Liability to Loan: Conversion Analysis

Personal Loan Campaign
Post Graduate AI/ML Business Applications

March 12, 2025

Contents / Agenda

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

Executive Summary

To **Identify potential customer** who have a **higher probability** of **purchasing a loan** using a Decision Tree Classifier as the AI/ML-powered predictive model

- ZIP Code Diversity encompasses a significant range of geographic locations, as evidenced by the presence of 467 unique ZIP Code combinations. This suggests that the customer base is geographically diverse

- Analysis of the ZIP Code data reveals that there are 7 unique prefixes when considering only the first two digits.

- Breakdown of **loan acceptance rates** (percentage of customers who accepted the loan) for each ZIP code prefix:
  - 90: 9.53%
  - 91: 9.73%
  - 92: 9.52%
  - 93: 10.31%
  - 94: 9.38%
  - 95: 9.82%
  - 96: 7.5%

Executive Summary

- Total Customer 5000
- 9,6 % Personal Loan Acceptance (480 borrowers)
- 90.4% Personal Loan Rejection (4520 borrowers)
- Average age of customer who purchase loan is 45 years of age. With the minimum age of 23 and maximum age of 67
- Income average is $73000 and maximum income is $224000
- Family size of 2 is the average and 4 the maximum family size
- The average mortgage is $101000 and maximum is $635000
- Online Banking Users 2984
- Non-Online Banking Users 2016

Executive Summary

- Professional Degree): 205 out of 1501 (13.7%) accepted the loan. This is the highest acceptance rate among the education levels
- Graduate Degree 182 out of 1403 (13.0%) accepted the loan.
- Undergraduate Degree 93 out of 2096 (4.4%) accepted the loan. This is the lowest acceptance rate.

Executive Summary

Insights

- **ZIP Code Prefixes** and Counts where customer Live:
  - **90:** 703 customers
  - **91:** 565 customers
  - **92:** 988 customers
  - **93**: 417 customers
  - **94:** 1472 customers
  - **95:** 815 customers
  - **96:** 40 customers
- 20 years of experience is the average for customer
- **Outliers** indicate that there are some people of very young or very old age that accepted the loan
- **2%** of customer have **income over** the threshold of **$186500**
- **Customers** that **use** a **Credit Card** is **41%** and **59% do not** use a credit card
- 62% of customer accepted a personal loan with only 38% who did not accept
- Customers with **higher professional Experience** were more likely to **accept the personal** loan.
- Only **6.5%** of **customer spend over** the **CCAvg** per month
- **5.8 % of Customers** have a **Upper Mortgage** Outlier of **$252500**
- **A family of 4** is identified as an **outlier**, as the **average family** size consists of **2 members**
- **110 Customer** are of the **age 85**, where the **average customer** age is **45**
- **50 Customer** have over **40 years** of Experience
- **Average Mortgage is $56000** where 75% of customers have a mortgage of **$101000**
- **42% customer** have **undergraduate, 28% graduate and 30%** are **professional degrees**

Executive Summary Insights

Insights

- ZIP code prefix 96 has the fewest customers (40)
- While there's a minor tendency for **older customers** to show slightly **more interest** in **purchasing loans**, the age distributions between loan purchasers and non-purchasers are remarkably similar. Therefore, age alone is not a strong determining factor for loan interest.
- Customers with **higher "CCAvg"** were much more likely to **accept** the **personal loan.**There is a strong positive relationship between credit card average spending per month ("CCAvg") and personal loan acceptance.
- Customers **Using Online Banking**
  - 291 out of 2984 **(9.75%) accepted** the personal loan.
  - 2693 out of 2984 (90.25%) did not accept the personal loan
- Customers **Not Using Online Banking**
  - 189 out of 2016 (**9.37%) accepted** the personal loan.
  - 1827 out of 2016 (90.63%) did not accept the personal loan.

Executive Summary

Insights

- **Customers with CD Accounts**:
  - 140 out of 302 **(46.4%) accepted** the **personal loan**.
  - 162 out of 302 (53.6%) did not accept the personal loan.

- **Customers without CD Accounts**:
  - 340 out of 4698 (**7.2%) accepted** the **personal loan**.
  - 4358 out of 4698 (92.8%) did not accept the personal loan.

- Customers with **Securities Accounts** have a slightly higher personal **loan acceptance rate (11.5%)** compared to those **without (9.4%)**

- **Family Size loan acceptance rates** (percentage of **Family** who **accepted** the Personal loan)
  - Family Size 1: 7.27% (107/1472)
  - Family Size 2: 8.18% (106/1296)
  - Family Size 3: 13.17% (133/1010)
  - Family Size 4: 10.97% (134/1222)

Business Problem Overview and Solution Approach

AllLife Bank, a leading US-based financial institution, is committed to strategically expanding its personal loan customer base by converting existing depositors into borrowers. Despite achieving a promising 9% conversion rate through a previous marketing campaign, the bank recognizes the potential for substantial improvement through more targeted efforts.

To achieve this objective, AllLife Bank plans to develop an advanced AI/ML-powered predictive model with the following key goals:

- **Customer Data Analysis:** Utilize comprehensive customer data to identify critical attributes and behaviors associated with higher personal loan conversion potential.
- **Enhanced Targeting:** Precisely target liability customers who demonstrate the highest propensity for personal loan uptake.
- **Resource Optimization:** Strategically allocate marketing resources to maximize efficiency and impact.
- **Campaign Effectiveness:** Increase conversion rates and optimize the return on marketing investment.

This initiative underscores AllLife Bank’s dedication to data-driven decision-making and its proactive strategy to expand the loan business and enhance overall profitability.

Business Problem Overview and Solution Approach

Recommendations to enhance AllLife Bank's personal loan targeting strategy

Prioritize Customers with High CCAvg

- **Insight:** There is a strong positive correlation between credit card average spending per month (CCAvg) and personal loan acceptance. Only 6.4% of customers spend over the average CCAvg per month.
- **Action:** Focus marketing efforts on customers with higher CCAvg. Develop targeted campaigns and offers specifically designed for this segment. Consider offering premium loan products or higher loan amounts to incentivize these customers.

Target Customers with CD Accounts

- **Insight:** Customers with CD accounts have a significantly higher personal loan acceptance rate (46.4%) compared to those without (7.2%).
- **Action:** Develop targeted campaigns specifically for customers with CD accounts. Leverage the existing relationship and offer personalized loan solutions based on their CD account details. Consider offering preferential interest rates or other benefits to incentivize loan uptake.

Business Problem Overview and Solution Approach

Recommendations to enhance AllLife Bank's personal loan targeting strategy

Focus on Larger Family Sizes

- **Insight:** Customers with larger family sizes (3 and 4) are more likely to accept personal loans.
- **Action:** Tailor marketing messages and loan offers to appeal to the needs and financial priorities of larger families. Highlight the benefits of personal loans for family-related expenses, such as education, home improvement, or travel.

Optimize Online Banking Channel

- **Insight:** Customers who use online banking have a slightly higher personal loan acceptance rate (9.75%) compared to those who do not (9.37%).
- **Action:** Enhance the online banking platform to promote personal loan offers and provide easy access to loan applications. Develop targeted online advertisements and personalized loan recommendations for online banking users.

Business Problem Overview and Solution Approach

Recommendations to enhance AllLife Bank's personal loan targeting strategy

Leverage Securities Account Information

- **Insight:** Customers with securities accounts have a slightly higher personal loan acceptance rate (11.5%) compared to those without (9.4%).
- **Action:** Explore opportunities to cross-sell personal loans to customers with securities accounts. Develop targeted offers and highlight the benefits of using personal loans for investment purposes.

De-emphasize Age as a Targeting Factor

- **Insight:** Age alone is not a strong determining factor for loan interest.
- **Action:** While age can be a factor in loan eligibility, avoid over-reliance on age as a primary targeting criterion. Focus on other factors, such as CCAvg, CD account ownership, and family size, which have stronger correlations with loan acceptance.

Business Problem Overview and Solution Approach

Recommendations to enhance AllLife Bank's personal loan targeting strategy

Refine Targeting Based on ZIP Code

- **Insight:** ZIP code prefixes provide insights into customer demographics and financial behavior. Analysis reveals the following loan acceptance rates by ZIP code prefix:
  - Highest Acceptance: 93: 10.31%
  - Medium Acceptance: 91: 9.73%, 95: 9.82%
  - Lowest Acceptance: 96: 7.5%

- **Action:**
  - **Prioritize:** Focus initial marketing efforts on customers residing in ZIP code prefix 93, where the loan acceptance rate is highest.
  - **Secondary Focus:** Target customers in ZIP code prefixes 91 and 95, which exhibit medium acceptance rates.
  - **Tailored Campaigns:** Develop marketing campaigns and loan offers tailored to the specific demographics and financial characteristics of customers within each targeted ZIP code prefix.

Business Problem Overview and Solution Approach

Recommendations to enhance AllLife Bank's personal loan targeting strategy

Target Customers with Professional and Graduate Degrees

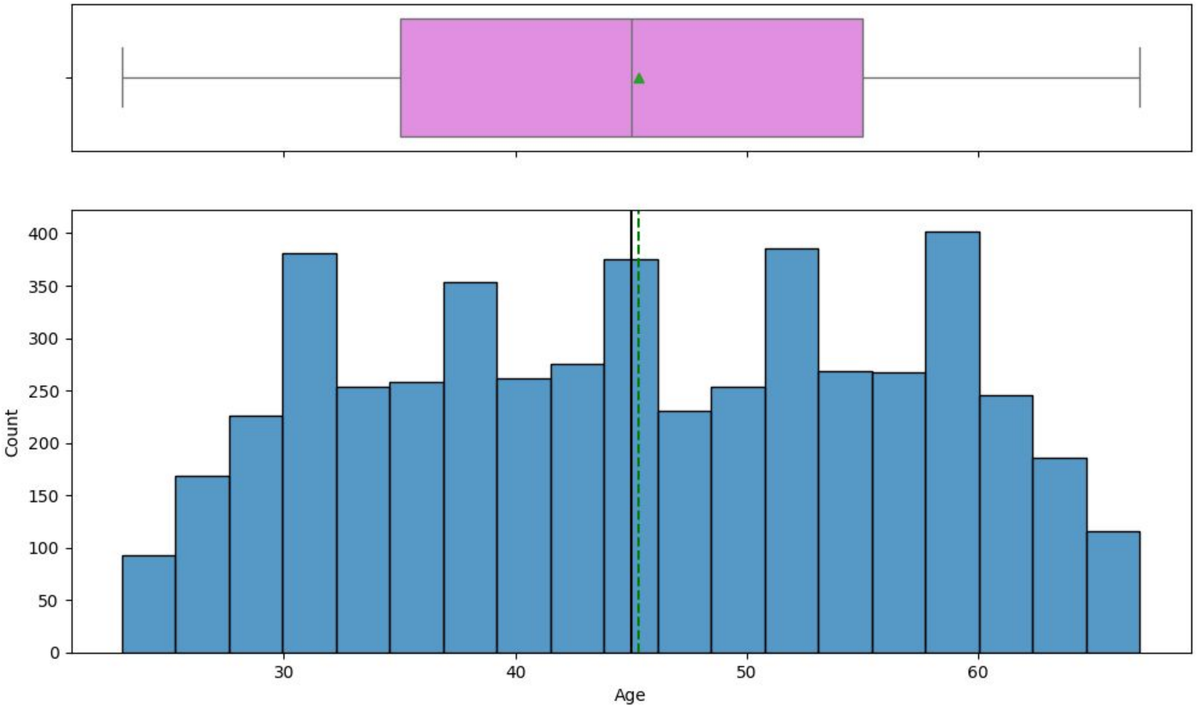
- **Insight:** Customers with Professional and Graduate degrees demonstrate significantly higher loan acceptance rates (13.7% and 13.0%, respectively) compared to those with Undergraduate degrees (4.4%).
- **Action:** Develop targeted marketing campaigns specifically for customers with Professional and Graduate degrees. Emphasize loan products and benefits that align with their financial goals and professional aspirations.

Monitor and Adapt

- **Insight:** The financial landscape and customer behavior are constantly evolving.
- **Action:** Continuously monitor the performance of marketing campaigns and loan products. Analyze customer data and feedback to identify new trends and adjust targeting strategies accordingly.

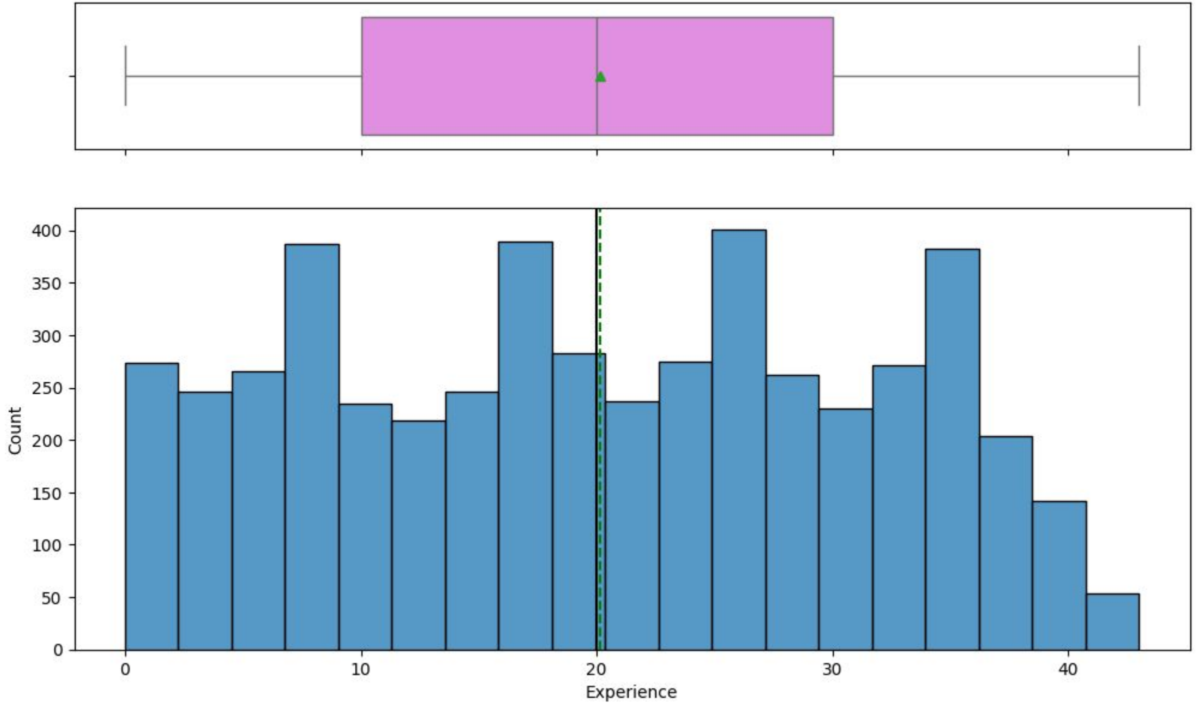
EDA Results

- **45** is the average age of **customer** with the minimum age is 23 years of age and maximum age of 67



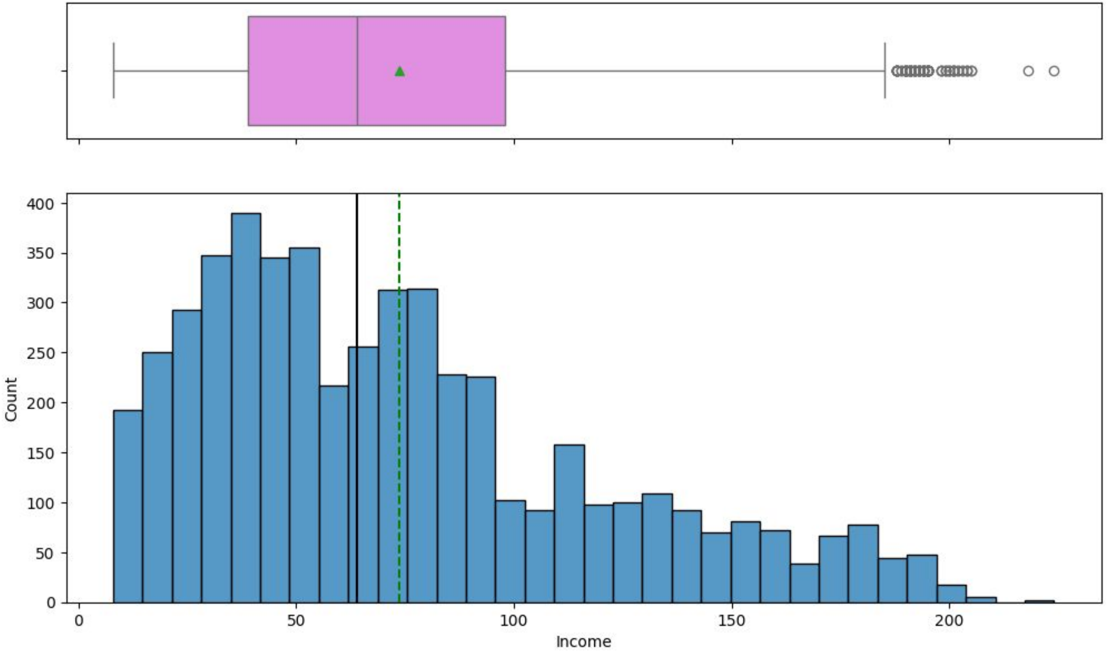
EDA Results

- **Experience** of 20 years is the average, with 43 years of experience is the maximum and 0 years of experience is the minimum



