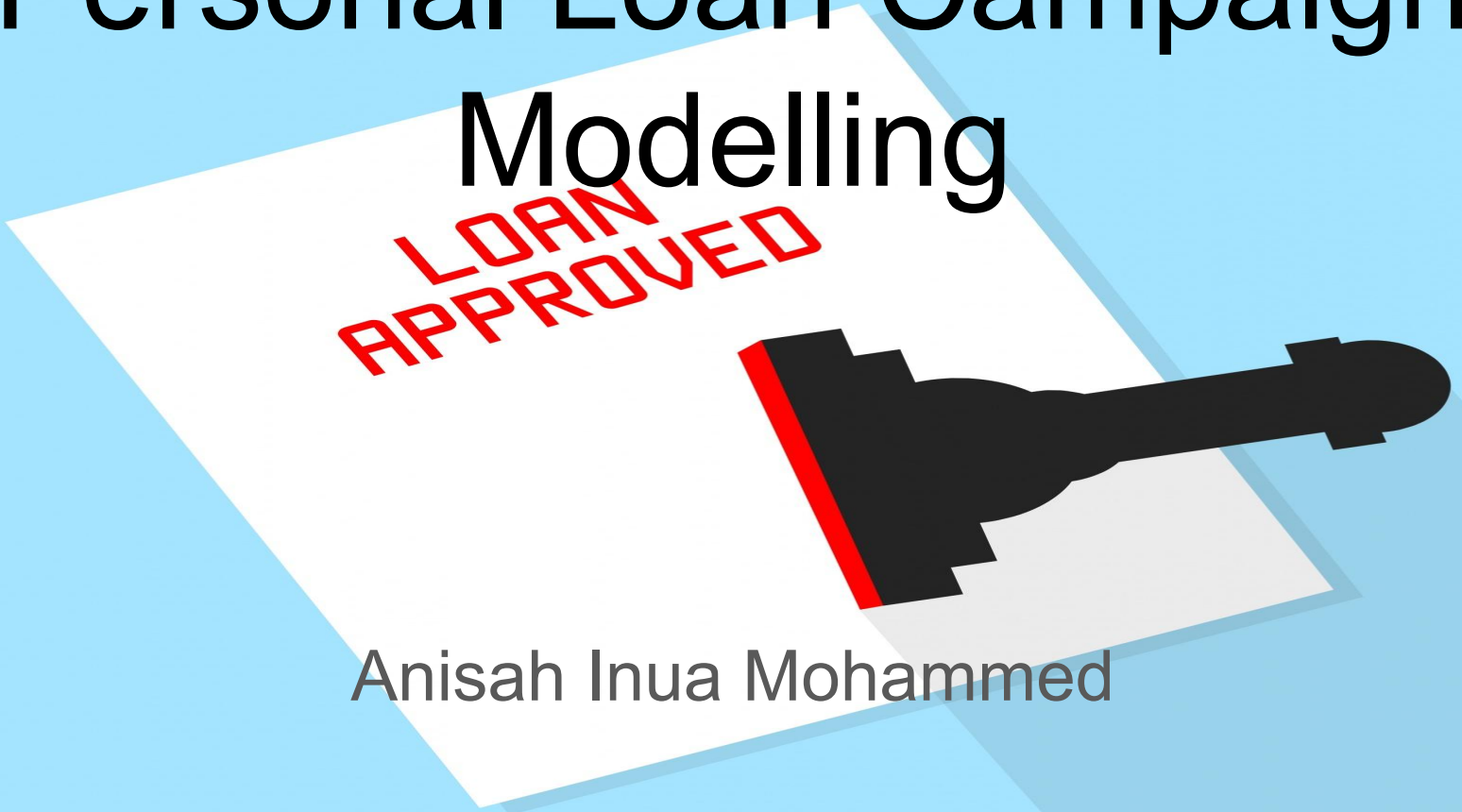
Personal Loan Campaign Modelling



Anisah Inua Mohammed

Description

The file contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Background

AllLife Bank has a growing customer base. Majority of these customers are liability customers (depositors) with varying size 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 with a minimal budget.

Objective from Analysis

The classification goal is to predict the likelihood of a liability customer buying personal loans which means we have to build a model which will be used to predict which customer will most likely to accept the offer for personal loan, based on the specific relationship with the bank across various features given in the dataset.

This means:

1. To predict whether a liability customer will buy a personal loan or not.
2. Which variables are most significant.
3. Which segment of customers should be targeted more.

Hypothesis Generation

Below are some of the factors which can affect the likelihood of a liability customer buying personal loans (dependent variable for this loan prediction problem):

Salary: Salary can be one of the major dependent variables as customers with high salaries are less feasible to buy personal loans while customers with medium or low salaries are more feasible for buying personal loans.

The number of family members: More the number of earning family members, less probability of buying personal loans.

Age: Customers with probably the age of 30–50 will buy personal loans.

Education of the customer: The customer is a graduate or under-graduate can affect the buying probability, people who are graduated or Advanced Professionals are more viable to buy personal loans from a bank rather than people who are under-graduated.

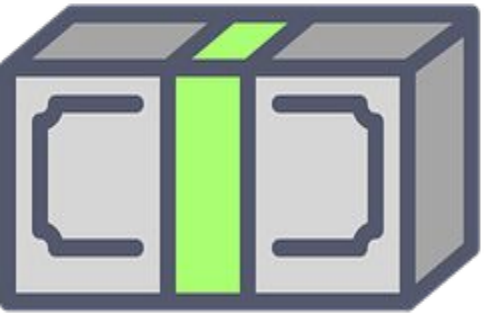
Data Dictionary

- 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: Avg. 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?
- Securities_Account: Does the customer have securities account with the bank?
- CD_Account: Does the customer have a certificate of deposit (CD) account with the bank?
- Online: Do customers use internet banking facilities?
- CreditCard: Does the customer use a credit card issued by Universal Bank?

Outcomes from the Dataset

- Exploratory Data Analysis
- Data Cleaning
- Data Visualization
- Preparing the data to train a model
- Training and making predictions using a classification model
- Model evaluation

Exploratory Data Analysis (EDA)



Viewing the first and Last rows for a brief overview

- The ID variable can be ignored as it will not any effect on our model.
- Target Variable is Personal Loan
- ZIP code of the customer, variable can also be ignored because we can not judge the customers based on their area or location.
- Education - Education level of the customer. In our dataset it ranges from 1 to 3 which are Under Graduate, Graduate and Postgraduate respectively.

View the first and last 5 rows of the dataset

```
In [5]: df.head()
```

Out[5]:

	ID	Age	Experience	Income	ZIPCode	Family	CCAvg	Education	Mortgage	Persor
0	1	25	1	49	91107	4	1.6	1	0	
1	2	45	19	34	90089	3	1.5	1	0	
2	3	39	15	11	94720	1	1.0	1	0	
3	4	35	9	100	94112	1	2.7	2	0	
4	5	35	8	45	91330	4	1.0	2	0	

```
In [6]: df.tail()
```

Out[6]:

	ID	Age	Experience	Income	ZIPCode	Family	CCAvg	Education	Mortgage	I
4995	4996	29	3	40	92697	1	1.9	3	0	
4996	4997	30	4	15	92037	4	0.4	1	85	
4997	4998	63	39	24	93023	2	0.3	3	0	
4998	4999	65	40	49	90034	3	0.5	2	0	
4999	5000	28	4	83	92612	3	0.8	1	0	

Shape of the Dataset

We have 13 independent variables and 1 dependent variable i.e. ‘Personal Loan’ in the data set. Also, we got 5000 rows which can be split into test & train datasets.

```
]: rows_count, columns_count = df.shape
   print('Total Number of rows :', rows_count)
   print('Total Number of columns :', columns_count)

   Total Number of rows : 5000
   Total Number of columns : 14
```

Datatype of each attribute

This shows that all the variables are numerical. But the columns 'CD Account', 'Online', 'Family', 'Education' , 'CreditCard' and 'Securities Account' can be categorical variables which should be converted into 'category' type.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 14 columns):
 #   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)
memory usage: 547.0 KB
```

Checking For Missing Values

It reveals that there are no missing values in the dataset.

```
df.isnull().sum()
```

```
ID                    0
Age                   0
Experience            0
Income                0
ZIPCode               0
Family                0
CCAvg                 0
Education             0
Mortgage              0
Personal_Loan         0
Securities_Account    0
CD_Account            0
Online                0
CreditCard            0
dtype: int64
```

