AllLife Bank

Project : Personal Loan Campaign
Course : Supervised Learning

Document Version : 1.0
Document Owner : Rahul Kulkarni
Document ID : Project 2 – Personal_Loan_Campaign.pdf
Submission Date : 22nd July 2023

Contents

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview & Analysis
- EDA - Univariate & Bivariate Analysis
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

Executive Summary – Business Context

- **Business Context:** AllLife, a US Bank, has a growing customer base. Majority of these customers are Liability Customers (Depositors) with varying sizes of deposits. The number of Asset customers (Borrowers) is quite small, and the bank is interested in expanding this customer base to bring in more loan business and thereby increase revenues by interests on loans.
- **The Problem Statement :** The bank ran a campaign last year for its liability customers, who showed a healthy conversion rate of over 9% success. The retail marketing department wants to devise campaigns with better target marketing to increase the above success ratio. In a nutshell, the management is interested in exploring different methods to convert its liability customers to personal loan customers, while still retaining them as depositors
- **Solution Approach:** In order to resolve the above problem, we will undertake the following key tasks:
  - Perform a deep-dive on the previous Loan Modelling dataset using libraries such as numpy and pandas for data manipulation, and seaborn and matplotlib for data visualisation
  - Perform exploratory data analysis on the dataset to deliver key findings and insights
  - Identify key customer attributes of the dataset that are most significant in driving purchases
  - Build a model that will be able to predict whether a lability customer will buy personal loans or not
  - Identify target customer segments in order to boost potential customer acquisition
  - Recommend opportunities for improvement, that will help the marketing team to lead a successful campaign

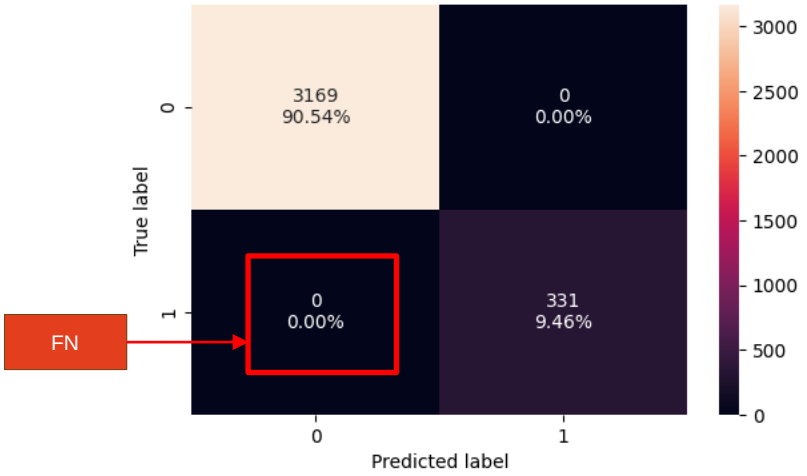
Executive Summary – Model, Evaluation Criteria & Approach

- **Model Evaluation Criteria :** The primary objective for building the model is to predict whether a liability customer will buy personal loan. Using the confusion matrix as guiding principle, it is imperative to focus on reducing the False Negatives i.e., predicting that a customer will not buy the personal loan, but eventually would have bought the personal loan. This would be a significant loss of opportunity and thereby loss of revenue. Besides the model will be used by Marketing to launch a campaign, which needs to have the maximum reach. So, if FN is high, that means we will be losing out on reaching potential customers. This implies that **reducing False Negatives** should be of utmost importance to the business
    - Key Criteria – Recall: The bank should therefore use Recall as the key model evaluation criteria – higher the Recall, greater are the chances of minimising False Negatives
- **Model Building Approach:** Based on the above, we have built the following 3 models and evaluated its performance, such that it satisfies our key criteria. We have built our model using the DecisionTreeClassifier function and used the default 'gini' criteria to split within our model. Also, based on the dataset, it is observed that we have approx., 10% of positive classes, which means if our model marks each sample as negative, then we would at least 90% accuracy. This implies that 'Accuracy' would not be a good metric to evaluate the performance of our model. This re-affirms our hypothesis that we should 'Recall' as the effective measure of performance for the below 3 models
    - Model 1: Initial Model
    - Model 2: Pre-Pruned Model using Hyper-Parameter Tuning
    - Model 3: Post Pruned Model using Cost Complexity
    -

Executive Summary – Model 1: Initial Model

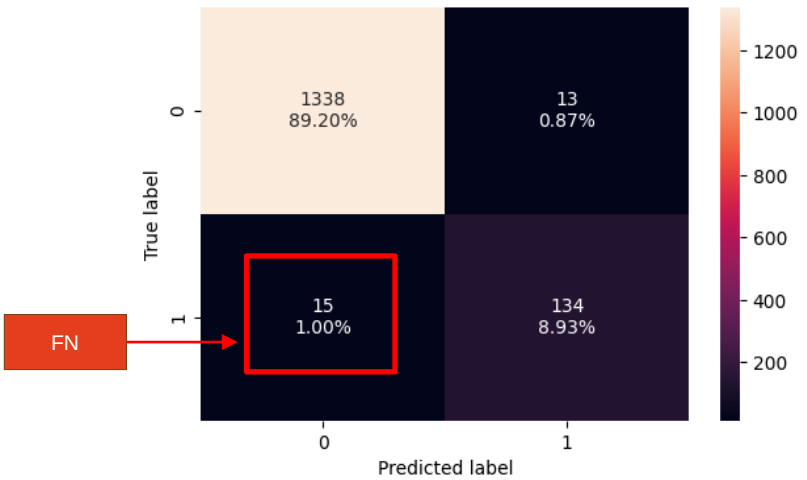
- **Model 1 – Initial Model:**
  - Recall on Training Dataset is 1 and on Testing Dataset is 0.89, which means the model is overfitting

Training Dataset



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

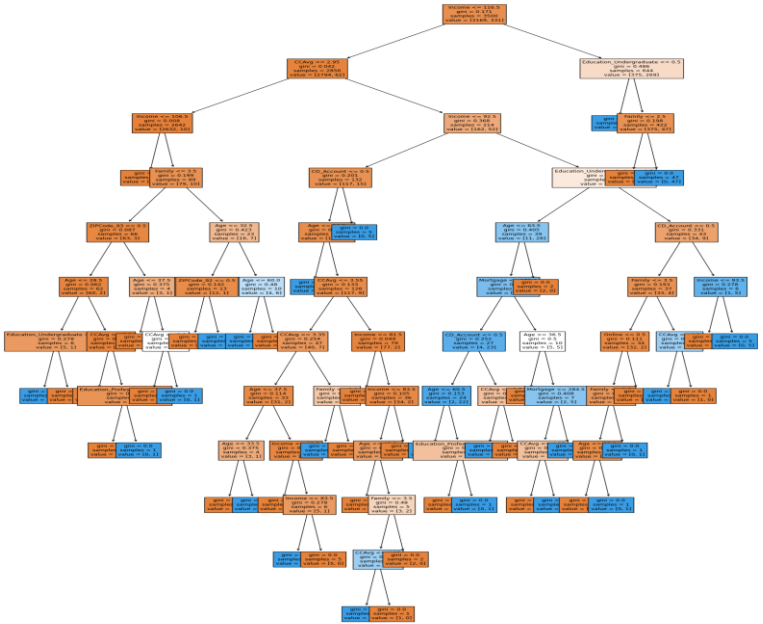
Testing Dataset



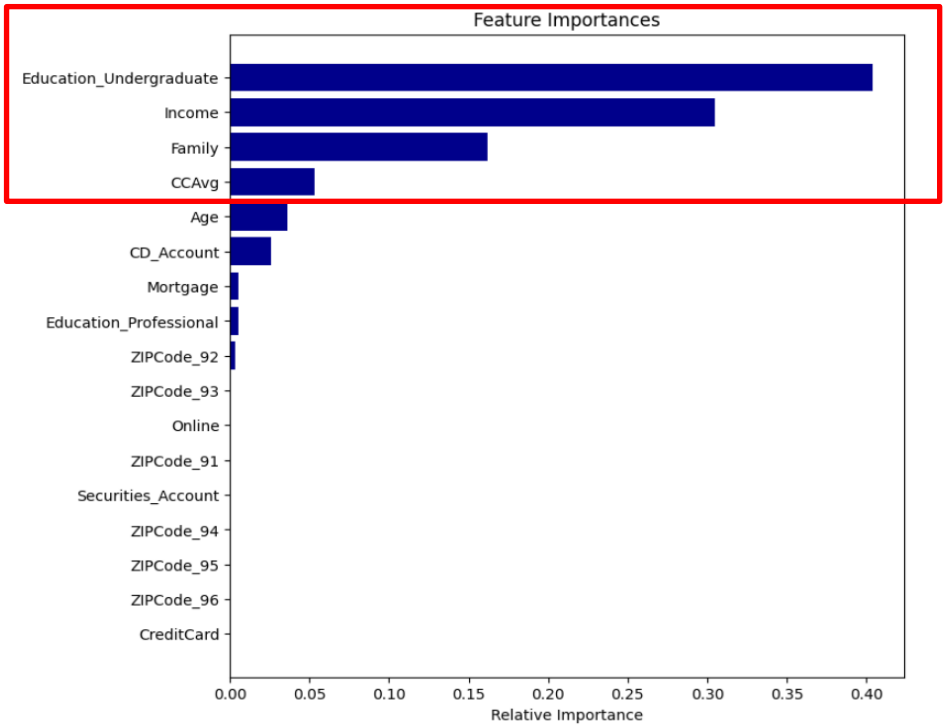
	Accuracy	Recall	Precision	F1
0	0.981333	0.899329	0.911565	0.905405

Model Building – Model 1: Initial Model (Cont'd)

- **Model 1 - Initial Model Decision Tree**



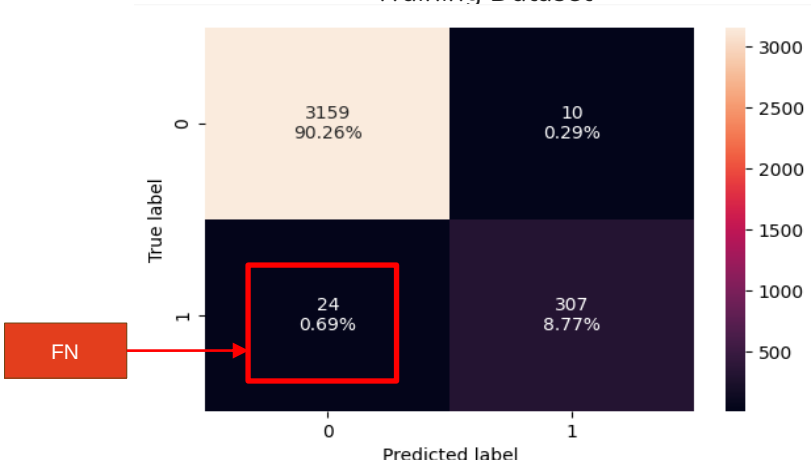
- **Most Important Features -** Education_Undergraduate, Income, Family and CCAvg at 40.3%, 30.4%, 16.1% & 5.3% respectively



Executive Summary – Model 2: Pre-Pruned (Hyper Parameter)

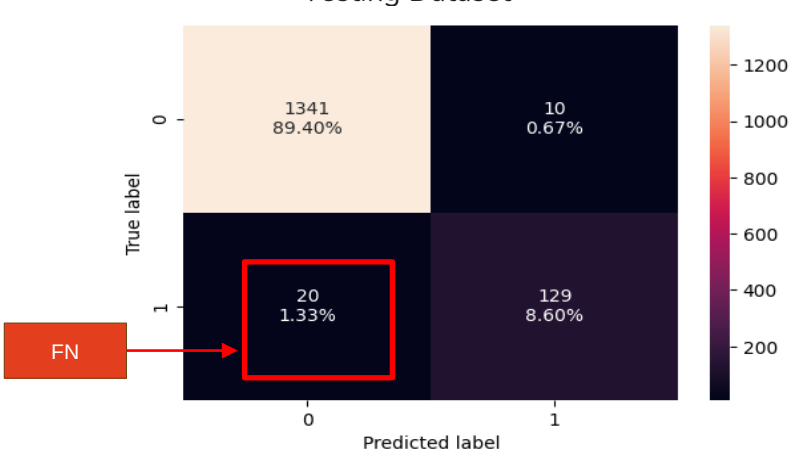
- **Model 2 – Pre-Pruned Model :**
  - We have built this model using GridSearchCV for hyper-parameter tuning.
  - Recall on Training Dataset is 0.92 and on Testing Dataset is 0.86, which means the model is not overfitting