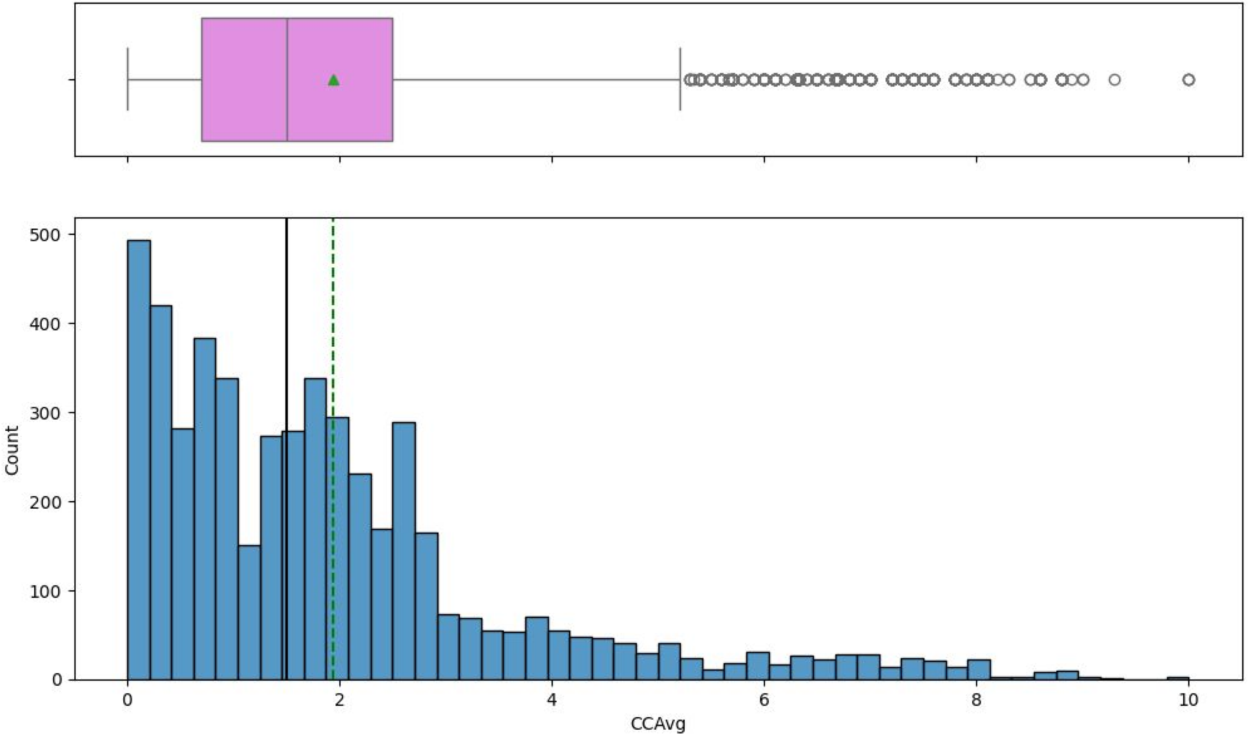
EDA Results

- **Income** is right **skewed with some outliers.** The minimum income is $8000 with the average $64000. The outlier are with income over $180000. The maximum outlier is $224000



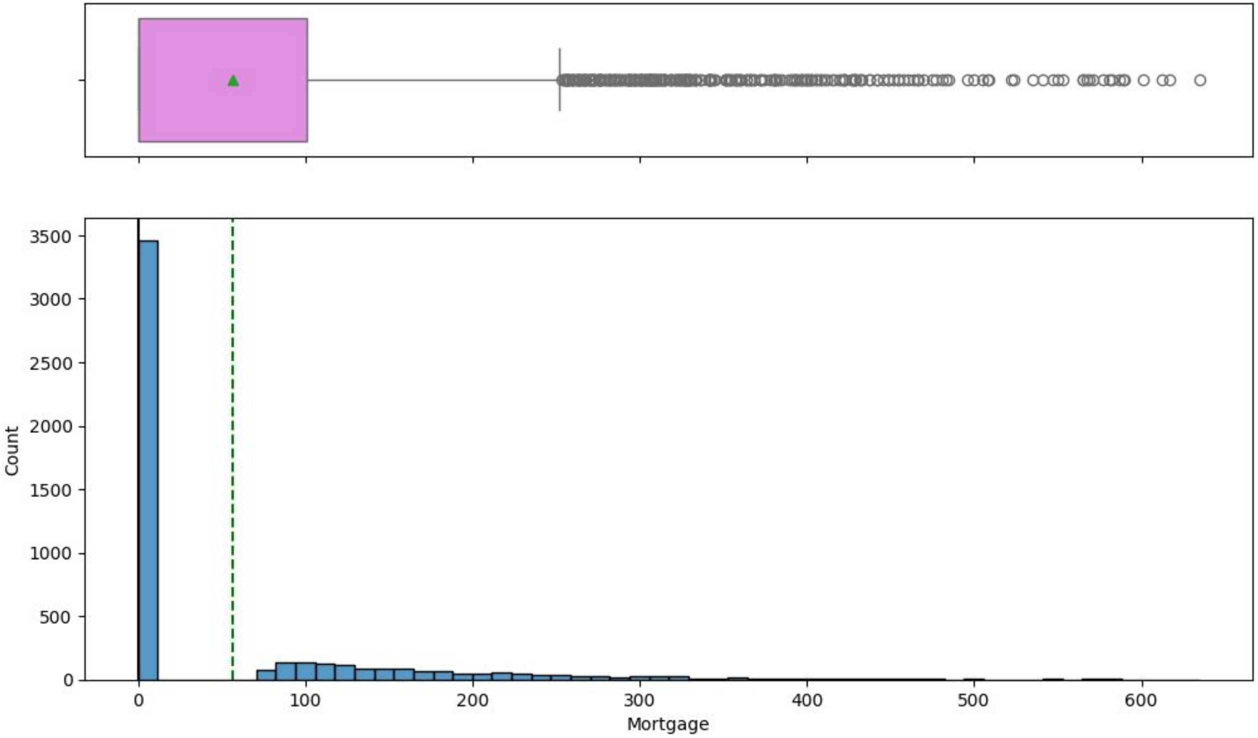
EDA Results

- Credit card spending (CCAvg) averages $1,500 per month, with a right-skewed distribution indicating a concentration of lower spending customers and a smaller segment with significantly higher spending, ranging from $0 (representing non-users) to $10,000.



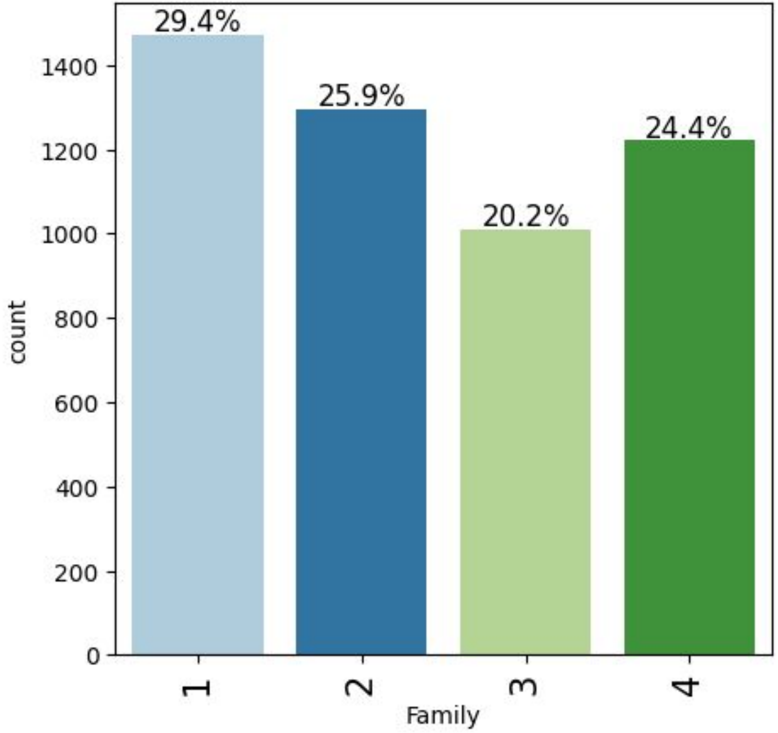
EDA Results

- The Mortgage data is heavily skewed to the right, indicating that most customers have low or no mortgages, with a significant number of outliers representing customers with substantial mortgage debt. **Outliers start at 250K, to max 634K**



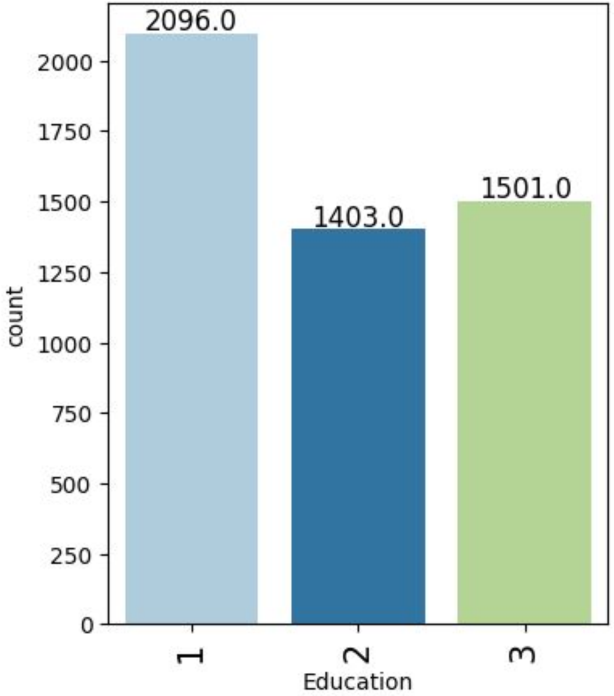
EDA Results

- The chart shows a relatively even distribution of **Family sizes** among customers, with single-person households being the most prevalent. While single person households have the largest percentage, there is a strong presence of families of 2, 3 and 4 as well.



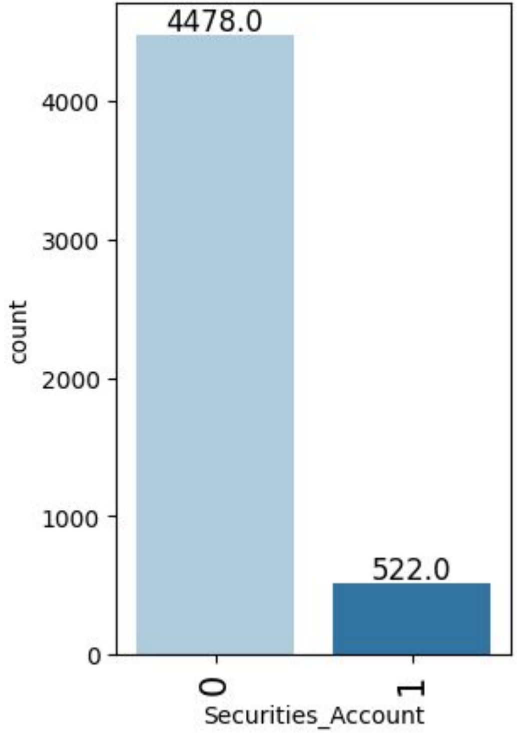
EDA Results

- The **Education** graph indicates that education level 1 is the most prevalent among the observed group, with significantly higher counts than levels 2 and 3.



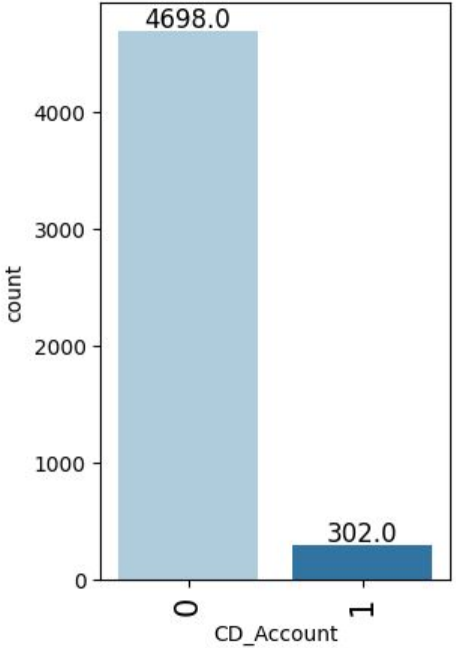
EDA Results

- The vast majority of customers do not have a **securities account**, as indicated by the significantly higher count for the "0" category compared to the "1" category.



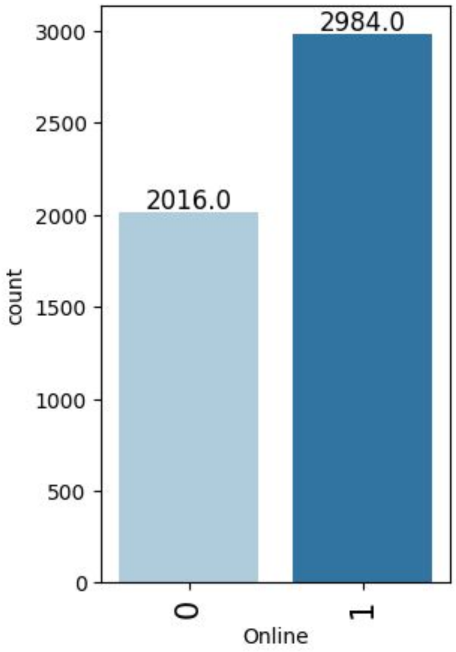
EDA Results

- Analysis of customer product holdings reveals a significant disparity in the adoption of **Certificate of Deposit (CD) accounts**. Only 302 customers hold a CD account, compared to 4,698 customers who do not. This indicates a low penetration rate for CD products within the customer base.



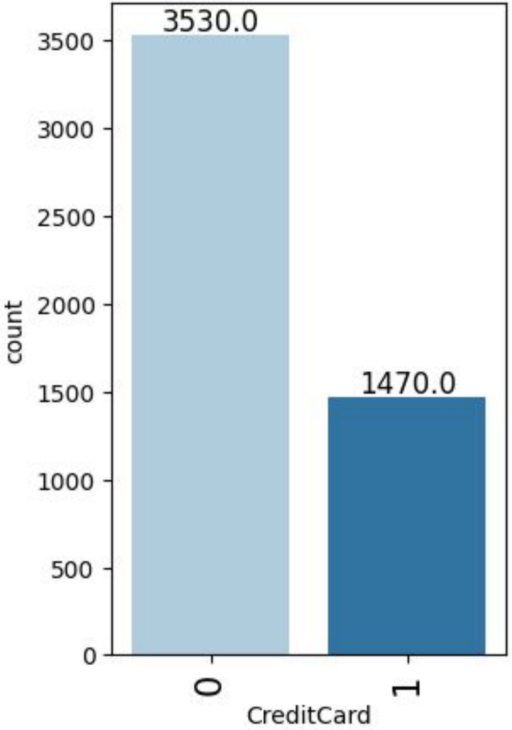
EDA Results

- A significantly higher number of customers utilize **Online Banking** facilities compared to those who do not, with approximately **2984 customers** using online banking and **2016 customers** not using it.