```
df.isnull().any()
```

```
ID                    False
Age                   False
Experience            False
Income                False
ZIPCode               False
Family                False
CCAvg                 False
Education             False
Mortgage              False
Personal_Loan         False
Securities_Account    False
CD_Account            False
Online                False
CreditCard            False
dtype: bool
```

Pairplot which includes all the columns in the dataframe

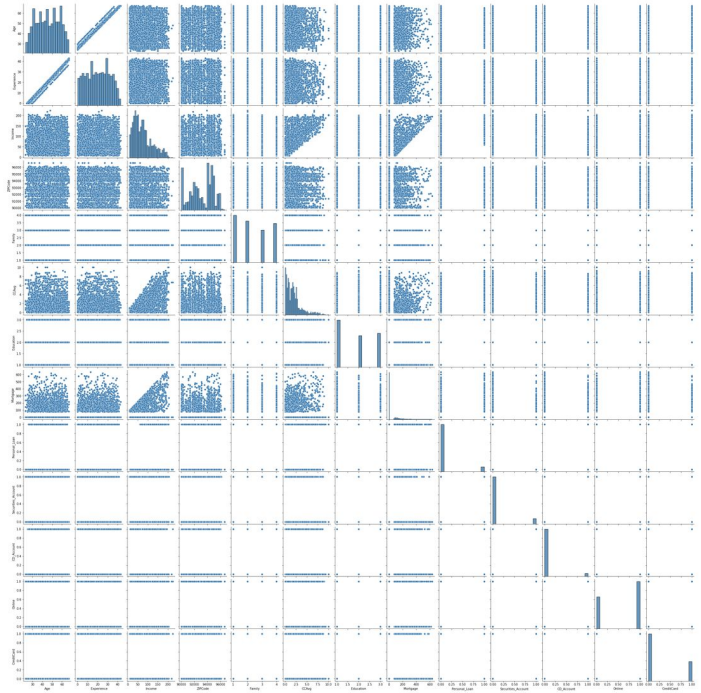
Observations

From the above pair plot we can infer the association among the attributes and target column as follows:

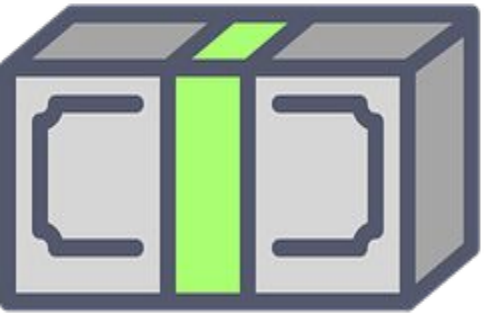
- ID: This attribute can be dropped.Though the data distribution is normal.
- Age:Three small peaks can be indicating three values of age would be slightly more in number.However, the mean and median of the attribute is equal.The distribution is normal. Most of the customers age is between 25 to 65 years.
- Education : Mean and median is almost equal. Data is finely distributed. A few peaks shows different values dominance.It also has low association with the 'Personal Loan'.
- Income : It is positively skewed and it will also have the outlier.Data for less income customers is more in the sample.
- ZIP Code: The attribute has sharp peaks telling the data from particular places are collected more.Spread is also less in the sample. More data from different places can be collected. Which means that 'Zip Code' does not really have any relationship with other variables.
- Family: It has 4 peaks(4 values) , families with least member is highest in the sample. It also has low association with the 'Personal Loan'.
- Mortgage: This attribute is highly left skewed with a very high peak on the left telling us that most customer are having least mortgage while a very few have some mortgage. It is also positively skewed. Majority of the individuals have a mortgage of less than 40K.
- Securities Account : This attributes tells us that majorly customers are not having Security account.
- CD account: Most of the customers don't have CD accounts.
- Online: Higher number of customers use online banking in the sample.
- Credit Card: This attribute has less customers using CC in comparison to the CC users.
- The distribution of CCAvg is also a positively skewed variable. Majority of the customers average monthly spending is between 1k to

```
sns.pairplot(df.iloc[:,1:])
```

```
<seaborn.axisgrid.PairGrid at 0x7f99742362e0>
```



Univariate & Multivariate Analysis

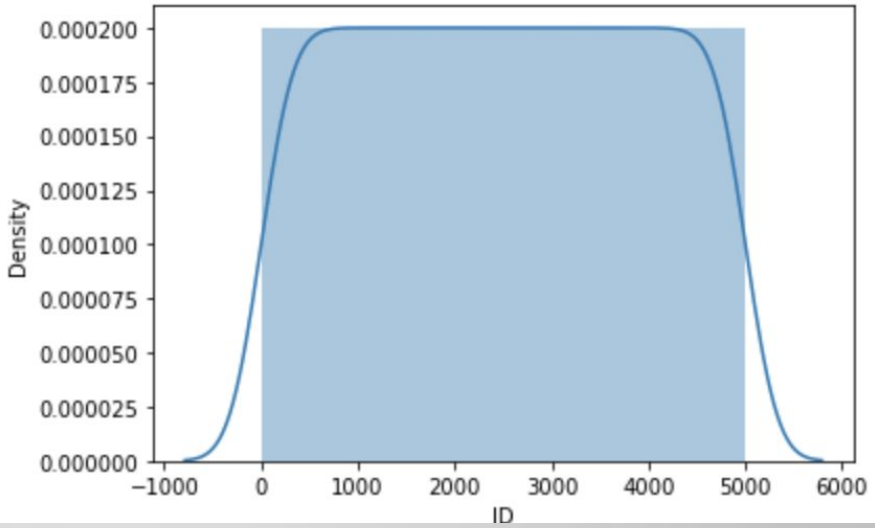


ID

From the graph it displays that 'ID' is uniformly distributed.

```
sns.distplot(df['ID'])
```

```
<AxesSubplot:xlabel='ID', ylabel='Density'>
```

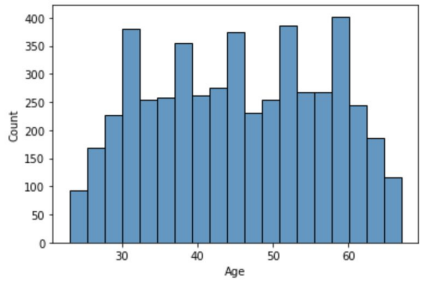


Age

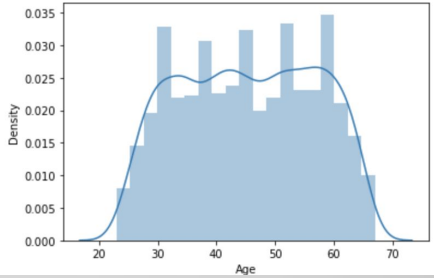
- This reveals that Age column is normally distributed
- Experience is highly correlated with Age (ρ = 0.994214857)
- The mean age of the customers is 45 with standard deviation of 11.5. Also, we had estimated the average age in hypothesis testing between 30–50. The curve is slightly negatively skewed (Skewness = -0.02934068151) hence the curve is fairly symmetrical
-

```
sns.histplot(df['Age'])
```

```
<AxesSubplot:xlabel='Age', ylabel='Count'>
```

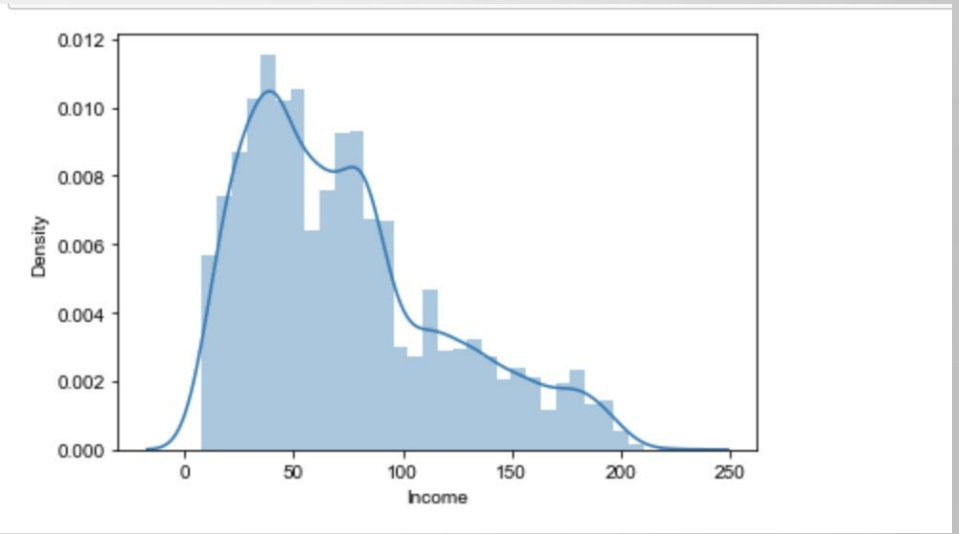