Training Dataset



	Accuracy	Recall	Precision	F1
0	0.990286	0.927492	0.968454	0.947531

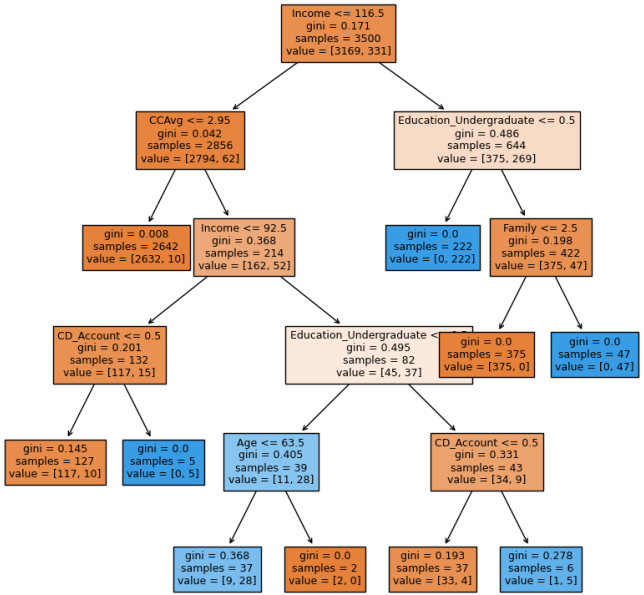
Testing Dataset



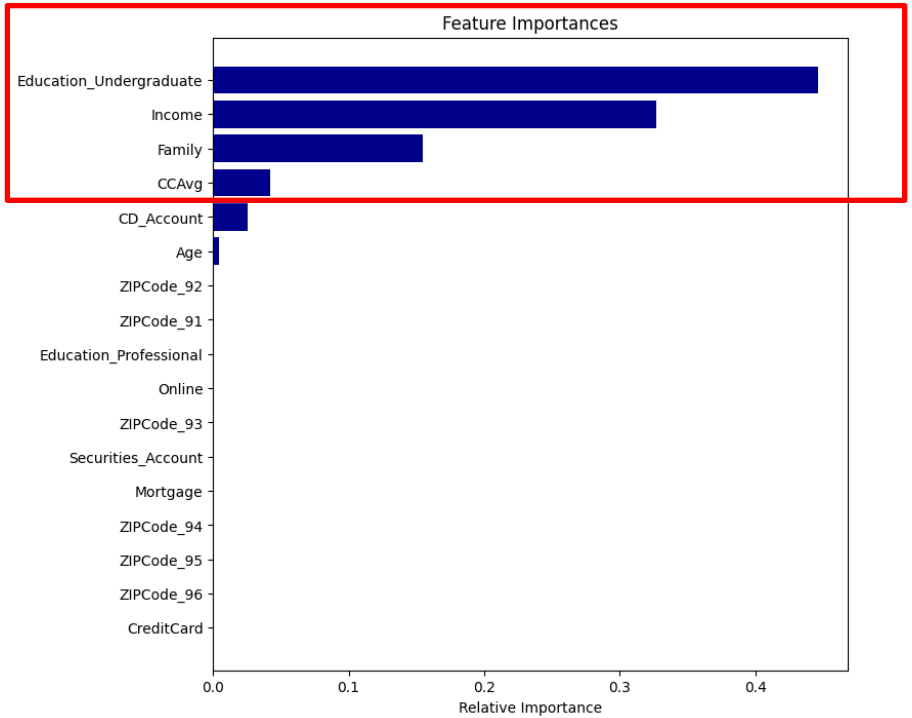
	Accuracy	Recall	Precision	F1
0	0.98	0.865772	0.928058	0.895833

Executive Summary – Model 2: Pre-Pruned (Cont'd)

- **Model 2 – Pre-Pruned Decision Tree**



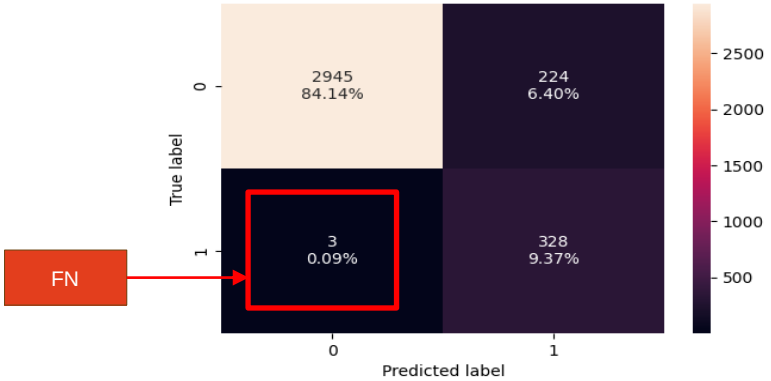
- **Most Important Features -** Education_Undergraduate, Income, Family and CCAvg at 44.6%, 32.7%, 15.5% & 4.2% respectively



Executive Summary – Model 3: Post-Pruned (Cost Complexity)

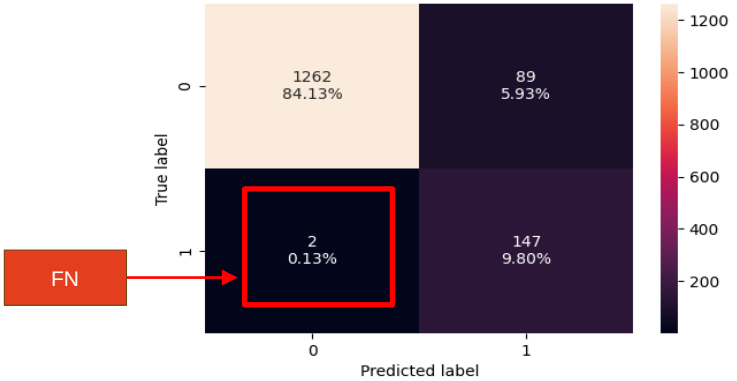
- **Model 3 – Post-Pruned Model (Cost Complexity) :**
  - We have built this model using DecisionTreeClassifier and set Cost Complexity Parameter Alpha to 0.010
  - Recall on Training Dataset is 0.99 and on Testing Dataset is 0.98, which means the model is fitting optimally

Training Dataset



	Accuracy	Recall	Precision	F1
0	0.935143	0.990937	0.594203	0.742922

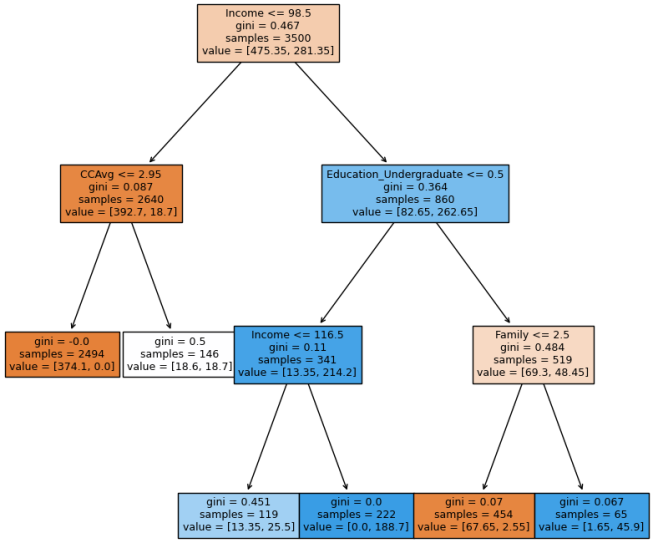
Testing Dataset



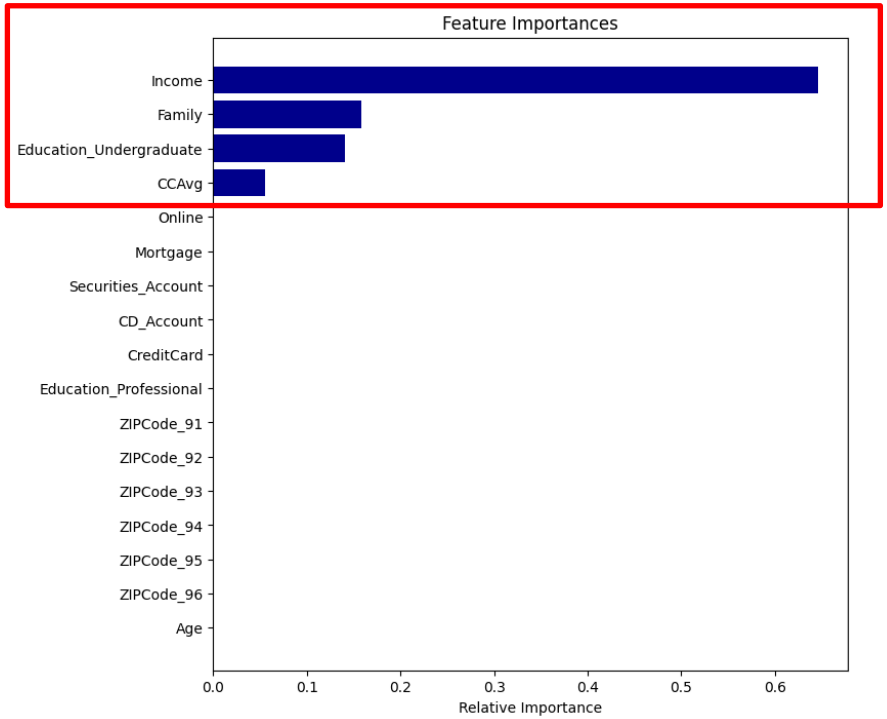
	Accuracy	Recall	Precision	F1
0	0.939333	0.986577	0.622881	0.763636

Executive Summary – Model 3: Post-Pruned (Cont'd)

- **Model 3 – Post-Pruned Decision Tree**



- **Most Important Features** – Income, Family, Education_Undergraduate, and CCAvg at 64.5%, 15.8%, 14.09%, and 5.5% respectively



Executive Summary – Model Performance Summary (All)

- **Model 3 Post Pruned - Best Fit Model:**
  - Model 3 – Post Pruned (Cost Complexity Alpha) seems to be the best fit model for Personal Loan Campaign since the Recall for Training and Testing Dataset is 0.99 and 0.98, respectively. The model is built using the DecisionTreeClassifier and the ccp_alpha parameter is set to 0.010 to achieve best results. Important features of the model are Income, Family, Education_Undergraduate, and CCAvg at 64.5%, 15.8%, 14.09%, and 5.5% respectively. With an effective Alpha of 0.010, the depth of the tree is 3 and the total nodes are near to 11. The model is easy to interpret and doesn't overfit the Training Dataset

- **Model 1 & Model 2:** The Pre-Pruned and the Post-Pruned models have reduced overfitting and the model is giving a generalized performance.

- **Performance Summary:** The table below summarizes the performance of all 3 models for their Accuracy, Recall, Precision and F1 scores on both – Training & Testing Datasets

#	Model Type	Accuracy		Recall		Precision		F1	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
1	Model – 1 Initial Model	1.0	0.981333	1.0	0.899329	1.0	0.911565	1.0	0.905405
2	Model 2 Pre-Pruned (Hyper Parameter)	0.990286	0.98	0.927492	0.865772	0.968454	0.928058	0.947531	0.895833
3	Model 3 Post Pruned (Cost Complexity Alpha)	0.994571	0.978667	0.990937	0.986577	0.945714	0.897959	0.9721	0.891892

Executive Summary - Conclusion

- Model 3 – Post Pruned (Cost Complexity Alpha) seems to be the best fit model for Personal Loan Campaign.
- With a high Recall score of 0.99 & 0.98, on the Training & Testing dataset, the model will minimise False Negatives, which is of utmost importance to the business.
- The model built can be used to predict whether a liability customer will buy personal loans or not. This will help the marketing department to identify the potential customers who have a higher probability of purchasing the loan.

Executive Summary - Key Actionable Business Insights

- **Summarised Key Insights :**
  - Customer attributes such as Income, Family, Education, and Credit Card spend are the most important features in predicting potential customers for a personal loan
  - Income seems to be the most important feature at 64.5%, followed by Family – 15.8%, Education - 14.09%, and Average Monthly Credit Card Spend – 5.5%
  - Customers with an annual income of less than $98.5K are less likely to have a personal loan, thus it is recommended to target potential customers in this segment.
  - Customers with an income greater than $98.5K and with a higher level of education (Graduate & Professionals) most likely already have a personal loan. It is therefore recommended to cross-sell other products of the bank
  - Customers with a growing family are more likely to avail of a personal loan. It is recommended to target customers with a family size of 3 and / or 4 family members
  - Customers using Credit Cards frequently are deemed highly credit-worthy and as such are potential buyers of personal loan.
  - Banks' existing customers holding Securities Accounts and Certificate Of Deposit Accounts are more likely to buy a personal loan and are a likely target for conversion
  - Customers using the online facilities are more likely to already have a personal loan

Executive Summary - Our Recommendation

Based on our key observations and insights, we recommend the following areas of improvement / opportunities that will drive business growth and lead to a better customer experience

- **Implement a Customer Centric Digital Channel to increase customer footprint:** Given that, customers who used the online facilities already have a loan, building a holistic user centric digital channel (Mobile App & Website with FAQs) which simplifies the existing loan to value application process from a customer perspective is likely to attract more prospects, thereby increasing the likelihood of selling loans to potential customers.
- **Implement Customer Incentivisation Scheme to cross sell products:** Incentivising existing customers (11.49% of Security Account Holders and 46.35% Certificate Of Deposit Account of the Bank) by offering them special interest rates / rebates on personal loans will drive customer gowth and increase revenue.

