Personal Loan Campaign

PGP-AIML-BA-UTA-Jun25-D

Yashpal Singh

Date: 08/25/2025

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix
- Post-Pruned Decesion Tree Rules

Executive Summary

AllLife Bank, while expanding its customer base, faces a challenge the majority of customers are depositors, but only a small fraction have personal loans. Since loans generate interest income, the bank needs to effectively identify liability customers who are most likely to convert into loan customers.

Using demographic and transactional data **from 5,000 customers**, I developed predictive models to classify depositors by their likelihood of loan uptake. Extensive exploratory data analysis (EDA) **revealed that Income and Credit Card Average Spending (CCAvg) are the strongest predictors of personal loan purchases, with education and family size also influencing loan adoption**.

Three Decision tree were evaluated

- Default Decision Tree – Showed perfect training performance but overfitted, failing to generalize well.
- Pre-Pruned Decision Tree – Avoided overfitting but underperformed due to oversimplification.
- **Post-Pruned Decision Tree** – Achieved the **best balance, delivering high precision, recall, and F1-score on both training and test data, making it the most reliable model.**

Key Findings

✓ **Higher-income customers** and those with higher average credit card spending are significantly more likely to take personal loans.
✓ **Customers with undergraduate education show greater loan adoption** compared to advanced degree holders.
✓ **CD account ownership is a strong indicator of loan interest**, while factors like age and experience are weak predictors.
✓ **Data is imbalanced, but model performance remained strong with post-pruning**.

Executive Summary

Three Decision tree were evaluated

- ✓ **Deploy the Post-Pruned Decision Tree model for targeted marketing, focusing on high-income, high-CCAvg segments.**
- ✓ **Optimize marketing ROI by prioritizing outreach to customers with CD accounts and undergraduate-level education**.
- ✓ **Adopt probability-based targeting:** instead of binary predictions, use the model's probability scores to rank customers by loan likelihood.
- ✓ **Refine campaign strategy:** customers below a defined threshold (e.g., <60% probability) should be routed for manual review, reducing false targeting costs.

Business Problem

Context:

AllLife Bank, a U.S.-based financial institution, is experiencing growth in its customer base. While the majority of its customers are depositors (liability customers), only a small segment holds loans (asset customers). Since loan products generate interest income, the bank's management aims to expand its loan customer base. The marketing team has previously run campaigns targeting depositors, achieving a conversion rate of ~9%. Encouraged by this, the bank now seeks to use data-driven strategies for more effective and personalized marketing, focusing on converting liability customers into personal loan customers while ensuring customer retention.

Problem Statement :

The bank needs a predictive model to identify depositors who are most likely to purchase a personal loan. By leveraging customer demographic and transactional data, the model should classify customers based on their probability of loan uptake. This will enable the marketing team to target high-potential customers, optimize campaign efforts, improve conversion rates, and maximize return on investment.

Objectives

1.To predict whether a liability customer will buy personal loans,
2.To understand which customer attributes are most significant in driving purchases, and identify which segment of customers to target more.

Solution Approach

Data Preparation

- **Load the Data**: Loading the customer data, which includes demographic details (e.g., Age, Experience) and other information
- **Handle Missing Values**: Check for any missing data points and decide on an appropriate imputation strategy, such as replacing them with the mean or median.
- **Feature Engineering**: If necessary, create new features that could be predictive. For example, handling Zip code

Split the Data:

- Divide the dataset into a training set and a test set The training set will be used to build the model, and the test set to evaluate its performance on unseen data.

Model Building :

- **Model Selection**: Choose a Decision Tree Classifier because it's highly interpretable, which is crucial for identifying key customer attributes to purchase a loan

- **Training**: Fit the Decision Tree model to training data. The model will learn a series of if-then-else rules from the features to predict the target variable (**Personal Loan**). And validate the performance metrics on test data set
- Based on the observation tune the model (Pre-Pruned or Post –Pruned)

- **Hyperparameter Tuning:** Adjust the model's hyperparameters on the training set to prevent overfitting. Key hyperparameters to tune include:

Solution Approach-Continued--

Performance Comparison

- **Performance Metrics:** Evaluate the model's effectiveness using metrics appropriate for an imbalanced dataset, such as Precision, Recall, F1-Score. A high Recall is especially important here to ensure you identify as many potential loan customers as possible.
- **Create a comparison table** for all performance matrices and select the best model based on the best performance Metrics

Model Evaluation and Insights

- **Prediction:** Use the final, tuned model to predict the probability of a personal loan purchase on the test set.
- **Feature Importance**: A key advantage of Decision Trees is their ability to show feature importance. The model will automatically rank the features based on how much they contribute to the purity of the splits. You can use this to identify the most significant customer attributes, such as Income or CCAvg, that drive loan purchases.
- **Targeted Segments:** Based on the feature importance and the rules derived from the tree, you can identify specific customer segments to target. For example, if the tree shows that high Income and high CCAvg are key indicators, you would recommend the marketing team focus on this specific customer group.

Data background and contents

1. **Data.shape=(5000,14):** Actual Data set is having 5000 record and 14 columns
2. **Data type** : float64(1), int64(13)
3. **Statistical Summary**---

   The 'Income' variable has a large standard **deviation (46.03) and a maximum value (224) that** is significantly higher than the **75th percentile (98).** This suggests that the income data is widely dispersed and likely includes a number of high-earning outliers.

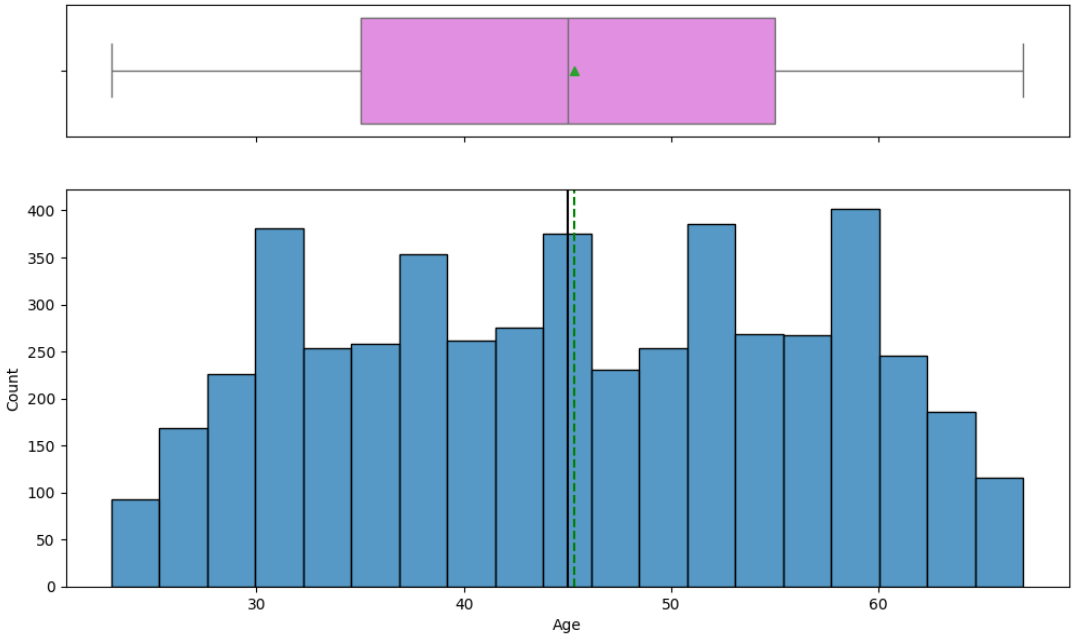
   The mean for the 'Online' variable is 0.5968, indicating that close to 60% of the individuals in the dataset use online banking. This makes it a very common feature compared to other Boolean variables like 'CreditCard' (29.4% adoption), 'Securities_Account' (10.4% adoption), and 'CD_Account' (6.04% adoption).

   The 'Mortgage' column shows a highly right-skewed distribution. While the mean mortgage is around 56.5, the median (50th percentile) is 0, and the 75th percentile is 101. This suggests that a significant portion of the population has no mortgage, while a smaller group has very high mortgage value
4. **Dropping column ID:** As ID is unique column and its not adding any value hence dropping it
5. **Experience is dropped** as it is perfectly correlated with Age
6. **Observation on Experience :** It is observed Experience column contains -1,-2 and -3 years of experience hence these values are made positive by using replace() function by using inplace=True as negative experience does not make sense
7. **Education** : Three level of educations are there in data set Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional and there is no observation for this column

Data Preprocessing

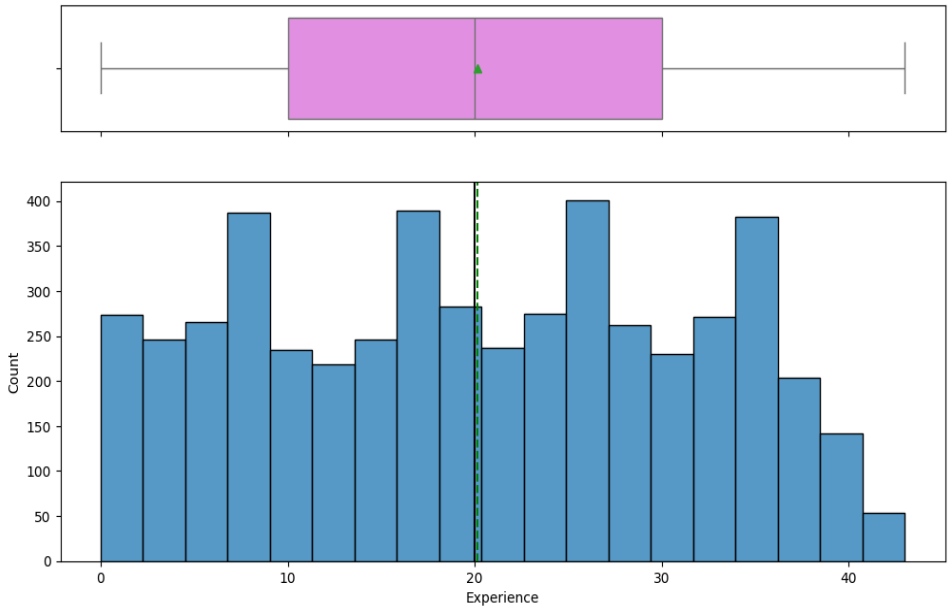
- **Duplicate value check :** Data does not have any duplicate value
- **Missing value treatment :** Data does not have any any missing value
- **Feature Engineering** : In this step transforms a high-cardinality numeric ZIP code (**Total unique 467 Zip code**)into a low-cardinality categorical region feature. This makes it interpretable and useful for the decision tree, while avoiding overfitting and meaningless splits, and because of it taking the first two digits groups ZIP codes into broader regions (**7 Unique regions**)
- **Outlier Check :** A significant number of outliers exist at the high end of the **Income range. Credit Card and high mortgage values,** as these are the important data hence, they won't be removed

EDA Result-Observations on the Age



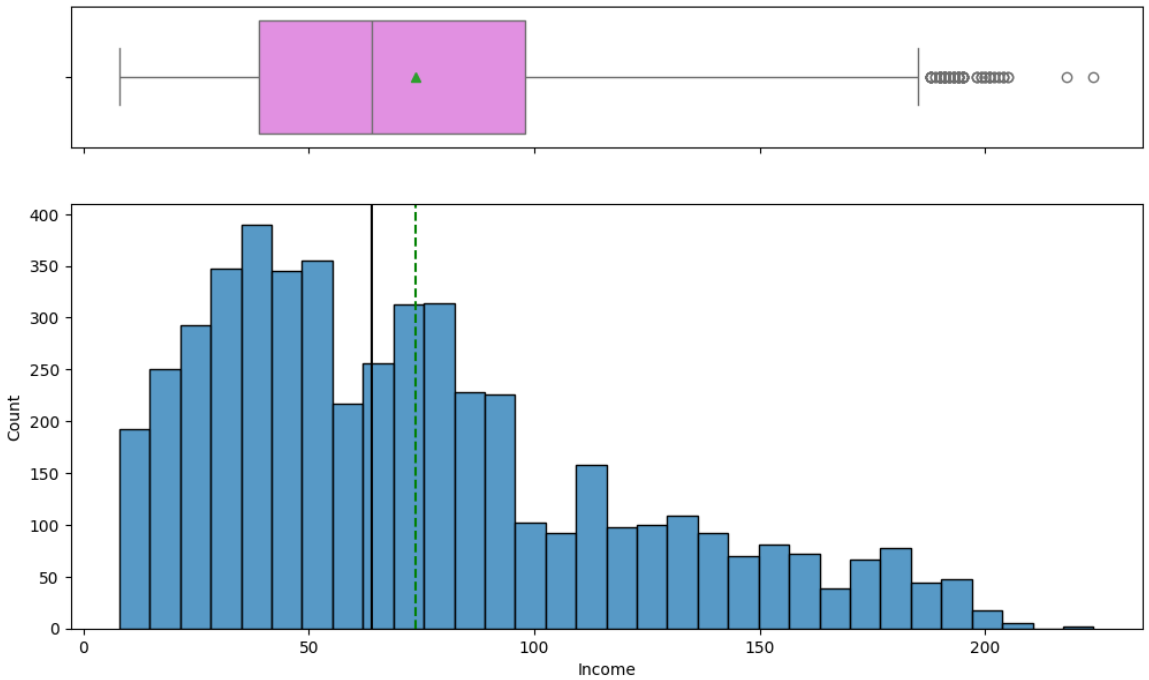
- **The distribution is generally bell-shaped but not perfectly normal**. The histogram shows that most ages are centered around the middle ranges, with fewer individuals at the youngest and oldest ages.
- **The data is approximately symmetrical**. The mean and median are very close, and the box plot shows balanced whiskers, indicating a lack of significant skew.
- **The median age is around 45.** This means that half of the individuals in the dataset are younger than 45, and the other half are older.
- **No outliers are present**. The box plot's whiskers extend to the minimum and maximum values, with no individual points plotted outside of the main data range.
- **The bulk of the population is between 35 and 55.** This central box in the box plot represents the middle 50% of the data, showing that most people are concentrated within this 20-year age span.

