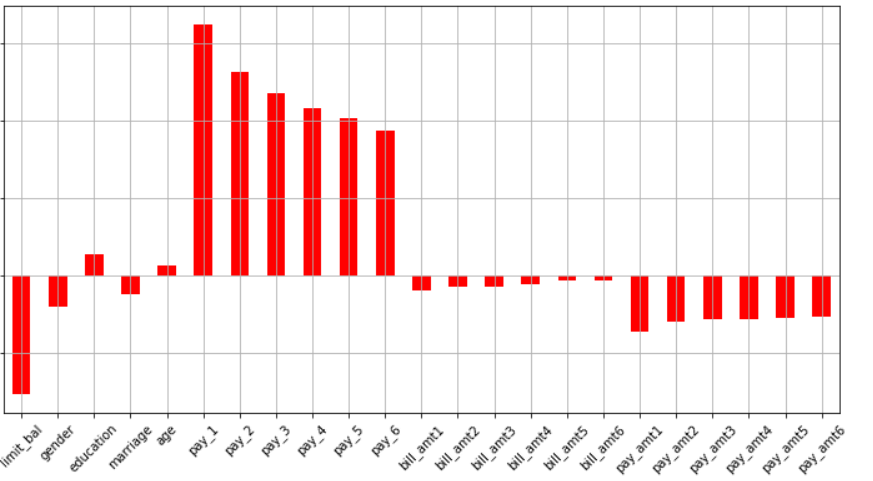
**minimum Limit balance is 10000 and maximum limit balance is 1000000**

**minimum age is 21 years and maximum age is 79 years**

we have plenty outliers in payments and billing colloms.

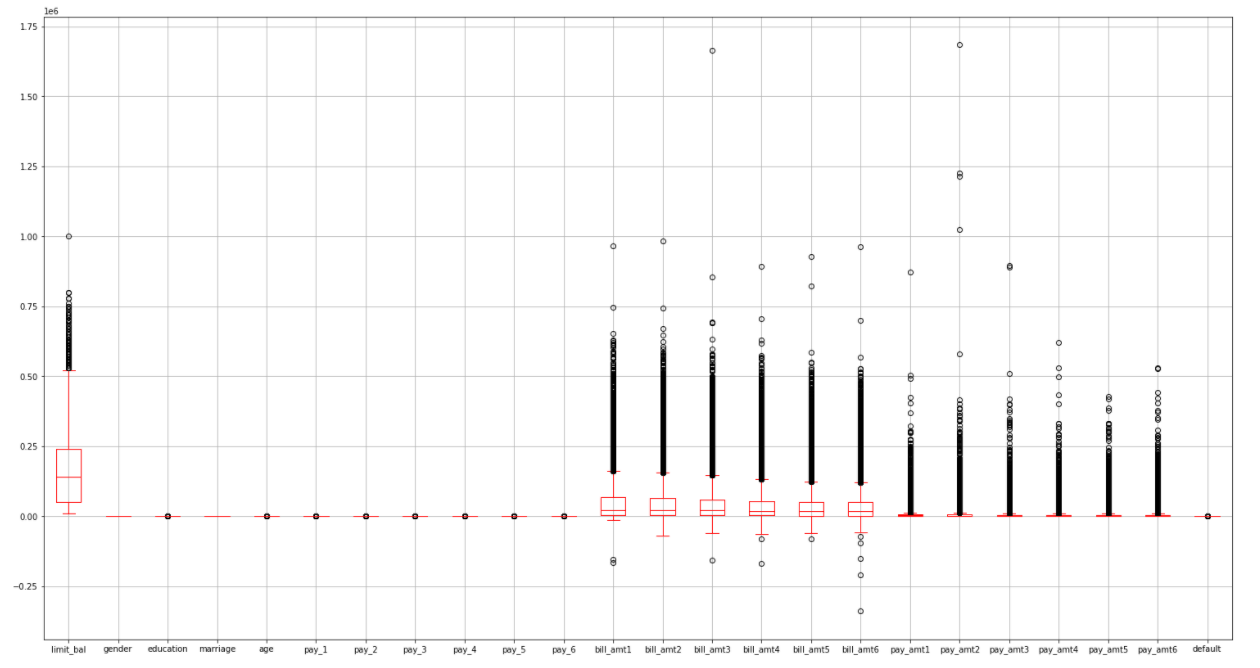
# Next I checked the correlation of each columns:-

