Personal Loan Campaign

MachineLearning Project 4th Dec 2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

In this project, it is predicted whether a liability customer will buy personal loans, to understand which customer attributes are most significant in driving purchases, and identify which segment of customers to target more.

Data set is first set to process , making sure there is no empty values and noises the can be misleading the prediction.
From the given data set., better understanding of data is primarily done with univariate analysis , bivariate analysis on exploratory design analysis.

Once the clear understanding is achieved on relation of each variable with others and its importance , the model building is started.
In the model building :

- model evaluation
- decision tree
-model performance improvement
and finally Comparison of the model performance

Business Problem Overview and Solution Approach

Context

All Life Bank is a US bank that has a growing customer base. The majority of these customers are liability customers (depositors) with varying sizes of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio.

You as a Data scientist at AllLife bank have to build a model that will help the marketing department to identify the potential customers who have a higher probability of purchasing the loan.

Business Problem Overview and Solution Approach

Objective

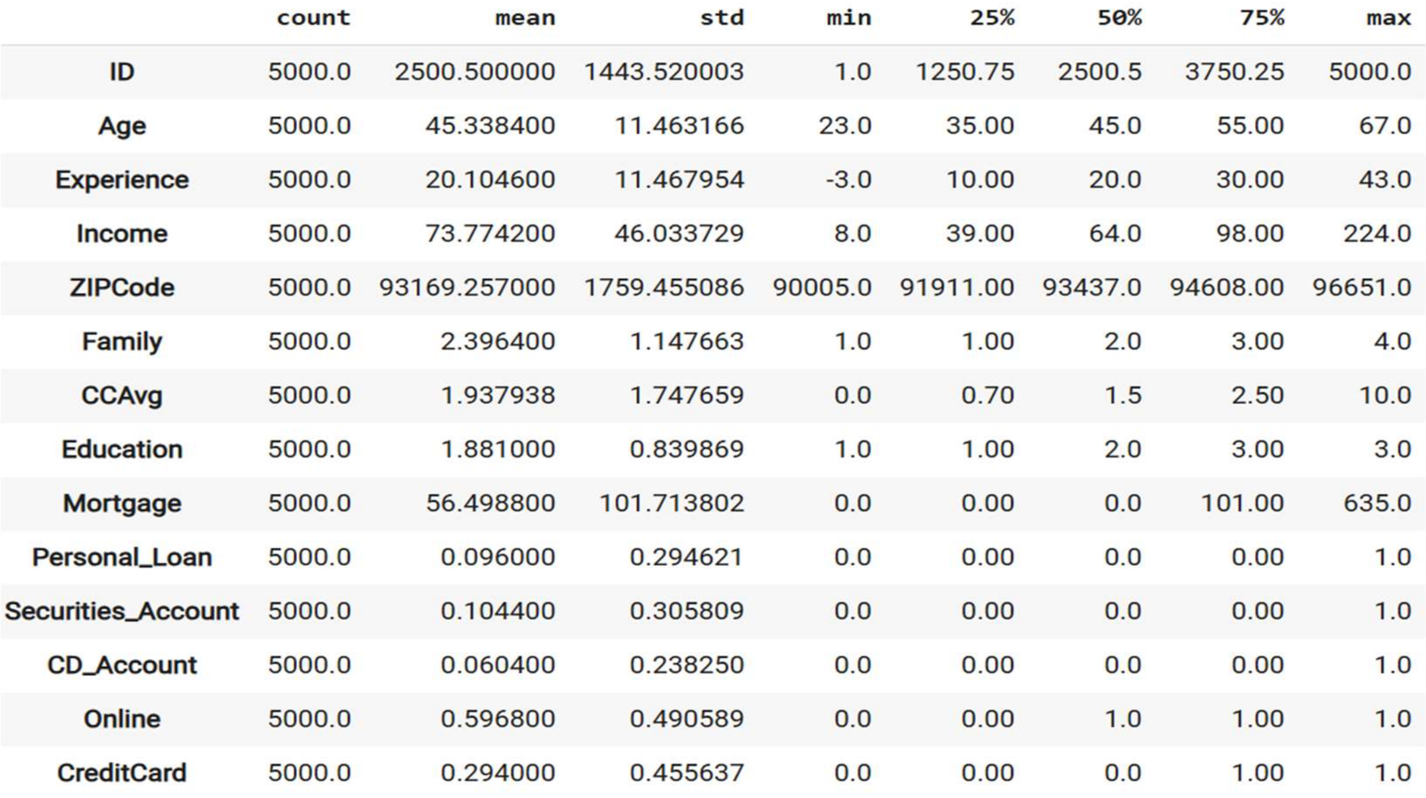
To predict whether a liability customer will buy personal loans, to understand which customer attributes are most significant in driving purchases, and identify which segment of customers to target more.

Solution Approach:

We need to have a clear understanding of the given data.
So before we can start on the problem ,Data set need to sorted.
There should not be any null values for the any features.
In this project, Exploratory data Analysis is used for better visualization of data set.
EDA is used not only for the individual data features , but also with the comparative
Analysis of the data set.
With this we can understand the relationship of the one feature with the other features.
And clear unify the problem requirement

Business Problem Overview and Solution Approach

Statistical analysis of the data helps for the clear understanding of the data set



	count	mean	std	min	25%	50%	75%	max
ID	5000.0	2500.500000	1443.520003	1.0	1250.75	2500.5	3750.25	5000.0
Age	5000.0	45.338400	11.463166	23.0	35.00	45.0	55.00	67.0
Experience	5000.0	20.104600	11.467954	-3.0	10.00	20.0	30.00	43.0
Income	5000.0	73.774200	46.033729	8.0	39.00	64.0	98.00	224.0
ZIPCode	5000.0	93169.257000	1759.455086	90005.0	91911.00	93437.0	94608.00	96651.0
Family	5000.0	2.396400	1.147663	1.0	1.00	2.0	3.00	4.0
CCAvg	5000.0	1.937938	1.747659	0.0	0.70	1.5	2.50	10.0
Education	5000.0	1.881000	0.839869	1.0	1.00	2.0	3.00	3.0
Mortgage	5000.0	56.498800	101.713802	0.0	0.00	0.0	101.00	635.0
Personal_Loan	5000.0	0.096000	0.294621	0.0	0.00	0.0	0.00	1.0
Securities_Account	5000.0	0.104400	0.305809	0.0	0.00	0.0	0.00	1.0
CD_Account	5000.0	0.060400	0.238250	0.0	0.00	0.0	0.00	1.0
Online	5000.0	0.596800	0.490589	0.0	0.00	1.0	1.00	1.0
CreditCard	5000.0	0.294000	0.455637	0.0	0.00	0.0	1.00	1.0

Business Problem Overview and Solution Approach

```
# Column Non-Null Count Dtype --- ------ ---
----------- -----
 0 ID 5000 non-null int64
1 Age 5000 non-null int64
2 Experience 5000 non-null int64
3 Income 5000 non-null int64
4 ZIPCode 5000 non-null int64
5 Family 5000 non-null int64
6 CCAvg 5000 non-null float64
7 Education 5000 non-null int64
8 Mortgage 5000 non-null int64
9 Personal_Loan 5000 non-null int64
10 Securities_Account 5000 non-null int64
11 CD_Account 5000 non-null int64
12 Online 5000 non-null int64
13 CreditCard 5000 non-null int64

dtypes: float64(1), int64(13)
```

There are total 5000 rows and 14 columns.
This means for the campaign of personal loans we have 14 features and 50000 comparative analysis.

Observations

The age ranges from 23 to 67 .

Mortgage ranges from 101k to 635k

Age 55 at the 75th percentile.

Some of the population are more than 55 or older.

Average Income of the population is 73.7K

Average Experience for the population is 11.46 years

50% of population make less than 65K per annual salary.

Business Problem Overview and Solution Approach

Data Dictionary

- ID: Customer ID
- Age: Customer’s age in completed years
- Experience: #years of professional experience
- Income: Annual income of the customer (in thousand dollars)
- ZIP Code: Home Address ZIP code.
- Family: the Family size of the customer
- CCAvg: Average spending on credit cards per month (in thousand dollars)
- Education: Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional
- Mortgage: Value of house mortgage if any. (in thousand dollars)
- Personal_Loan: Did this customer accept the personal loan offered in the last campaign? (0: No, 1: Yes)
- Securities_Account: Does the customer have securities account with the bank? (0: No, 1: Yes)
- CD_Account: Does the customer have a certificate of deposit (CD) account with the bank? (0: No, 1: Yes)
- Online: Do customers use internet banking facilities? (0: No, 1: Yes)
- CreditCard: Does the customer use a credit card issued by any other Bank (excluding All life Bank)? (0: No, 1: Yes)

These are the features collect for the campaign of personal loan.From the given data we need to sort the required data. **The unique values can be dropped like ID**

Data Processing

Anomalies often skew the two most important characteristics of distributions: mean and standard deviation. It is important as identifies suspicious activity that falls outside of your established normal patterns of behavior.

In the given data set, there is Experience column with some Anomalous data like -1, -2 and -3.

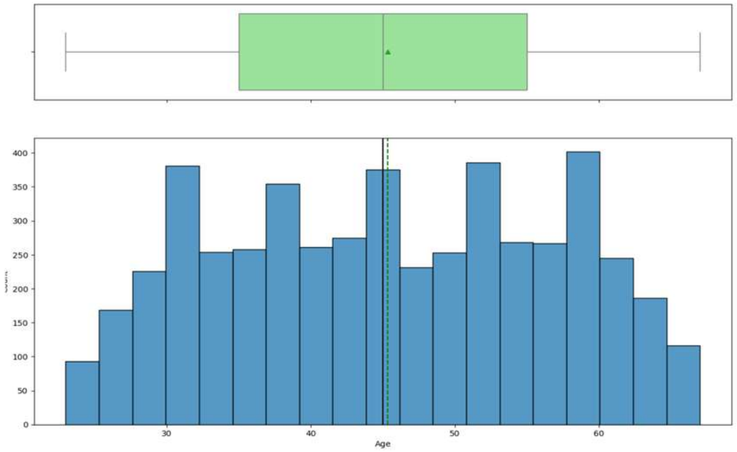
This has to be handled **to ensure dataset accuracy and reliability**. So all the -1,-2, -3 are replaced with the respective 1, 2, 3

Data should be clearly defined under categorical and Numerical types for better understanding.

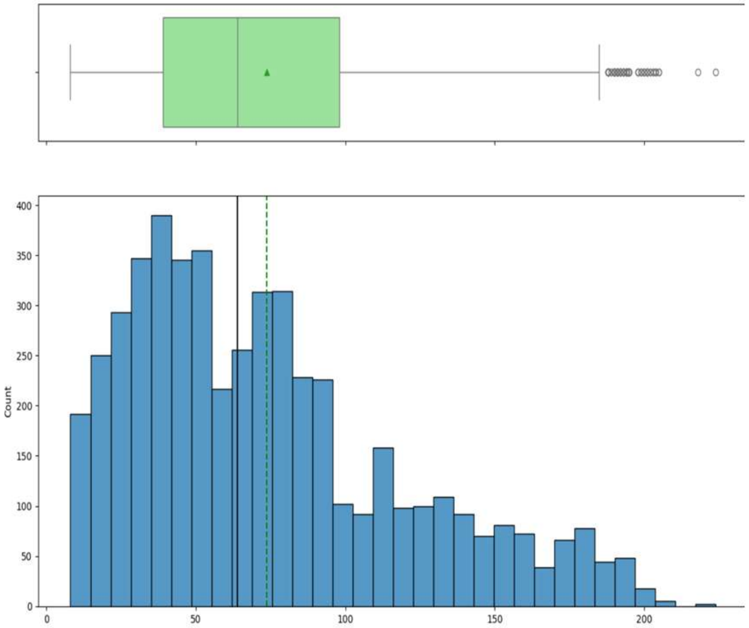
To Optimize storage and computation when analyzing or processing data .The zip code data is converted into categorical data.
Counting and displaying the unique first two-digit prefixes to understand geographic representation.

