Personal Loan Campaign

Project 2: Machine Learning (AI/ML course) by Sarah Choi

Date: 9/15/23

Contents / Agenda

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

Executive Summary - Conclusions

	Conclusion
Decision Tree Model	• Analyzed the dataset from AllLife Bank for their liability customers by using CART algorithm and Decision Tree Classifier to build a predictive model. • The model built can be used to predict if an existing liability customer is going to purchase personal loan with the bank. • Visualized different trees and their confusion matrix to get a better understanding of the model. Easy interpretation is one of the key benefits of Decision Trees.
Model performance	• The performance was increased by tuning the hyper-parameters. • Cross-validation of the tree is performed by splitting the data into different ratio the training and test sets. Afterward, we iterate over the folds. we use all the folds but to train a tree for each combination in the grid, validating the fitted tree on the reserved fold . That way, we get trees and validation scores for each grid combination.
Feature Importance	• Income, Size of Family, Education, are the most important variable in predicting the customers who will purchase personal loan.
Outliers	• There are outliers in many variables (Income, Mortgage, Credit Card spending), some of which has nearly 8% of the data. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results. This type of outlier can be a problem in regression analysis
Customer Segment	• From the data analysis the highest number of customers who have the personal loan in the current dataset fell into the customer population that have income between 125K to 175K. • Most customers who have relationship with the bank (~29.4%) are singles in their household, and 25.9% has a family size of 2, 24.4% has a family size of 4 and 20.2% has a family size of 3. • Majority of the customers (42%) in the current dataset have undergraduate degree, 30% have professional degree and the rest of 28% have graduate degree.

Executive Summary - Recommendations

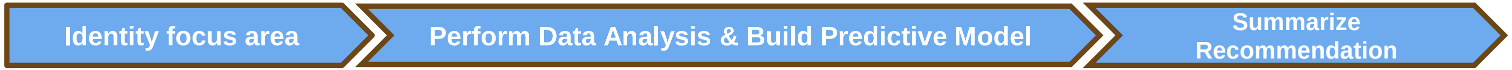
	Recommendations
Personal Loan Campaign Strategy	According to the decision tree model predication: a) If a customer has income less than $98.5K there's a very high chance the customer will not buy a personal loan offer. b) If a customer has an income greater than $98.5K with a family size less than 3 and has an undergraduate education, then there is a very high potential to buy a personal loan. These criteria is the most significant in driving personal loan purchases, the bank should use them to drive the marketing campaign.
Most significant customer attributes	According to the decision tree model: • Income is the most important feature, and since Income has a high positive correlation with customers average credit card spending (0.65), this indicates the higher the customers income is, the more they spent on their credit card. The bank should target the customers who have high average credit card spending to be the population for the personal loan marketing campaign. • Income also shows a positive correlation with value of house Mortgage, The bank should identify the customers who have mortgage accounts with them and target those customers who have high value of house Mortgage to be the population for the personal loan marketing campaign. • Education is another important attribute for predicating potential sales of personal loan, and particularly the customer who has Undergraduate education who have working experience of 20 years. The bank should conduct consider this as one of their customer segment for the marketing campaign.
Outliers	• It is observed that there are large amount of outliers in many variables (Income, Mortgage, Credit Card spending). In any case where a large amount of the data is identified as "outliers", it is likely either that the outlier test has been incorrectly applied, or the outlier test is based on a distributional assumption that assumes much thinner tails than the data and is therefore falsified by the data. There may have been an error in data transmission or transcription. Alternatively, an outlier could be the result of a flaw in the assumed theory, calling for further investigation by the researcher. • The proper action depends on what causes the outliers. In broad strokes, there are three causes for outliers—data entry or measurement errors, sampling problems and unusual conditions, and natural variation. • Recommended approach: analyze how the initial machine learning or data science problem was framed, how the target population and sample were chosen, and how so many outliers made into the dataset. • The stakeholders (people who directly benefit or lose from the project) should be informed of the decision. Preferably, you should present two results: one with outliers present and one without.

Business Problem Overview and Solution Approach

Problem Statement

AllLife 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. The objective of this analysis is to build a model to help AllLife bank marketing department to predict whether a liability customer will buy personal loans, and to understand which customer attributes are most significant in driving purchases, and to identify which segment of customers to target more.

Solution Approach



To extract actionable insights from the data that AllLife Bank has collected to identify key areas of focus.

- Perform EDA on Variables that influence the personal loan sales
- Build a Decision Tree Model using CART algorithms by training the model and to maximize the performance so that we can predict whether a liability customer will buy personal loans

Summarize the business recommendations for the bank's marketing department

Data Overview

The data contains the different data related to a customer at AllLife Bank. The detailed data dictionary is given below.

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)
ZIPCode	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)

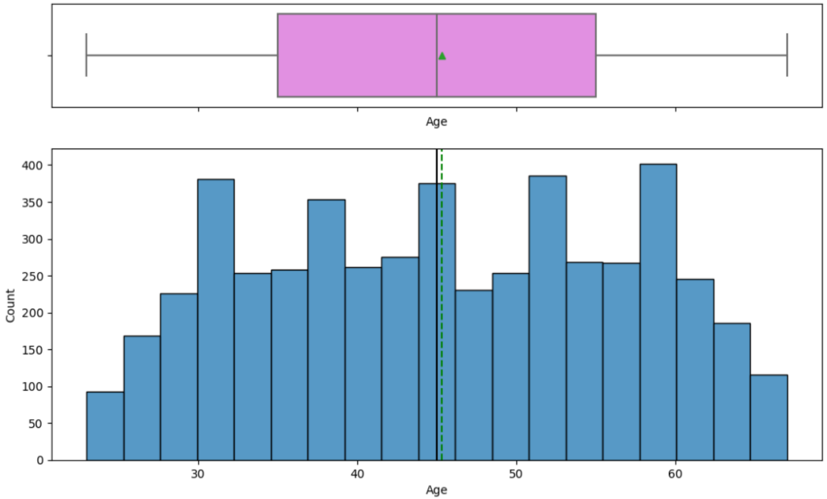
Observations	Variables
5000	14

Note:

- There is a total number of 5,000 unique customer records in the data frame with 14 columns for the data analysis
- There is no missing value in the data
- During data pre-processing phase, we assumed that for some those entries in the Experience column that have negative signs are data input errors, we replaced them with positive signs

EDA - Univariate Data Analysis

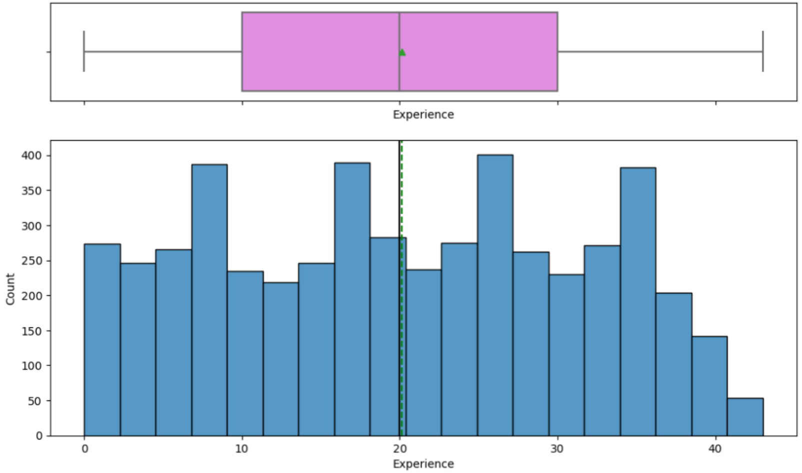
Customer’s Age



Observations:

- Age has symmetrical and normal distribution, the number of customers who have accounts above the median age of 45 are slightly higher than those customers who are aged below 45

Customer’s Year of Professional Experience

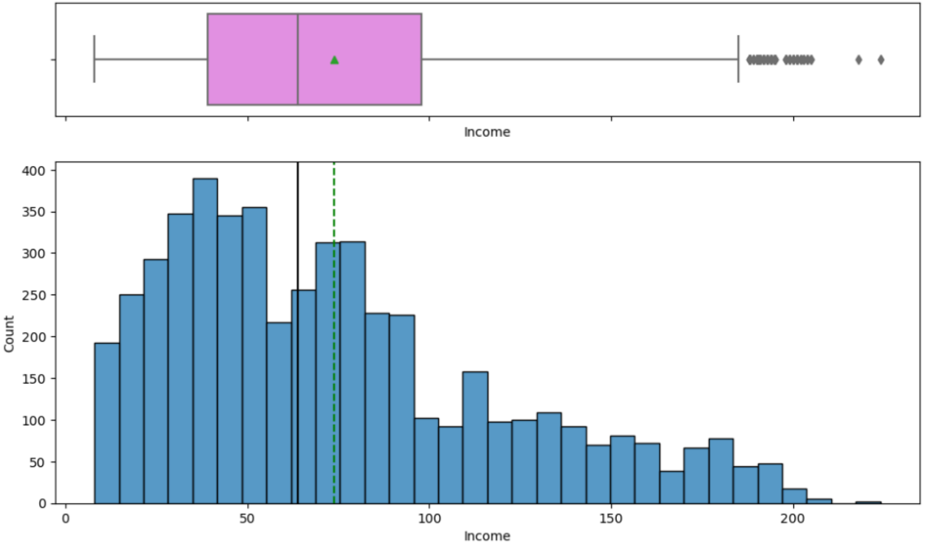


Observations:

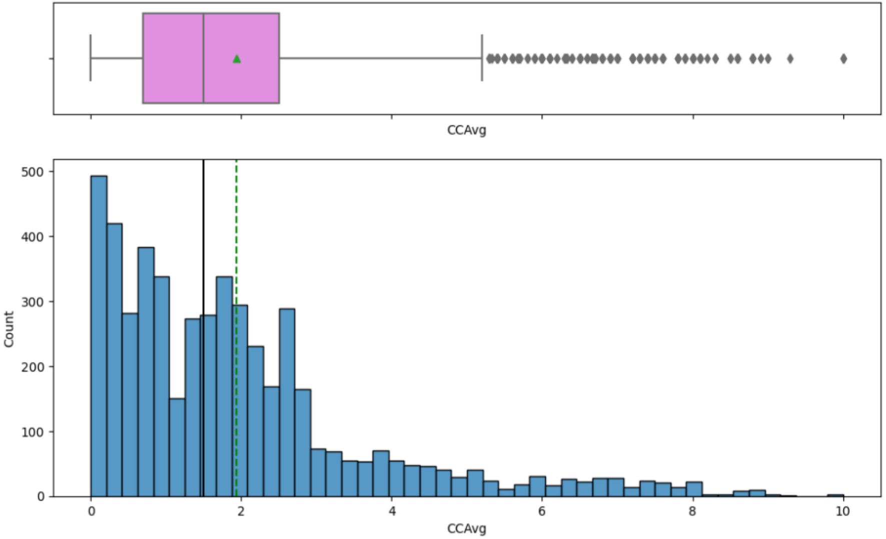
- Average customers year of professional working experience is 20
- Work experience has symmetrical and normal distribution