# 

****

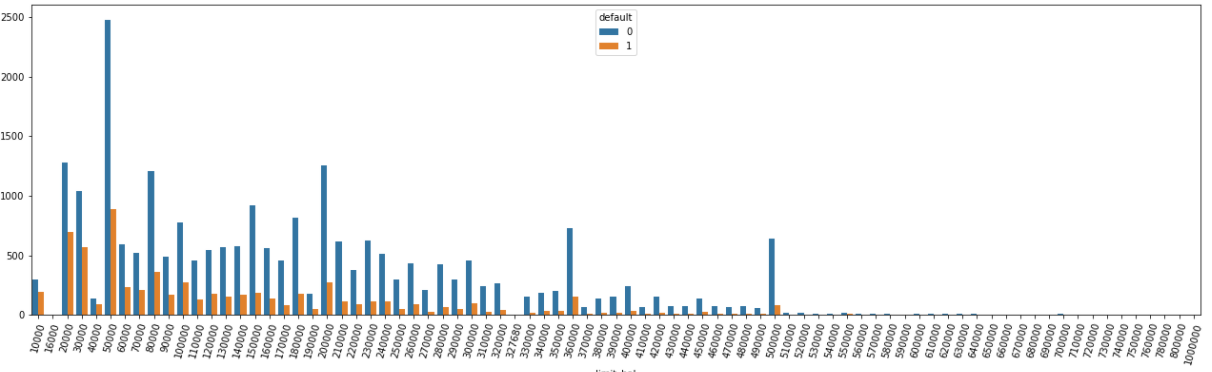
**From the above charts I got to know that many of columns are -negetively correlated with the target variable.**

**After checked the Correlation am going for Checking the outliers:**

****

**We can see the maximum numbers of columns are having the outliers.**

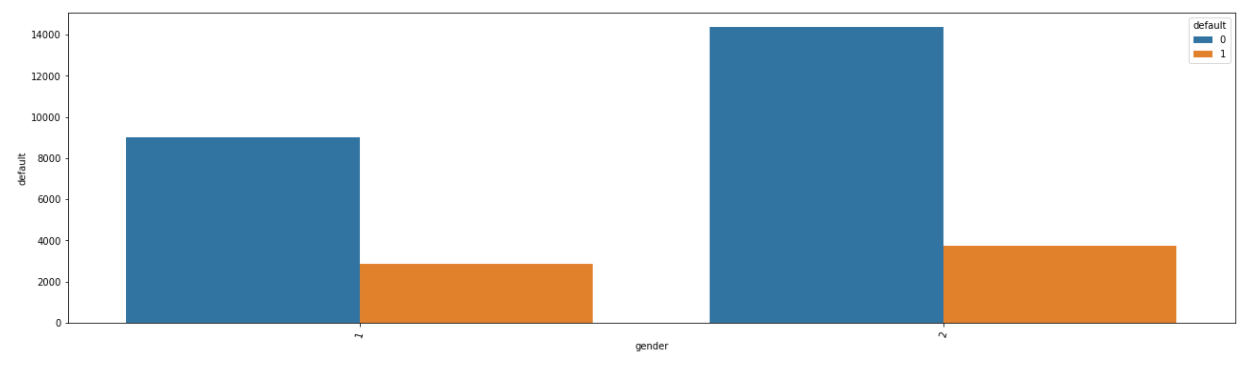
**EDA**

****

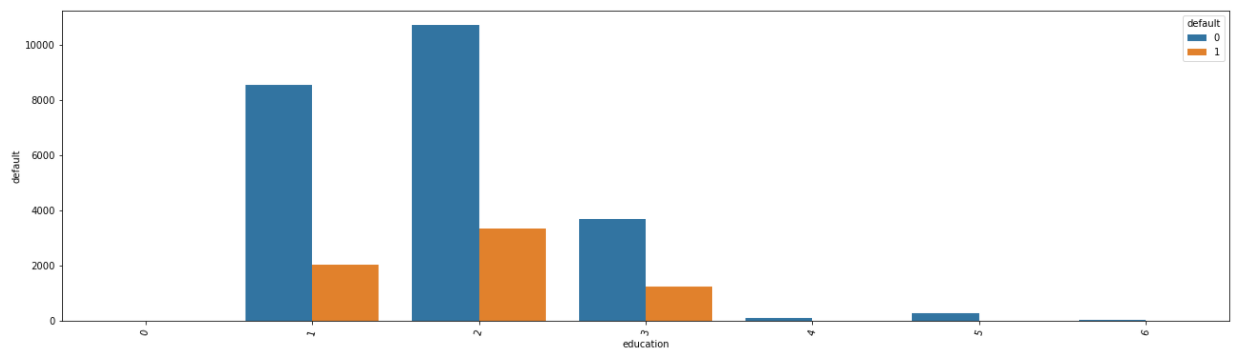
**We can see from the above chart that whe the limit balance is 20k to 40k**

