Personal Loan Campaign Project 4

Objectives

- Explore the dataset and extract actionable insights that will enable growth in the market.
- Explore and visualize the dataset.
- To predict whether a liability customer will buy a personal loan or not.
- Identify which variables are most significant.
- Identify which segment of customers should be targeted more.

Data Information

The data contains the following information:

Variable	Description
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?
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 any other Bank (excluding All life Bank)?

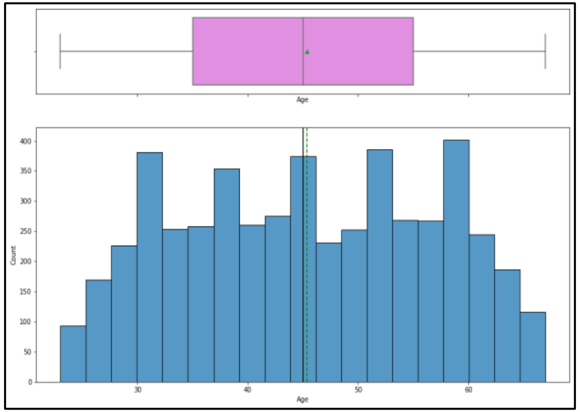
Observations	Variables
5000	14

Note:

- There are no missing values in the dataset
- All variables are numeric values
- Zip code is a float that will later be converted to City object type
- We also consolidate cities with counts less than or equal to 30 counts into a different category.
- Since all the values in ID column are unique we decide to drop the column

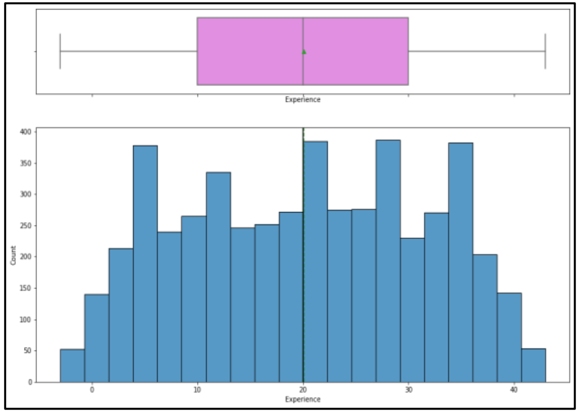
Exploratory Data Analysis – Age, Experience, Zip Code

Age



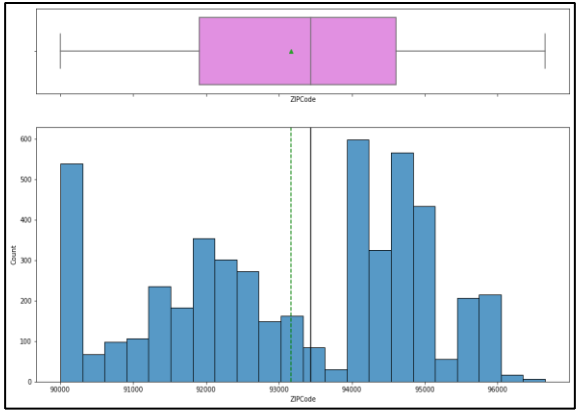
- The distribution of the age doesn't appear to be skewed
- The boxplot does not show any outliers
- The mean and the median are both very close in the distribution

Experience



- The distribution of the experience doesn't appear to be skewed
- The boxplot does not show any outliers
- The mean and the median are both very close in the distribution

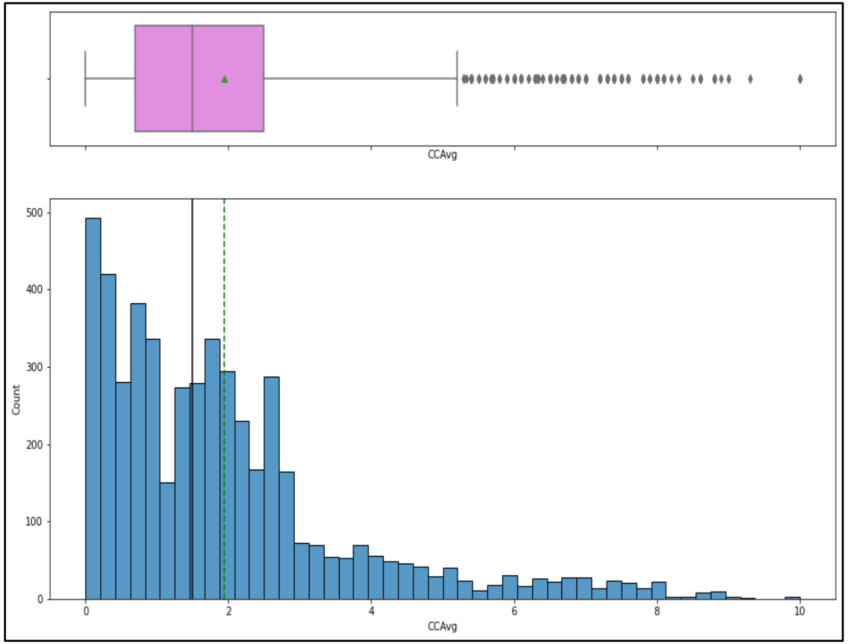
Zip Code



- Zip code does not appear to have a trend.
- There is not much that can be determined based on this trend.
- We will be converting the zip code to cities later on.