For this reason , the following categorical features are converted into category
Securities_Account, CD_Account, Online, CreditCard, ZIPCode, Education and Personal_Loan.

EDA Results

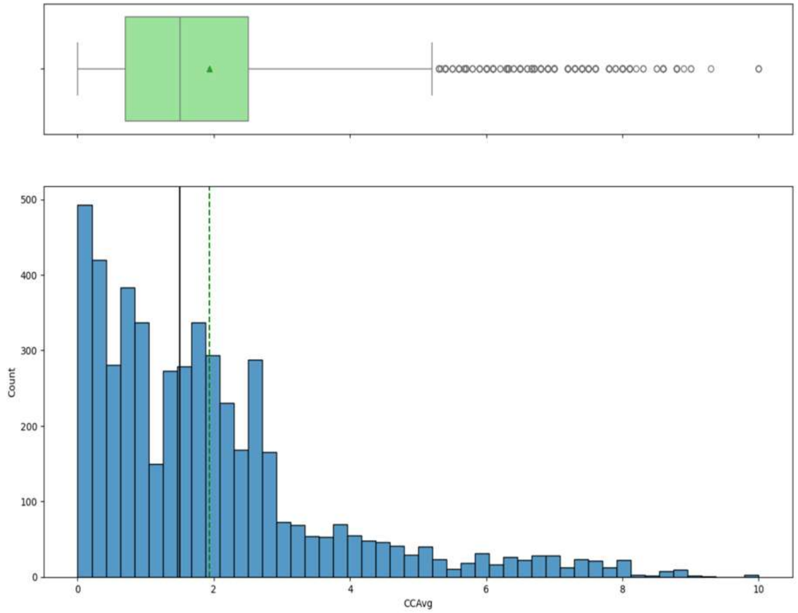


INCOME:
The income distribution is skewed to the right with many outliers on the upper quartile.
•Majority of the population income is between 15K - 95 K



AGE:
The age distribution looks slightly right skewed with the average age which is 45.
•There are no outliers present.

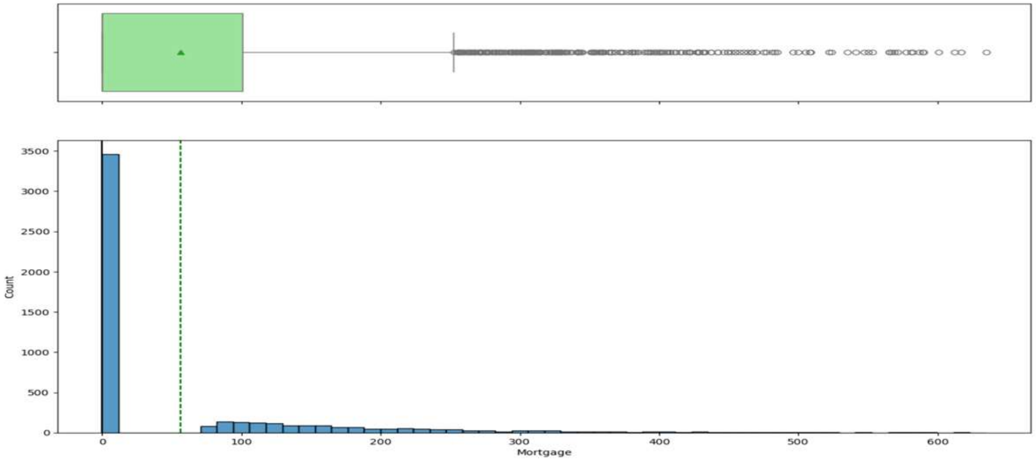
EDA Results



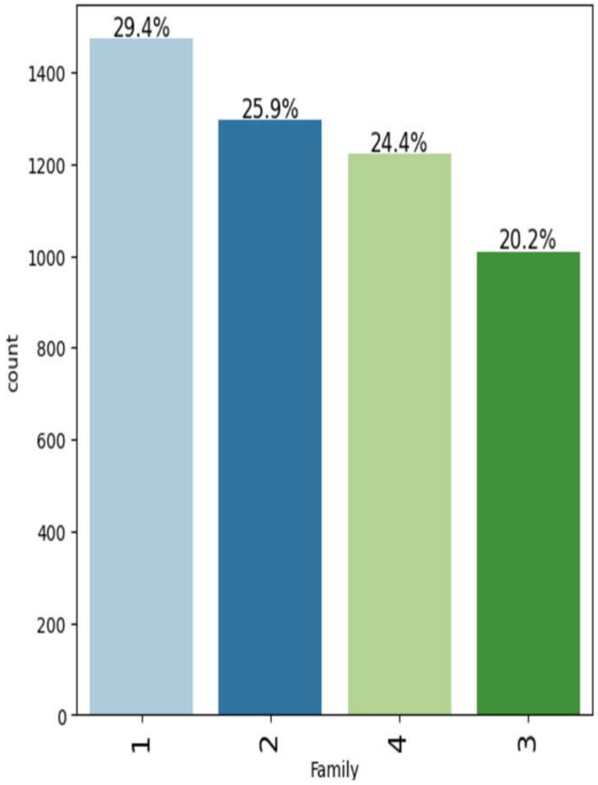
•Credit Card Balance:

The Credit Card balance is right skewed with many outliers on the upper quartile.
•Majority of the population have less than 4K Credit Card Debt.

•MORTGAGE:
Most of the population has no Mortgage. Either they have paid off the mortgage or renting a property to live. This information is unavailable in this dataset.
•So the distribution of the Mortgage is heavily right skewed with several outliers in the upper quartlie.

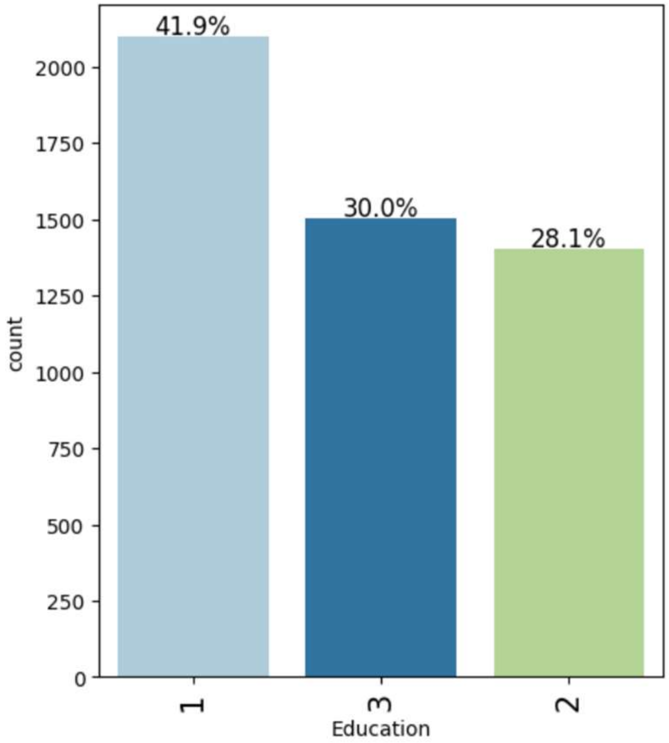


EDA Results

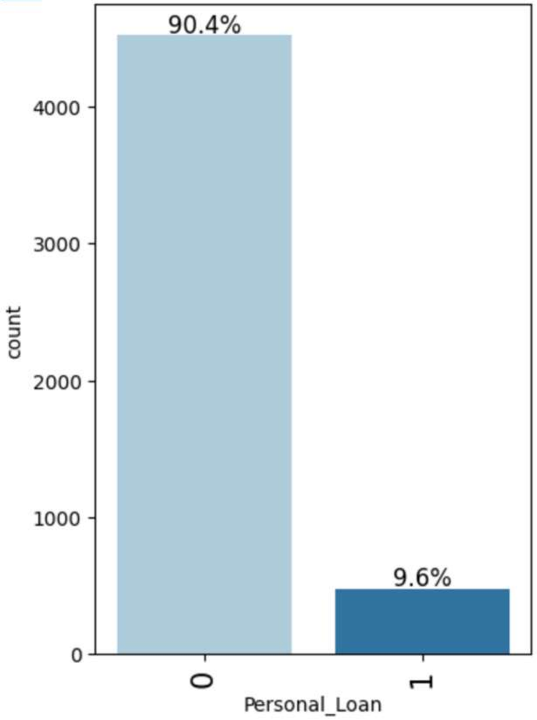


•FAMILY: ⇦
29% of the family size is 1, also they are largest portion of this data set.
•Family Size with individuals 3 has the lowest among the population at 20%

EDUCATION: ⇨
most of the population in this data set are under graduate.
•30% of the population has an advanced degree and 28% are Graduates.



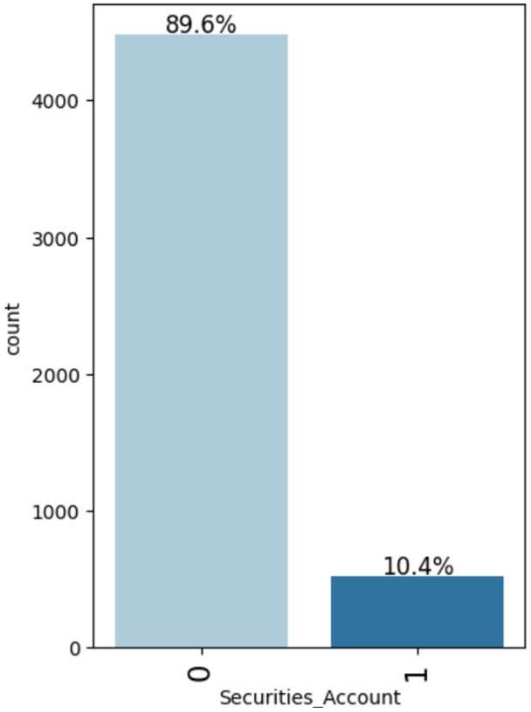
EDA Results

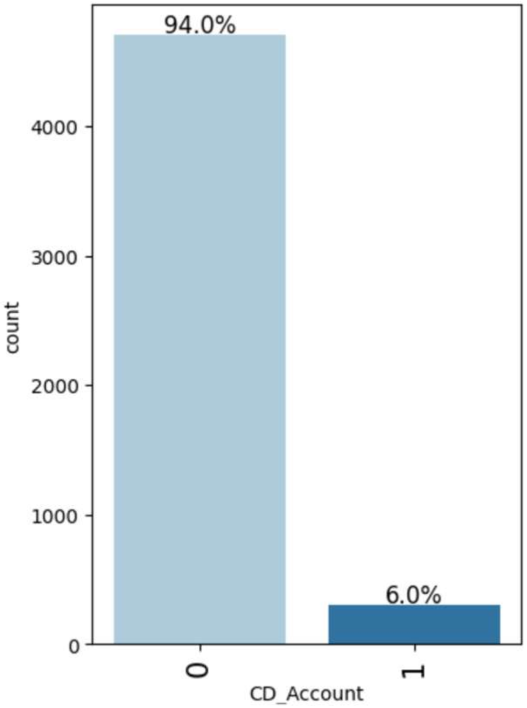


PERSONAL LOAN:

Most of the population does not have a personal loan.