Business Problem Overview & Solution Approach

Business Problem Overview and Solution Approach

- **Business Context:** AllLife, a US Bank, has a growing customer base. Majority of these customers are Liability Customers (Depositors) with varying sizes of deposits. The number of Asset customers (Borrowers) is quite small, and the bank is interested in expanding this customer base to bring in more loan business and thereby increase revenues by interests on loans.
- **The Problem Statement :** The bank ran a campaign last year for its liability customers, who showed a healthy conversion rate of over 9% success. The retail marketing department wants to devise campaigns with better target marketing to increase the above success ratio. In a nutshell, the management is interested in exploring different methods to convert its liability customers to personal loan customers, while still retaining them as depositors
- **Solution Approach:** In order to resolve the above problem, we will undertake the following key tasks:
  - Perform a deep-dive on the previous Loan Modelling dataset using libraries such as numpy and pandas for data manipulation, and seaborn and matplotlib for data visualisation
  - Perform exploratory data analysis on the dataset to deliver key findings and insights
  - Identify key customer attributes of the dataset that are most significant in driving purchases
  - Build a model that will be able to predict whether a lability customer will buy personal loans or not
  - Identify target customer segments in order to boost potential customer acquisition
  - Recommend opportunities for improvement, that will help the marketing team to lead a successful campaign

Data Overview & Analysis

Data Overview & Analysis

- The Loan Modelling dataset has the following Data-Structure:

#	Columns	Data-type	Total Rows	Description
1	ID	Integer 64	5000	Customer ID
2	Age	Integer 64	5000	Customer's age in completed years
3	Experience	Integer 64	5000	Number of years of professional experience
4	Income	Integer 64	5000	Annual income of the customer (in thousand dollars)
5	ZIPCode	Integer 64	5000	Home Address ZIP code
6	Family	Integer 64	5000	Family size of the customer
7	CCAvg	Float 64	5000	Average spending on credit cards per month (in thousand dollars)
8	Education	Integer 64	5000	Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional
9	Mortgage	Integer 64	5000	Value of house mortgage if any. (in thousand dollars)
10	Personal_Loan	Integer 64	5000	Did this customer accept the personal loan offered in the last campaign?
11	Securities_Account	Integer 64	5000	Did this customer accept the personal loan offered in the last campaign?
12	CD_Account	Integer 64	5000	Does the customer have a certificate of deposit (CD) account with the bank
13	Online	Integer 64	5000	Do customers use internet banking facilities?
14	CreditCard	Integer 64	5000	Does the customer use a credit card issued by any other Bank (excluding All life Bank)

Data Overview & Analysis (Cont'd)

- Total No. Of Columns: 14 | Total No. Of Rows: 5000
- Column Data-types: Float (1), Integer 64 (13)
- Missing Values: There are No missing values in the data-set

Data Overview & Analysis (Cont'd)

- Statistical Summary: Following is the statistical summary of the dataset

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

Data Overview & Analysis – Key Observations & Insights

- **Key Observations:**
  - There are no missing values in the dataset
  - There are no null values in dataset
  - There are no duplicate values in the dataset
  - The minimum and maximum Age is 23 years and 67 years respectively, whereas the mean Age is 45 years
  - The minimum Experience is -3 years, which probably seems to be a data-entry error and needs to be investigated
  - The minimum and maximum Income is USD $8K and USD $224K respectively, whereas the mean is approx. USD $73K
  - The minimum and maximum Family size is 1 and 4 respectively, whereas the mean Family size is approx. 2
  - The minimum and maximum CCAvg is USD $0 and USD $10K respectively, with a mean CCAvg of USD $1.93K
  - The minimum and maximum Mortgage is USD $0 and USD $635K respectively, with a mean of approx. USD $56K

Data Overview & Analysis – Key Observations & Insights

- **Key Insights:**
  - The Personal_Loan is an integer 64 variable but can be converted to a categorical variable where 1 is if the customer accepted the personal loan offered in the last campaign and 0 is if the customer did not accept the loan
  - The Education has an integer 64 value: 1: Undergrad; 2: Graduate;3: Advanced/Professional but can be converted to a categorical variable
  - The Securities_Account is an integer 64 variable but can be converted to a categorical variable where 1 is if the customer has a Securities_Account and 0 is if the customer does not have a Securities_Account
  - The CD_Account is an integer 64 variable but can be converted to a categorical variable where 1 is if the customer customer has a CD_Account and 0 is if the customer did not have a CD_Account
  - The Online is an integer 64 variable but can be converted to a categorical variable where 1 is if the customer uses Online banking and 0 is if the customer does not use online banking
  - The CreditCard is an integer 64 variable but can be converted to a categorical variable where 1 is if the customer uses a credit card issued by any other Bank and 0 is if the customer does not use a credit card issued by any other Bank
  - The Zip code locations are integer 64 variables, that would need further data pre-processing. These can be converted to categorical variable so that they can be used for our model to predict effectively.

Data Preprocessing

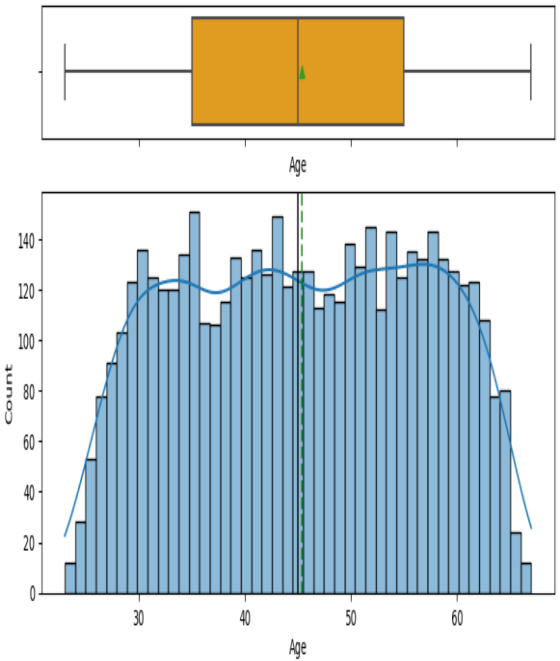
Data Preprocessing – Key Observations & Insights

- **Duplicate Value Check :** There are no duplicate values in the dataset
- **Missing Value Treatment :**There are no missing values in the dataset
- **Outlier Check & Treatment :**
  - Experience has negative values of -1,-2 and -3, which probably seems to be a data-entry error. These have been replaced with 1, 2 and 3. There are approx. 49 records that have negative values (32 of -1, 14 of -2, and 3 of -3)
- **Feature Engineering:**
  - ZipCode has interger values - there are 467 unique Zip codes and 4533 duplicates. By considering only the first 2 digits of the ZipCode, we have categorised these into 7 unique categories (90, 91, 92, 93, 94, 95, and 96)across the dataset
- **Data Pre-processing for Modelling:**
  - Education has integers values and have been pre-processed to categorical values as following : 1: Undergraduate; 2: Graduate; and 3: Professional
  - Personal_Loan, Securities_Account, CD_Account, Online, and CreditCard have been converted to categorical variables
  - ID column has been dropped-off from the data set for effective modelling

EDA - Univariate Analysis

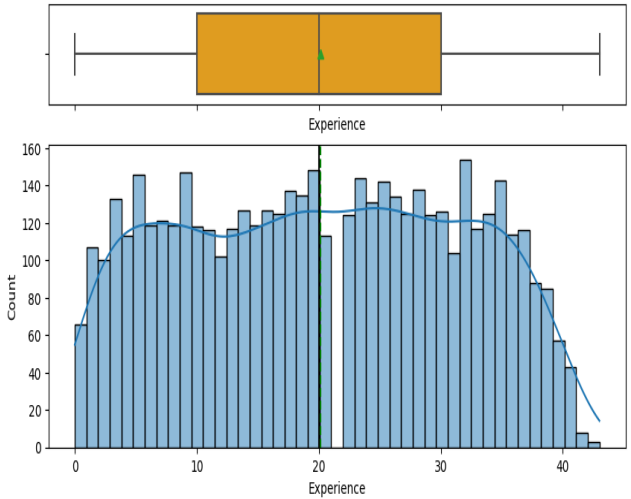
EDA - Univariate Analysis - Age

- **Age:**
  - Highest Age Group: There are circa 151 people of approx. 35 years of age
  - Minimum: The minimum age is 23
  - Q1: 25% of the population are less 35 years of age
  - Q3: 75% of the population are less 55 years of age
  - Maximum: The maximum age is 67
  - Median & Mean: The median and the mean age is approx. 45 years of age
  - Outliers: There are no outliers
  - Skewness: From the plot, it can be observed that the graph has a normal distribution



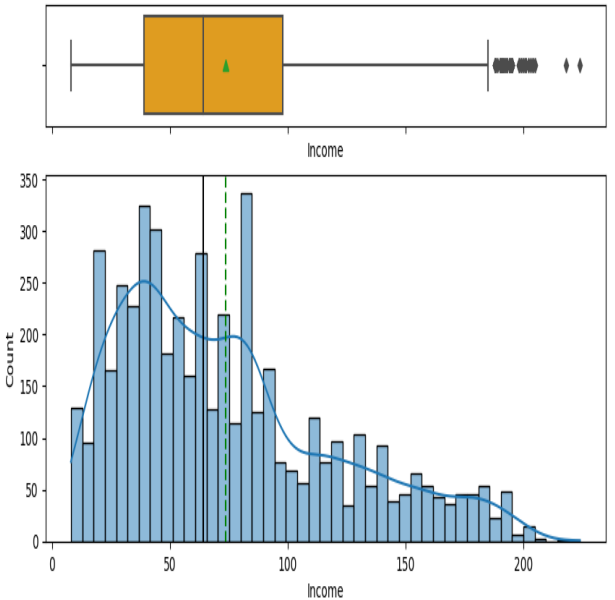
EDA - Univariate Analysis - Experience

- **Experience:**
  - Highest: 151 people have approx. 32 years of work experience
  - Minimum: The minimum work experience is 1 year
  - Q1: 25% of the population have less than 10 years of work experience
  - Q3: 75% of the population have less than 35 years of work experience
  - Maximum: The maximum work experience is 43 years
  - Median & Mean: The median and mean work experience is circa. 20 years
  - Outliers: Some negative values (-1,-2 & -3) have been observed but these are potential data entry errors. Approx. 1% (49 out of 5000) anomalous negative values (-1:32, -2:14, -3:3) are observed. These are potential data entry errors and would need to be rectified in the data pre-processing stage. Also, none of the customers related to the negative values had accepted any loans from the previous campaigns, which implies that these would not have a huge impact to the model.
  - Skewness: From the plot, it can be observed that the graph has a near to normal distribution



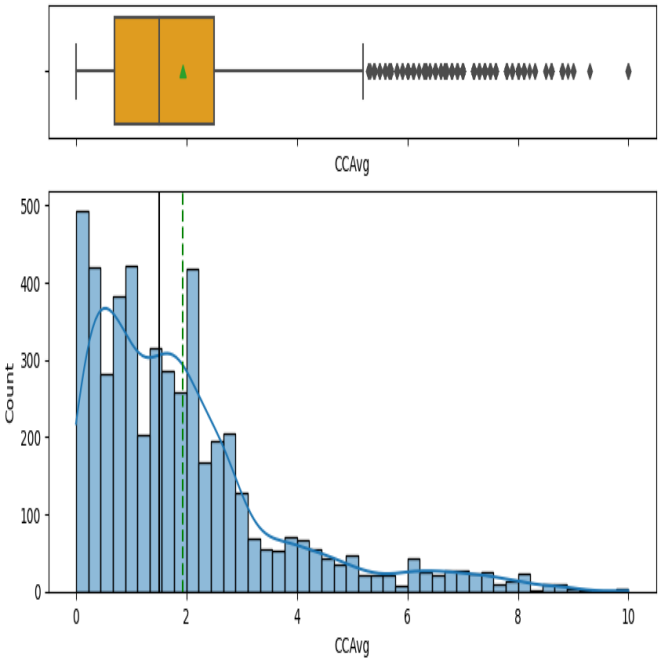
EDA - Univariate Analysis - Income

- **Income:**
  - Highest Income Group: Approx. 392 people have an income between $40K- $44K
  - Minimum: The minimum income is $8K
  - Q1: 25% of the population have circa. less than $39K of income
  - Q3: 75% of the population have circa. less than $98K of income
  - Maximum: The maximum income is $185K with a few outliers
  - Median: The median income is $64K
  - Mean: The mean income is circa. $73K
  - Outliers: There are approx. 18 outliers ranging between $188K-$224K
  - Skewness: From the plot, it can be observed that the graph right skewed



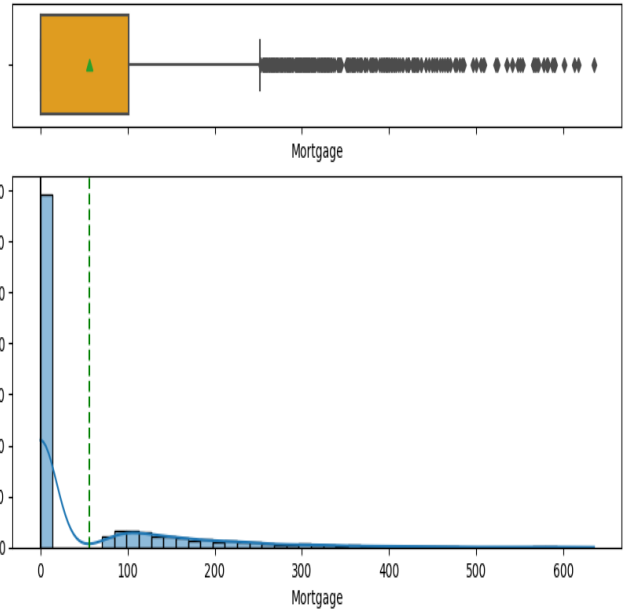
EDA - Univariate Analysis - CCAvg

- **CCAvg:**
  - Highest CCAvg: Approx. 445 people have the highest CCAvg monthly spend between $0.2K-$0.39K
  - Minimum: The minimum CCAvg monthly spend is $0.0
  - Q1: 25% of the CCAvg monthly spend is less than $0.7K
  - Q3: 75% of the CCAvg monthly spend is less than $2.5K
  - Maximum: The upper fence and maximum CCAvg monthly spend is $5.3K and $10.0K respectively
  - Median: The median CCAvg monthly spend is $1.5K
  - Mean: The mean CCAvg monthly spend is $1.93K
  - Outliers: There are approx. 40 outliers ranging between $5.3K-$10K
  - Skewness: From the plot, it can be observed that the graph highly right skewed