```
sns.distplot(df['Age'])
plt.show()
```



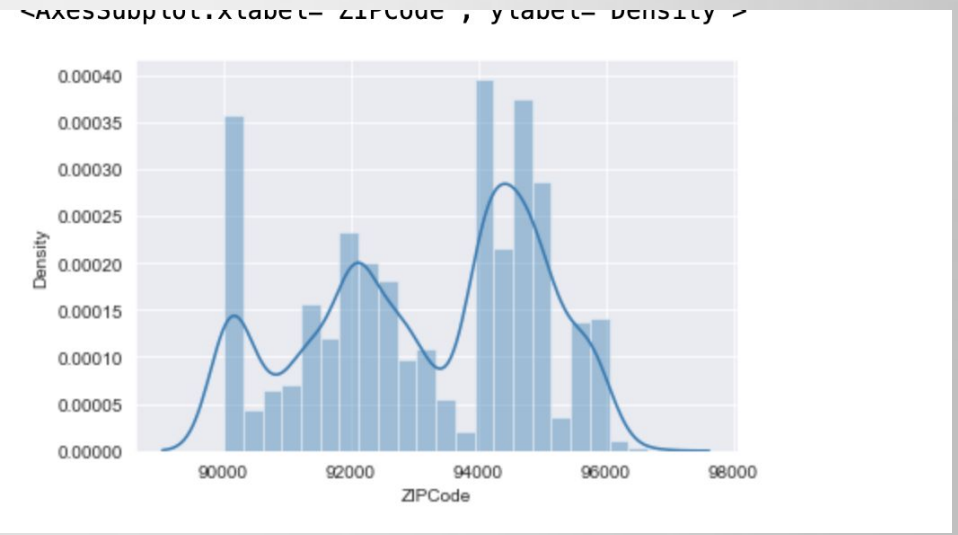
Income

- The distribution is right skewed distribution because the tail goes to the right.
- The mean annual income of the customer is 73.77 with standard deviation of 46. The curve is moderately positive skewed (Skewness = 0.8413386073)



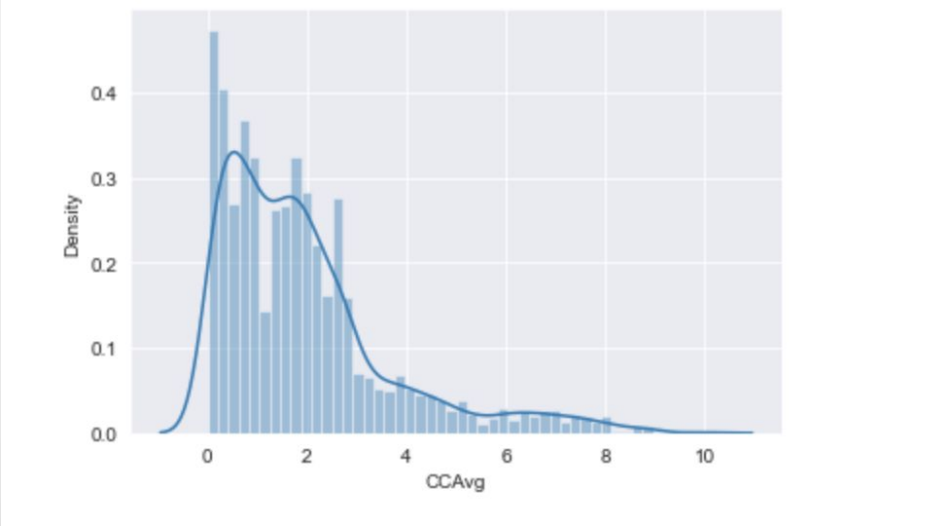
Zip Code

- The is uniformly distributed. Data points are more with family size 1 and 2.
-



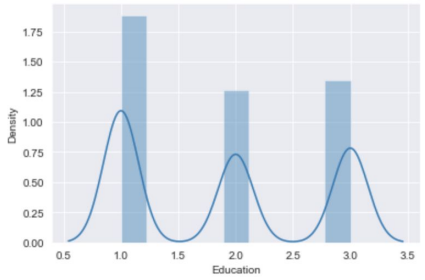
CCAvg

It shows that the distribution is rightly skewed because the tails is at the right. This means that most of the customers' monthly average spending on credit cards is usually between 1k to 2.5k. However, there are few customers whereby their monthly average spending on credit card is >8k.

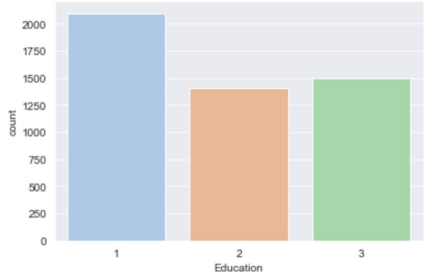


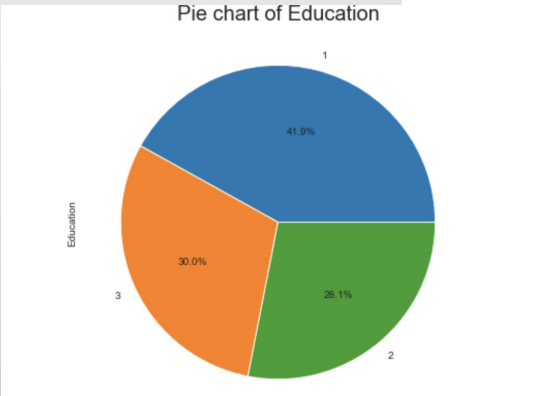
Education

- 1 stands for undergrad level customers, 2 stands for Graduate level customers and 3 stands for Advanced/Professionals level customers.
- The above graphs show that Undergrad level customers are more than the Graduate and Advanced/Professional customers.
- 42% of the candidates are graduated, while 30% are professional and 28% are Undergraduate.
-



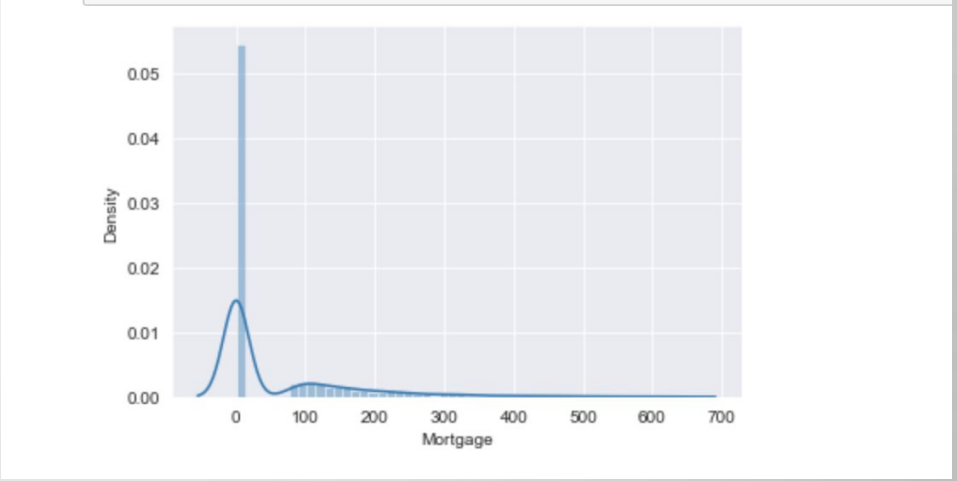

```
[36]: sns.countplot(df['Education'],palette='pastel')
      plt.show()
```





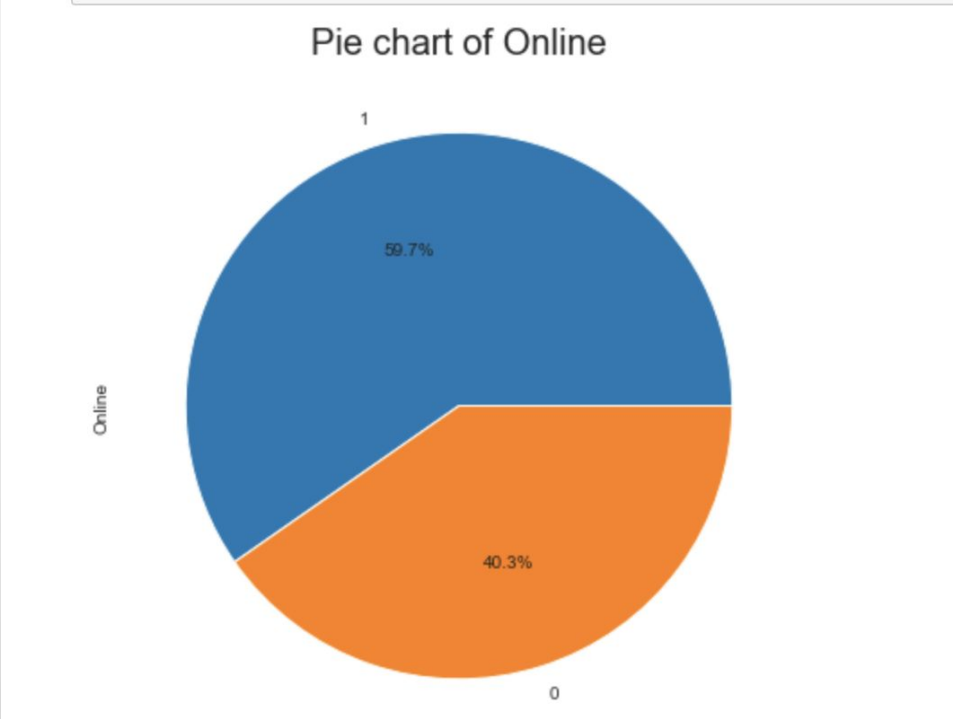
Mortgage

- The above distribution is also right skewed distribution because the tail goes to the right. Most of the customers do not have mortgage. There are more customers whose mortgage amount is between 80000 to 150000 . Very few customers whos mortgage amount is more than $600000.
- The mean value of house mortgage is 56.5 with standard deviation of 101.71. The curve is highly positive skewed (Skewness = 2.104002319) and there are a lot of outliers present (Kurtosis = 4.756796669)
-



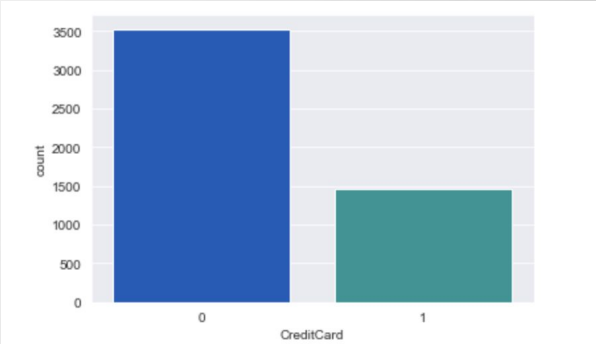
Online

- This is a bit of an unusual distribution but I can deduce that the number of customers who own an Online account is greater than of those customers who do not own online accounts.
- Around 60% of customers use internet banking facilities.
-



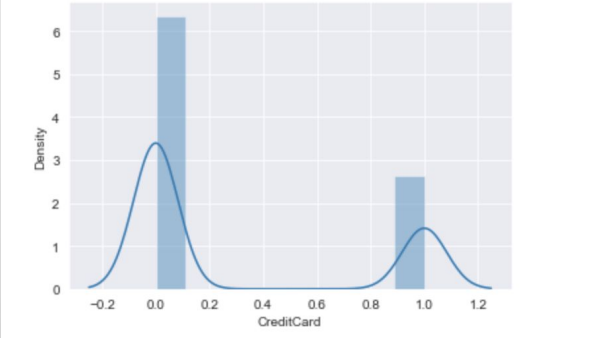
Credit Card

- The number of customers without Credit Cards way more than the number of customer with Credit Cards
- Around 71% of the customer doesn't use a credit card issued by UniversalBank.
- The mean of Avg. spending on credit cards per month is 1.93 with standard deviation of 1.75. The curve is highly positive skewed (Skewness = 1.598443337)
-