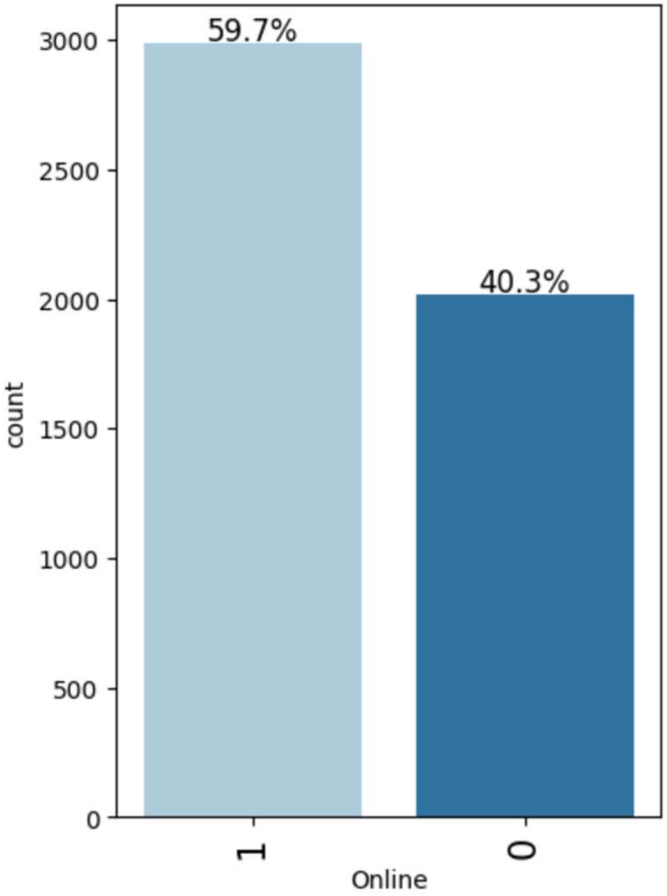
POPULATION:
Only 10% of the population have Securities Account



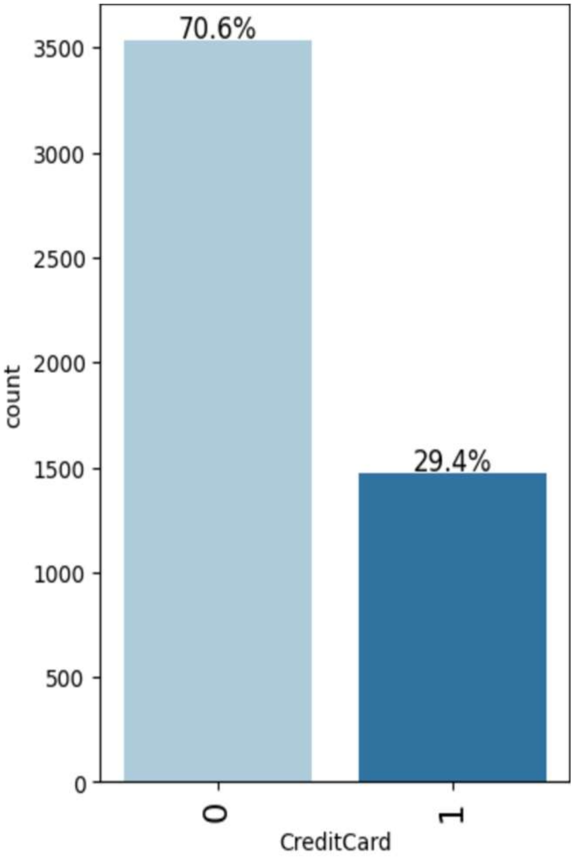


CD_Account:
Only 6% of the population have a CD Account.

EDA Results

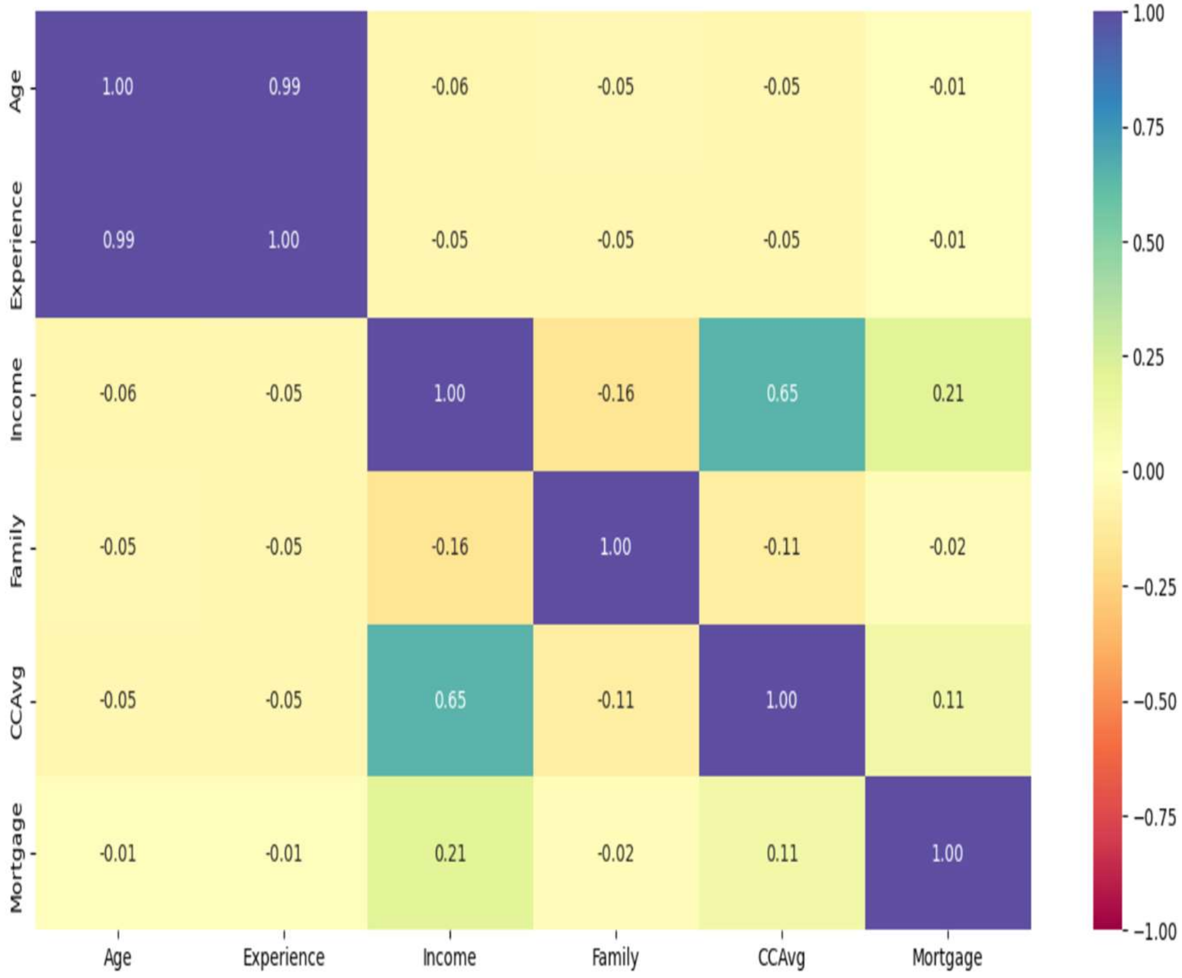


CREDIT CARD:
70% of the population does not have a Credit Card from another bank, only 39 percent have a Credit Card issued by another bank.



ONLINE ACCOUNT
60% of the population have an Online Bank Account.

EDA Results



•Experience and Age has a strong positive correlation.

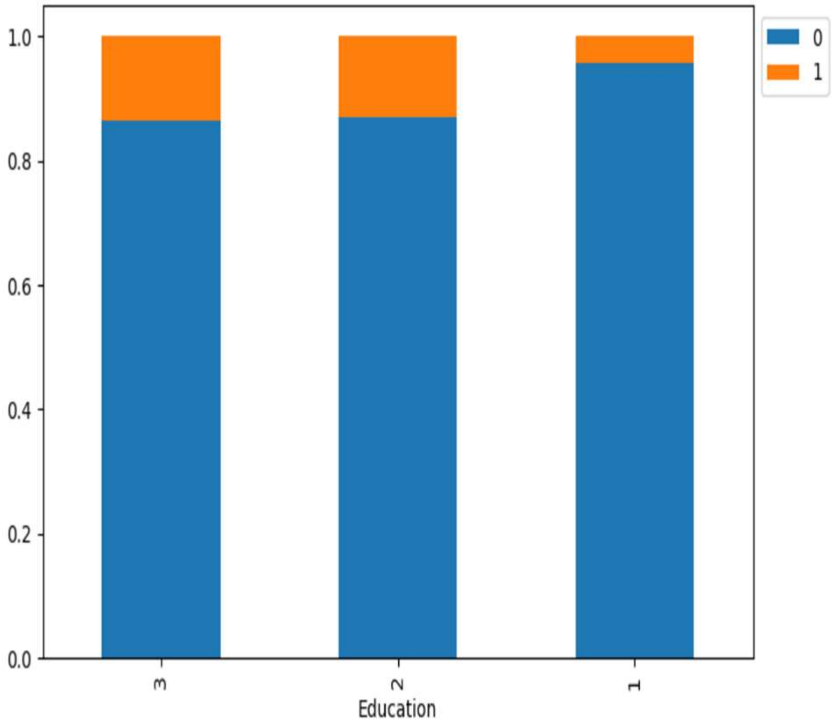
•Income and Credit Card Average Balance has a positive correlation.

•What are the attributes that have a strong correlation with the target attribute (personal loan)?

