All Life Bank ML Project

(PGP_AIML_BA_UTA_JAN24_B)

April 20.2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview
- EDA Results
- Correlation Check
- Data Preprocessing
- Model Building
- Model Performance
- Actionable Insights & Business Recommendations
- Appendix

Executive Summary

Supervised Machine Learning models recommends to implement income-centric marketing strategies, focusing on high-income individuals, and integrating secondary indicators like education and family size can enhance marketing precision. Broadening market insights to include demographic factors facilitates personalized approaches, particularly for lower-income segments, optimizing customer engagement and driving profitability.

By implementing below recommendations, All Life bank can enhance its marketing efficacy, optimize customer engagement, and drive higher conversion rates, thereby fostering sustainable growth and profitability.

- **Marketing Strategy:**
  - Prioritize marketing efforts towards customers with incomes exceeding $100K, recognizing income as a pivotal predictor of loan product interest.
  - Intensify marketing endeavors aimed at individuals with both high incomes and advanced educational qualifications. Develop premium loan offerings aligning with the financial sophistication of this demographic to enhance engagement and conversion rates.
  - Augment segmentation and targeting strategies by considering secondary factors such as education level, family size, and credit card spending habits while income remains primary.
  - Broaden market insights to encompass factors like age, CD account holdings, and regional demographics indicated by ZIP codes. Leveraging these insights will facilitate nuanced segmentation and personalized marketing approaches.

Executive Summary

- **Promotions and Discounts:**
  - Implement promotions or discounts for depositors who purchase personal loans, incentivizing loan uptake and deposit activity synergy.
  - Tailor marketing strategies to cater to customers with incomes below $100K, offering specialized promotions, discounts or financial products to address their unique needs and circumstances.
  - Offer interest rate discounts on personal loan purchases exclusively for depositors, enhancing customer retention and loan product attractiveness.
- **Minimizing Missed Opportunities:**
  - With the primary campaign goal being customer base expansion, minimizing missed opportunities is paramount. The focus should be on identifying and targeting every prospective client, with priority given to cases overlooked in previous campaigns.
  - Utilize Machine learning models (Decision Tree) which has been developed, performance tuned and trained on existing bank customer data to minimize missed opportunities and perform correct predictions where marketing efforts can be efficiently targeted for conversion.

Business Problem Overview

- AllLife Bank recognizes a significant reservoir of potential loan customers within its existing depositor base.
- The bank aims to implement data-driven decision-making models to refine loan product offerings and maximize ROI on marketing campaigns.
- Targeted marketing strategies will prioritize high-potential customers, thereby increasing overall conversion rates and driving loan sales.
- Despite a previous marketing campaign's success with a 9% conversion rate for depositors to loan customers, AllLife Bank faces challenges in customer conversion.
- To address these challenges and elevate the loan conversion rate, the bank seeks to develop a model capable of identifying depositors with a high probability of transitioning into loan customers.

Solution Approach

- **Data Collection & Preparation:** Gather and clean customer data.
- **Feature Selection:** Identify important features for predicting loan conversion.
- **Model Development:** Select and develop a machine learning model (Decision Tree).
- **Model Training:**Train the model using existing customers data.
- **Model Evaluation:** Evaluate model performance using metrics.
- **Customer Segmentation:** Use the model to segment customers into high-potential groups.
- **Campaign Strategy:** Develop targeted marketing strategies based on customer segments.
- **Campaign Execution & Monitoring:** Implement and monitor the marketing campaign.
- **Iterative Improvement:**Continuously refine the model and strategies based on results.
- **ROI Assessment:**Measure the effectiveness of the campaign in terms of ROI.

Data Dictionary

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

Data Overview

- **Checking Dataset** - AllLife Bank loan dataset has 5000 rows & 14 columns, indicating there are 13 features and 1 target
- **Data Types** - There are a total of 14 columns in the dataset, with 13 columns being of data type Int64 and only the CCAvg column being of data type Float. There are no columns with the object data type.
- **Checking for Null Values -** There are no null values in the dataset.
- **Data Duplicates** - There are no duplicate values in the dataset.

Data Overview

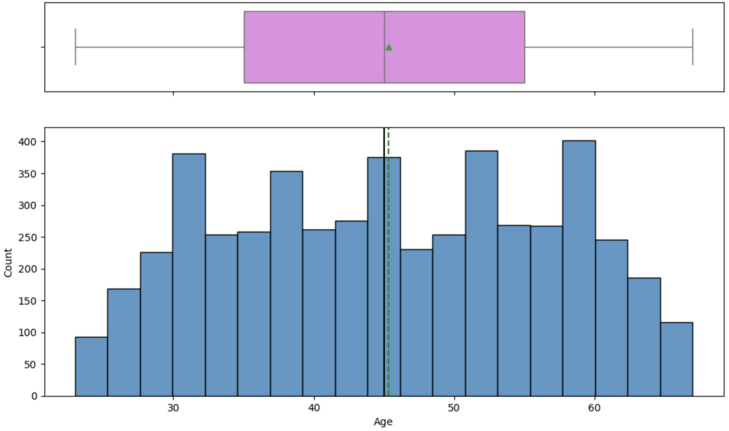
Statistical Summary:

- **ID** - Ranges from 1 to 5000, with a mean of 2500.5, suggesting it's a simple sequential identifier.
- **Age** - Average customer age is 45. Customer Age range is from 23 - 67 yrs.
- **Experience** - Average customer's experience is 20 years. Maximum customer experience is 43 years. Minimum Experience value is negative (-3 may indicate data issues) so need further analysis.
- **Income** - Average customer income is 73.77 K. There is huge difference between min income(8k) & Max income (224k), it means outliers needs to be treated. Median income is 64 K.
- **Zip Code** -Highest number of customers have zip code starting with 94, means majority of customers are from same area, rest of the data points are scattered. Very less customers are from zip code starting with 96.
- **Family** - Family sizes range from 1 to 4. Average customer's family size is 2 members. Maximum family size is 4.
- **CCAvg** - Average customer credit card spending is 1.93k,There are few customers who doesn't use credit cards for spending money. Max CCAvg is 10K.
- **Education** - Education levels range from 1 to 3, representing undergraduate to advanced/professional degrees. Majority customers are Graduates or Undergraduates.
- **Mortgage** - Very few customers have taken mortgage. More than 50% percentile customers didn’t opt for mortgage.
- **Personal_loan** - About 9.6% of customers accepted the personal loan offered in the last campaign (1 = Yes, 0 = No).
- **Securities_Account**- More than 75% customers doesn’t have securities account, only 10.4% have a securities account
- **CD_Account** - More than 75% customers doesn’t have CD accounts, only 6.04% have CD accounts.
- **Online** - More than 59% customers use internet banking.
- **Credit card** - 29.4% customers have other bank credit cards (1 = Yes, 0 = No).

EDA Results - Univariate Analysis

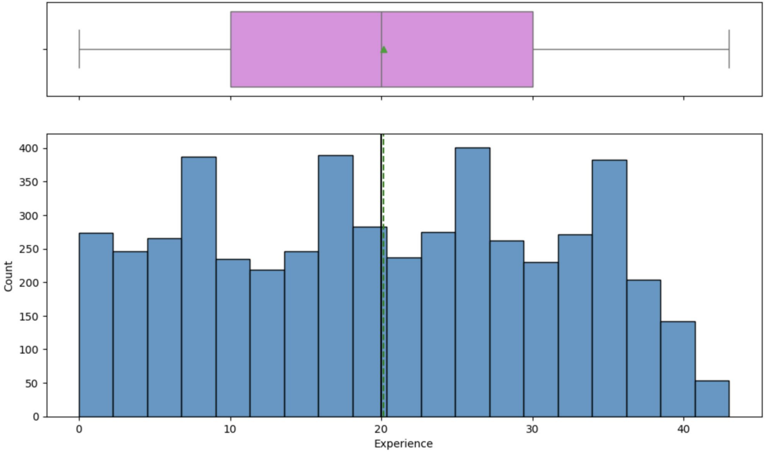
Age:

- 50% customers are in the age range of 35 -55 yrs.
- 25% customers are younger than 35 & 25% customers are 55 & older.
- Median age is 45 yrs.
- Minimum age of customer is 23 yrs and maximum age of customer is 67 yrs.



Experience:

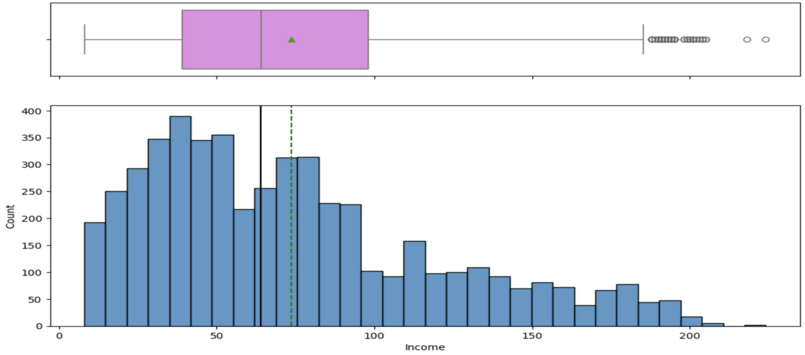
- The minimum **Experience** is showing as **-3** years which does not seems to be right and need correction
- Average experience is 20 yrs.
- Very few customers are high experienced compared to rest data points.
- Maximum experience is 43 yrs.



EDA Results - Univariate Analysis

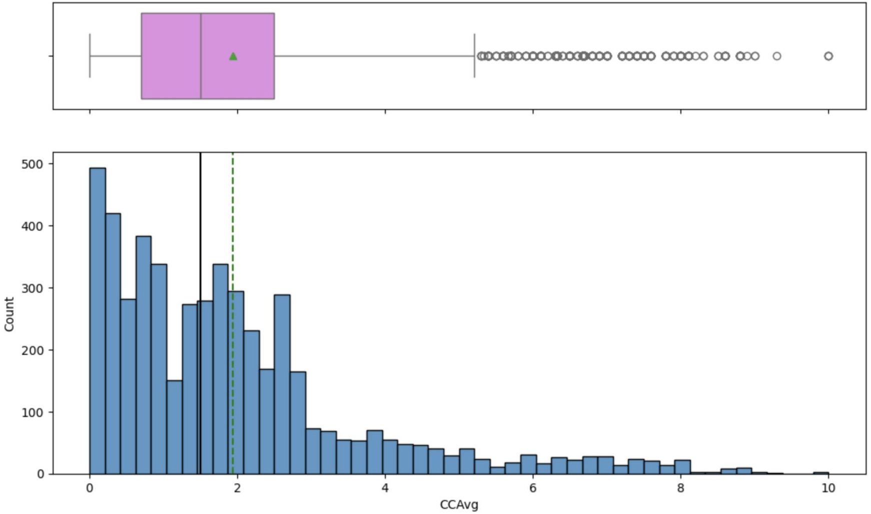
Income:

- The average income is approx $73.77K
- The standard deviation is about $46.03k, indicating a wide variation in income levels.
- Income ranges from a minimum of $8k to a maximum of $224k.
- The median income (50th percentile) is $64k, suggesting that half of the customers earn less than this amount and half earn more.
- The first quartile (25th percentile) is at $39k, and the third quartile (75th percentile) is at $98k, highlighting that the middle 50% of customers' incomes fall within this range



CCAvg:

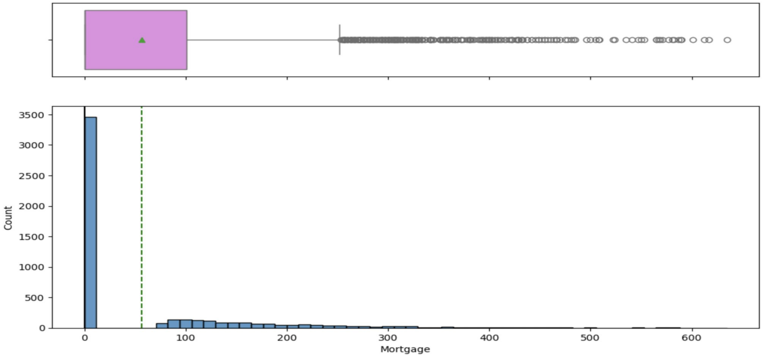
- CCAvg data is right skewed.
- The median spending (50th percentile) is $1.5k, meaning half of the customers spend less than this amount and half spend more.
- Monthly spending ranges from a minimum of $0 to $10k
- Average Credit card spending is approx $1.94K.



