AllLife Bank Personal Loan Campaign

SUPERVISED LEARNING CLASSIFICATION PROJECT

AIML UNIVERSITY OF TEXAS AUSTIN

JULY 21, 2023

PREPARED BY GREG WENZEL

Problem Statement

Context

AllLife Bank is a US bank that has a growing customer base. The majority of these customers are liability customers (depositors) with varying sizes of deposits. The number of customers who are also borrowers (asset customers) is quite small, & the bank is interested in expanding this base rapidly to bring in more loan business & in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio.

You as a Data scientist at AllLife bank have to build a model that will help the marketing department to identify the potential customers who have a higher probability of purchasing the loan.

Problem Statement Continued

Objective

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

Data Dictionary

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

Contents / Agenda

Executive Summary

- AllLife Bank is a growing financial institution with a focus on liability customers (depositors).
- To further expand its business & increase interest earnings, the bank aims to convert liability customers into personal loan customers while retaining them as depositors.
- A previous campaign targeting liability customers resulted in a healthy conversion rate of over 9%, encouraging the bank's retail marketing department to improve target marketing strategies for higher success ratios.
- To address this business objective, a data science project was undertaken to build a predictive model that identifies potential customers with a higher probability of purchasing a personal loan. The project aimed to understand the significant customer attributes driving loan purchases & determine the target segment for effective marketing.

Executive Summary Key Findings & Insights

- **Target Market Identification:**
  - Through data analysis & modeling, we identified specific customer attributes that significantly influence the likelihood of purchasing a personal loan.
  - Age, income, education level, family size, average credit card spending, & previous banking activities (such as having a certificate of deposit or securities account) emerged as important factors in predicting loan acceptance.
  - By targeting customers within specific age groups, income brackets, education levels, & family sizes, AllLife Bank can optimize its marketing efforts & improve the conversion rate.
- **Feature Importance:**
  - The analysis revealed that income is the most significant feature in driving loan purchases, followed by family size & education level.
  - Customers with higher incomes & larger families showed a higher tendency to accept personal loan offers.
  - By leveraging these insights, AllLife Bank can allocate its marketing resources effectively & tailor its loan offers to target customers who are more likely to convert.

Executive Summary Recommendations

1. Targeted Marketing Strategies:
   - Based on the identified key customer attributes, AllLife Bank should tailor its marketing campaigns to specific customer segments.
   - Develop personalized offers & incentives that align with the financial needs, preferences, & lifestyles of the target customers.
   - Utilize digital marketing channels & personalized communication to reach out to potential loan customers effectively.
2. Enhanced Customer Profiling:
   - Continuously collect & update customer data to refine customer profiles & segmentations.
   - Leverage data analytics techniques to identify new patterns & trends, enabling the bank to refine its target audience & enhance marketing strategies.
3. Customer Engagement & Relationship Management:
   - Foster strong customer relationships by providing personalized & superior customer experiences.
   - Offer exceptional customer service, convenient digital banking options, & personalized loan recommendations based on individual financial goals & needs.

Executive Summary Conclusion

Significant features providing potential for personal loan consumption and targeting for current AllLife Bank customers, as represented by the provided data set are:

1. Income : Gini Importance = 0.674927
2. Family Size : Gini Importance = 0.171953
3. Professional Education Sector : Gini Importance = 0.153126

Income is by far the most significant predictive feature in determining the target variable. Family Size and Education rank in the top three though their importance is not as significant as Income.

- By leveraging data-driven insights & implementing targeted marketing strategies, AllLife Bank can effectively convert liability customers into personal loan customers. The identified customer attributes & feature importance provide a solid foundation for personalized marketing campaigns, allowing the bank to maximize loan acceptance rates & achieve business growth.
- It is crucial for AllLife Bank to continually evaluate & refine its marketing strategies based on customer feedback, market dynamics, & emerging trends to stay competitive in the industry. By embracing data science & leveraging customer data effectively, the bank can drive business success & enhance customer satisfaction.

	Imp
Income	0.674921
Family	0.171953
Education_Undergraduate	0.153126
ZIPCode_92	0.000000
Education_Professional	0.000000
Education_Graduate	0.000000
ZIPCode_96	0.000000
ZIPCode_95	0.000000
ZIPCode_94	0.000000
ZIPCode_93	0.000000
Age	0.000000
ZIPCode_91	0.000000
CreditCard	0.000000
Online	0.000000
CD_Account	0.000000
Securities_Account	0.000000
Mortgage	0.000000
CCAvg	0.000000
ZIPCode_90	0.000000

Business Problem Definition & Solution Approach

1. The problem:

- The problem is to build a predictive model that can identify customers who are more likely to accept personal loan offers.
- The model will assist the marketing department in targeting the right customers & allocating marketing resources effectively.

2. Solution Approach:

- The solution approach involves employing supervised machine learning techniques, specifically classification algorithms, to develop the predictive model.
- The model will be trained on historical data that includes various customer attributes such as age, income, education level, family size, credit card spending, & previous banking activities.
- The target variable will be the acceptance or rejection of the personal loan offer.

By following this approach, AllLife Bank will be equipped with a predictive model that can effectively identify potential customers who are more likely to accept personal loan offers. This will enable the bank to focus its marketing efforts on the right target audience, increasing the conversion rate & achieving its business objectives.

Business Problem Overview Methodology

1. Data Exploration & Preprocessing:
   - The dataset will be explored to understand data distribution, identify missing values, & handle any outliers or anomalies.
2. Feature Selection & Engineering:
   - The importance of features will be analyzed to determine which attributes have the most significant impact on loan acceptance.
3. Model Training & Evaluation:
   - The dataset will be divided into training & testing sets to train & evaluate the performance of the predictive model.
   - Classification algorithms, such as decision trees, roman forests & logistic regression will be considered & compared.
   - Evaluation metrics, including accuracy, recall, precision, & F1 score, will be used to assess the model's performance.
4. Model Fine-Tuning & Optimization:
   - Hyperparameter tuning techniques, such as grid search or random search, will be employed to optimize the model's performance.
5. Model Deployment & Monitoring:
   - The final trained model will be deployed to predict the likelihood of loan acceptance for new customers.
   - Regular monitoring of model performance & feedback will be essential to maintain its accuracy & relevance.

Exploratory Data Analysis : Data Summary

Data Shape : 5000 Rows , 14 Columns (Including ID) = Original Values
Data Types : int64, float64 = Converted Values
Range Index : 0 to 4999