EDA Result-Observations on the Experience



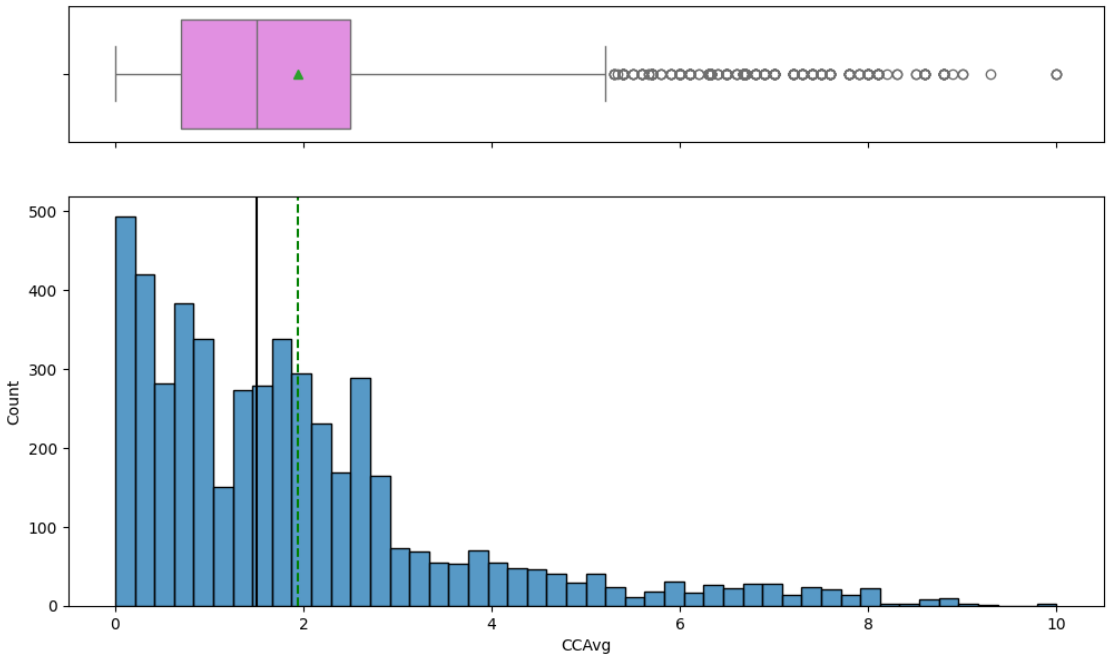
- **The median experience is around 20 years**. The median, is located at approximately 20 on the x-axis, suggesting that half of the individuals have more than 20 years of experience and half have less.
- **The data is somewhat symmetric with a slight right skew**. The median is positioned almost centrally within the box, indicating a roughly symmetric distribution. However, the whisker on the right side of the box plot is slightly longer, and the histogram shows a longer tail to the right, suggesting a minor positive skew.
- **The interquartile range (IQR) is roughly between 10 and 30 years of experience**. The box in the box plot spans from the first quartile (Q1) to the third quartile (Q3). The left edge of the box is near 10 and the right edge is near 30, meaning the middle 50% of the data falls within this 20-year range.
- **The distribution is multi-modal, with peaks around 5, 10, 20, 25, and 35 years.** The histogram shows several distinct peaks, indicating that a large number of people have experience concentrated at these specific years rather than a single, dominant peak.
- **The range of experience is from approximately 0 to 45 years**.

EDA Result-Observations on the Income



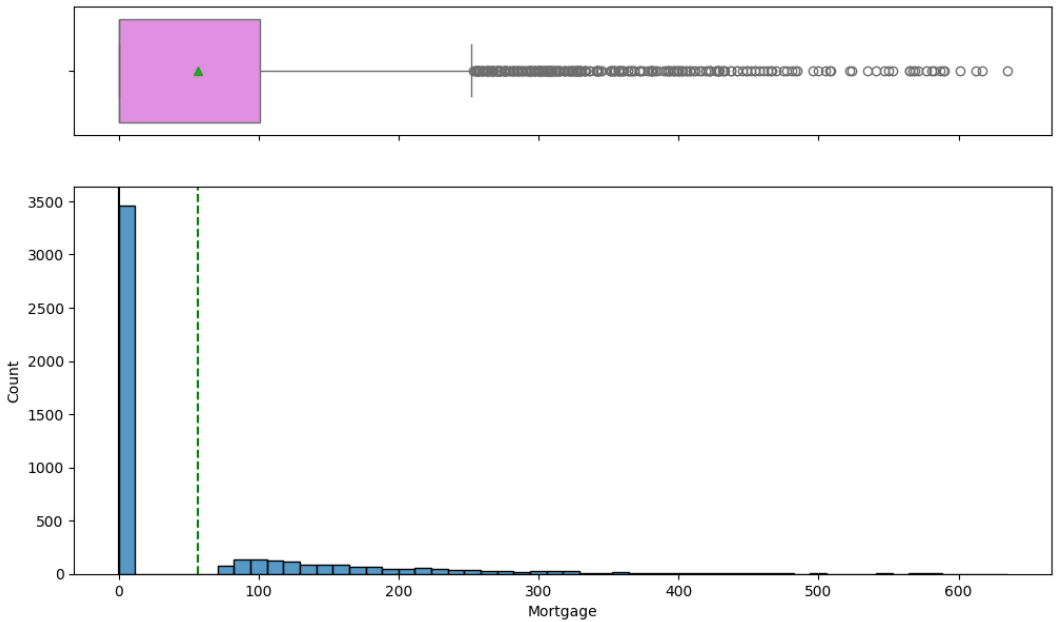
- **The median income is approximately 75**. The green triangle within the box plot and the dashed green line on the histogram both indicate that the median is around 75.
- **The data is positively skewed**. The long tail extending to the right in the histogram and the group of outliers on the right side of the box plot both show that the distribution is skewed to the right.
- **A significant number of outliers exist at the high end of the income range.** The individual points to the right of the upper whisker in the box plot represent a number of incomes that are significantly higher than the rest of the data.
- **The most frequent income is around 60**. The tallest bar in the histogram is near the 60 mark, indicating a high concentration of individuals with an income around this value.
- **The majority of incomes fall between approximately 40 and 100.** The central box in the box plot, which represents the interquartile range (IQR), spans from about 40 to 100, showing that 50% of the data lies within this range.

EDA Result-Observations on the CCAvg



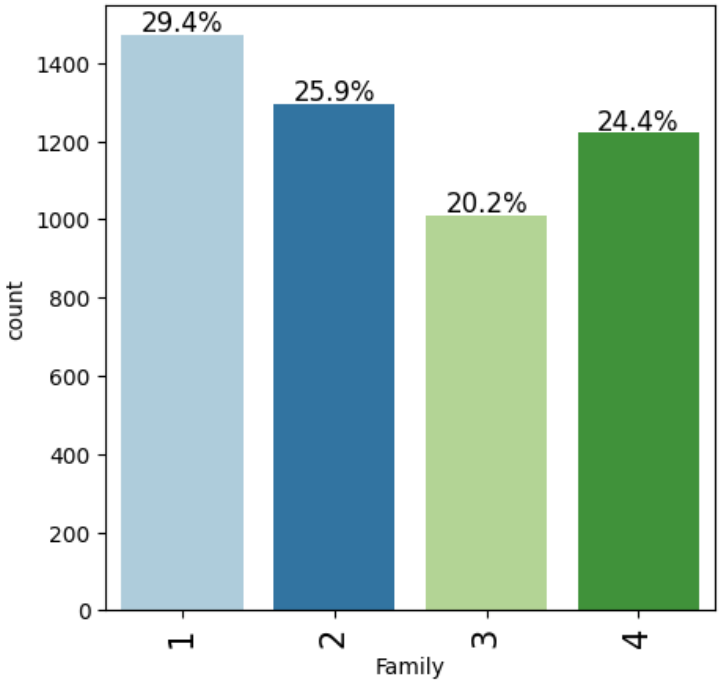
- **The data is highly positively skewed**. The histogram shows a long tail to the right, with a high concentration of data points at lower values and a few at very high values.
- **The median CCAvg is approximately 1.5**. The green triangle in the box plot and the dashed green line on the histogram both indicate that the median is around 1.5.
- **The majority of individuals have a low credit card average spending.** The histogram shows a large number of people with a CCAvg between 0 and 2.
- **There is a significant number of outliers with high credit card average spending**. The individual circles to the far right of the box plot represent a number of data points with high CCAvg values.
- **The most common CCAvg is near zero**. The tallest bar in the histogram is at the very beginning of the scale, indicating a peak in the count of people with a very low or zero average credit card spending.

EDA Result-Observations on the Mortgage



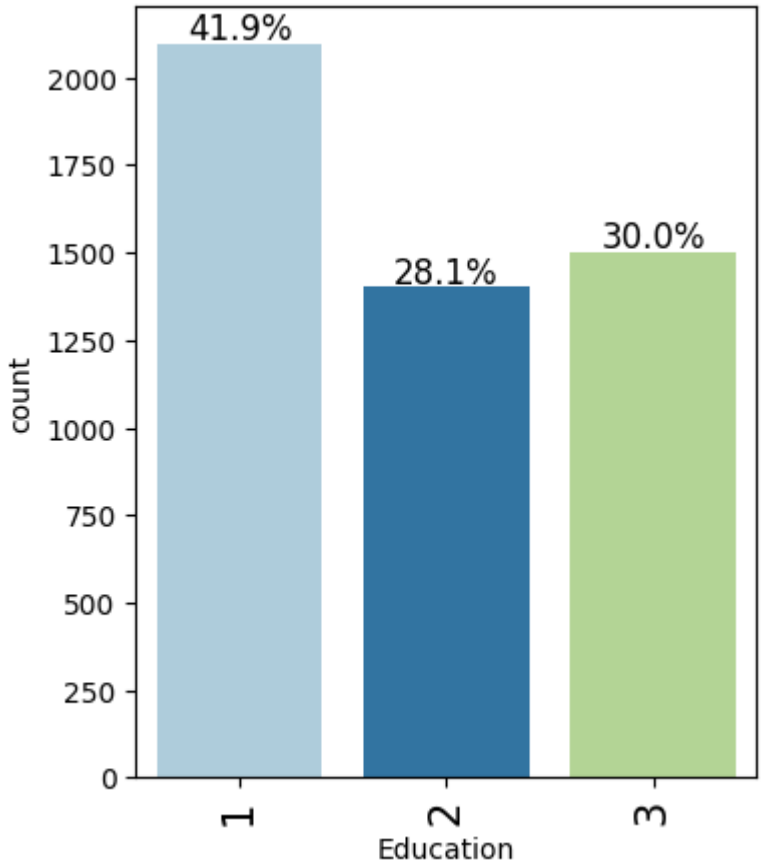
- **The data is highly positively skewed**. The histogram shows a very long tail to the right, with most of the data clustered at the lower end of the mortgage values.
- **The median mortgage value is zero**. The green triangle in the box plot and the dashed green line in the histogram are both at the zero mark. This indicates that more than half of the individuals have a mortgage value of zero.
- **The majority of individuals have no mortgage**. The first bar of the histogram is significantly taller than all others, indicating a large number of people with a mortgage value of zero.
- **There is a large number of outliers with high mortgage** values. The individual circles to the right of the box plot's whisker represent many data points with mortgage values much higher than the majority.
- **The range of mortgage values is from zero up to approximately 650**. The histogram and the box plot show that while most values are low, the data extends all the way to around 650.

EDA Result-Observations on the Family



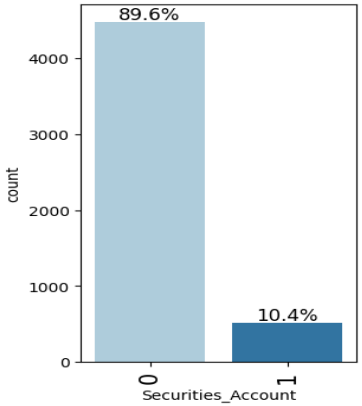
- **The majority of individuals are from families of size 1 and 2**. Families with one person make up the largest group at 29.4%, followed closely by families of two at 25.9%.
- **Families of size 3 are the least common.** This group accounts for the smallest percentage of the data, at 20.2%.
- **Family sizes 1, 2, and 4 are the most frequent**. The counts for these family sizes are all relatively high, ranging from about 1200 to 1400.
- **The distribution is not uniform**. The percentages of the different family sizes vary, with a notable difference between the most and least frequent categories.
- **Family of size 4 is more common than family of size 3.** The count for family size 4 is over 1200 (24.4%), which is significantly higher than the count for family size 3, which is just over 1000 (20.2%).