EDA Results - Univariate Analysis

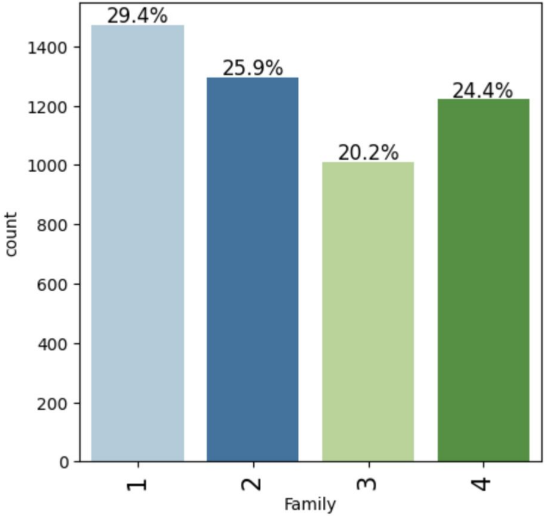
Mortgage:

- Mortgage column is highly right skewed.
- The average mortgage value is approximately `$56.5k`
- Mortgage values range from `$0` to `$635k`.
- There are so many outliers. Only One data point shows maximum value compared with rest of the data points.



Family:

- Average family size of customers is 3 members which is 20%. Majority of customers have one member /family.
- The largest category of the family column is 1, with a percentage of 29.44%.
- The second largest category is a family size of 2, with 25.92%, followed by a size of 4 with 24.44%.
- A family size of 3 constitutes the smallest portion of our dataset, with 20.20%



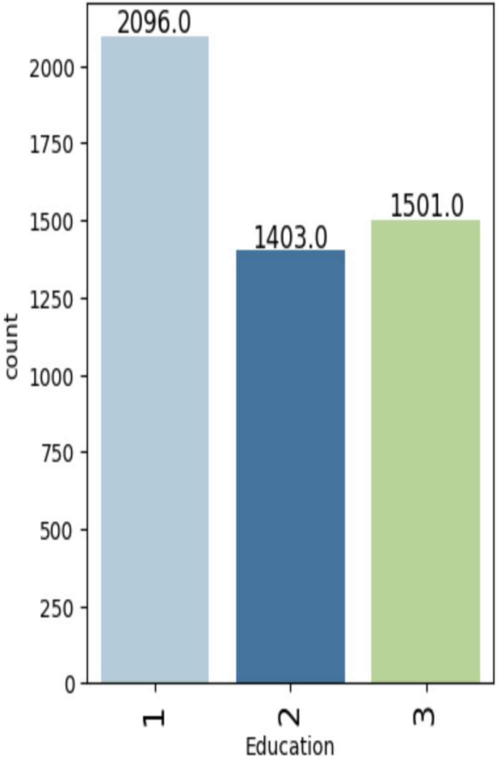
EDA Results - Univariate Analysis

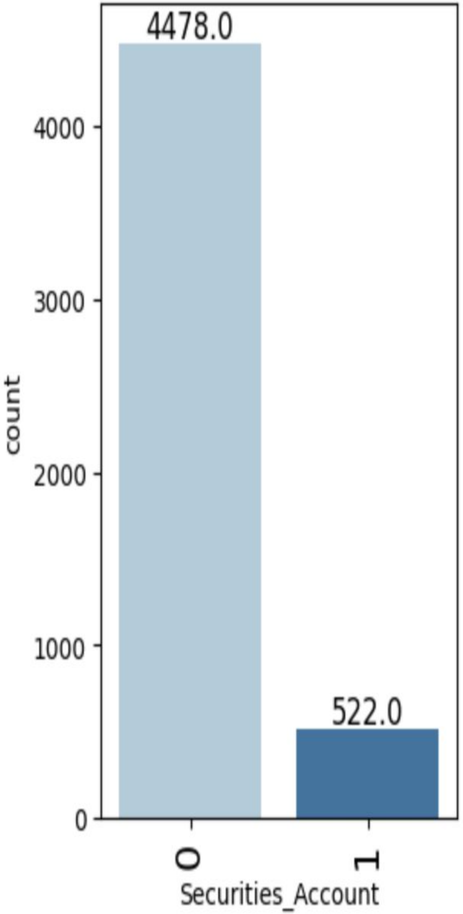
Education -

- The largest educational category is level 1 (Undergraduate), making up 41.92% of the datase
- Level 3 (Advanced/Professional) is the second largest category, comprising 30.0%.
- Level 2 (Graduate) is the smallest educational category, accounting for 28.1% of the dataset.
- There are less number of graduate customers compared to advanced professionals.

Securities_Account -

- Majority of customers(89.6%) doesn't have securit account, Negligible customers have securities account.
- Only 10.4% of customers have a securities accour





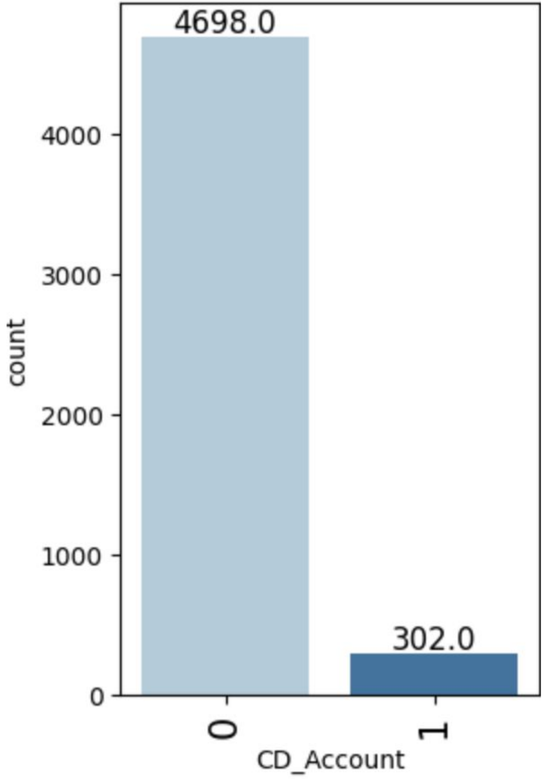
EDA Results - Univariate Analysis

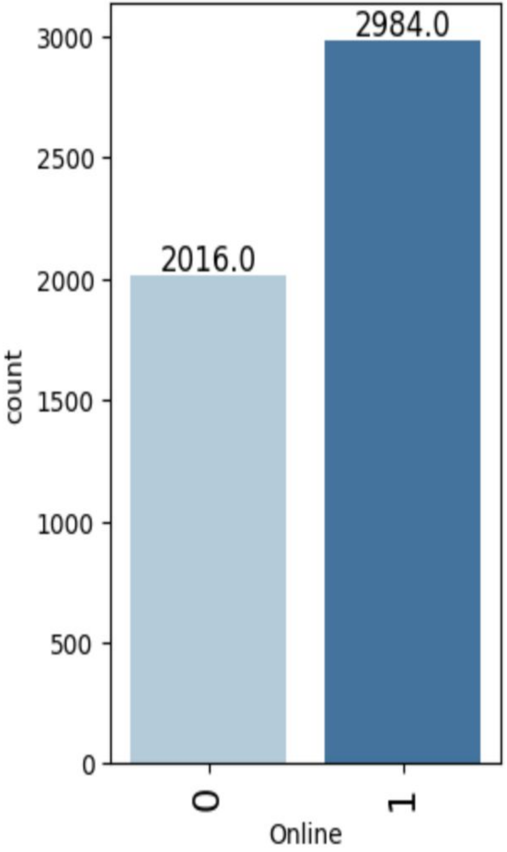
CD_Account-

- Majority of customers doesn’t have Certificate of deposits accounts with the bank
- Negligible amount of customers have CD account.

Online -

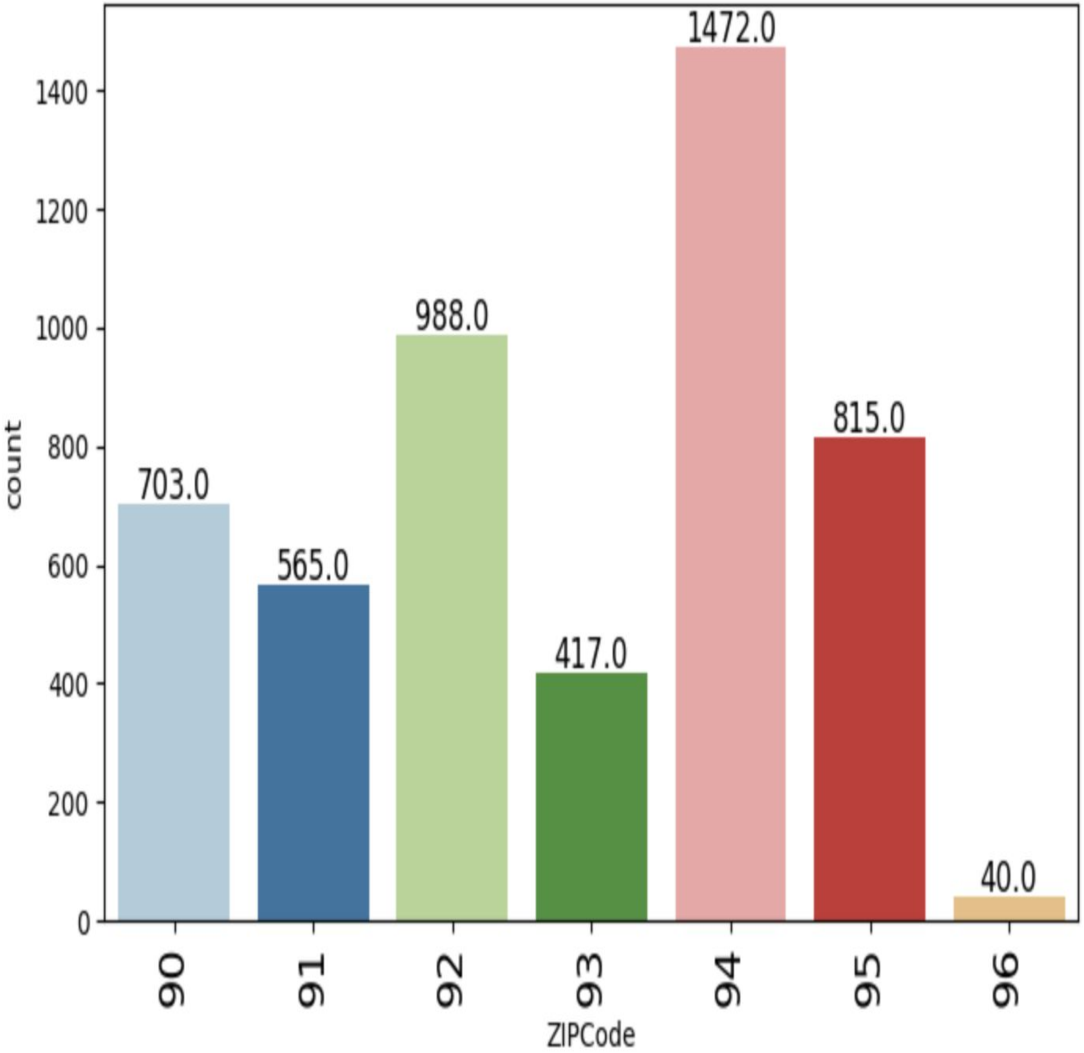
- Majority of customers (2984) uses online banking.
- 2016 (40%) customers doesn't use online banking.