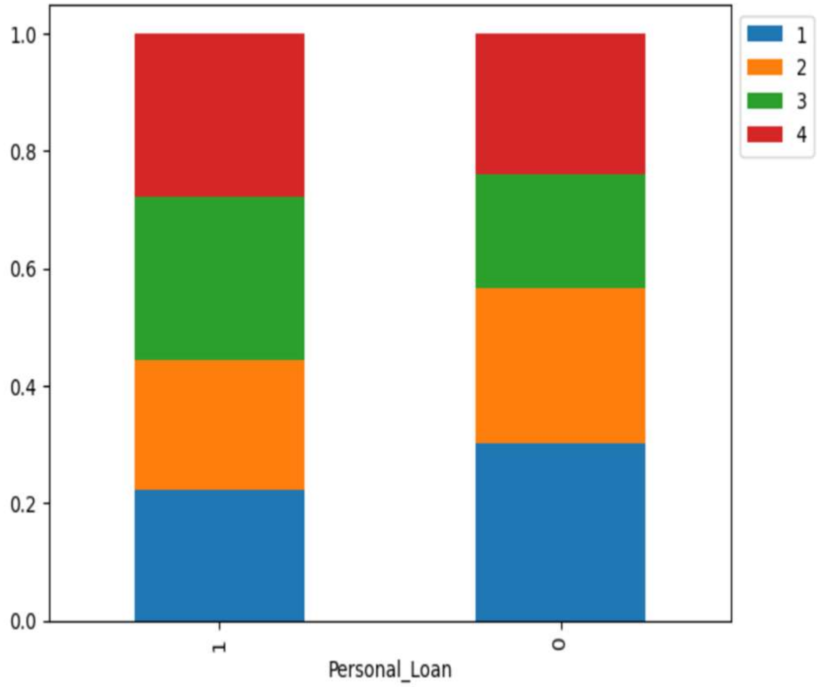
- Income, Credit Card Average Balance and CD Account has strong coorelation with the target attribute (personal loan).

EDA Results

Education Vs Personal_Loan



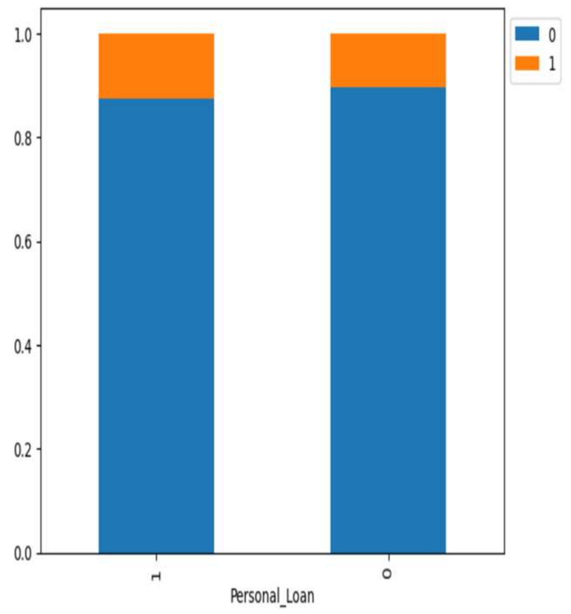
- Less than 20% of the population have personal loan when the have an advanced degree.
- Less than 20% of the population have personal loan when the have an graduate degree.
- Less than 10% of the population have personal loan when they have a Under Graduate degree.

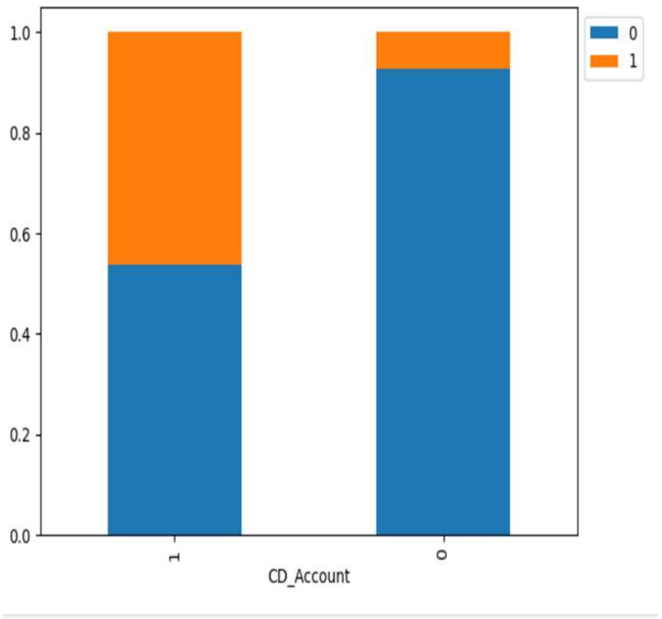


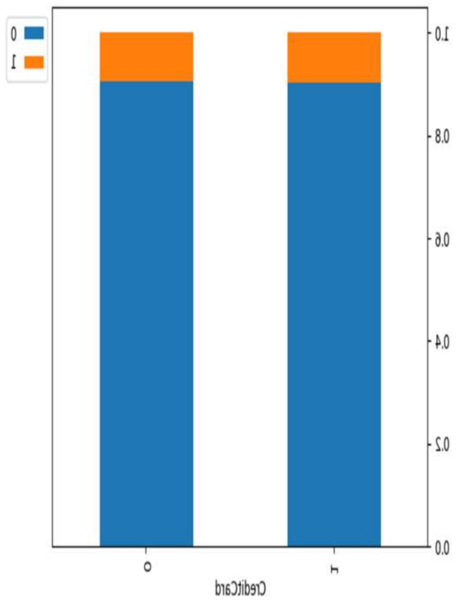
Personal_Loan Vs Family

The single person family mostly got the loan un-accepted. Three and four member family most got the loan.

EDA Results





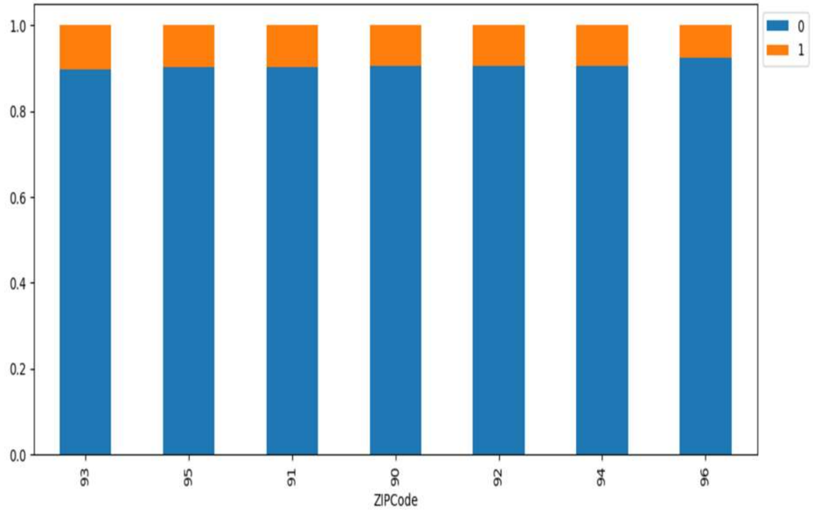


Personal_Loan Vs Securities_Account
Most of population does have a personal loan whether they have a securities account or not.

CD_Account Vs Personal_Loan
Maximum CD_account does not have the personal loan.

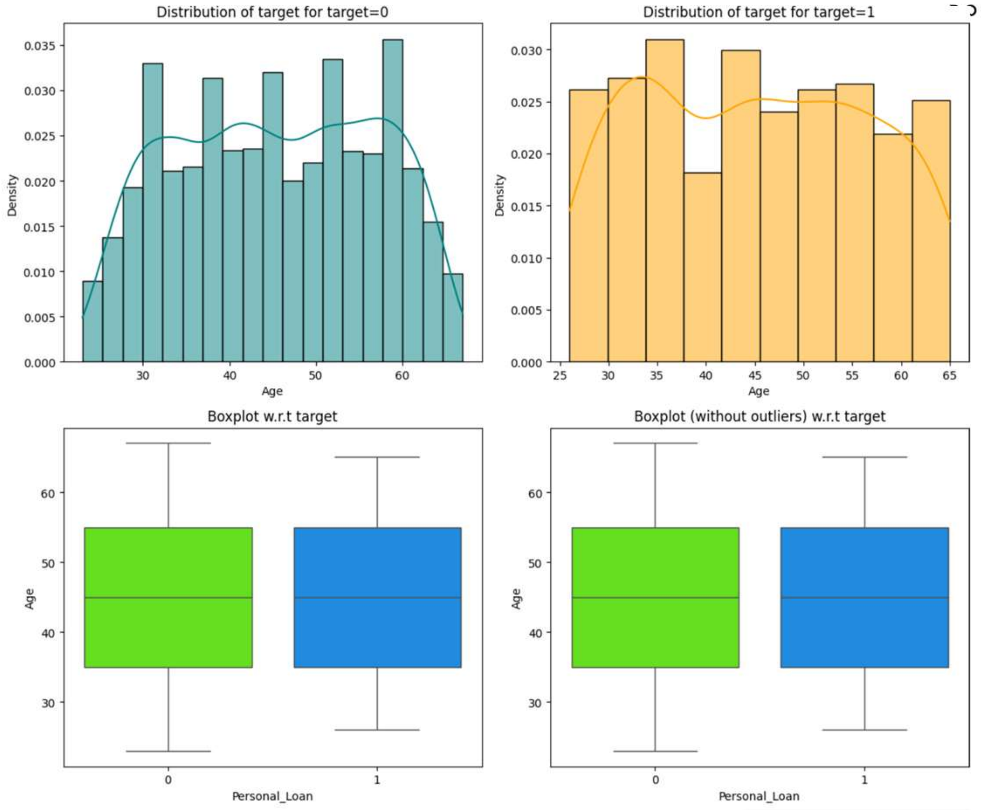
CreditCard Vs Personal_Loan
Most of population does have a personal loan whether they have a creditcard or not.

EDA Results



Personal_Loan Vs Zipcode

Most of population does have a personal loan



Age Vs Personal_Loan

- The population who have personal loans are predominantly over 25 years old and less than 65 years old.
- There is a sizable population between the ages of 25 and 65 who do not have personal loans.
- Individuals in their late 30s to early 40s possess fewer personal loans compared to other demographics.

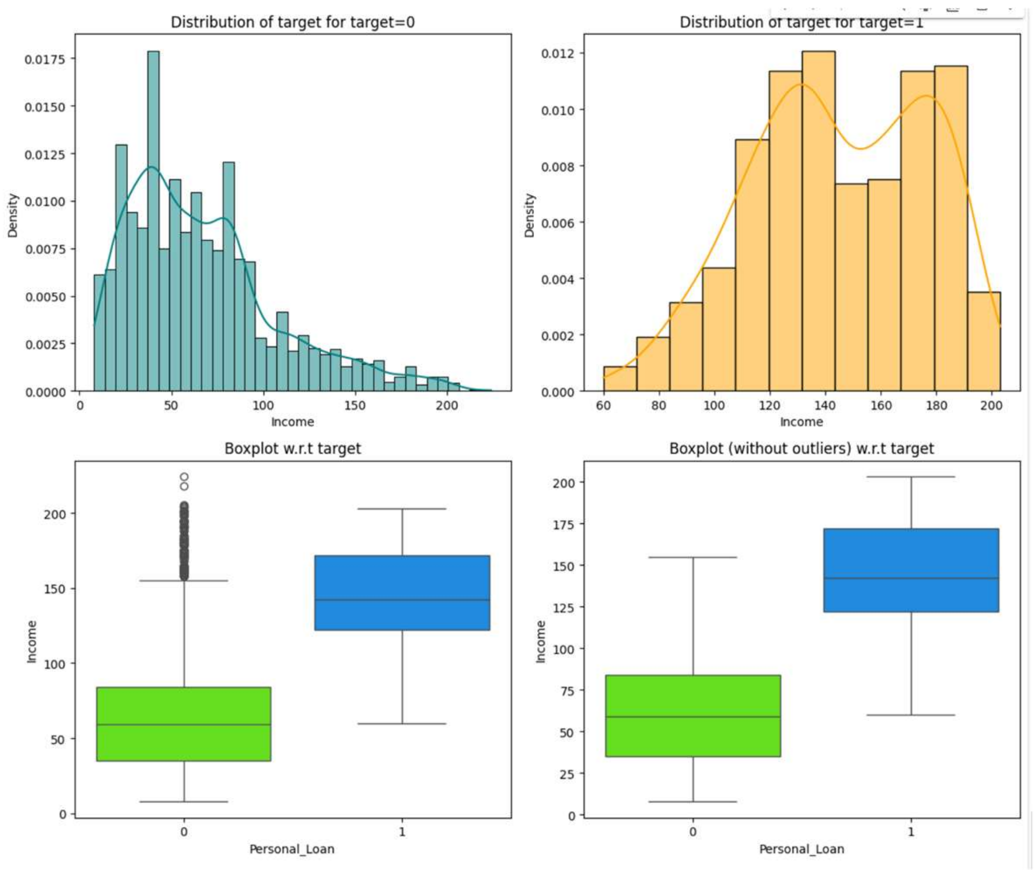
EDA Results

Income Vs Personal_Loan

There is some outlier for the approved personal loans with Income.