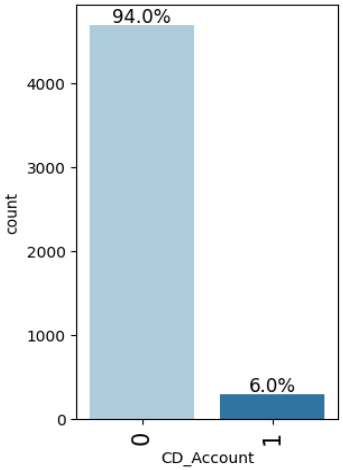
EDA Result-Observations on the Education



- **Level 1 Education is the most common**. This group accounts for the largest share of the data, at 41.9%.
- **Level 2 Education is the least common**. This group has the lowest count, representing 28.1% of the total.
- **There is a significant difference in frequency between the education levels.** The count for Level 1 is much higher than for Levels 2 and 3.
- **Level 3 Education is more frequent than Level 2**. The percentage for Level 3 is 30.0%, which is higher than the 28.1% for Level 2.
- **The total count of individuals with Education Level 1 is over 2000.** The bar for Education 1 extends past the 2000 mark on the y-axis, indicating a high number of individuals in this category.

EDA Result-





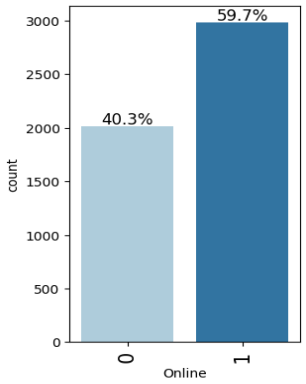
Observations on Security Account

- The vast majority of individuals do not have a securities account. The bar for a Securities Account value of 0 is significantly taller, representing 89.6% of the total count.
- Only a small percentage of individuals have a securities account. The bar for a Securities Account value of 1 is very short, indicating that only 10.4% of the people in the dataset have one.
- The data is highly imbalanced. There's a massive disparity between the number of people who have a securities account and those who don't, with the latter being almost nine times more common.
- The number of people without a securities account is over 4500. The count for the "0" category is close to 4500 on the y-axis, while the count for the "1" category is only around 500.

Observations on CD Account

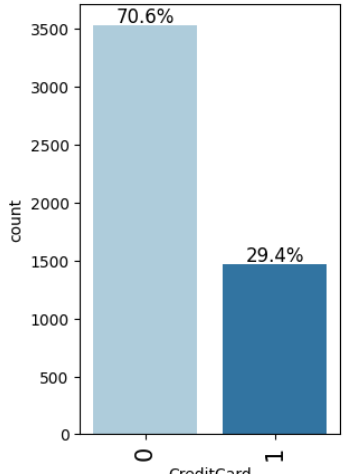
- The vast majority of individuals do not have a CD account. The bar for the CD_Account value of 0 is significantly taller, representing 94.0% of the total count.
- Only a small percentage of individuals have a CD account. The bar for the CD_Account value of 1 is very short, indicating that only 6.0% of the people in the dataset have one.
- The data is highly imbalanced. There is a massive disparity between the number of people who have a CD account and those who don't, with the latter being more than fifteen times more common.
- The number of people without a CD account is over 4500. The count for the "0" category is close to 4700 on the y-axis, while the count for the "1" category is only around 300.

EDA Result-



Observations on Online

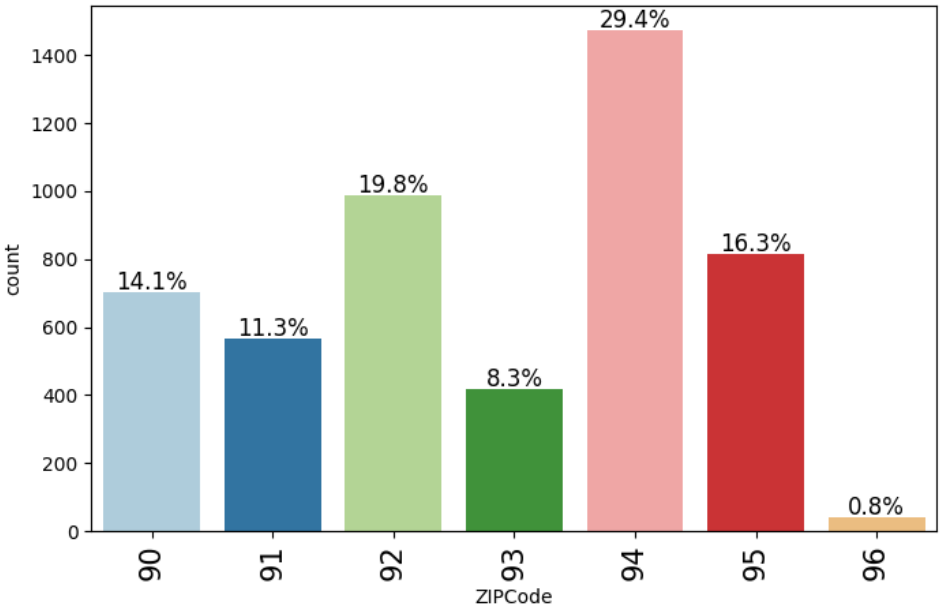
- The majority of individuals are online banking users. The bar for the "1" category is taller, representing 59.7% of the total count.
- A significant portion of individuals are not online banking users. The bar for the "0" category shows that 40.3% of the population does not use online banking.
- The data shows a moderate imbalance. While online users are the majority, the difference is not extreme, with the non-online users still making up a substantial portion of the dataset.
- The count of online users is nearly 3000. The bar for the "1" category is just below the 3000 mark on the y-axis, indicating a high number of online banking customers



Observations on Credit Card

- The majority of individuals do not have a credit card. The bar for the "0" category is significantly taller, representing 70.6% of the total count.
- A substantial portion of individuals do have a credit card. The bar for the "1" category shows that 29.4% of the population has a credit card.
- The data shows an imbalance. The number of people without a credit card is more than twice the number of those who have one.
- The count of people without a credit card is over 3500. The bar for the "0" category is just above the 3500 mark on the y-axis, while the count for the "1" category is about 1500.

EDA Result-Observation on Zip Code



- The most common ZIP code is 94. This ZIP code represents the largest proportion of the data, with 29.4% of the total count.
- **The least common ZIP code is 96**. This ZIP code has the lowest count, accounting for only 0.8% of the data.
- **The distribution is not uniform**. The percentages of the different ZIP codes vary widely, ranging from under 1% to nearly 30%.
- **ZIP code 92 is the second most common**. With 19.8%, this ZIP code has the second-highest count in the dataset.
- **The three most common ZIP codes (94, 92, and 95) collectively represent a majority of the data**. Their combined percentage is 65.5% (29.4% + 19.8% + 16.3%), showing a concentration of individuals in these three areas.

Bivariate Analysis –Heatmap Key Observations

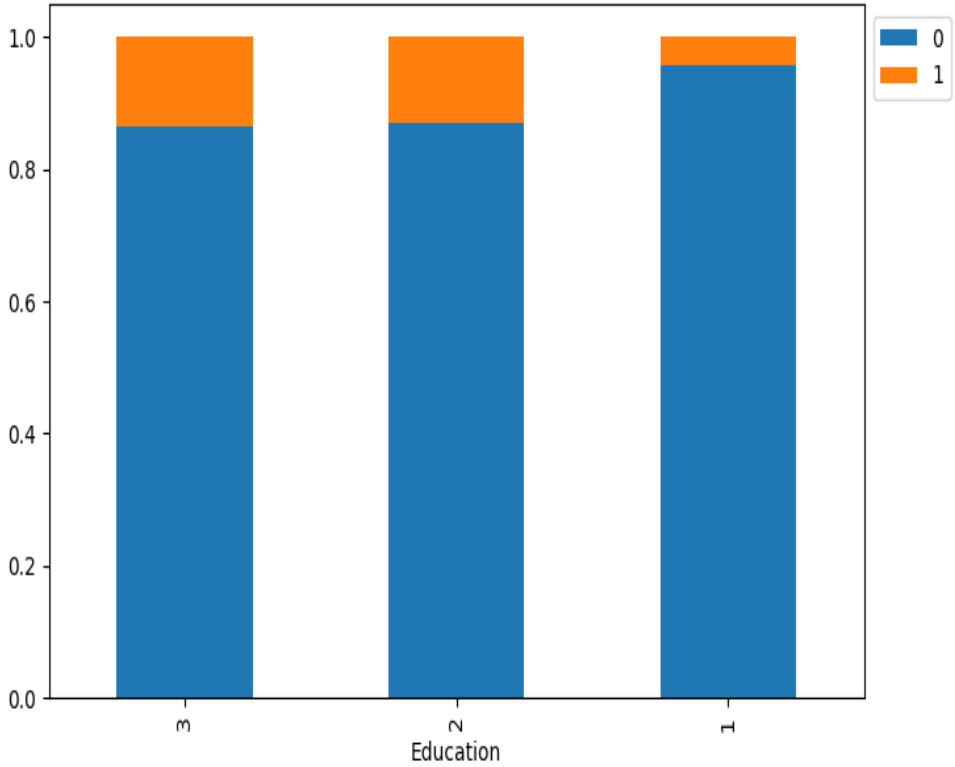
Strong Positive Correlation between Age and Experience: The correlation coefficient between Age and Experience is 0.99, which is very close to 1. This indicates an extremely strong positive relationship, meaning as one's age increases, their years of experience also increase. This is an expected and intuitive relationship.
Experience is dropped as it is perfectly correlated with Age

Moderate Positive Correlation between Income and CCAvg: There is a moderate positive correlation of 0.65 between Income and CCAvg (Credit Card Average Spending). This suggests that as a person's income increases, their average credit card spending tends to also increase. This is a logical relationship, as higher earners often have greater spending capacity.

Weak to No Correlation for Most Variables: The majority of the other variable pairs, such as Age with Income (-0.06) or Family with Mortgage (-0.02), show very low correlation coefficients, with values close to 0. This means there is little to no linear relationship between these variables.

No Strong Negative Correlations: There are no strong negative correlations present in the dataset. The only notable negative correlations are weak, such as between Income and Family at -0.16, suggesting that as family size increases, income slightly decreases, though the relationship is not strong.

Bivariate Analysis –Customer's interest in purchasing a loan varies with their education



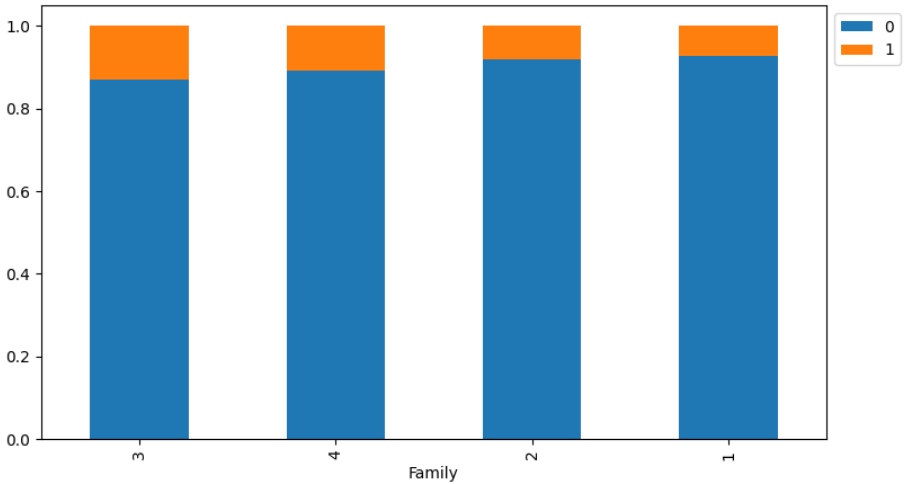
Undergraduate (1): Customers with an undergraduate degree show the highest proportion of loan purchases. The orange bar for this group is the largest of the three, indicating a greater interest in loans compared to the other groups.

Graduate (2): The proportion of customers who bought a loan is lower for graduates compared to undergraduates. The orange segment is smaller, suggesting a reduced interest in purchasing a loan.

Advanced/Professional (3): Customers with an advanced or professional degree have the lowest proportion of loan purchases among the three groups. The orange segment is smallest here, indicating that this group is the least likely to purchase a loan.

This trend suggests a negative correlation between education level and the likelihood of purchasing a loan, with the highest educated group showing the least interest.

Bivariate Analysis –Personal Loan Vs Family



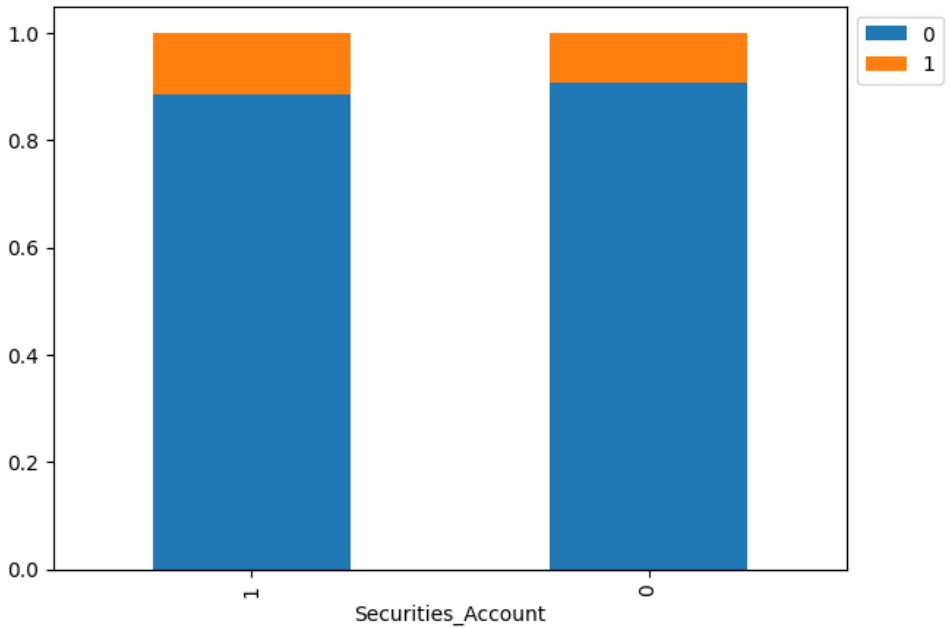
Families of size 1 have the lowest proportion of "1"s. The bar for family size 1 has the smallest orange section at the top, indicating the lowest relative share of the "1" category.