EDA - Univariate Data Analysis

Customer’s Annual Income (in Thousand Dollars)



Customer’s Average Credit Spending per month



Observations:

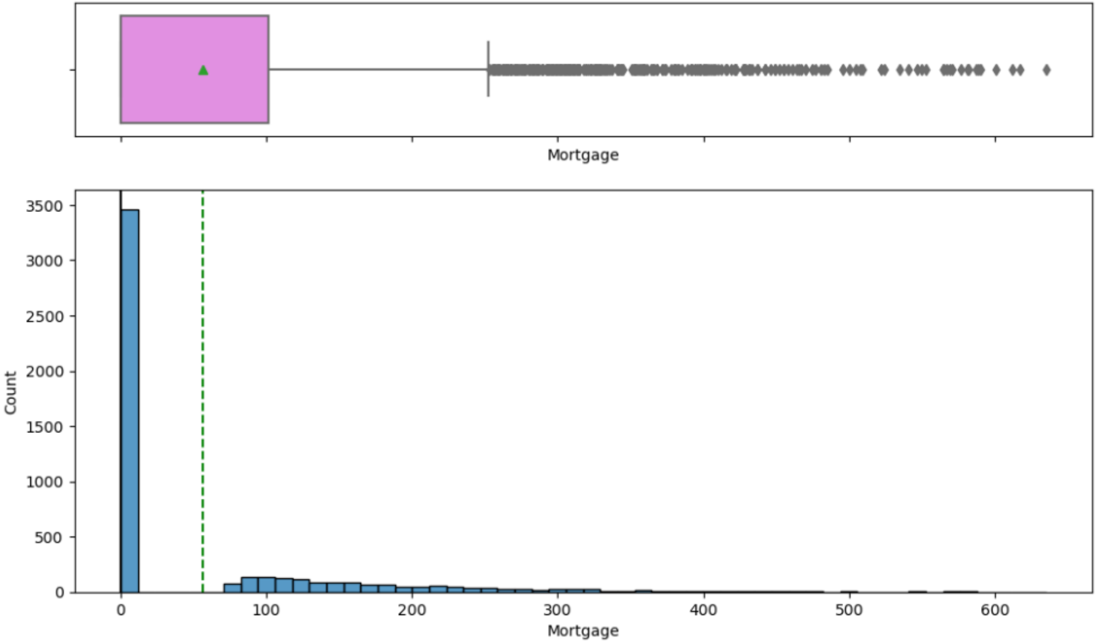
- The distribution of the Income is skewed towards the right.
- There are many outliers in this variable, the income values above 175K are being represented as outliers by the boxplot.
- There are 195 customer records that have more than 175K income values.

Observations:

- The distribution of the Average spending on Credit Card is skewed towards the right.
- There are many outliers in this variable and the values above $5,000 are being represented as outliers by the boxplot.

EDA - Univariate Data Analysis

Customer's Mortgage (if Any)

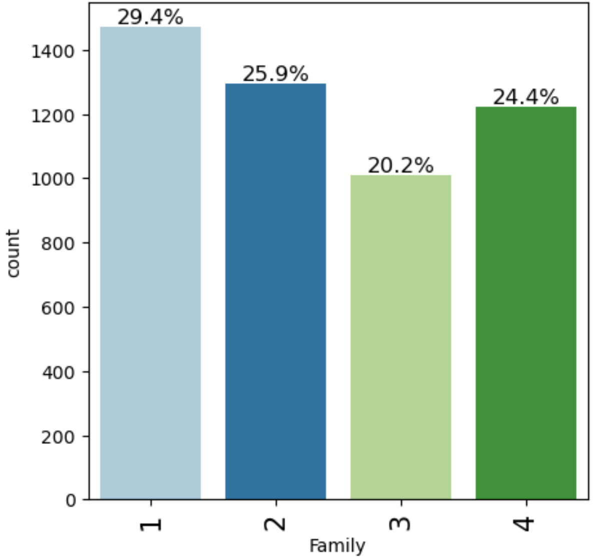


Observations:

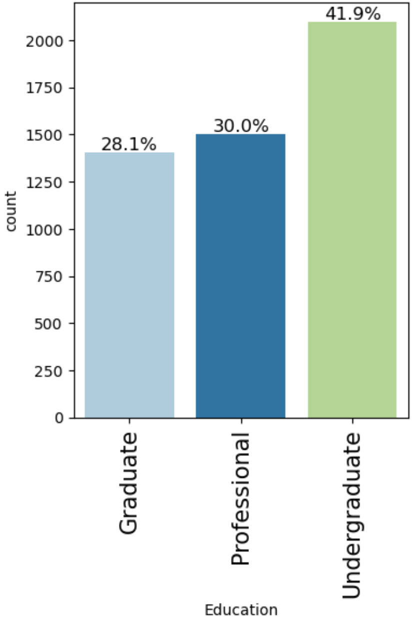
- The distribution of the value of Mortgage is skewed towards the right due to majority of the customer 3,462 (70% of total) do not have mortgage.
- There are many outliers in this variable and the values above $203K are being represented as outliers by the boxplot.

EDA - Univariate Data Analysis

Family Size of the Customer



Customer’s Education Level



Observations:

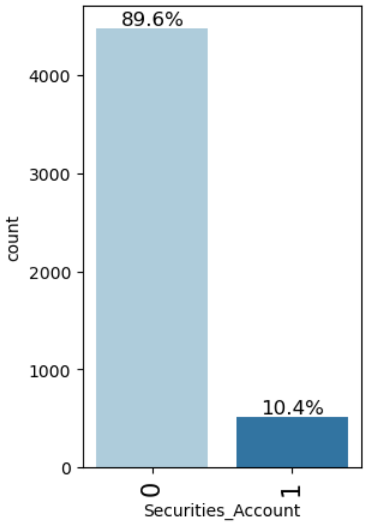
- Most customers who have relationship with the bank (~29.4%) are singles in their household.
- 25.9% has a family size of 2, 24.4% has a family size of 4 and 20.2% has a family size of 3.

Observations:

- majority of the customers (42%) in the data frame have undergraduate degree, 30% have professional degree and the rest of 28% have graduate degree.

EDA - Univariate Data Analysis

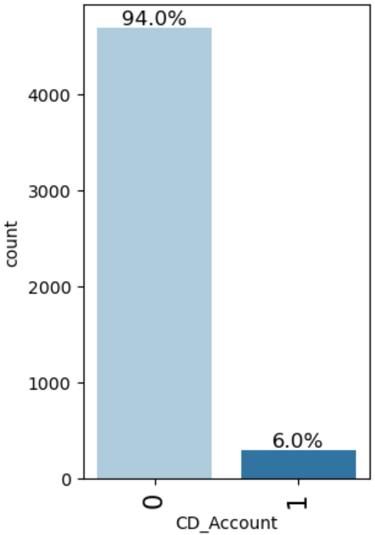
Securities Accounts



Observations:

- 89.6% of the customers do not have securities accounts with the bank

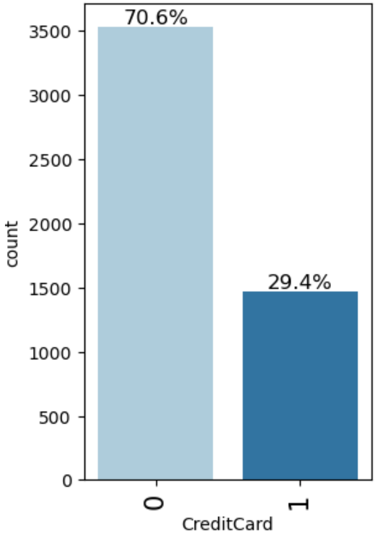
CD Accounts



Observations:

- 94% of the customers do not have CD accounts with the bank

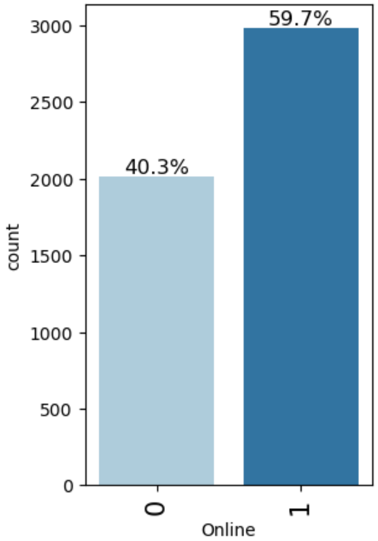
Credit Card Accounts



Observations:

- 70.6% of the customers do not use a credit card issued by any other Bank (excluding AllLife Bank)