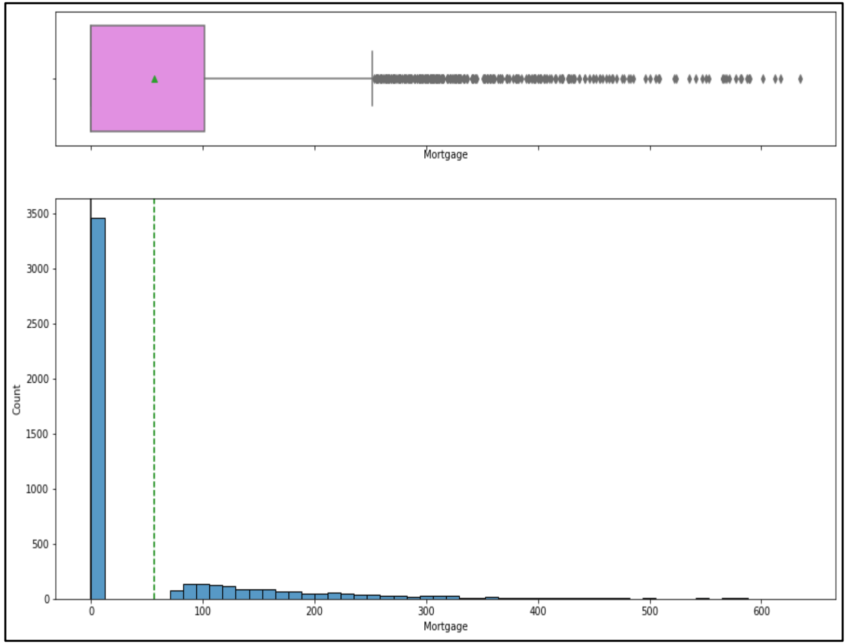
Exploratory Data Analysis – CCAvg and Mortgage

CCAvg Type



- The distribution of the credit card average is very skewed to the left
- The boxplot shows outliers on the upper end

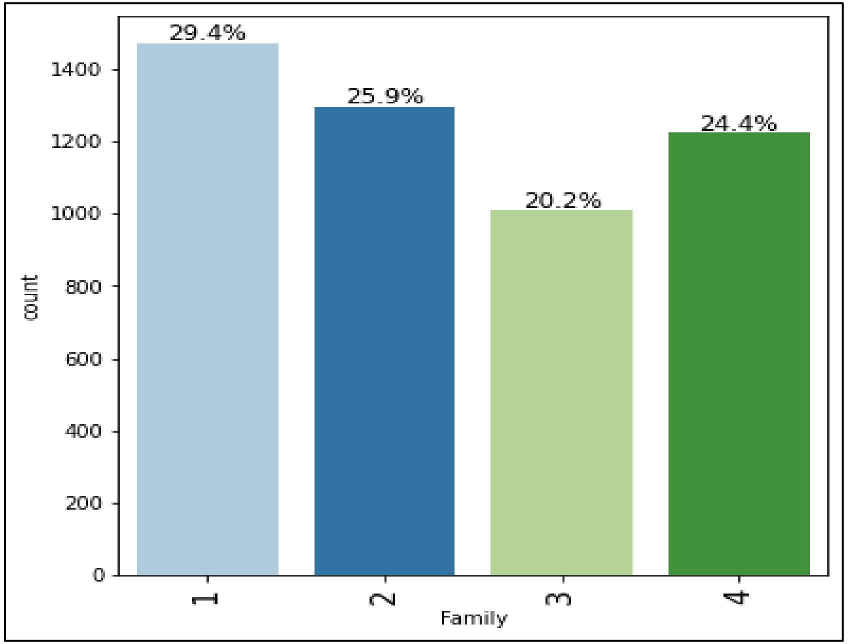
Mortgage



- The distribution of the mortgage is heavily skewed to the left
- The boxplot shows outliers on the upper end