Families of size 3 and 4 have the highest proportion of "1"s. The bars for family sizes 3 and 4 have the largest orange sections, showing the highest relative share of the "1" category.

The proportion of "1"s varies across family sizes. The percentage of "1"s is not consistent for each family size, with a notable difference between the largest and smallest family sizes.

The majority of individuals in all family sizes are "0"s. The blue portion of all the bars is much larger than the orange portion, indicating that the "0" category is dominant across all family sizes.

Bivariate Analysis –Personal Loan Vs Security Amount



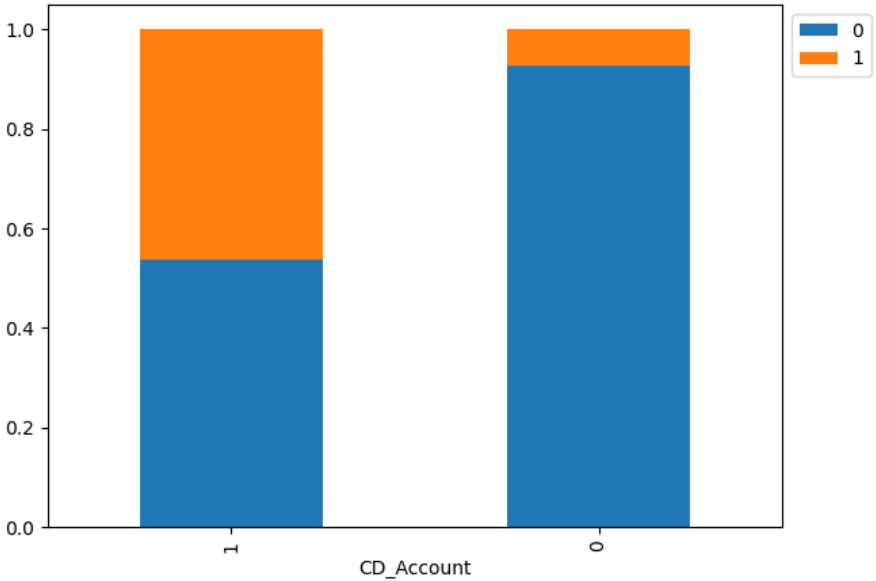
The vast majority of individuals, regardless of whether they have a securities account, fall into the "0" category. The blue part of both bars is significantly larger, showing that the outcome represented by "0" is far more common.

Individuals with a securities account ("1") have a slightly higher proportion of the "1" outcome. The orange segment for the "Securities_Account" category of "1" is slightly taller than the orange segment for the "0" category. This suggests a weak positive relationship between having a securities account and the "1" outcome.

The difference in the proportion of the "1" outcome between the two groups is very small. The heights of the orange sections are very similar, indicating that having a securities account has a minimal impact on the likelihood of the "1" outcome.

The outcome represented by "1" is a rare event for both groups. The orange segment for both "Securities_Account" categories is very small, representing less than 15% of the total in both cases.

Bivariate Analysis –Personal Loan Vs CD Account



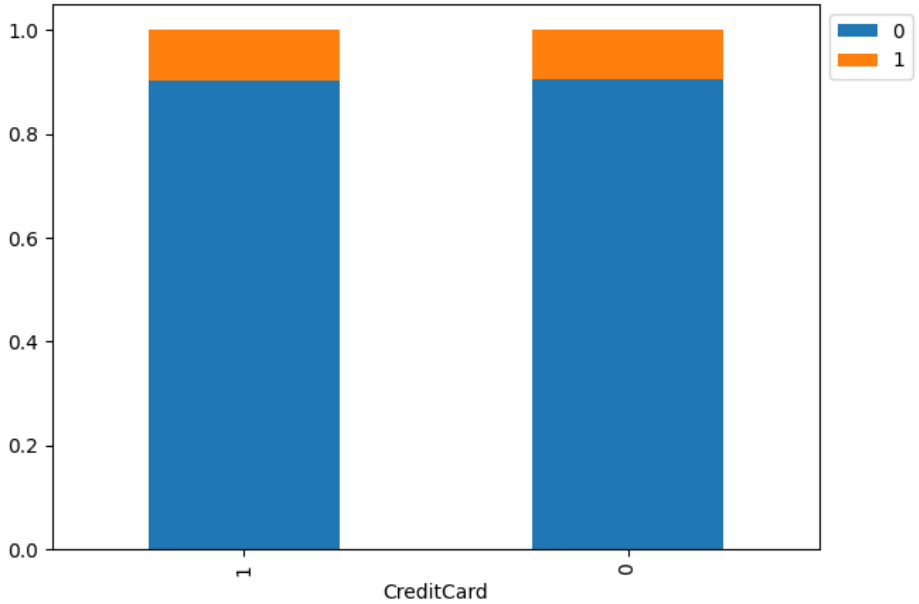
Having a CD account significantly increases the likelihood of the "1" outcome. The orange segment for individuals with a CD account (CD_Account = 1) is much larger, representing nearly half of the group.

Most individuals without a CD account fall into the "0" category. The blue portion of the bar for those without a CD account (CD_Account = 0) is very large, showing that the "0" outcome is highly dominant.

The proportion of the "1" outcome is much higher for those with a CD account. The height of the orange bar for "CD_Account" value "1" is almost half of the total bar height, while for value "0" it is very small.

The outcome represented by "0" is much more common overall. Even though having a CD account changes the proportions, the majority of the population in both groups still falls into the "0" category.

Bivariate Analysis –Personal Loan Vs Credit Card



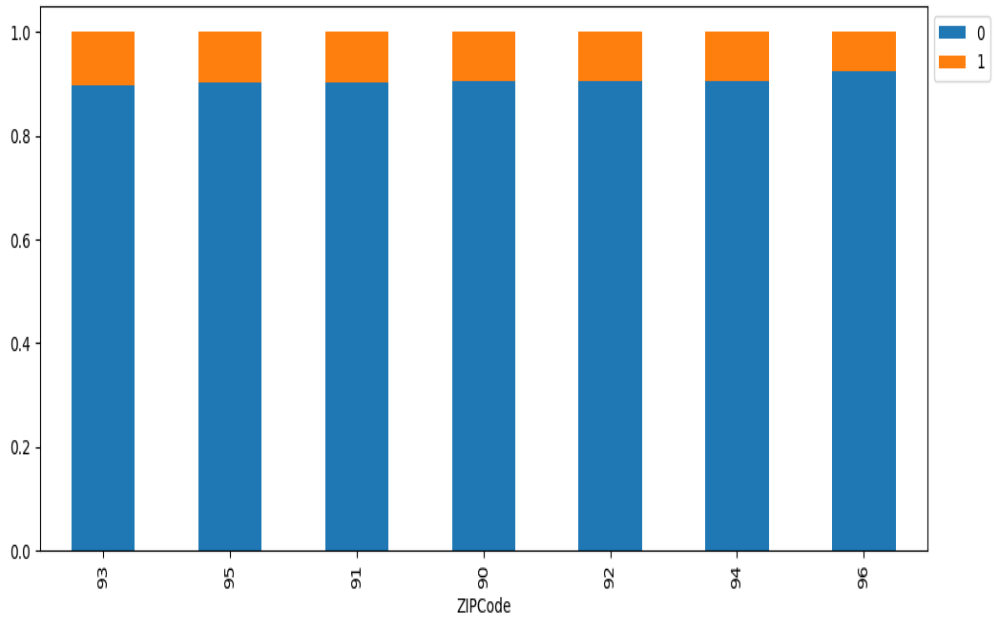
The vast majority of individuals, regardless of credit card ownership, fall into the "0" category. The blue section of both bars is significantly larger, indicating that the outcome represented by "0" is far more common for both groups.

Having a credit card does not seem to affect the outcome. The proportion of the "1" outcome (the orange section) is nearly identical for both people who have a credit card (CreditCard=1) and those who do not (CreditCard=0).

The outcome represented by "1" is a rare event for both groups. The orange segment for both categories is very small, representing less than 15% of the total in both cases.

The plot suggests no correlation between credit card ownership and the represented outcome. The lack of a notable difference in the proportion of "1"s indicates that having a credit card is not a strong predictor of the outcome.

Bivariate Analysis –Personal Loan Vs Zip Code



The vast majority of individuals in all ZIP codes fall into the "0" category. The blue section of every bar is significantly larger, showing that the outcome represented by "0" is far more common across all ZIP codes.

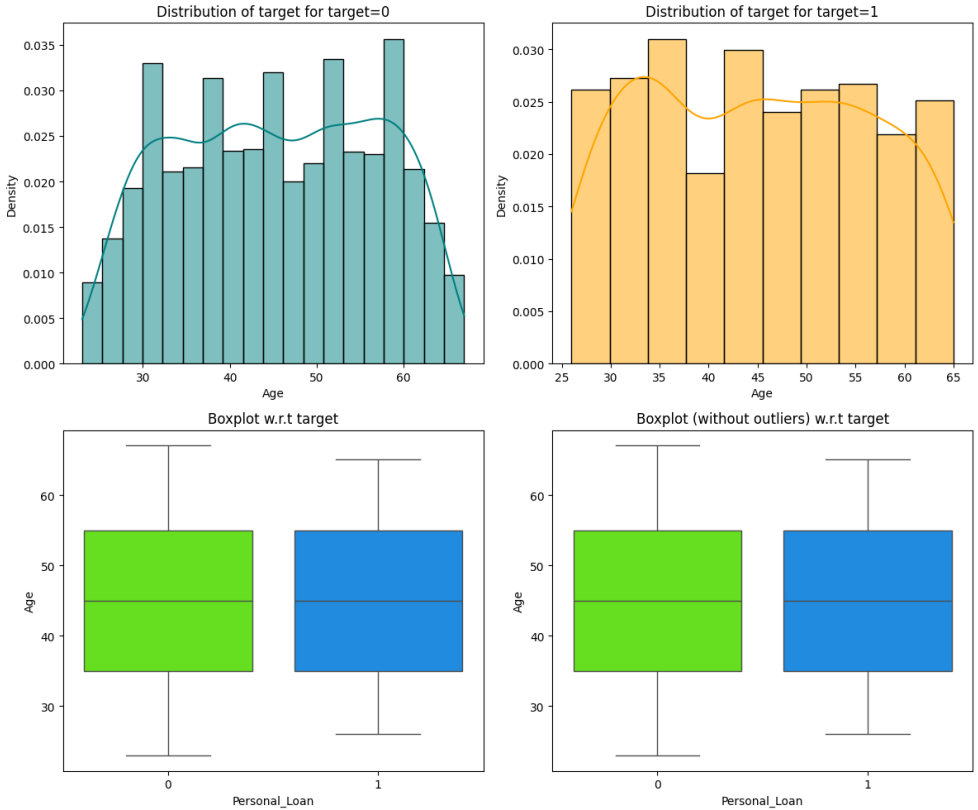
The proportion of the "1" outcome is highest for ZIP code 96. The orange segment for this ZIP code is the largest, indicating that individuals in ZIP code 96 are most likely to be in the "1" category.

The proportion of the "1" outcome is lowest for ZIP code 93. The orange segment for this ZIP code is the smallest, showing the lowest relative share of the "1" category.

There is a slight variation in the proportion of the "1" outcome across ZIP codes. While the overall trend is consistent, there are minor differences in the heights of the orange sections, suggesting a weak relationship between ZIP code and the outcome.

The outcome represented by "1" is a rare event for all ZIP codes. The orange segments are small across all bars, indicating that the "1" outcome is uncommon regardless of location.

Bivariate Analysis –Personal Loan Vs Age



The age distributions are quite similar for both target groups. Both the histogram for target=0 and target=1 show a roughly normal-like distribution, centered around similar age ranges (approximately 40 to 60). The shapes of the distributions are not dramatically different from each other.

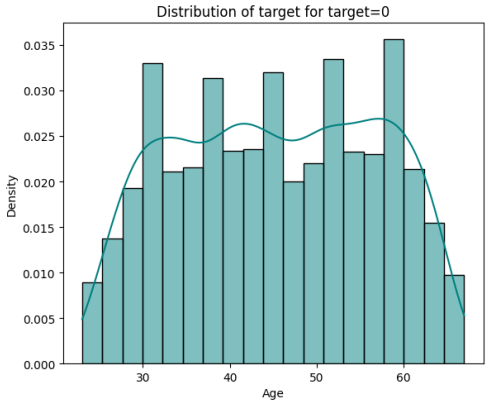
The median age is almost identical for both groups. The boxplots show that the horizontal line inside the boxes, which represents the median, is at approximately the same age for both Personal_Loan 0 and 1, suggesting that the median age is not a distinguishing factor.