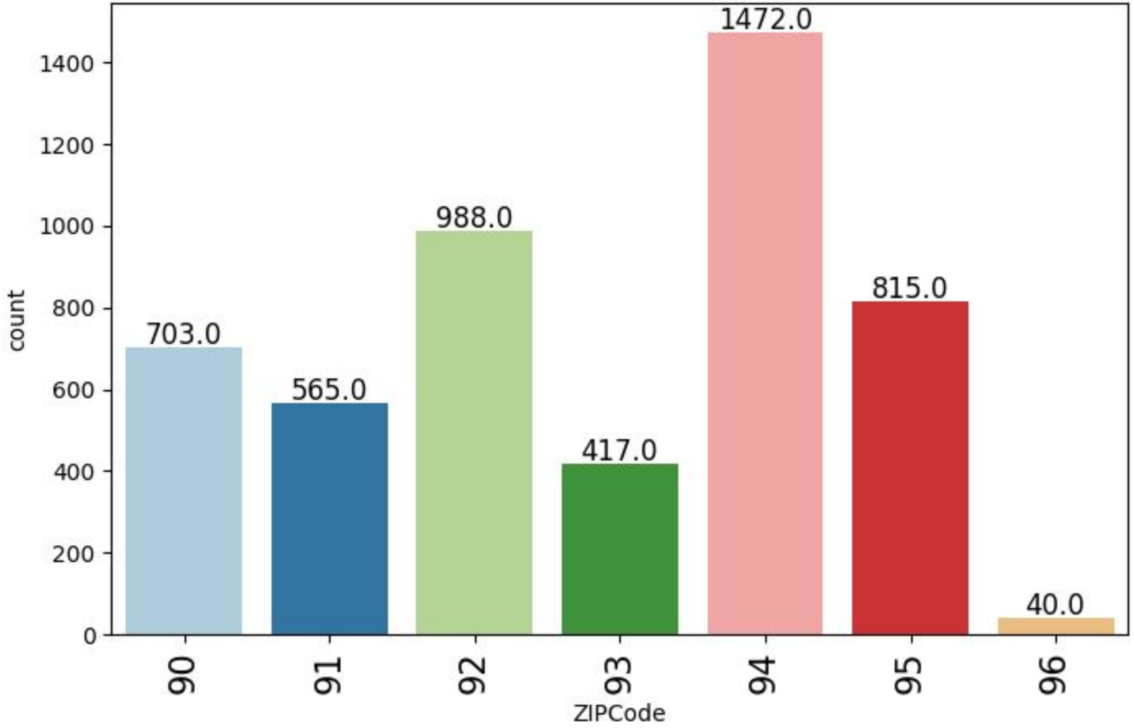
EDA Results

- **3530 Customer** use another **Credit Card** from another bank compared to only **1470 customer** who does not use another bank credit card.



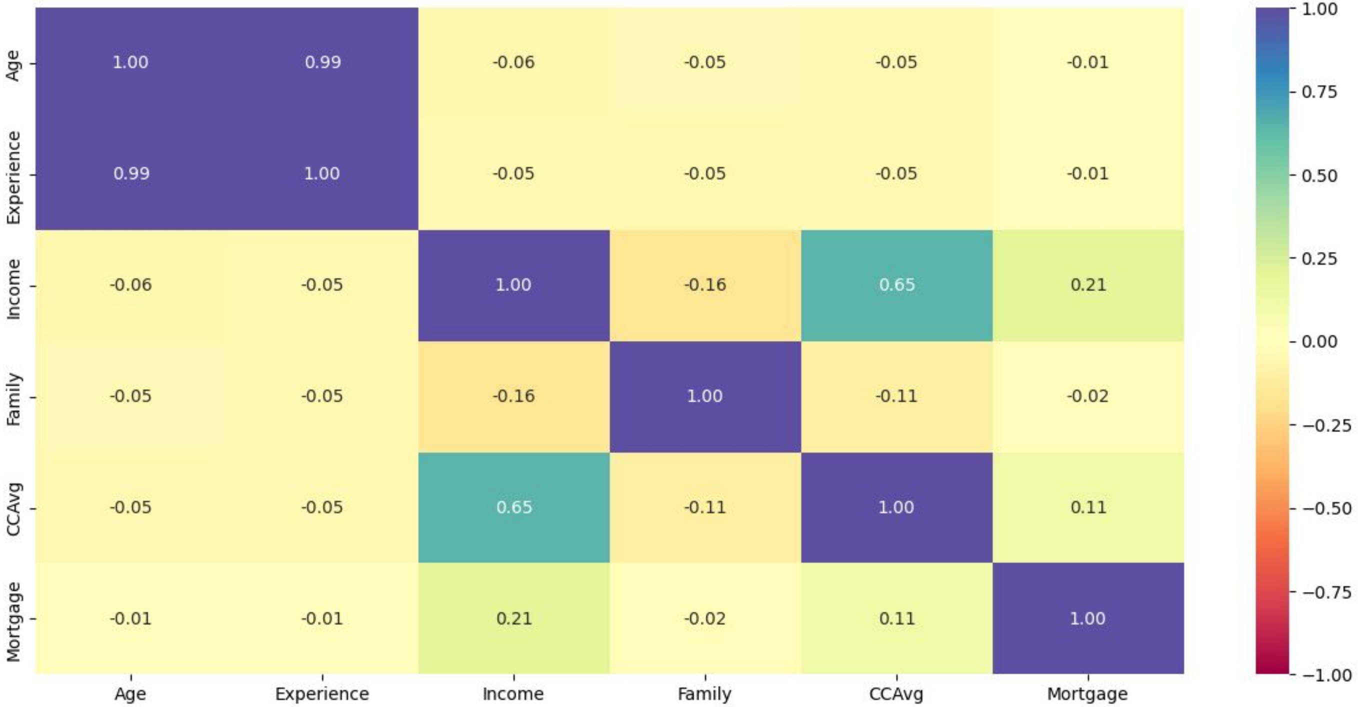
EDA Results

- There are **7** unique **Zip Code Prefixes** where the Customer lives. Zip Code prefix 94 is where the vast majority lives.



EDA Results Bi Rate

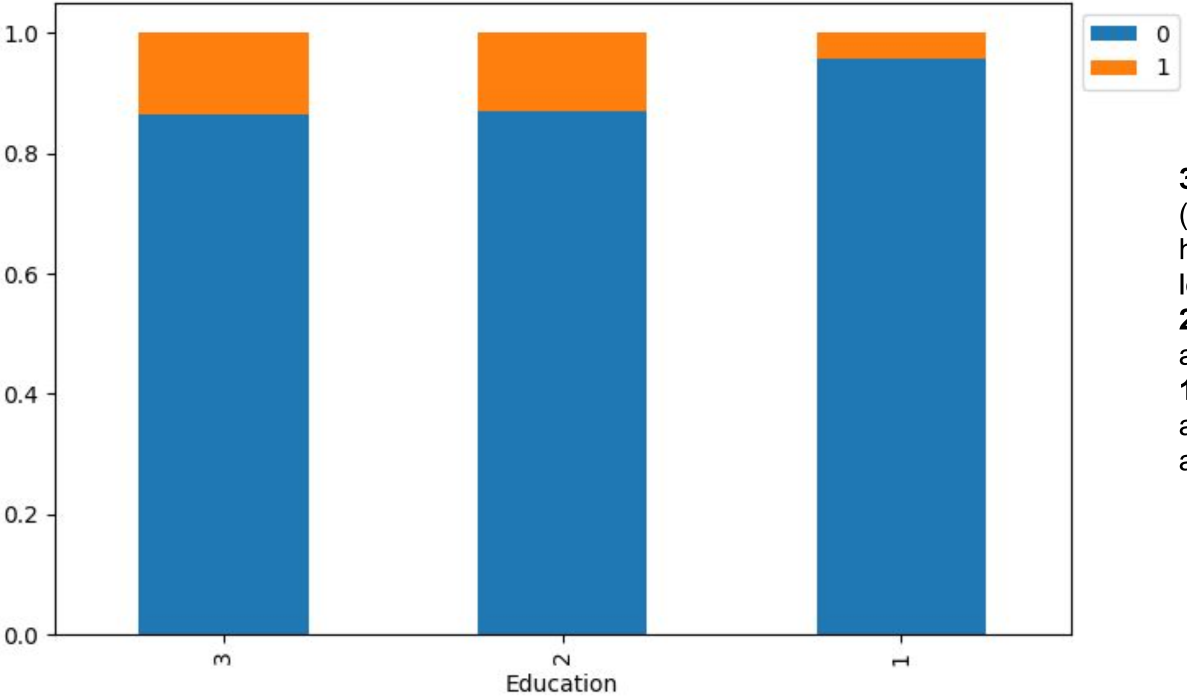
The strongest correlation is observed between **Age and Experience (0.99)**, followed by **Income and Credit Card Spending Average (0.65)**. Other correlations are either weak or negligible, indicating that these variables do not strongly influence each other. Understanding these relationships can help in decision-making related to financial behaviors, credit risk assessments, and marketing strategies



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

EDA Results Bi Rate

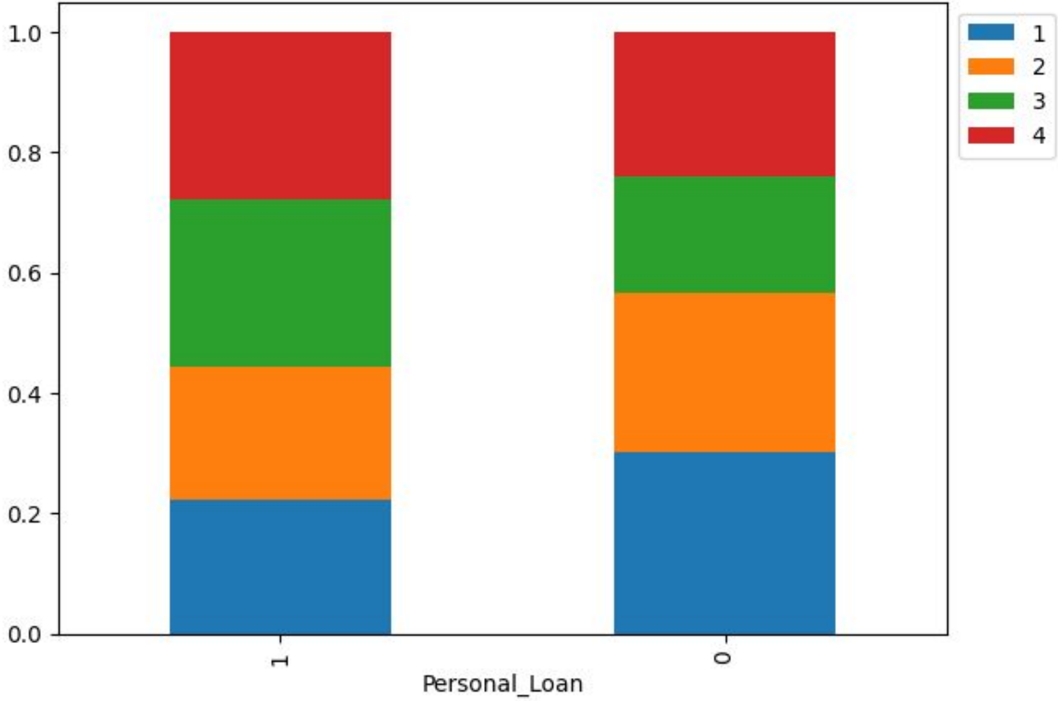
There is a **positive correlation** between **Education Level and Loan Acceptance**. Customers with higher education levels, particularly those with **professional degrees**, are more likely to **accept personal loan** offers. This insight could be valuable for targeted marketing campaigns and risk assessment



3 (Professional Degree): 205 out of 1501 (13.7%) accepted the loan. This is the highest acceptance rate among the education levels.
2 (Graduate): 182 out of 1403 (13.0%) accepted the loan.
1 (Undergraduate): 93 out of 2096 (4.4%) accepted the loan. This is the lowest acceptance rate.

EDA Results Bi Rate

Larger family sizes correlate with a higher probability of loan acceptance. The numerical data reveals that families of size 3 and 4 have a higher acceptance rate than families of size 1 and 2.

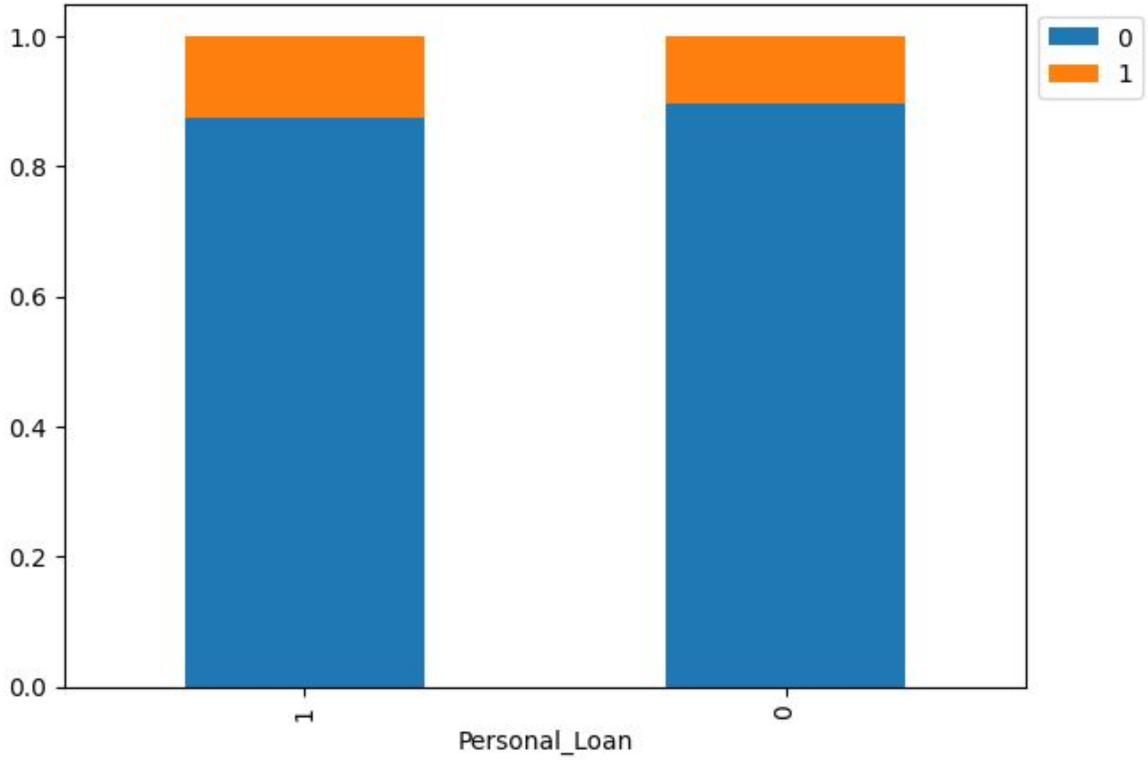


Loan Acceptance by Family Size:

- **Family 1:** 107/1472 (7.3%) accepted the loan.
- **Family 2:** 106/1296 (8.2%) accepted the loan.
- **Family 3:** 133/1010 (13.2%) accepted the loan.
- **Family 4:** 134/1222 (11.0%) accepted the loan.

EDA Results Bi Rate

The presence of a **Securities Account** shows a very slight positive correlation with **Personal Loan acceptance**. Overall **acceptance rate remains low.**

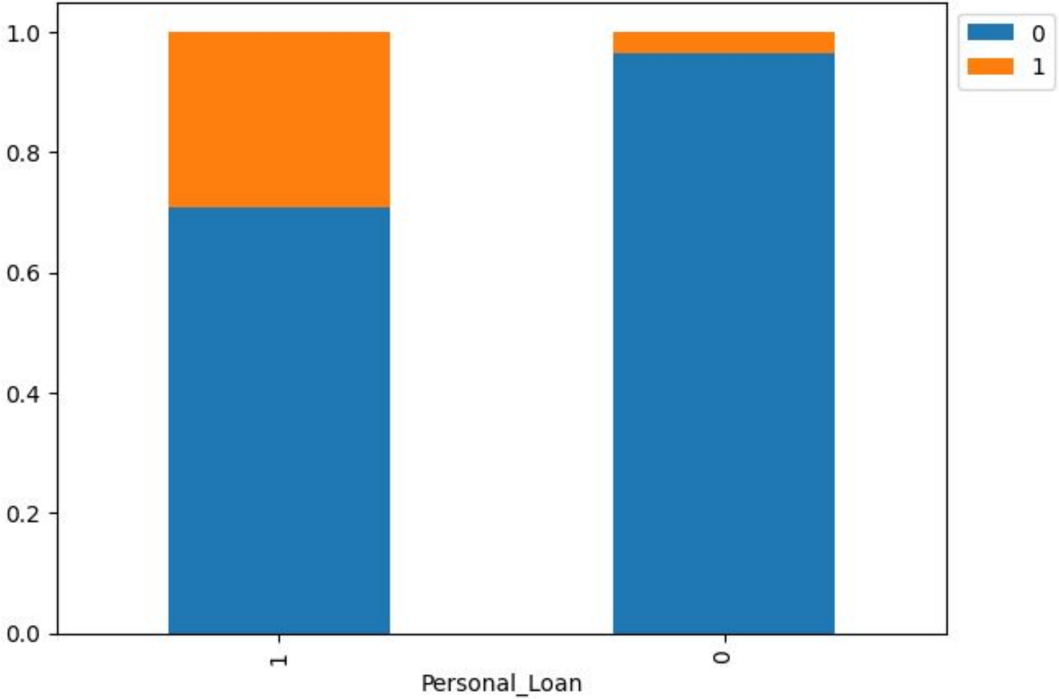


Loan Acceptance by Securities Account:

- **Customers with Securities Accounts:**
  - 60 out of 522 (11.5%) accepted the personal loan.
  - 462 out of 522 (88.5%) did not accept the personal loan.
- **Customers without Securities Accounts:**
  - 420 out of 4478 (9.4%) accepted the personal loan.
  - 4058 out of 4478 (90.6%) did not accept the personal loan.