**The default cases are high, On the same way if the limit balance is 50k or more than 50k then there is very low risk for default.**

**Gender (1=male, 2=female)**

****

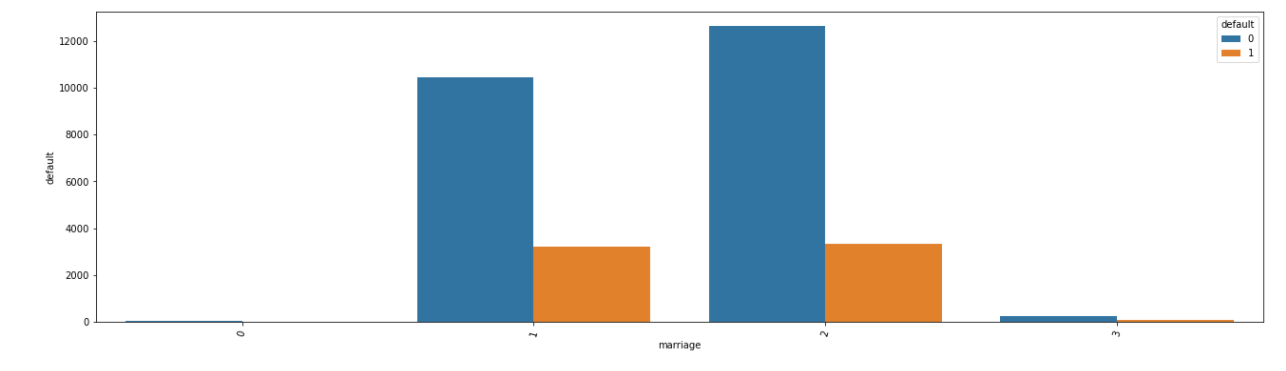
**EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)**

****

**From the above count plot,We can see from above bar chart that number of customers having university education is highest,   followed by customers having graduate school and high school in both the categories**

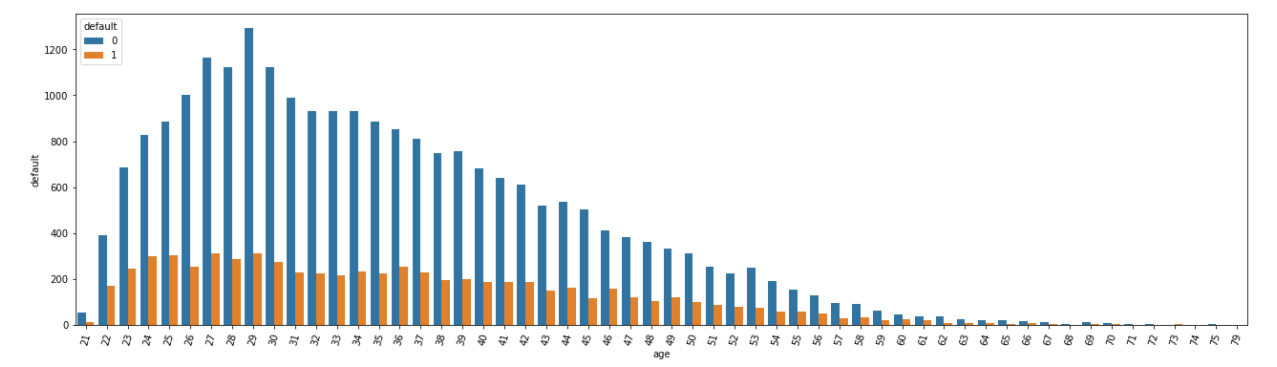
**Next I thought visualize the Material status via count plot:**

**Marital status (1=married, 2=single, 3=others)**

****

**We can see from the above figure that Single persons are using the most of the credit card and their default risky also very low.**