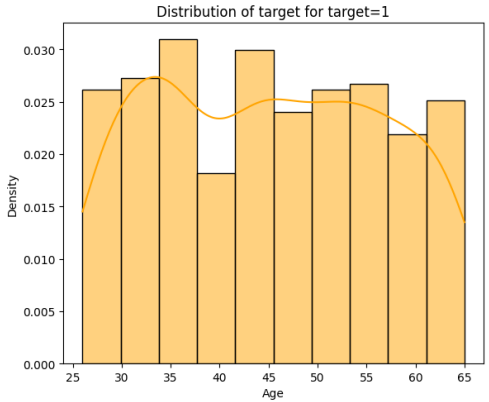
The interquartile ranges (IQR) are also very similar for both groups. The boxes in both boxplots are of comparable height, indicating that the middle 50% of the ages for both Personal_Loan 0 and 1 are spread over a similar range.

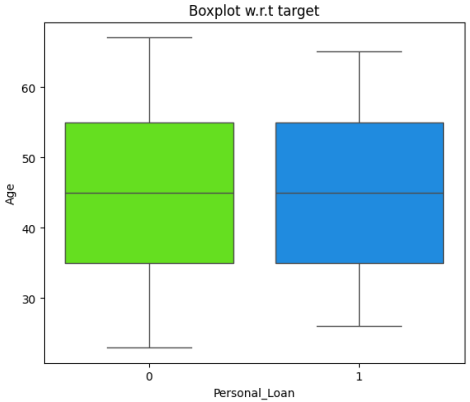
There is little to no difference between the two Personal_Loan groups in terms of age. All four plots consistently show that age, including its central tendency, spread, and overall distribution, is not a strong differentiating factor between people who take a personal loan and those who don't.

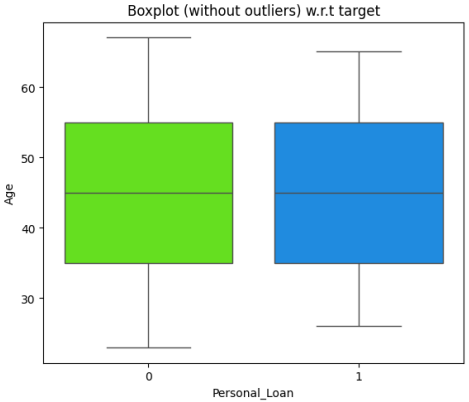
The distributions appear to have multiple peaks. The histograms show several peaks in the age distribution for both target groups, rather than a single, clear bell shape. This suggests that the data may contain a mix of different age cohorts.

Bivariate Analysis –Personal Loan Vs Experience









The experience distributions are very similar for both target groups. The histograms for Personal_Loan 0 and 1 have comparable shapes, indicating that experience isn't a strong distinguishing feature between the two groups.

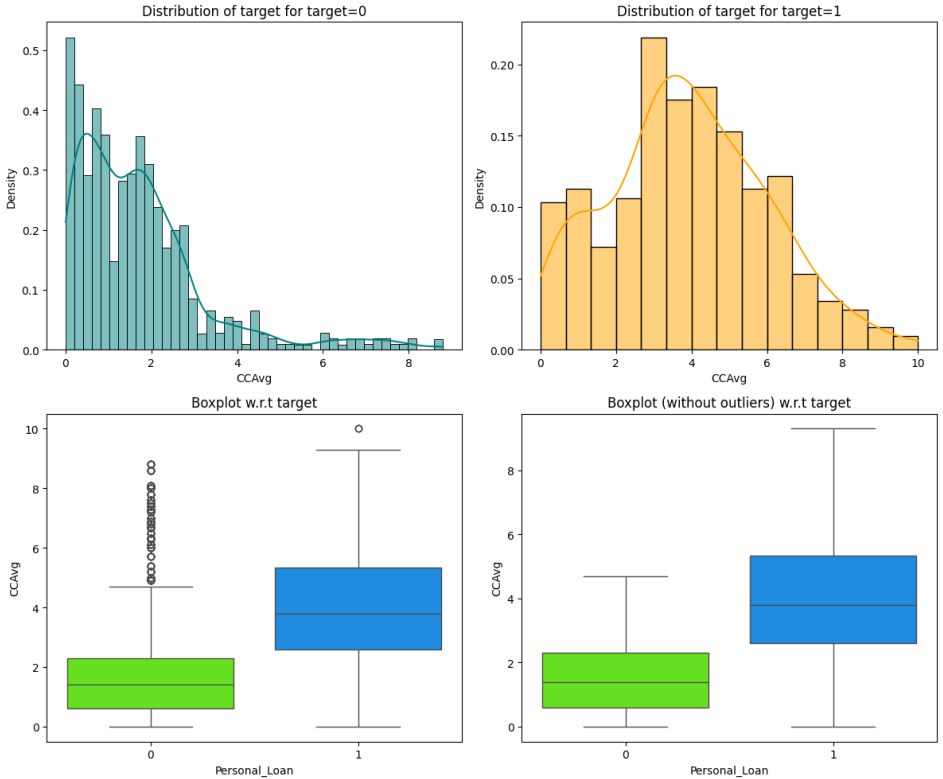
The median experience is almost identical for both groups. The boxplots show that the median line is at approximately the same level for both Personal_Loan 0 and 1, suggesting that the central tendency of experience is not a key differentiator.

The interquartile ranges (IQR) are also very similar for both groups. The boxes in the boxplots are of comparable size, indicating that the middle 50% of the experience data is spread over a similar range for both groups.

Overall, there is little to no difference in the distributions of experience for both Personal_Loan groups. All four plots consistently demonstrate that a person's years of experience is not a significant predictor of whether they will take a personal loan.

Both experience distributions are multi-modal. The histograms for both Personal_Loan 0 and 1 show several peaks rather than a single dominant one, suggesting a clustering of experience levels in the dataset.

Bivariate Analysis –Personal Loan Vs Income



Income is a strong predictor of taking a personal loan. Individuals who took a loan (Target = 1) have a significantly higher median and overall income distribution compared to those who didn't (Target = 0).

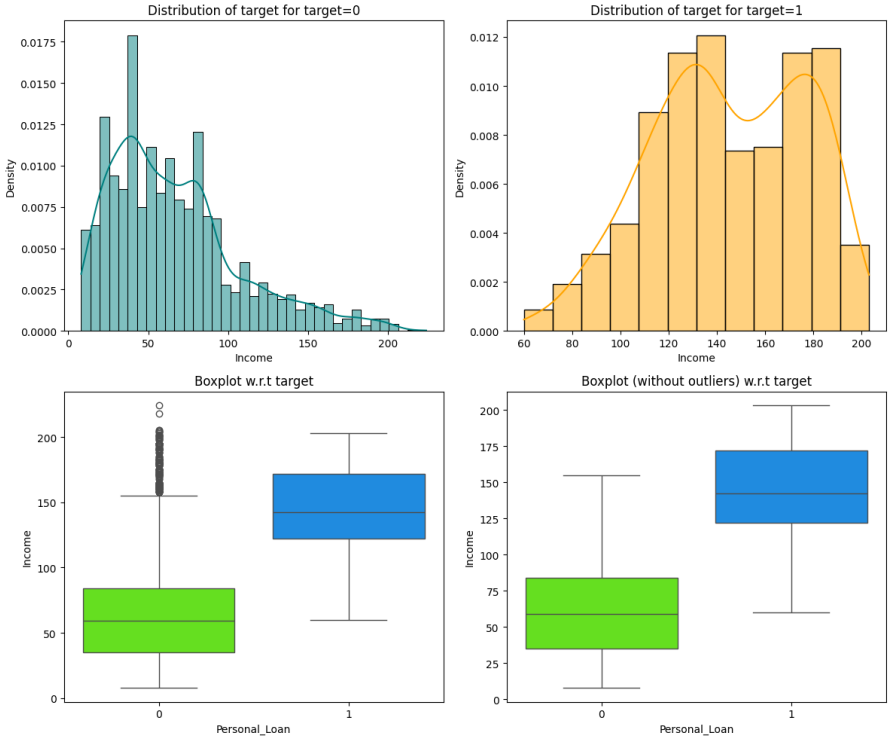
The median income for the Target=1 group is much higher. The boxplot shows the median income for the Personal_Loan group "1" is around 150, while the median for group "0" is only about 50.

The income distributions are very different for the two groups. The histogram for Target=0 is heavily skewed to the right with a long tail, while the histogram for Target=1 is more centered at higher income values.

There is greater variability and a presence of outliers in the Target=0 group. The boxplot for this group is much wider, and it contains numerous outliers, while the Target=1 group is more concentrated at higher values with fewer outliers.

The majority of individuals with high income belong to the Target=1 group. The histogram for Target=1 shows most of the data is above an income of 100, whereas for Target=0, the data is clustered below 100.

Bivariate Analysis –Personal Loan Vs Income



CCAvg is a strong predictor of a personal loan. The distribution of credit card spending is significantly different between those who took a loan (Target = 1) and those who didn't (Target = 0).

The median CCAvg for the Target=1 group is much higher. The boxplot shows the median for the Personal_Loan group "1" is about 4.0, while the median for group "0" is only about 1.5. This indicates that people with higher credit card spending are more likely to take a personal loan.

The distribution for Target=0 is highly skewed to the right. The histogram for this group shows a heavy concentration of data at low CCAvg values with a long tail, indicating that most people who didn't take a loan have low credit card spending.

The distribution for Target=1 is more symmetrical and centered at a higher value. The histogram for those who took a loan is centered around a higher CCAvg value, suggesting that this group's spending habits are generally higher.

There is a notable presence of outliers in the Target=0 group. The boxplot for this group shows numerous outliers with very high CCAvg values, which could be mislabeled or represent a small group of high spenders who did not take a personal loan.

EDA Questions

What is the distribution of mortgage attribute? Are there any noticeable patterns or outliers in the distribution?

- **The data is highly positively skewed**. The histogram shows a very long tail to the right, with most of the data clustered at the lower end of the mortgage values.
- **The median mortgage value is zero**. The green triangle in the box plot and the dashed green line in the histogram are both at the zero mark. This indicates that more than half of the individuals have a mortgage value of zero.
- **The majority of individuals have no mortgage**. The first bar of the histogram is significantly taller than all others, indicating a large number of people with a mortgage value of zero.
- **There is a large number of outliers with high mortgage** values. The individual circles to the right of the box plot's whisker represent many data points with mortgage values much higher than the majority.
- **The range of mortgage values is from zero up to approximately 650**. The histogram and the box plot show that while most values are low, the data extends all the way to around 650.

How many customers have credit cards?

- **Total 1470 Customers have credit cards**

What are the attributes that have a strong correlation with the target attribute (personal loan)?

- **Higher-income customers and those with higher average credit card spending** are significantly more likely to take personal loans. Customers with **undergraduate education show greater loan adoption compared to advanced degree holders and CD account ownership is a strong indicator of loan interest,**

EDA Questions

How does a customer's interest in purchasing a loan vary with their age?

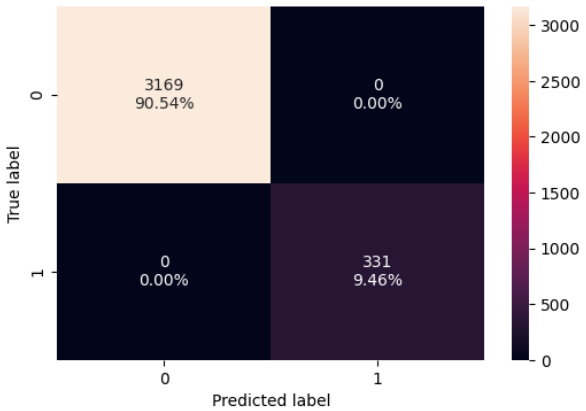
- There is no significant variation in customer interest in purchasing a loan with their age
- The data does not show a strong relationship between a customer's age and their likelihood of purchasing a loan

How does a customer's interest in purchasing a loan vary with their education?

- **Undergraduate (1):** Customers with an undergraduate degree show the highest proportion of loan purchases
- **Graduate (2):** The proportion of customers who bought a loan is lower for graduates compared to undergraduates.
- **Advanced/Professional (3):** Customers with an advanced or professional degree have the lowest proportion of loan purchases among the three groups.
- This trend suggests a negative correlation between education level and the likelihood of purchasing a loan, with the highest educated group showing the least interest.

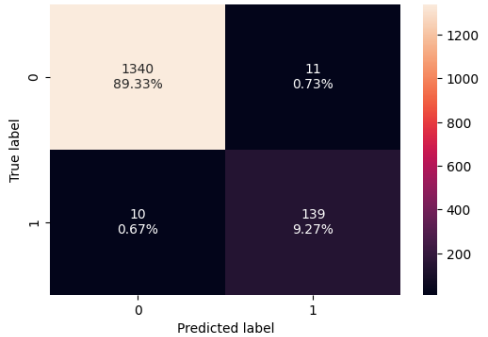
Default Model- Confusion Metrics

Model Performance on Training Data



- **High Accuracy TN :90 .54 %**
- **FP=0.00%**
- **FN: 0.00%**
- **TP : 9.46 %.**

Model Performance on Test Data



- **High overall Accuracy TN :89 .33 %**
- **TP:9.27 %**
- **Excellent recall for class 0**
- **Low False positive rate : 0.73 %**
- **Low False Negative rate FN: 0.67%** this indicate the model is doing good job and not missing the positive cases
- **Data is imbalance :**The model's strong performance on the majority class (0) and good, but lower, performance on the minority class (1) is a typical pattern in such unbalanced datasets.