Internet Banking

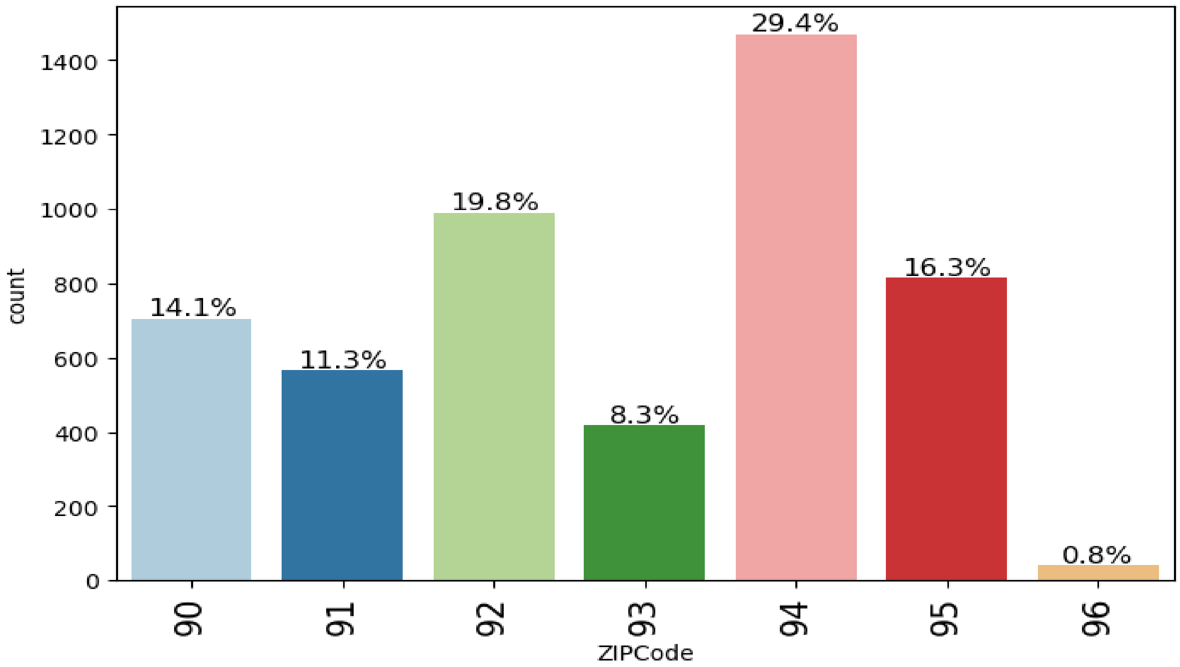


Observations:

- 59.7% of the customers use internet banking facilities

EDA - Univariate Data Analysis

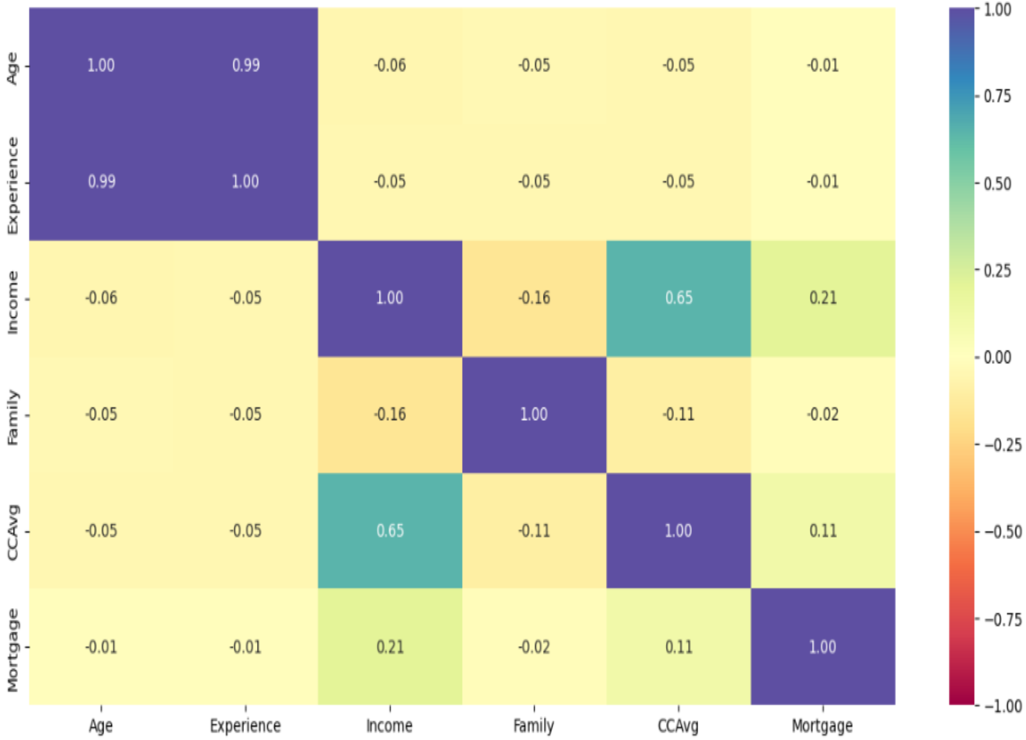
Customer's Home Address ZIP Code



Observations:

- Top 3 zip code (first 2 digits) from Customers home address : 94 (29.4% in Northern California), 92 (19.8% in Southern California), 95 (16.3% in Northern California)

Bivariate Analysis - Correlation Matrix



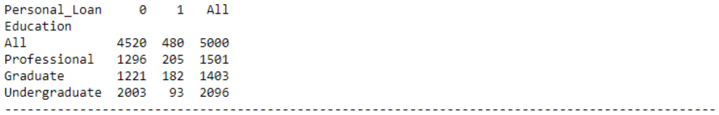
	Age	Experience	Income	Family	CCAvg	Mortgage
Age	1.00	0.99	-0.06	-0.05	-0.05	-0.01
Experience	0.99	1.00	-0.05	-0.05	-0.05	-0.01
Income	-0.06	-0.05	1.00	-0.16	0.65	0.21
Family	-0.05	-0.05	-0.16	1.00	-0.11	-0.02
CCAvg	-0.05	-0.05	0.65	-0.11	1.00	0.11
Mortgage	-0.01	-0.01	0.21	-0.02	0.11	1.00

Observations:

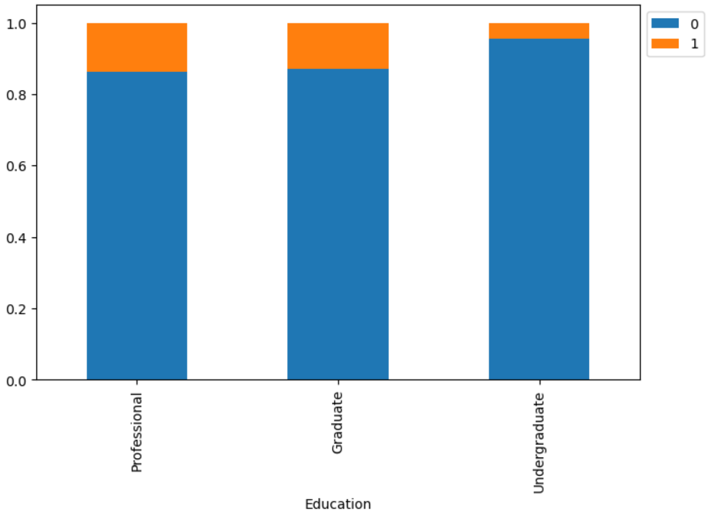
- Customer's age has the strongest positive correlation with # of years of professional experience (0.99), this indicates as age goes up, the number of years of experience goes up.
- Income has a high positive correlation with customers average credit card spending (0.65), this indicates the higher the customers income is, the more they spent on their credit card.
- Income also shows a positive correlation with Value of house Mortgage which indicates the higher the customer’s income is, the more expensive property they can afford to purchase.
- There is no significant correlation between Age and Mortgage, credit card spending, income and family size.
- There is no significant correlation between professional experience and Mortgage, credit card spending, income and family size.

EDA - Bivariate Analysis

Personal Loan vs Education

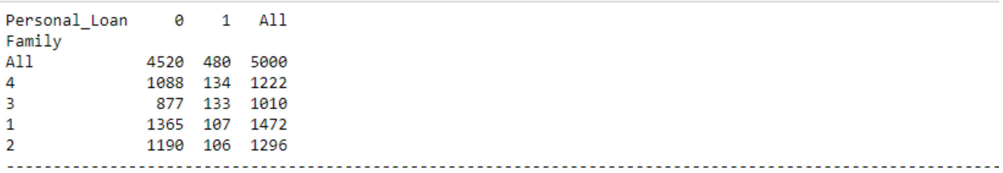


```
Personal_Loan     0    1   All
Education
All            4520  480  5000
Professional   1296  205  1501
Graduate       1221  182  1403
Undergraduate  2003   93  2096
------------------------------------------------------------------------------------------------------------------------
```

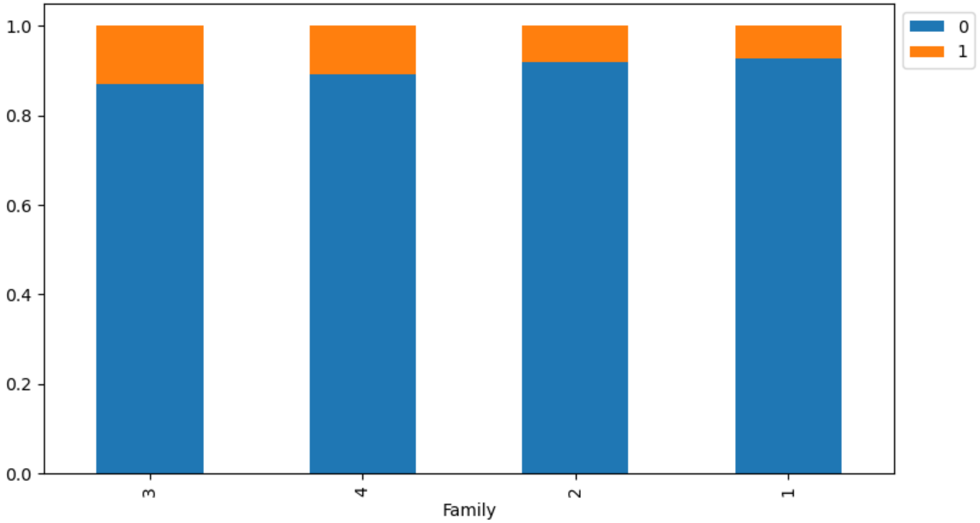


Observations: The number of customers who have professional degree accepted the most total number of personal loan offered as compared to the graduate and undergraduate