```
sns.distplot(df['CreditCard'])
```

```
<AxesSubplot:xlabel='CreditCard', ylabel='Density'>
```



Personal Loan (Target Variable)

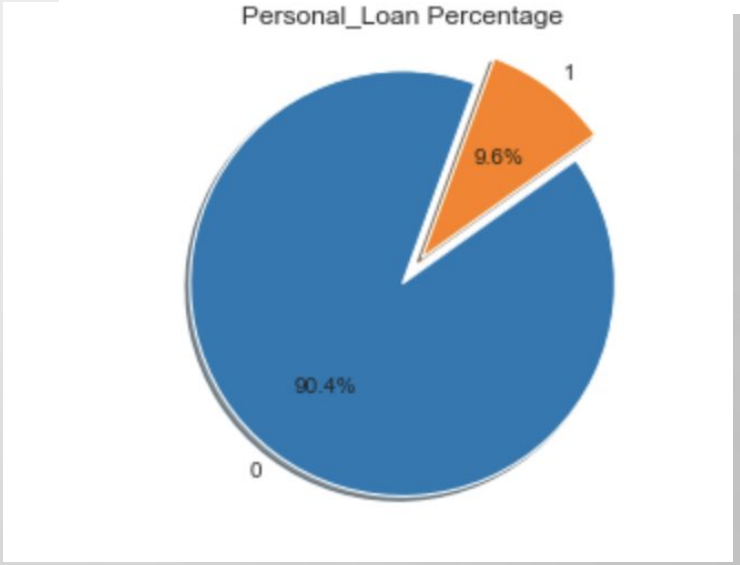
This reveals that out of 5000 data points:

- 4520 are being labelled as 0 and 480 are being labelled as 1.

This could mean that the percentage of customers who took the loan is significantly greater than customers who did take the loan.

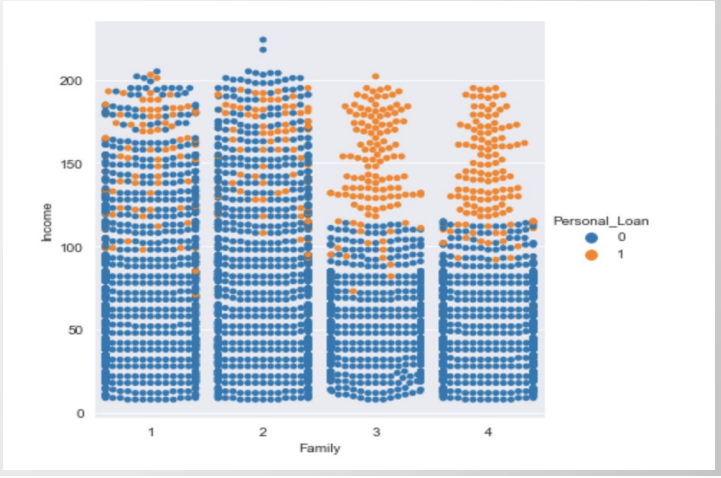
From the pie chart it shows that the data is really biased (almost 1:10) in respect to the customers in category of not accepting personal loan. Therefore we could build an assessment model which could perform better towards predicting which customers will either be accepting personal loans or not. Hence, our goal should be to identify the customers who can accept personal loans based on the given features.

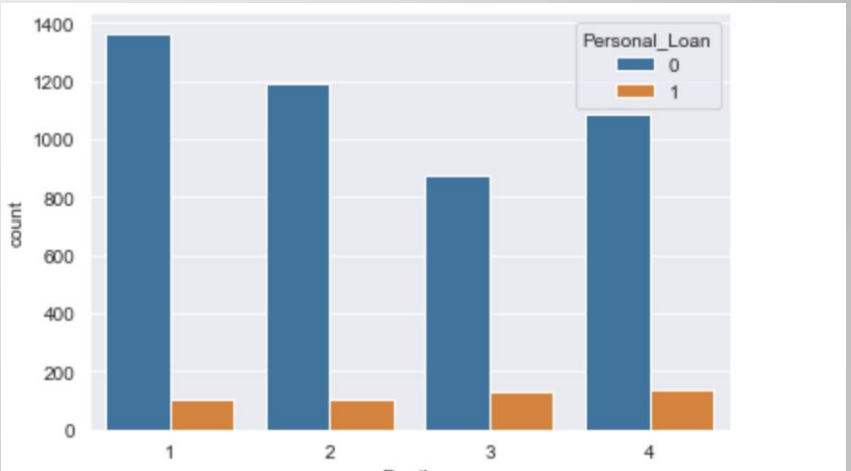
	Labels	Personal_Loan
0	0	4520
1	1	480



Family vs Income with hue:Personal Loan

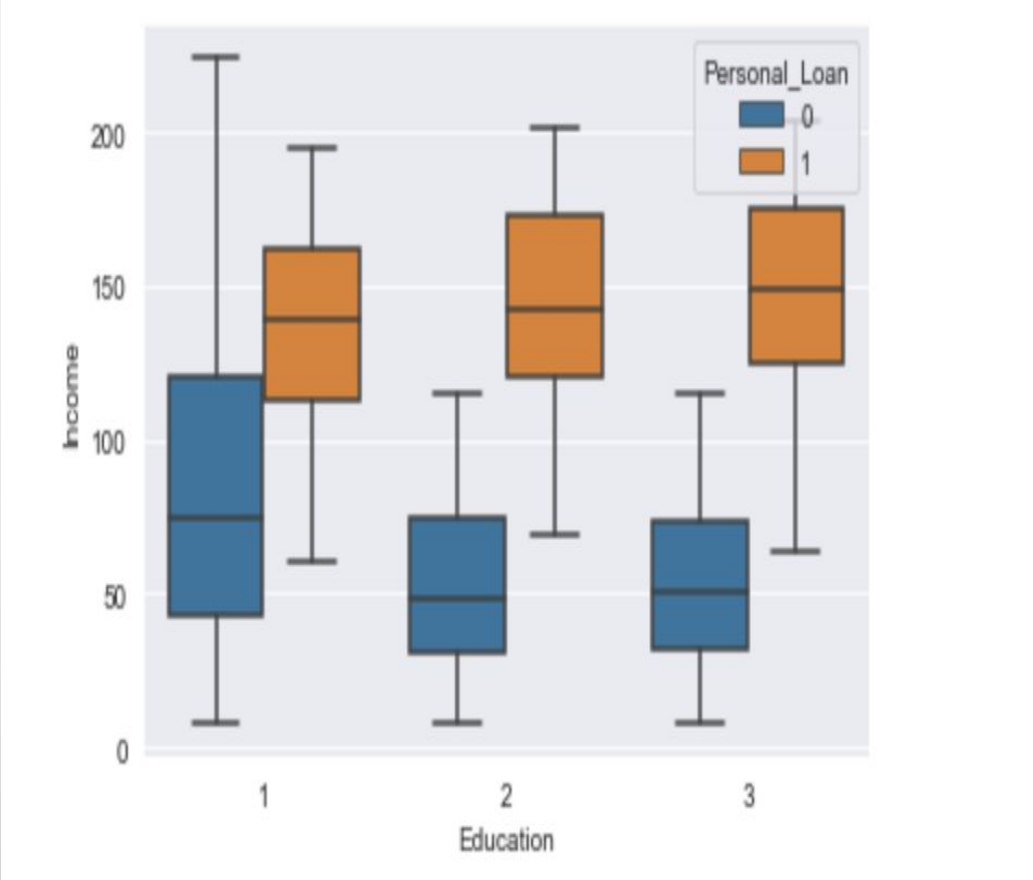
- Customers who have family size 3 or greater with higher income between 100k to 200k are more likely to take loan.
- Family size does not have any impact in personal loan. But it seems families with size of 3 and 4 are more likely to take loan.
- The number of family members not significantly affect probability. Hence it contradicts our hypothesis that the number of family members will affect the probability.





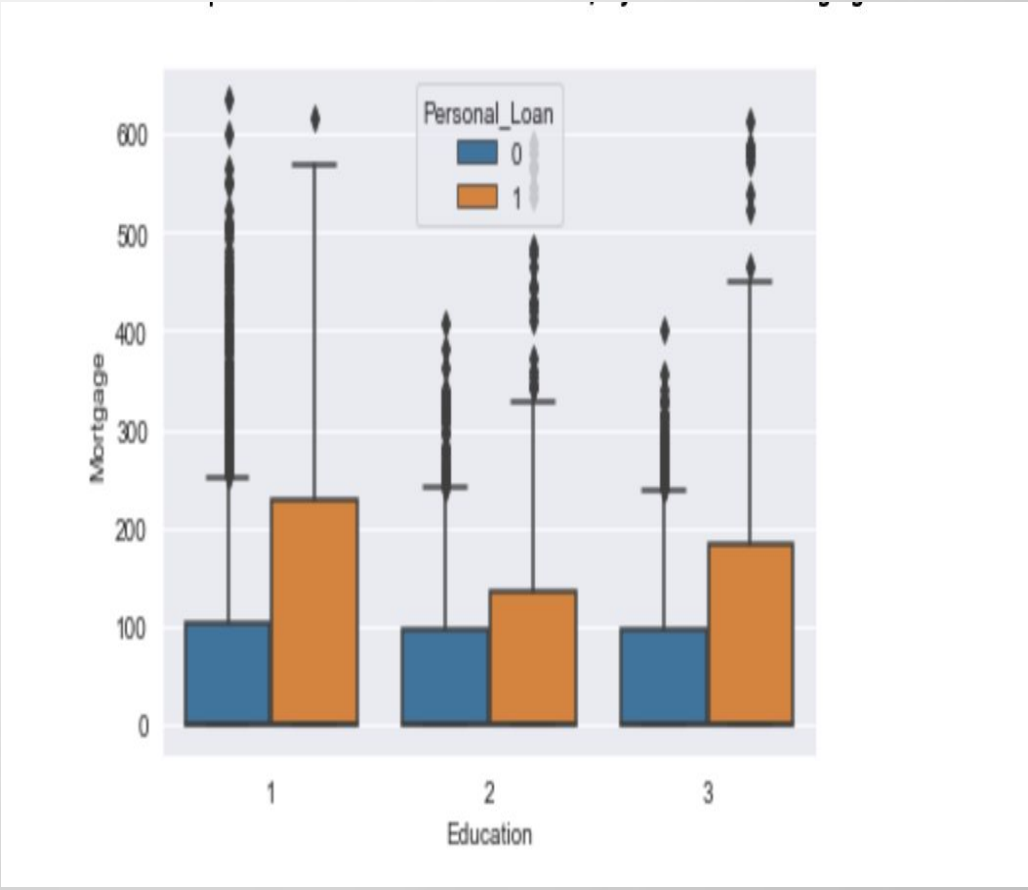
Education vs Income with hue: Personal Loan

- It displays that customers with undergraduate level of education and family greater than 3 are good customers who took loan. Customer who took loan have same income range irrespective of education level. Education of Graduate and above have more chance to take loan.
- The customer is a graduate or undergraduate can affect the buying probability, people who are graduated or Advanced Professionals are more viable to buy personal loans from a bank rather than people who are undergraduate.
-



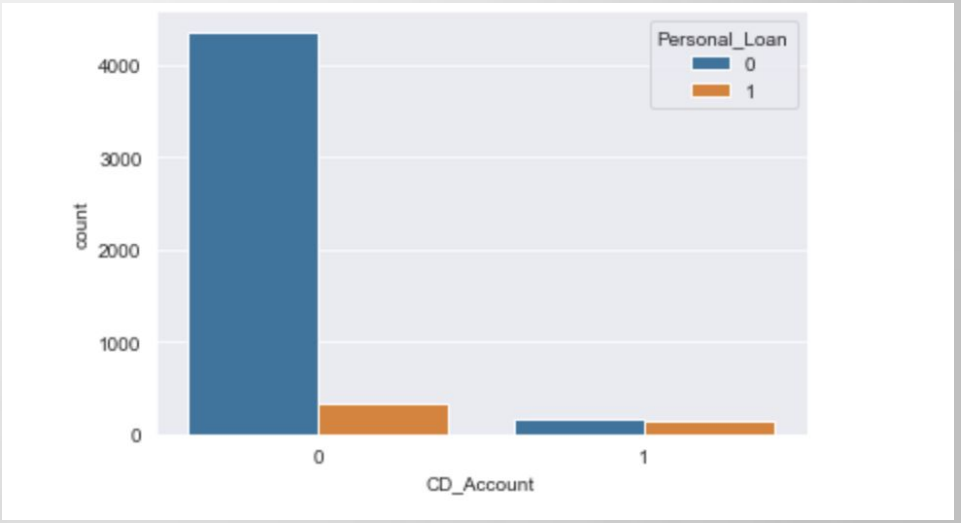
Education vs Mortgage with hue: Personal Loan

- It can be deduced that that customers whose education level is 1 and did not take loan has higher mortgage than customers who take loan of same education level. Customers whose education level is 2 and 3 and did not take loan has lesser mortgage than customers who take loan of same education level.
- Cross tabulation bar plot we can infer that customers who are more educated have a higher probability of buying personal loans. Hence our hypothesis was true.