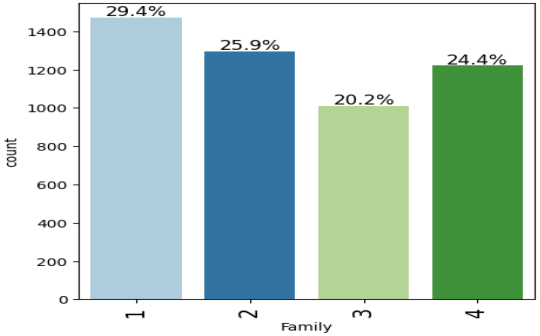
EDA - Univariate Analysis - Mortgage

- **Mortgage:**
  - Highest: Approx. 3,462 customers have $0 value of house mortgage
  - Minimum: The minimum house mortgage value is $0
  - Q1: 25% of the house mortgage value is $0
  - Q3: 75% of the house mortgage value is less than $101K
  - Maximum: The upper fence and maximum house mortgage value is $252K and $635K respectively with several outliers
  - Median: The median house mortgage value is $0
  - Mean: The mean house mortgage value is $56.49K
  - Outliers: There are several outliers ranging between $253K-$635K
  - Skewness: From the plot, it can be observed that the graph highly right skewed

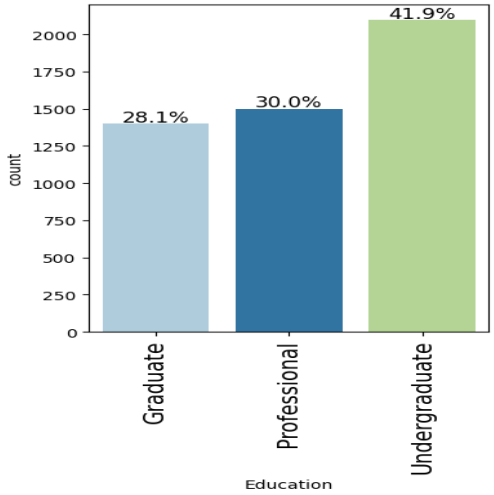


EDA - Univariate Analysis – Family & Education

- **Family:**
  - Majority of customers - approx. 29.4% (1,472 of 5000) - have only 1 person in their family, followed by 25.9% (1,296 of 5000) and 24.4% (1,222 of 5000) who have 2 & 4 people in their families, respectively
  - Lowest Family Group: 20.2% (1,010 out of 5000) have 3 people in their family

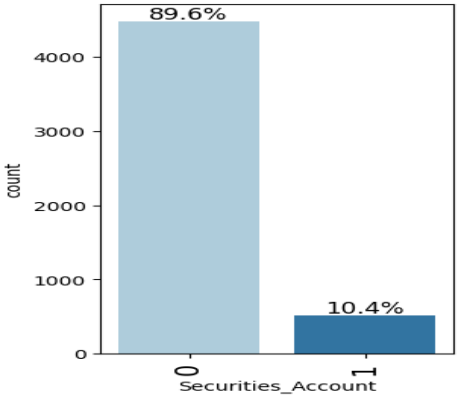


- **Education:**
  - There are higher number of undergraduates than graduates and professionals. There are approx. 41.9% (2,096 of 5000) of undergraduates, followed by 30.0% (1,501 of 5000) of professionals, and 28.1% (1,403 of 5000) of graduates

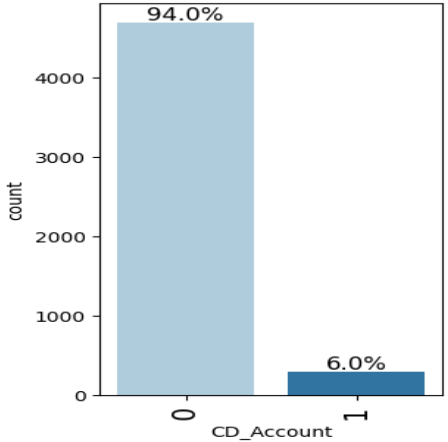


EDA - Univariate Analysis – Securities & CD Account

- **Securities Account:**
  - Majority of customers do not have a securities account with the bank. Approx. 89.6% (4,479 of 5000) of customers do not have a Securities Account with the bank vis-à-vis 10.4% (523 of 5000) of customers that do hold a securities account

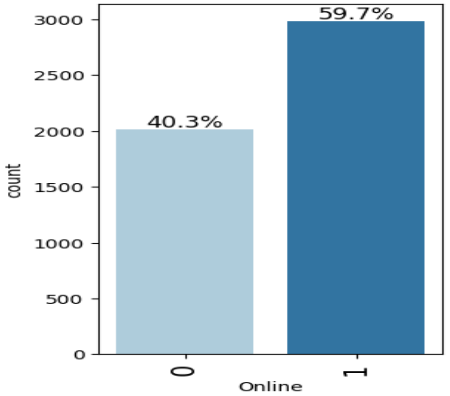


- **Certificate Of Deposit Account:**
  - Majority of customers do not have a Certificate Of Deposit account with the bank. Approx. 94% (4,698 of 5000) of customers do not have a Certificate Of Deposit account with the bank vis-à-vis 6% (302 of 5000) of customers that do hold a Certificate Of Deposit account with the bank

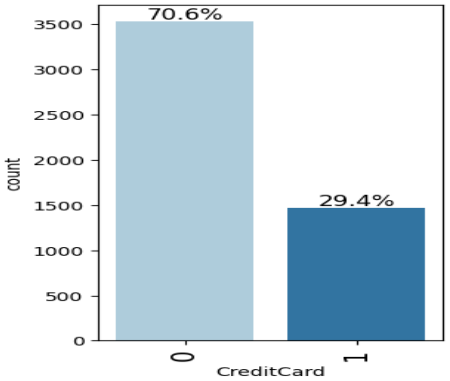


EDA - Univariate Analysis – Online & Credit Cards

- **Online:**
  - Almost over a half of customers use the internet banking facilities provided by the bank. Approx. 59.7% (2,984 of 5000) of customers do avail of internet banking provided by the bank vis-à-vis 40.3% (2,016 of 5000) of customers that do not use the internet banking provided by the bank

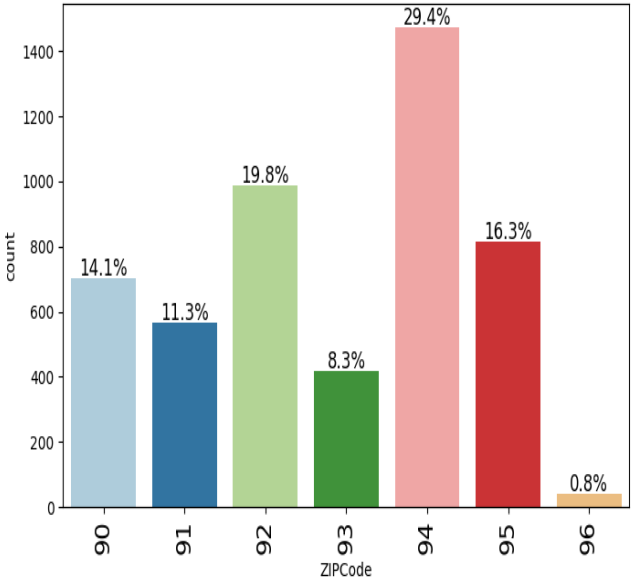


- **Credit Cards:**
  - Majority of customers do not use a Credit Card issued by any other bank. Approx. 70.6% (3,530 of 5000) of customers do not use a Credit Card issued by any other bank than compared to 29.4% (1,470 of 5000) of customers that do use a Credit Card issued by another bank



EDA - Univariate Analysis – ZIPCode

- **ZipCode:**
  - Pre-Data Processing, there were 467 unique ZIP codes out of 5000. However, for the Decision Tree to perform optimally, the post code has been classified in 7 different categories: 90, 91, 92, 93, 94 95 and 96. Of which, 29.4% of customers reside at post code 94, 19.8% of customers reside at 92, 16.3% of customers reside at 95, 14.1% of customers reside at 90, 11.3% of customers reside at 91, 8.3% of customers reside at 93, and a small fragment of customers – 0.8% reside at post code 96.



EDA - Univariate Analysis – Key Observations & Insights

- **Key Observations :**
  - The median and the mean age of customers across the dataset is approx. 45 years of age, whereas the median and mean work experience is circa. 20 years
  - The median income of customers is approx. $64K and the mean CCAvg monthly spend is $1.93K
  - Some negative values (-1,-2 & -3) have been observed within Experience but these could be potential data entry errors.
  - The minimum CCAvg monthly spend is $0.0, which could imply that some customers do not use credit cards at all.
- **Key Insights :**
  - Majority of customers - approx. 29.4% (1,472 of 5000) - have only 1 person in their family
  - There are higher number of undergraduates than graduates and professionals
  - Majority of customers, approx. 89.6% (4,479 of 5000), do not have a securities account with the bank
  - Majority of customers, approx. 94% (4,698 of 5000) , do not have a Certificate Of Deposit account with the bank
  - Almost over a half, approx. 59.7% (2,984 of 5000), of customers use the internet banking facilities provided by the bank
  - Majority of customers, approx.70.6% (3,530 of 5000), do not use a Credit Card issued by any other bank

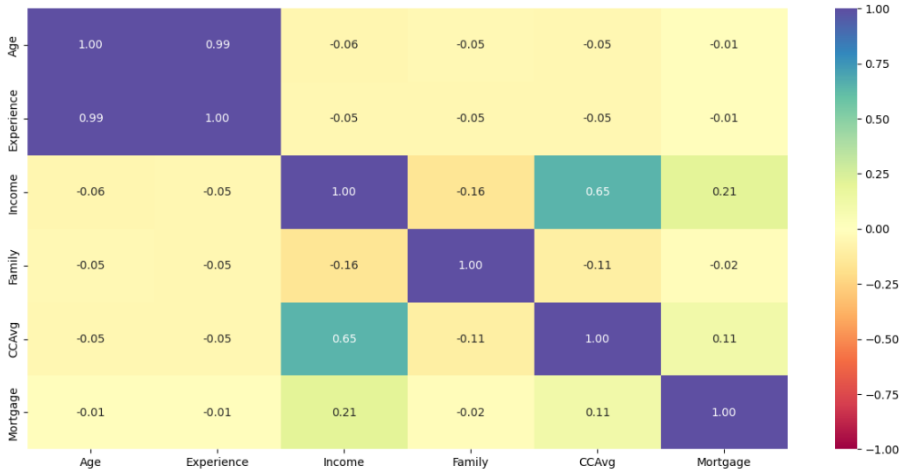
EDA - Bivariate Analysis

EDA - Bivariate Analysis – Correlation Check

- **Correlation Amongst Variables :**

  - There is significant positive correlation of 0.99 between Age and Experience. With growing Age, customers have gained significant years of experience.

  - There is an insignificant negative correlation of -0.01, -0.05, -0.05, and -0.06 between Age and Mortgage, CCVAvg, Family and Income, respectively.
    This implies that with increasing Age, other factors such as house value of the mortgage, credit card spend, size of the family dependents, and Income decreases and vice-versa.



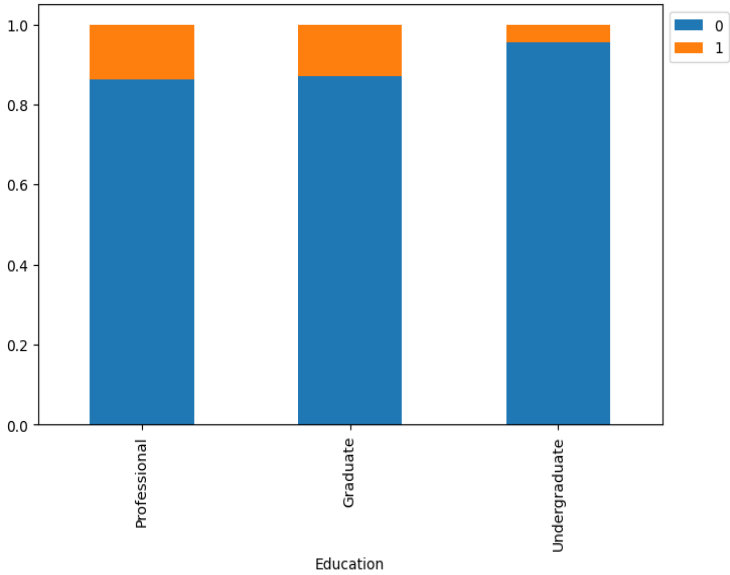
  - There is an insignificant negative correlation of -0.01, -0.05, -0.05, and -0.05 between Experience and Mortgage, CCVAvg, Family and Income respectively. This implies that with higher years of experience, which yields directly to Age, other factors such as house value of the mortgage, credit card spend, size of the family dependents, and Income decreases and vice-versa.