Data Description

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

Data Info

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  5000 non-null   int64
 1   Age                 5000 non-null   int64
 2   Experience          5000 non-null   int64
 3   Income              5000 non-null   int64
 4   ZIPCode             5000 non-null   int64
 5   Family              5000 non-null   int64
 6   CCAvg               5000 non-null   float64
 7   Education           5000 non-null   int64
 8   Mortgage            5000 non-null   int64
 9   Personal_Loan       5000 non-null   int64
 10  Securities_Account  5000 non-null   int64
 11  CD_Account          5000 non-null   int64
 12  Online              5000 non-null   int64
 13  CreditCard          5000 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 547.0 KB
```

Exploratory Data Analysis : General Observations

- ✓ Mean Age : 45.34 Years
- ✓ Mean Years Experience : 20.11 Years
- ✓ **Mean Income :** $73.77 Thousand
- ✓ **Mean Family Size :** 2.1 Members
- ✓ Mean Credit Card : $1.94 thousand Spent per month
- ✓ Education : Categories = Graduate , Undergraduate , Advanced / Professional
- ✓ Mean Mortgage : $56.500 value of home mortgage
- ✓ **Mean Personal Loan :** 9.6% accepted a personal loan in previous campaign
- ✓ Mean Securities Account : 10.44% of customers maintain a bank's securities account
- ✓ Mean CD Account : 6.04% or customers maintain a bank's certificate of deposit
- ✓ Mean Online Banking : 59.86% of customers have an online banking account
- ✓ Mean Credit Card : 29.4% of customers possess alternative bank credit card
- ✓ ZIP Code : 467 zip codes (Grouped by 1st two numbers = 7 zip categories)

*Securities, CD's &, Online Banking represent AllLife Bank products owned by customers

Exploratory Data Analysis Summary

Age:
The age distribution of customers ranges from around 23 - 67 years, with a mean age of approximately 45. The distribution appears to be fairly symmetric.

Experience:
The professional experience of customers ranges from 0 to 43 years, with a mean experience of approximately 20 years. There were negative values initially, but they have been corrected to positive values.

Income:
The annual income of customers ranges from 8 to $224 ,0000, with a mean income of approximately $74 ,000. The distribution is right-skewed, indicating there are relatively few customers with high incomes.

Family & Zip Code:
The family size of customers ranges from 1-4, with a mean size of approximately 2.4. The distribution shows a significant number of customers have a family size of 1 or 2. The 1st two digits of the ZIP Code have been extracted, resulting in 7 unique categories.

Credit Card Average Spend & Credit Card:
The average spending on credit cards / month by customers ranges from 0 to $10 ,000 , with a mean of approximately $1,940. The distribution is right-skewed, indicating that most customers have relatively low spending on credit cards. Approximately 29.4% of bank customers maintain credit cards with outside banks.

Education:
The education level of customers is categorized into three levels: Undergraduate, Graduate, and Professional.

Mortgage:
The value of house mortgages ranges from 0 to $635,000, with a mean of approximately $56.500. The distribution is right-skewed, indicating that most customers have relatively low mortgage values.

Personal Loan:
This variable represents whether the customer accepted a personal loan offered in the last campaign. Approximately 9.6% of customers accepted a personal loan from the last campaign

Securities Account & CD Account:
This variable represents whether the customer has a securities / cd account with the bank. 10.4% of customers maintain a securities account with the bank and approximately 6% of customers maintain a certificate of deposit with the bank.

Data Preprocessing

Duplicate Value Check

- No duplicates present

Missing Value Check

- No missing values

Missing Value Treatment

- No missing value treatment necessary

```
duplicate_rows = data.duplicated()
num_duplicates = duplicate_rows.sum()
print("Number of duplicate rows:", num_duplicates)
```

```
Number of duplicate rows: 0
```

```
missing_values = data.isnull().sum()
print("Missing values per column:")
print(missing_values)
```

```
Missing values per column:
ID                    0
Age                   0
Experience            0
Income                0
ZIPCode               0
Family                0
CCAvg                 0
Education             0
Mortgage              0
Personal_Loan         0
Securities_Account    0
CD_Account            0
Online                0
CreditCard            0
dtype: int64
```

Data Preprocessing

Outlier check

- Income : 1.92%
- CCAvg : 6.48%
- Mortgage : 5.82%
- Personal_Loan : 9.60
- Securities_Account : 10.44%
- CD_Account : 6.04%

```
Outlier Percentage by Feature:

Age                   0.00
Experience            0.00
Income                1.92
ZIPCode               0.00
Family                0.00
CCAvg                 6.48
Education             0.00
Mortgage              5.82
Personal_Loan         9.60
Securities_Account   10.44
CD_Account            6.04
Online                0.00
CreditCard            0.00
dtype: float64
```

- Personal_Loan = Independent Variable
- Training Data : 70% | Training Data Shape : (3500 , 11)
- Training Data - Personal_Loan Acceptance : 9.46%
- Testing Data : 30% | Testing Data Shape : (1500 , 11)
- Testing Data - Personal_Loan Acceptance : 9.93%

Data Preprocessing

Feature Engineering

- Data features are converted to int64 and float64
- A count for ZIPCode identified 467 unique variables. The ZIPCode was categorized into 7 features based on first two numbers of the ZIPCode.
- Data features were converted to [cat_cols].astype‘category’

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 5000 non-null   int64
 1   Experience          5000 non-null   int64
 2   Income              5000 non-null   int64
 3   ZIPCode             5000 non-null   int64
 4   Family              5000 non-null   int64
 5   CCAvg               5000 non-null   float64
 6   Education           5000 non-null   int64
 7   Mortgage            5000 non-null   int64
 8   Personal_Loan       5000 non-null   int64
 9   Securities_Account  5000 non-null   int64
 10  CD_Account          5000 non-null   int64
 11  Online              5000 non-null   int64
 12  CreditCard          5000 non-null   int64
dtypes: float64(1), int64(12)
memory usage: 507.9 KB
```

Data Preprocessing

Data preprocessing for modeling

- Experience values converted from negative to positive integers
- Education mapped into 3 separate values

1. Undergraduate
2. Graduate
3. Professional

```
# Correcting the experience values
data["Experience"].replace(-1, 1, inplace=True)
data["Experience"].replace(-2, 2, inplace=True)
data["Experience"].replace(-3, 3, inplace=True)
```

```
# Map the values to 1: Undergraduate; 2: Graduate 3: Advanced/Professional
data['Education'].replace(1 , 'Undergraduate' , inplace = True)
data['Education'].replace(2 , 'Graduate' , inplace = True)
data['Education'].replace(3 , 'Professional' , inplace = True)
```

Model Building Summary

- The model_performance_classification_sklearn function is utilized to check the model performance of models
- The confusion_matrix_sklearnfunction is utilized to plot the confusion matrix.

Function to compute different metrics & check classification performance

- model: classifier
- predictors: independent variables
- target: dependent variable

Function to compute the confusion_matrix with percentages

- model: classifier
- predictors: independent variables
- target: dependent variable

Model Building Criterion

Evaluation Criterion

A possibility exists that a learning model can make a wrong prediction resulting in:

- Loss of Resources : Predicting a bank customer will purchase a personal loan when in reality they are not interested.
- Loss of Opportunity : Predicting a bank customer will not purchase a personal loan when in reality they need and will buy a personal loan.

In a bank situation where interest is generated from the issuance of personal loans, wrongly predicting a customer won't buy a loan is a loss of opportunity.

The bank should reduce the loss by minimizing or eliminating `False Negatives' using the maximum `Recall' rate. A higher `Recall' rate will minimize `False Negatives' which should be the focus to minimize opportunity loss.

Model Building Process I

Objective : Determine if a bank customer will purchase a personal loan

- Split data into 1. training set, and 2. testing and validation set for evaluation
- Define functions for metrics, performance & accuracy of predictions
- Encode categorical features and scale numeric values
- Build a model using the training data and check performance
- The Decision Tree model is utilized in this data analysis
  1. Attributes of data are selected and all possible splits in data are made
  2. The Gini impurity is calculated after each split
  3. Steps are repeated for every attribute present in the data
  4. Decisions for the best split are based on the lowest Gini impurity
  5. Complete process repeats: the stopping criterion is reached or the leaves achieve homogeneity.
- Performance on training model is checked
- Model is visualized : Plot and / or textual reporting

Model Building Process II

- Importance of features in model are printed, plotted, and evaluated
- Model performance is checked on test data using confusion_matrix
- New values are assigned for additional testing
- Accuracy of 98.6% indicating the model correctly predicts the outcome of Personal Loan Acceptance for a large majority of instances in the test data.
- Recall of 0.919 indicates the sensitivity the model correctly identified approximately 91.9% of customers who actually accepted the personal loan.
- Precision of 0.901indicates approximately 90.1% of loan accepetance were True Positive with a low False Positive rate.
- F1 Score of 0.910 reflects a good balance between Precision & Recall
- Conclusion of Testing : This is an accurate model