EDA Results - Univariate Analysis

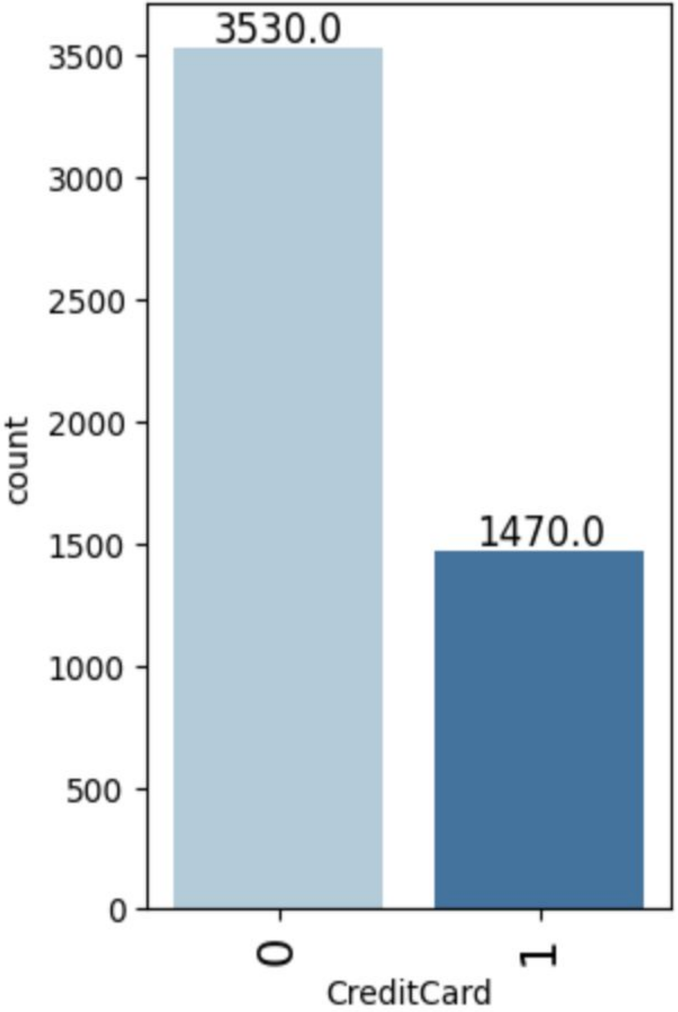
- **Zip Code** - Highest number of customers have zipcode starting with 94, means majority of customers are from same area, rest of the data points are scattered. Very less customers are from Zip code starting with 96.
- **Personal_Loan**- Max number of customers didn’t buy personal loans during last campaign. Only 480 customers bought loans during last campaign.



EDA Results - Univariate Analysis

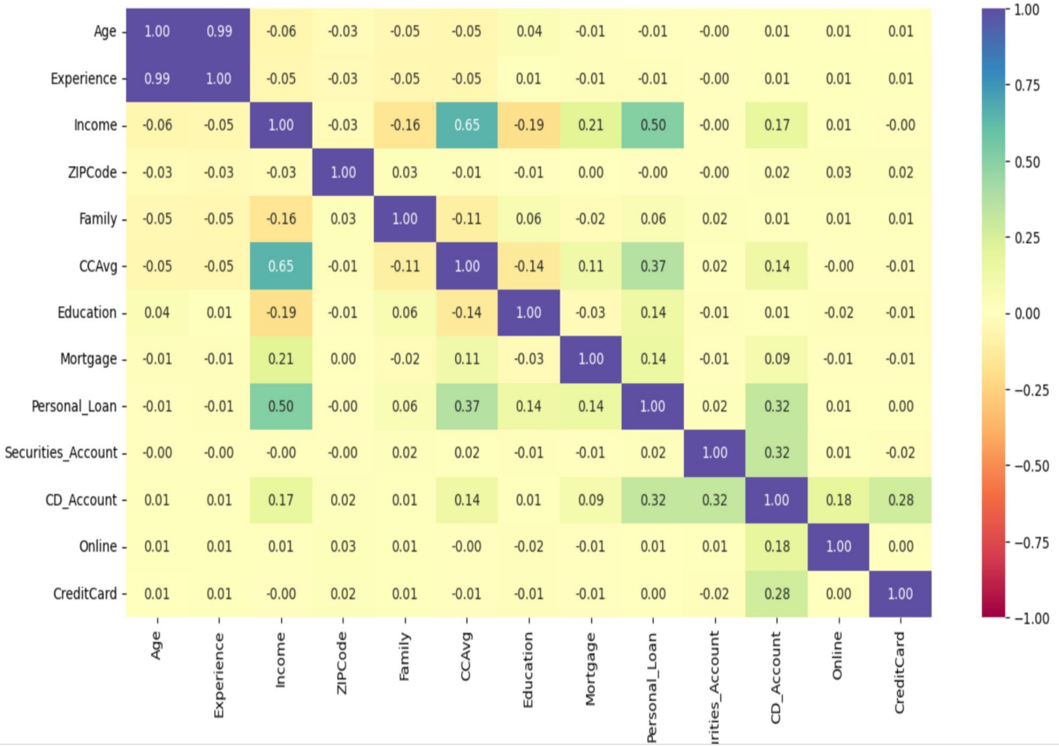
Credit card -

- Majority of customers doesn’t have other bank credit cards, it means there is scope that customer can get personal loans.
- Less number of customers have other bank credit cards.



Correlation Check

- The correlation between Age and Experience is very high (value = 0.99)
- The correlation between Income and CCAvg is low (value = 0.65)
- All other correlation values are quiet small for consideration.
- Family & Income shows negative correlation -0.16
- Income: Positive correlation (0.50), indicating higher income levels are associated with a greater likelihood of accepting a personal loan.
- CCAvg: Positive correlation (0.37), suggesting that higher average credit card spending per month is associated with a higher probability of taking a personal loan.
- CD_Account: Positive correlation (0.32), meaning customers with a certificate of deposit (CD) account are more likely to accept a personal loan.
- Education: Positive correlation (0.14), implying higher education levels are somewhat associated with a higher likelihood of taking a personal loan.



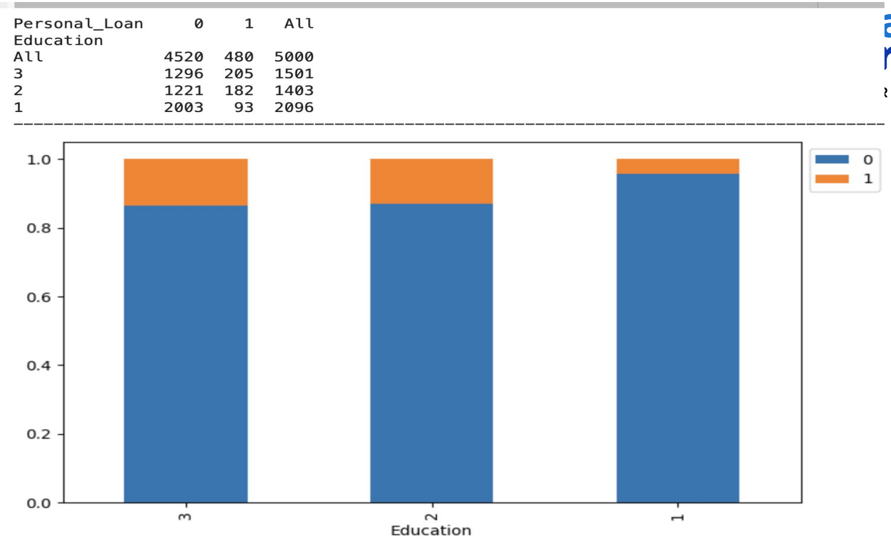
	Age	Experience	Income	ZIPCode	Family	CCAvg	Education	Mortgage	Personal_Loan	rities_Account	CD_Account	Online	CreditCard
Age	1.00	0.99	-0.06	-0.03	-0.05	-0.05	0.04	-0.01	-0.01	-0.00	0.01	0.01	0.01
Experience	0.99	1.00	-0.05	-0.03	-0.05	-0.05	0.01	-0.01	-0.01	-0.00	0.01	0.01	0.01
Income	-0.06	-0.05	1.00	-0.03	-0.16	0.65	-0.19	0.21	0.50	-0.00	0.17	0.01	-0.00
ZIPCode	-0.03	-0.03	-0.03	1.00	0.03	-0.01	-0.01	0.00	-0.00	-0.00	0.02	0.03	0.02
Family	-0.05	-0.05	-0.16	0.03	1.00	-0.11	0.06	-0.02	0.06	0.02	0.01	0.01	0.01
CCAvg	-0.05	-0.05	0.65	-0.01	-0.11	1.00	-0.14	0.11	0.37	0.02	0.14	-0.00	-0.01
Education	0.04	0.01	-0.19	-0.01	0.06	-0.14	1.00	-0.03	0.14	-0.01	0.01	-0.02	-0.01
Mortgage	-0.01	-0.01	0.21	0.00	-0.02	0.11	-0.03	1.00	0.14	-0.01	0.09	-0.01	-0.01
Personal_Loan	-0.01	-0.01	0.50	-0.00	0.06	0.37	0.14	0.14	1.00	0.02	0.32	0.01	0.00
Securities_Account	-0.00	-0.00	-0.00	-0.00	0.02	0.02	-0.01	-0.01	0.02	1.00	0.32	0.01	-0.02
CD_Account	0.01	0.01	0.17	0.02	0.01	0.14	0.01	0.09	0.32	0.32	1.00	0.18	0.28
Online	0.01	0.01	0.01	0.03	0.01	-0.00	-0.02	-0.01	0.01	0.01	0.18	1.00	0.00
CreditCard	0.01	0.01	-0.00	0.02	0.01	-0.01	-0.01	-0.01	0.00	-0.02	0.28	0.00	1.00

Education Vs Personal_Loan

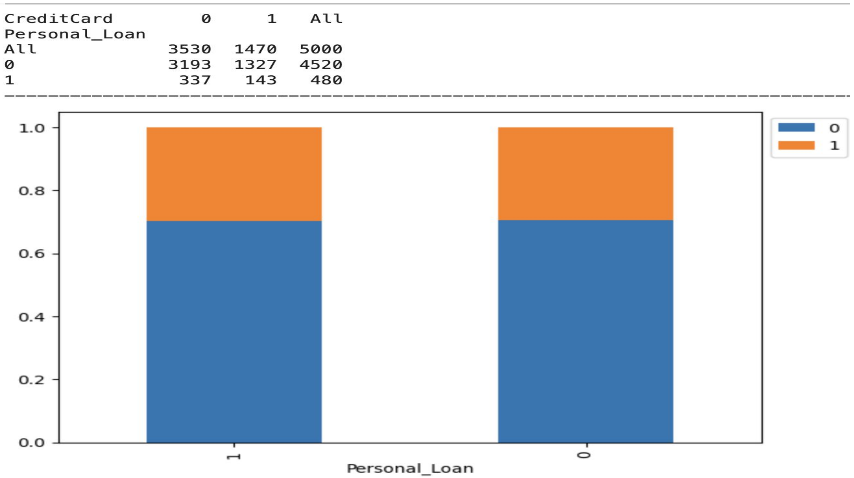
- Customers with Advanced professional education(level 3) purchased maximum loan(205)during last campaign, compared with graduate / undergraduate customers.
- Customers with 'Education' Graduate (level 2) have shown similar loan purchase like Advanced Professionals (level 3)during last campaign.
- Customers whose Education was undergraduates(level1) have shown very less purchase of loan.