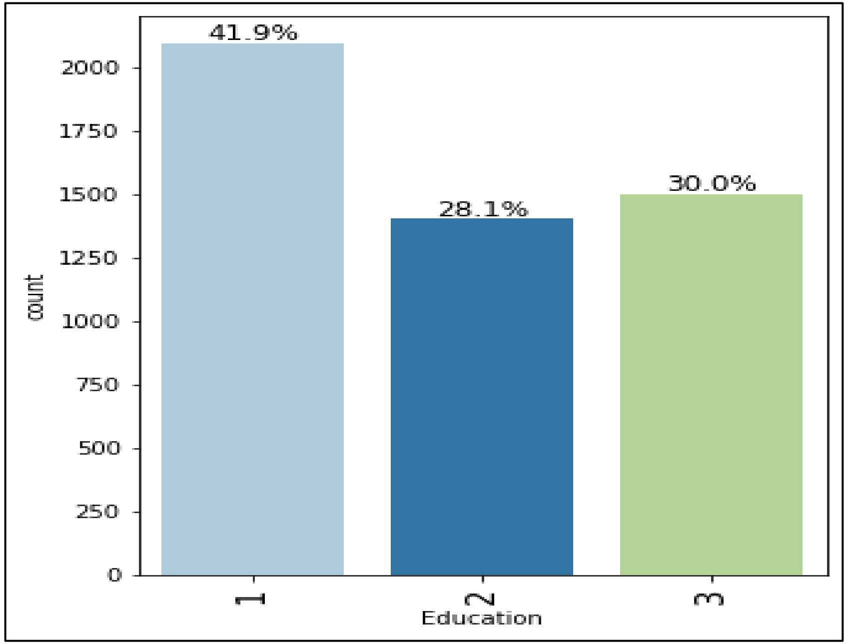
Exploratory Data Analysis – Family and Education

Family



- A majority of the customers are single (in a 1 person family)
- The next highest counts are 2 and 4 people families

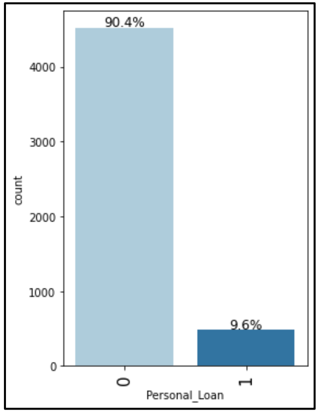
Education



- Most customers have an undergraduate degree at 41.9%
- 28.1% of customers have a graduate degree while 30.0% have advanced/professional education

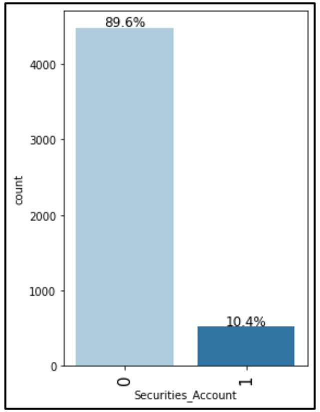
Exploratory Data Analysis – Personal Loan, Securities Account, CD Account, Online, Credit Card

Personal Loan



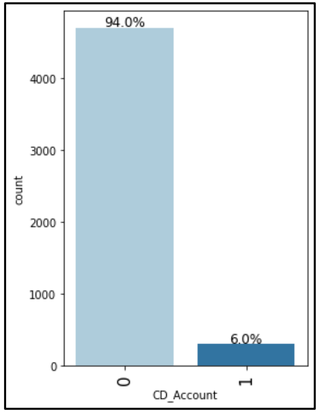
- 90.4% of customers did not the personal loan offered to them during the last campaign

Securities Account



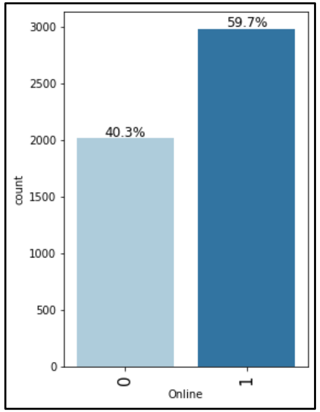
- 10.4% have securities accounts with the bank

CD Account



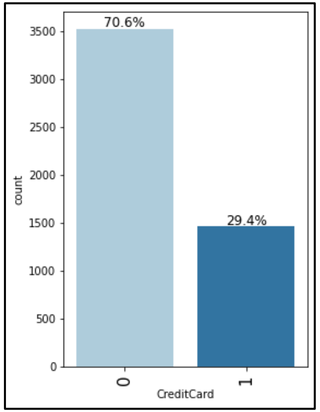
- 6% of customers have a certificate of deposit account with the bank

Online



- 59% of customers use internet banking facilities

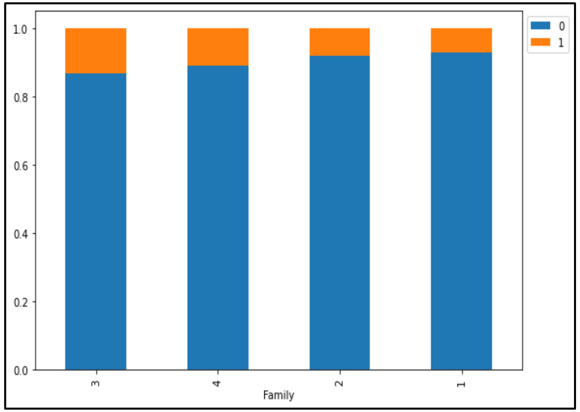
Credit Card



- 29.4% of customers use a credit card issued by another bank

Exploratory Data Analysis – Family, Education, and Securities Account vs Personal Loan