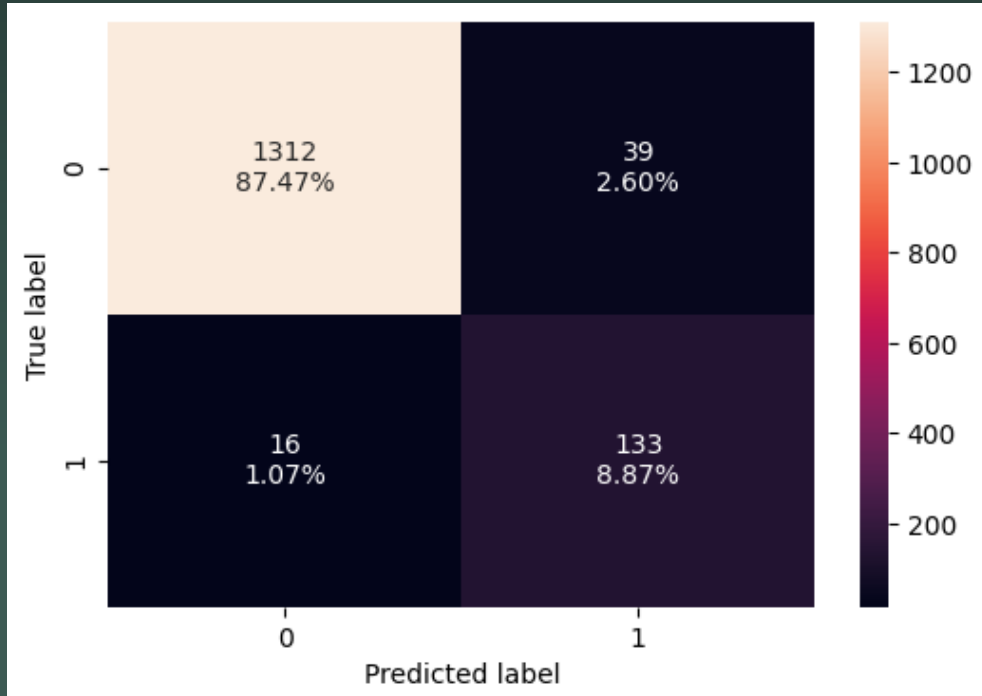
Model Building Process III

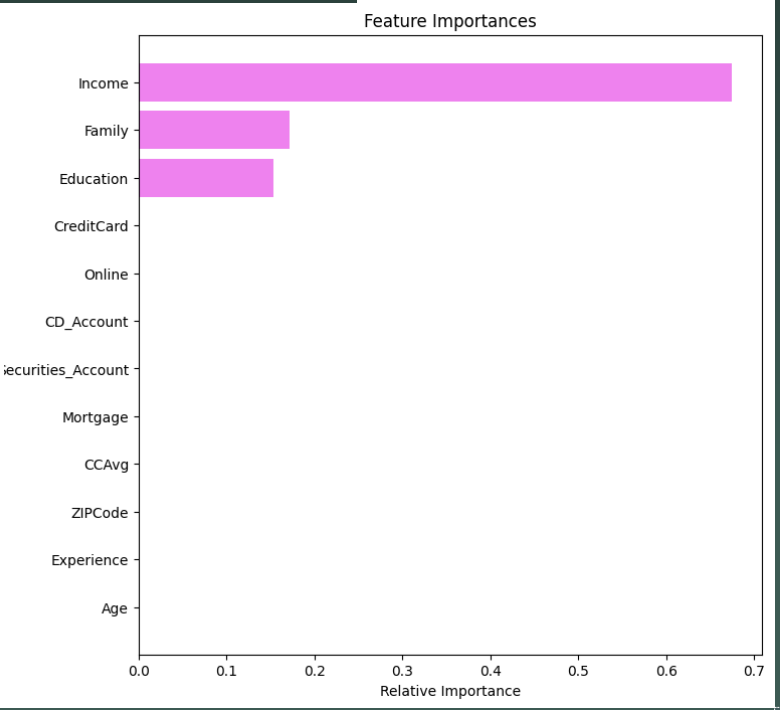
- Perform Pre-Pruning for model improvement
- Plot confusion matrix to check performance of training data
- Evaluate tuned performance of model
- Visualize the model : plot and/or textual
- Identify the importance of features in model via print and visualization
- Check Performance of test data
- Perform Cost-Complexity Pruning & plot impurities
- Train Decision Tree using effective Alpha values
- Check, plot and evaluate Recall vs Alpha for training and testing sets
- Apply Alpha to balance complexity and performance of model
- Check performance on training data via confusion matrix

Model Building Process IV

- Visualize Model via plot and/or text
- Calculate and plot Gini value & feature importance
- Check performance of test data
- Compare performance of testing & training models

	Imp
Income	0.674921
Family	0.171953
Education	0.153126
Age	0.000000
Experience	0.000000
ZIPCode	0.000000
CCAvg	0.000000
Mortgage	0.000000
Securities_Account	0.000000
CD_Account	0.000000
Online	0.000000
CreditCard	0.000000





Model Building Performance Summary

- Training Performance Comparison:
  - The decision tree model with Pre-Pruning provided perfect accuracy, recall, and precision on the training data (values of 1.0). This indicates that the model is likely overfit on training data.
  - The Decision Tree (Post-Pruning) model performs well on the training data but not as well as the Pre-Pruning model. It achieves accuracy of 0.964, which is slightly lower than the Pre-Pruning model's accuracy of 1.0. This indicates the Post-Pruning model generalizes better to unseen data compared to the Pre-Pruning model.

- Testing Performance Comparison:
  - The Decision Tree (Post-Pruning) model shows strong performance on the test data with an accuracy of 0.964. This indicates that the model is able to make accurate predictions on new, unseen data.
  - The Recall and Precision of the Post-Pruning model perform reasonably well, indicating the model is effective in correctly identifying positive instances (Recall) and minimizing false positives (Precision).
  - The F1 score of 0.830 suggests a good balance of precision & recall for the Post-Pruning model.

- Overall, the Decision Tree (Post-Pruning) model appears to be the best choice among the three models presented. It shows strong performance on both the training and test data, indicating good generalization ability and avoidance of overfitting. However, more context about the specific problem and business requirements is necessary to make a final decision.

- It's worth noting that the Decision Tree sklearn model also performed well, with an accuracy of 0.981 on the training data. However, the model's perfect performance on the training data (accuracy, recall, and precision of 1.0) raises concerns about potential overfitting.

Model Performance Summary

- **Decision Tree Model Overview:** This classification model was used to predict whether a customer will buy a personal loan with AllLife Bank. The model is trained on a dataset containing various features such as "Income," "Family," "Education," "Age," "Experience," "ZIPCode," "CCAvg," "Mortgage," "Securities_Account," "CD_Account," "Online," and "CreditCard." The model employs the Gini impurity criterion for splitting and is designed to be interpretable.
- **Pruning and Complexity Parameter (ccp_alpha):** Pruning reduces overfitting in the model by removing nodes that do not significantly contribute to the model's performance on unseen data. The ccp_alpha parameter (complexity parameter) controls the amount of pruning applied to the tree. A higher ccp_alpha value results in more aggressive pruning, leading to simpler trees with fewer nodes.
- **Class Weight (class_weight):** Class Weight is used to address class imbalance in the dataset. In this case, the class_weight is set to {0: 0.15, 1: 0.85}, indicating that the "Not taking a loan" class (0) has a lower weight (15%) compared to the "Taking a loan" class (1), which has a higher weight (85%). This weighting scheme is useful when one class is significantly underrepresented, as it ensures that the model pays more attention to the minority class during training.
- **Data Insights and Interpretation:** Based on the model's decision rules, the most significant feature for predicting whether a customer will take a personal loan is "Income." If a customer's income is less than or equal to 98.50, the model predicts that the customer will not take a loan (class 0). On the other hand, if a customer's income is greater than 98.50, the model further considers the "Education" and "Family" features to make the final prediction.
- **Feature Importance:** Feature importance ranking indicates that "Income" is the most crucial feature in predicting whether a customer will take a personal loan. This aligns with common intuition as income is often a significant factor influencing borrowing decisions. "Family" and "Education" are also important features in the decision tree model. The remaining features ("Age," "Experience," "ZIPCode," "CCAvg," "Mortgage," "Securities_Account," "CD_Account," "Online," and "CreditCard") do not contribute significantly to the model's predictions, as they have zero importance.