Default Model-Performance Comparison

Training Data

Accuracy	Recall	Precessi on	F1 Score
1,0	1.0	1.0	1.0

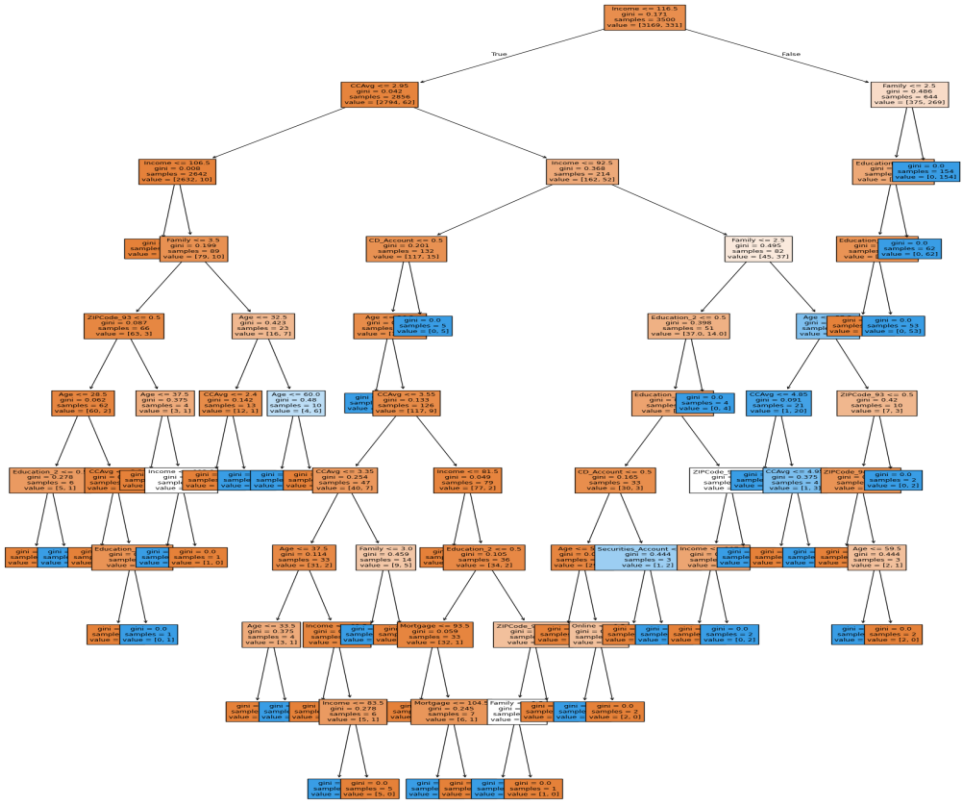
Test Data

Accuracy	Recall	Precessi on	F1 Score
0.986	0.932886	0.926667	0.929766

Observations

- **Perfect Training Performance:** The model achieved a perfect score of 1.0 across all metrics (Accuracy, Recall, Precision, and F1-Score) on the training data. This indicates that the model has fully learned or even memorized the training dataset.
- **Slight Drop in Overall Accuracy:** The model's accuracy on the test data is 0.986, which is slightly lower than the perfect 1.0 achieved on the training data. While this is an excellent score, the drop suggests the model is not a perfect fit for unseen data.
- **Significant Drop in Recall and Precision**: The most notable observation is the decrease in the individual class-based metrics. Recall drops from 1.0 on training to 0.933 on the test data, and Precision drops from 1.0 to 0.927. This indicates the model is making more errors (false positives and false negatives) on new data compared to the training data.
- **Decline in the F1-Score:** The F1-Score, which provides a balanced view of both precision and recall, drops from a perfect 1.0 to 0.930. This confirms that the model's overall performance has degraded on the test set, reinforcing the observation that the drop in recall and precision is significant.
- **Lead to Overfitting:** The combination of perfect scores on the training data and the subsequent drop in all metrics on the test data is a classic sign of overfitting. The model has learned the training data too well, including its noise, and is failing to generalize to new, unseen data as effectively.

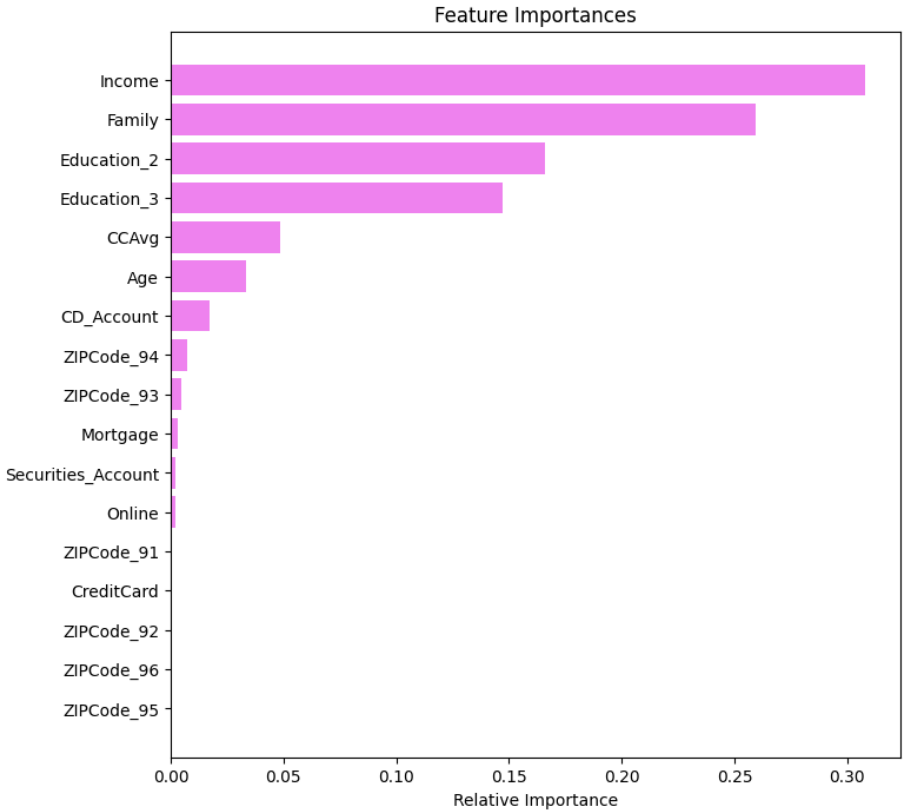
Default Model –Decision Tree



Key Observations :

- **Income is the most important feature**. The root node, which is the starting point of the tree, makes its first split based on the Income <= 116.5 feature. This indicates that income is the single most significant factor in classifying the data, as it provides the greatest reduction in Gini impurity from the very beginning.
- **The model is complex and may be overfitted**. The tree is very deep, with numerous splits and branches
- **Gini impurity is significantly reduced in later stages**. While the root node starts with a Gini of 0.5, many of the leaf nodes have a Gini of 0.0. This means that the decision paths are effectively creating very pure, homogeneous groups, where all samples in a leaf node belong to a single class.
- **A variety of features are used. Besides Income,** the tree uses a wide range of features like CCAvg (Average Credit Card Spending), Family, Education, CD_Account, Age, Mortgage, ZIPCode, Online, and Securities Account. This indicates that the final classification is determined by a complex combination of various factors, not just a few key variables.
- Some leaf nodes are based on very specific, narrow conditions. The tree has created numerous small, highly specific leaf nodes (e.g., nodes with only 1 or 2 samples). For example, there's a leaf node with samples = 1 and value = [1, 0]. **These narrow conditions are a strong indicator of overfitting, as the model has created rules that are too tailored to individual data points in the training set.**

Default Model –Feature Importance

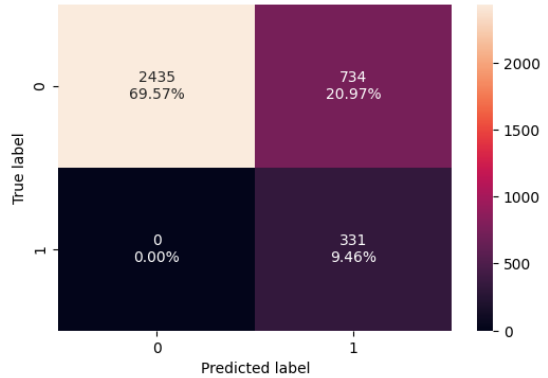


Observations

- ✓ In default model Income Family and education_2 are the three important features

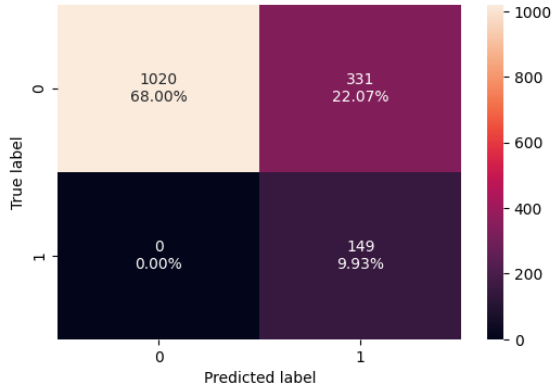
Pre-pruned Model-Confusion Metrics

Model Performance on Training Data



- **True Positive (TP):** The model correctly predicted the positive class. **TP=9.46%**
- **True Negative (TN)**: The model correctly predicted the negative class. **TN =69.57%.**
- **False Positive (FP): 20.97%**
- **False Negative (FN): 0.00 %**

Model Performance on Test Data

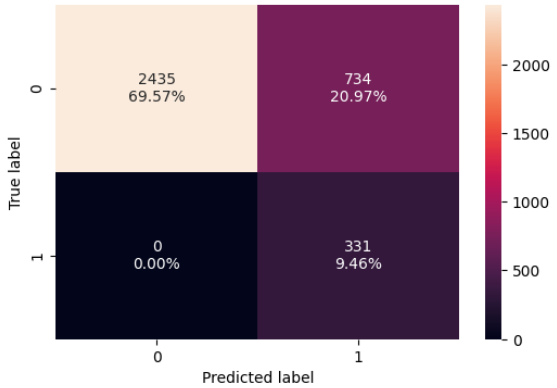


- **True Positive (TP):** The model correctly predicted the positive class. **TP=9.93%**
- **True Negative (TN)**: The model correctly predicted the negative class. **TN =68.00%.**
- **False Positive (FP):** The model incorrectly predicted the positive class. **FP = 22.07 %**
- **False Negative (FN): = 0.00 %**
- This leads , the model is highly predicting the negative value and imbalances nature of the data on high accuracy on class 0 and poor performance on class 1

Pre-pruned Model- Performance

Model Performance on Training Data

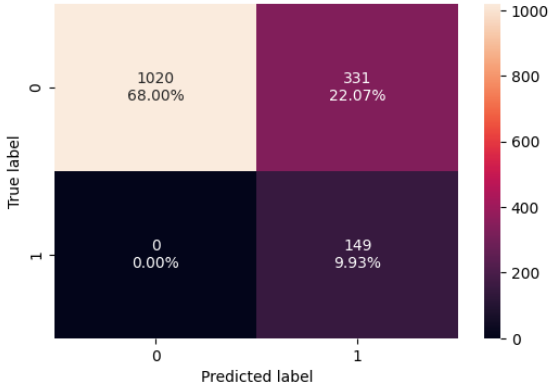
Confusion Metrix



Accuracy	Recall	Precession	F1 Score
0.790286	1.0	0.310798	0.474212

Model Performance on Test Data

Confusion Metrix

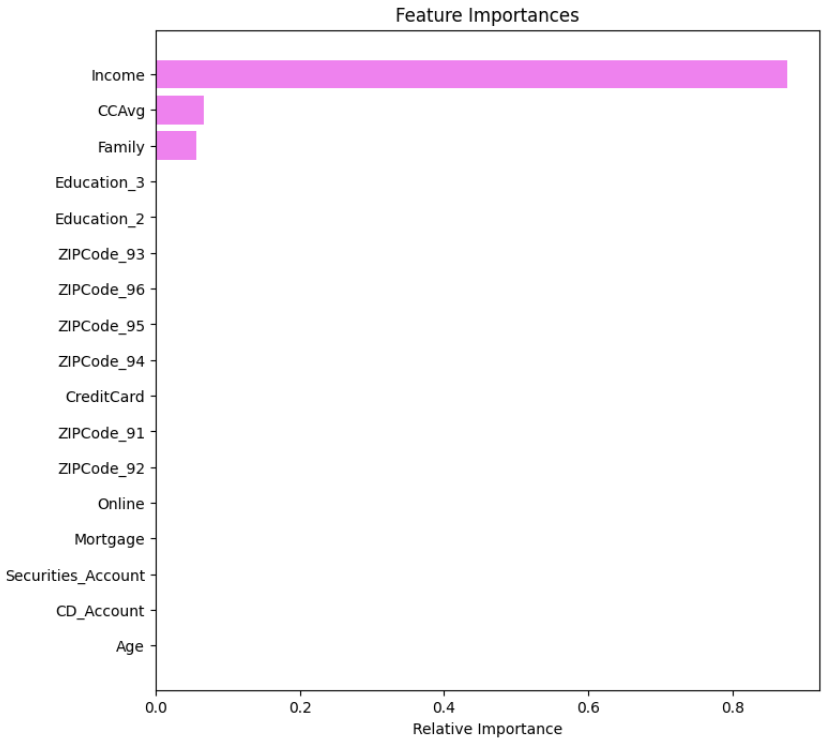


Accuracy	Recall	Precessi on	F1 Score
0.779333	1.0	0.310417	0.473768

Observations

- This model shows signs of **underfitting**.
- Its performance on both the training and **test sets is poor**,
- especially its **precision** (0.310798 on training and 0.310417 on test).
- While its recall is perfect at 1.0, this can be misleading. A perfect recall can be achieved by simply predicting every instance as positive, but this would lead to a very low precision, which is what we see here.
- This model is **too simple and cannot accurately distinguish between classes**