The income vs personal loan has many other factors influencing Like:
Income and loan Approval
Income vs Loan Amount
Income Distribution
Risk factor

The **Income vs. Personal Loan** analysis can reveal useful insights about financial behavior, loan approval patterns, and the factors influencing loan amounts and repayment. It's important to understand that **income** alone may not be sufficient for making loan-related decisions. Therefore, combining income data with other financial metrics can provide a fuller picture.

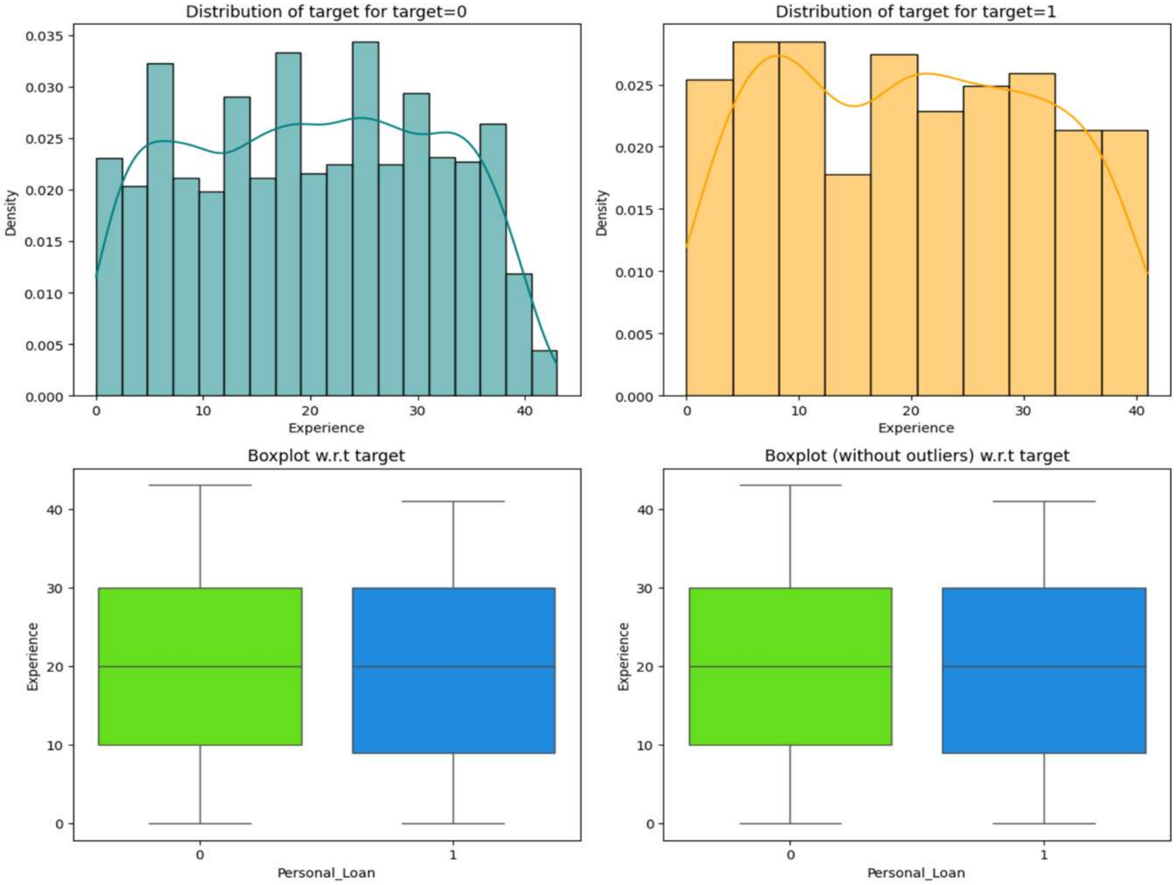


EDA Results

Experience Vs Personal_Loan

Insights from the Analysis:
These are the factor classifying the personal loan based on the experience
1.Approved loan is very random
2.Approval is independent of Experience
3.Approval includes different other variables

The **Experience vs. Personal Loan** analysis helps us understand how the number of years an individual has worked (professional experience) relates to their borrowing behavior. This can provide insights into financial decision-making, loan approval processes, and how people with different levels of experience approach personal loans.

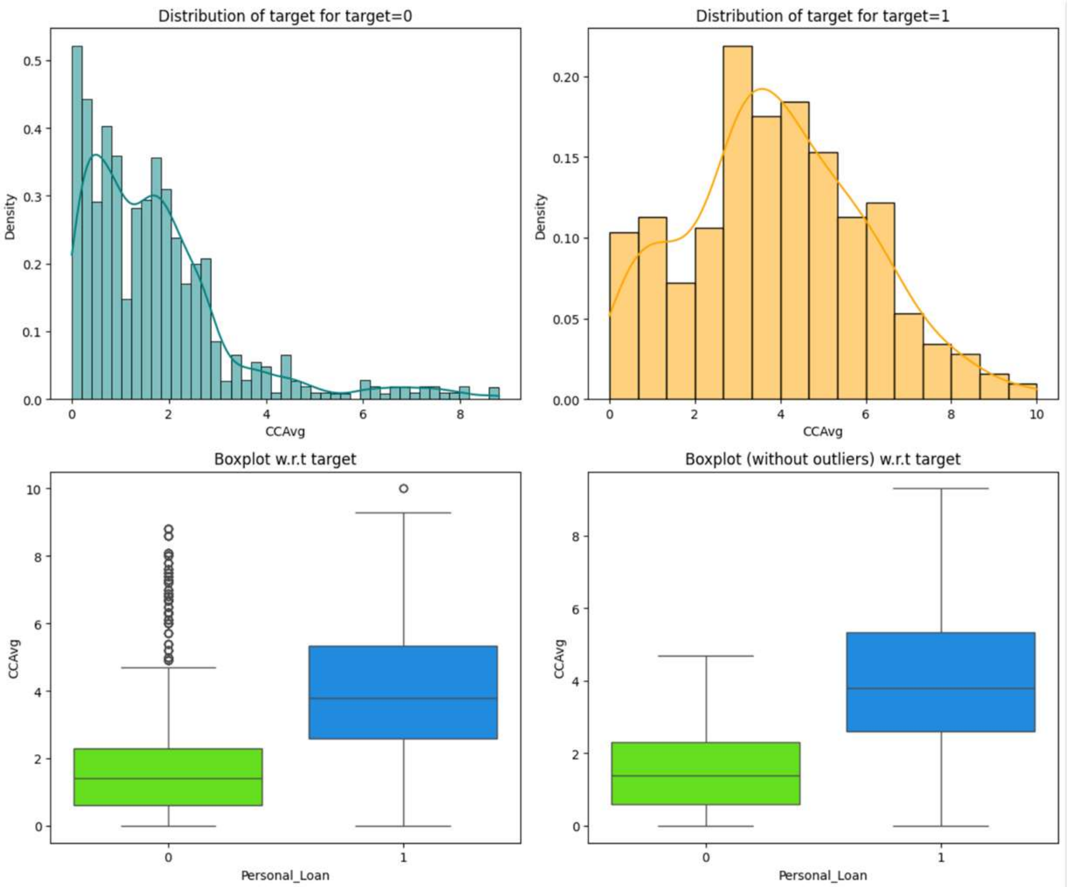


EDA Results

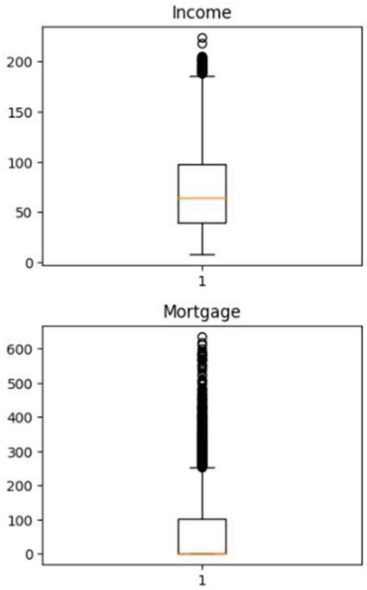
CCAvg Vs Personal_Loan

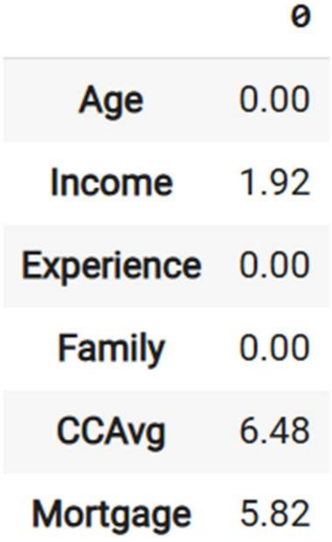
Insights from the Analysis:

- Approved personal loan is positive skewed means it is right skewed.
- There are few outlier are also seen with the approval segment on the bases of credit card usage.
- CCAVG is very high level classification which includes credit usuage,credit balance and average balance .
- Customer with low credit card average is under personal loan approval then high credit card average.
- There is no outlier on the rejection of personal loan with the credit card average.
- Analyzing these relationships can help lenders assess the risk and determine the likelihood of loan approval or default based on an individual’s credit behavior.

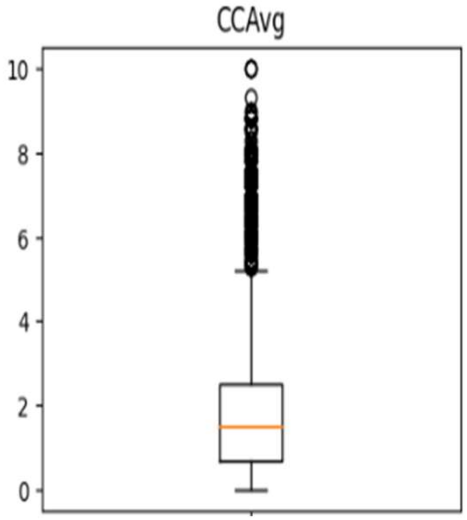


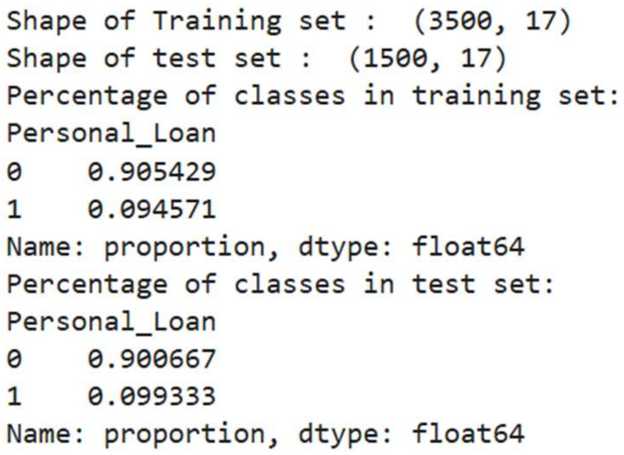
Data Preprocessing (contd.)