Personal_Loan vs Credit Card

- Customers who use credit cards issued by other banks and those who do not are almost equally accepted personal loans during last campaign.



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

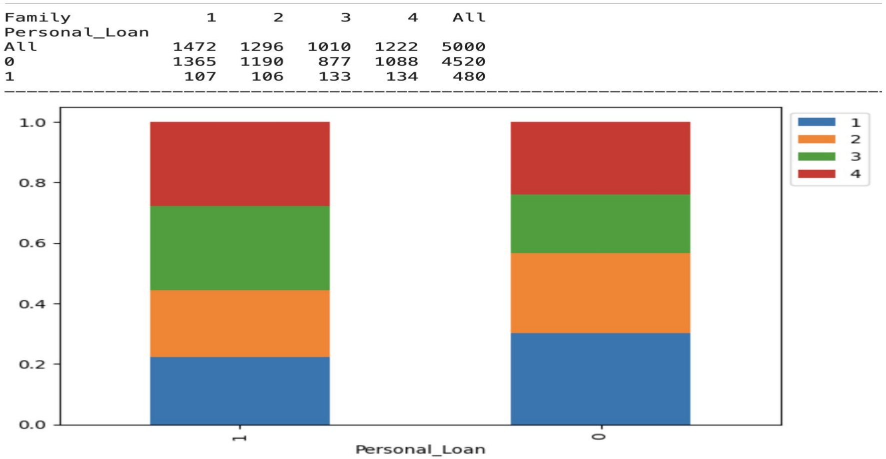


```
CreditCard          0      1    All
Personal_Loan
All              3530   1470   5000
0                3193   1327   4520
1                 337    143    480
------------------------------------------------------------------------------------------------------------------------
```

Bivariate Analysis -Results

Personal_Loan Vs Family

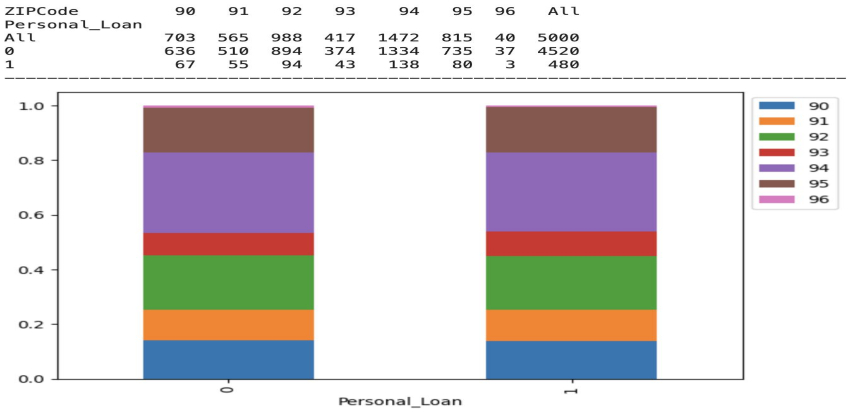
- Customers with 3 or 4 family members bought personal loan during last campaign.
- As family size increases customers are inclined to accept personal loans



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

Personal_Loan Vs ZIP Code

- Customers with zip code starts with '94' accepted the highest personal loans in last campaign, compared with other ZIP Codes.



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

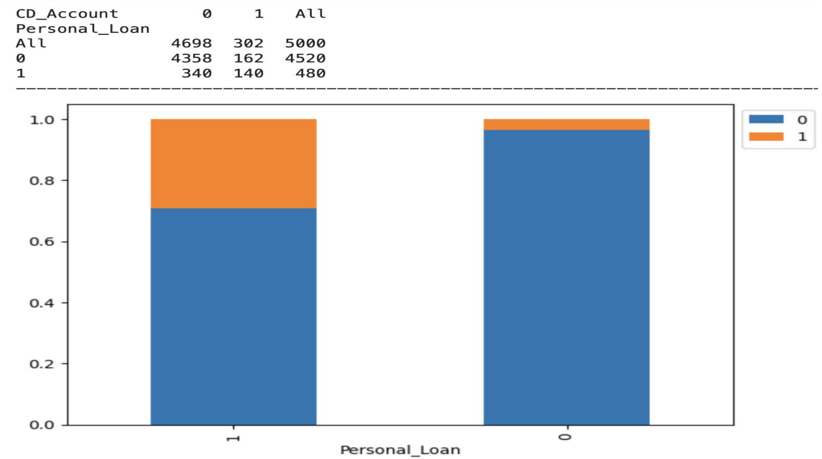
Bivariate Analysis -Results

Personal_Loan Vs CD_Account

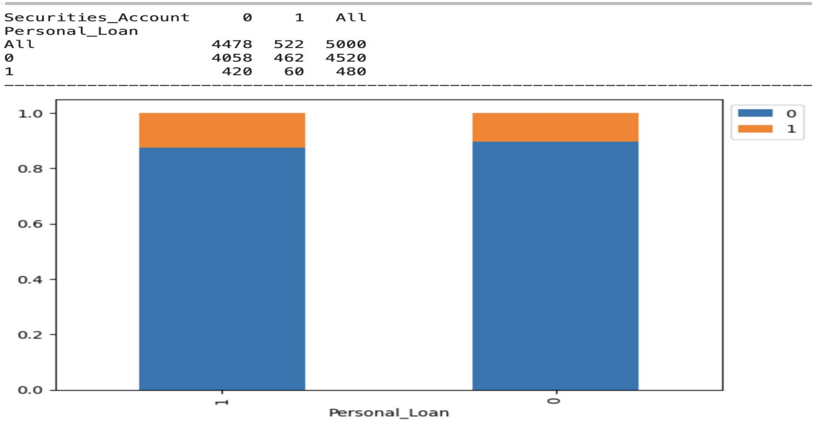
- Customers with CD account are more inclined to accept personal loans.

Personal_Loan Vs Securities_Account

- Customers with security accounts shown more interested in buying personal loans.
- Total 522 customers have securities account, out of them 60 customers with security account opted loans in last campaign,462 customers did not buy personal loans , they can be potential customers.



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

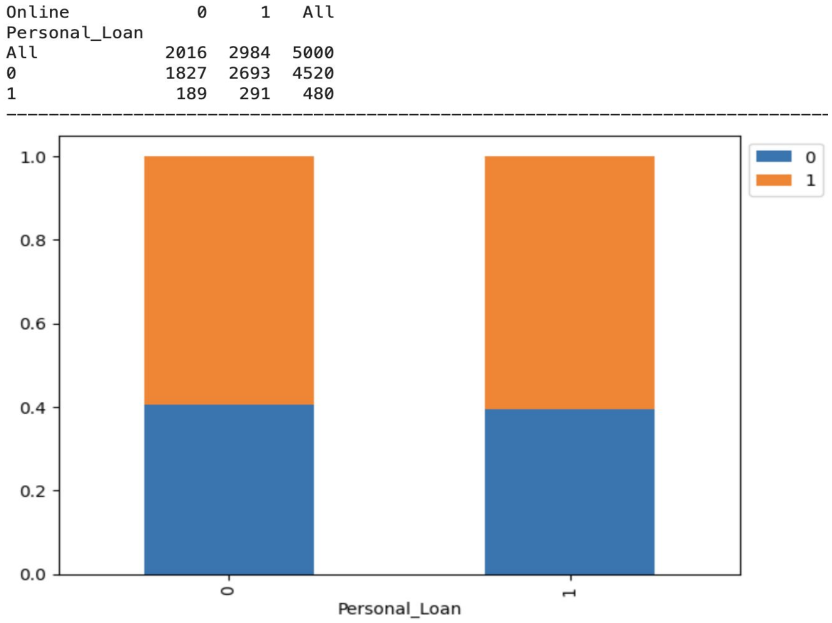


```
Securities_Account     0    1   All
Personal_Loan
All                 4478  522  5000
0                   4058  462  4520
1                    420   60   480
------------------------------------------------------------------------------------------------------------------------
```