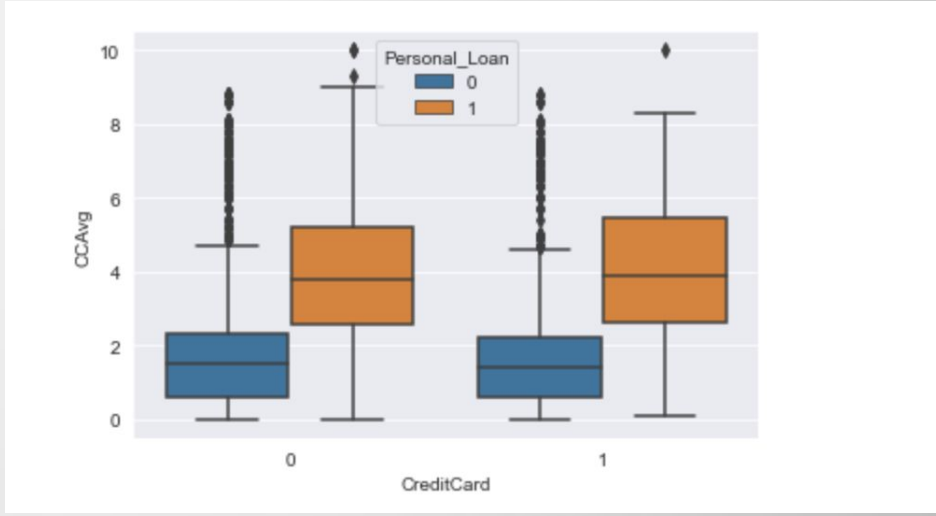
CD Account vs Personal Loan

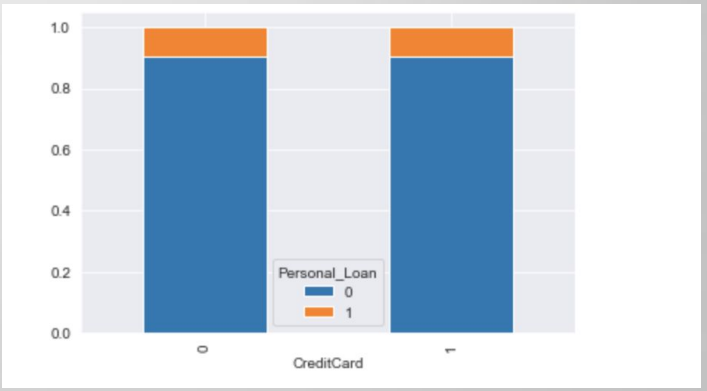
- Customers who do not have CD account , do not have loan as well. This seems to be majority, but almost all customers who has CD account has loan as well.
- The customer who has a certificate of deposit (CD) account with the bank seems to buy personal loans from the bank.



Credit Card vs CCAvg with hue: Personal Loan

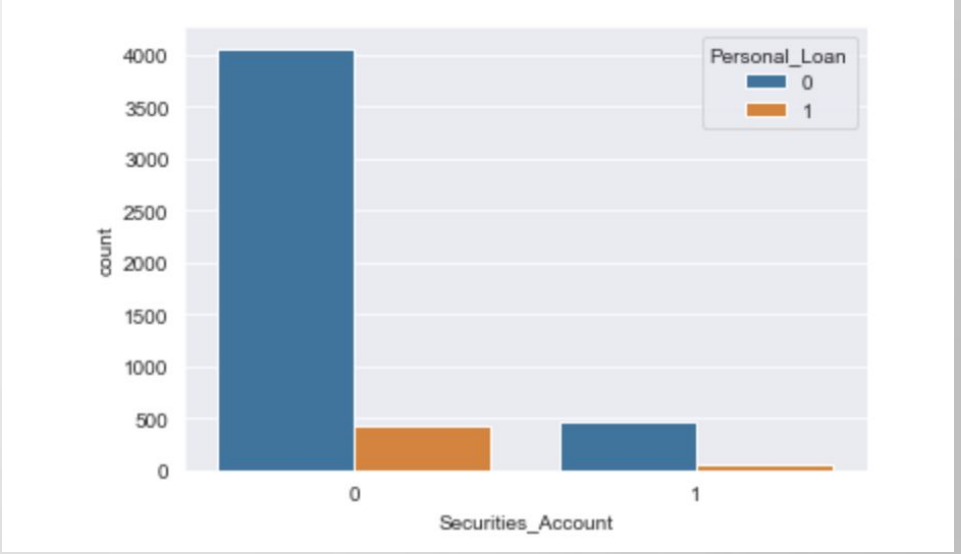
- Customers who have credit card and monthly spending is higher are more likly to take loan.
- The customer who uses or doesn't use a credit card issued by UniversalBank doesn't seem to affect the probability of buying a personal loan.





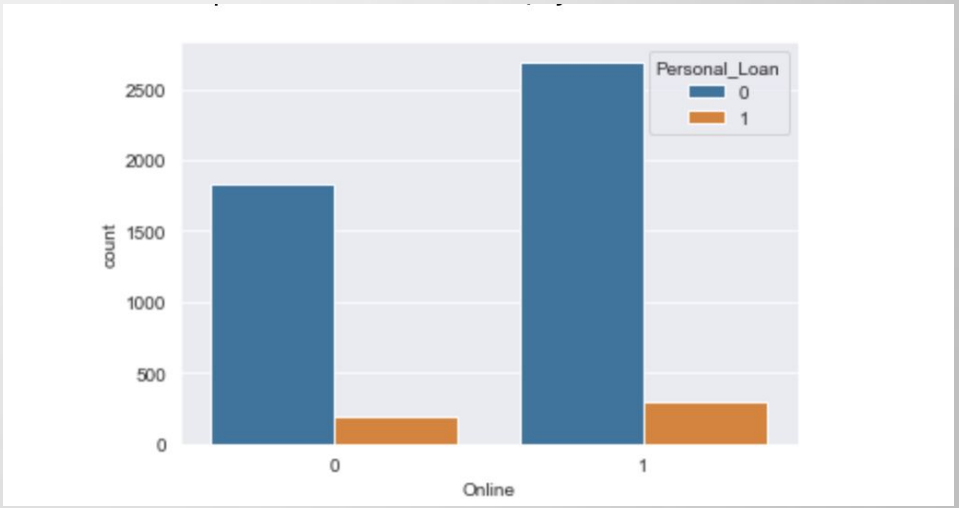
Securities Account vs Personal Loan

- Customers who has securities account are more likely to take loan
- Majority of customers who does not have loan do not have securities account.
- The customers who have or don’t have securities account with the bank do not affect the probability of buying a personal loan.



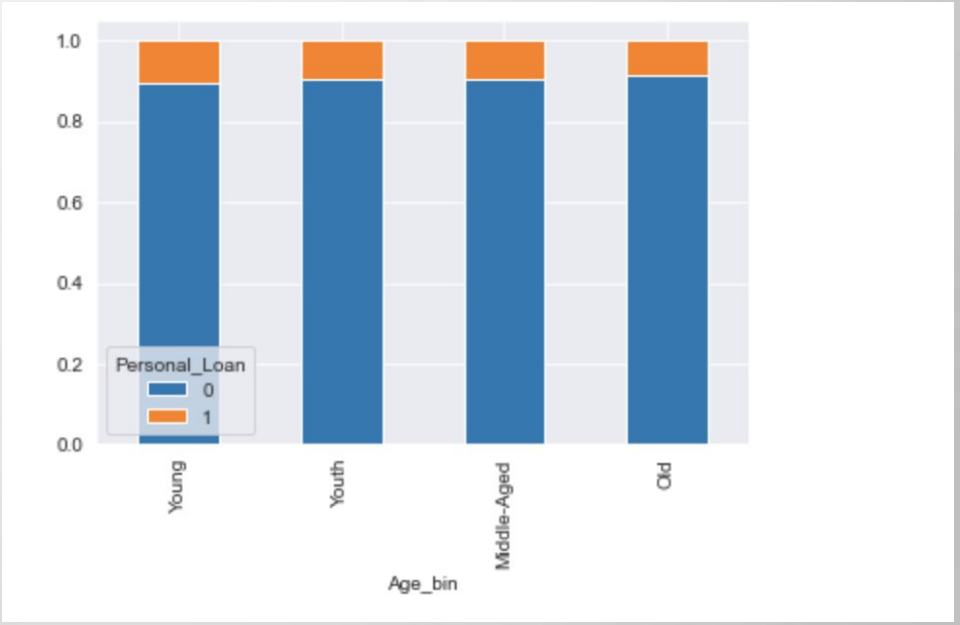
Online vs Personal Loan

- The customer who uses or doesn’t use internet banking facilities seems to not affect the probability of buying personal loans.



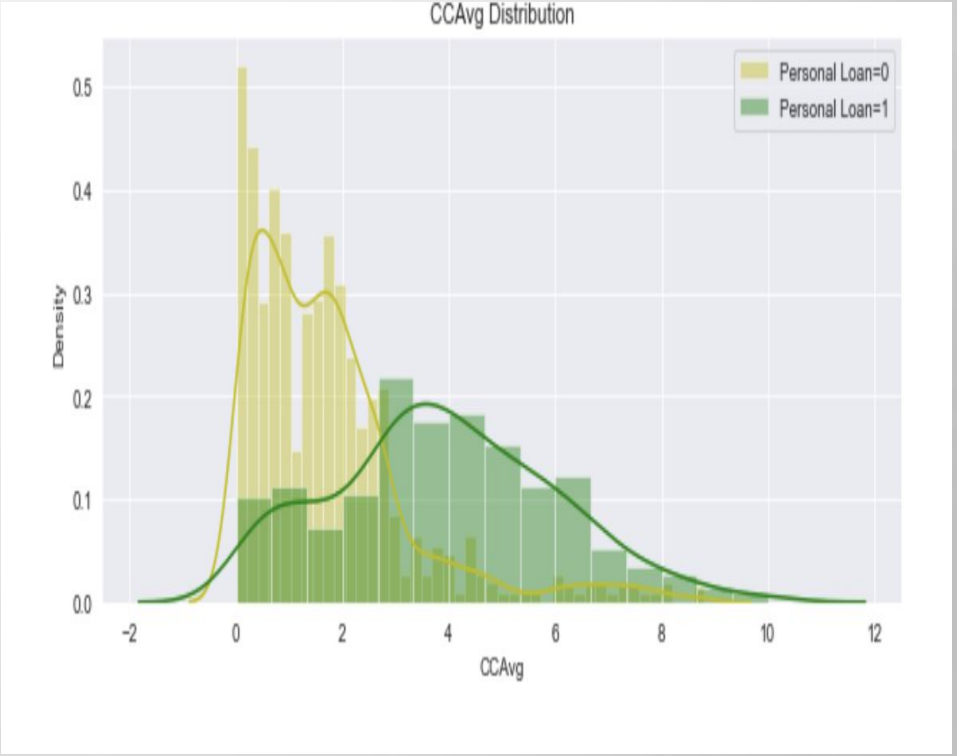
Age vs Personal Loan

- There is not much change in the mean of age. Which is why I decided to create bins for the customers to have a clearer view and to make analysis for the loan status for each age bin.
- From the second graph it displays Age and Experience have a strong correlation. This could mean as Age increases, Experience will also increases.
- It could be deduced that the Age of the customer applying for a Personal Loan will not affect the chances of acquiring a Personal Loan which counters the hypothesis which we assumed that the Age of people applying for Personal Loan is a major factor while actually buying a Loan.



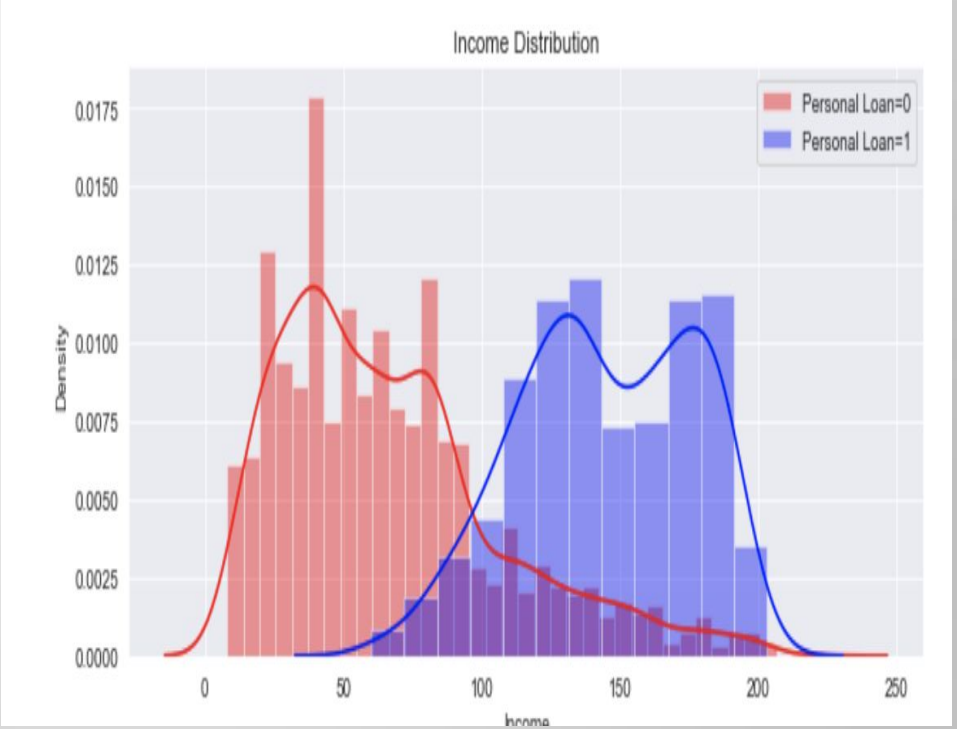
CCAvg vs Personal Loan

- This shows that customers who have taken personal loan have higher credit card average than those who did not take loan. So high credit card average seems to be good predictor of whether or not a customer will take a personal loan.
- Shows that customers who have taken Personal Loans have a high Credit Card Average
- Average credit card spending with a median of 3800 dollar indicates a higher probability of personal loan