EDA Results Bi Rate

Customers who have a CD account are significantly more likely to accept a personal loan offer. There is a strong positive correlation between having a CD account and accepting a personal loan. This is a very important finding.

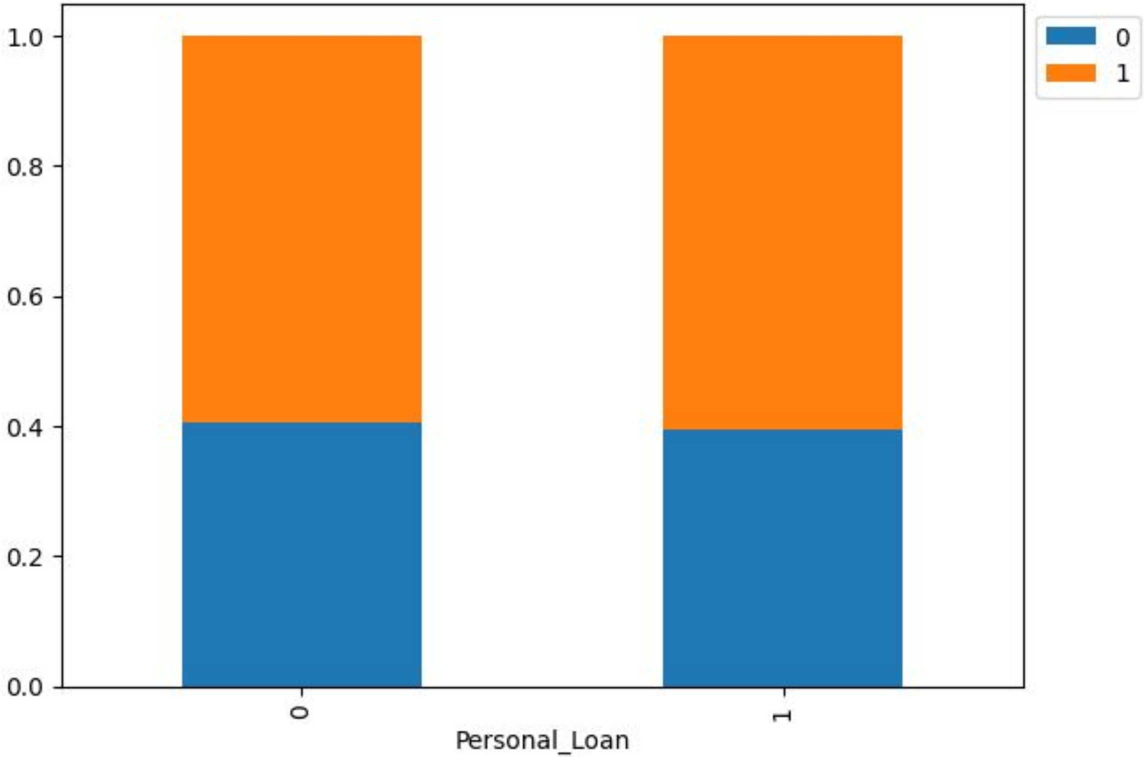


Loan Acceptance by CD Account:

- **Customers with CD Accounts:**
  - 140 out of 302 (46.4%) accepted the personal loan.
  - 162 out of 302 (53.6%) did not accept the personal loan.
- **Customers without CD Accounts:**
  - 340 out of 4698 (7.2%) accepted the personal loan.
  - 4358 out of 4698 (92.8%) did not accept the personal loan.

EDA Results Bi Rate

Online banking usage has a **very weak** positive **correlation** with **Personal Loan** acceptance.

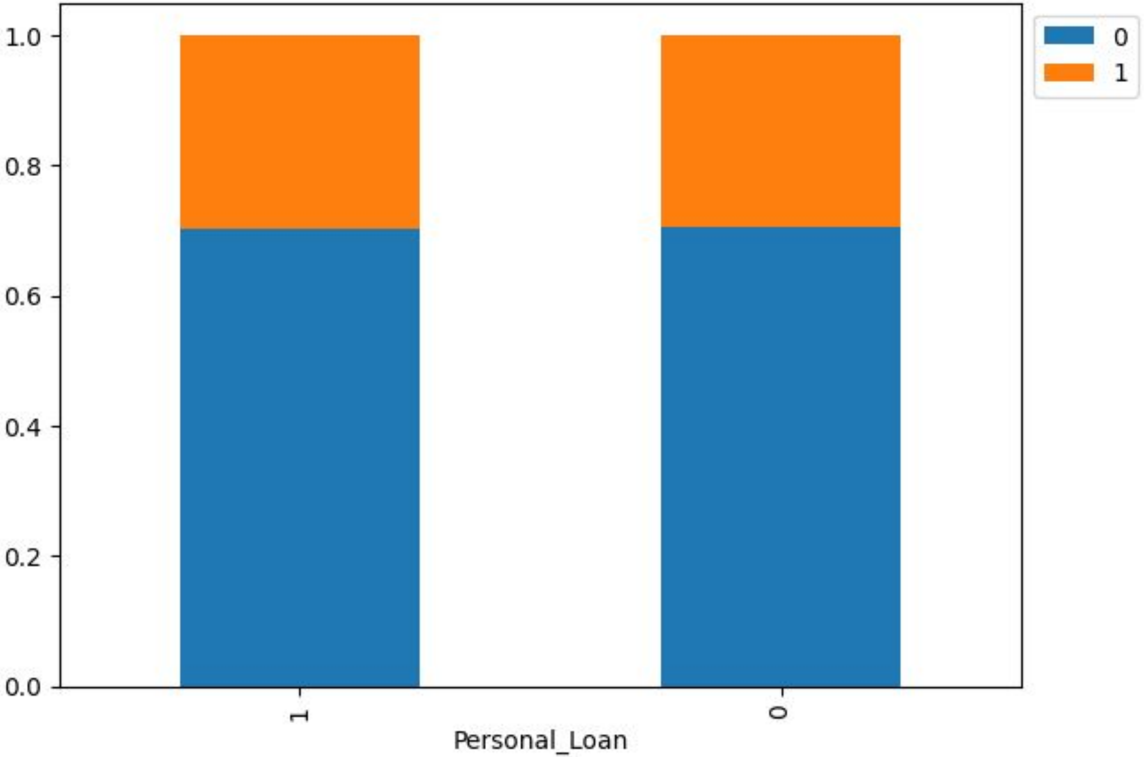


Loan Acceptance by Online Banking Usage:

- **Customers Using Online Banking:**
  - 291 out of 2984 (9.75%) accepted the personal loan.
  - 2693 out of 2984 (90.25%) did not accept the personal loan.
- **Customers Not Using Online Banking:**
  - 189 out of 2016 (9.37%) accepted the personal loan.
  - 1827 out of 2016 (90.63%) did not accept the personal loan.

EDA Results Bi Rate

Customers who use a **Credit Card from another bank** are slightly more likely to accept a **Personal Loan** offer. Indicates a **very weak positive correlation** with personal loan acceptance.

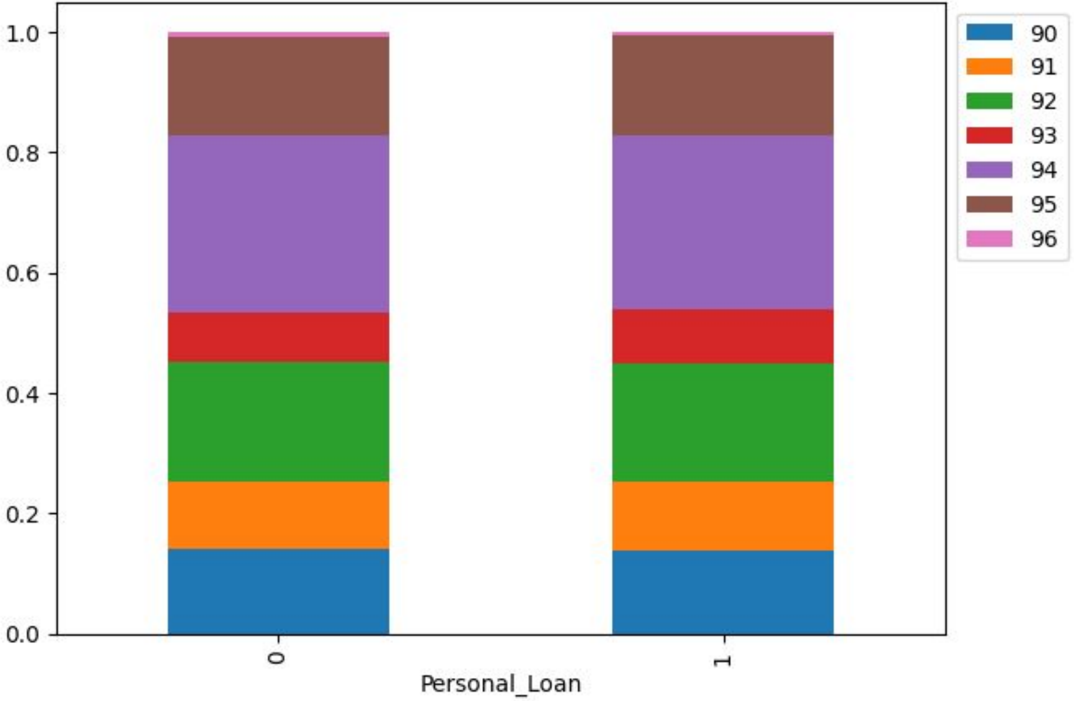


Loan Acceptance by Credit Card Usage (Other Bank):

- **Customers with Credit Card (Other Bank):**
  - 143 out of 1470 (9.73%) accepted the personal loan.
  - 1327 out of 1470 (90.27%) did not accept the personal loan.
- **Customers without Credit Card (Other Bank):**
  - 337 out of 3530 (9.55%) accepted the personal loan.
  - 3193 out of 3530 (90.45%) did not accept the personal loan.

EDA Results Bi Rate

Personal Loan acceptance rates across different **ZIP Code Prefixes** are quite similar, ranging from 7.5% to 10.31%. Acceptance of the l**oan is not influence** by the **zip code** prefix.



Here's a breakdown of loan acceptance rates (percentage of customers who accepted the loan) for each ZIP code prefix:

- 90: 9.53%
- 91: 9.73%
- 92: 9.52%
- 93: 10.31%
- 94: 9.38%
- 95: 9.82%
- 96: 7.5%

EDA Results Bi Rate Age vs Personal Loan

Age Distribution Similarity:

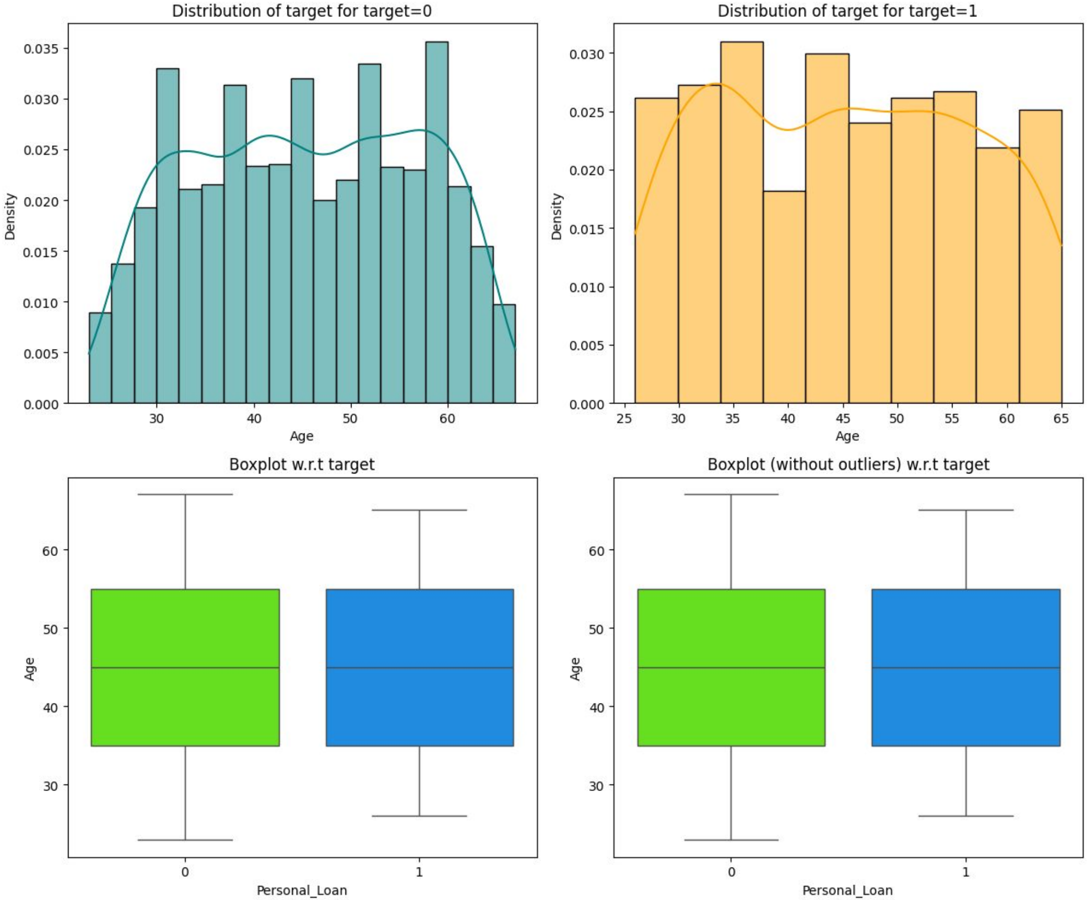
- The age distributions for customers who purchased a loan and those who did not are quite similar. This suggests that age alone is not a strong predictor of loan purchase interest.

Slight Trend Towards Older Customers:

- There is a subtle trend indicating that older customers may be slightly more likely to purchase a loan. This is visible in the right histogram and its density plot.

Outliers:

- The outliers in the box plots, indicate that there are some people of very young or very old age in both groups.



Distribution of target for target=0

Distribution of target for target=1

Boxplot w.r.t target

Boxplot (without outliers) w.r.t target

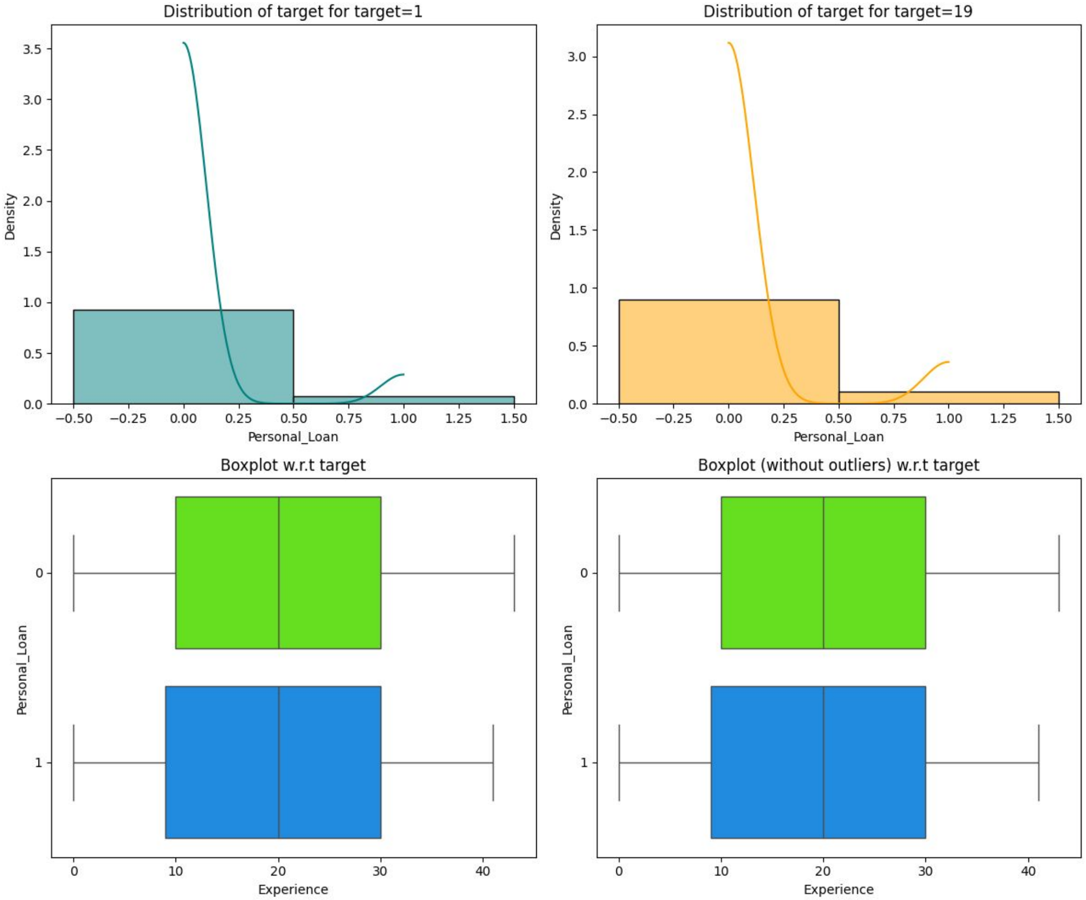
EDA Results

Personal Loan vs Experience

Key Insight: Both distributions are skewed towards lower experience, but those that accepted the loan had slightly higher experience on average

Experience and Loan Acceptance:

- Customers with slightly higher professional experience were more likely to accept the personal loan.