EDA - Bivariate Analysis – Correlation Check (Con’td)

- **Correlation Amongst Variables :**
  - There is a strong positive correlation of 0.65 between Income and CCVAvg. This implies that higher the income, greater is the monthly spend and vice-versa. There is also a weak positive correlation of 0.21 between Income and Mortgage. This implies that as income increases, the affordability towards buying a house with a higher mortage also increases and vice-versa. There is an insignificant negative correlation of -0.16, -0.05 and -0.06 between Income and Family, Experience and Age respectively.
  - There is an insignificant negative correlation of -0.02, -0.11, -0.16, -0.05 and -0.05 between Family and Mortgage, CCVAvg, Income, Experience and Age respectively.
  - There is an insignificant positive correlation of 0.11 between CCVAvg and Mortgage. Given the strong positive correlation of 0.65 between CCVAvg and Income, customers have a high degree of affordability towards monthly spending, thereby increasing the credit worthiness of customers availing house mortgages.
  - There is an insignificant negative correlation of -0.11, -0.05 and -0.05 between CCVAvg and Family, Experience and Age respectively.
  - There is an insignificant positive correlation of 0.21 between Mortgage and Income. This implies higher the income, greater is the affordability towards a house with a higher mortgage value. There is also an insignificant negative correlation of -0.02, -0.01 and -0.01 between Mortgage and Family, Experience and Age, respectively.

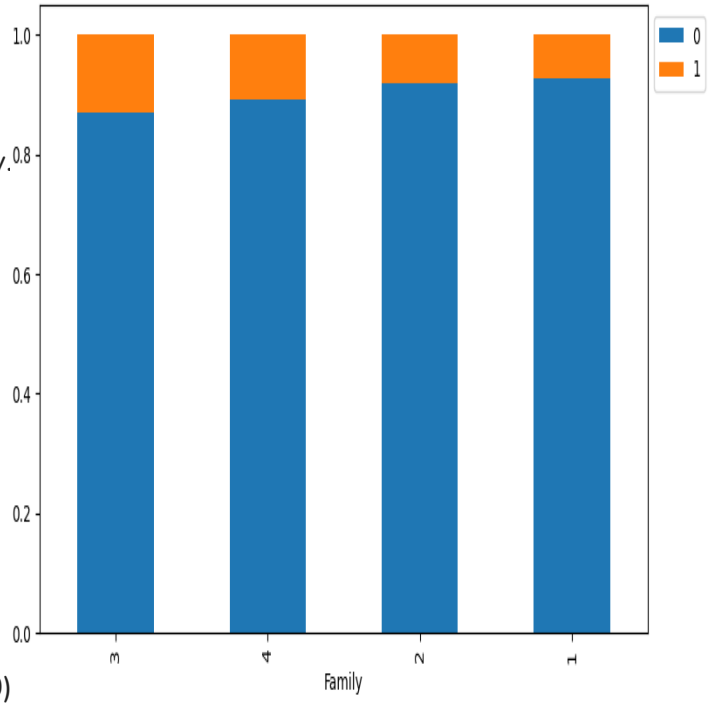
EDA - Bivariate Analysis – Education vs Personal Loan

- **Education vs Personal Loan:**
  - Professionals are more interested in availing of a personal loan, followed by Graduates and Undergraduates.
  - 13.65 % (205 of 1501) of Professionals were interested in availing of a personal loan than compared to 86.34% (1296/1501) who were not interested.
  - 12.92 % (182 of 1403) of Graduates were interested in availing of a personal loan than compared to 87.02% (1221 of 1403) who were not interested.
  - 4.43 % (93/2096) of Undergraduates were interested in availing of a personal loan than compared to 95.56% (2023 of 2096) who were not interested.



EDA - Bivariate Analysis – Personal Loan vs Family

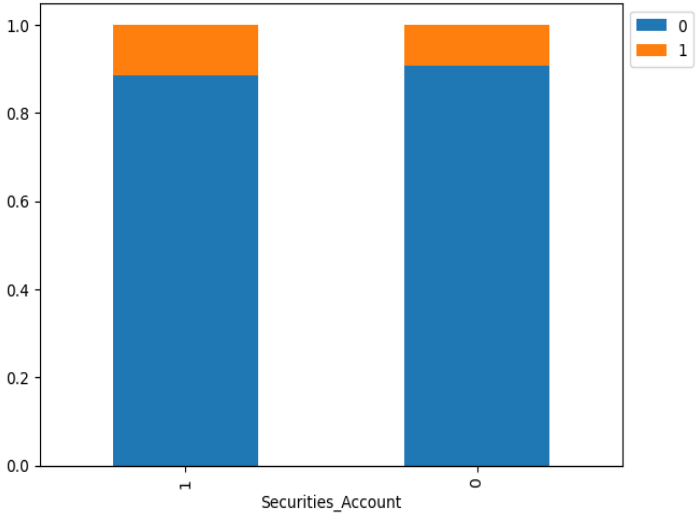
- **Personal Loan vs Family:**
  - Families with 3 members show high interests in availing a personal loan followed by families of 4, 2 and 1 member, respectively. 9.6% (480 of 5000) of Families of all sizes were interested in availing a personal loan than compared to 90.4% (4520 of 5000) that were not interested in procuring the personal loan
  - 7.26% (107/1472) of Family with 1 member were interested in availing of a personal loan than compared to 92.73% (1365/1472) who were not interested in procuring the personal loan
  - 8.17% (106 of 1296) of Family with 2 members were interested in availing of a personal loan than compared to 91.82% (1190 of 1296) who were not interested in procuring the loan
  - 13.16% (133 of 1010) of Family with 3 members were interested in availing of a personal loan than compared to 86.83% (877 of 1010)



  - 10.96% (134 of 1222) of Family with 4 members were interested in availing of a personal loan than compared to 89.03% (1088 of 1222) who were not interested in procuring the loan

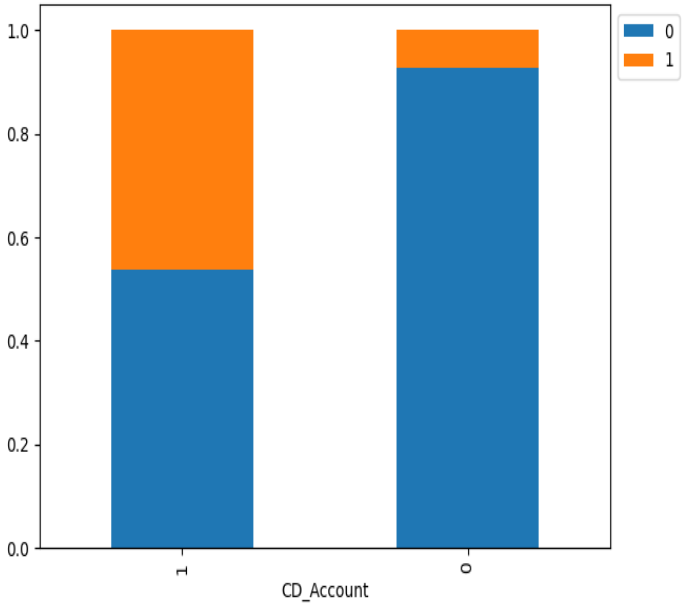
EDA - Bivariate Analysis – Personal Loan vs Securities Account

- **Personal Loan vs Securities_Account:**
  - Existing Security Account holders are more interested in availing of a personal loan than Non-Security Account holders.
  - 9.6% (480 of 5000) of combined Security & Non-Security Account holders are interested in availing a personal loan than compared to 90.4% (4520 of 5000) that are not interested in the personal loan
  - 11.49% (60 of 522) of existing Security Account holders are interested in availing a personal loan than compared to 88.50% (462 of 522) who are not interested in the personal loan
  - 9.3% (420 of 4478) who do not hold a Security Account are interested in availing a personal loan than compared to 90.62% (4058 of 4478) that are not interested in the personal loan



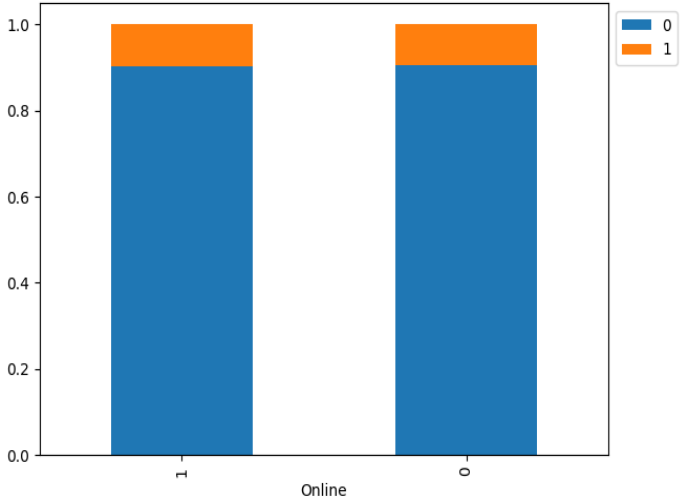
EDA - Bivariate Analysis – Personal Loan vs CD_Account

- **Personal Loan vs Certificate of Deposit (CD_Account):**
  - Existing CD Account holders are more interested in availing of a personal loan than Non-CD Account holders.
  - 9.6% (480 of 5000) of existing CD Account holders are interested in availing a personal loan than compared to 90.4% (4520 of 5000) that are not interested in the personal loan
  - 46.35% (140 of 302) of existing CD Account holders are interested in availing a personal loan than compared to 53.64% (162 of 302) of existing CD Account holders that are not interested in the personal loan
  - 7.23% (340 of 4698) who do not hold a CD Account are interested in availing a personal loan than compared to 92.76% (4358 of 4698) who do not hold a CD Account and are not interested in the personal loan



EDA - Bivariate Analysis – Personal Loan vs Online

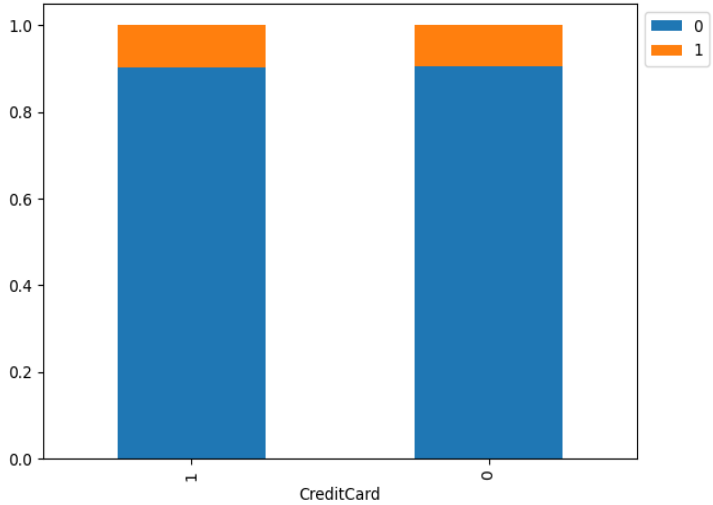
- **Personal Loan vs Online:**
  - Online customers are marginally more interested in availing of a personal loan than offline customers.
  - 9.6% (480 of 5000) of combined Online & Offline customers are interested in availing a personal loan than compared to 90.4% (4520 of 5000) that are not interested in the personal loan
  - 9.75% (291 of 2984) of online customers are interested in availing a personal loan than compared to 90.24% (291 of 2984) in the personal loan



  - 9.37% (189 of 2016) offline customers are interested in availing a personal loan than compared to 90.62% (1872 of 2016) that are not interested in the personal loan

EDA - Bivariate Analysis – Personal Loan vs Credit Card

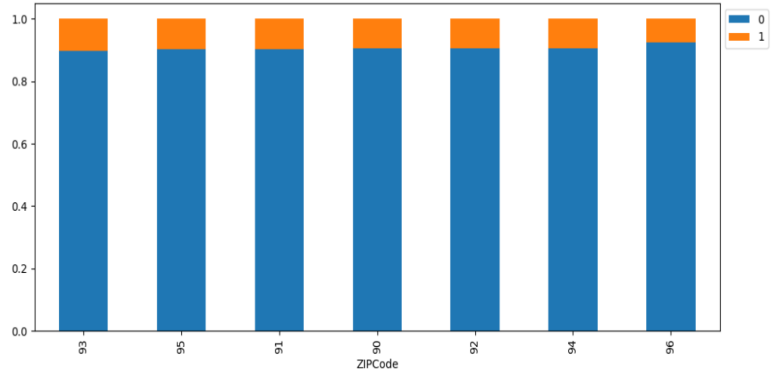
- **Personal Loan vs Credit Card:**
  - Customers having an existing Credit Card issued by another bank are marginally more interested in availing of a personal loan than those that do not have a Credit Card.
  - 9.6% (480 of 5000) customers, including the ones who have and do not have an existing credit card issued by another bank are interested in availing a personal loan than compared to 90.4% (4520 of 5000) that are not interested in the personal loan
  - 9.72% (143 of 1470) of customers having an existing credit card issued by another bank are interested in availing a personal loan than compared to 90.27% (1327 of 1470) who are not interested in procuring a personal loan



  - 9.54% (337 of 3530) of customers who do not have an existing credit card issued by another bank are interested in availing a personal loan than compared to 90.45% (3193 of 3530) who are not interested in procuring a personal loan