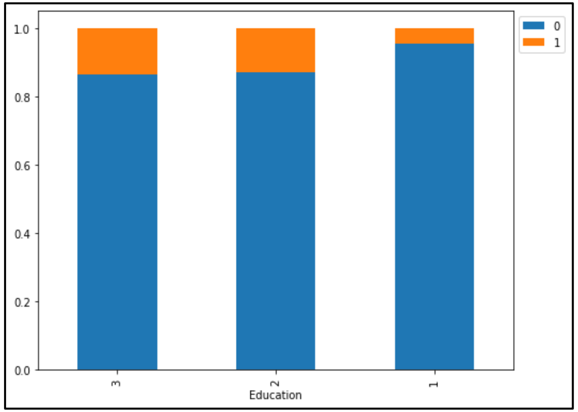
Family vs Personal Loan



Personal_Loan	0	1	All
Family			
All	4520	480	5000
4	1088	134	1222
3	877	133	1010
1	1365	107	1472
2	1190	106	1296

- The distribution of personal loans by the number of family members is almost evenly distributed

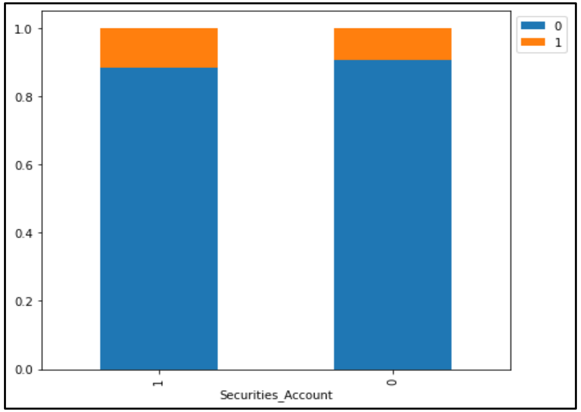
Education vs Personal Loan



Personal_Loan	0	1	All
Education			
All	4520	480	5000
3	1296	205	1501
2	1221	182	1403
1	2003	93	2096

- The distribution of the personal loans is mostly evenly distributed between graduate and advanced/ professional
- There are fewer personal loans bought by customers with only undergraduate education

Securities Account vs Personal Loan

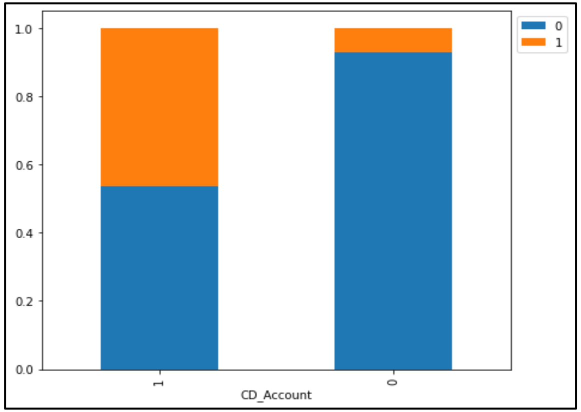


Personal_Loan	0	1	All
Securities_Account			
All	4520	480	5000
0	4058	420	4478
1	462	60	522

- There are more customers who do not have securities accounts that bought personal loans than customers without securities accounts who bought personal loans

Exploratory Data Analysis – Family, Education, and Securities Account vs Personal Loan

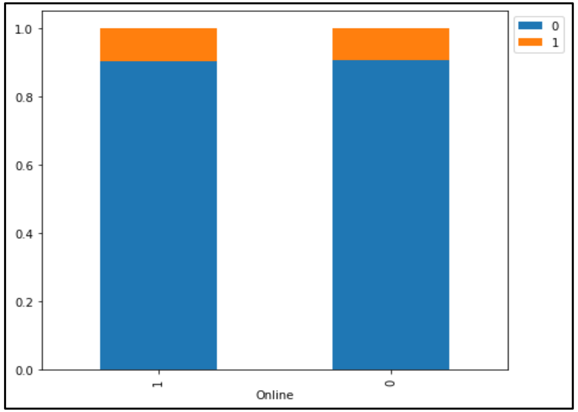
CD Account vs Personal Loan



Personal_Loan	0	1	All
CD_Account			
All	4520	480	5000
0	4358	340	4698
1	162	140	302

- There were more customers who had CD account with the bank that bought a personal loan

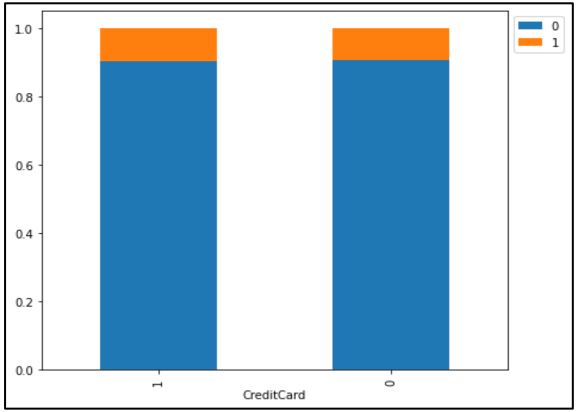
Online vs Personal Loan



Personal_Loan	0	1	All
Online			
All	4520	480	5000
1	2693	291	2984
0	1827	189	2016

- There were slightly more customers who bought personal loans that used the online services