Model Performance Summary

- Conclusion: The final decision tree model, with the chosen pruning technique and class weighting, provides a good balance between performance and interpretability. It is crucial to note that the provided results and conclusions are based on the specific dataset used for training and testing the model. The model's performance and conclusions may vary depending on the data's quality, size, and the specific business problem it aims to address. Therefore, further validation on new data and consideration of the model's context in real-world scenarios are essential before drawing concrete conclusions or making critical business decisions based on the model's predictions

- Summary of most important features used by the decision tree model for prediction

- Feature Importance: The decision tree model identified the most important features used to predict whether a customer will assume a personal loan. The top three most important features, in descending order of importance, are:

    1. Income: Income has the highest importance, indicating that it is the most influential factor in determining whether a customer will take a personal loan. Customers with higher incomes are more likely to be considered as potential loan-takers.

    2. Family: The "Family" feature is the second most important, suggesting that family size may also play a significant role in loan decisions. Customers with larger families might be more inclined to consider taking a loan for various financial needs.

    3. Education: Education is the third most important feature, indicating that the level of education of the customer could be a relevant factor in deciding whether they will take a personal loan.

- Conclusion: Based on the results and insights from the decision tree model, we can conclude that income, family size, and education are crucial factors in predicting whether a customer will take a personal loan. Other demographic and financial features in the dataset appear to have limited impact on the loan decision. This model provides valuable insights into customer behavior and can be useful for making informed decisions related to loan marketing strategies, targeted advertising, and customer segmentation. However, it is essential to validate the model's performance on new data and consider real-world implications before implementing recommendations based on its predictions.

Model Performance Summary

Summary of key performance metrics for training and test data

Metric	Decision Tree sklearn	Decision Tree(Pre-Pruning)	Decision Tree(Post-Pruning)
Training Accuracy	0.980667	1.0	0.964286
Training Recall	0.892617	1.0	0.924471
Training Precision	0.910959	1.0	0.753695
Training F1-Score	0.901695	1.0	0.830393
Test Accuracy	0.963333	N/A	N/A
Test Recall	0.892617	N/A	N/A
Test Precision	0.773256	N/A	N/A
Test F1-Score	0.82866	N/A	N/A

Model	Accuracy	Recall	Precision	F1
Decision Tree sklearn	0.981	0.893	0.911	0.902
Decision Tree (Pre-Pruning)	1.000	1.000	1.000	1.000
Decision Tree (Post-Pruning)	0.964	0.924	0.754	0.830
Decision Tree (Post-Pruning)	0.963	0.893	0.773	0.829

Model Performance Improvement

Commentary on improvement of model performance via pruning techniques

1. Pre-Pruning: The decision tree model with Pre-Pruning achieved perfect accuracy, recall, and precision on the training data, which is an indicator of potential overfitting. While the model may fit the training data extremely well, it is likely to have limited generalization ability on unseen data, leading to poor performance on the test set.

2. Post-Pruning: The decision tree model with Post-Pruning improved its generalization ability on unseen data, as evidenced by its performance on the test data. The Post-Pruning model achieved high accuracy, recall, precision, and F1-score on the test data, indicating that it performs well on previously unseen instances. By pruning back some of the decision nodes, the model avoids overfitting and results in a more balanced and reliable model.

3. Importance of Pruning: The improvement in model performance after applying Post-Pruning demonstrates the significance of pruning techniques in decision tree models. Pruning helps prevent overfitting by reducing the complexity of the tree, ensuring that it captures meaningful patterns in the data without memorizing noise or outliers. As a result, the pruned model is more likely to generalize well to new data, making it more valuable in real-world applications.

- Overall, the final decision tree model with Post-Pruning offers a more reliable and balanced prediction mechanism for the Personal_Loan target variable. It showcases the importance of carefully tuning hyperparameters, such as ccp_alpha, and using appropriate pruning techniques to enhance the model's performance and ensure its ability to make accurate predictions on unseen data.

Model Performance Improvement

Commentary on the decision rules and checking the feature importance

1. Decision Rules: The model provides decision rules that can be easily understood. These decision rules outline the conditions based on specific features for the target variable, Personal_Loan. For example:

- If the Income is less than or equal to 98.50, the model predicts class 0, indicating that customers with lower income are less likely to accept a personal loan offer.
- If the Income is greater than 98.50, the model further considers the Education feature. If Education is less than or equal to 1.50, the model predicts class 1, suggesting that customers with higher income and lower education are more likely to accept a personal loan offer.
- If the Income is greater than 98.50 and Education is greater than 1.50, the model then considers the Family feature. If Family is less than or equal to 2.50, the model predicts class 0, implying that customers with higher income, higher education, and smaller family sizes are less likely to accept a personal loan offer.
- If the Income is greater than 98.50, Education is greater than 1.50, and Family is greater than 2.50, the model predicts class 1, indicating that customers with higher income, higher education, and larger family sizes are more likely to accept a personal loan offer.

1. Feature Importance: The feature importance indicates the significance of each feature in the decision-making process of the model. The descending order of feature importance is as follows:

- Income: Income plays the most critical role in determining the likelihood of accepting a personal loan. Higher income individuals are more likely to accept the offer.
- Family: Family size is the second most important feature. Smaller family sizes increase the likelihood of accepting a personal loan.
- Education: Education level is the third most important feature. Customers with lower education levels are more likely to accept the loan offer.
- Age, Experience, ZIPCode, CCAvg, Mortgage, Securities_Account, CD_Account, Online, CreditCard: These features have an importance of 0.0, indicating that they have little or no impact on the model's predictions. It suggests that these features do not significantly contribute to the decision-making process for the Personal_Loan target.

- Overall, the decision tree model provides valuable insights into the behavior of customers regarding personal loan acceptance. It identifies the key features that drive the decision and highlights the importance of income, family size, and education level in determining whether a customer will accept a personal loan offer. The model's simplicity and interpretability make it a useful tool for businesses to understand customer preferences and tailor their marketing strategies accordingly. Additionally, the post-pruning technique employed in the final model improves its generalization ability, ensuring better performance on unseen data and making it a reliable tool for real-world applications.

Data Background & Contents

Data Background : The dataset is from AllLife Bank, a US-based financial institution with a growing customer base. The majority of customers are liability customers, meaning they are depositors with varying deposit sizes. However, there is a small group of asset customers who are also borrowers. The bank aims to increase its loan business by converting liability customers into personal loan customers while retaining them as depositors.

Context : AllLife Bank previously ran a campaign for liability customers to promote personal loans and achieved a healthy conversion rate of over 9% success. This success has motivated the retail marketing department to devise better-targeted campaigns to increase the conversion rate.

Objective : The objective of this project is to build a predictive model that can identify potential liability customers who have a higher probability of purchasing personal loans. The model aims to understand which customer attributes are most significant in driving loan purchases and to identify specific segments of customers that should be targeted more effectively.

The dataset includes various demographic, financial, and behavioral attributes of the bank's customers, along with a binary target variable (Personal_Loan) indicating whether a customer accepted a personal loan during the last campaign.

The data is used to build a supervised learning classification model that to help the marketing department effectively target potential customers for personal loan offers based on their characteristics and behaviors.

Actionable Insights and Business Recommendations I

- Targeted Marketing Campaign: Based on the decision tree model, the bank can identify specific segments of customers that are more likely to accept personal loans. Customers with higher incomes and family size, as well as those with advanced/professional education, are more likely to purchase personal loans. The marketing department can tailor their campaigns to focus on these specific segments to increase the success ratio.

- Importance of Income: The decision tree model highlights that income is the most significant predictor for personal loan acceptance. The bank can focus on customers with higher incomes and offer personalized loan products that match their financial needs and goals.

- Online Banking Engagement: Customers who use internet banking facilities (Online = 1) might be more receptive to digital marketing efforts. The bank can leverage this information to promote personal loan offers through online channels, such as email campaigns or targeted online advertisements.