	0
Age	0.00
Income	1.92
Experience	0.00
Family	0.00
CCAvg	6.48
Mortgage	5.82





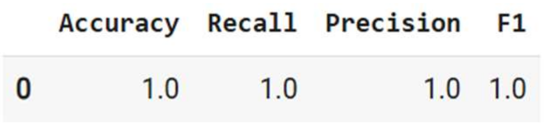
```
Shape of Training set :  (3500, 17)
Shape of test set :  (1500, 17)
Percentage of classes in training set:
Personal_Loan
0    0.905429
1    0.094571
Name: proportion, dtype: float64
Percentage of classes in test set:
Personal_Loan
0    0.900667
1    0.099333
Name: proportion, dtype: float64
```

90% are passed for the personal loan while only 9% failed in the training set .Same for the test set data also

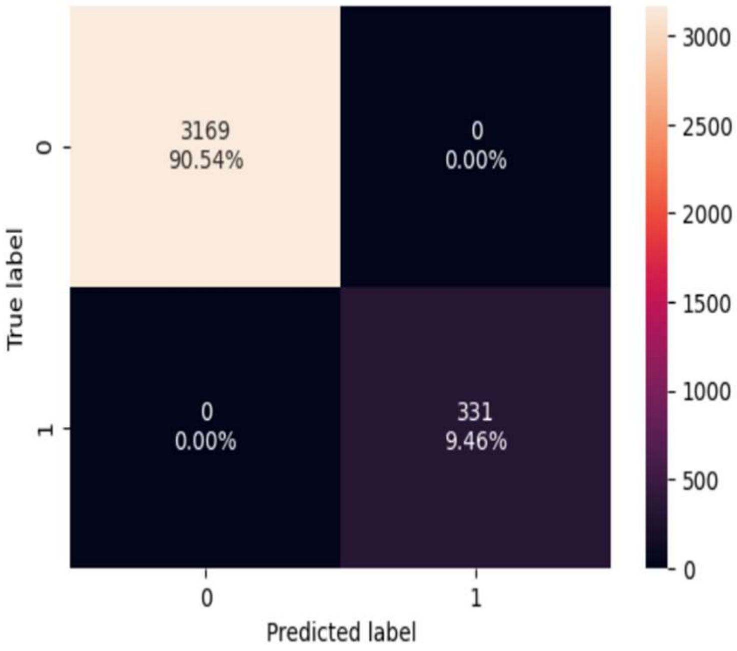
- There are quite a few outliers in the data.
- CCAVG, Income and Mortgage only has the outliers in their data.
- However, we will not treat them as they are proper values

Model Performance Summary

Confusion matrix from Training set data

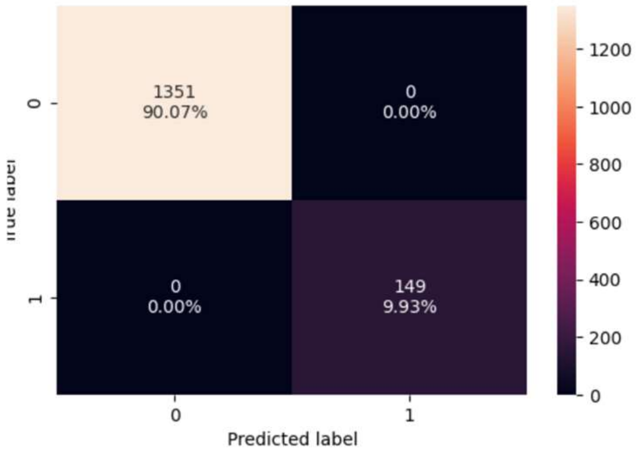


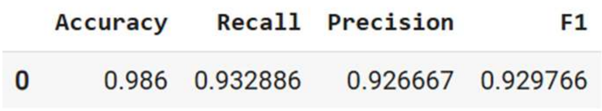
	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0



- Model is able to perfectly classify all the data points on the training set.
- 0 errors on the training set, each sample has been classified correctly.
- As we know a decision tree will continue to grow and classify each data point correctly if no restrictions are applied as the trees will learn all the patterns in the training set.
- This generally leads to overfitting of the model as Decision Tree will perform well on the training set but will fail to replicate the performance on the test set.