- Lower credit card spending with a median of 1400 dollars is less likely to take a loan.



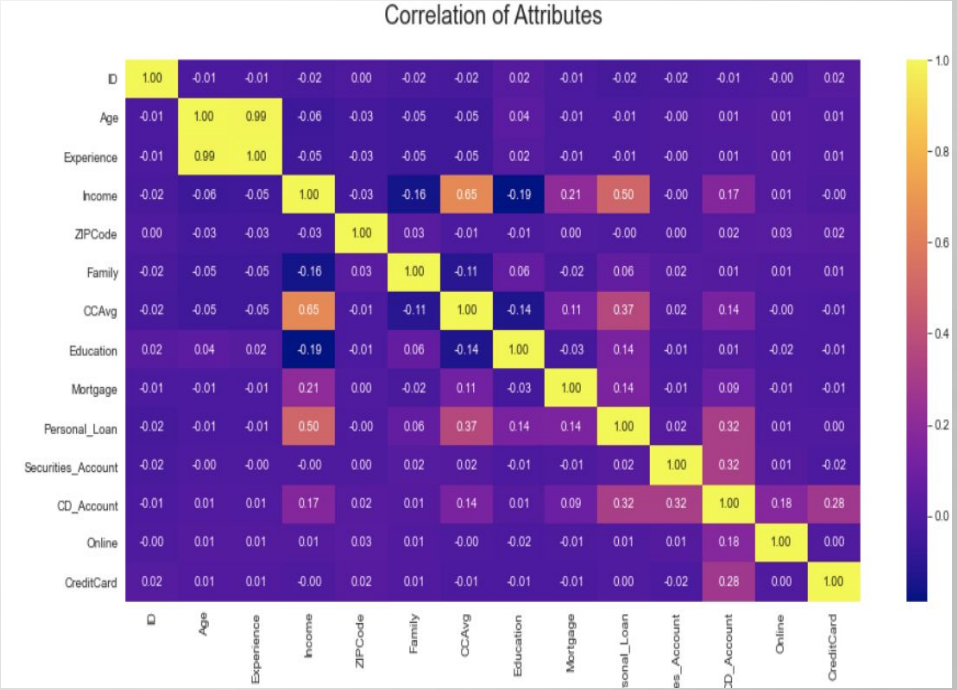
Income vs Personal Loan

- This still just proves the point made in the graph above clearer which is customers who have taken personal loan have income than those who did not take. So high income seems to be good predictor of whether or not a customer will take a personal loan.
-

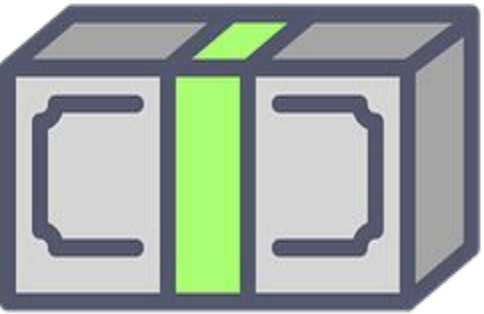


Correlation Using Heatmap

- Age and Experience are highly correlated and the correlation is almost 1.
- 'Income' and 'CCAvg' is moderately correlated.
- Personal Loan has maximum correlation with 'Income', 'CCAvg', 'CD Account', 'Mortgage', and 'Education'.
- It can be deduce in above heat map there is association of 'CD Account' with 'Credit Card', 'Securities Account', 'Online', 'CCAvg' and 'Income'.
- 'Mortgage' has moderate correlation with 'Income' which is about 12%.
- 'Income' influences 'CCAvg', 'Personal Loan', 'CD Account' and 'Mortgage'.



Model Building



Logistic Regression

Age' and 'Experience' are highly correlated so we will build our model **with 'Experience'** and **without 'Experience'** after that we will compare the accuracy which will lead us to the conclusion that with 'Experience' or without 'Experience' which model is better for prediction.

I have chosen to use experience because it is preferably easier for me because it is categorised into TWO that is:

- Customers taking Personal Loan with Experience
- Customers taking Personal Loan without Experience

Created a model Creating two new dataframes for the model building which are 'With Experience' and 'Without Experience' mutually.

Then separated the target variable the independent variable

Then splitting the dataset into a test and training Dataset

In [113]:
```
# From Expreneece Dataframe:
X_Expr = ploan_with_experience.drop('Personal_Loan', axis=1)
Y_Expr = ploan_with_experience[['Personal_Loan']]
```

In [114]:
```
# From Without Expreneece Dataframe:
X_Without_Expr = ploan_without_experience.drop('Personal_Loan', ax
Y_Without_Expr = ploan_without_experience[['Personal_Loan']]
```

Spliting the data into training and test set in the ratio of 70:30 respectively

In [115]:
```
# From Experience Dataframe:
X_Expr_train, X_Expr_test, y_Expr_train, y_Expr_test = train_test_
print('x train data {}'.format(X_Expr_train.shape))
print('y train data {}'.format(y_Expr_train.shape))
print('x test data  {}'.format(X_Expr_test.shape))
print('y test data  {}'.format(y_Expr_test.shape))
```