Pre-pruned model Feature Importance

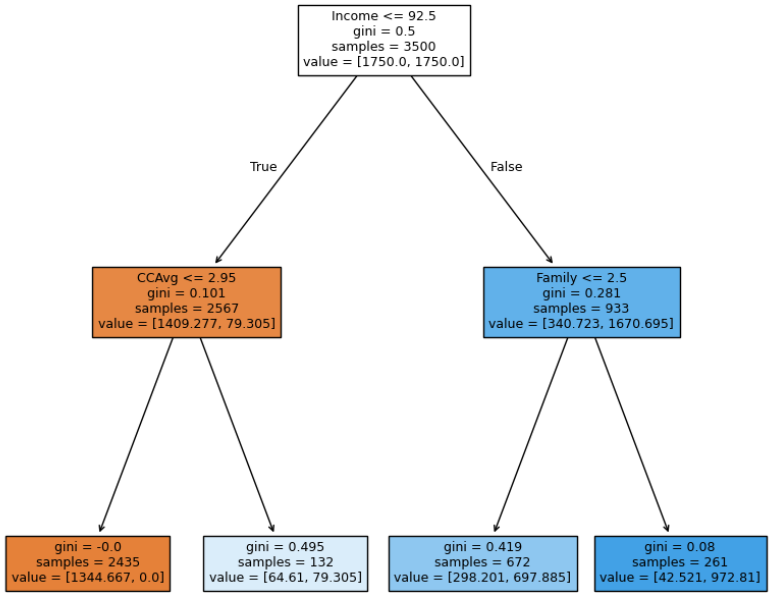


Observations

In Pre-Pruned model below two are the important features

- ✓ **Income**
- ✓ **CCAvg**

Pre- Pruned – Decision Tree



Key Observations

- ✓ **Income is the most important feature**. The tree's root node uses Income **<= 92.5** to make the first split. This signifies that Income is the single most influential factor in classifying the data, as it provides the greatest reduction in impurity from the very beginning.

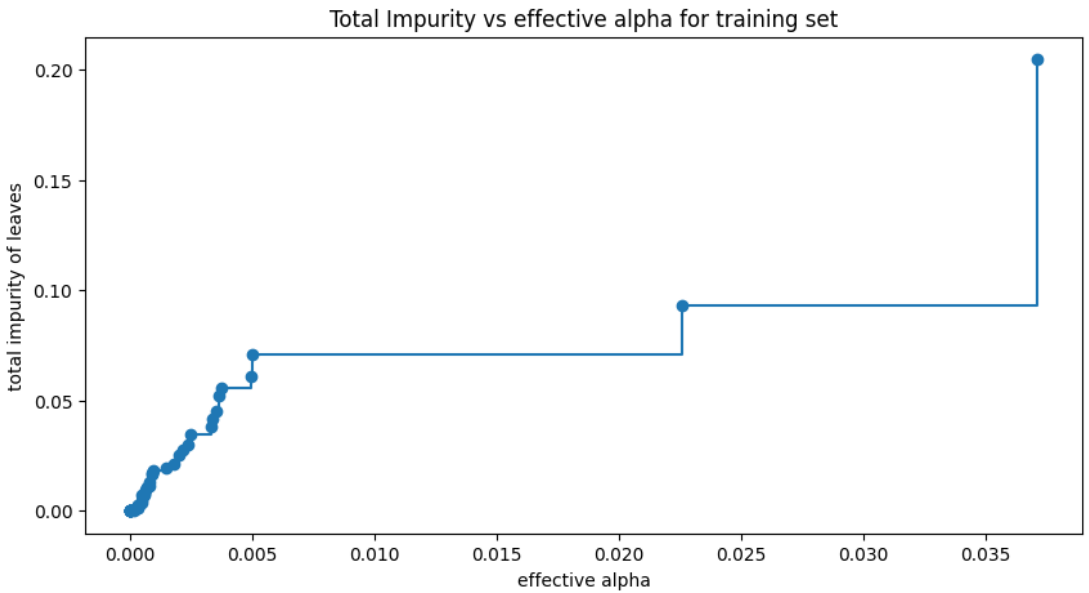
- ✓ The tree is relatively shallow. Unlike a complex tree that branches out many times, this tree has a maximum depth of three. This suggests that the model is less prone to overfitting because it's not creating overly specific rules for the training data. The model is capturing the main patterns without memorizing noise.

- ✓ The root node has a maximum Gini impurity. The gini value at the root node is 0.5. This indicates a perfectly mixed or random distribution of the two classes within the initial 3500 samples, which is expected before any splits are made.

- ✓ Significant impurity reduction occurs at the first split. After the initial split on Income, the Gini impurity drops considerably in the subsequent nodes. The "True" branch has a gini of 0.101, and the "False" branch has a gini of 0.281. This shows that the Income split was highly effective at separating the two classes.

- ✓ CCAvg and Family are the next most important features. After the initial Income split, the next level of **splits uses CCAvg <= 2.95 and Family <= 2.5.** This indicates that average credit card spending and the number of family members are the next most predictive features for the target variable.

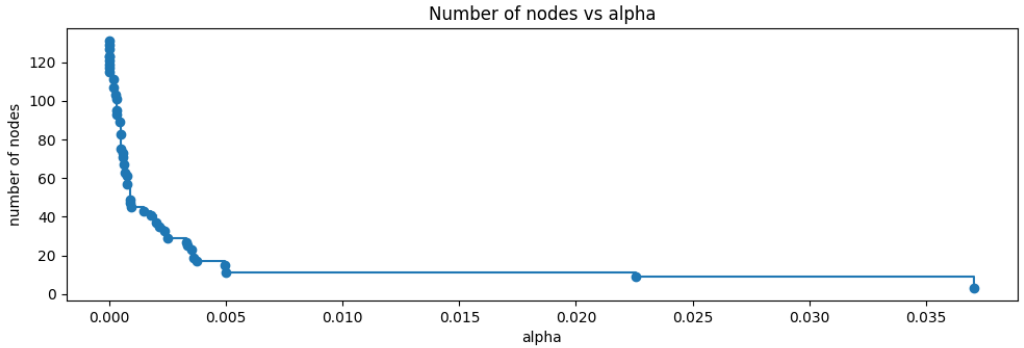
Post-Pruned Total Impurity vs alpha

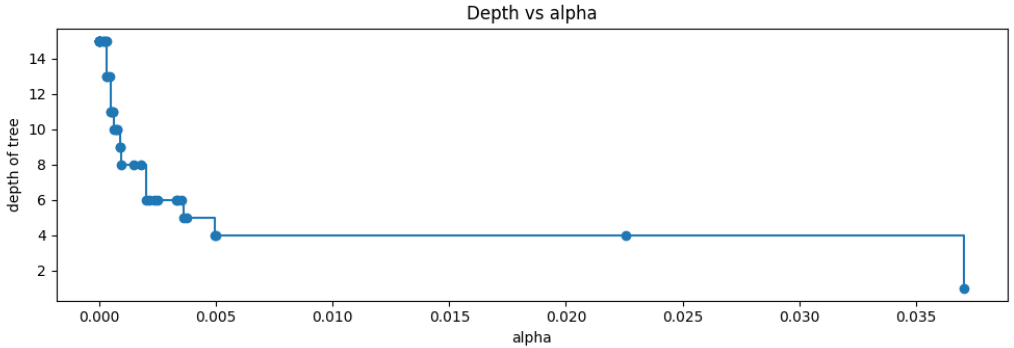


Key Observations

1. Total impurity of the leaves decreases as effective alpha decreases.
2. The impurity drops sharply at a specific effective alpha value of approximately 0.022.
3. The plot has distinct steps, indicating that the impurity remains constant over a range of effective alpha values.
4. The lowest impurity is achieved at effective alpha values close to zero.
5. A high effective alpha value (e.g., > 0.035) results in a high total impurity.

Post-pruned Number of nodes and depth vs alpha





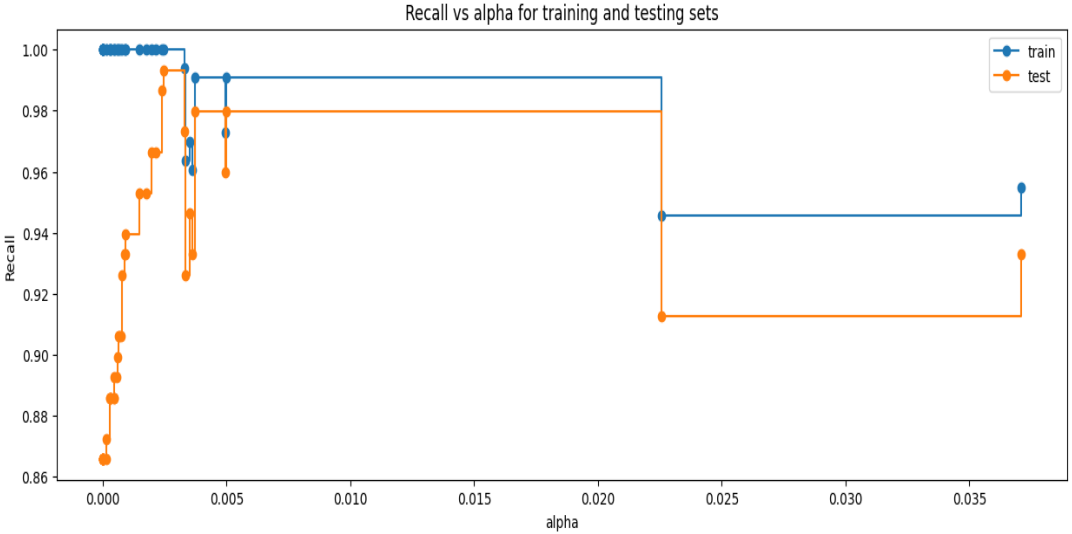
Number of nodes vs alpha

- As alpha increases, the number of nodes in the decision tree sharply decreases.
- The most significant reduction in nodes occurs at very small values of alpha.
- The plot shows a staircase-like pattern, indicating that the number of nodes remains constant over a range of alpha values.
- The tree prunes heavily with even a slight increase in alpha, especially at the beginning of the plot.
- The tree is pruned down to a single node (the root) at the largest alpha value shown.

Depth Vs Alpha

- As alpha increases, the depth of the tree also decreases, similar to the number of nodes.
- The tree's depth drops rapidly at smaller alpha values, from its maximum depth down to a much smaller value.
- The depth remains constant over certain intervals of alpha, creating a step function.
- The tree depth is reduced from its maximum of over 14 down to a depth of around 4 relatively quickly, and
- The final large value of alpha reduces the tree to its minimum depth of 1 (a single root node)

Post-Pruned Recall Vs alpha for training and test data

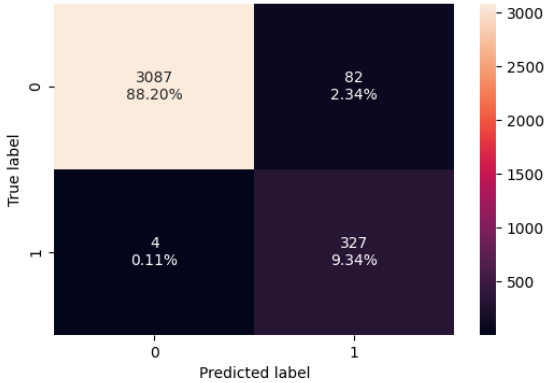


Key Observations

- The training recall is consistently higher than the testing recall for all alpha values, indicating some degree of overfitting.
- For small values of alpha (close to 0), the training recall is nearly perfect (1.0), while the testing recall shows significant fluctuations.
- The testing recall is highly sensitive to small changes in alpha at the beginning, with sharp drops and rises.
- A large increase in alpha to about 0.022 causes a significant drop in both training and testing recall, suggesting a substantial loss in true positive identification.
- The testing recall is more stable for alpha values between approximately 0.0026 and 0.022

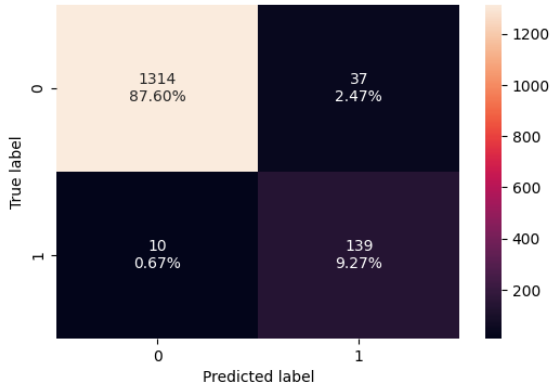
Post-pruned Model- Confusion Metrics

Model Performance on Training Data



- **True Positive (TP):** The model correctly predicted the positive class. **TP=9.34%**
- **True Negative (TN)**: The model correctly predicted the negative class. **TN =88.20%.**
- **False Positive (FP):** The model incorrectly predicted the positive class**. FP = 2.34 %**
- **False Negative (FN):** The model incorrectly predicted the negative class **FN = 0.11 %**