Distribution of target for target=1

Distribution of target for target=19

Boxplot w.r.t target

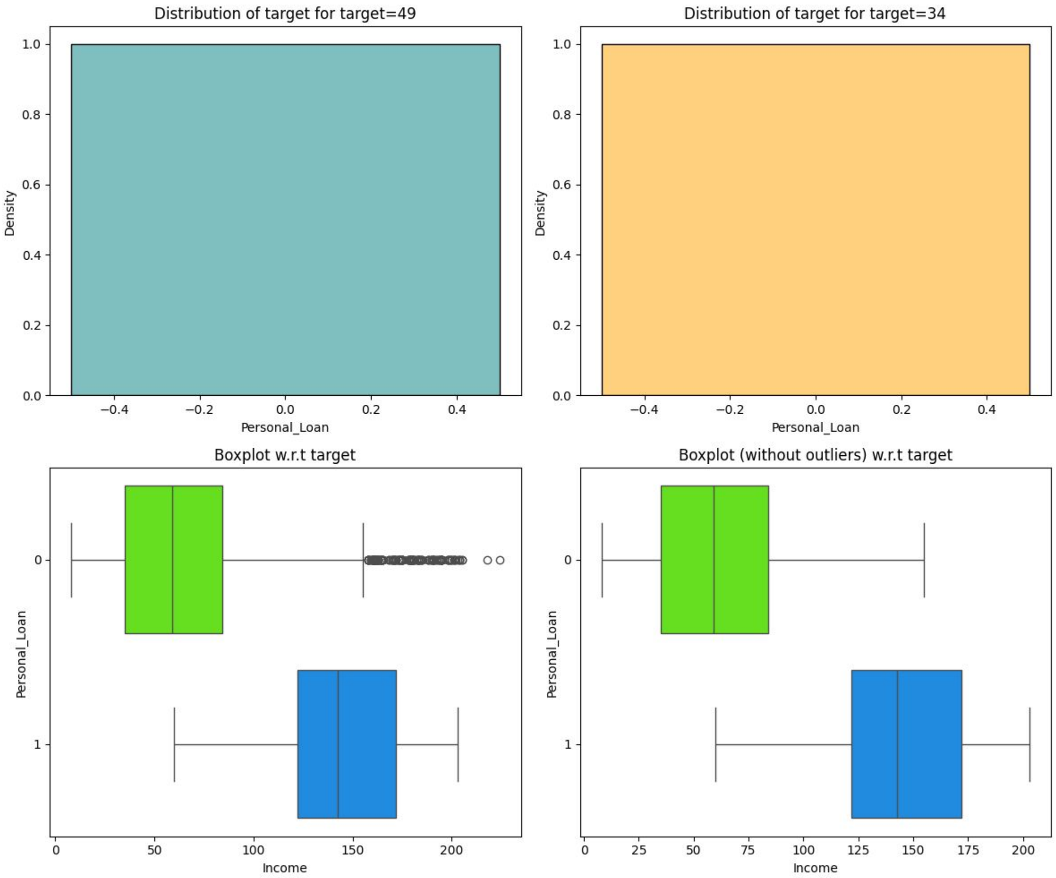
Boxplot (without outliers) w.r.t target

EDA Results Bi Rate

Personal Loan vs Income

Key Observations and Interpretations:

1. **Income and Loan Acceptance:**
   - Customers with significantly higher incomes were much more likely to accept the personal loan.
2. **Clear Separation:**
   - The boxplots show a clear separation between the income distributions of the two groups.
3. **Strong Correlation:**
   - This indicates a strong positive correlation between income and personal loan acceptance.
4. **Top Graphs:**
   - The top graphs displays the personal loan binary data


Distribution of target for target=49

Distribution of target for target=34

Boxplot w.r.t target

Boxplot (without outliers) w.r.t target

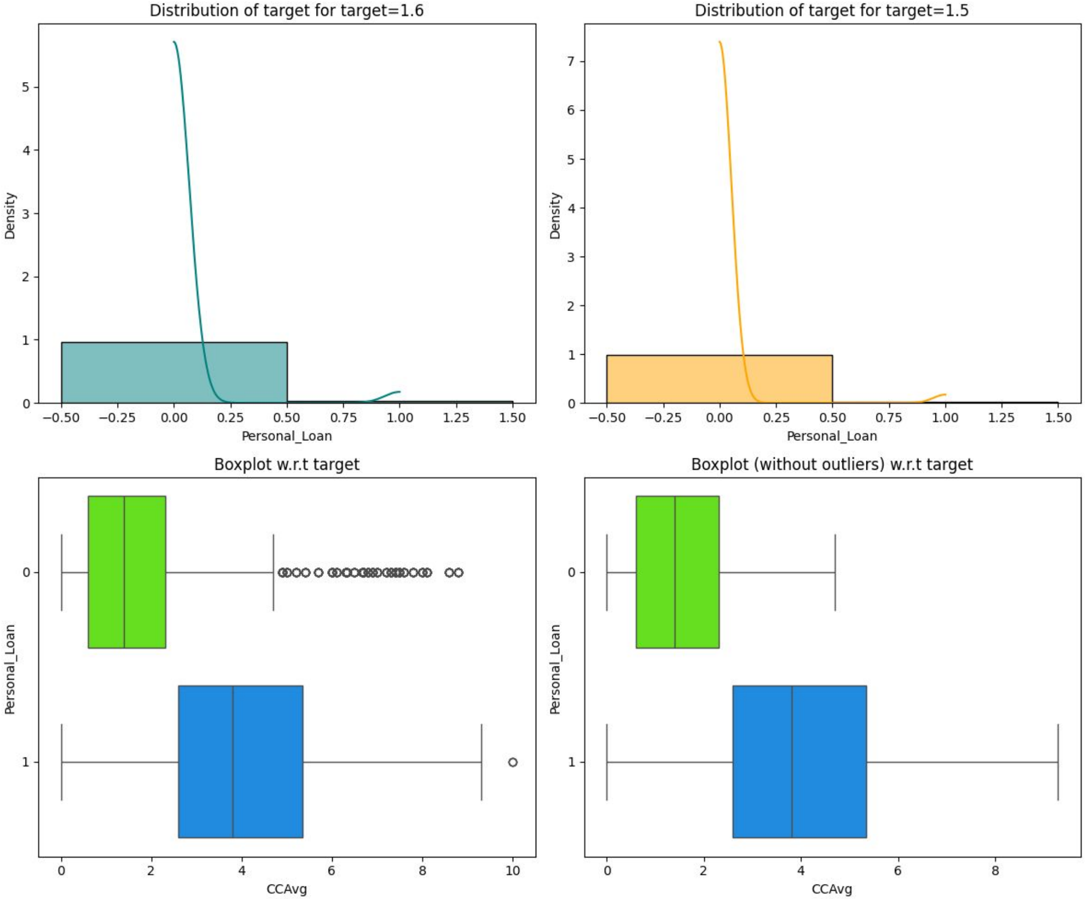
EDA Results Bi Rate

Personal loan vs CCAvg

- There is a **strong positive relationship** between **Credit Card Average Spending** per month (**CCAvg**) and **Personal Loan** acceptance.

- Customers with higher CCAvg are significantly more likely to accept personal loan offers.

- **CCAvg** is a strong indicator of personal loan acceptance.



Data Preprocessing

- **Are there any duplicate data**:
  - No Duplicate data exists
- **Missing value treatment**
  - No Missing value exists in the dataset. An effective check was performed for verification

Data Preprocessing

Percentage of **Outliers** in each numerical feature of the dataset that found data outside the upper and lower bounds.

Feature	**Percentage of Outliers**
Income	1.92
CCAvg	6.48
Mortgage	5.82

Data Preprocessing

Comprehensive Data Quality Check was conducted on the **Experience** variable, resulting in the identification and correction of 52 data integrity issues. Specifically, negative values within the **Experience** variable were rectified by converting them to positive values, ensuring data accuracy and consistency.

Experience	Count
-1	33
-2	15
-3	4

Data Preprocessing

- **Feature Engineering Zip Code Prefix Extraction**
    - To reduce dimensionality and potentially improve model performance, feature engineering was applied to the ZIP Code variable. Initially, the dataset contained **467 unique** ZIP Code **combinations.**
    - **To address this high cardinality,** a prefix extraction technique was employed, where the **first two digits of each ZIP Code were extracted** to create a new ZIP Code Prefix feature.
    - This resulted in a **total** of **7 unique prefixes,** effectively reducing the dimensionality of the ZIP Code information.

Data Preprocessing

To **optimize** data storage and facilitate efficient **categorical data handling**, the data types of the following features were explicitly **converted** to the **'category'** data type

● Education	● Online
● Personal_Loan	● CreditCard
● Securities_Account	● ZIPCode
● CD_Account	

This conversion enables more memory-efficient representation of categorical variables and can enhance the performance of subsequent analytical and modeling operations.

Data Preprocessing

Feature Selection:

- Dropping Redundant Features: The Experience variable was removed due to perfect correlation with Age, eliminating redundancy.

Data Transformation:

- One-Hot Encoding: Categorical features ZIP Code and Education were one-hot encoded to enable their use in numerical models.
- Data Type Conversion: All features were converted to float format for consistency.

Data Splitting: The dataset was split into training and testing sets using a 70/30 ratio with a random state of 1 to ensure reproducibility.

Model Building

Three different **Decision Tree Classifier models** was built and performance evaluated. The first model built was a **Default model** without setting the minimum samples per leaf**,** maximum leaf nodes, or depth constraints.

1. To predict loan purchase
2. Prior to building a model a Clean dataset is required, Feature Engineer Categories and perform encoding categorical features as necessary
3. Partition dataset into train and test sets to facilitate model training and evaluation, ensuring robust training on the train dataset
4. Construct a DecisionTreeClassifier model utilizing Gini impurity as the criterion for assessing node impurity and guiding the decision-making process
5. Performance Evaluation is done on the train set and test set to verify model accuracy, recall, precision and F1 score.

Model Building

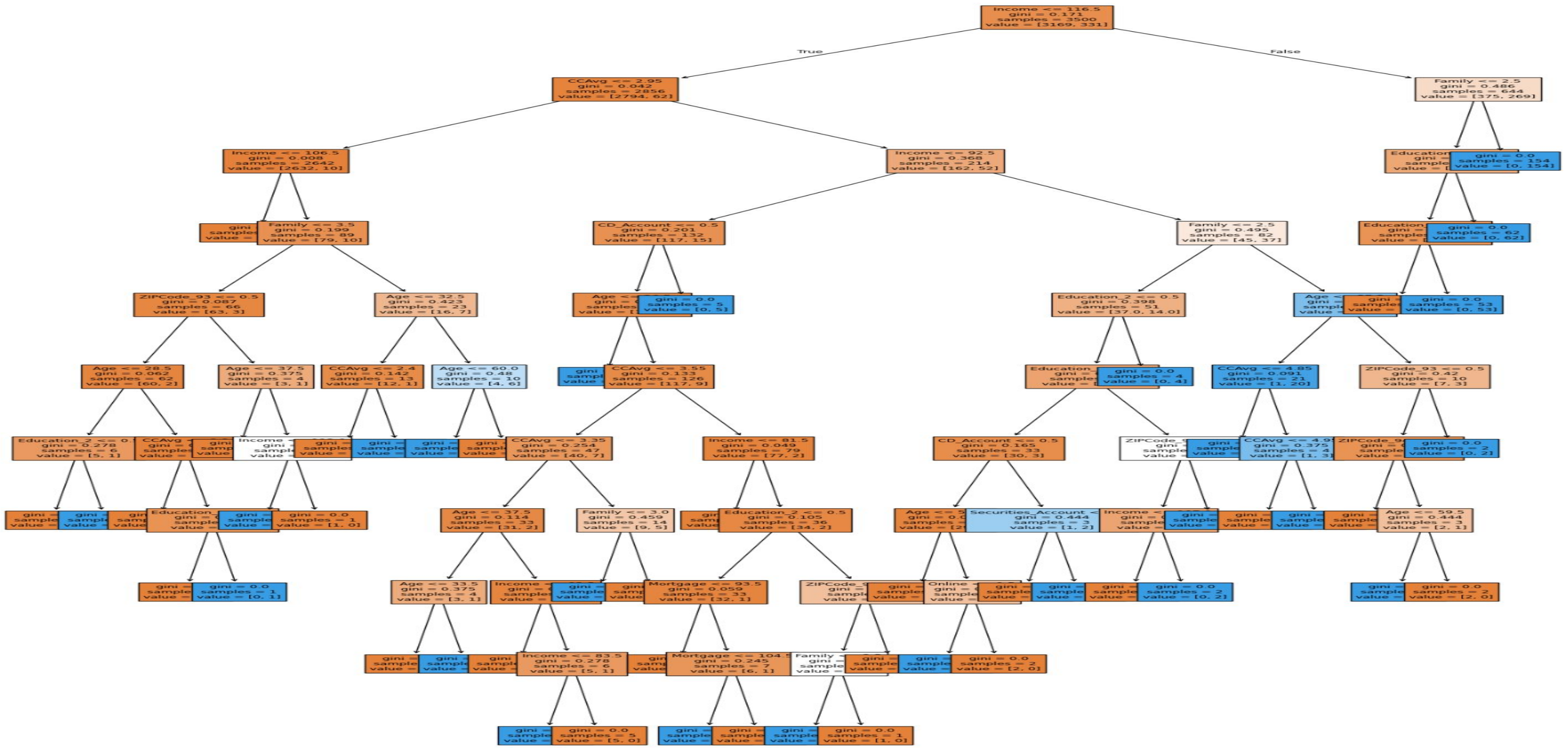
- **Pre Prune Model**
  - Define parameter such as **maximum depth**, **maximum leaf node** and **minimum sample split** values, **weight** with a random state
  - The decision tree classifier model will iterate over the data looking for **the best** score based off the **recall score metrics**.
  - After finding the best algorithm we use that to Train the model on the data and check correlation with the **Test Data** to make **prediction.**
  - Finally check the Performance on the model

Model Building

- **Post Pruning** Decision Tree Classifier **model building**
    - **Compute Cost Complexity** pruning path for the model using the training data
    - After computing cost complexing and verify the alphas, we **training the model** on the **training data set** using the best cost complexity value with a random state so the model can generalize
    - After training the model we can make predictions on the test data and compute the recall score to determine model performance.