Personal Loan vs Family



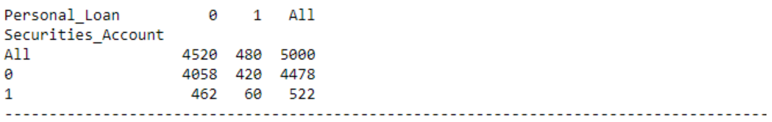
```
Personal_Loan     0    1   All
Family
All            4520  480  5000
4              1088  134  1222
3               877  133  1010
1              1365  107  1472
2              1190  106  1296
------------------------------------------------------------------------------------------------------------------------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



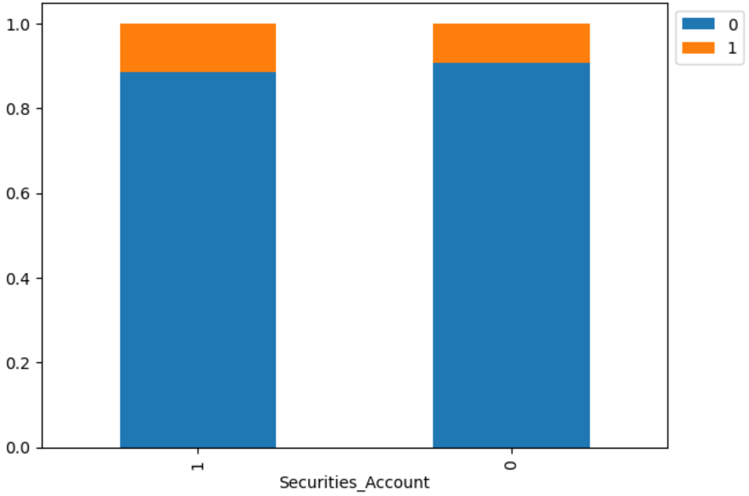
Observations: The number of customers who accepted personal loan offered are higher in the family size of 3 and 4 as compared to the customers in size of 1 and 2. And in each group of the two groups (3,4) or (1,2) there is only one customer count difference.

EDA - Bivariate Analysis

Personal Loan vs Securities Account



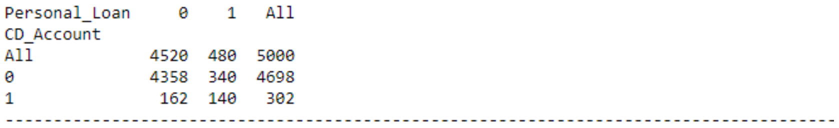
```
Personal_Loan          0    1   All
Securities_Account
All                 4520  480  5000
0                   4058  420  4478
1                    462   60   522
------------------------------------------------------------------------------------------
```



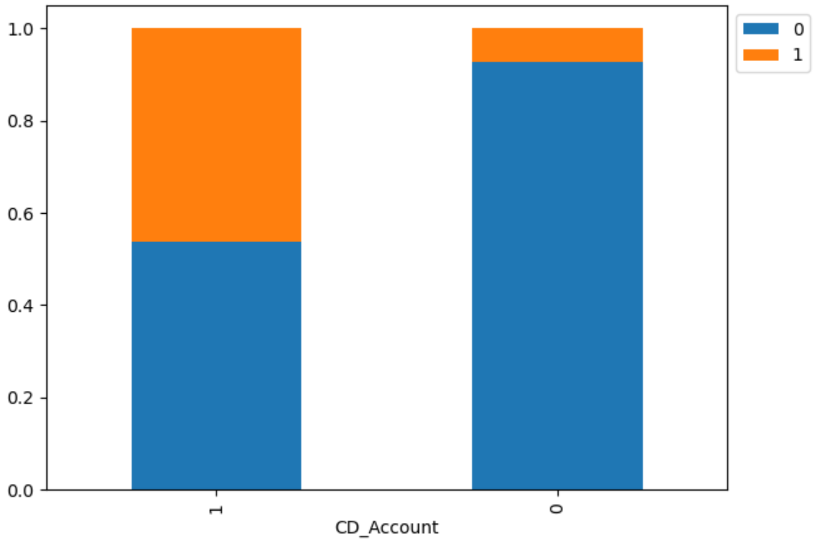
Observations:

- 60 out of the 522 customers who have securities account with the bank accepted the personal loan offered in the last campaign
- 420 out of the 4478 customers who do not have securities account with the bank accepted the personal loan offered in the last campaign

Personal Loan vs CD Account



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

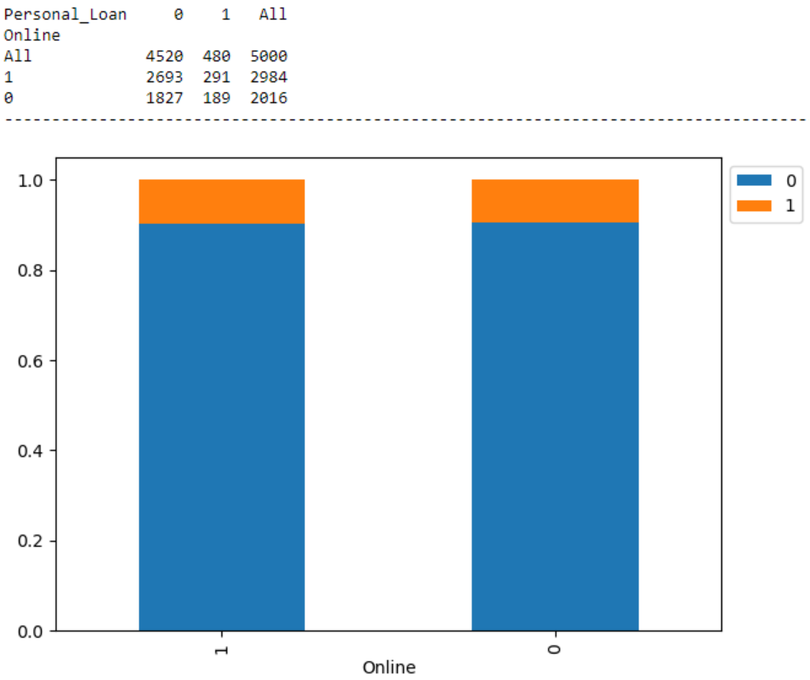


Observations:

- 140 out of 302 customers who have CD account with the bank accepted the personal loan offered in the last campaign
- 340 out of 4698 customers who do not have CD account with the bank accepted the personal loan offered in the last campaign

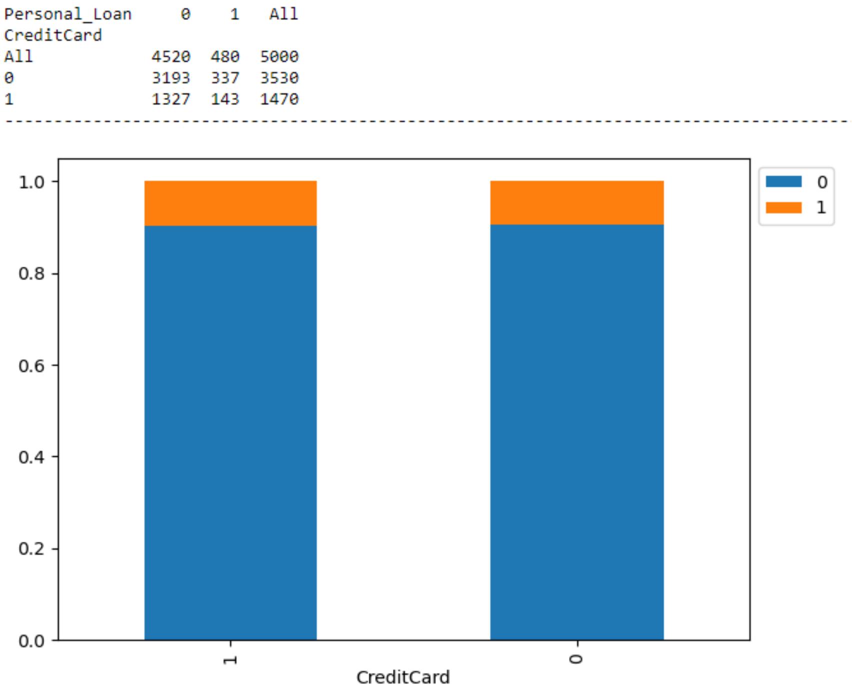
EDA - Bivariate Analysis

Personal Loan vs Online Banking



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

Personal Loan vs Credit Card Account



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

Observations:

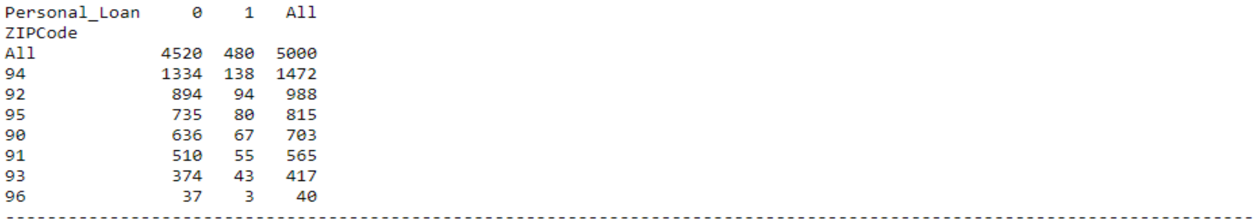
- 291 out of 2984 customers who use online banking accepted the personal loan offered in the last campaign
- 189 out of 1827 customers who do not use online banking accepted the personal loan offered in the last campaign

Observations:

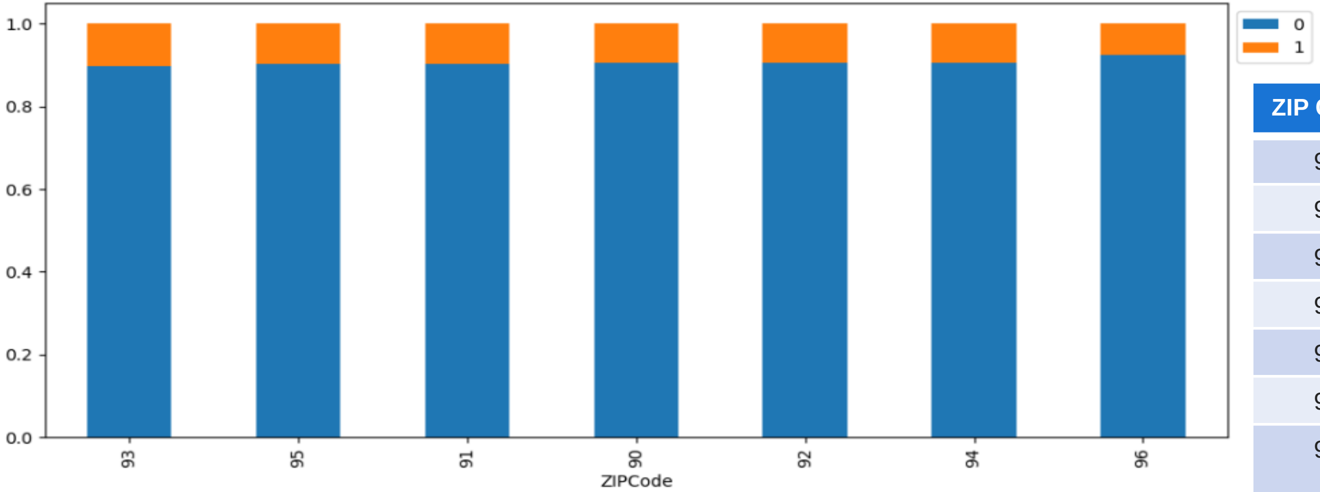
- 143 out of 1470 customers who use credit card issued by any other Bank accepted the personal loan offered in the last campaign
- 337 out of 3530 customers who do not use online banking accepted the personal loan offered in the last campaign

EDA - Bivariate Analysis

Personal Loan vs ZIP Code