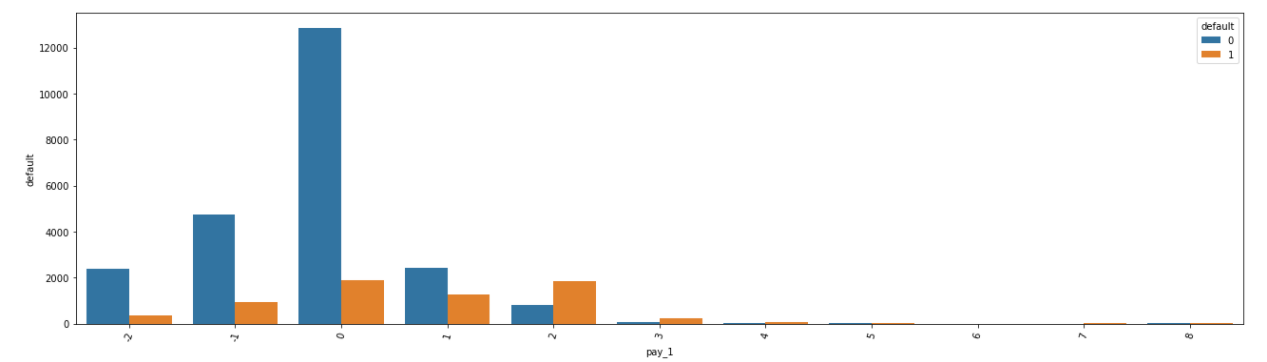
**Next I go for the Age:**

****

**From the above figure we got the information that most of the credit card holders age is 29 ,and its similar for the age of 28.Here we can notice one more think that most of the card holders age is below 40years.**

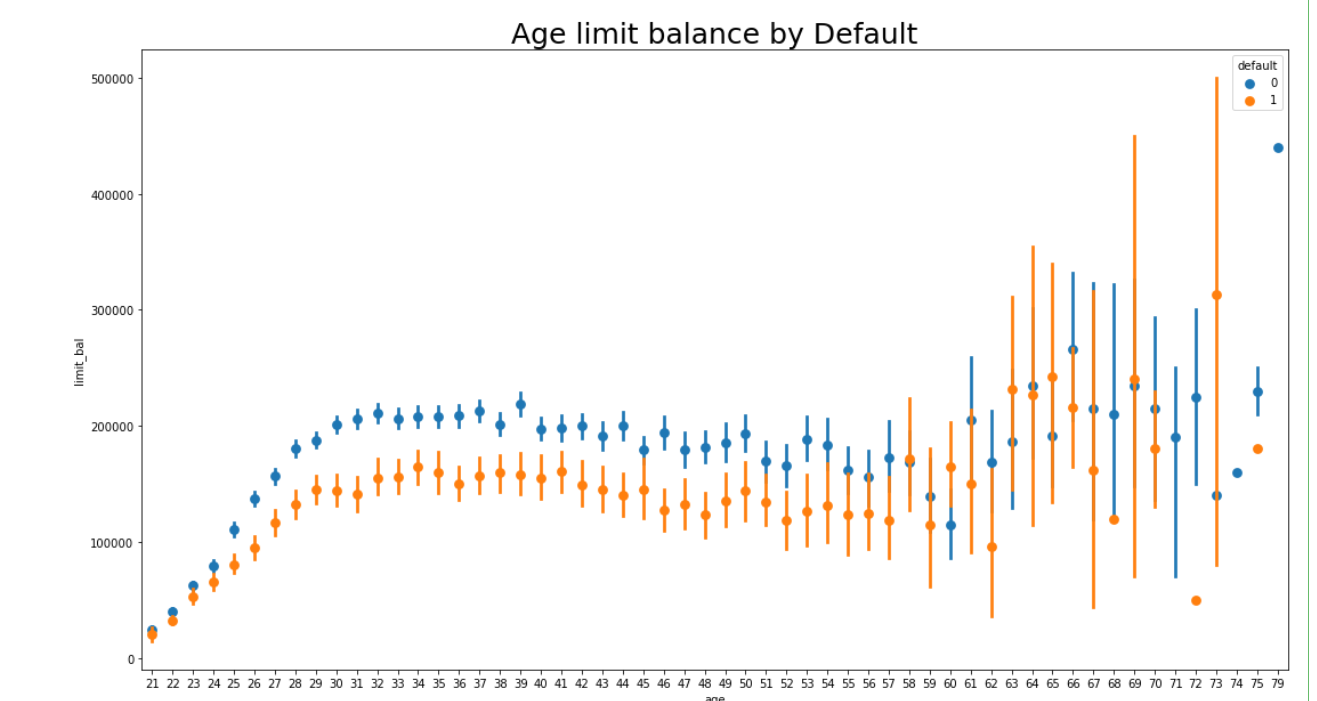
**-1=pay duly, 1=payment delay for one month, 2=payment delay for two months,8=payment delay for eight months, 9=payment delay for nine months and above)**