Bivariate Analysis -Results

- **Personal_Loan vs Online**

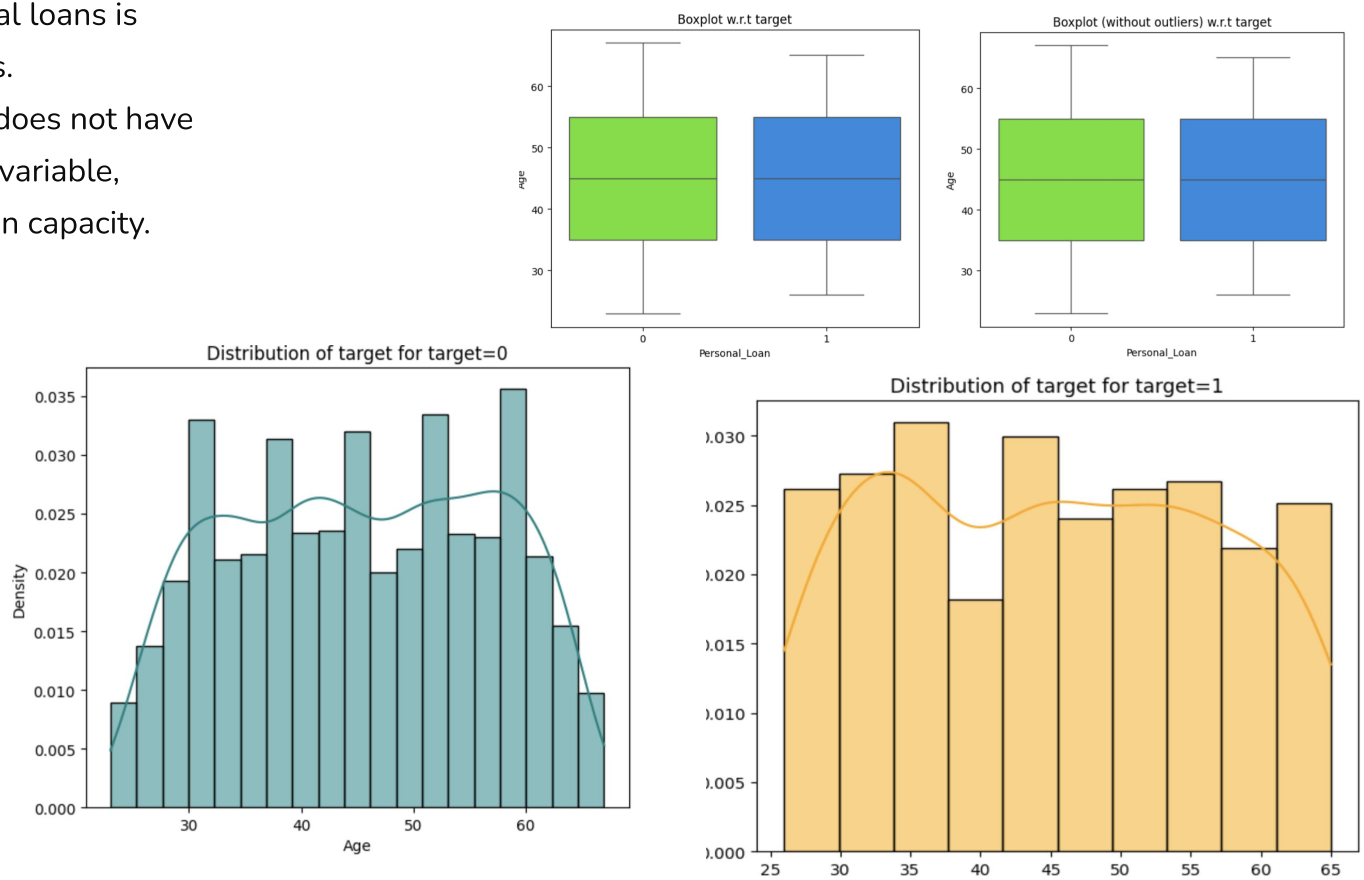
  Customers who uses online banking & customers who did not use online banking opted almost equally accepted personal loans during last campaign.



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

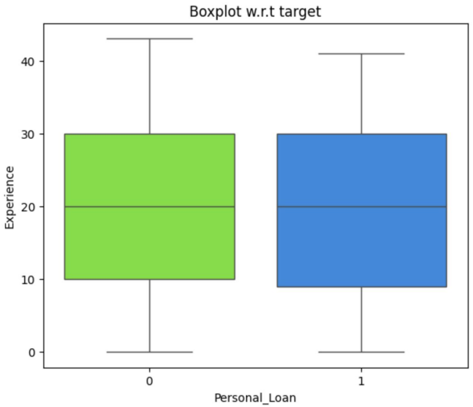
Personal_Loan Vs Age

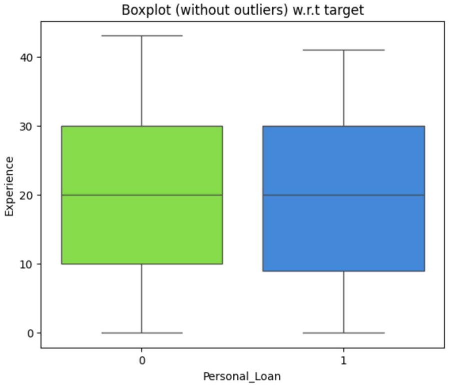
- The mean age for customers accepting and not accepting personal loans is almost similar i.e 45 years.
- As shown in graphs,Age does not have a relation with our target variable, hence negligible prediction capacity.

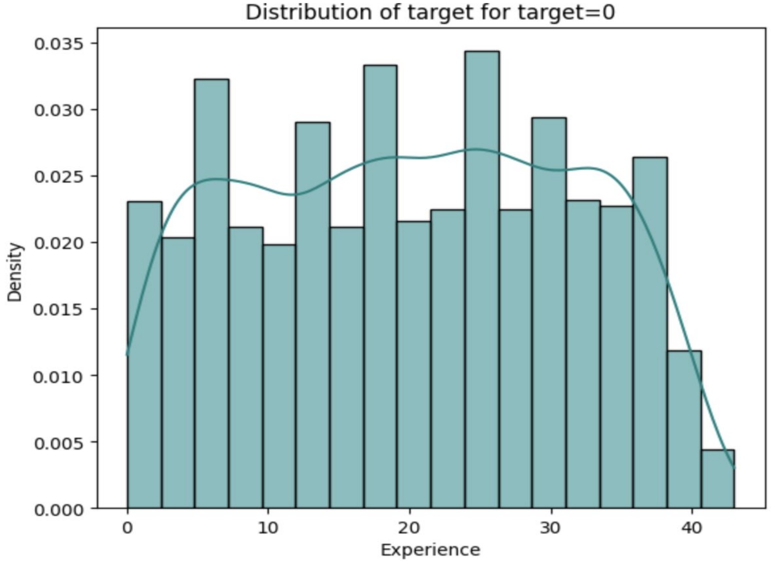


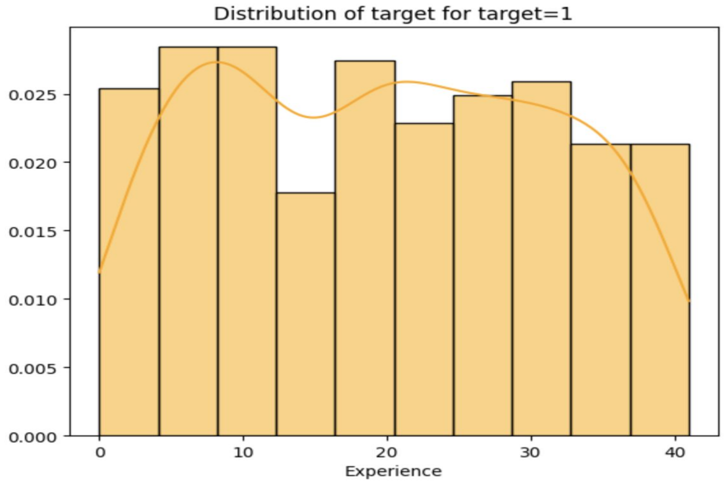
Personal Loan Vs Experience

- The mean Experience for customers accepting and not accepting personal loans is very close i.e 20 years.
- As shown in graphs, Experience does not have a relation with target variable hence a prediction power on our target variable is less.





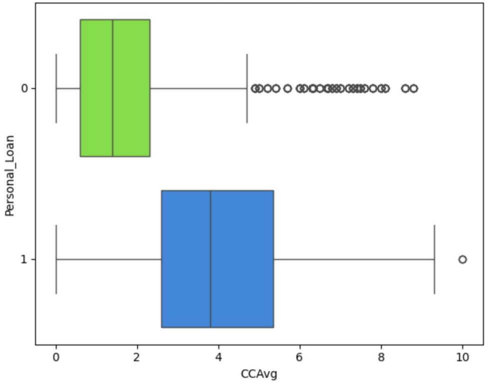




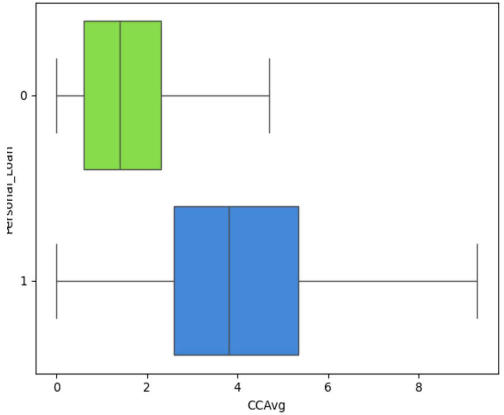
Personal Loan Vs CCAvg -

- The mean value of CCAvg of customers not accepting a personal loan is approx 1.6K which is much less than the mean value of customers accepting personal loan of approx 4K.
- It infers that the customers who accepted personal loans spends more money on credit card on monthly basis.

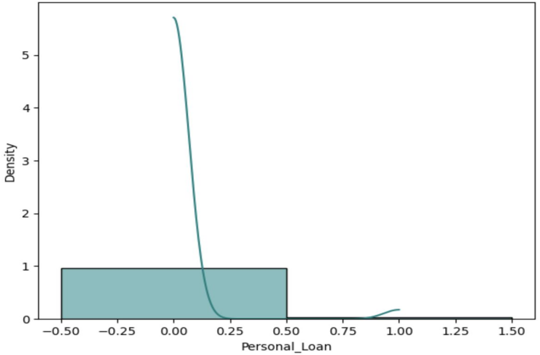
Boxplot w.r.t target



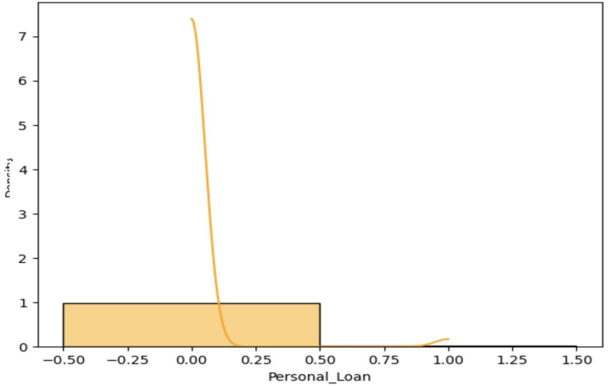
Boxplot (without outliers) w.r.t target



Distribution of target for target=1.6

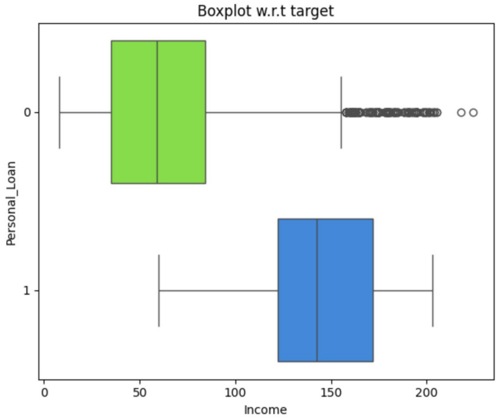


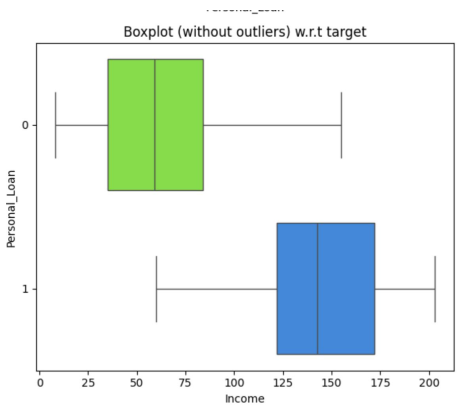
Distribution of target for target=1.5

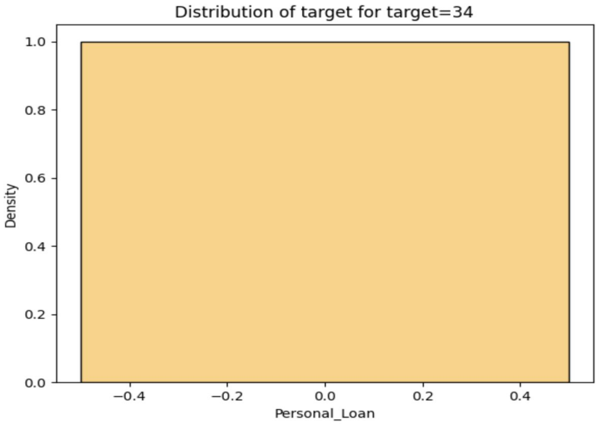


Personal Loan vs Income

- The mean income for the customers who did not accept the loan is approx $65K
- The mean income for the customers who did accept the loan is approx 145K
- This shows that the income level of the customer has a high impact on the customer decision to accept a personal loan, The higher the income, the more chances the customer will accept a personal loan.