```
Personal_Loan      0    1   All
ZIPCode
All             4520  480  5000
94              1334  138  1472
92               894   94   988
95               735   80   815
90               636   67   703
91               510   55   565
93               374   43   417
96                37    3    40
------------------------------------------------------------------------------------------------------------------------
```



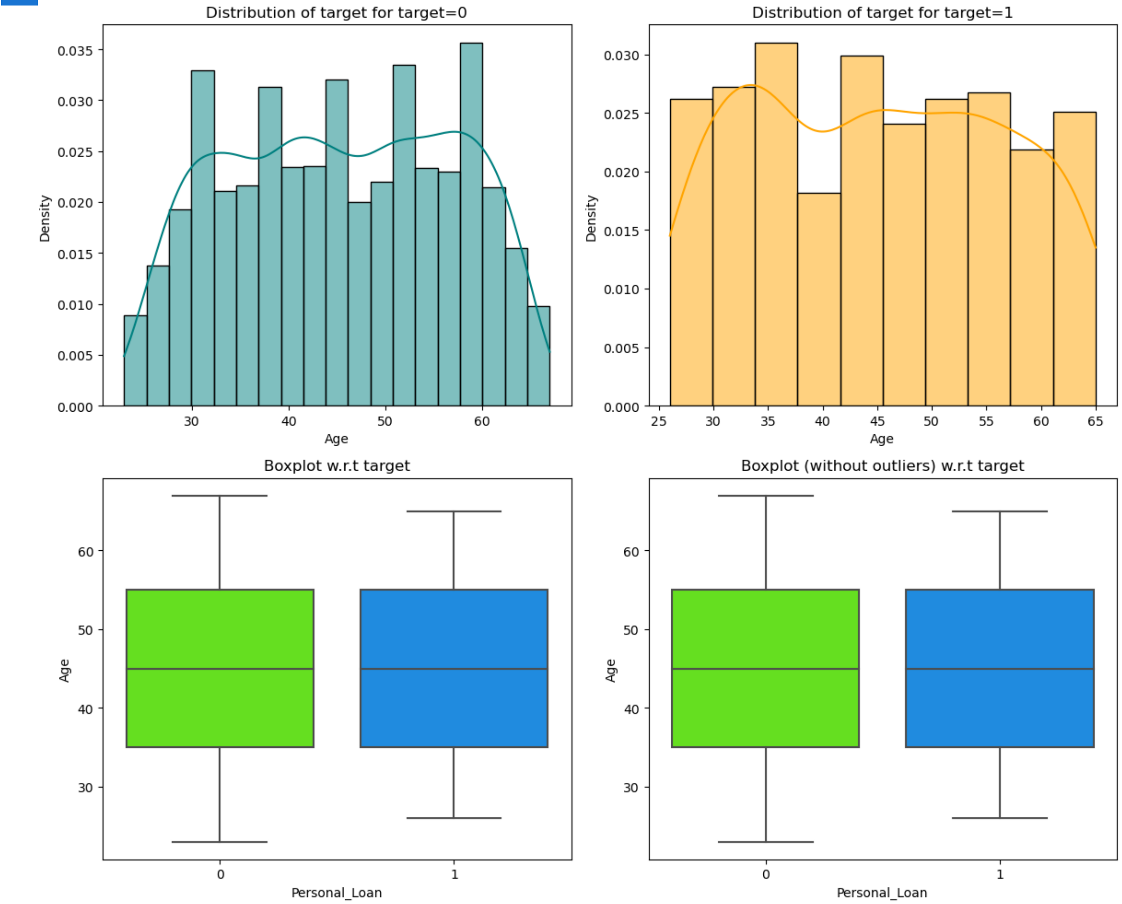
ZIP Code	Accept personal Loan (%)
90	9.5%
91	9.7
92	9.5
93	10
94	9.3
95	9.8
96	7.5

Observations:

- There are 467 unique ZIP code, and there are seven ZIP code that have the same first two digit

EDA - Bivariate Analysis

Personal Loan vs Age

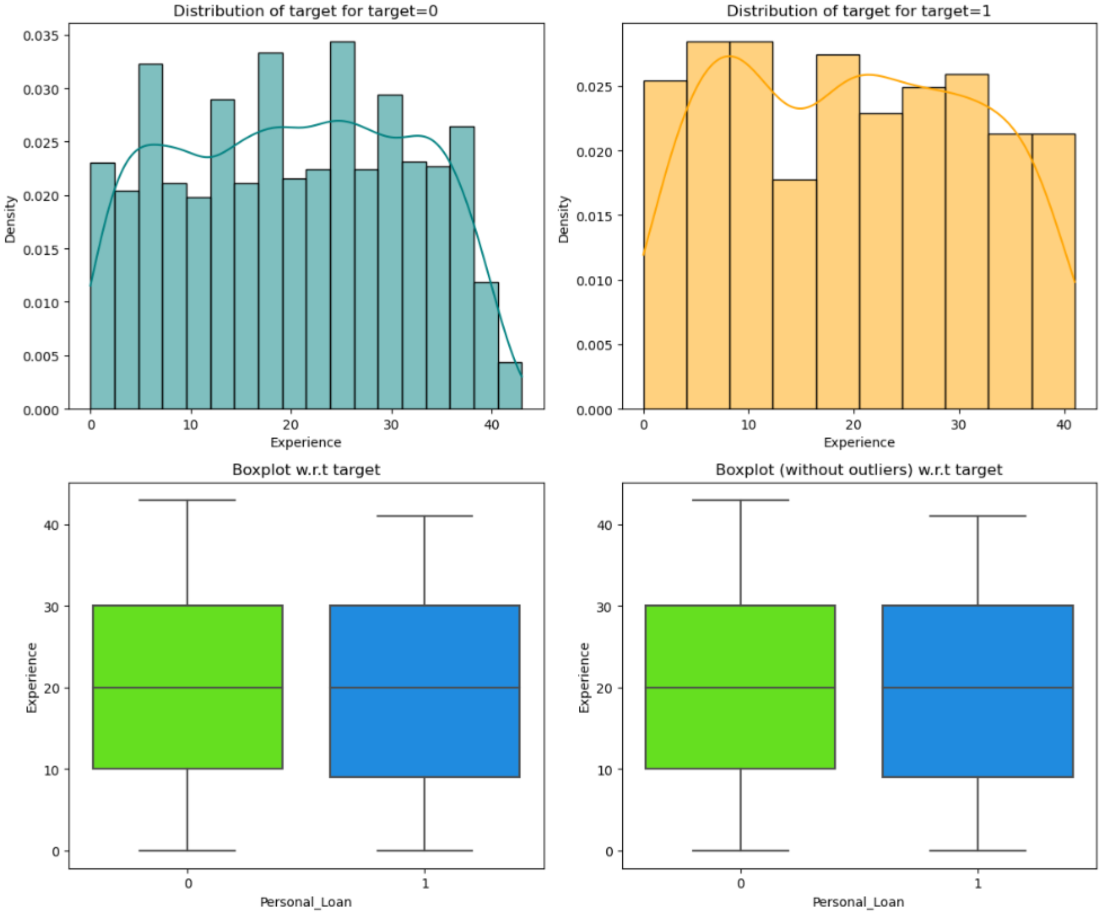


Observations:

- The customer age group between 45 and 65 who accepted the last personal loan campaign is very evenly distributed as compared to the age group between 20 and 40
- It is best used when comparing a variable's distribution between groups of another variable, a concept known as segmented univariate distribution.
- The distribution of the customer age for those customers who do not accept the last personal loan campaign is even among all ages, there's not a particular customer segment we can define
- The two boxplots, one with outliers and the other without outliers, plotted on Age vs Person Loan show almost the same indicates that there is no outlier data point in this variable.

EDA - Bivariate Analysis

Personal Loan vs Experience

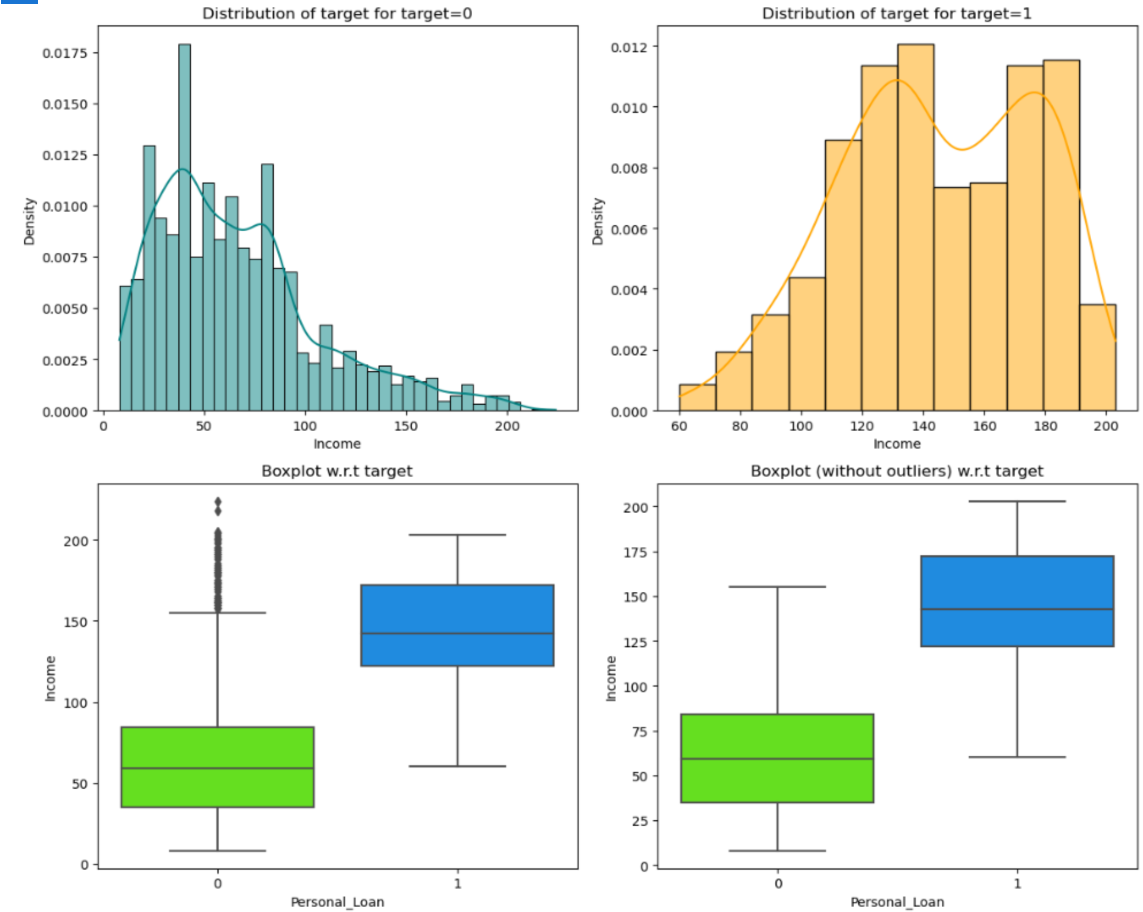


Observations:

- When compare the plot for years of professional experience vs personal loan acceptance, when compare the two distributions for the one that customer acceptance the loan vs rejected the loan, the two distribution are similar with one (accepted offers) slight skewed to the right
- The two boxplots, one with outliers and the other without outliers, plotted on Age vs Person Loan show almost the same indicates that there is no outlier data point in this variable

EDA - Bivariate Analysis

Personal Loan vs Income

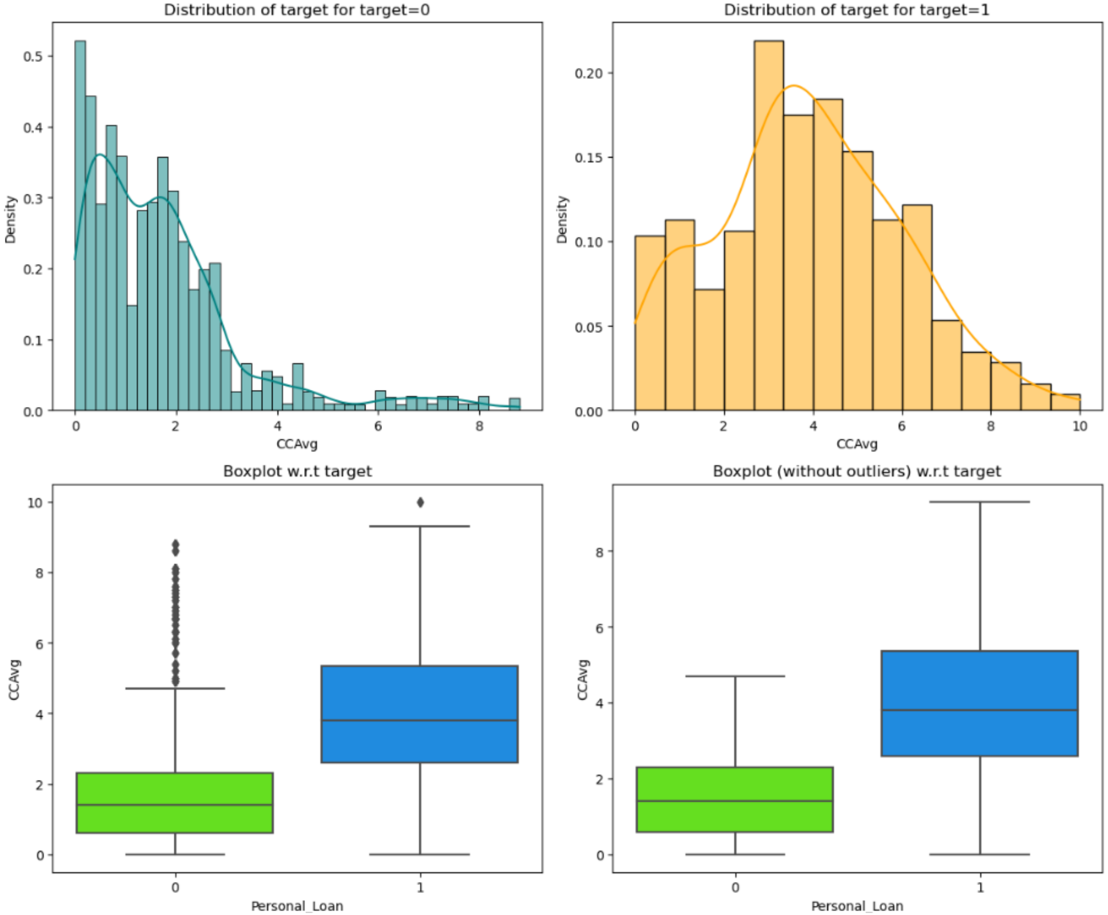


Observations:

- When compare the distribution of the customer incomes between those customers who accepted the personal loan offer vs those who rejected them, the result shows quite a big difference between the two output.
- shows that the median income of the customer who accepted the offer is greater than that of those who rejected.
- This difference also reflected in the Boxplot between the two variables as the median line of box (0) lies outside of the box of a comparison box (1)
- The distribution of the income is skewed to the right for the group of customers rejected the offers, mainly due to the outliers found in that dataset
- the highest number of customers who accepted the personal loan offer fell into two customer segments that have income between 125K to 175K. The distribution is also negatively skewed from the boxplot tail when the median is closer to the box's higher value.
- the boxplot that was executed without the outliers shows more dispersed which indicates that it reduced some statistical significance and distort the result slightly

EDA - Bivariate Analysis

Personal Loan vs Credit Card Spending

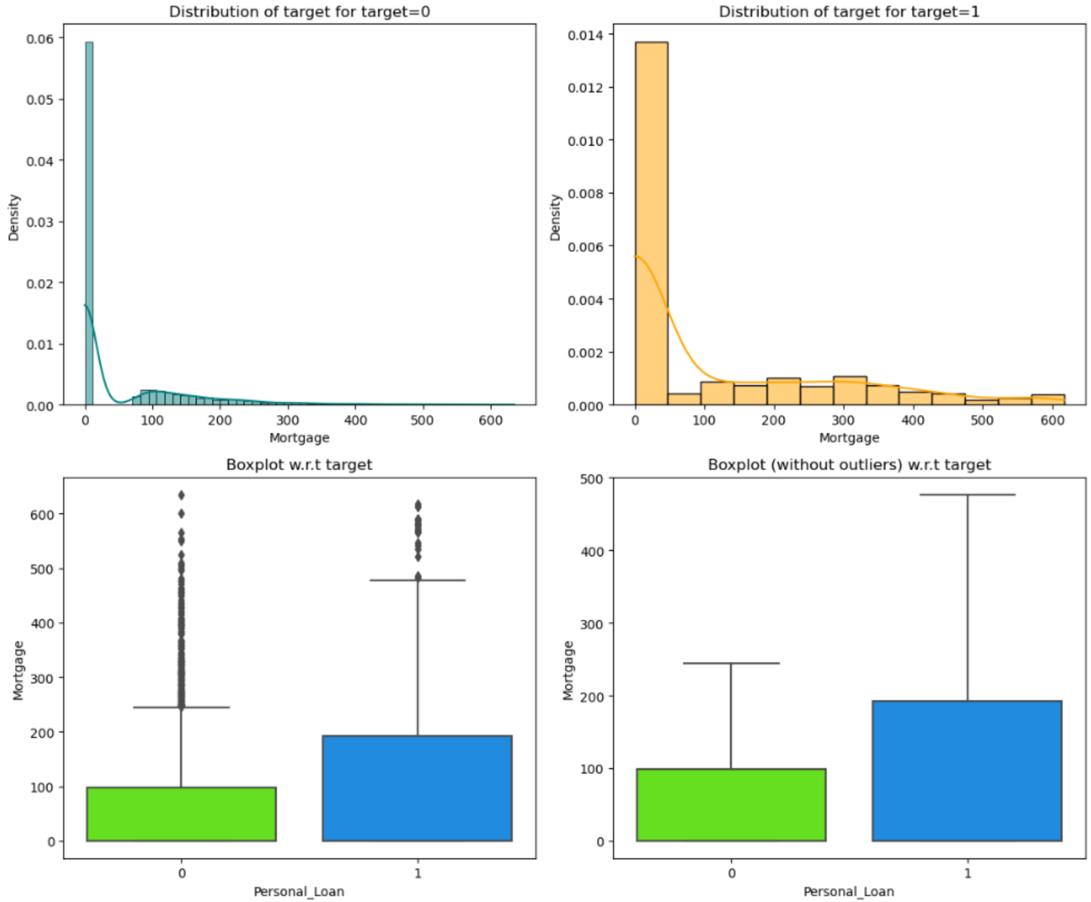


Observations:

- When compare the distribution of the customer credit card spending between those customers who accepted the personal loan offer vs those who rejected them, the result shows difference between the two output.
- This difference also reflected in the Boxplot between the two variables as the median line of box (0) lies outside of the box of a comparison box (1)