Model Performance on Test Data

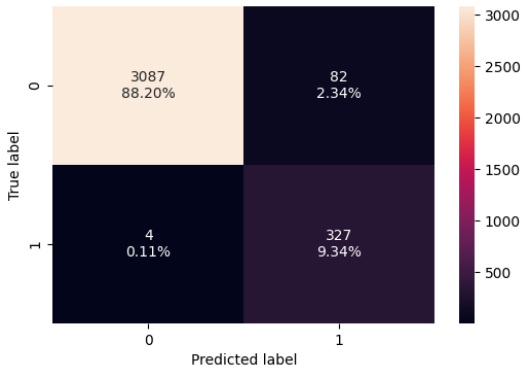


- **True Positive (TP):** The model correctly predicted the positive class. **TP=9.27%**
- **True Negative (TN)**: The model correctly predicted the negative class. **TN =87.60%.**
- **False Positive (FP):** The model incorrectly predicted the positive class**. FP = 2.47 %**
- **False Negative (FN):** The model incorrectly predicted the negative class **FN = 0.67 %**

Post-pruned Model-Model Performance

Model Performance on Training_Data

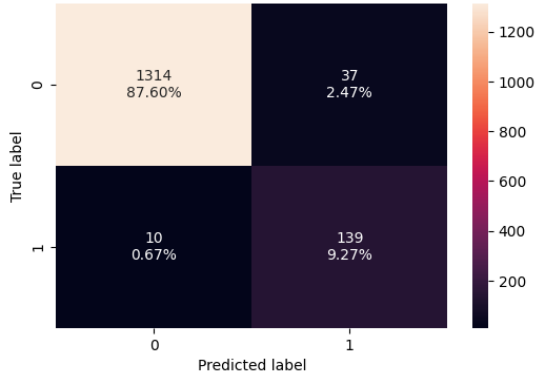
Confusion Metrix



Accuracy	Recall	Precession	F1 Score
0.975429	0.987915	0.799511	0.883784

Model Performance on Test_Data

Confusion Metrix



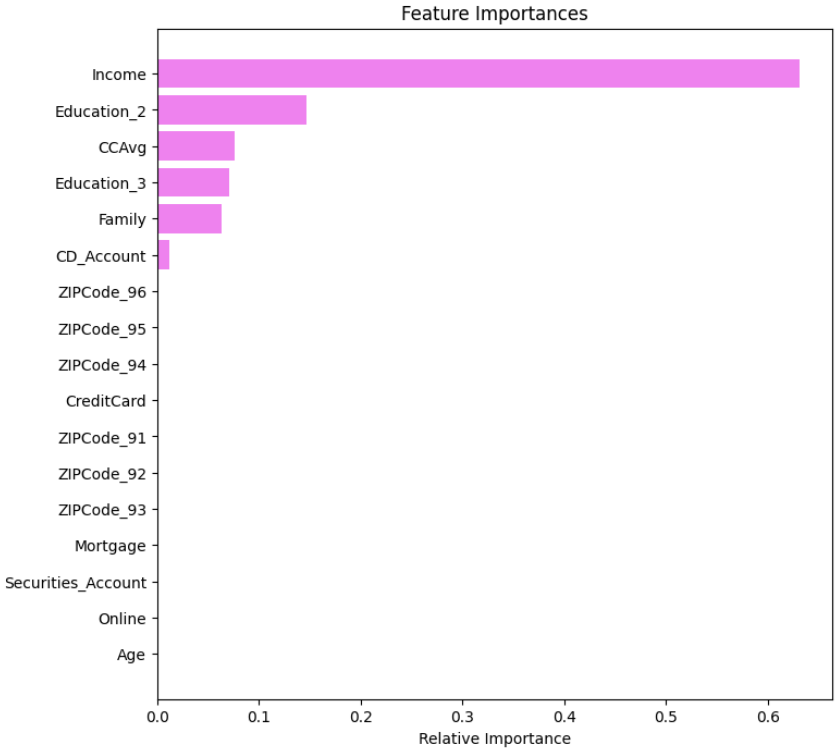
Accuracy	Recall	Precession	F1 Score
0.968667	0.932886	0.789773	0.855385

Best ccp_alpha =0.0025
Selected based on the **argmax(recall_test)** result

Observations

- ✓ Accuracy : Very little difference between test and training data (0.006762)
- ✓ Recall : Little difference between test and training data (0.055029)
- ✓ **Precession :** Very little difference between test and training data (0.009738)
- ✓ F1 Score-Little difference between test and training data (0.028399)

Post-Pruned model Feature Importance

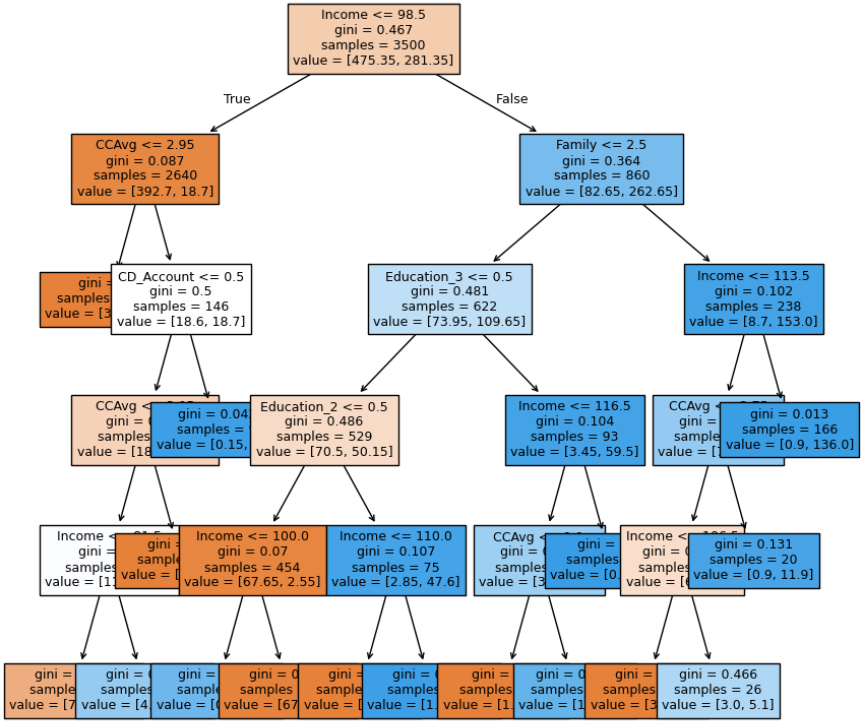


Key Observations

In post pruned decision tree below are the two most important features

- ✓ **Income**
- ✓ **Education_2**

Post- Pruned – Decision Tree



Key Observations

- Income is the most important feature. The root node, which makes the first split, is based on the **Income <= 98.5 feature**. **This indicates that income is the single most significant factor in classifying the data**.
- The tree is relatively deep and complex. The tree extends to multiple levels, suggesting that the model is using several features and decision rules to make its predictions.
- Gini impurity is significantly reduced with each split. The gini value at the root node is 0.467, which is close to the maximum possible impurity (0.5 for a two-class problem). Subsequent splits, such as the one at the root, result in a significant drop in Gini impurity for **the child nodes (e.g., 0.087 and 0.364).** This shows that the splits are effective at creating more homogeneous groups.
- CCAvg and Family are also important features. While Income is the root, features like CCAvg (Average Credit Card spending) and Family are used in the first level of splits.
- Some leaf nodes are very pure. Several leaf nodes (the nodes at the bottom of the tree) have a gini value of 0. This indicates that these nodes contain samples that belong to a single class, representing a very confident and pure prediction. **For example, the leaf node with a value of [4, 0] and gini=0 is 100% composed of one class**.

Models Comparison & Recommendations

Training Performance Comparison

Performance Metrics	Decision Tree (Sklearn Default)	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	1.0	0790286	**0.975429**
Recall	1.0	1.00000	**0.987915**
Precession	1.0	0.310798	**0.799511**
F1-Score	1.0	0474212	**0.883784**

Test Performance Comparison

Performance Metrics	Decision Tree (Sklearn Default)	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.986000	0.779333	**0.968667**
Recall	0.932886	1.00000	**0.932886**
Precession	0.926667	0.310417	**0.789773**
F1-Score	0.929766	0.473768	**0.855385**

Recommendation

Based on the comparison **Post-Pruned is recommended model**

Why Post-Pruning is the Best Model1. Decision Tree (sklearn default)

- ✓ This model has a perfect **training performance** with 100% accuracy, recall, precision, and F1 score. This is a classic sign of **overfitting**. The model has essentially memorized the training data, including its noise, and is not learning the underlying patterns. This is confirmed by its significantly lower performance on the test set ,a recall of **0.932886** and a precision of **0.926667**, and F1 score 0.929776 a clear drop from its perfect training scores.

2. Decision Tree (Pre-Pruning)

- ✓ This model shows signs of **underfitting**. Its performance on both the training and test sets is poor, especially its **precision** (0.310798 on training and 0.310417 on test). While its recall is perfect at 1.0, this can be misleading. A perfect recall can be achieved by simply predicting every instance as positive, but this would lead to a very low precision, which is what we see here. This model is too simple and cannot accurately distinguish between classes.

3. Decision Tree (Post-Pruning)

- ✓ This model strikes the best balance between **bias** and **variance** (underfitting and overfitting).
- ✓ It has a high training recall (**0.987915**) and precision (**0.799511**), showing it learned the data well without memorizing it perfectly.
- ✓ Crucially, its test set performance is very close to its training performance, with a test recall of **0.932886** and a test precision of **0.789773**. The small difference between training and test scores indicates that **the model generalizes effectively to new data**.
- ✓ This model has the highest **F1 score** on training set (**0.883784**) and little lower on the test set (**0.855385**), which is often the most important metric for evaluating a model. The F1 score is the harmonic mean of precision and recall and provides a single, balanced metric for a model's performance, especially when there's an imbalance between classes.

Actionable Insights and Business Recommendations

- Deploy the **Post-Pruned Decision Tree model for targeted marketing, focusing on high-income, high-CCAvg segments.**
- Optimize marketing ROI by prioritizing outreach to customers **with CD accounts and undergraduate-level education.**
- Adopt probability-based targeting: instead of binary predictions, use the model's probability scores to rank customers by loan likelihood.
- Refine campaign strategy: customers below a defined threshold (**e.g., <60% probability) should be routed for manual review, reducing false targeting costs**

APPENDIX

Post –Pruned Tree Decesion rules

```
|--- Income <= 98.50
|   |--- CCAvg <= 2.95
|   |   |--- weights: [374.10, 0.00] class: 0
|   |--- CCAvg >  2.95
|   |   |--- CD_Account <= 0.50
|   |   |   |--- CCAvg <= 3.95
|   |   |   |   |--- Income <= 81.50
|   |   |   |   |   |--- weights: [7.35, 2.55] class: 0
|   |   |   |   |--- Income >  81.50
|   |   |   |   |   |--- weights: [4.35, 9.35] class: 1
|   |   |   |--- CCAvg >  3.95
|   |   |   |   |--- weights: [6.75, 0.00] class: 0
|   |   |--- CD_Account >  0.50
|   |   |   |--- weights: [0.15, 6.80] class: 1
|--- Income >  98.50
|   |--- Family <= 2.50
|   |   |--- Education_3 <= 0.50
|   |   |   |--- Education_2 <= 0.50
|   |   |   |   |--- Income <= 100.00
|   |   |   |   |   |--- weights: [0.45, 1.70] class: 1
|   |   |   |   |--- Income >  100.00
|   |   |   |   |   |--- weights: [67.20, 0.85] class: 0
|   |   |   |--- Education_2 >  0.50
|   |   |   |   |--- Income <= 110.00
|   |   |   |   |   |--- weights: [1.80, 0.00] class: 0
|   |   |   |   |--- Income >  110.00
|   |   |   |   |   |--- weights: [1.05, 47.60] class: 1
|   |   |--- Education_3 >  0.50
|   |   |   |--- Income <= 116.50
|   |   |   |   |--- CCAvg <= 1.10
|   |   |   |   |   |--- weights: [1.95, 0.00] class: 0
|   |   |   |   |--- CCAvg >  1.10
|   |   |   |   |   |--- weights: [1.50, 6.80] class: 1
|   |   |   |--- Income >  116.50
|   |   |   |   |--- weights: [0.00, 52.70] class: 1
|   |--- Family >  2.50
|   |   |--- Income <= 113.50
|   |   |   |--- CCAvg <= 2.75
|   |   |   |   |--- Income <= 106.50
|   |   |   |   |   |--- weights: [3.90, 0.00] class: 0
|   |   |   |   |--- Income >  106.50
|   |   |   |   |   |--- weights: [3.00, 5.10] class: 1
|   |   |   |--- CCAvg >  2.75
|   |   |   |   |--- weights: [0.90, 11.90] class: 1
|   |   |--- Income >  113.50
|   |   |   |--- weights: [0.90, 136.00] class: 1
```

Great Learning

Happy Learning !