Data Pre-Processing

- **Dropping columns** (data = data.drop(['ID'], axis=1)) Dropped 'customer ID' column, as it's unique numbering for customers & doesn't lead to any valuable inference.
- **Checking for Anomalous Values**-   Few negative values observed in the 'Experience' column. Treated negative values [-1, -2, -3] with (.replace() function)
- **Feature Engineering** - Converted 'Zip code' column to Categorical feature.Converted data types of all categorical features(Education,Personal_Loan ,Securities_Account ,CD_Account, Online ,Credit Card , ZIP Code) to 'Category' .
- **Outlier Detection** - For Outlier detection verified numerical columns & calculated 25th percentile and 75th percentile using.quantile function on Q1 / Q3 values.Verified IQR values for Age, Experience, Income,Family, CCAvg, Mortgage columns. Verified lower and upper bounds for all values. Only Income, CCAvg, Mortgage columns has outliers.
- **Missing Value Treatment -** There are no missing values in the dataset.
- **Data Preparation for Modeling -**
- Dropping 'Experience' column data as it is perfectly correlated with 'Age' using data.drop().
- Dropping 'Personal loan' column from X variable, as 'Personal loan' is our final target variable. So We are assigning it to Y variable, We need to find all other features (X) which impacts Personal Loan.
- **Creating dummy variables -**
- Creating dummies with (pd.get_dummies) for 'ZIPCode' & 'Education' column. These 2 columns are categorical columns with various features, so need to create dummies for covering all features.
- **Splitting data**-Splitting data into 70% train - 30% test data & setting random state=1(This sets a random seed for the splitting process.) to ensures same split after running code multiple times.

Model Evaluation Criterion

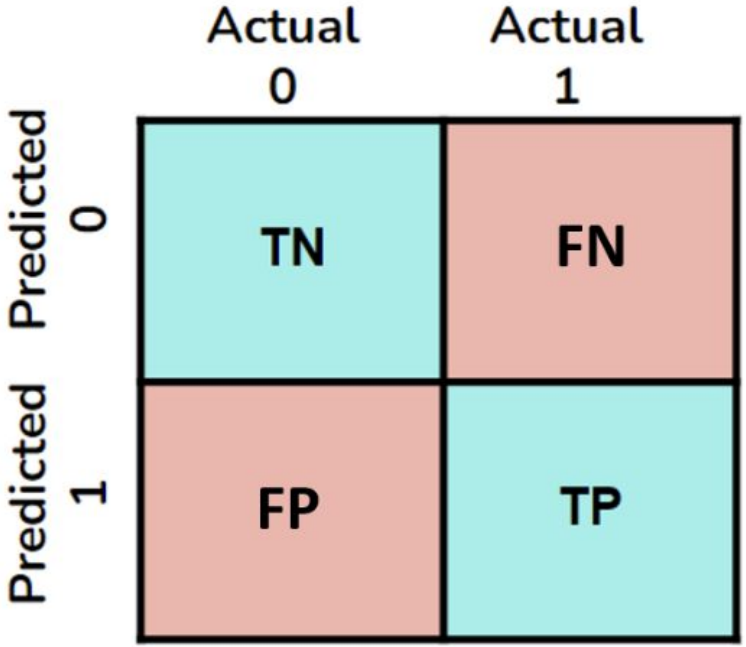
We can use **Confusion Matrix** to provide a comprehensive breakdown of the model's predictions Vs the actual outcomes across different classes (in our case, loan acceptance or rejection).

True Positive (TP): Represents a correct prediction where marketing efforts can be efficiently targeted for conversion.

True Negative (TN): Indicates accurate identification of customers unlikely to convert, saving resources on unnecessary marketing.

False Positive (FP): Wastes resources on customers unlikely to convert, leading to inefficient marketing spending.

False Negative (FN): Misses potential conversion opportunities, resulting in lost revenue and missed customer engagement.



	Actual 0	Actual 1
Predicted 0	TN	FN
Predicted 1	FP	TP

Model Building Criterion

In case of All life bank we should focus on Recall & avoid below cases -

- **False Positive :Loss of Resources**

Predicting a customer will accept a loan when they would not actually accept one results in a loss of resources.

- **False Negative :Missed Opportunity**

Predicting a customer will not accept a loan when they would have actually accepted one represents a missed opportunity.

How to reduce this loss (False Negatives)?

- **Recall should be maximized, the greater the recall higher the chances of minimizing the false negatives.**hence,
- The recall_score function will be used to check the model performances.
- **Strategy to Minimize Missed Opportunities** - **Recall**
- **Recall Formula -**

$$\boxed{Recall = \frac{TP}{TP+FN}}$$

A higher Recall rate improves the model's ability to accurately identify potential customers, thereby reducing missed opportunities.

Decision Tree -Model Building

- Used below steps for decision tree model building-
  - **Data Preparation** - Gathered and preprocessed the loan modelling dataset, handled missing values, encoded categorical variables, and splitted categorical variable (all features plotted on (X)) and target variable (y)i.e 'Personal_loan'.
  - **Splitting Data-** Divided the dataset into train and test data sets using '`train_test_split`' from `sklearn.model_selection`.
  - **Instantiate the Model**- Created instances of the decision tree classifier using `DecisionTreeClassifier` from `sklearn.tree` at each time building & improving models(baseline/ pre-pruned / post-pruned)
  - **Model Training -** Fit the decision tree classifier to the training data using `fit(X_train, y_train).`
  - **Model Evaluation -**Predicted the target variable (Personal_loan) on the test data. Evaluated the model performance using metrics like accuracy, precision, recall, and F1-score.
  - **Hyperparameter Tuning-** Performed hyperparameters tuning (e.g., `class_weights`, `max_depth`, `min_samples_split`) using techniques like Grid search CV (`Gini, Entropy`, `GridSearchCV`) to optimize model performance.(Built Decision tree model on train data using '**Gini Impurity**' as criterion., setting seed random state =1)
  - **Feature Importance -**Extracted feature importance scores using (`feature_importances)` attribute to identify key predictor features of loan acceptance.
  - **Visualization -**Visualized the decision tree structure using `plot_tree` to interpret decision rules and node splits.

Baseline - Model Building on Train Data

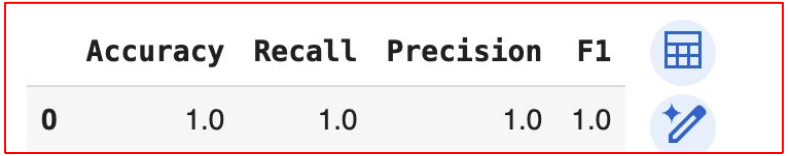
Defining Functions -

Used below two functions for model building & model performance evaluation -

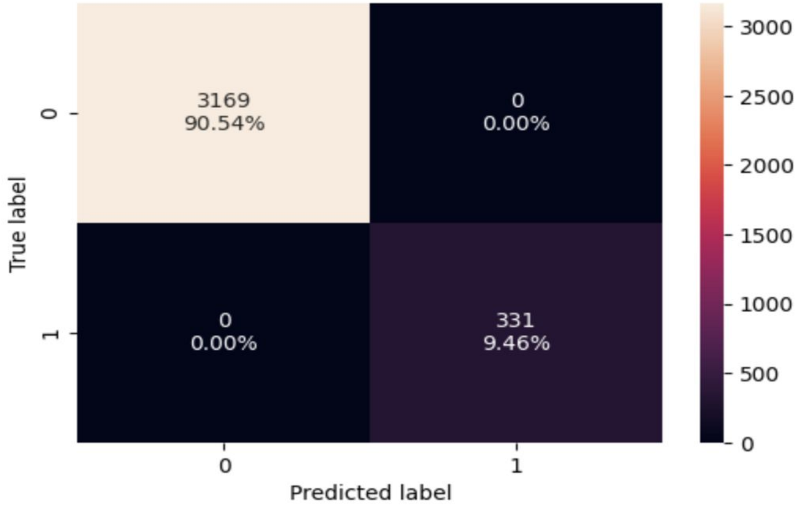
- ‘model_performance_classification_sklearn’ function to check the model performance of models.
- ‘confusion_matrix_sklearn’ function to plot confusion matrix.

Checking model performance - Train Data

Baseline Train data model performance is as below -



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0



Visualizing Decision tree on Train data model -

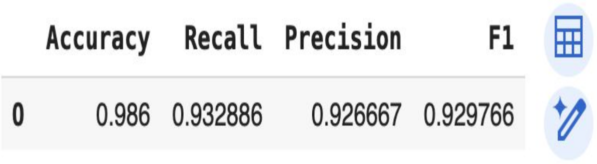
Observations -

- The first split was on the "Income" variable, so it’s the most important feature to be considered.
- **Variable Importance** – The major important variables observed are Income, Family,Education_2,   Education_3, CCAvg, Age.

Baseline - Model Performance on Test Data

Checking Model Performance -Test data

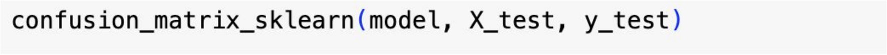
Baseline(Test data) model performance is as below -



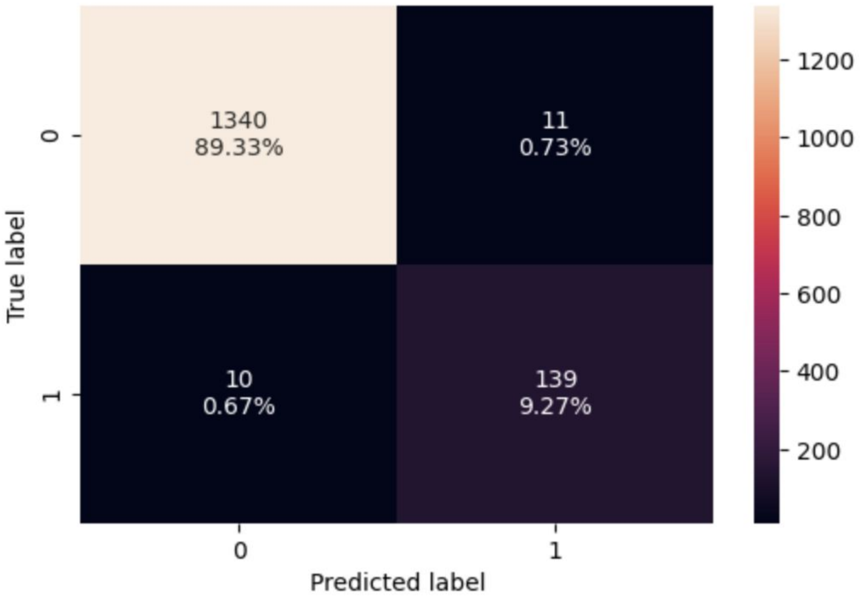
	Accuracy	Recall	Precision	F1
0	0.986	0.932886	0.926667	0.929766