- The distribution of the credit card spending is extremely skewed to the right for the group of customers rejected the offers, mainly due to the outliers found in that dataset
- the highest number of customers who accepted the personal loan offer fell into two customer segments that have credit card spending between 3K to 5K
- the boxplot that was executed without the outliers shows more dispersed which indicates that it reduced some statistical significance and distort the result slightly

EDA - Bivariate Analysis

Personal Loan vs House Mortgage



Observations:

- The distribution of the Mortgage value is extremely skewed to the right for the group of customers, mainly due to the high number of customers who does not have any house mortgage which skewed the entire data population.
- The boxplot that was executed without the outliers shows more dispersed which indicates that it reduced some statistical significance and distorted the result.

Data Preprocessing

- **Duplicate value check:** There's one row with duplicate value found and the row was dropped from the data frame
- **Missing value treatment:** There is no missing value in the dataset
- **Outlier check (treatment if needed):** Outliers were identified as follows, they were removed as part of the EDA.

Columns	% of outlier
Age	0
Experience	0
Income	1.92
Family	0
Credit Card	6.48
Mortgage	5.82

- **Feature Engineering:** Converted 7 ZIP code columns to "category" data type
- **Data preprocessing for modeling:** checking for anomalous Values

➢ Mappes the values in Education columns as the following 1: Undergrad; 2: Graduate 3: Advanced/Professional

Model Building

Comment on the model performance

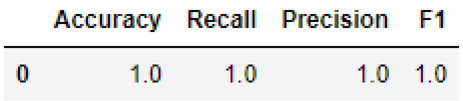
The Followings are the model building steps of Decision Tree :

- EDA (Univariate and Bivariate) was performed on the data frame, detail is described in slides 6- 21 to get good evaluation of the data before building the model, also data preprocessing was performed to ensure all anomalies are handled.
- Feature Engineering was also performed on ZIP code to convert them to Category data type to support a better build model
- Before proceeding with building the model, the data was split the data into train data (70%) and test data (30%), test and validation to be able to evaluate the model that you build on the train data, encode categorical features and scale numerical values.
- The model was built using the training data and the performance was checked. It used the DecisionTreeClassifier function in Sklearn library and used default 'gini' criteria to split. CART algorithm is used to restrict the test conditions to binary splits only.
- Two functions were created for reused throughout the model building process: 1) The model_performance_classification_sklearn function will be used to check the model performance of models. 2) The confusion_matrix_sklearnfunction will be used to plot confusion matrix.
- To avoid the decision tree to become biased toward a dominant classes, a dictionary {0:0.15,1:0.85} was passed to the model to specify the weight of each class and the decision tree will give more weightage to class 1, class_weight is a hyperparameter for the decision tree classifier.

Model Building – initial Build

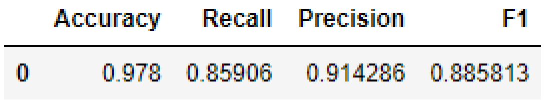
- During **initial built of the model**, Model can perfectly classify all the data points on the training set with 100% accuracy and recall score. 0 errors on the training set, each sample has been classified correctly.
- However, since there's no restriction applied, the trees continue to grow with all possible patterns in the training set. The result on the performance of the test set is 97% accuracy and 89% Recall. This huge disparity in performance of model on training set and test set, which suggests that the model is overfitting.