Credit Card vs Personal Loan



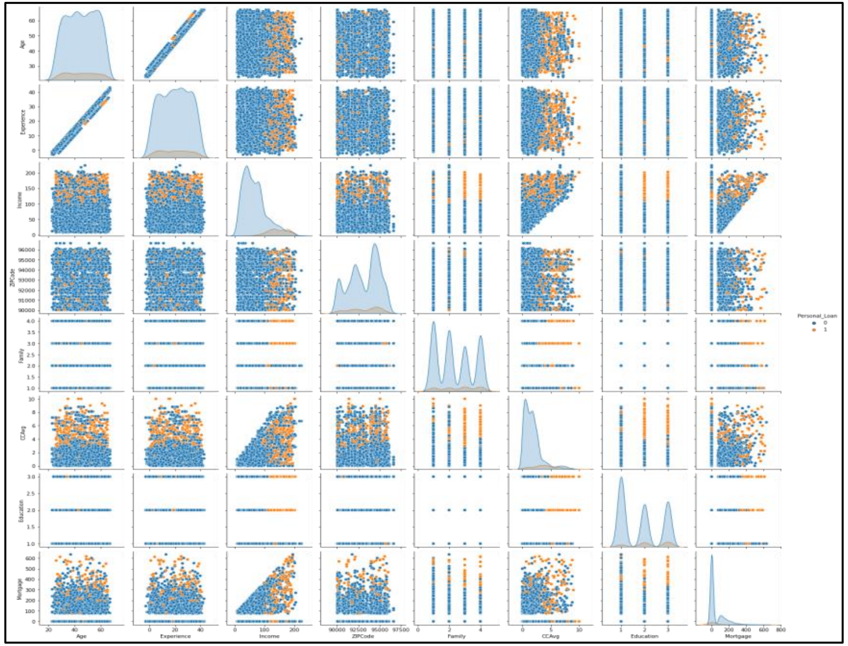
Personal_Loan	0	1	All
CreditCard			
All	4520	480	5000
0	3193	337	3530
1	1327	143	1470

- More customers who bought personal loans did not have credit cards at other banks

Exploratory Data Analysis – Correlation

Heat Map Correlation

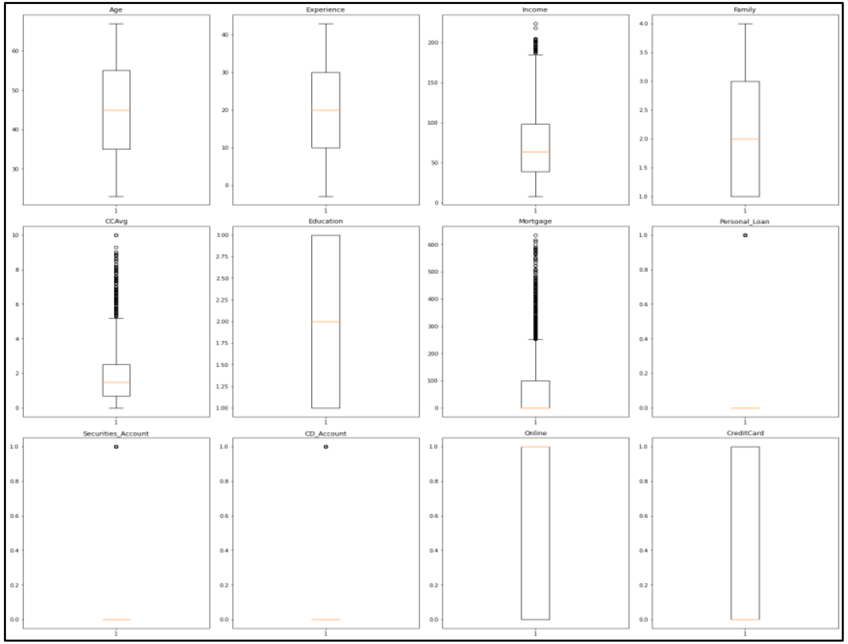
Pair Plot Correlation



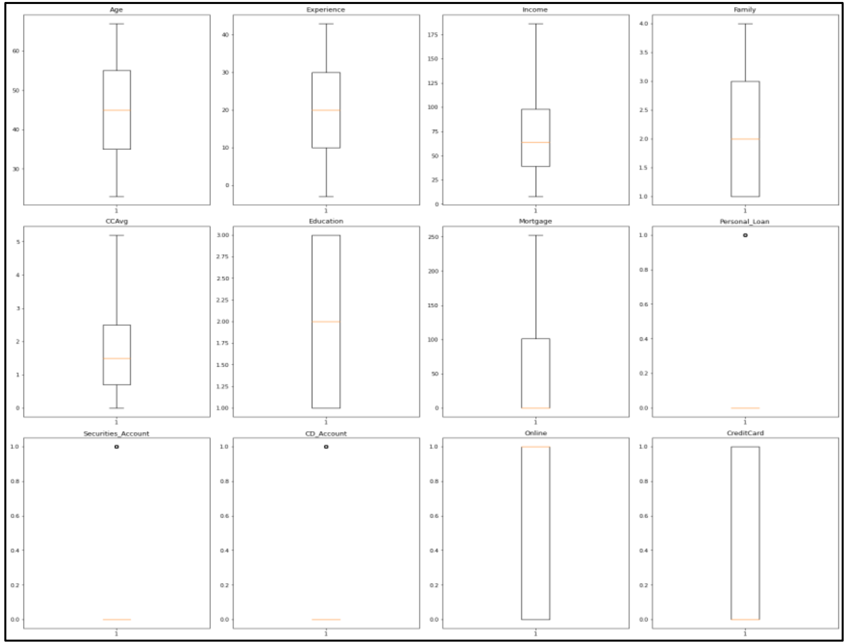
- CD account seems to have some influence if the customer is likely to purchase the personal loan.
- Most other comparisons seem equally distributed.
- There is a very strong correlation between experience and age based on the pair plot.
- There are also some correlation between CCAvg and Income.