**Next I visualize the pay\_1**

**We can see the above chart In the month of September(Pay\_1)**

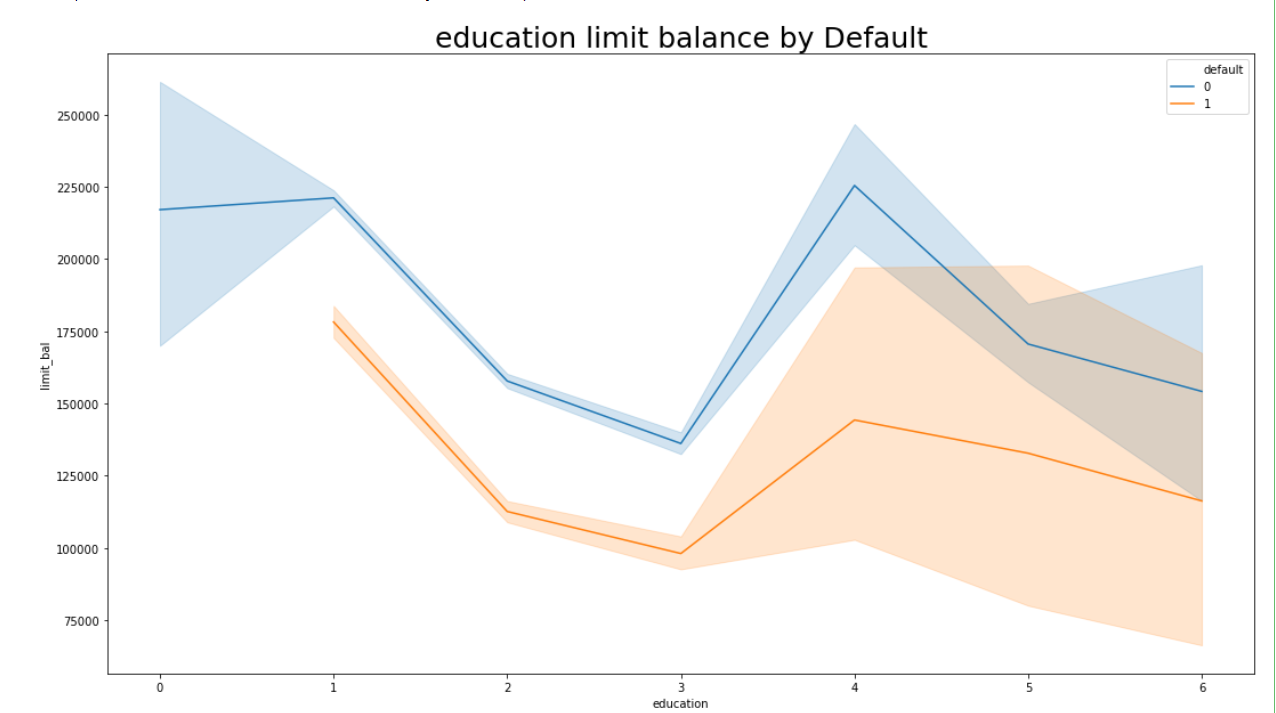
**Highest number of transaction is . no delay for the payment. On this month .**