Recall Comparison of train & test data -

Recall on Train Data = 1.0
Recall on Test Data = 0.932886



```
confusion_matrix_sklearn(model, X_test, y_test)
```

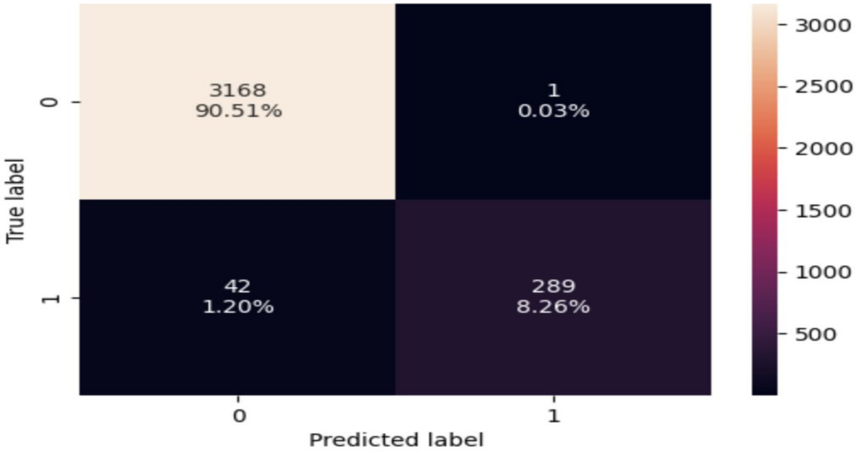


Observations -

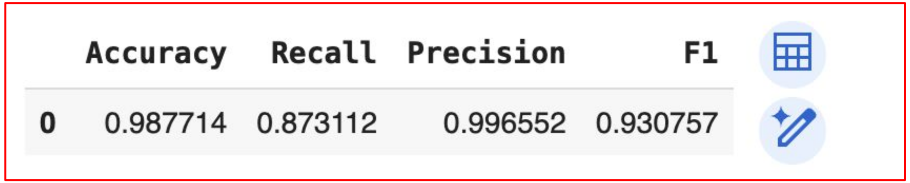
- After computing tree complexity observed the max depth of model is 10, node count is 97 and number of leaves are 49
- A quiet big mismatch is observed between train and test sets performance, hence observed that Baseline model is overfitting the data.
- The original decision tree is quite complex and is overfitting the training data set, hence pre-pruning and post pruning is required to check improved model performance.

Pre-Pruning Model Building on Train Data

Pre-Pruning Decision Tree Model with train data



Checking PrePrun model performance - Train Data



	Accuracy	Recall	Precision	F1
0	0.987714	0.873112	0.996552	0.930757

Visualizing Decision tree on Train data model -

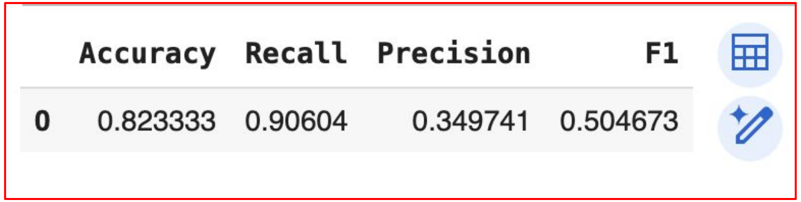
Refer Appendix for Decision tree

Observations -

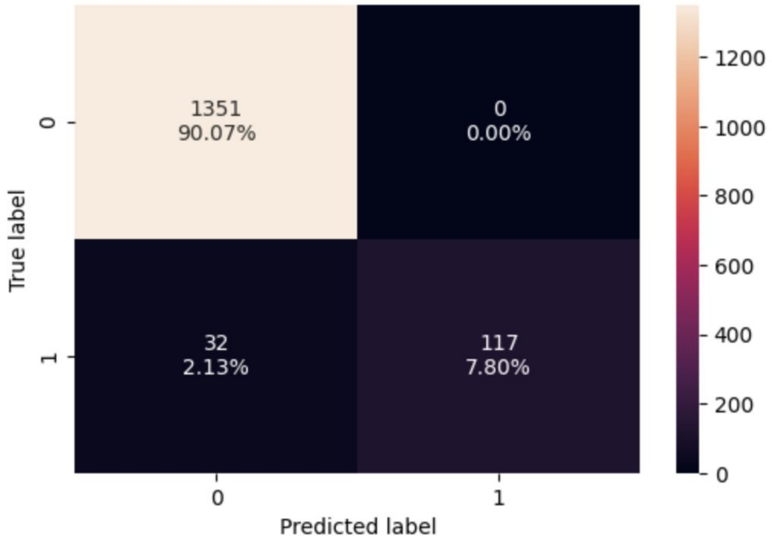
- After computing tree complexity observed the max depth of model is 5, node count is 19 and number of leaves are 10
- The model shows Income is the most important feature, followed by family & Education(Graduate), CCAvg, and Age

Pre-Pruning Model Performance Test Data

Visualizing Decision tree on Test data model -



	Accuracy	Recall	Precision	F1
0	0.823333	0.90604	0.349741	0.504673



Recall Comparison of train & test data -

Recall on Train Data = 0.873112
Recall on Test Data = 0.90604

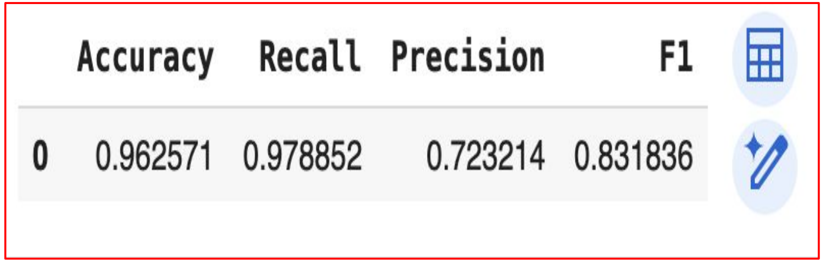
Observations -

- The Pre-pruned model is showing improved Recall on Test data.
- GridSearchCV optimizes model hyperparameters by exhaustively searching through a predefined grid of parameters, assessing each combination via cross-validation. It automates the selection process to identify the best-fit model estimator, enhancing model performance and accuracy

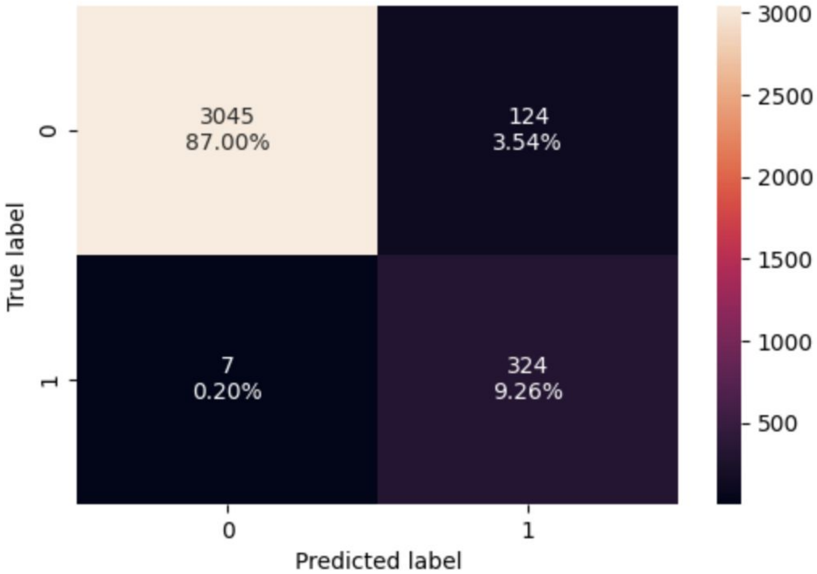
Post-Pruning Model Building on Train Data

Checking model performance - Train Data

Post -Pruned Train data model performance is as below



	Accuracy	Recall	Precision	F1
0	0.962571	0.978852	0.723214	0.831836



Visualizing Decision tree on Train data model -

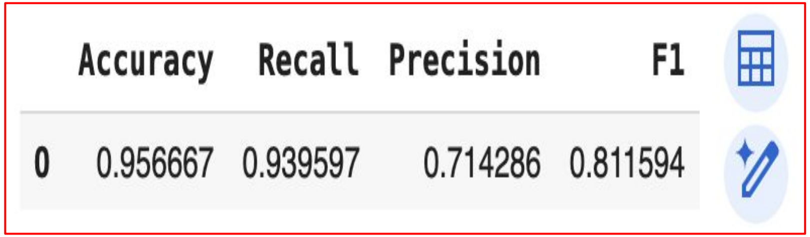
Refer Appendix for Decision tree

Observations -

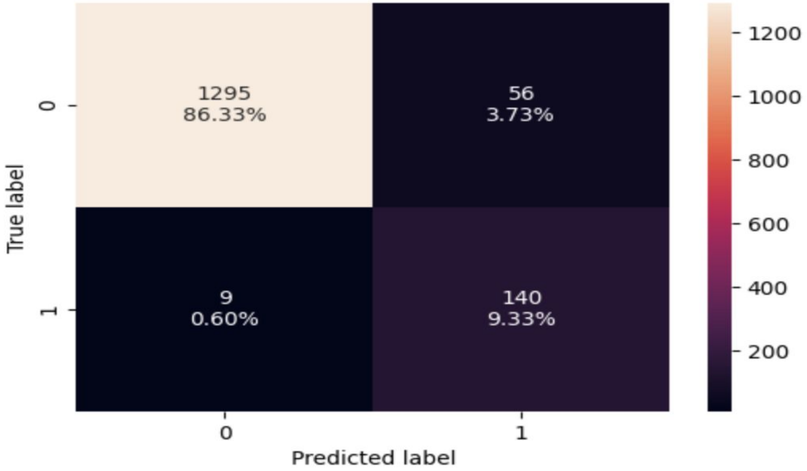
- After computing tree complexity observed the max depth of model is 5, node count is 23 and number of leaves are 12.
- The first split was on the "Income" variable, so it's the most important feature to be considered.
- The model shows Income is the most important feature, followed by Education(Graduate), CCAvg, and family
- Cost complexity pruning reduces overfitting in decision trees by trimming less significant nodes, controlled by the ccp_alpha parameter. Increasing ccp_alpha leads to a simpler tree, enhancing its ability to generalize to new data.

Post-Pruning Model Performance on Test Data

Visualizing Decision tree on Test data model -



	Accuracy	Recall	Precision	F1
0	0.956667	0.939597	0.714286	0.811594



Recall Comparison of train & test data -

Recall on Train Data = 0.978852
Recall on Test Data = 0.939597

Observations -

- Visualizing the decision tree helped in understanding the decision rules used by the model.
- Feature importance provided insights into which features are most influential in determining loan acceptance or rejection.(Income , Family, Education_2,Education_3,CCAvg, Age) are important features to be considered.
- The tree is very optimal with a depth of **5** and **23** nodes.