```
x train data (3426, 11)
y train data (3426, 1)
x test data  (1469, 11)
y test data  (1469, 1)
```

In [116]:
```
ithout Experience Dataframe:
X_test, y_train, y_test = train_test_split(X_Without_Expr, Y_Withe
train data {}'.format(X_train.shape))
train data {}'.format(y_train.shape))
test data  {}'.format(X_test.shape))
test data  {}'.format(y_test.shape))
```

```
x train data (3426, 10)
y train data (3426, 1)
x test data  (1469, 10)
y test data  (1469, 1)
```

Logistic Regression Model Accuracy and Confusion Matrix With 'Experience' and Without 'Experience'

- The above accuracy results that show that the accuracy is higher With Experience which is 94.82% than Without Experience which was 94.69%
- We can also see from the confusion matrix that the prediction of customers who do not have accept loan and the customers who accept loan is better With Experience.
- Therefore, to improve the accuracy by scaling the attributes
- I will not be considering the second Data Frame which is "Without Experience" for further iteration
- Type 1 (False Positive) and Type 2(False Negative) errors is less with experience.
-

```
Logistic Regression Model Accuracy Score W/O Experience  : 0.946
903
Logistic Regression Model Accuracy Score With Experience : 0.948
264

Logistic Regression Confusion Matrix W/O Experience:
 [[1323   18]
 [  60   68]]

True Possitive    =  68
True Negative     =  1323
False Possive     =  18
False Negative    =  60

Logistic Regression Confusion Matrix With Experience:
 [[1325   16]
 [  60   68]]

True Possitive    =  68
True Negative     =  1325
False Possive     =  16
False Negative    =  60
```

Final Analysis of Logistic Regression

	precision	recall	f1-score	support
0	0.94	0.98	0.96	893
1	0.62	0.40	0.48	86
accuracy			0.93	979
macro avg	0.78	0.69	0.72	979
weighted avg	0.92	0.93	0.92	979

0.9254341164453525

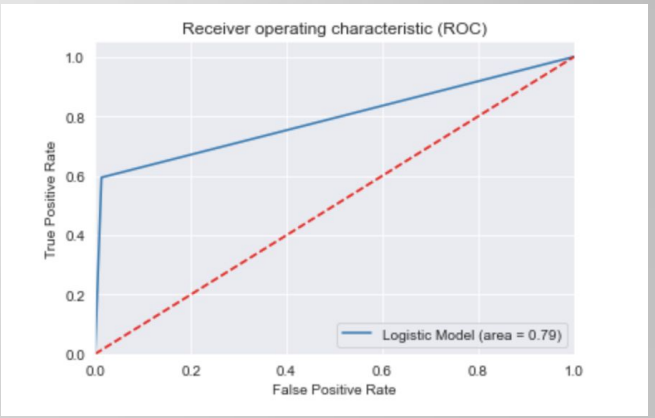
After Scalling Logistic Regression Model Accuracy Score with Experience: 0.952349

After Scalling Logistic Regression Confusion Matrix With Experience:

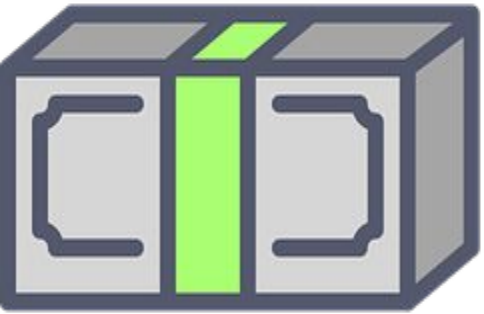
```
 [[1323   18]
 [  52   76]]
```

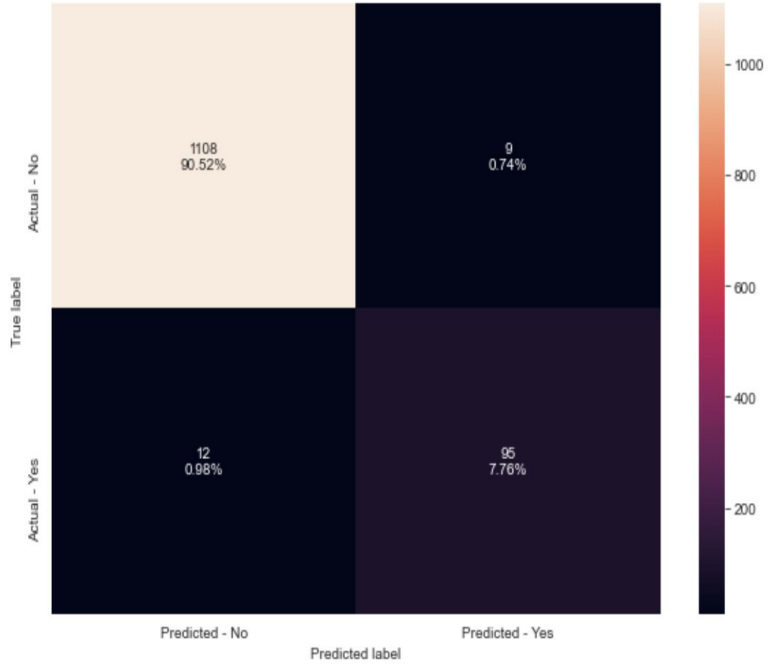
True Possitive = 76
True Negative = 1323
False Possive = 18
False Negative = 52

- With the recall value of 40%, which means our model did much better in predicting True Positives.
- The area under the curve is around 954%, much higher than previously
- It is better to analyze other models with only scaled data.



Decision Trees





Accuracy on train set 1.0
Accuracy on test set 0.9795751633986928

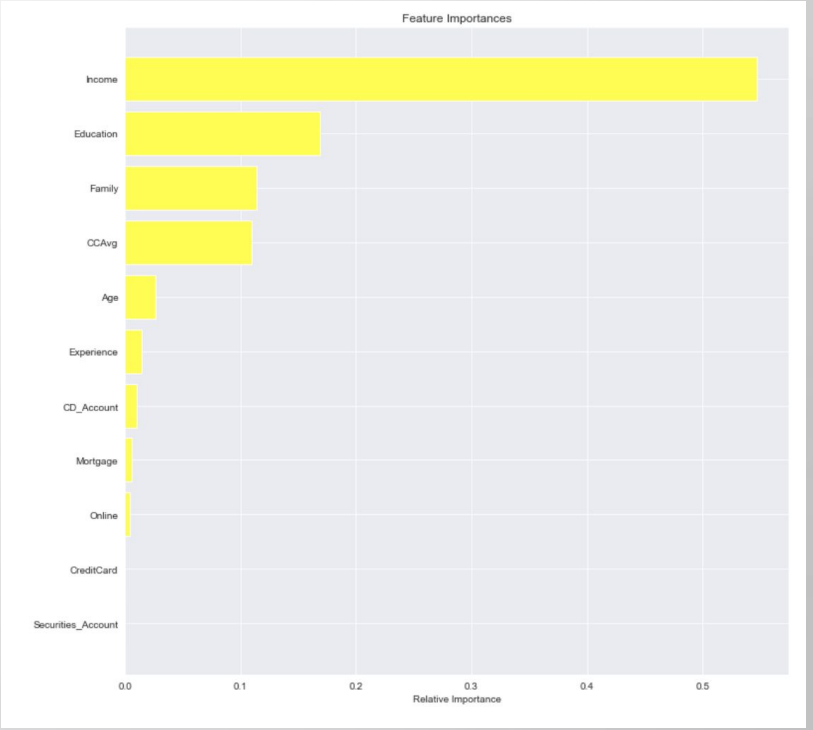
	precision	recall	f1-score	support
0	0.99	0.98	0.99	1117
1	0.84	0.92	0.88	107
micro avg	0.98	0.98	0.98	1224
macro avg	0.91	0.95	0.93	1224
weighted avg	0.98	0.98	0.98	1224

0.9771241830065359