- Pruning Techniques: The comparison of models before and after pruning shows the importance of using appropriate pruning techniques to avoid overfitting. The Post-Pruning model performs better on the test data, indicating improved generalization and reliability in real-world scenarios.

- Investment in Education Campaigns: Customers with higher education levels (Graduate and Advanced/Professional) have a higher likelihood of accepting personal loans. The bank can design marketing campaigns that highlight the benefits of personal loans for career growth, education, or other significant life events that align with customers' educational aspirations.

Actionable Insights and Business Recommendations II

- **Customer Segmentation:** The decision tree segments customers based on their attributes and behavior, providing valuable insights into customer profiles. The bank can use this information to create targeted marketing campaigns for different customer segments, ensuring that each campaign resonates with the specific needs and preferences of each group.

- **Credit Card Usage:** Customers who use credit cards issued by other banks might be more open to exploring loan options beyond their primary bank. The bank can consider cross-selling opportunities by offering competitive loan products to these customers and encouraging them to consolidate their financial services with AllLife Bank.

- **Continuous Monitoring:** To ensure the model's performance remains robust, continuous monitoring and validation should be carried out on new data. As customer preferences and behaviors evolve, the model should be periodically updated to reflect these changes.

- **Customer Experience and Relationship:** While promoting personal loans, the bank should prioritize maintaining a positive customer experience and building strong customer relationships. Transparent and personalized communication regarding loan offers, terms, and conditions will foster trust and loyalty among customers.

- **Compliance and Ethical Considerations:** When conducting targeted marketing campaigns, the bank should adhere to regulatory guidelines and ensure ethical practices are followed in data usage and customer engagement. Protecting customer privacy and data security is of utmost importance.

- Overall, the insights from the decision tree model can guide AllLife Bank in formulating effective marketing strategies to boost personal loan sales, enhance customer engagement, and achieve sustainable business growth.

APPENDIX I

Univariate Analysis

APPENDIX II

Bivariate Analysis

Correlation Matrix

Loan Interest vs Education

Personal Loan vs Family

Personal Loan vs Securities Account

Personal Loan vs CD Account

Loan Interest by Online Bank Account

Loan Interest by Zip Code Category

Loan Interest by Age

Personal Loan Interest by Experience

Personal Loan Interest vs Income

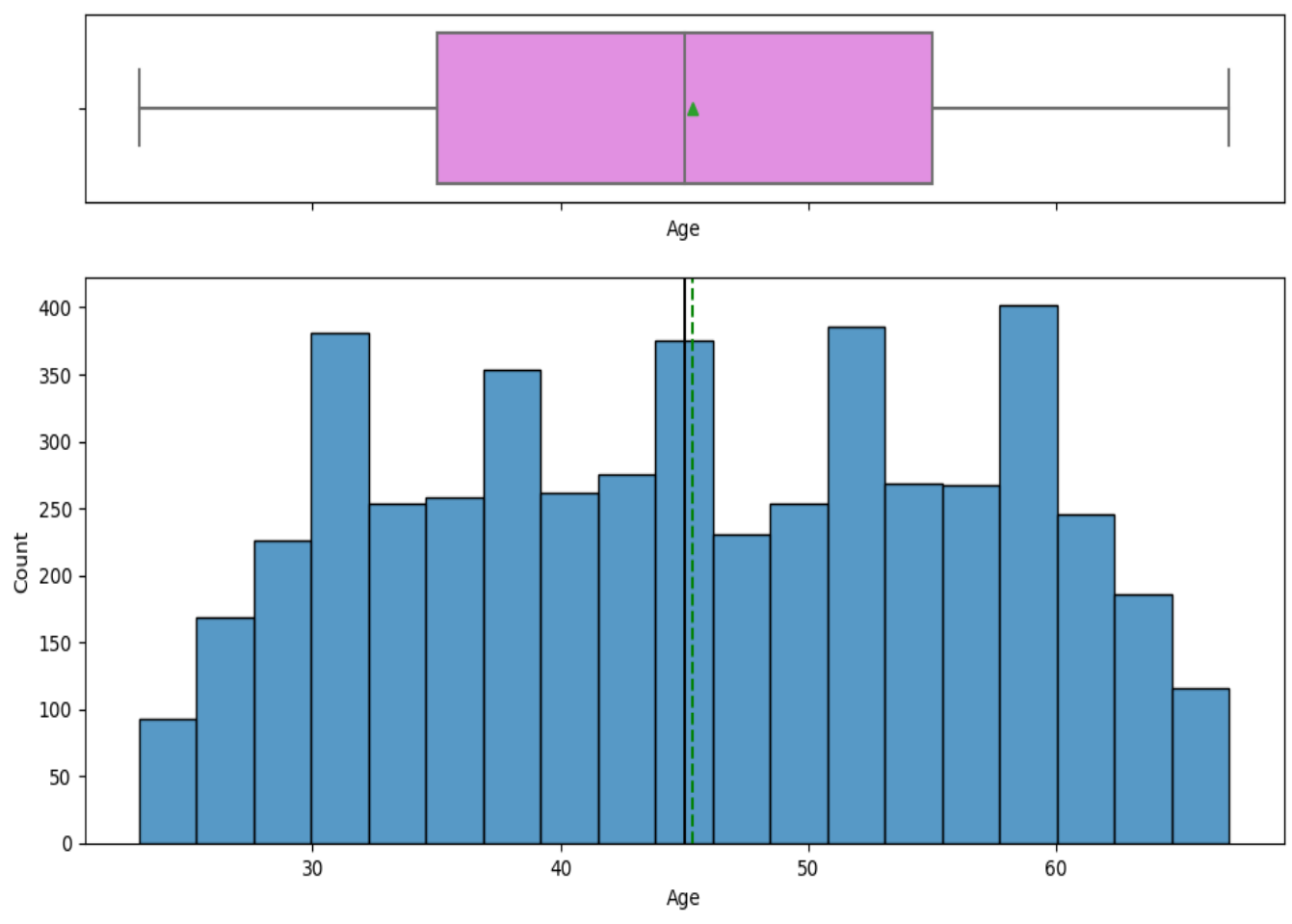
Personal Loan vs Monthly Credit Card Spend