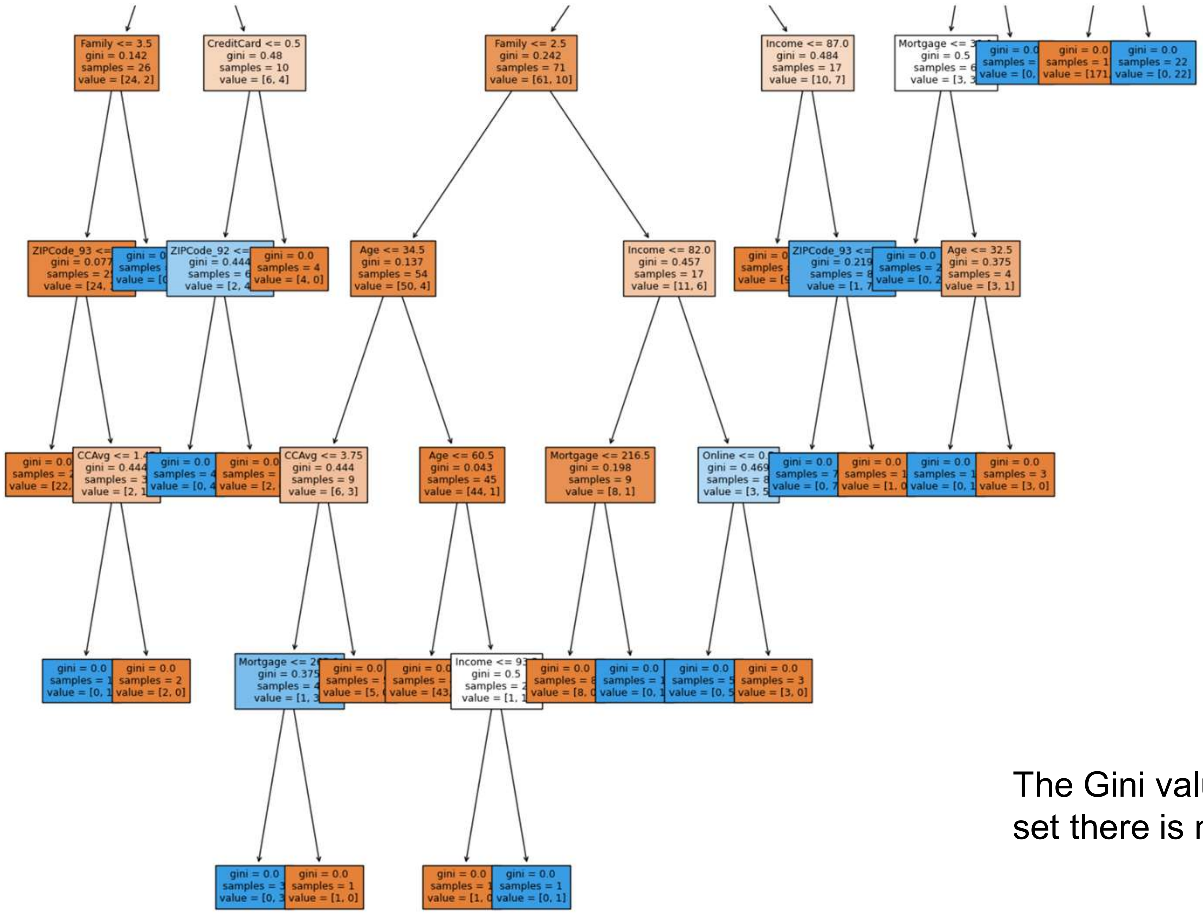
Confusion matrix from Test set data

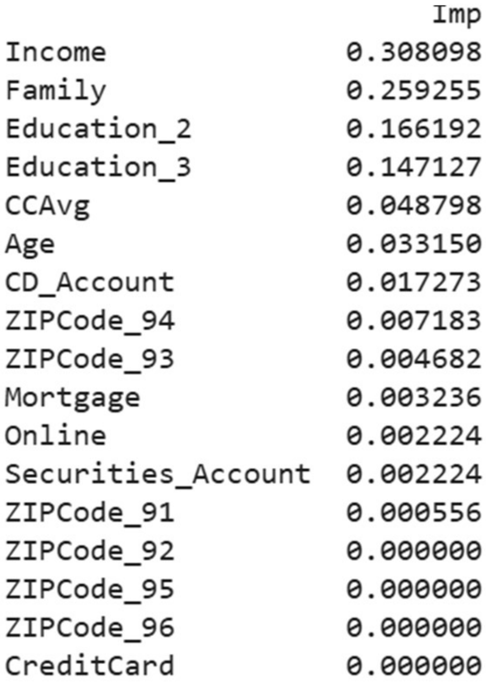




	Accuracy	Recall	Precision	F1
0	0.986	0.932886	0.926667	0.929766

Model Performance Summary

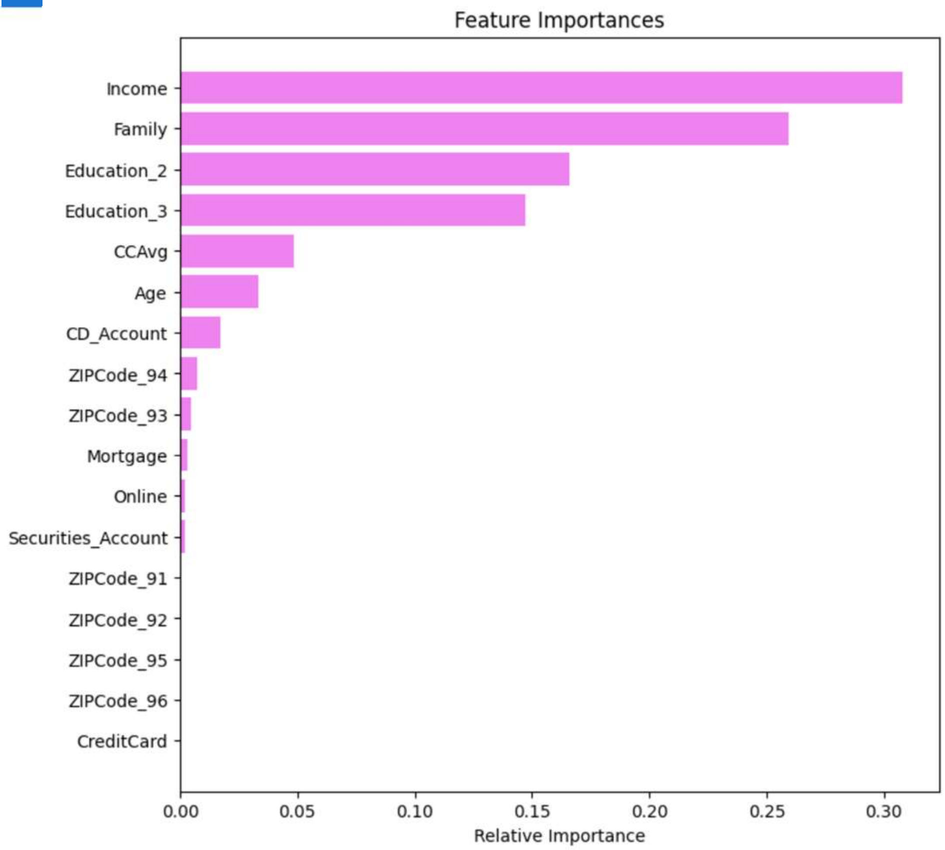




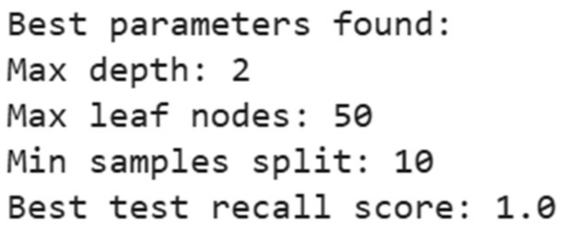
```
                              Imp
Income                   0.308098
Family                   0.259255
Education_2              0.166192
Education_3              0.147127
CCAvg                    0.048798
Age                      0.033150
CD_Account               0.017273
ZIPCode_94               0.007183
ZIPCode_93               0.004682
Mortgage                 0.003236
Online                   0.002224
Securities_Account       0.002224
ZIPCode_91               0.000556
ZIPCode_92               0.000000
ZIPCode_95               0.000000
ZIPCode_96               0.000000
CreditCard               0.000000
```

The Gini values lies between 0 to 1.From the given data set there is no values to 1.

Model Performance Summary



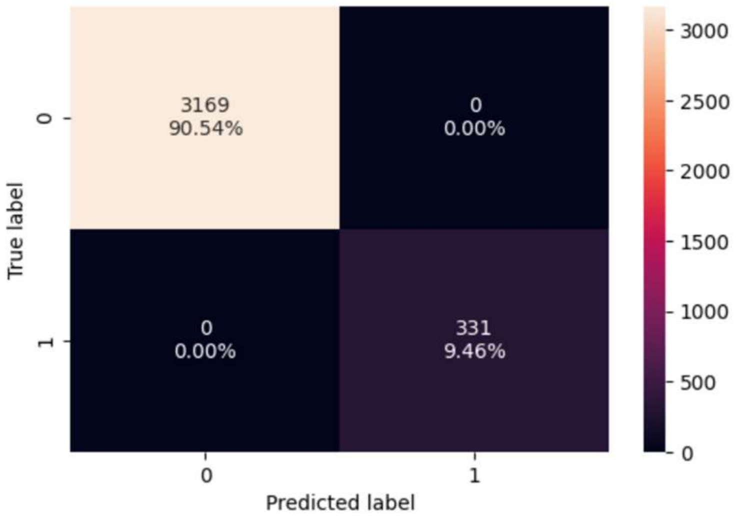
From the plot, it demonstrate the relative importance of individual feature for the personal loan . It shows that income is the most dependent variable for the loan.

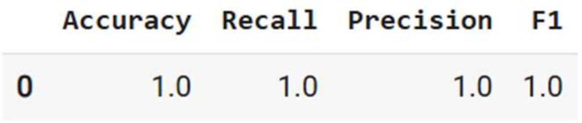


```
Best parameters found:
Max depth: 2
Max leaf nodes: 50
Min samples split: 10
Best test recall score: 1.0
```

After post pruning , the best parameters are collected to illustrate the
Confusion matrix on the test set data.

Model Performance Summary

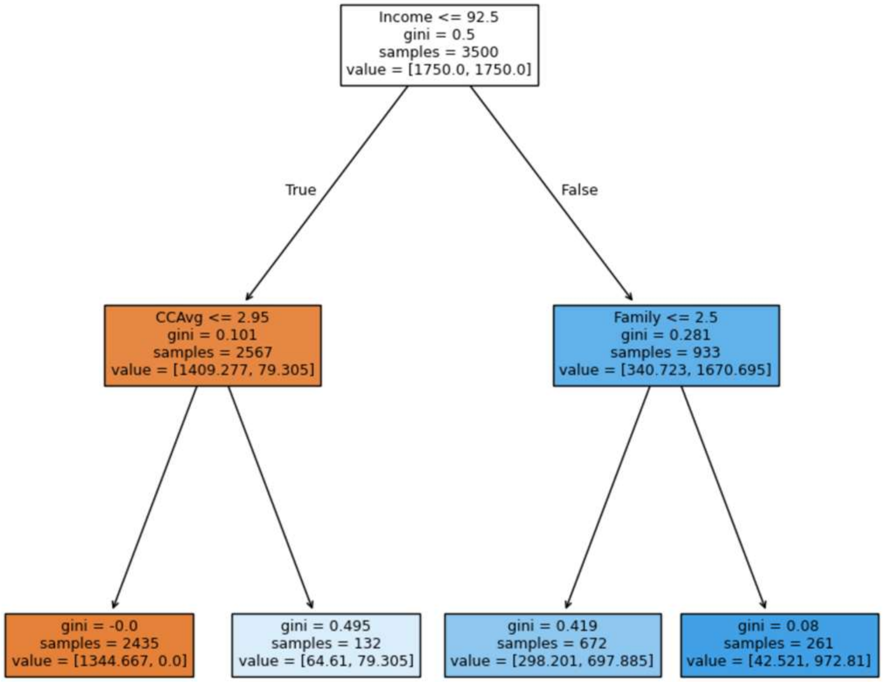




	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

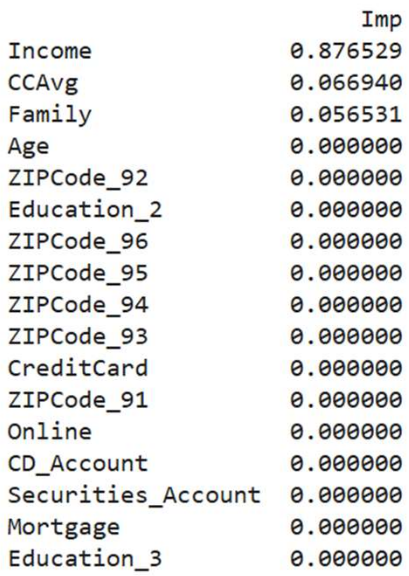
Post pruning the data set, the test set has better performance on the model.
Form the result it is clear that accuracy , recall .precision and F1 score are 1.

Post pruning Tree

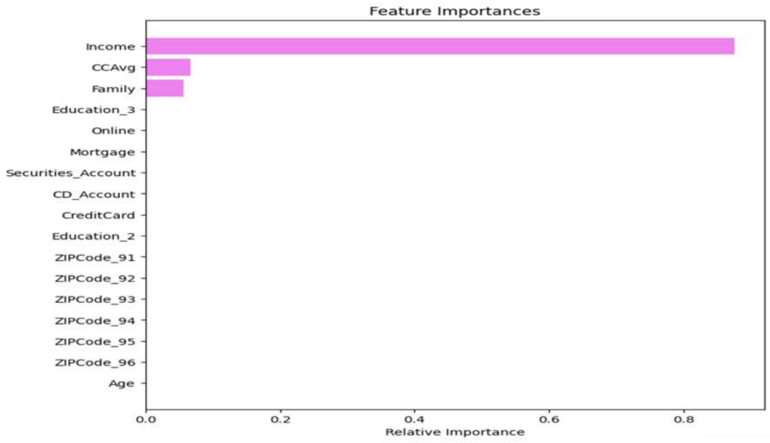


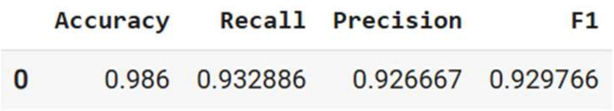
Model Performance Summary

Mean decrease in impurity or GINI importance is a way to measure how important a feature is in making decisions

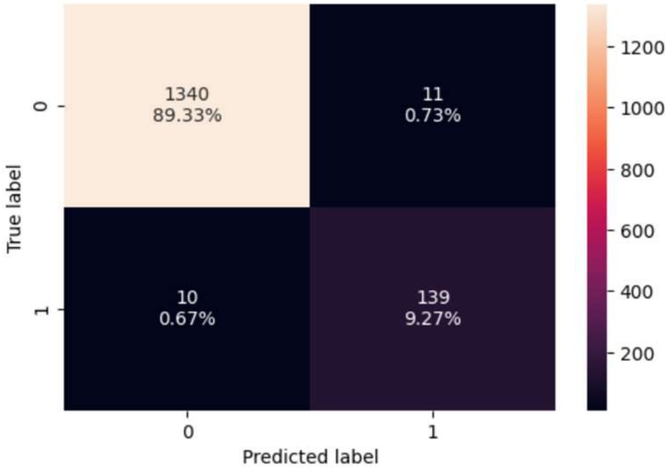


```
                         Imp
Income              0.876529
CCAvg               0.066940
Family              0.056531
Age                 0.000000
ZIPCode_92          0.000000
Education_2         0.000000
ZIPCode_96          0.000000
ZIPCode_95          0.000000
ZIPCode_94          0.000000
ZIPCode_93          0.000000
CreditCard          0.000000
ZIPCode_91          0.000000
Online              0.000000
CD_Account          0.000000
Securities_Account  0.000000
Mortgage            0.000000
Education_3         0.000000
```