EDA - Bivariate Analysis – Personal Loan vs ZIP Code

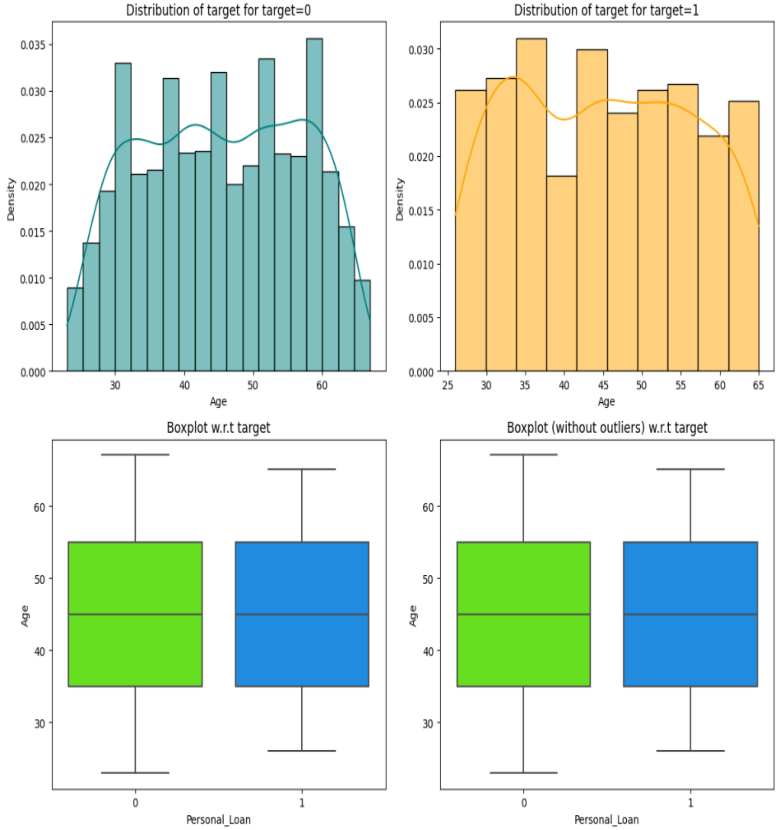
- **Personal Loan vs ZIP Code:**
  - Of the 7 ZIP Codes, customers living in 93 are more in favour of availing a personal loan, followed by 95, 91, 90, 92, 94, and 96, respectively. 10.31% (43 of 417) of customers residing at 93 are interested in availing a personal loan than compared to 89.68% (374 of 417) that are not, whereas 9.81% (80 of 815) of customers residing at 95 are interested in availing a personal loan than compared to 90.18% (735 of 815) who are not interested in the loan



  - 9.73% (55 of 565) of customers residing at 91 are interested in availing a personal loan than compared to 90.26% (510 of 565) that are not, whereas 9.53% (67 of 703) of customers residing at 90 are interested in availing a personal loan than compared to 90.46% (636 of 703) that are not interested in the loan
  - 9.51% (94 of 988) of customers residing at 92 are interested in availing a personal loan than compared to 90.48% (894 of 998) that are not, whereas 9.51% (94 of 988) of customers residing at 92 are interested in availing a personal loan than compared to 90.48% (894 of 998) that are not interested in the loan
  - 9.37% (138 of 1472) of customers residing at 94 are interested in availing a personal loan than compared to 90.62% (1334 of 1472) that are not interested in the loan, whereas 7.5% (3 of 40) of customers residing at 96 are interested in availing a personal loan than compared to 92.5% (37 of 40) that are not interested in the loan

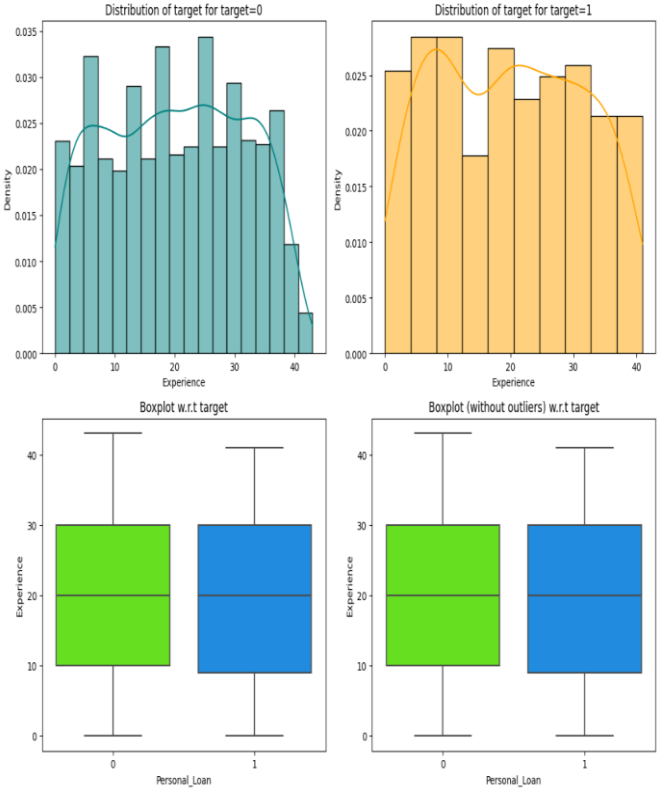
EDA - Bivariate Analysis – Personal Loan vs Age

- **Personal Loan vs Age:**
  - Minimum: Customers that are highly likely to avail of a personal loan are at least 26 years, whereas those that are likely to reject the loan are at least 23 years
  - Maximum: Customers that are highly likely to avail of a personal loan are at most 65 years, whereas those that are likely to reject the loan are at most 67 years
  - Q1: 25% of the customers that would accept or reject the loan are approx. under 35 years of age
  - Q3: 75% of the of the customers that would accept or reject the loan are approx. under 55 years of age
  - The mean age of customers interested in personal loans (45.36 years) to those that are not interested in personal loans (45.07 years) are very similar
  - The median age of customers interested in personal loans and those that are not interested is the same (45 years)



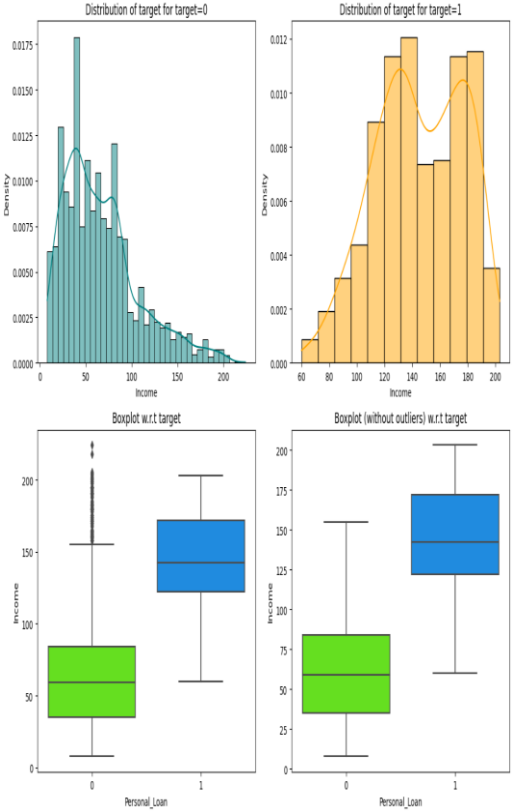
EDA - Bivariate Analysis – Personal Loan vs Experience

- **Personal Loan vs Experience:**
  - Minimum: Customers that are highly likely to avail or reject of a personal loan have at least 1-year experience.
  - Maximum: Customers that are highly likely to avail of a personal loan have at most 41 years of experience, whereas those that are likely to reject the loan have at most 43 years of experience
  - Q1: 25% of the customers that would accept the loan have under 9 years of experience, whereas those that would reject the loan have approx. under 10 years of experience
  - Q3: 75% of the customers that would accept or reject the loan have under 30 years of experience
  - The mean number of years of experience for customers interested in personal loans (19.84 years) to those that are not interested in personal loans (2013 years) is very similar
  - The median number of years of experience for customers interested in personal loans and those that are not interested is the same (20 years)



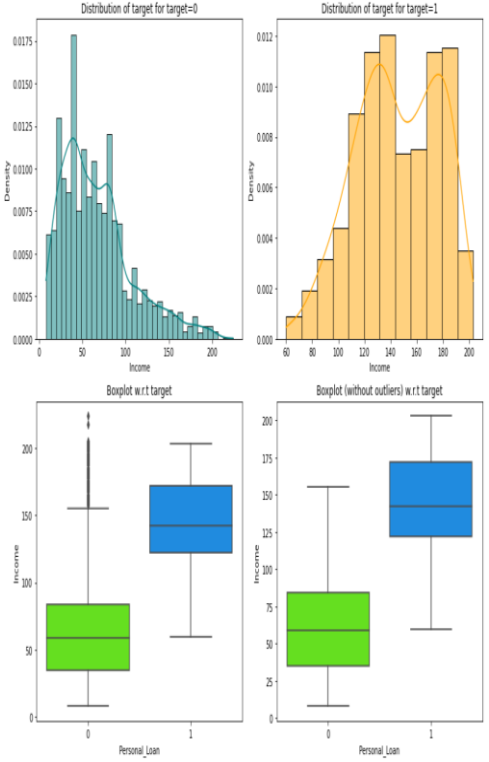
EDA - Bivariate Analysis – Personal Loan vs Income

- **Personal Loan vs Income:**
  - Customers that accepted the personal loan had a high income than those who did not accept the personal loan
  - Minimum: Customers that accepted the personal loan had an income of under $60K, that those that rejected the loan, who had an income of under $8K
  - Maximum: Customers that accepted the personal loan had an income of under $203K, that those that rejected the loan, who had an income of under $224K
  - Q1: 25% of the customers that accepted the personal loan had an income of under $122K, that those that rejected the loan, who had an income of under $35K
  - Q3: 75% of the customers that accepted the personal loan had an income of under $172K, that those that rejected the loan, who had an income of under $84K
  - The mean income for customers who accepted the personal loan is $144.75K to those that did not accept the personal loan is $66.24K.
  - The median income for customers who accepted the personal loan is $142.5K to those that did not accept the personal loan is $59K



EDA - Bivariate Analysis – Personal Loan vs CCAvg

- **Personal Loan vs CCAvg:**
  - Customers that accepted the personal loan had a high monthly CCAvg spend than those who did not accept the personal loan
  - Minimum: Customers that accepted the personal loan had a CCAvg of under $2.6K, compared to those that rejected the loan with a CCAvg of under $0.6K
  - Maximum: Customers that accepted the personal loan had a CCAvg of under $9.3K, compared to those that rejected the loan, with a CCAvg of under $8.8K
  - Q1: 25% of the customers that accepted the personal loan had a CCAvg of under $2.6K, compared to those that rejected the loan with a CCAvg of under $0.6K
  - Q3: 75% of the customers that accepted the personal loan had a CCAvg of under $5.3K, compared to those that rejected the loan with a CCAvg of under $2.3K
  - The mean CCAvg for customers who accepted the personal loan is $3.91K to those that did not accept the personal loan is $1.73K
  - The median CCAvg for customers who accepted the personal loan is $3.8K to those that did not accept the personal loan is $1.4K



EDA - Bivariate Analysis – Key Observations & Insights

- **Key Observations :**
  - There is significant positive correlation of 0.99 between Age and Experience. With growing Age, customers have gained significant years of experience
  - There is an insignificant negative correlation of -0.01, -0.05, -0.05, and -0.06 between Age and Mortgage, CCVAvg, Family and Income, respectively. This implies that with increasing Age, other factors such as house value of the mortgage, credit card spend, size of the family dependents, and Income decreases and vice-versa.
  - There is an insignificant negative correlation of -0.01, -0.05, -0.05, and -0.05 between Experience and Mortgage, CCVAvg, Family and Income respectively. This implies that with higher years of experience, which yields directly to Age, other factors such as house value of the mortgage, credit card spend, size of the family dependents, and Income decreases and vice-versa.
  - There is a strong positive correlation of 0.65 between Income and CCVAvg. This implies that higher the income, greater is the monthly spend and vice-versa. There is also a weak positive correlation of 0.21 between Income and Mortgage. This implies that as income increases, the affordability towards buying a house with a higher mortage also increases and vice-versa. There is an insignificant negative correlation of -0.16, -0.05 and -0.06 between Income and Family, Experience and Age respectively.
  - There is an insignificant positive correlation of 0.11 between CCVAvg and Mortgage. Given the strong positive correlation of 0.65 between CCVAvg and Income, customers have a high degree of affordability towards monthly spending, thereby increasing the credit worthiness of customers availing house mortgages.

EDA - Bivariate Analysis – Key Observations & Insights (Cont'd)

- **Key Observations :**
  - There is an insignificant positive correlation of 0.21 between Mortgage and Income. This implies higher the income, greater is the affordability towards a house with a higher mortgage value. There is also an insignificant negative correlation of -0.02, -0.01 and -0.01 between Mortgage and Family, Experience and Age, respectively.
- **Key Insights :**
  - High concentration of customers who accepted a personal loan are observed at:
    - Higher Income level (approx. above $98.5K)
    - Higher CCAvg (approx. above $3K and above)
    - Existing customers with Certificate Of Deposit Accounts
    - Existing customers with Securities Accounts
    - Customers who use credict cards issued from other banks
    - Customers with Education of Graduate (2) or Professional (3)
  - Families with size of 3 and 4 members show high interests in availing a personal loan.