APPENDIX III

Model Visualization

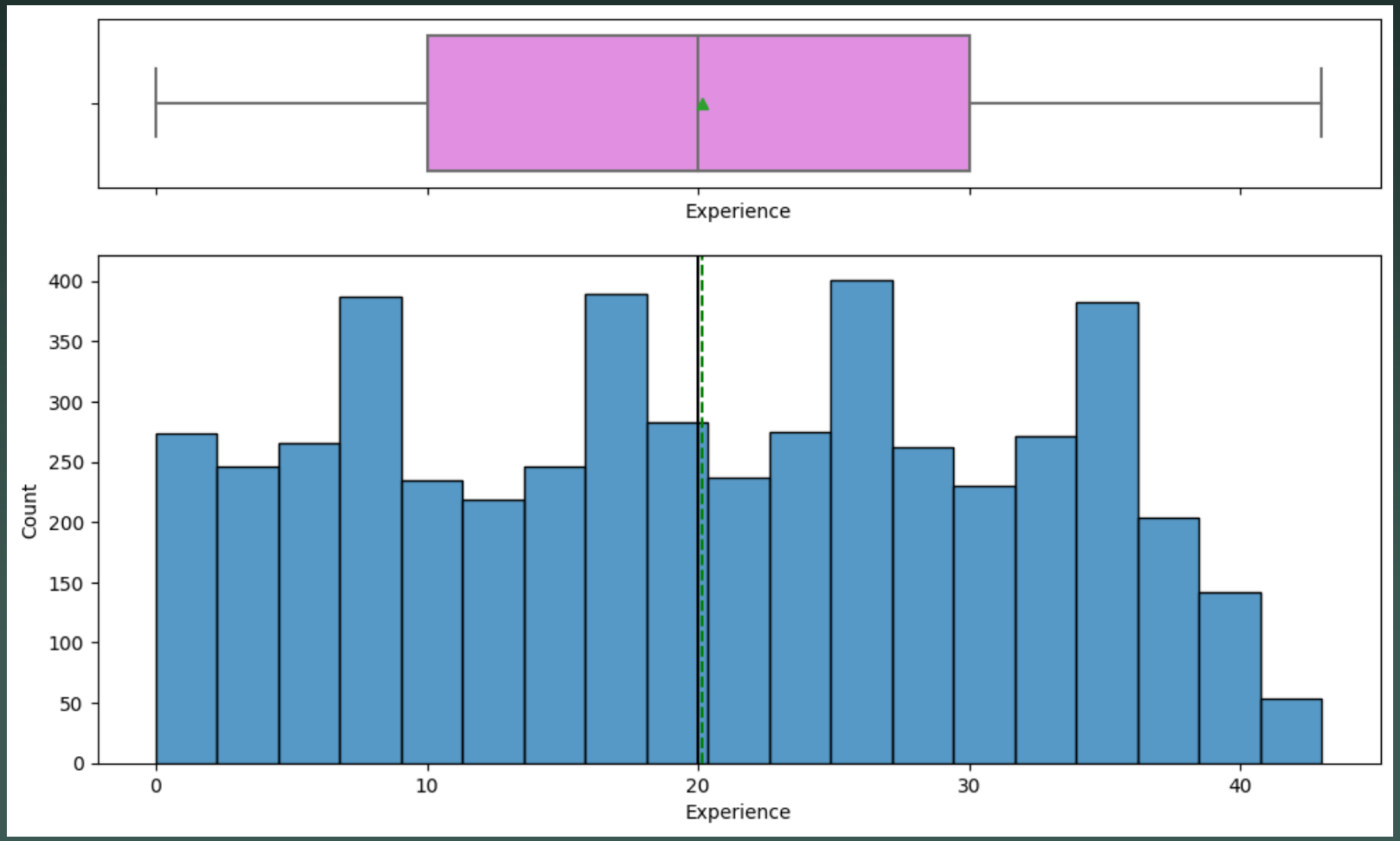
Appendix I

Mean Age in Years = 45.34



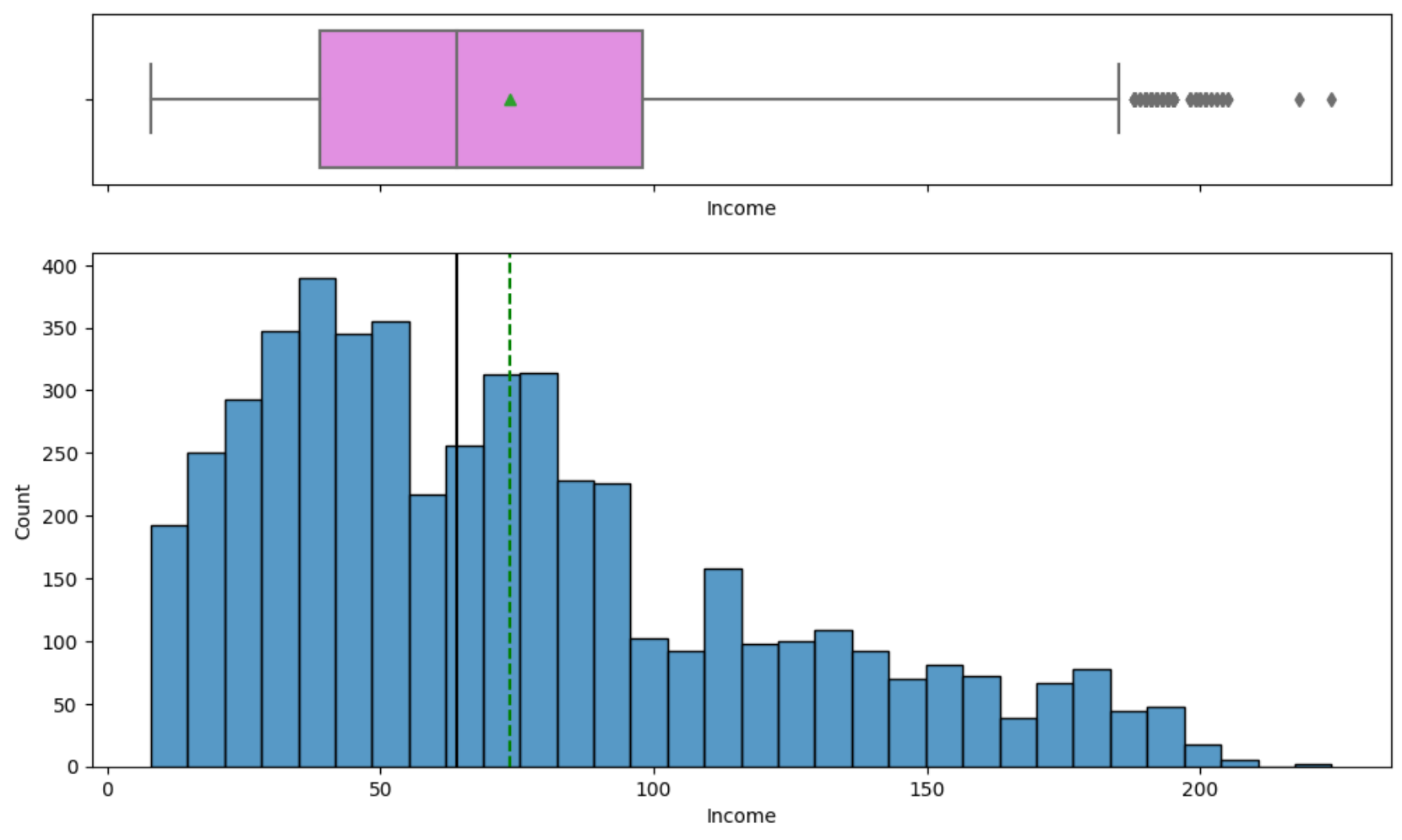
Appendix I

Mean Job Experience in Years = 20.10



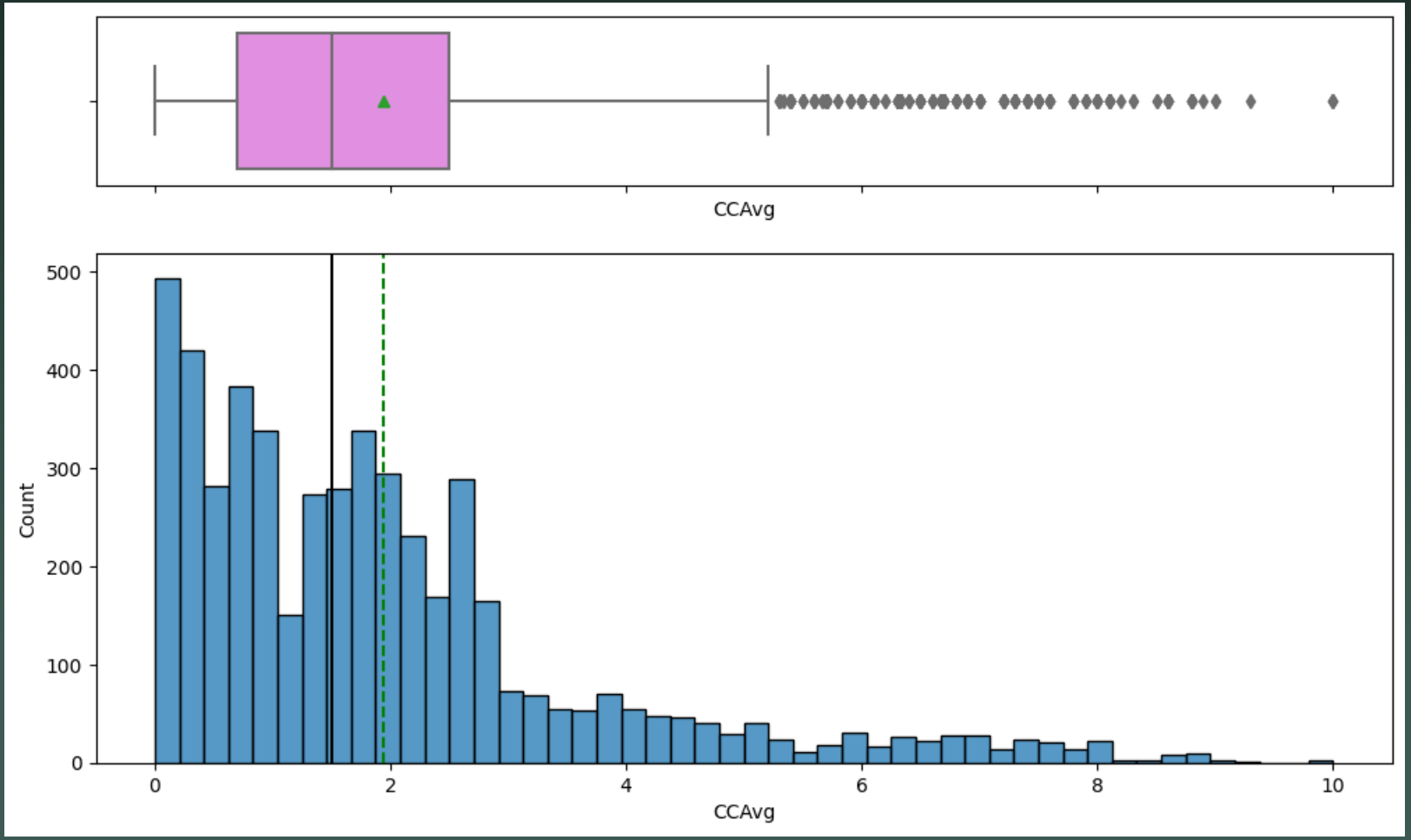
Appendix I