```
[[1098   19]
 [   9   98]]
```

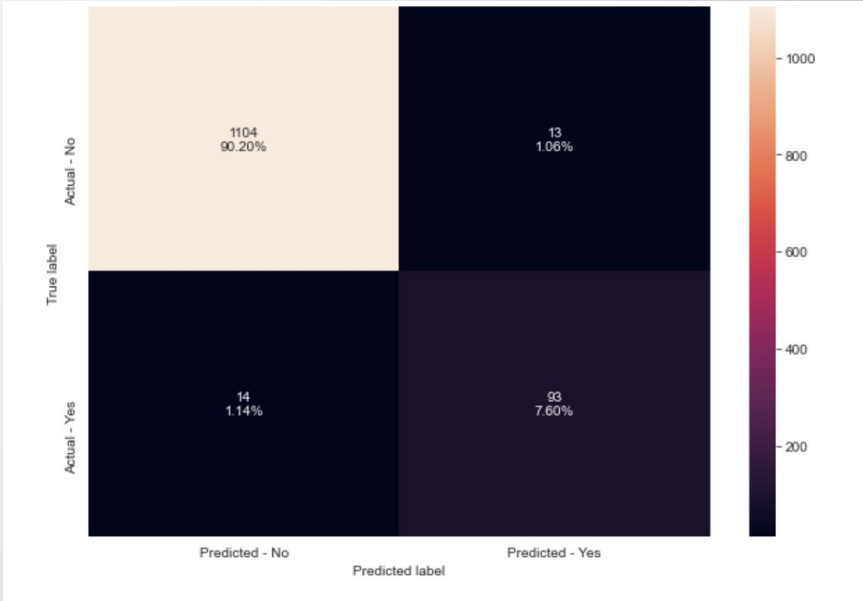
Observations

- We used decision tree with criterion as entropy can nail it down to higher recall value or with criterion as gini.
- We got a 98% accuracy score while 89% recall value
- Decision tree seems like a really good fit model for this dataframe
- We got a 98% accuracy score while an astonishing 92% recall value but the area under the ROC curve is much smaller in this case
- From the above observation it gives us the reason being the ‘overfitting’ of the data. Which is why we checked the accuracy of the test and training data
- This means we have to prune the branches to overcome overfitting
- Income, Education, Family and CCAvg are the top important features.



Pruning

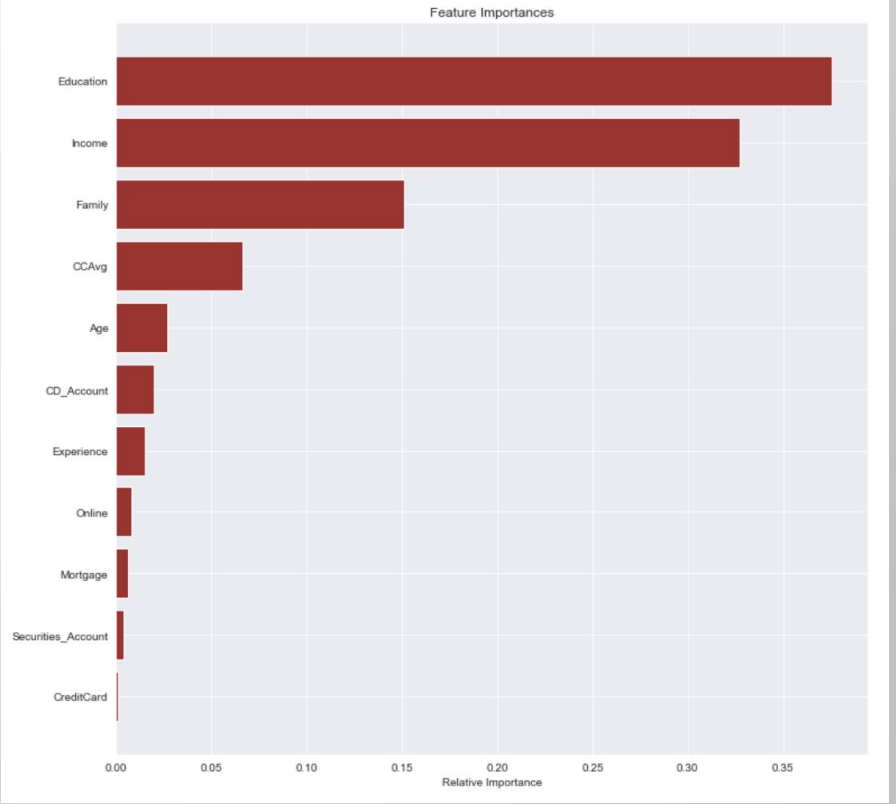
- The visualisation reveals that the model has performed better because has improved the performance in my opinion.
- The hyperparameter turning actually helped our model because it displays the leaf nodes that need to be pruned off
-



Post Pruning

Income, Education, Family and CCAvg are still the top important features. However, there has been a shift in which one has more relative importance and initially Income used to be the most important feature but after the pruning exercise, it reveals that Education has more importance relatively.

Therefore, Monthly Income and Education is the most significant factor that decides personal loan



- Decision tree with post-pruning is giving the highest recall on the test set, even though we got recall as 1 with hyperparameter tuning but that model wasn't a generalized one
- It reveals that 94% recall value with a 95% accuracy level, also AUC is approximately 97% which is fairly good.

	Model	Train_Recall	Test_Recall
0	Initial decision tree model	0.90	0.89
1	Decision treee with hyperparameter tuning	1.00	1.00
2	Decision tree with post-pruning	0.93	0.92

```
              precision    recall  f1-score   support

           0       0.99      0.95      0.97      1117
           1       0.64      0.94      0.77       107

   micro avg       0.95      0.95      0.95      1224
   macro avg       0.82      0.95      0.87      1224
weighted avg       0.96      0.95      0.95      1224

0.9493464052287581
[[1061   56]
 [   6  101]]
```

```
Area under the ROC curve :  0.9679799864456696
```

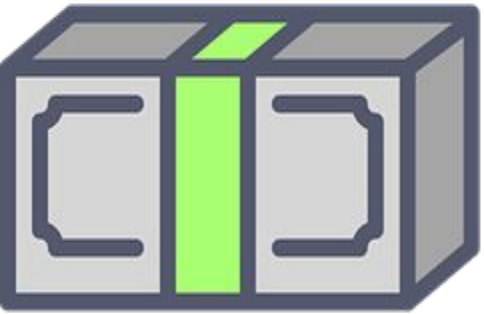
Comparing The Two Models

```
Logistic Regression Confusion Matrix \:
 [[1325   16]
 [  60   68]]

 Decision Tree Confusion matrix :
 [[1093   24]
 [  48   59]]
```

	Training Accuracy	Testing Accuracy	F1Score
Logistic Regression	0.951143	0.954000	0.693333
DecisionTrees	0.899429	0.914667	0.000000

Final Observations



Recommendations For determining if Personal Loan

- Customers with high credit card average should be a potential customers to get a loan
- When a pearsons educational level is high, has better chances of receiving a loan
- High income is a key factor for determining eligibility of receiving loans
- Family size should also be a determining factor for customers

Final Observation from Dataset

- There is a strong relationship between income and number of relationships.
- There is also a correlation between average credit card balance per month and bank relationships.
- The relationship between relationships and value of mortgage is flat when relationships are less than 2, but once it reaches 80K, it spikes and splits along the 55−75K range.
- All groups with any relationship are very active online, though there is a slight correlation with increase in online activity.
- Those with loans, securities accounts, and credit card accounts tend to have strong relationships with the bank, validating what the decision tree showed.
- Those with CD accounts tend to have strong overall relationships with the bank.
- Among the 2 models that we have implemented DecisionTreeClassifier give the best F1 Score and accuracy score with almost accuracy of 98% and F1-Score of 91%
- I would not really recommend the the use of Logistic Regression to test the accuracy of this model and we can also use better models such as KNN, RandomForest would can provide better insights from the dataset

Conclusion

The aim of the universal bank is to convert there liability customers into loan customers. They want to set up a new marketing campaign; hence, they need information about the connection between the variables given in the data. Two classification algorithms were used in this study. From the above graph , it seems like **Decision Tree** algorithm have the highest accuracy and we can choose that as our final model