Exploratory Data Analysis – Family and Education

Pre Treatment



Post Treatment



- Three of our variables appear to have outliers: Income, CCAvg, and Mortgage
- We will treat only these three variable for outliers.
- We will exclude the other variables as they either do not have outliers or they have values of 1 or 0.

- All numerical variables that had outliers have been treated.
- Only the target variables have been treated for outliers.

Model and Regression Outputs

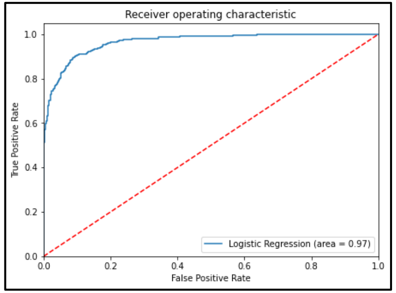
Odds from coefficients

	const	Income	Family	CCAvg	Education	Securities_Account	CD_Account	Online	CreditCard
odds	4.836004e-07	1.057445	2.152445	1.474822	6.059412	0.312396	48.021466	0.516076	0.310117

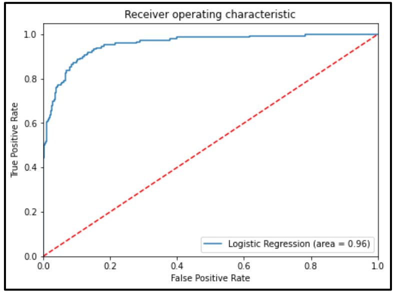
Percentage change in odds

	const	Income	Family	CCAvg	Education	Securities_Account	CD_Account	Online	CreditCard
change_odds%	-99.999952	5.744541	115.244464	47.482201	505.94125	-68.760433	4702.146644	-48.392401	-68.988328

ROC-AUC on training set



ROC-AUC on test set



Coefficient interpretations:

- ***Income***: Holding all other features constant a unit change in Income will increase the odds of a customer buying a personal loan by 1.05 times or a 5.74% increase in odds.
- ***Family***: Holding all other features constant a unit change in Family will increase the odds of a customer buying a personal loan by 2.15 times or a 115.24% increase in the odds.
- Interpretation for other attributes can be done similarly

Coefficient interpretations:

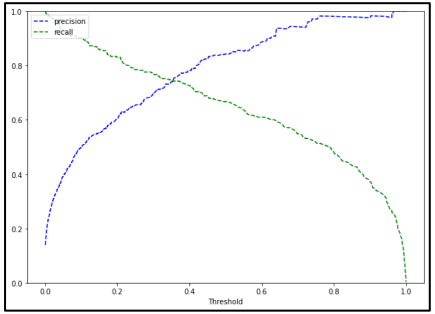
- Coefficients of Income, Family, CCAvg, Education, and CD_Account are positive; an increase in these will lead to an increase in chances of a customer buying a personal loan.
- Coefficients of Securities_Account, Online, and Credit Card are all negative; an increase in these will lead to a decrease in chances of a customer buying a personal loan.

Model and Regression Outputs

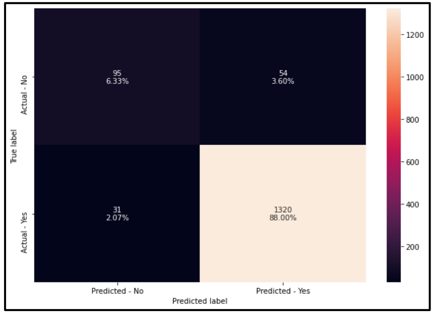
Model performance summary

	Model	Train_Accuracy	Test_Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1	Test F1
0	Logistic Regression Model - Statsmodels	0.956857	0.948667	0.667674	0.610738	0.843511	0.827273	0.745363	0.702703
1	Logistic Regression - Optimal threshold = 0 .09	0.904571	0.909333	0.900302	0.865772	0.497496	0.526531	0.640860	0.654822
2	Logistic Regression - Optimal threshold = 0 .39	0.954571	0.943333	0.731118	0.637584	0.775641	0.753968	0.752722	0.690909