Mean Income * 1000 US Dollars = 73.74



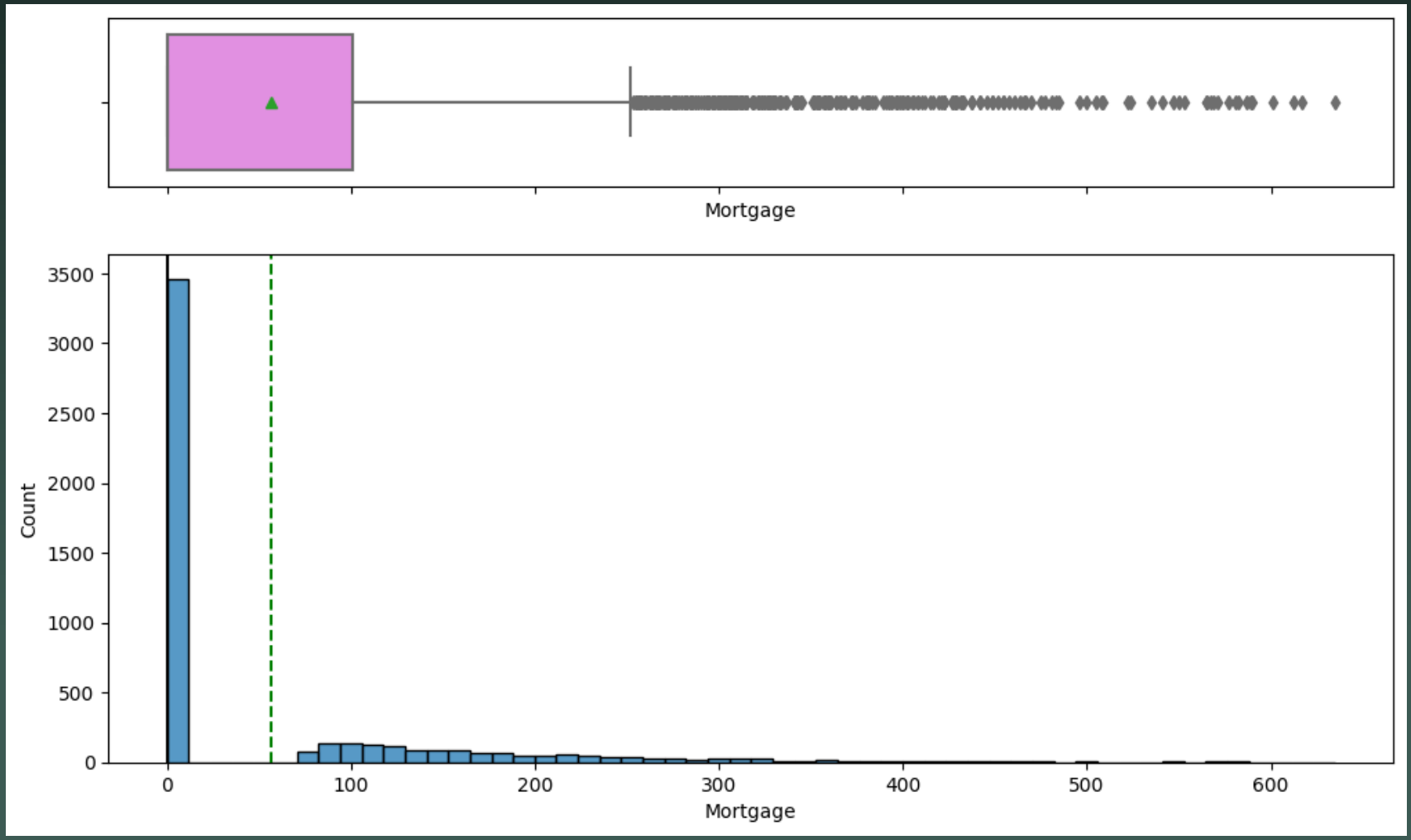
Appendix I

Credit Card Average Monthly Spend * 1000 US Dollars = 1.94



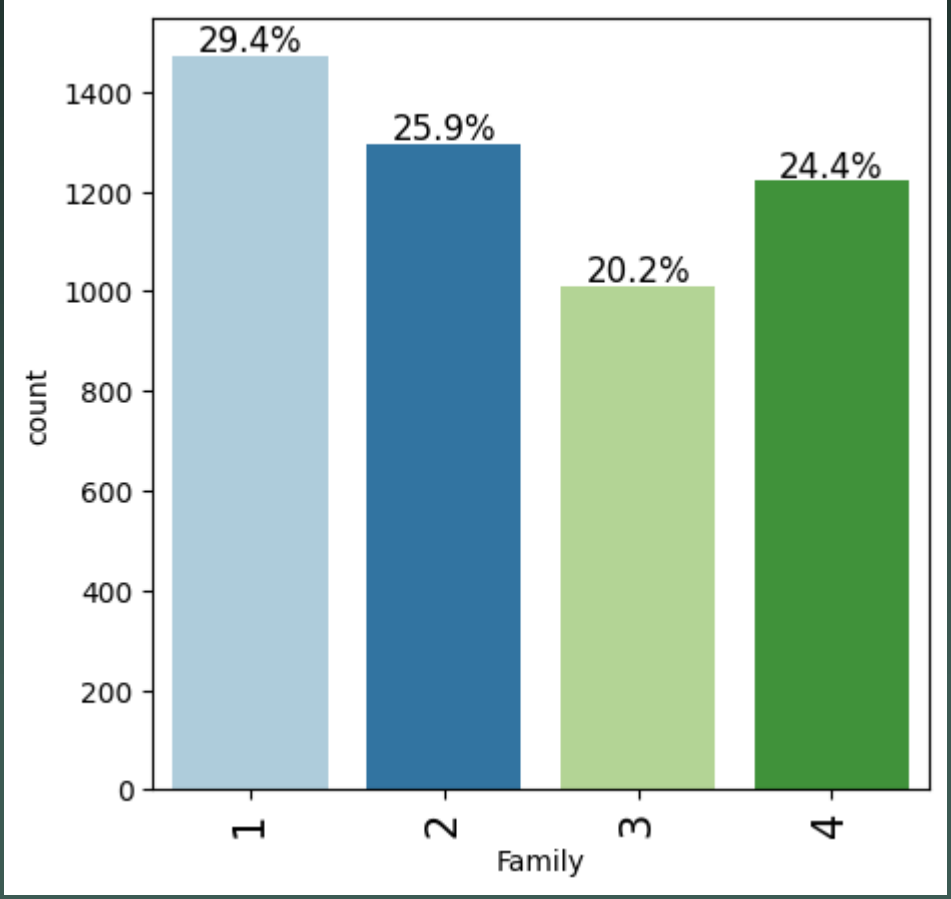
Appendix I

Average Mortgage Held * 1000 US Dollars = 56.50



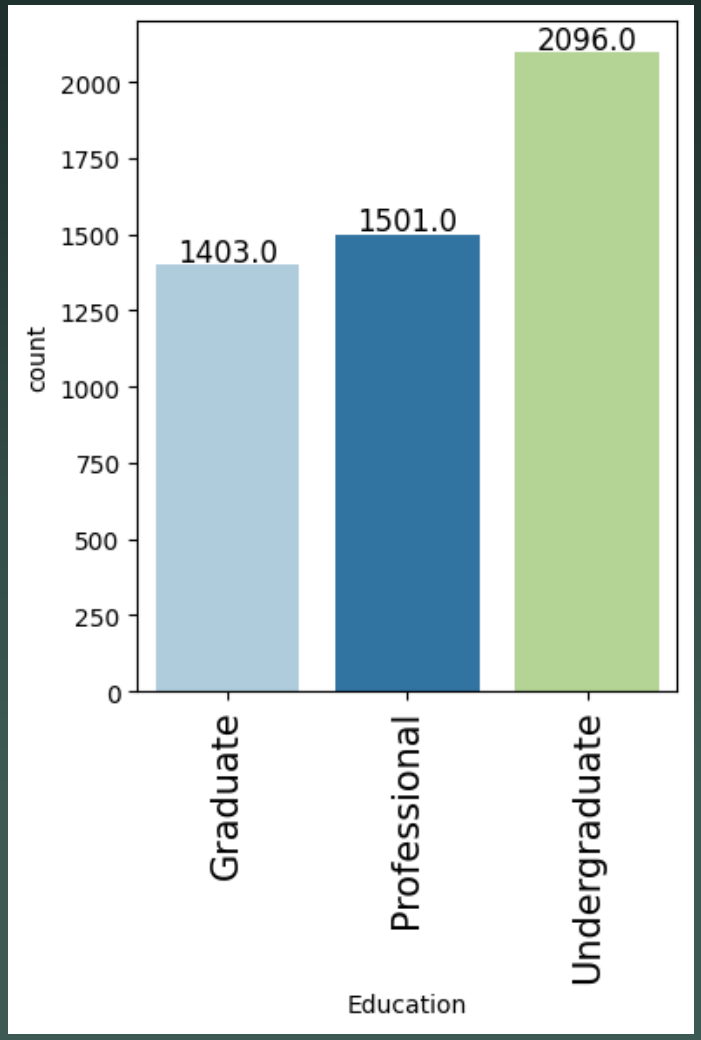