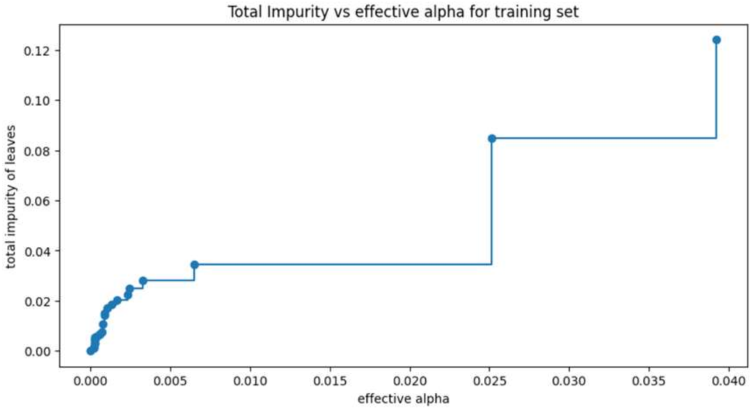




	Accuracy	Recall	Precision	F1
0	0.986	0.932886	0.926667	0.929766



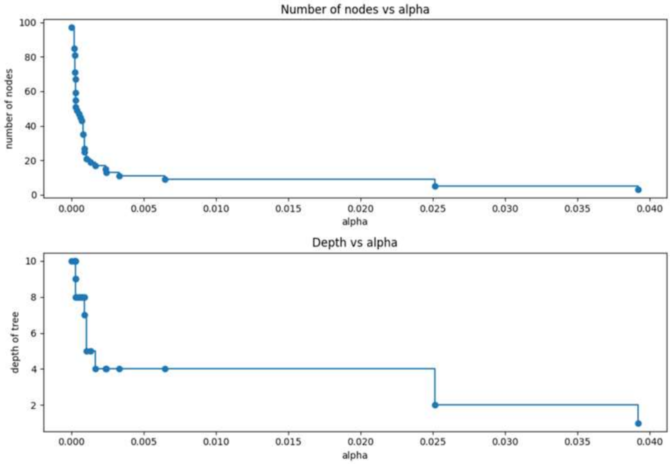
Model Performance Summary



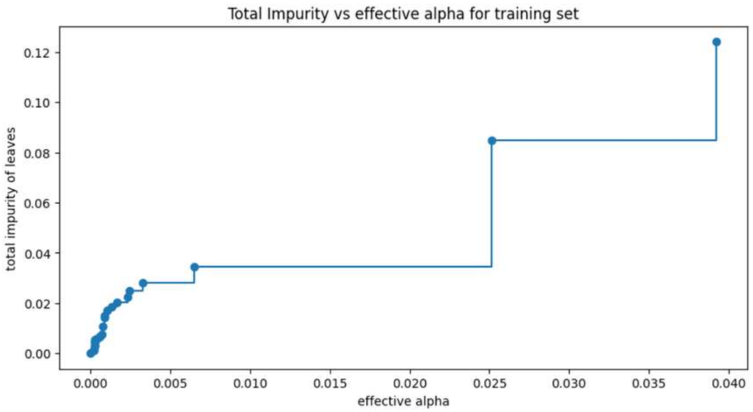
The **Number of Nodes vs. Alpha** plot provides a visual representation of how pruning impacts the size of the decision tree. Analyzing this plot, determine the level of pruning that minimizes tree size while still retaining a model that performs well on the data. It is a valuable tool for managing model complexity and preventing both overfitting and underfitting.

The **impurity vs. effective alpha plot** helps to visualize how pruning affects the complexity and performance of decision tree.

The point on the plot where the impurity curve flattens or starts to increase sharply is a good candidate for choosing ccp_alpha.
From the plot , 0.049 is taken as the effective alpha value to get a better generalizing model.



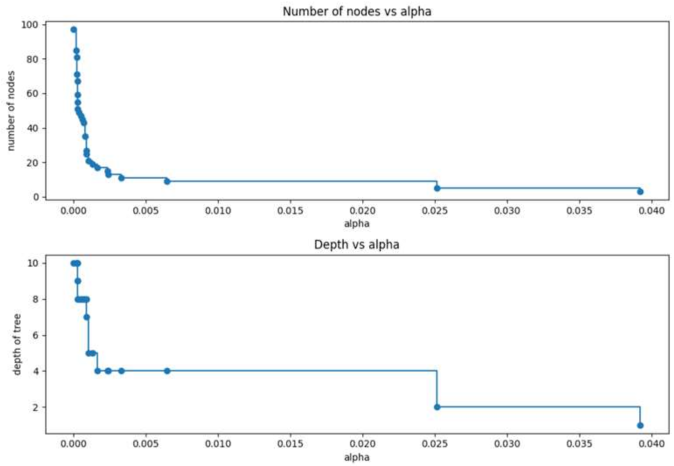
Model Performance Summary



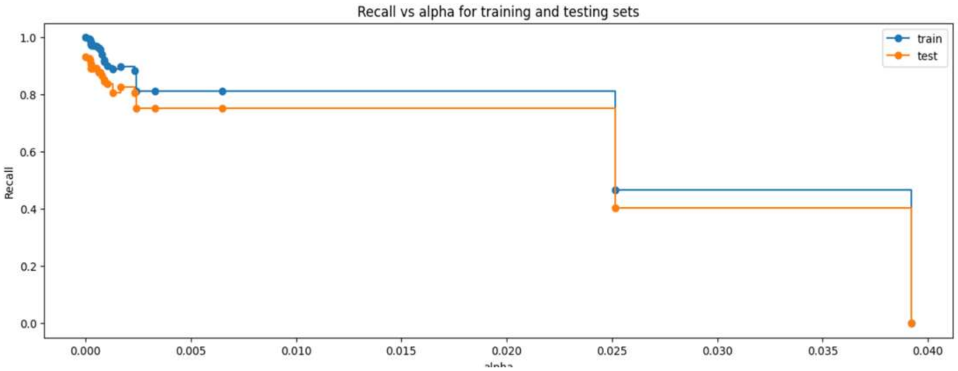
The **Number of Nodes vs. Alpha** plot provides a visual representation of how pruning impacts the size of the decision tree. Analyzing this plot, determine the level of pruning that minimizes tree size while still retaining a model that performs well on the data. It is a valuable tool for managing model complexity and preventing both overfitting and underfitting.

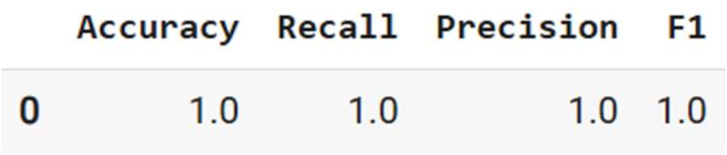
The **impurity vs. effective alpha plot** helps to visualize how pruning affects the complexity and performance of decision tree.

The point on the plot where the impurity curve flattens or starts to increase sharply is a good candidate for choosing ccp_alpha.
From the plot , 0.049 is taken as the effective alpha value to get a better generalizing model.



Model Performance Summary



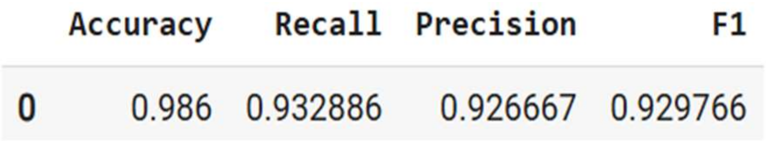


	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Decision_Tree_Tune_Post_Train

These values are used for the better understanding of the model performance .And creating a final model

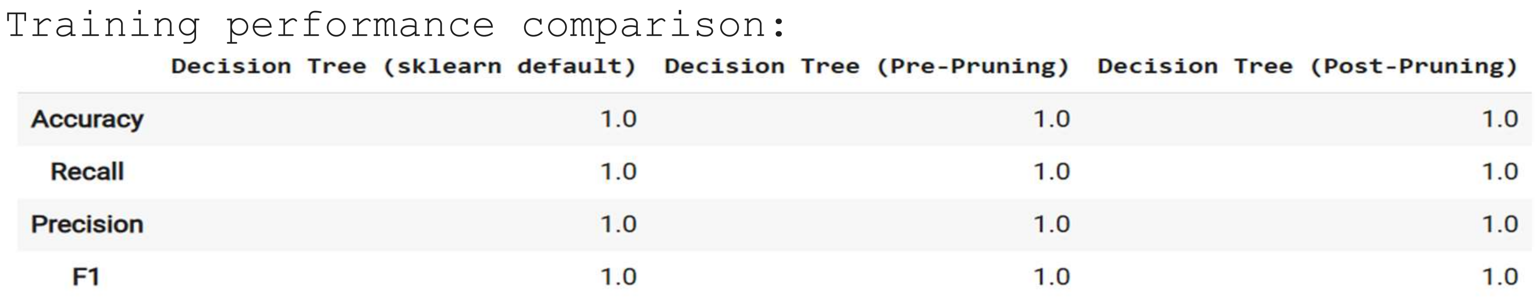
The **Recall vs. Alpha plot** is an essential tool for understanding how pruning affects the recall of your decision tree model. Analyzing this plot, can find the optimal amount of pruning (in terms of ccp_alpha) that maximizes recall on the test set without significantly underfitting the data. This process is vital for achieving a balanced model that generalizes well to new, unseen data while avoiding both overfitting and underfitting.



	Accuracy	Recall	Precision	F1
0	0.986	0.932886	0.926667	0.929766

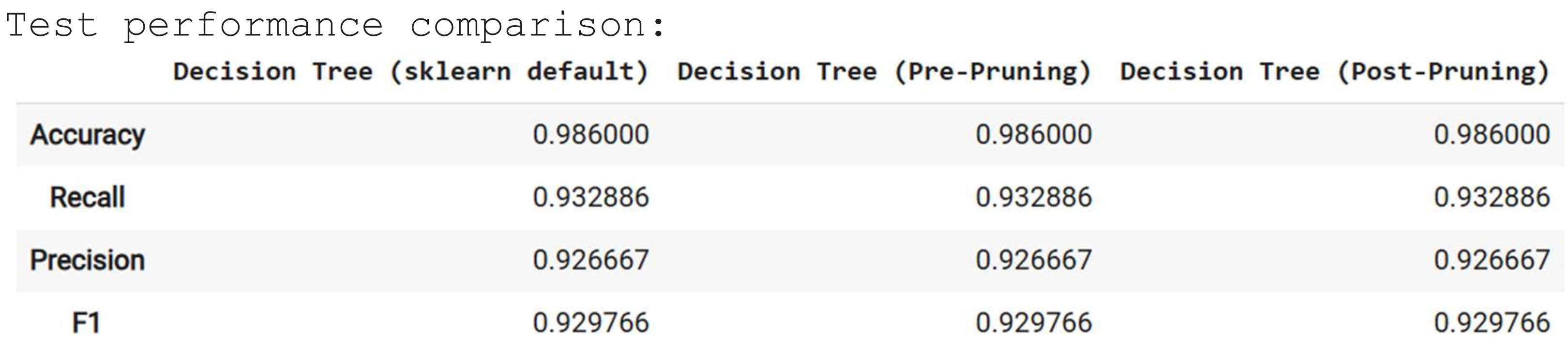
Decision_Tree_Tune_Post_Test

Model Performance Summary



Training performance comparison:

	Decision Tree (sklearn default)	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	1.0	1.0	1.0
Recall	1.0	1.0	1.0
Precision	1.0	1.0	1.0
F1	1.0	1.0	1.0



Test performance comparison:

	Decision Tree (sklearn default)	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.986000	0.986000	0.986000
Recall	0.932886	0.932886	0.932886
Precision	0.926667	0.926667	0.926667
F1	0.929766	0.929766	0.929766

From the performance matrix of the model, it is ready for validating the customer for loan.
This model is suitable to predict whether a liability customer will buy personal loans, to understand which customer attributes are most significant in driving purchases, and to identify which segment of customers to target more.

APPENDIX

Data Background and Contents

Data Background:

Background data is not directly seen but has the influence on the data collected.

-CCAVG: It has many hidden factors. Credit card usage, credit card balance.

-Family : It includes theat dependent family people and the independent family person.

-Credit card: Only holding another credit is not important .But the credit value in that card is also an influence factor for the loan.

Content:

"Content" is the actual information or data itself that is readily visible or presented to the user.
-**Education**: It demonstrate that the education level of the person has the direct impact on the financial income of the person . So to the credit card value directly.

-Experience: As the experience increases , there is a greater possibility of increase in the salary of the person. But at the same time , after certain age there will be either steady income or dependent income.This factor ahs to be considered.

Great Learning

Happy Learning !