Precision-Recall Curve



Final model performance

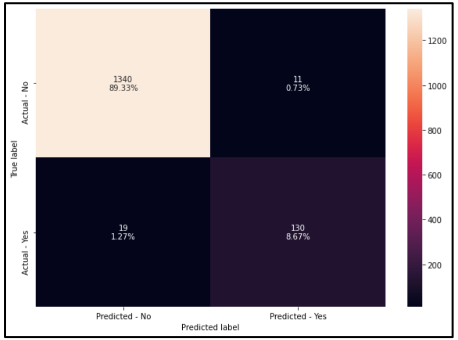


Observations:

- After initial interpretations of coefficients we decided to check if the F1 score can be improved further.
- To do so we changed the model threshold by using AUC-ROC Curve.
- We calculated the optimal cutoff which yielded us a threshold of 0.0997 which increased the recall significantly on both the test and training set.
- The best test recall is 86% but the test precision is low i.e ~52% at the same time. This means that the model is not good at identifying prospective customers, therefore the bank can lose many opportunities of campaigning personal loans to prospective customers.

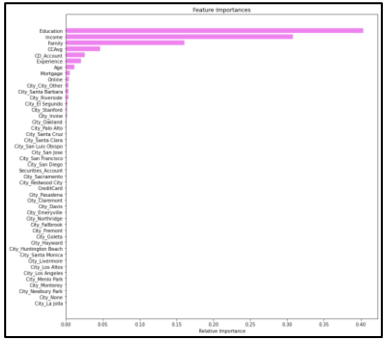
Decision Tree Outputs

Initial Model output

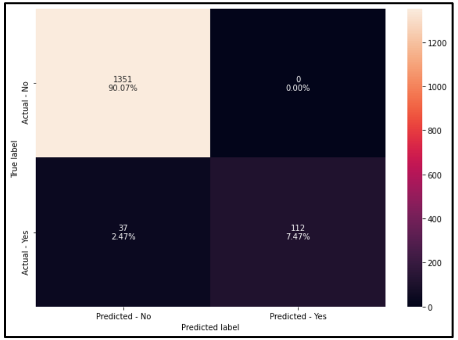



```
Recall on training set :  1.0
Recall on test set :  0.87248322147651
```

Feature Importance

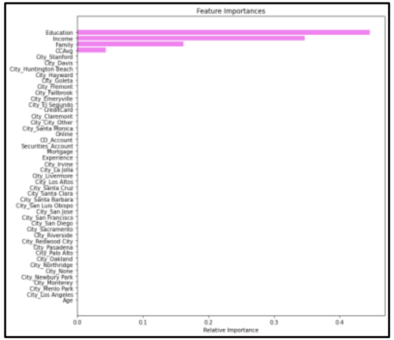


Model output with depth restricted to 3




```
Recall on training set :  0.8126888217522659
Recall on test set :  0.7516778523489933
```

Feature Importance



Observations:

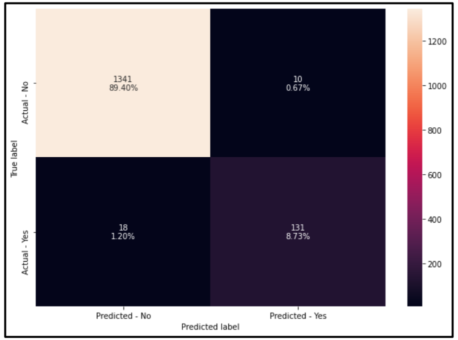
- According to the decision tree model, Education is the most important variable for predicting the customer default.

Observations:

- Recall on training set has reduced from 1 to 0.81 but this is an improvement because now the model is not overfitting and we have a generalized model.
- In important features of previous model, Education was on top.
- Here Education is still on top as the top important feature.

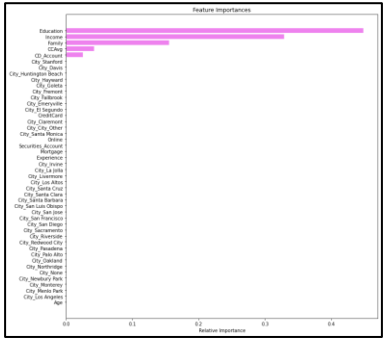
Decision Tree Outputs – Tuned Hyperparameters

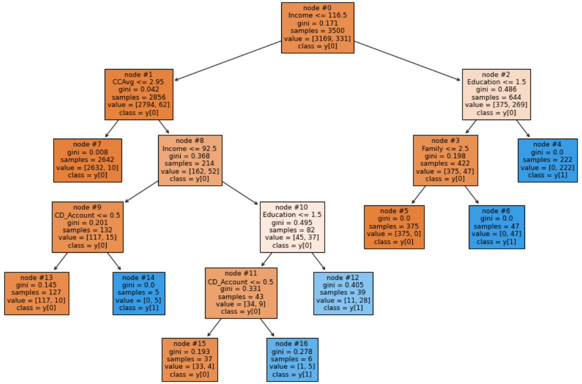
Model output with tuned hyperparameters




```
Recall on training set :  0.9274924471299094
Recall on test set :  0.8791946308724832
```