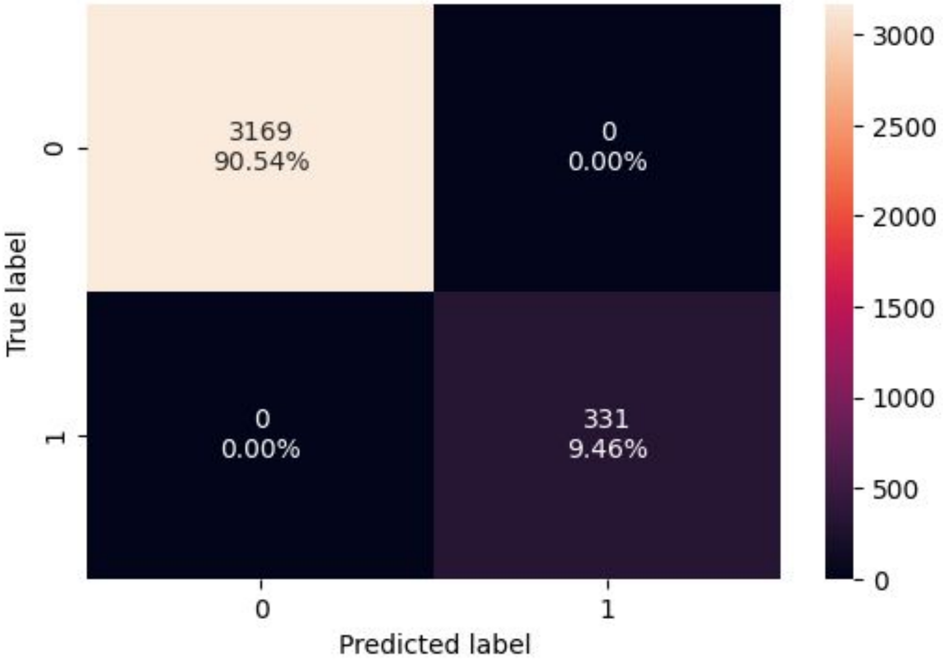
Model Building

Default Model performance built with only Gini Impurity the Tree is very complex with many levels


Income <= 116.5 gini = 0.171 samples = 3500 value = [3169, 331]
True
False
CCAvg <= 2.95 gini = 0.042 samples = 2856 value = [2794, 62]
Family <= 2.5 gini = 0.486 samples = 644 value = [375, 269]
Income <= 106.5 gini = 0.008 samples = 2642 value = [2632, 10]
Income <= 92.5 gini = 0.368 samples = 214 value = [162, 52]
gini = 0.0 samples = 154 value = [0, 154]
Family <= 3.5 gini = 0.199 samples = 89 value = [79, 10]
CD_Account <= 0.5 gini = 0.201 samples = 132 value = [117, 15]
Family <= 2.5 gini = 0.495 samples = 82 value = [45, 37]
gini = 0.0 samples = 62 value = [0, 62]
ZIPCode_93 <= 0.5 gini = 0.087 samples = 66 value = [63, 3]
Age <= 32.5 gini = 0.423 samples = 23 value = [16, 7]
gini = 0.0 samples = 5 value = [0, 5]
Education_2 <= 0.5 gini = 0.398 samples = 51 value = [37.0, 14.0]
gini = 0.0 samples = 53 value = [0, 53]
Age <= 28.5 gini = 0.062 samples = 62 value = [60, 2]
Age <= 37.5 gini = 0.375 samples = 4 value = [3, 1]
CCAvg <= 2.4 gini = 0.142 samples = 13 value = [12, 1]
Age <= 60.0 gini = 0.48 samples = 10 value = [4, 6]
CCAvg <= 3.55 gini = 0.133 samples = 126 value = [117, 9]
gini = 0.0 samples = 4 value = [0, 4]
CCAvg <= 4.85 gini = 0.091 samples = 21 value = [1, 20]
ZIPCode_93 <= 0.5 gini = 0.42 samples = 10 value = [7, 3]
Education_2 <= 0. gini = 0.278 samples = 6 value = [5, 1]
CCAvg <= 3.35 gini = 0.254 samples = 47 value = [40, 7]
Income <= 81.5 gini = 0.049 samples = 79 value = [77, 2]
CD_Account <= 0.5 gini = 0.165 samples = 33 value = [30, 3]
CCAvg <= 4.9 gini = 0.375 samples = 4 value = [1, 3]
gini = 0.0 samples = 2 value = [0, 2]
gini = 0.0 samples = 1 value = [1, 0]
Age <= 37.5 gini = 0.114 samples = 33 value = [31, 2]
Family <= 3.0 gini = 0.459 samples = 14 value = [9, 5]
Education_2 <= 0.5 gini = 0.105 samples = 36 value = [34, 2]
Securities_Account < gini = 0.444 samples = 3 value = [1, 2]
Age <= 59.5 gini = 0.444 samples = 3 value = [2, 1]
gini = 0.0 value = [0, 1]
Age <= 33.5 gini = 0.375 samples = 4 value = [3, 1]
Mortgage <= 93.5 gini = 0.059 samples = 33 value = [32, 1]
gini = 0.0 samples = 2 value = [0, 2]
gini = 0.0 samples = 2 value = [2, 0]
Income <= 83.5 gini = 0.278 samples = 6 value = [5, 1]
Mortgage <= 104.5 gini = 0.245 samples = 7 value = [6, 1]
gini = 0.0 samples = 2 value = [2, 0]
gini = 0.0 samples = 5 value = [5, 0]
gini = 0.0 samples = 1 value = [1, 0]

Model Building

Default Model performance built with Gini Impurity on Training Data. **Potential Overfitting**, perfect results are rare and might indicate overfitting.



Imbalance:

- The data is imbalanced, with a much larger number of instances belonging to class 0 (3169) compared to class 1 (331)

High Accuracy:

- The overall accuracy of the model is 100% because all predictions are correct

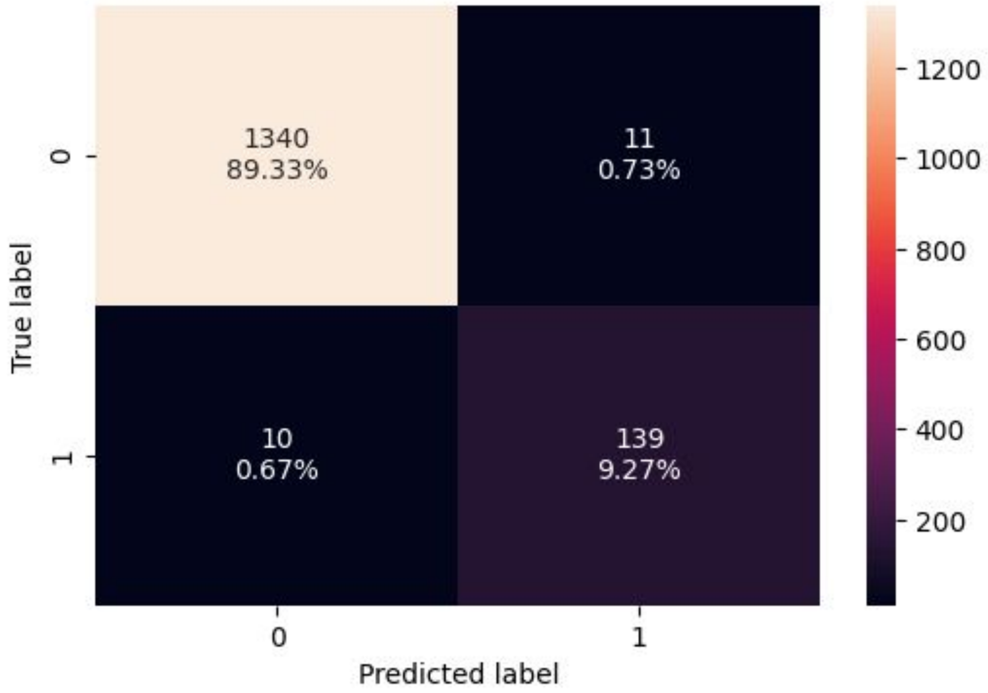
Recall: 100%

Precision: 100%

F1 Score: 100 %

Model Building

Default Model performance built with only Gini Impurity on the **Testing data.** While the data is imbalanced, the model is able to make reasonably accurate predictions.



Good Performance: The model performs well on the test data, with high accuracy and low error rates.

Accuracy:	**98%**
Recall:	**93%**
Precision:	**92%**
F1 Score:	**92%**

Model Building

To **optimize** model **performance** and **mitigate overfitting**, a **Pre-Pruning Decision Tree Classifier model strategy** was **implemented** with the following parameters

1. **Maximum Depth:** The maximum depth of the decision tree was restricted to prevent excessive complexity.

2. **Maximum Leaf Nodes:** A limit was placed on the number of leaf nodes to further control tree growth.

3. **Minimum Sample Split:** A minimum number of samples was required to split an internal node, ensuring that splits are based on sufficient data.

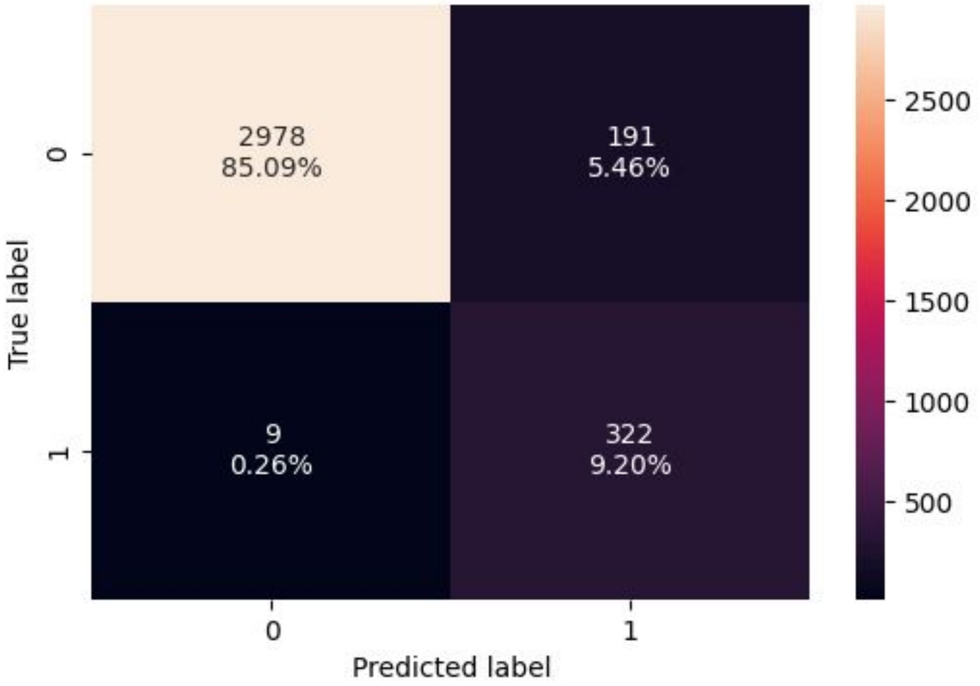
4. **Balanced Weights:** Class weights were automatically adjusted to address potential imbalances in the dataset.

5. **Random Permutation:** Data was randomly permuted at each split to reduce bias and improve generalization.

The model was trained on the training data and subsequently used to generate predictions on the test data. To assess performance, recall scores were calculated for both the training and test sets. The absolute difference between these scores was computed to quantify the extent of overfitting. The model with the highest recall score and minimal overfitting was selected as the optimal model.

Model Building

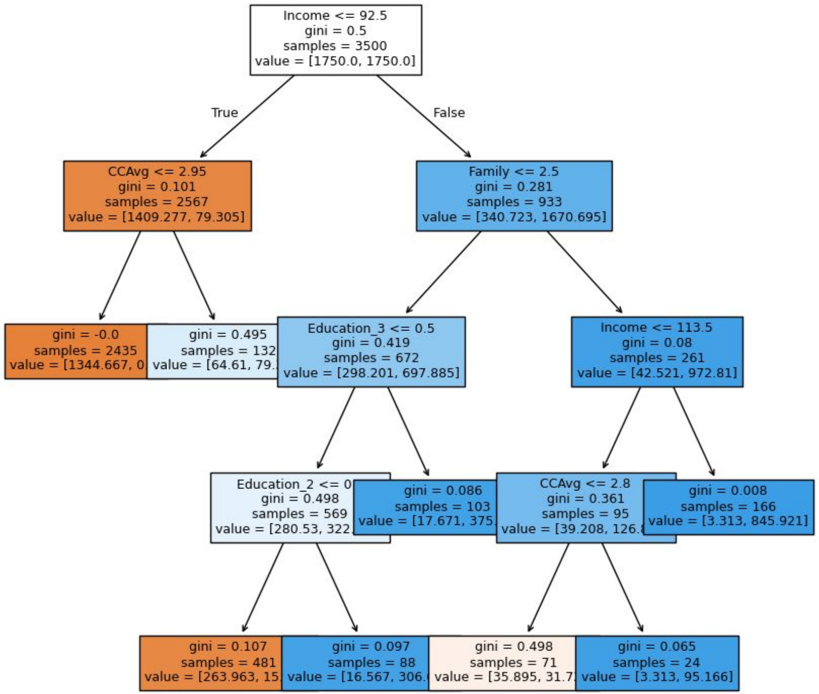
Pre-prune the model setting **minimum** samples **per leaf, maximum leaf** nodes, and **depth** constraints. The **training data** results are similar to Default model without fine tuning.



- **Top Left (TN):** 2978 (85%) - The model correctly predicted 2542 instances as class 0.
- **Top Right (FP):** 191 (5%) - model incorrectly predicted as class 1
- **Bottom Left (FN):** 0 (0.3%) - Incorrectly predicted as class 0
- **Bottom Right (TP):** 320 (9%) - The model correctly predicted 322 instances as class 1.

Model Building

Pre Pruned model effectively uses "**Income**," "**CCAvg**," and "**Family**" **as key features**. The pruning parameters successfully controlled the tree's complexity

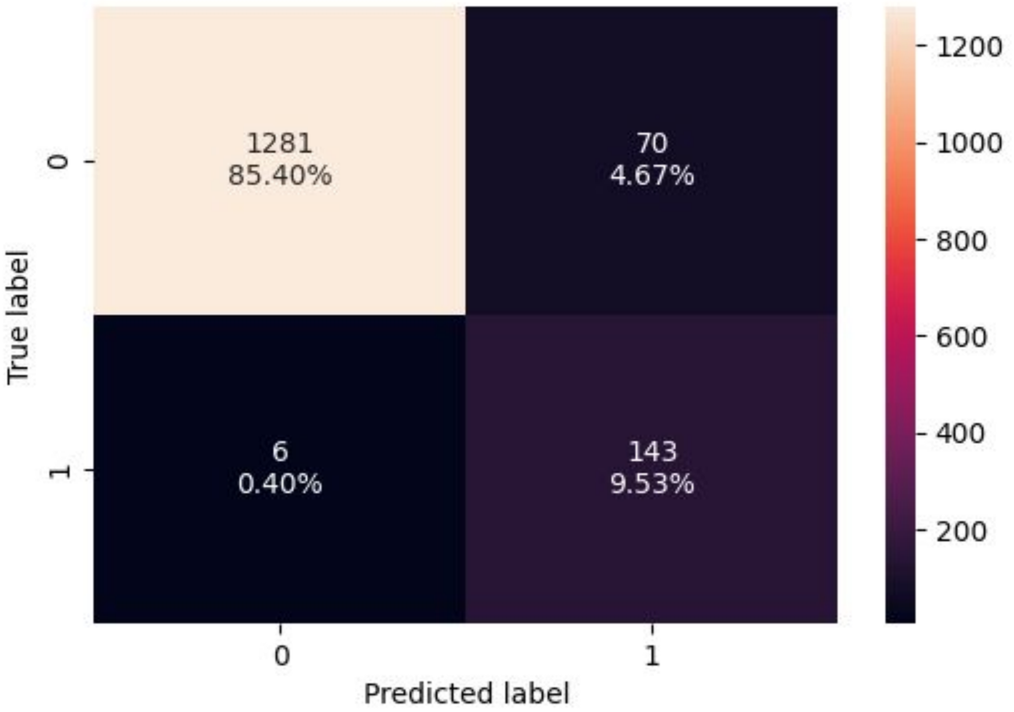


Observations:

- **Key Features:** The tree highlights "Income," "CCAvg", “Education” and “Family” as important features.
- **Depth:** The tree's depth of 4 suggests that simpler models were favored during pruning.
- **Impurity:** The Gini impurity decreases as you move down the tree, indicating that the splits are effective in separating the classes.
- **Pruning Impact:** The pruning parameters have likely prevented the tree from growing too complex, which is beneficial for generalization.