Model performance on training data

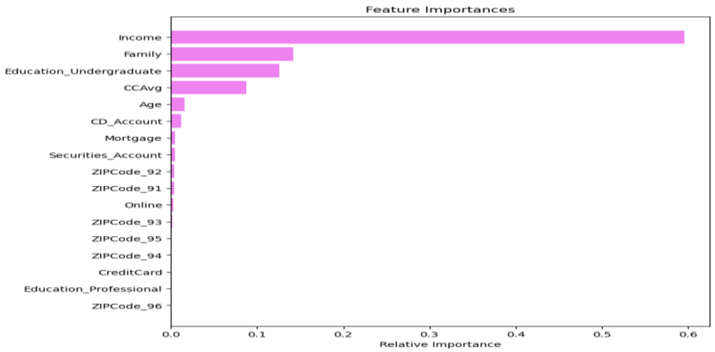


	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

Model performance on test data

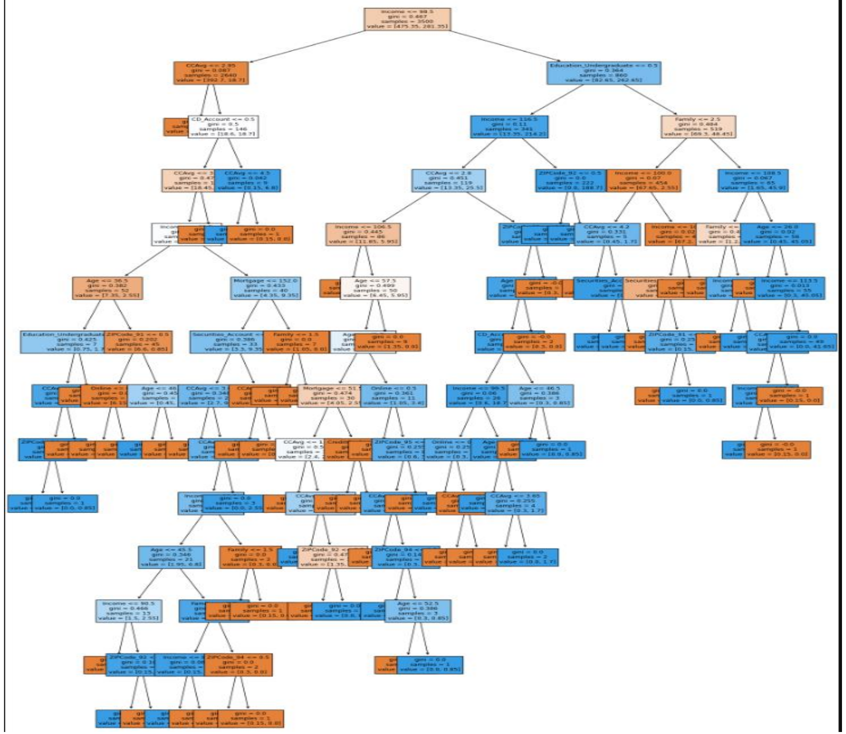


	Accuracy	Recall	Precision	F1
0	0.978	0.85906	0.914286	0.885813



- Income, Family and Education are the top 3 important features.
- Decision rules – IF (condition) –THEN (prediction) statement at each tree node is used for making predication based on the algorithm the model uses

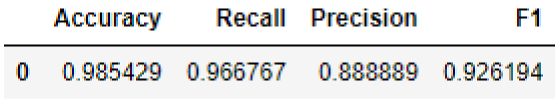
Visualizing the Decision Tree on Train data



Model Performance Improvement – Pre-Pruning

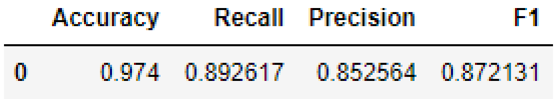
- To reduce overfitting, **pre-pruning method** was applied by using **GridSearch** was used for Hyperparameter tuning of the tree model to compute the optimum values of hyperparameters. Grid parameters that were passed in are “max-depth”, “criterion”, “splitter” and “mini_impurity_decrease”
- Recall on the test set has improved from 0.85906 to 0.865772 and this is an improvement because now the model is not overfitting, and we have a generalized model. Visualization of the tree (see appendix) was observed. The most important Feature is “Income” followed by Family and Education.

Model performance on training data

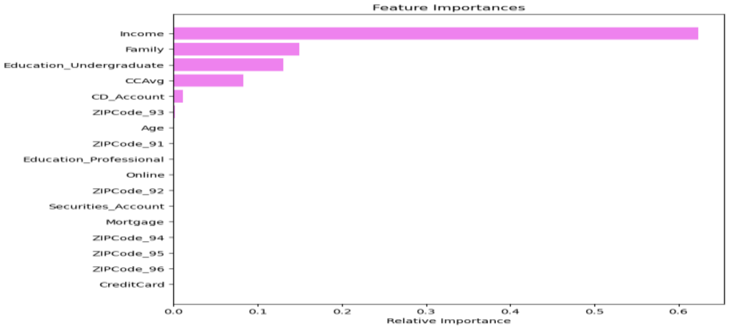


	Accuracy	Recall	Precision	F1
0	0.985429	0.966767	0.888889	0.926194

Model performance on test data

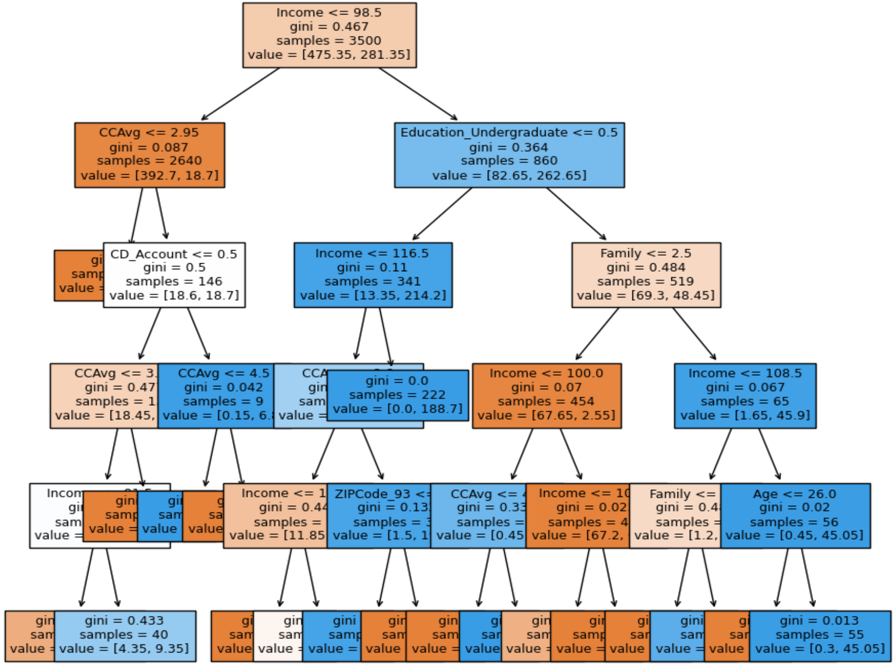


	Accuracy	Recall	Precision	F1
0	0.974	0.892617	0.852564	0.872131



- Income, Family and Education are the top 3 important features.
- Decision rules – IF (condition) –THEN (prediction) statement at each tree node is used for making predication based on the algorithm the model uses

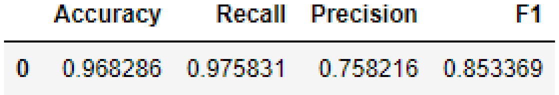
Visualizing the Decision Tree on training data



Model Performance Improvement – Cost-Complexity Pruning

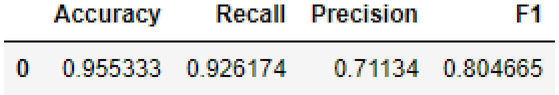
Cost complexity pruning was used to prune and control the size of the tree to further prevent the model from overfitting. Minimal cost complexity pruning recursively finds the node with the "weakest link". The weakest link is characterized by an effective alpha, where the nodes with the smallest effective alpha are pruned first. Greater values of ccp_alpha increase the number of nodes pruned. Here we only show the effect of ccp_alpha on regularizing the trees and how to choose a ccp_alpha based on validation scores.

Model performance on training data

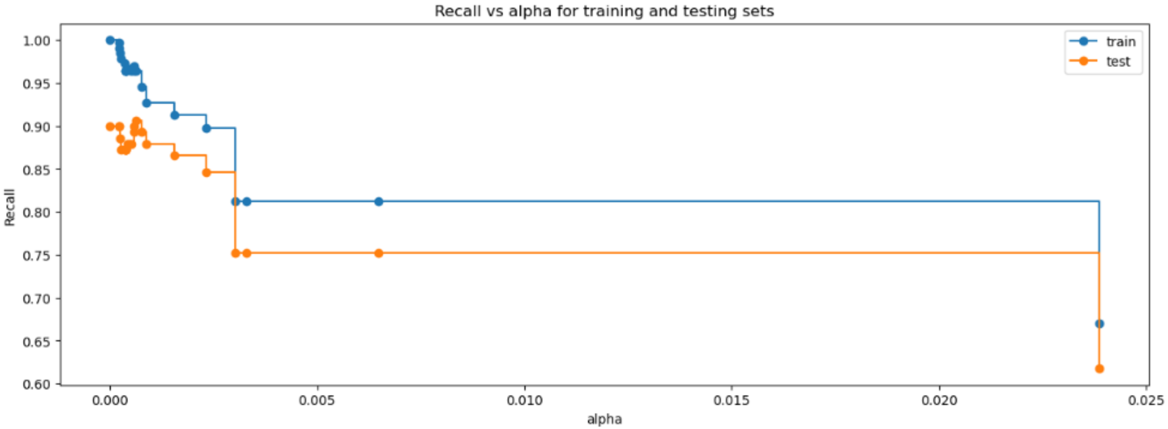


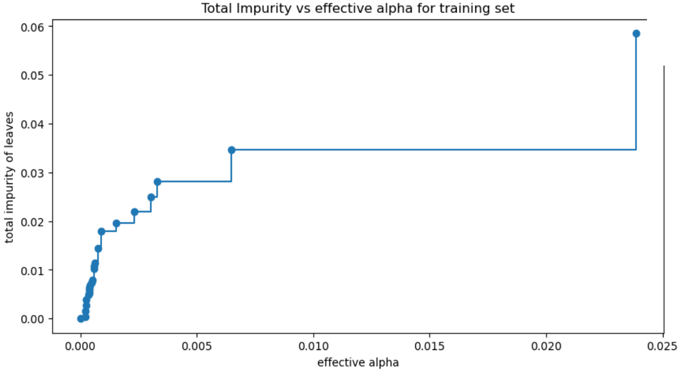
	Accuracy	Recall	Precision	F1
0	0.968286	0.975831	0.758216	0.853369

Model performance on test data



	Accuracy	Recall	Precision	F1
0	0.955333	0.926174	0.71134	0.804665





- As alpha increase on the training data, the most complex tree has a very good fit but it loose as it goes to the right
- as alpha increase on the test data, the most complex tree has the worst performance but as the tree becomes simpler, it becomes better and then it decrease
- therefore, take the best model in the middle and create decision tree classifier with that alpha for the best model
- Maximum value of Recall is at 0.0006 alpha, however, instead I decided to choose alpha 0.003 to retain information and get higher recall
- Recall on the test set has improved reduced from 0.892617 to 0.926174

- Income, Family and Education are the top 3 important features.

Model Performance Summary

Model evaluation criterion

- The objective is to predict whether a liability customer will buy personal loans.

Model can make wrong predictions as:

1) Predicting a customer will take the personal loan but in reality the customer will not take the personal loan **-** Loss of resources

2) Predicting a customer will not take the personal loan but in reality the customer was going to take the personal loan **-** Loss of opportunity

Which case is more important?

- Losing a potential customer by predicting that the customer will not be taking the personal loan but in reality the customer was going to take the personal loan.

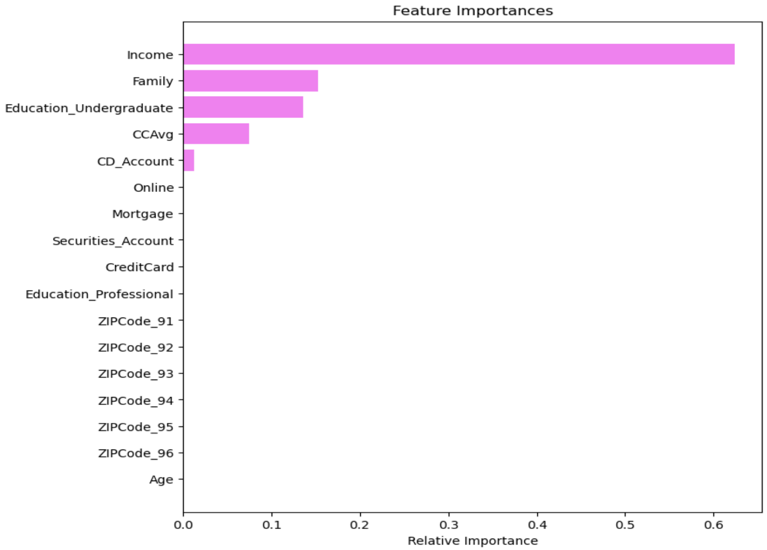
How to reduce this loss i.e need to reduce False Negatives?

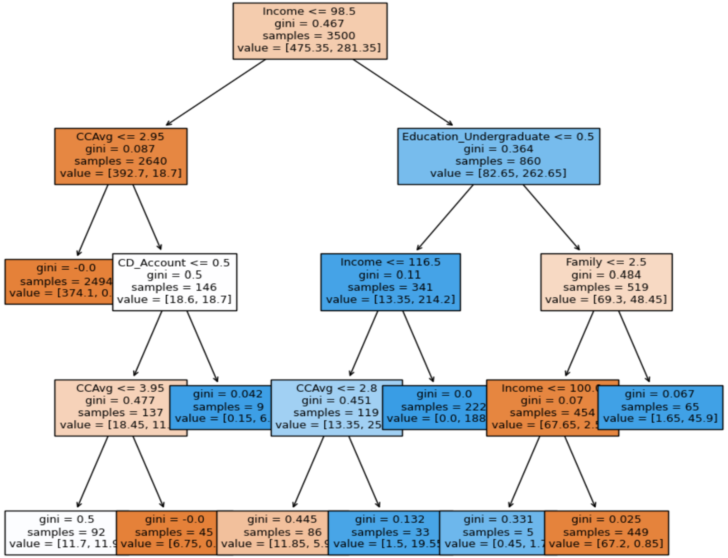
- Bank would want “**Recall” to be maximized**, greater the Recall higher the chances of minimizing false negatives. Hence, our focus for the model building is on increasing Recall or minimizing the false negatives.
- Cross-validation technique is used in the model evaluation from a predictive point of view is executed solely based on observed data while making modifications in order to preserve the accuracy of parameter estimation as much as possible.

Model Performance Summary – con’t

Summary of the most important features for prediction

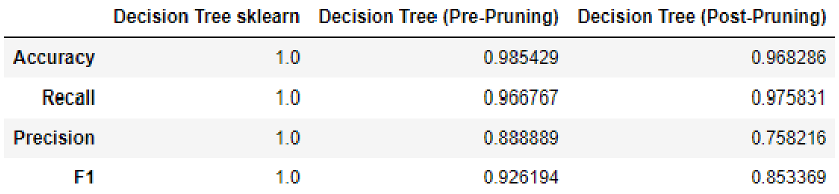
The overall importance of a feature in a decision tree can be computed by going through all the splits for which the feature was used and measure how much it has reduced Gini index compared to the parent node, Inspect the feature indicated by the root and visit one of its children depending on the feature’s value. Then, we **repeat the process until we reach a leaf node and read the decision.** The root node in a decision tree is our starting point, If we were to use the root node to make predictions, it would predict the mean of the outcome of the training data. The maximum allowed depth for this tree is set to 5. The root Income feature which represent the income of the customer to be used to predict future possibility of purchasing bank loan and With the next split, the Education, Credit Card balance have been selected we either subtract or add a term to this sum, depending on the next node in the path. To get to the final prediction, we must follow the path of the data instance that we want to explain and add the contributions for each of the features and get an interpretation of how much each feature has contributed to a prediction. In the final model the feature importance measure shows that the Education is far more important than CCAvg in the next split hence the node is selected in the path.





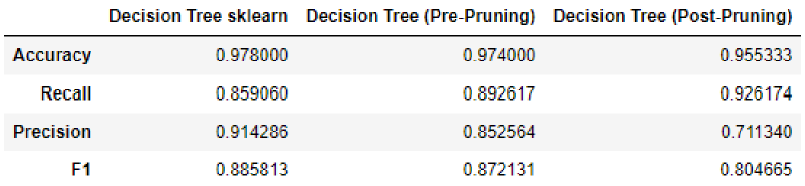
Model Performance Summary – Model Permance Comparison

Model performance on training data in the final Decision Tree



	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	1.0	0.985429	0.968286
Recall	1.0	0.966767	0.975831
Precision	1.0	0.888889	0.758216
F1	1.0	0.926194	0.853369

Model performance on test data in the final Decision Tree



	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.978000	0.974000	0.955333
Recall	0.859060	0.892617	0.926174
Precision	0.914286	0.852564	0.711340
F1	0.885813	0.872131	0.804665

Observation:

- The initial decision tree, the recall is at 100% in the train set, the recall is 86% in the test data which suggests that the model is overfitting.
- When used hyperparameter tuning we improved the recall for both set a little bit.
- But when we did post pruning, we improved the training recall but most importantly we improved the test recall by quite a lot.
- The pre-pruned and the post-pruned models have reduced overfitting, and the model is giving a generalized performance.
- The final tree model has given the best recall score, and it is able to identify 93% of the customers who would purchase the personal loan from the Bank.