Feature Importance



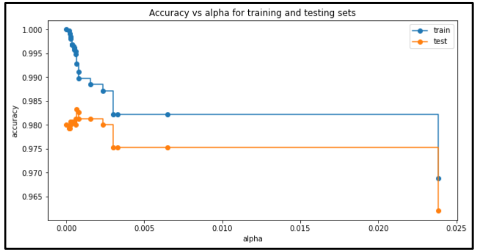


Observations:

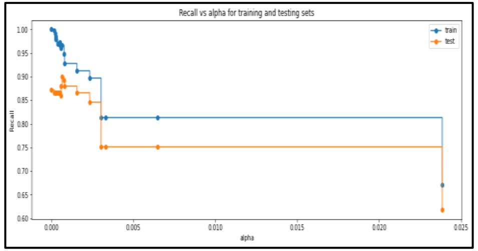
- After tuning hyperparameters, the performance of the model has become more generalized.
- Recall has increased from 0.81 to 0.92
- Feature importance is still Education for this model

Cost Complexity Pruning

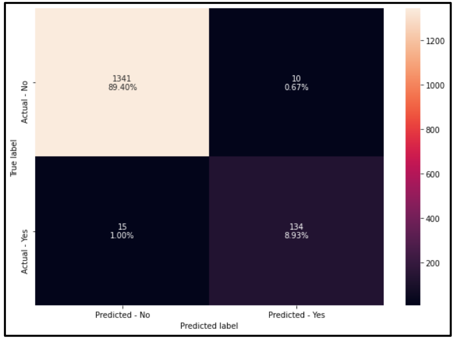
Accuracy vs Alpha



Recall vs Alpha

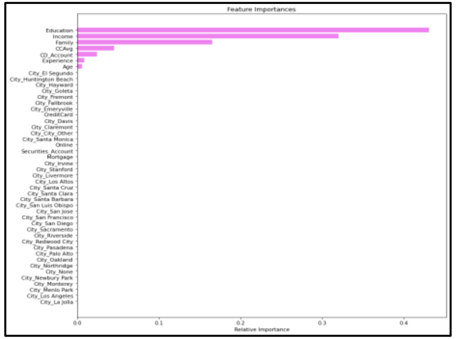


Post pruned model




```
Recall on training set :  0.9667673716012085
Recall on test set :  0.8993288590604027
```

Feature Importance



Observations:

- With post-pruning we get the high recall on both training and test set
- The initial decision tree model gives the highest test recall
- We did not see much improvement in test recall as a result of our pruning methods

	Model	Train_Recall	Test_Recall
0	Initial decision tree model	1.00	0.98
1	Decision tree with restricted maximum depth	1.00	0.87
2	Decision treee with hyperparameter tuning	0.81	0.75
3	Decision tree with post-pruning	0.96	0.89

Conclusion

After all the analysis, we have been able to conclude:

- Having CD accounts had some influence on if the customer bought a personal loan.
- Zip code was not going to be very useful for our analysis as is so we converted to cities and limited the cities with counts greater than 30
- Income, CC Average, and Mortgage had outliers that were treated.
- The model evaluation criterion was based on the following:
    - Predicting a liability customer is not going to buy a personal loan but they do - Loss of opportunity
    - Predicting a liability customer is going to buy a personal loan but they don't - Loss of resources
- Loss of opportunity would be the greater loss
- The bank would want to reduce false negatives, this can be done by maximizing the Recall.
    - The greater the recall lesser the chances of false negatives.
- Age and Experience have high VIF but the rest of the variables in the summary appear to be reliable.
- All the categorical levels of Age, Experience, Mortgage, and City have a high p-value. Hence, the variable can be dropped.
- Holding all other features constant a unit change in Income will increase the odds of a customer buying a personal loan by 1.05 times or a 5.74% increase in odds.
- Holding all other features constant a unit change in Family will increase the odds of a customer buying a personal loan by 2.15 times or a 115.24% increase in the odds.
- Based on our coefficient interpretation, having securities accounts, using online feature, and having credit cards at other banks decrease the odds of customers buying a personal loan.
- The best test recall is 86% but the test precision is low i.e ~52% at the same time. This means that the model is not good at identifying prospective customers, therefore the bank can lose many opportunities of campaigning personal loans to prospective customers.
- According to the decision tree model, Education is the most important variable for predicting the customer default.

Recommendations

After all the analysis, we suggest the following recommendations:

- We saw our analysis that customers who use the online banking feature are less likely to purchase a personal loan. The bank can improve its online presence or perhaps campaign via other means.
- We saw that customers who have more credit cards less likely to purchase a personal loan while customers with more monthly credit card payments are more likely to purchase a personal loan. The bank should focus more on customers with fewer credit cards and that have higher monthly payments.
- Our analysis showed that families with more members are more likely to purchase a personal loan. The bank can focus more on customers with larger families.
- Our analysis showed that customers with security accounts are less likely to purchase a personal loan. This implies that the bank has good security for its customers. The bank should focus its campaigns to customers who are not their customers.

greatlearning
Power Ahead

Happy Learning !