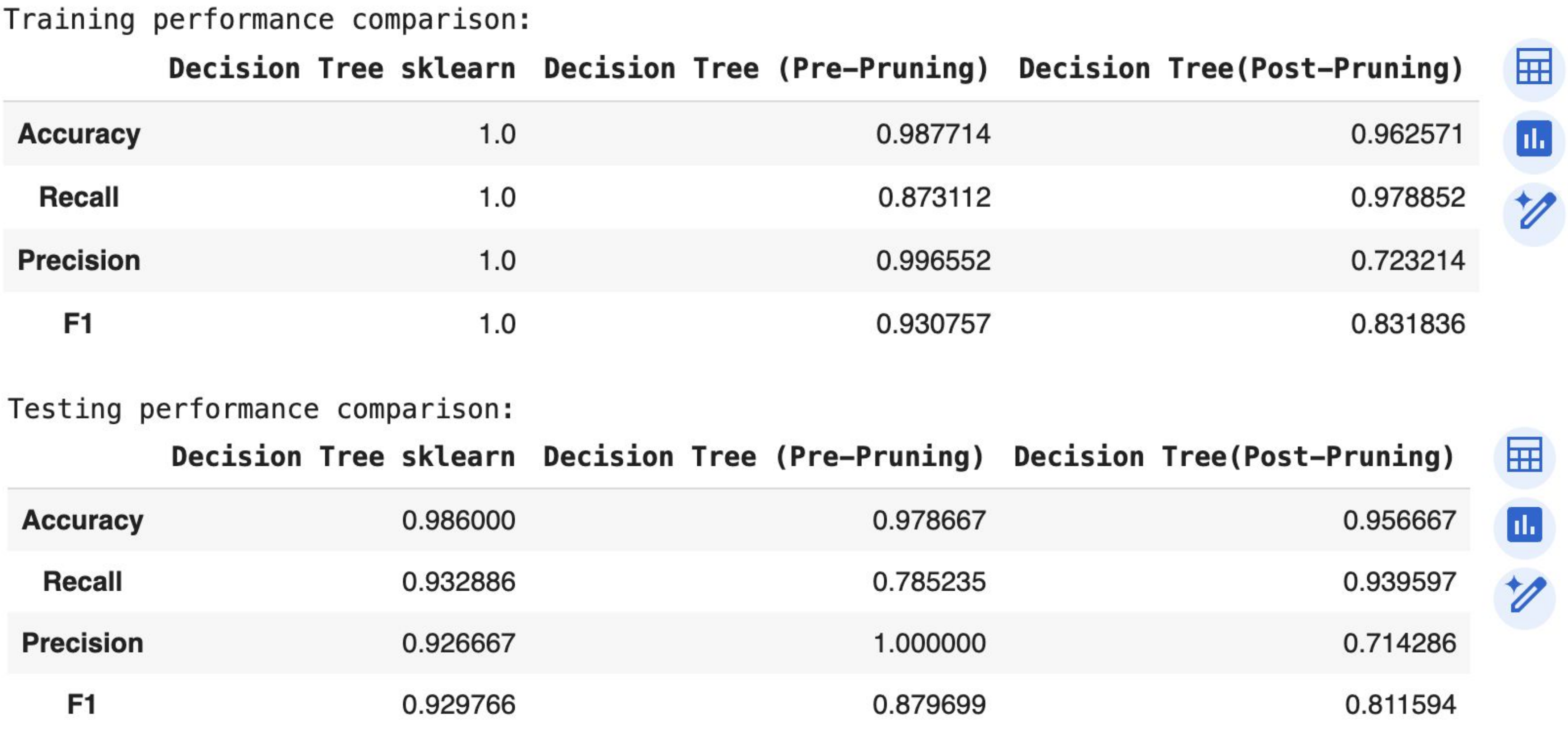
Summary of Most Important Features

Based on the summary of ‘Feature Importances’ we can infer that -

- Income is the first most important feature to be considered.
- Education_2(Grad) is the second important feature selected by model algorithm.
- Family is also important feature in the dataset.
- Education 3 , CC Avg, Age, CD_Account are little less important features compared with Income, Education & Family.

Baseline Model	Pre-Pruned Model	Post-Pruned Model
Income 0.308098 Family 0.259255 Education_2 0.166192 Education_3 0.147127 CCAvg 0.048798 Age 0.033150 CD_Account 0.017273	Income 0.337681 Family 0.275581 Education_2 0.175687 Education_3 0.157286 CCAvg 0.042856 Age 0.010908	Income 0.628714 Education_2 0.150473 Education_3 0.072248 CCAvg 0.071238 Family 0.065218 CD Account 0.012109

Model Performance Comparison



Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree(Post-Pruning)
Accuracy	1.0	0.987714	0.962571
Recall	1.0	0.873112	0.978852
Precision	1.0	0.996552	0.723214
F1	1.0	0.930757	0.831836

Testing performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree(Post-Pruning)
Accuracy	0.986000	0.978667	0.956667
Recall	0.932886	0.785235	0.939597
Precision	0.926667	1.000000	0.714286
F1	0.929766	0.879699	0.811594

Model Selection

The Recall Score calculation for all models is as below -

1. Decision Tree (sklearn): Recall = 1.0 (100%)
2. Decision Tree (Pre-Pruning): Recall = 0.873112 (87.31%)
3. Decision Tree (Post-Pruning): Recall = 0.978852 (97.89%)

- Pre-pruned model shows good balance between recall and overall performance, making it suitable for scenarios where identifying as many true positive cases as possible is critical.
- **The best model based on Recall Score is the post pruning model as it shows highest Recall among all three model**.
- The post-pruned decision tree model is most effective at identifying and capturing a high proportion of customers who would actually accept a loan (true positives) out of all positive instances in the dataset.

Model Performance Improvement

Below steps has been taken for the model performance improvement using different pruning techniques -

- Pruning techniques (Pre-Pruning and Post-Pruning) were applied to decision tree models to improve performance and generalization.
- Results show that the decision tree with Pre-Pruning achieved a good balance of accuracy, precision, and recall compared to the baseline model.
- Post-Pruning further refined the model, significantly improving recall (ability to capture positive instances) while maintaining acceptable precision.
- Pruning techniques effectively prevented overfitting, enhancing the model's ability to generalize to new data and perform well in real-world scenarios.
- The decision tree with Post-Pruning is recommended for applications where maximizing recall (identifying positive instances) is crucial without sacrificing overall performance.

Actionable Insights

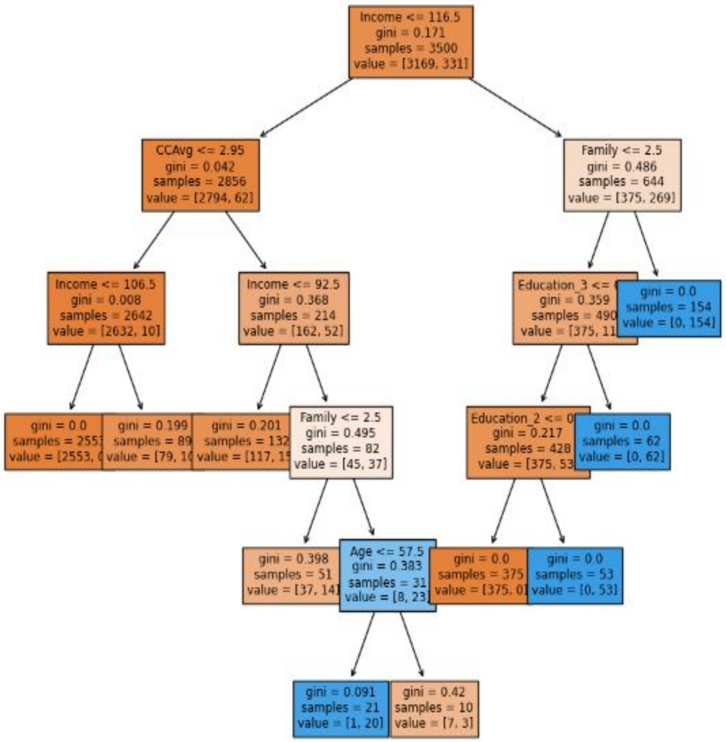
- The best performing model was derived from the Decision Tree Modelling technique where the original tree was post pruned via ccp_alpha=0.0.003 and gave the below Recall values for test and training data sets:
- Recall for Post pruned model on Train Data = 0.978852
- Recall for Post pruned model on Test Data = 0.939597
- The statistical evidence show the Features that most affects the client decision to accept a personal loan are listed in below table with priority levels (1 being the highest priority and 5 being the lowest priority:
- Priority Feature Effect on customer
  - Income : The higher the income, the more chances the customer will accept a personal loan
  - Education_2 Customers with Education level 2 are more willing to accept a personal loan than levels 1 & 3
  - CCAvg As the monthly spending of customers increase, the more they are willing to accept personal loan
  - Education_3 Customers with Education level are more willing to accept a personal loan than level 1
  - Family As family size grows, customers are more willing to accept personal loan

Business Recommendations

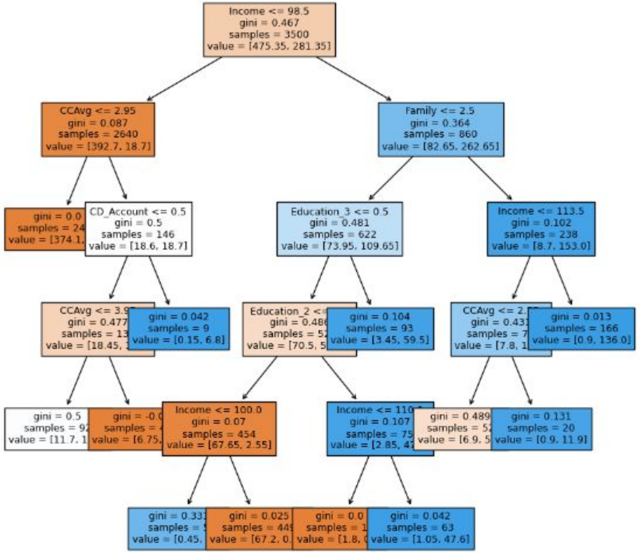
- **Leverage Income Data:** Focus marketing initiatives on targeting customers with an income of $100K and above, as income is a significant predictor of a customer's likelihood to be interested in personal loan products. Engage this high-income segment to maximize the potential for loan product uptake.
- **Capitalize on High Income and Education:** Intensify marketing campaigns towards individuals earning above $100K and with advanced educational qualifications. Develop premium loan offerings that align with the financial capabilities and sophistication of this customer base.
- **Incorporate Secondary Indicators:** Include considerations of education level, family size, and credit card spending habits in segmenting and targeting, while recognizing their impact is secondary to income.
- **Broader Market Insights:** Do not overlook the moderate influences of age, CD account holdings, and regional factors indicated by ZIP codes. These variables should inform a more nuanced segmentation and personalized marketing strategy.
- **Develop a Personalized Targeting Approach:** Craft personalized marketing strategies for customers with incomes below $100K. This group presents an opportunity for market expansion and could benefit from targeted financial products designed to meet their specific circumstances.
- The marketing team is recommended to study the customers profiles first before approaching them for a personal loan offer.
- The top 5 features stated in the features (Income, Education_2, CCAvg,Education_3,Family) need to be considered as the target customer profile for a personal loan campaign.

Appendix

Pre-pruning - Decision Tree



Post-pruning - Decision Tree



Great Learning

Happy Learning !