**Next I thought to check the Age limit balance by Default.**

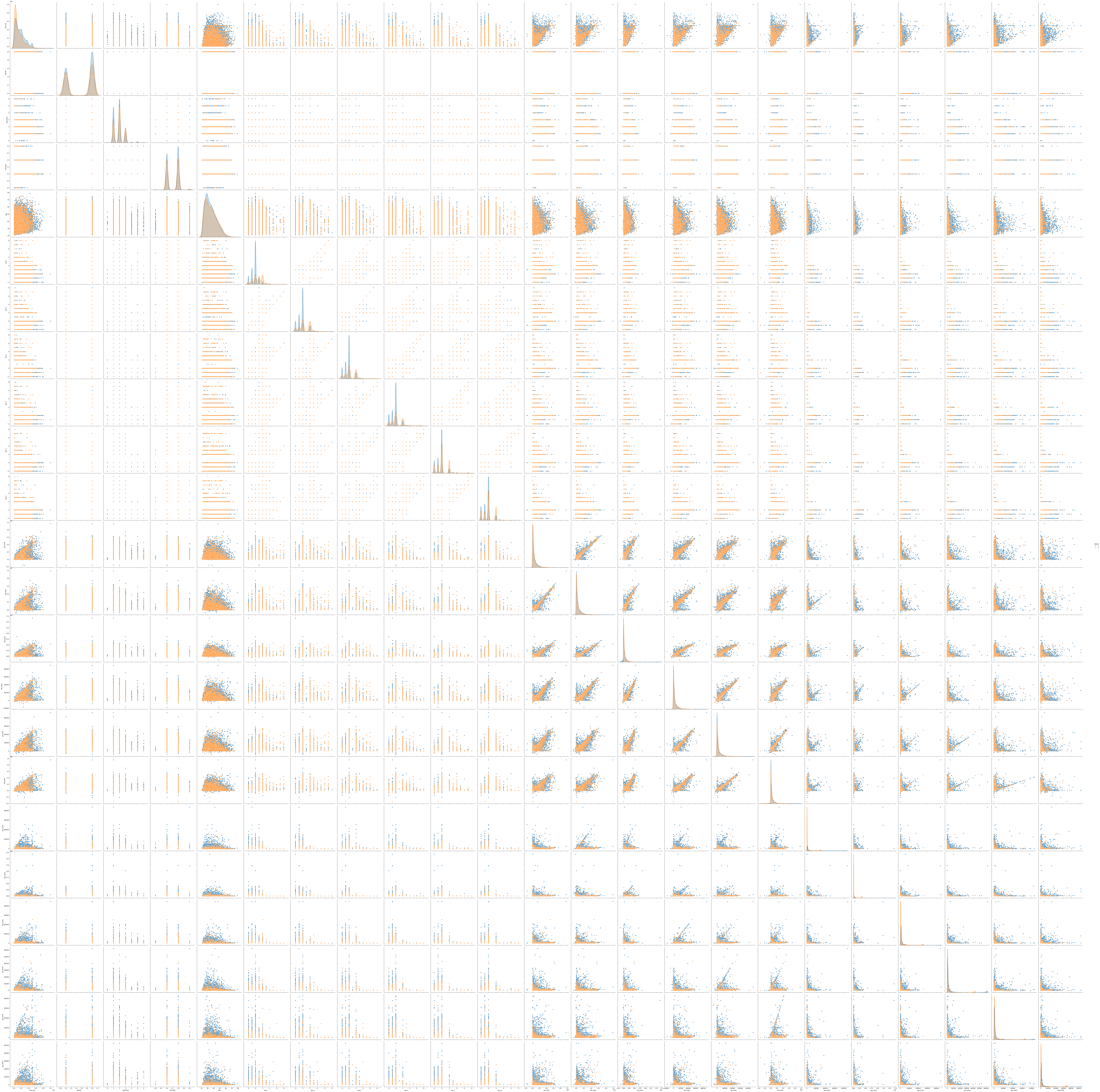
****

**We can see from the above chart that when the age is below 40 and the Limit balance of the card is below 100k card holders are finding default in CC payment. We can see many thinks here like ,If The CC holder’s age is 69 or more than 69 maximum times the CC holder’s default to payment for CC.**

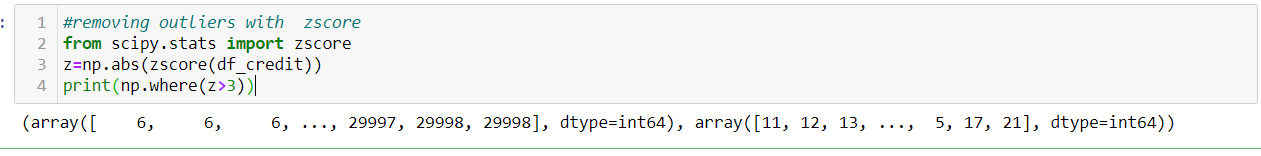
**Next I go for the education limit balance by Default:-**

**we can see from the above figure that**

**Others qualified CC holder’s are getting the highest limit balance ,even this category cc holder’s are finding the most default payment for cc.**

 jhjvfvhbnbhgbjhn

Next I tried to found the outliers and remove it via zscore



On the above code I try to see which zscore value are more than the Threshold value 3.

next, I removed all the outliers and make a new DataFrame with below code 

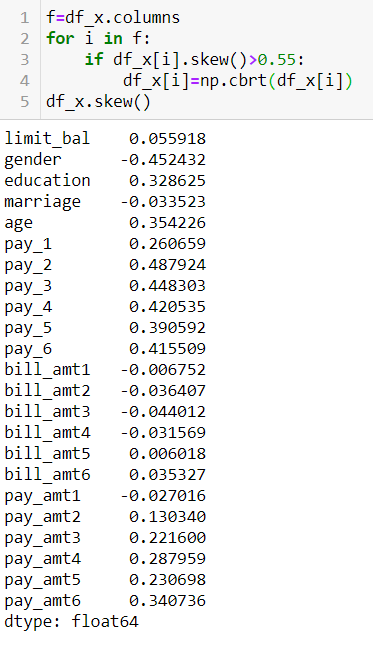
finally we make a dnew data frame with out any Outliers

Next I split the data set input and output variables with the below codes:-

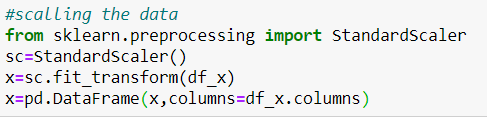


Now we split the DataFrame

Next I try to remove the skewness from my input data via below codes:

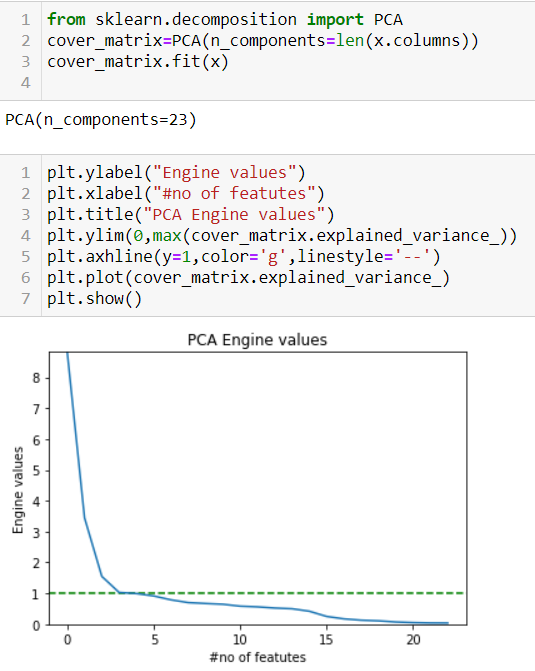


Next I standardize the data via standard scaler.



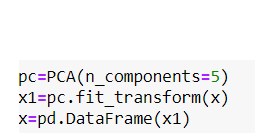
Now I standardize the data

Next I make a scree plot for finding the bests features



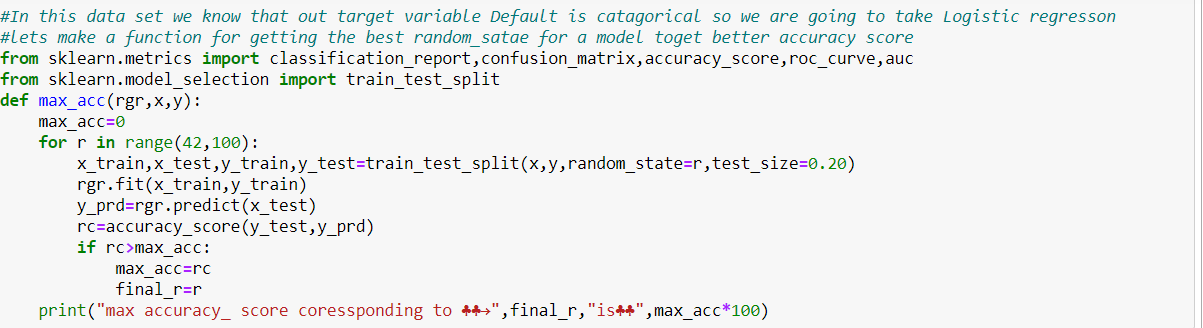
I got from the above figure that 5 features are explaining good

So I take n\_components is 5 for PCA.

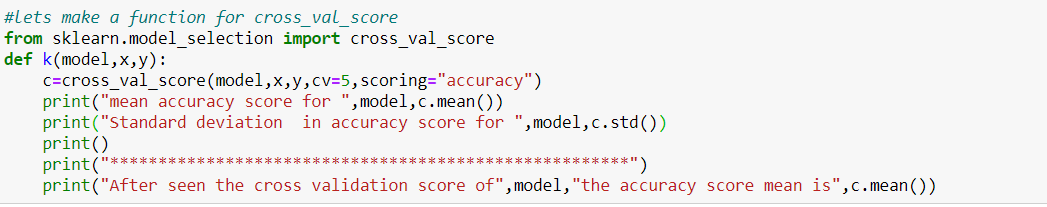


**Building Machine Learning Models**

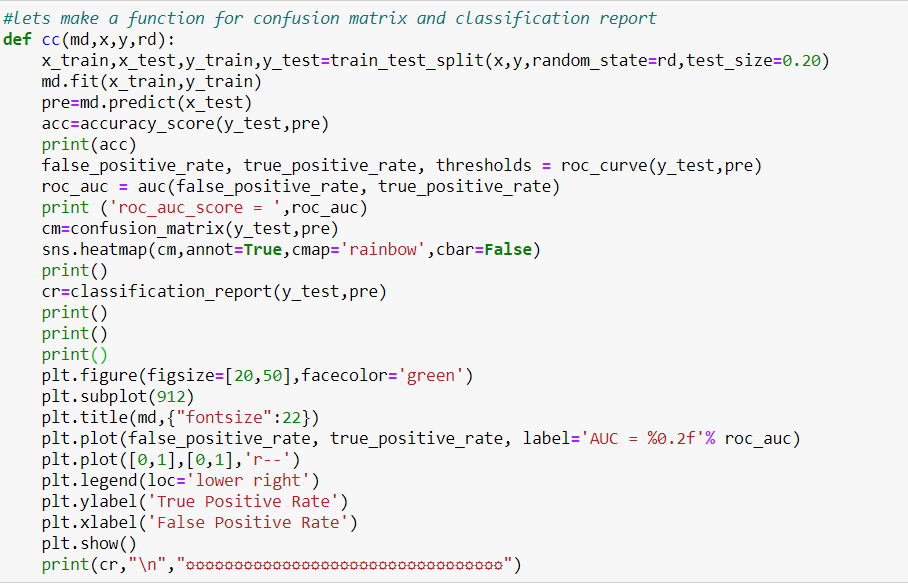
Next I create a function for finding the best random\_state with the below codes:



Similarly I make function for cross validation with below codes:



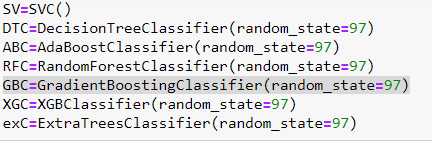
An finally one more function for checking the confusion matrix and ROC\_AOC curve.

Now we make 3 function max\_acc for finding the best Random\_state for better accuracy

K for cross validation of that model

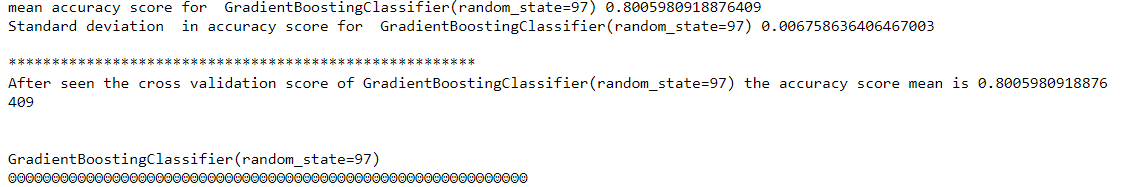
Cc for confusion matrix and roc\_auc curve.

After see below accuracy I final my model with Gridient boost classifiers



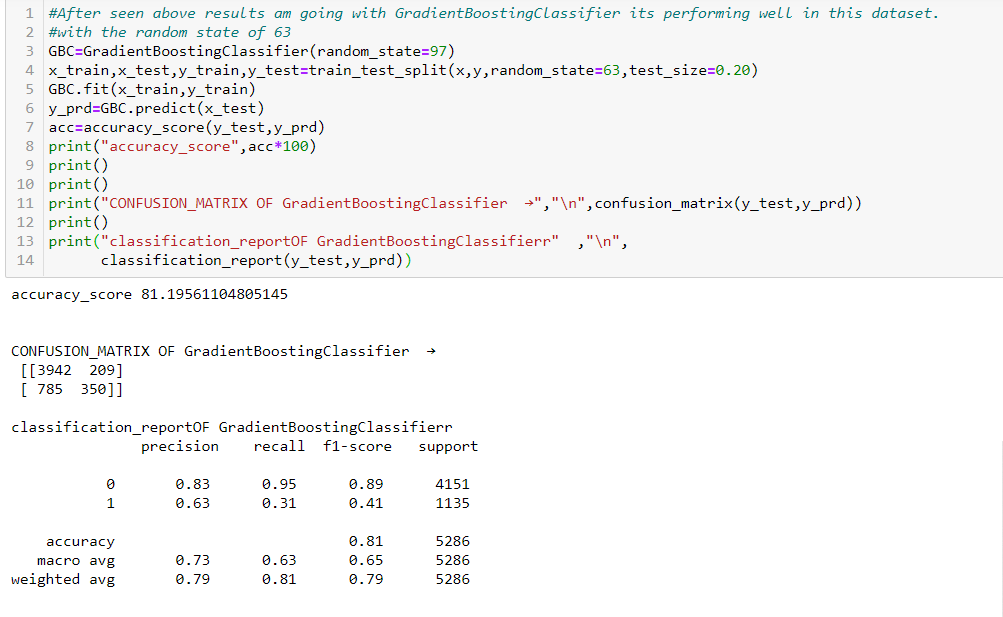




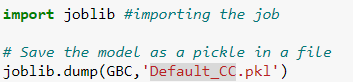


Here we can see from above’s pic that Gridiandboosting classifier’s accuracy score is 81.19 .

And cross validation score is 80.00.



Finally I save the model as pkl.



Conclusion:

When the Companies providing the CC to the customers, they should be recognize the Age, Material Status and Education Also.Bacause Default of a CC is a huge damage for the business.

Thanks

Source:- <https://github.com/RanjitM007/ML_projects/blob/main/%20Default%20of%20credit%20card%20clients%24.ipynb>