Model Building

Pre-prune model performance on **test data** The model is effective at identifying positive cases (class 1) while maintaining high accuracy

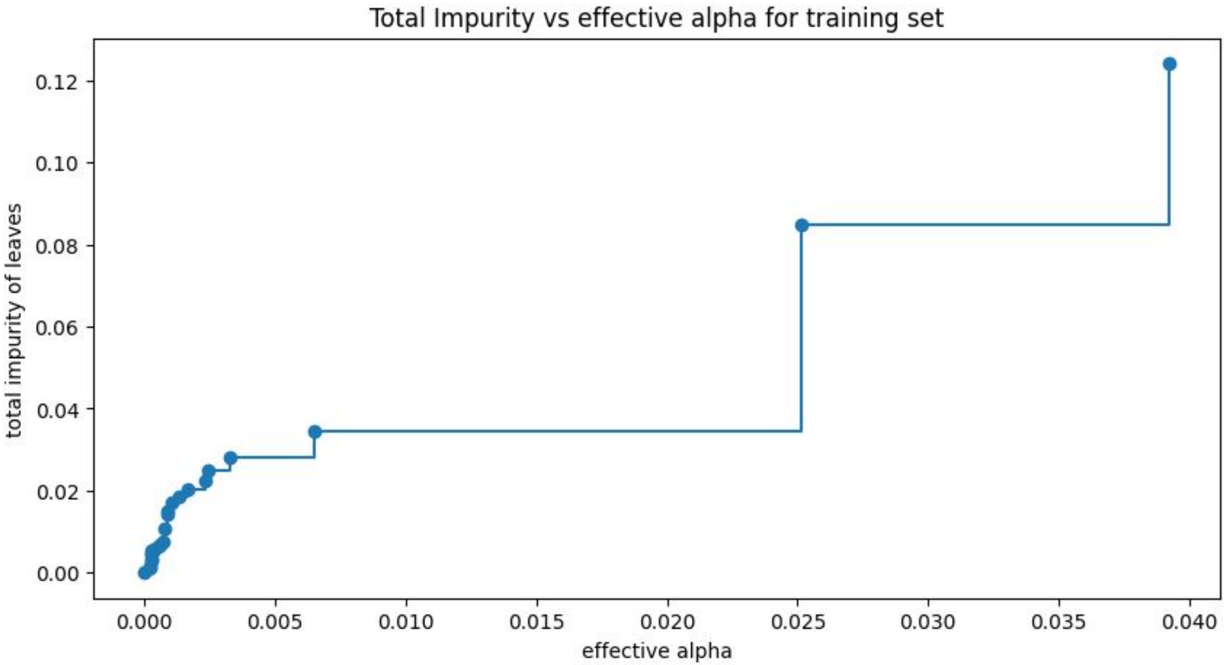


Generalization

- The model is generalizing well to the test data, suggesting that it has learned meaningful patterns from the training data
- **Accuracy: 94%**
- **Recall: 95%**
- **Precision: 67%**
- **F1 Score: 79%**.

Model Building

Post Prune Decision Tree Classifier was **built** with **Cost Complexity Pruning** to simplify the Tree and prevent overfitting. Cost Complexity trims the tree to make it simpler.



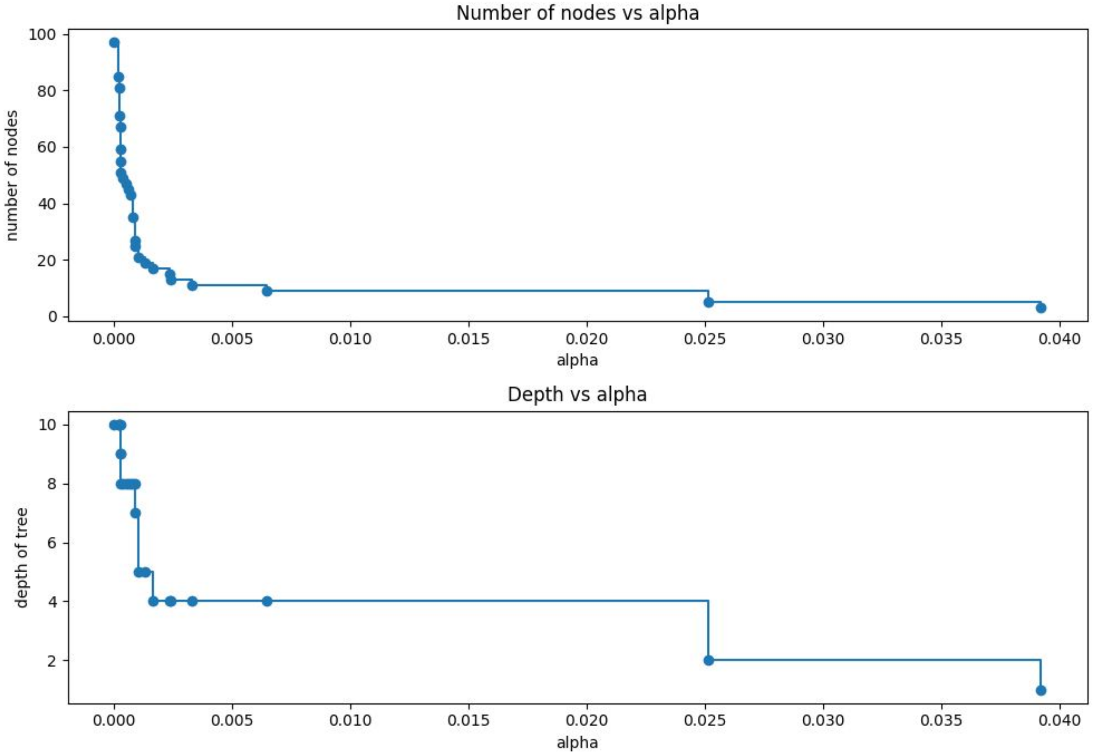
Inverse Relationship: There is a general inverse relationship between "effective alpha" and "total impurity." As alpha increases, the total impurity tends to increase as well.

Low Alpha, Low Impurity: At very low alpha values (close to 0), the total impurity is also very low. This indicates that the tree is complex and fits the training data very well.

Sharp Increase in Impurity: There are a few points where the impurity increases sharply with a small increase in alpha. This means that even a small amount of pruning can have a significant impact on the tree's impurity.

Model Building

Understanding how alpha controls the tree's complexity and aid in selecting an appropriate alpha value to balance overfitting and underfitting is very important



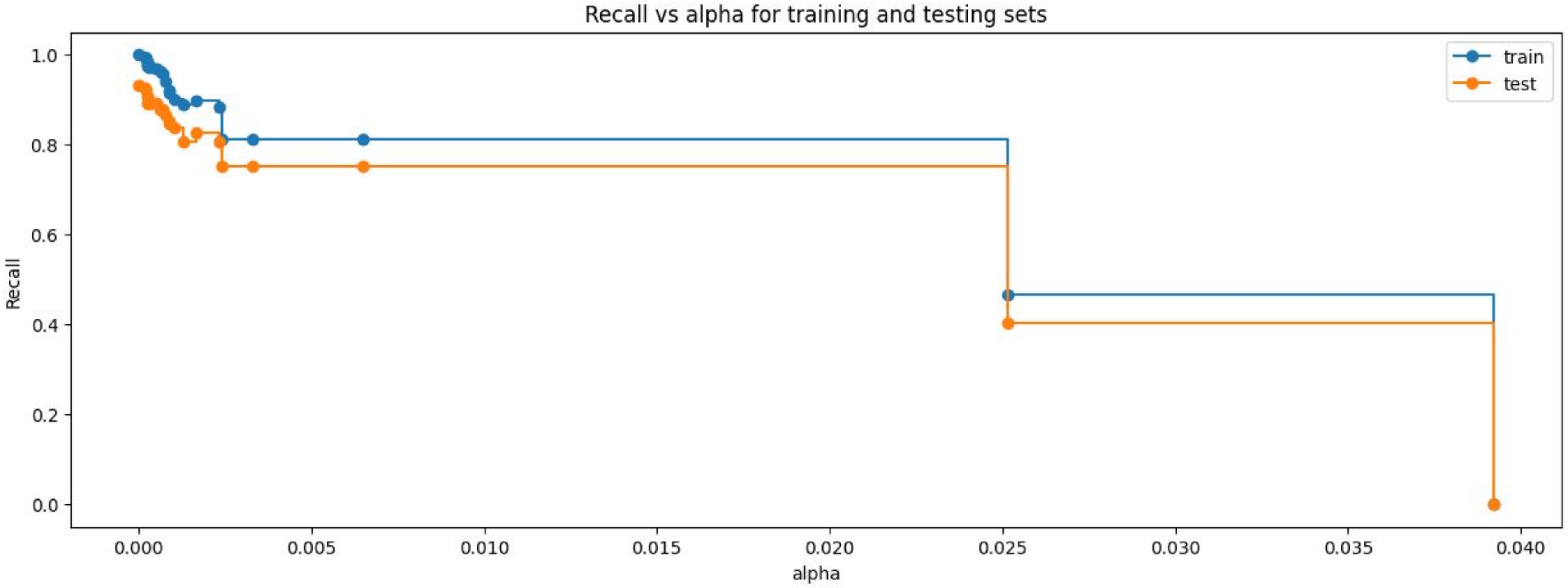
Overfitting vs. Underfitting:

- A very low alpha results in a complex tree (high number of nodes and depth), which may overfit the training data.
- A very high alpha results in a simple tree (low number of nodes and depth), which may underfit the data.

Choosing the right alpha: The plots show that there are drastic changes in the tree structure with very small changes in the alpha value, at the lower end of the alpha scale.

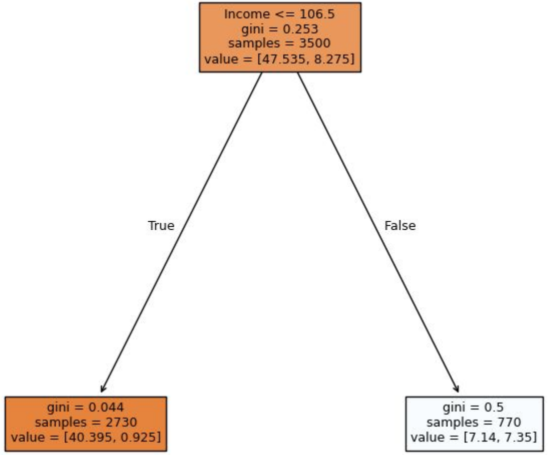
Model Building

The plot highlights the risks of overfitting at low alpha values and underfitting at high alpha values. The "**elbow**" or point where the testing recall starts to decline sharply indicates the **optimal alpha** value. In this case, it appears to be somewhere **before 0.025**.



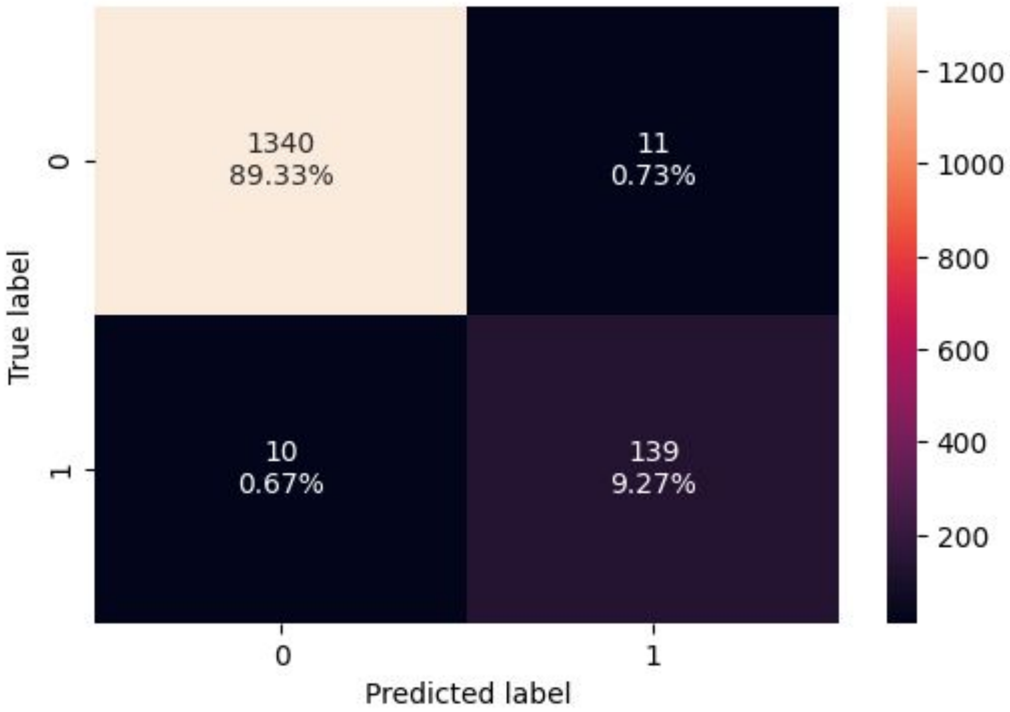
Model Building

The decision tree graph shows a simple split based on "**Income,**" effectively separating a portion of the data with low impurity. However, the right node remains mixed, and the data is imbalanced



Model Building

Post Prune model performance evaluation on **Test Data** Set. Reveals similar performance on Train Dat set.



- **Accuracy:** **98%**
- **Recall:** **93%**
- **Precision:** **92%**
- **F1 Score:** **92%**.

Model Performance Summary

Model Evaluation Criterion

- Given that all three models **exhibit identical performance** across all the listed metrics (**accuracy, recall**, **precision**, and **F1-score**), the optimal model selection prioritize simplicity and interpretability.
    - **Identical Performance** with no differences in their predictive capabilities on the test set, there is no advantage in terms of accuracy, recall, precision, or F1-score to favor one model over another.
    - **Simplicity and Interpretability** where predictive performance is equivalent, the model's complexity becomes a crucial factor. Simpler models are easier to understand
    - **Potential for Overfitting** more complex models have a higher risk of overfitting the training data, even if they perform similarly on the test set.

Model Performance Summary

Model Selection: Pre-Pruned Decision Tree Classifier

Based on analysis, the **pre-pruned decision tree classifier** has been identified as the **most effective algorithm** for this classification task. The model was optimized using the following hyperparameters:

- **Maximum Depth:** A range of values were evaluated (5, 10, and 20) to determine the optimal depth for the decision tree. This hyperparameter limits the complexity of the tree by restricting the number of levels.
- **Maximum Leaf Nodes**: The maximum number of leaf nodes (terminal nodes) was also explored, with values of 2, 8, 16, and 32. This hyperparameter further controls the tree's complexity by limiting the number of final decision points.
- **Minimum Samples Split**: To prevent overfitting and ensure meaningful splits, the minimum number of samples required to split an internal node was set to 2, 5, and 10. This constraint prevents the creation of very small leaves that may be sensitive to noise in the data.

1.0

Model Performance Summary

- Summary of key performance metrics for **Training Data Metrics** the models correctly classified all instances and demonstrated excellent balance between identifying true positives and minimizing false positives and false negatives.