Model Building

Model Building – Evaluation Criteria

- **Model Evaluation Criteria :**
  - The primary objective for building the model is to predict whether a liability customer will buy personal loan. Following are the 4 scenarios that will have an impact on the model's prediction:
    - True Positive (TP): The model predicts the customer **will buy** the loan and the customer **buys** the loan
    - True Negative (TN): The model predicts the customer **will not buy** the loan and the customer **does not buy** the loan
    - False Positive (FP): The model predicts the customer **will buy** the loan, but the customer **does no**t buy the loan
    - False Negative (FN): The model predicts the customer **will not buy** the loan, but the customer **would have bought** the loan
  - Key Business Outcomes: The following outcomes would be the key to focus from the business perspective
    - False Positive (FP): Predicting that a customer will buy the personal loan but eventually doesn't buy the loan would lead to massive loss of resources (FP). On the other hand, predicting that a customer will not buy the personal loan, but eventually would have bought the personal loan (FN), would be a significant loss of opportunity and thereby loss of revenue. Besides the model will be used by Marketing to launch a campaign, which needs to have the maximum reach. So, if FN is high, that means we will be losing out on reaching potential customers. This implies that **reducing False Negatives** should be of utmost importance to the business.
  - Key Evaluation Criteria:
    - Recall: The bank should use Recall as the key model evaluation criteria – higher the Recall, greater are the chances of minimising False Negatives

Model Building – Evaluation Criteria ('Contd)

- **Building the Model:**
  - We have taken the following approaches towards building our model and evaluating its performance, such that it satisfies our key criteria - Increase the Recall to significantly minimise False Negatives.
    - Model 1: Initial Model
    - Model 2: Pre-Pruned Model using Hyper-Parameter Tuning
    - Model 3: Post Pruned Model using Cost Complexity
  - We have built our model using the DecisionTreeClassifier function
  - We have used the default 'gini' criteria to split within our model
  - Based on the dataset, it is observed that we have approx., 10% of positive classes, which means if our model marks each sample as negative, then we would at least 90% accuracy. This implies that 'Accuracy' would not be a good metric to evaluate the performance of our model. This re-affirms our hypothesis that we should 'Recall' as the effective measure of performance for the above 3 models

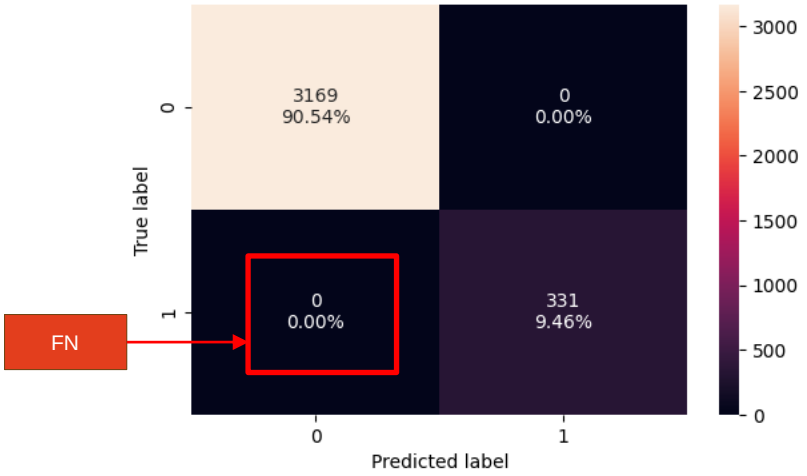
```
Percentage of classes in training set:
0    0.905429
1    0.094571
Name: Personal_Loan, dtype: float64
Number of rows in Y-Training set :  (3500,)

Percentage of classes in test set:
0    0.900667
1    0.099333
Name: Personal_Loan, dtype: float64
Number of rows in Y-Testing set :  (1500,)
```

Model Building – Model 1: Initial Model

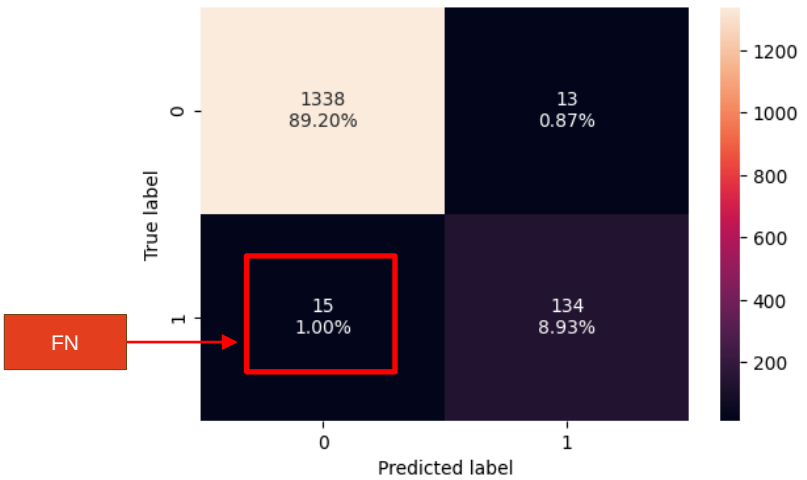
- **Model 1 – Initial Model:**
  - Recall on Training Dataset is 1 and on Testing Dataset is 0.89, which means the model is overfitting

Training Dataset



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0

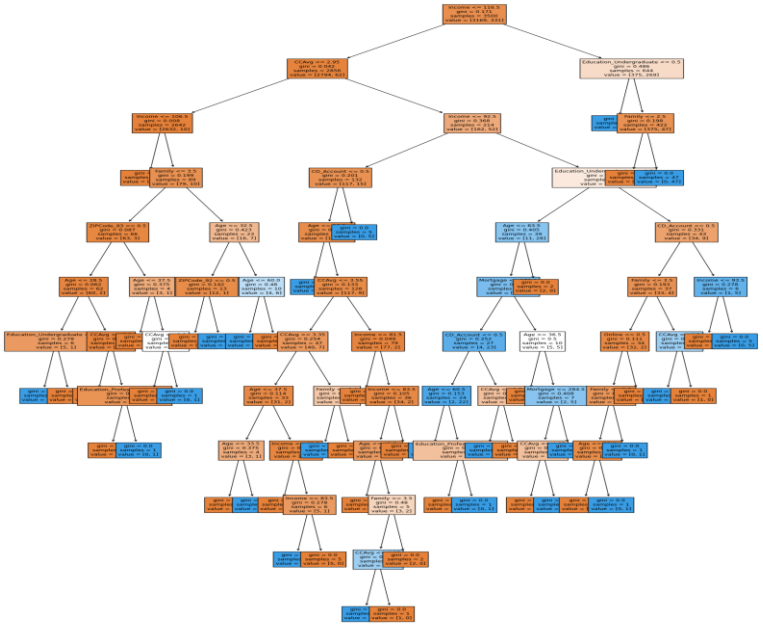
Testing Dataset



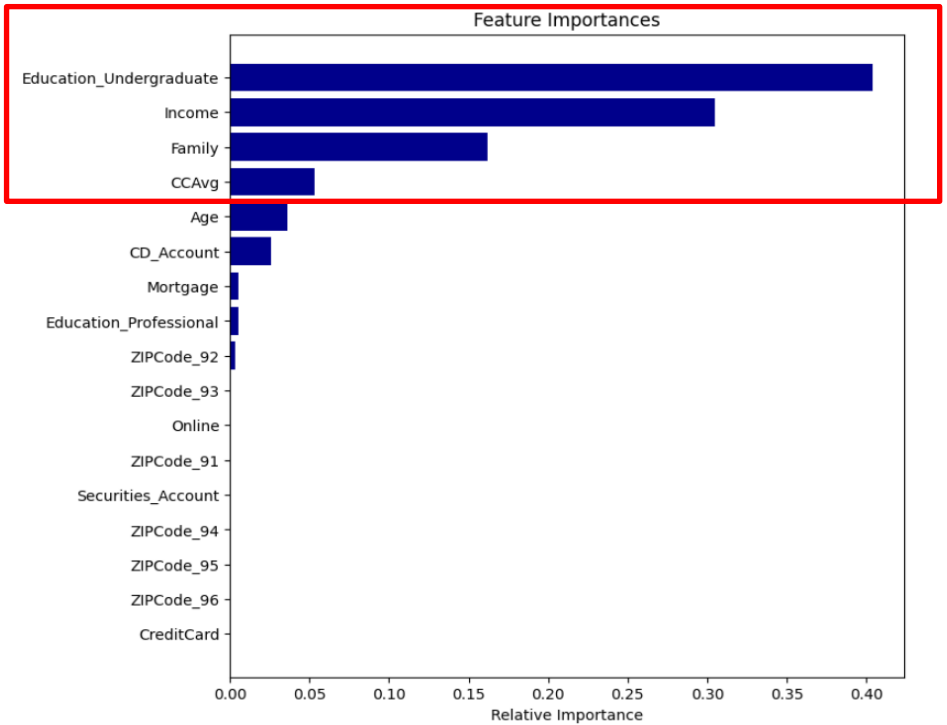
	Accuracy	Recall	Precision	F1
0	0.981333	0.899329	0.911565	0.905405

Model Building – Model 1: Decision Tree & Important Features

- **Model 1 - Initial Model Decision Tree**



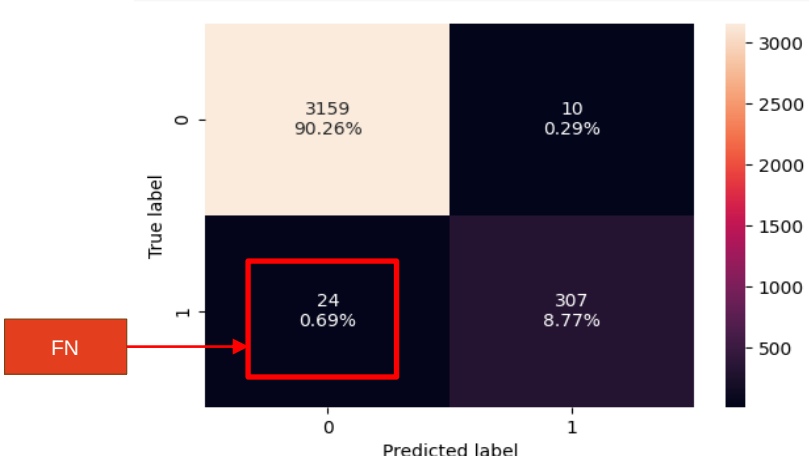
- **Most Important Features -** Education_Undergraduate, Income, Family and CCAvg at 40.3%, 30.4%, 16.1% & 5.3% respectively



Model Building – Model 2: Pre-Pruned (Hyper Parameter)

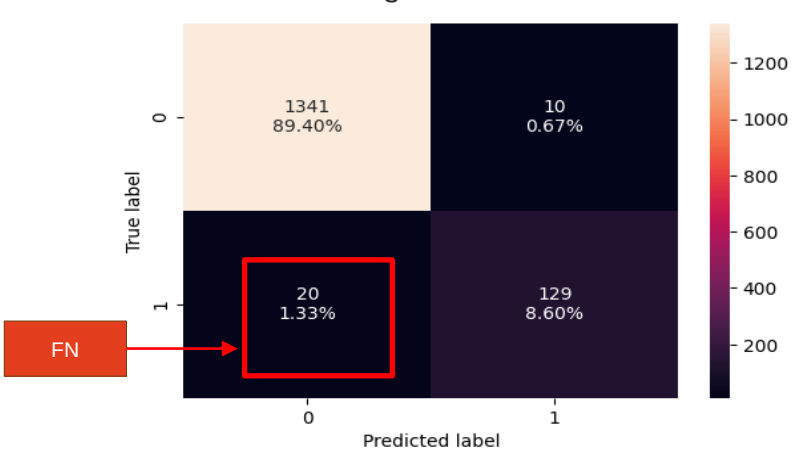
- **Model 2 – Pre-Pruned Model :**
  - We have built this model using GridSearchCV for hyper-parameter tuning.
  - Recall on Training Dataset is 0.92 and on Testing Dataset is 0.86, which means the model is not overfitting

Training Dataset



	Accuracy	Recall	Precision	F1
0	0.990286	0.927492	0.968454	0.947531

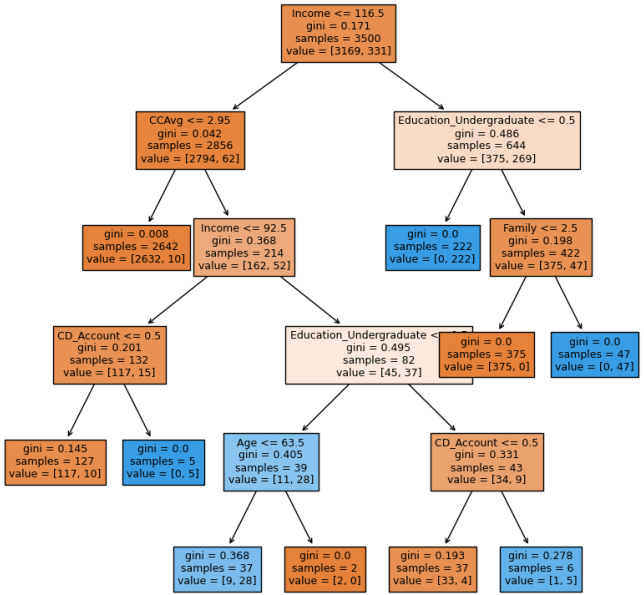
Testing Dataset



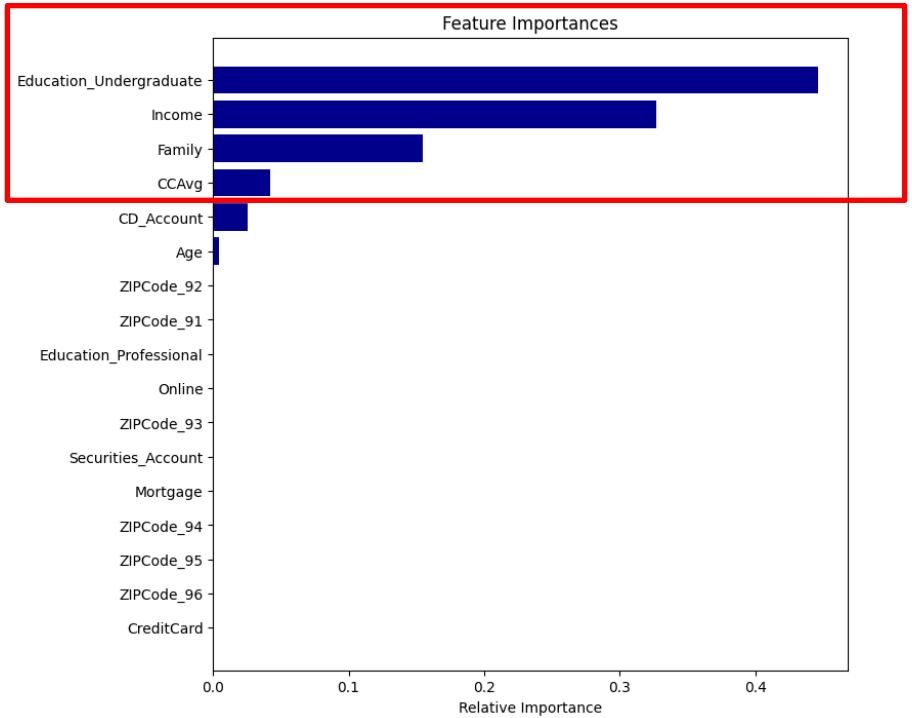
	Accuracy	Recall	Precision	F1
0	0.98	0.865772	0.928058	0.895833