Decision Tree Model	**Accuracy**	**Recall**	**Precision**	**F1**
Default Model	1.0	1.0	1.0	1.0
Pre Pruning	0.94	97	0.62	0.76
Post Pruning	1.0	1.0	1.0	1.0

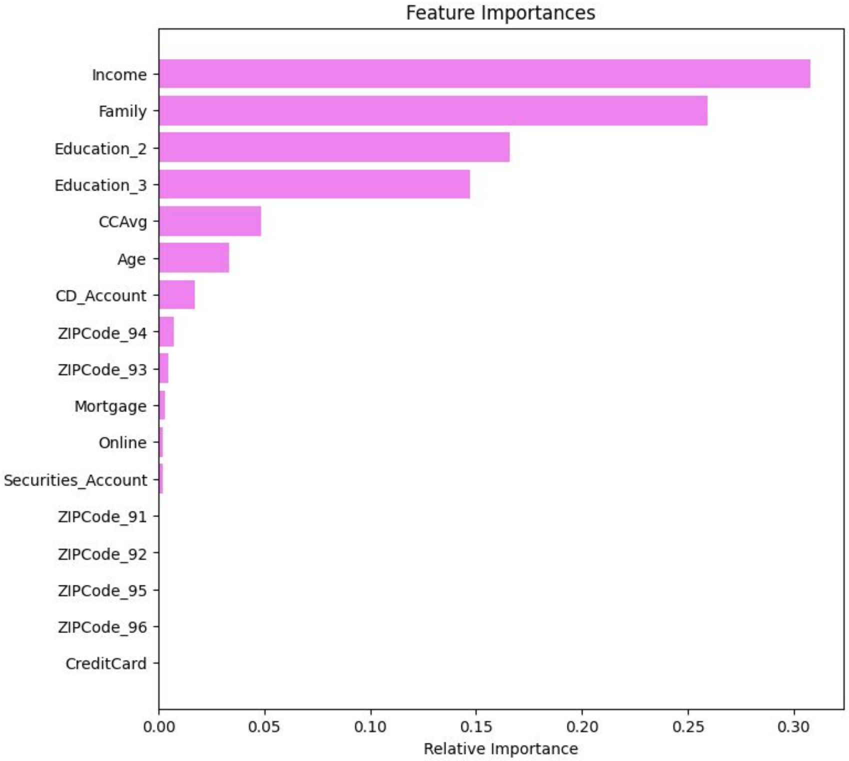
Model Performance Summary

- The **Default** and **Post-Pruning models** show identical and strong performance across all metrics, with high accuracy, recall, precision, and F1-scores. In contrast, the **Pre-Pruning** model **exhibits a trade-off**, showing a decrease in precision and F1-score despite maintaining high recall

Decision Tree Model	Accuracy	Recall	Precision	F1
Default Model	0.986	0.933	0.927	0.930
Pre Pruning	0.949	0.959	0.671	0.790
Post Pruning	0.986	0.933	0.927	0.930

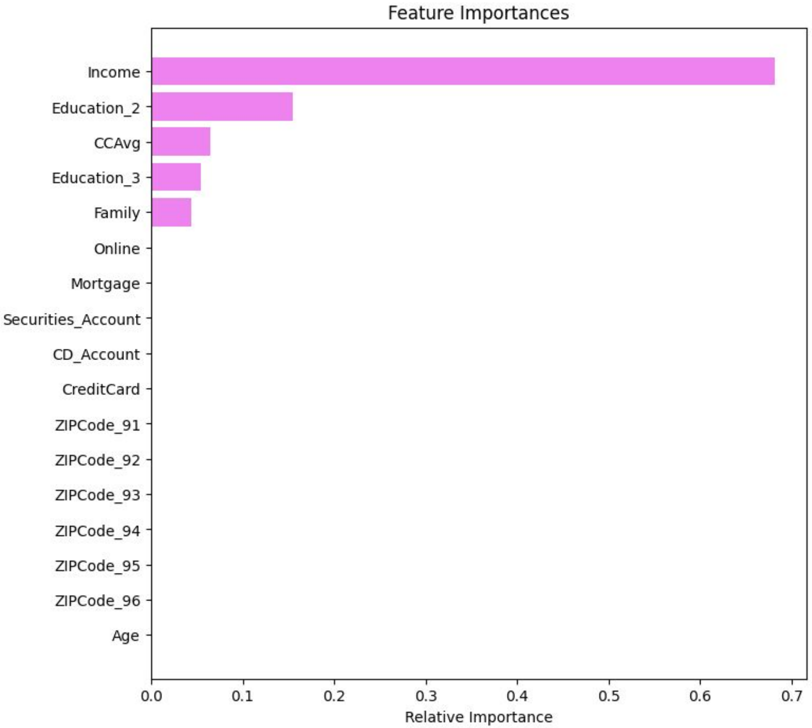
Model Performance Summary

Default Model Feature Importance Income is the most important feature **in the model**, followed by **Family** size and **education levels**. Other features have varying levels of importance, on the model's predictions



Model Performance Summary

The **"Income"** feature is overwhelmingly the most important feature in the **Pre Prune model,** with a significantly longer bar than any other feature. Its relative importance is very high, close to 0.9.

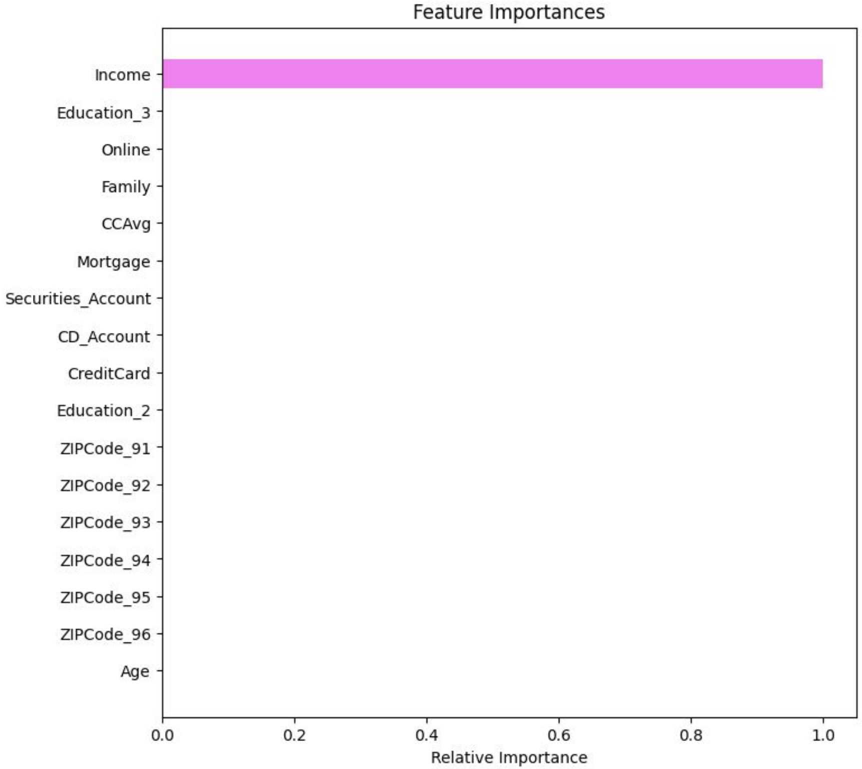


Moderate Importance:

- **Income:** has high importance, but its relative importance
- **Education:** Graduate, Professional degrees
- **CCAvg:** Credit Card Average Spending
- **Family:** Family is moderately important

Model Performance Summary

The **Post Pruning** process has resulted in a highly simplified **model** that ignores all other features. This suggests that the tree may have been over-pruned. Income determines the predictions



Model Performance Improvement

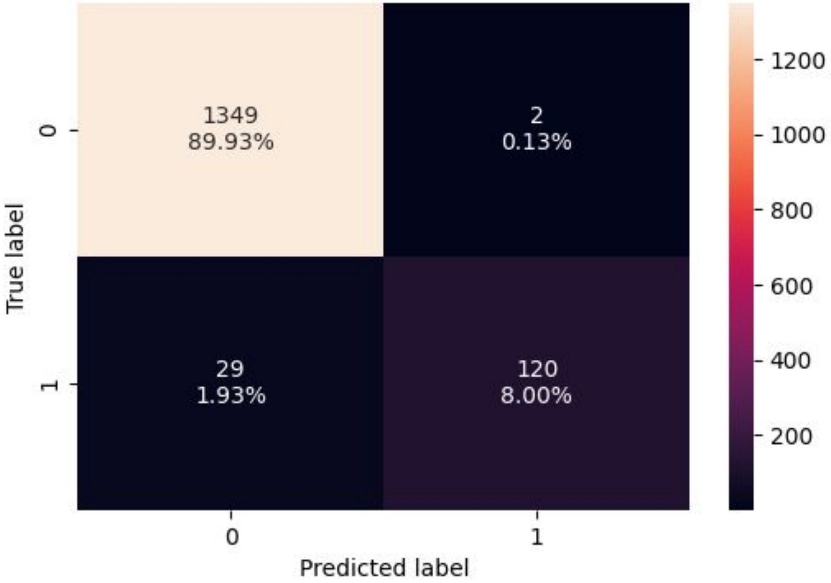
- The **decision rules for RandomizedSearchCV** strategies and principles is to randomly sample a specified number of combinations from a given distribution.and check the feature importance.
- **Pre Pruning Decision Tree Classifier** stops the tree from growing too deep during training phase. It sets the rules to limit the size and complexity of the tree.Defined by the hyperparameters.
- **Post-pruning Decision Tree Classifier** is applied after the full tree is built, and it prunes back nodes that do not provide sufficient improvement. It balances model complexity and performance.

Cost Complexity Pruning (CCP)

- Uses a **cost complexity parameter (α)** to find the optimal trade-off between the complexity of the tree and its performance

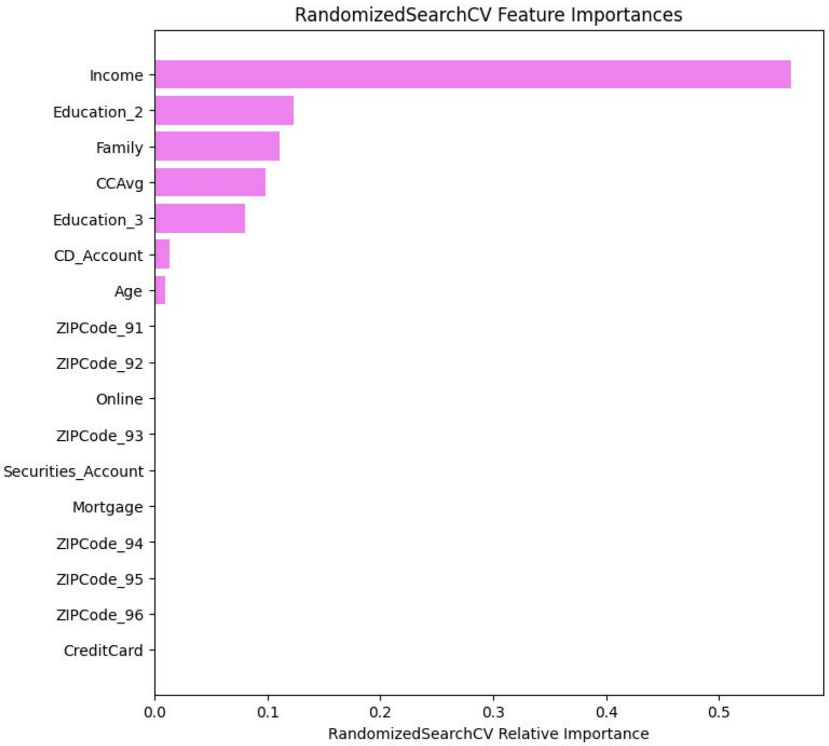
Model Performance Improvement

Randomized Search Model correctly classified **97.93%** of the instances as accuracy. The model is highly precise when predicting positives with a **98.36%** on precision score.



Model Performance Improvement

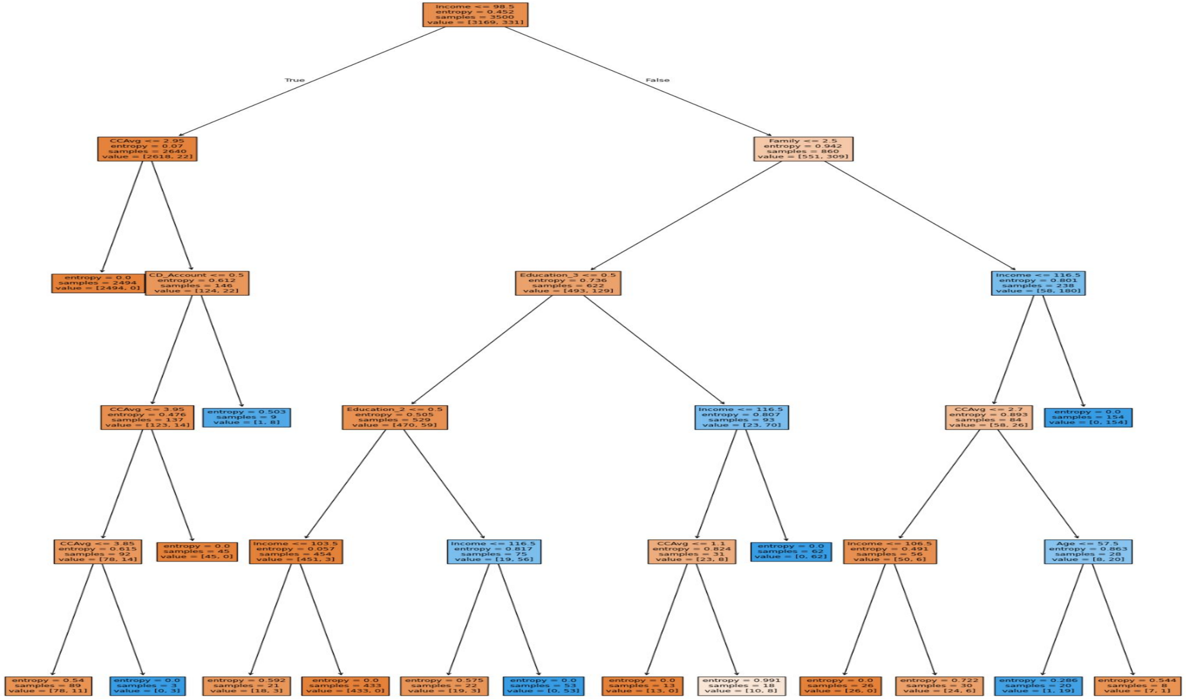
Pre Purned model with Randomized Search CV, the importance feature is Income, Education 2, Family and CCAvg, Education 3. Income is over 0.5 as the dominant feature



	Imp
Income	0.564204
Education_2	0.123596
Family	0.111099
CCAvg	0.098022
Education_3	0.079924
CD_Account	0.013459
Age	0.009696
Online	0.000000
Securities_Account	0.000000
ZIPCode_91	0.000000
ZIPCode_92	0.000000
ZIPCode_93	0.000000
ZIPCode_94	0.000000
ZIPCode_95	0.000000
ZIPCode_96	0.000000
Mortgage	0.000000
CreditCard	0.000000

Model Performance Improvement

- Randomized Search CV utilization of entropy to build tree,



Model Performance Improvement

- **Randomized Search CV model** has **high precision** but **lower recall**, the pattern may indicate the model is conservative labeling positive instances. **Confident predicting positives**. If recall is more critical than precision than adjustment is required

Metric	Score
Accuracy	0.979333
Recall	0.805369
Precision	0.983607
F1 Score	0.885609

APPENDIX

Data Background and Contents

- 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)
- CreditCard: Does the customer use a credit card issued by any other Bank (excluding All life Bank)? (0: No, 1: Yes)

Data Background and Contents

What are the datatypes of the different columns in the dataset

ID	int64
Age	int64
Experience	int64
Income	int64
ZIP Code	int64
Family	int64
CCAvg	float64
Education	int64
Mortgage	int64
Personal_Loan	int64

Data Background and Contents

What are the datatypes of the different columns in the dataset

Securities_Account	int64
CD_Account	int64
Online	int64
CreditCard	int64

Data Background and Contents

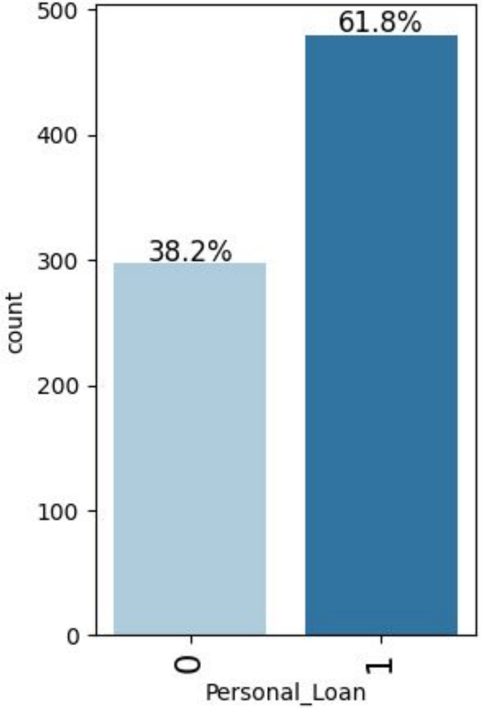
- **How many rows and columns** are there in the dataset:
  - **Row:** 5000
  - **Columns:** 14

- **Statistical summary of the data**

	Minimum	Average	Maximum
Age	23	45	67
Income	8000	64000	224000
Family	1	2	4
Education	Undergraduate	Graduate	Professional

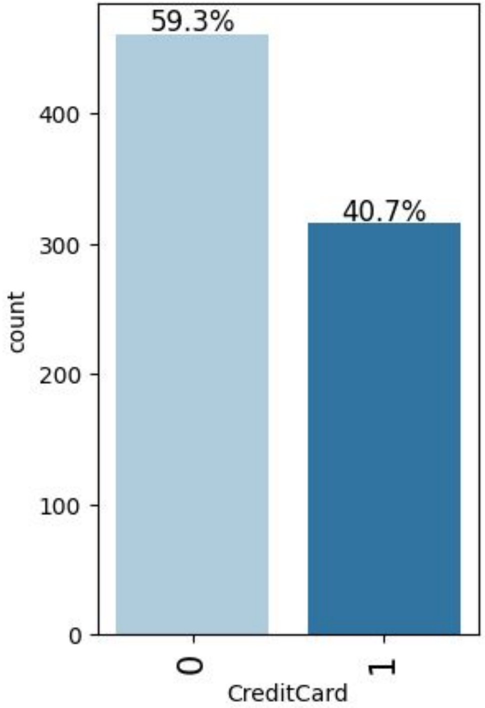
Personal Loan Accepted Last Campaign

- Percentage of Personal Loan Accepted in Last Campaign. Showing that **62%** of Customer **accepted Loan**, compared to 38% who did not accept Loan



Credit Card Analysis

- **Credit Card** Analysis shows that 59% of customer use a credit card from another bank, and only 40% does not use a credit card from another bank.



Great Learning

Happy Learning !