Model 2: Pre-Pruning - Decision Tree & Important Features

- **Model 2 – Pre-Pruned Decision Tree**



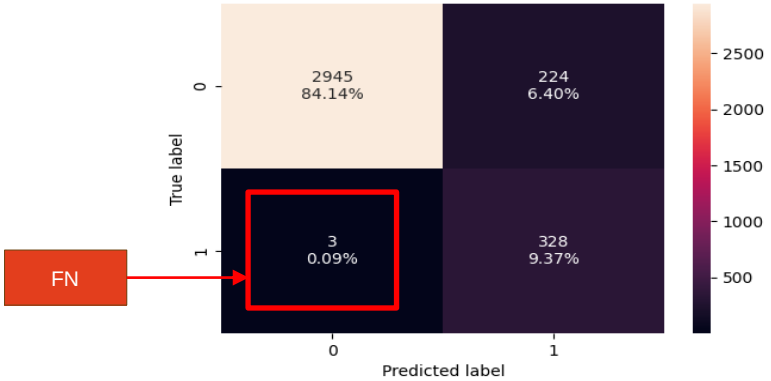
- **Most Important Features -** Education_Undergraduate, Income, Family and CCAvg at 44.6%, 32.7%, 15.5% & 4.2% respectively



Model Building – Model 3: Post-Pruned (Cost Complexity)

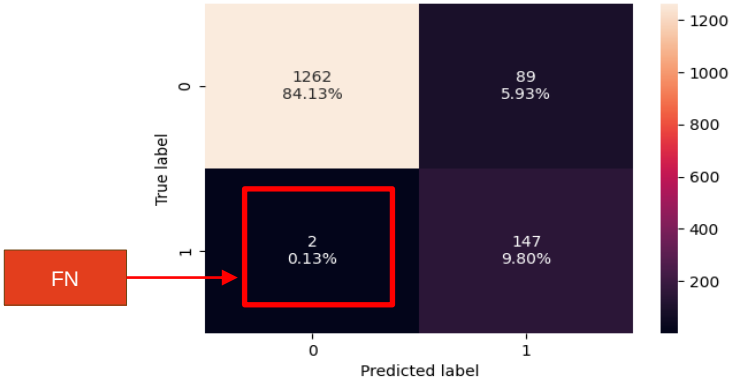
- **Model 2 – Post-Pruned Model (Cost Complexity) :**
  - We have built this model using DecisionTreeClassifier and set Cost Complexity Parameter Alpha to 0.010
  - Recall on Training Dataset is 0.99 and on Testing Dataset is 0.98, which means the model is fitting optimally

Training Dataset



	Accuracy	Recall	Precision	F1
0	0.935143	0.990937	0.594203	0.742922

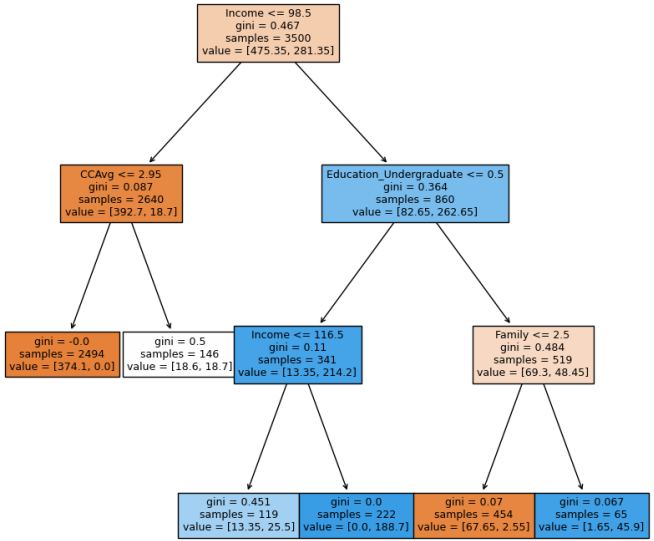
Testing Dataset



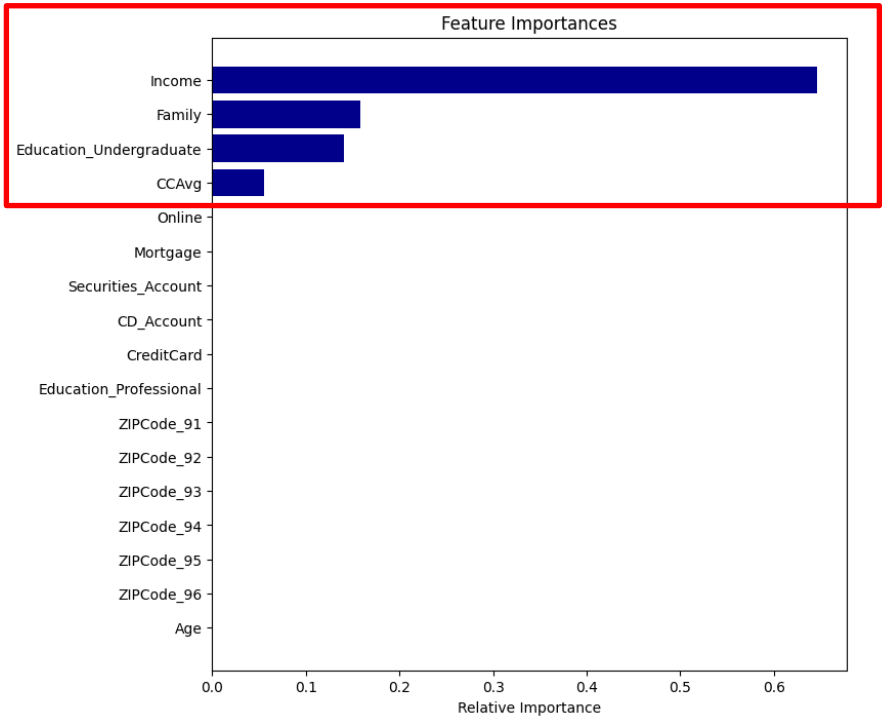
	Accuracy	Recall	Precision	F1
0	0.939333	0.986577	0.622881	0.763636

Model 3: Post-Pruning - Decision Tree & Important Features

- **Model 3 – Post-Pruned Decision Tree**



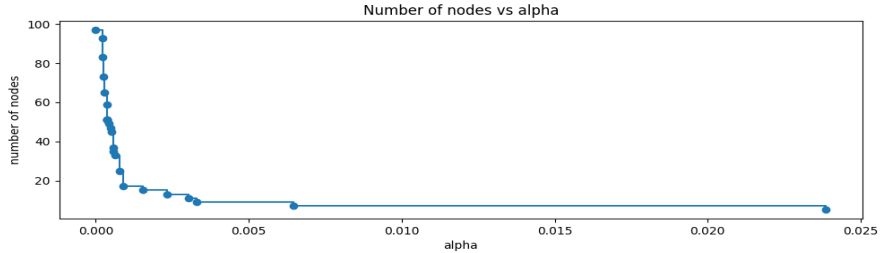
- **Most Important Features** – Income, Family, Education_Undergraduate, and CCAvg at 64.5%, 15.8%, 14.09%, and 5.5% respectively

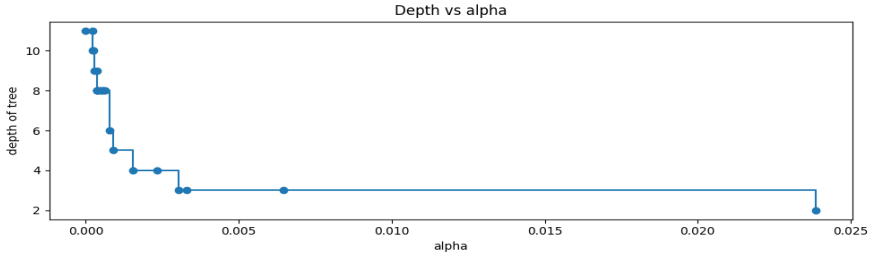


Model 3: Post-Pruned (Cost Complexity)

- **Recall vs Alpha :**
  - Recall on Training and Testing Dataset is best when the value of Alpha lies between 0.004 and 0.015
  - In order to achieve optimal results i.e., retain information on the tree and get a high recall, we have chosen alpha to be 0.010, which yields a Recall of 0.98 on the Testing Dataset, thereby minimizing the False Negatives.

- **Tree Depth & Nodes vs Alpha :**
  - As Alpha increases, more nodes are pruned thereby decreasing the depth of the tree and vice-versa.
  - With an effective Alpha of 0.010, the depth of the tree is 3 and the total nodes are near to 11





Model Performance Summary – All Models

- **Model 3 - Best Fit Model:**
  - Model 3 – Post Pruned (Cost Complexity Alpha) seems to be the best fit model for Personal Loan Campaign since the Recall for Training and Testing Dataset is 0.99 and 0.98, respectively. The model is built using the DecisionTreeClassifier and the ccp_alpha parameter is set to 0.010 to achieve best results. Important features of the model are Income, Family, Education_Undergraduate, and CCAvg at 64.5%, 15.8%, 14.09%, and 5.5% respectively. With an effective Alpha of 0.010, the depth of the tree is 3 and the total nodes are near to 11. The model is easy to interpret and doesn't overfit the Training Dataset

- **Model 1 & Model 2:** The Pre-Pruned and the Post-Pruned models have reduced overfitting and the model is giving a generalized performance.

- **Performance Summary:** The table below summarizes the performance of all 3 models for their Accuracy, Recall, Precision and F1 scores on both – Training & Testing Datasets

#	Model Type	Accuracy		Recall		Precision		F1	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
1	Model – 1 Initial Model	1.0	0.981333	1.0	0.899329	1.0	0.911565	1.0	0.905405
2	Model 2 Pre-Pruned (Hyper Parameter)	0.990286	0.98	0.927492	0.865772	0.968454	0.928058	0.947531	0.895833
3	Model 3 Post Pruned (Cost Complexity Alpha)	0.994571	0.978667	0.990937	0.986577	0.945714	0.897959	0.9721	0.891892

Conclusion

- Model 3 – Post Pruned (Cost Complexity Alpha) seems to be the best fit model for Personal Loan Campaign.
- With a high Recall score of 0.99 & 0.98, on the Training & Testing dataset, the model will minimise False Negatives, which is of utmost importance to the business.
- The model built can be used to predict whether a liability customer will buy personal loans or not. This will help the marketing department to identify the potential customers who have a higher probability of purchasing the loan.

Key Actionable Business Insights

- **Key Business Insights :**
  - Customer attributes such as Income, Family, Education, and Credit Card spend are the most important features in predicting potential customers for a personal loan
  - Income seems to be the most important feature at 64.5%, followed by Family – 15.8%, Education - 14.09%, and Average Monthly Credit Card Spend – 5.5%
  - Customers with an annual income of less than $98.5K are less likely to have a personal loan, thus it is recommended to target potential customers in this segment.
  - Customers with an income greater than $98.5K and with a higher level of education (Graduate & Professionals) are most likely already having a personal loan. It is recommended to cross-sell other products of the bank
  - Customers with a growing family are more likely to avail of a personal loan. It is recommended to target customers with a family size of 3 and / or 4 family members
  - Customers using Credit Cards frequently are deemed highly credit-worthy and as such are potential buyers of personal loan.
  - Banks' existing customers holding Securities Accounts and Certificate Of Deposit Accounts are more likely to buy a personal loan and are a likely target for conversion
  - Customers using the online facilities are more likely to already have a personal loan

Our Recommendation

Based on our key observations and insights, we recommend the following areas of improvement / opportunities that will drive business growth and lead to a better customer experience

- **Implement a Customer Centric Digital Channel to increase customer footprint:** Given that, customers who used the online facilities already have a loan, building a holistic user centric digital channel (Mobile App & Website with FAQs) which simplifies the existing loan to value application process from a customer perspective is likely to attract more prospects, thereby increasing the likelihood of selling loans to potential customers.
- **Implement Customer Incentivisation Scheme to cross sell products:** Incentivising existing customers (11.49% of Security Account Holders and 46.35% Certificate Of Deposit Account of the Bank) by offering them special interest rates / rebates on personal loans will drive customer gowth and increase revenue.

APPENDIX

Appendix - Notes

- Further analysis would be required on a comprehensive dataset to provide customer segmentation strategies

Great Learning

Happy Learning !