Appendix I

Average Number of Family Members= 2.40



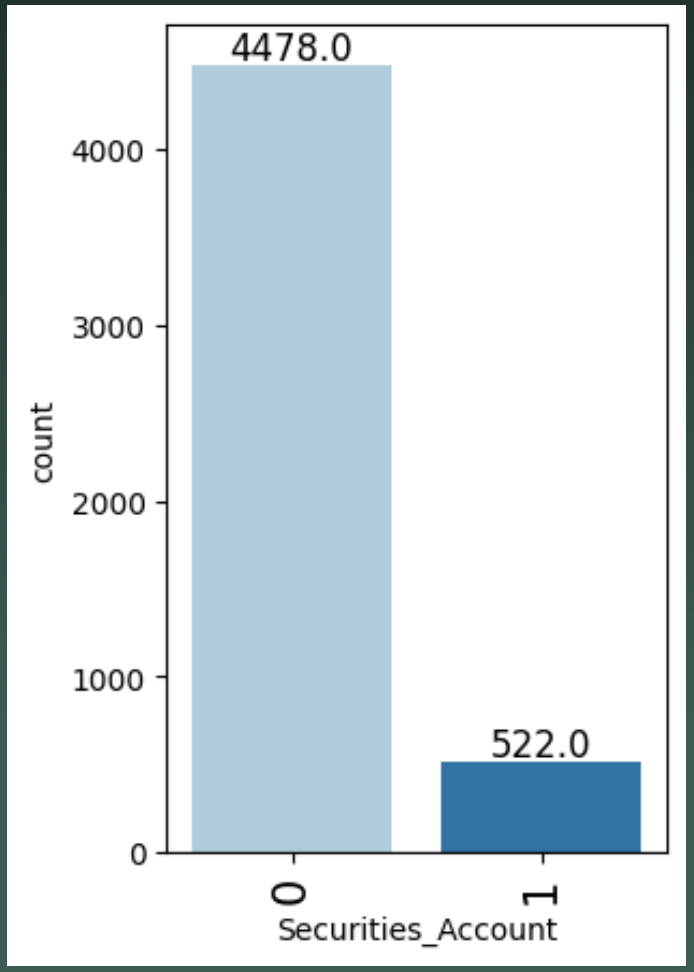
Appendix I

Education Level = Graduate | Undergraduate | Professional



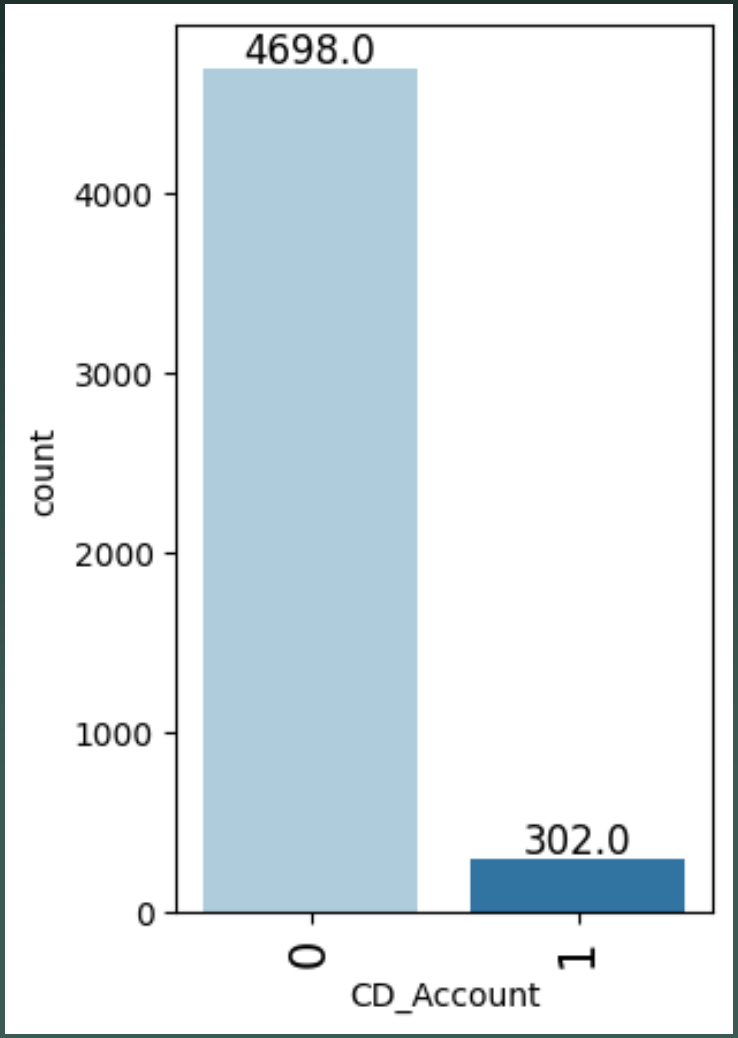
Appendix I

Bank Members Maintaining an AllLife Securities Account



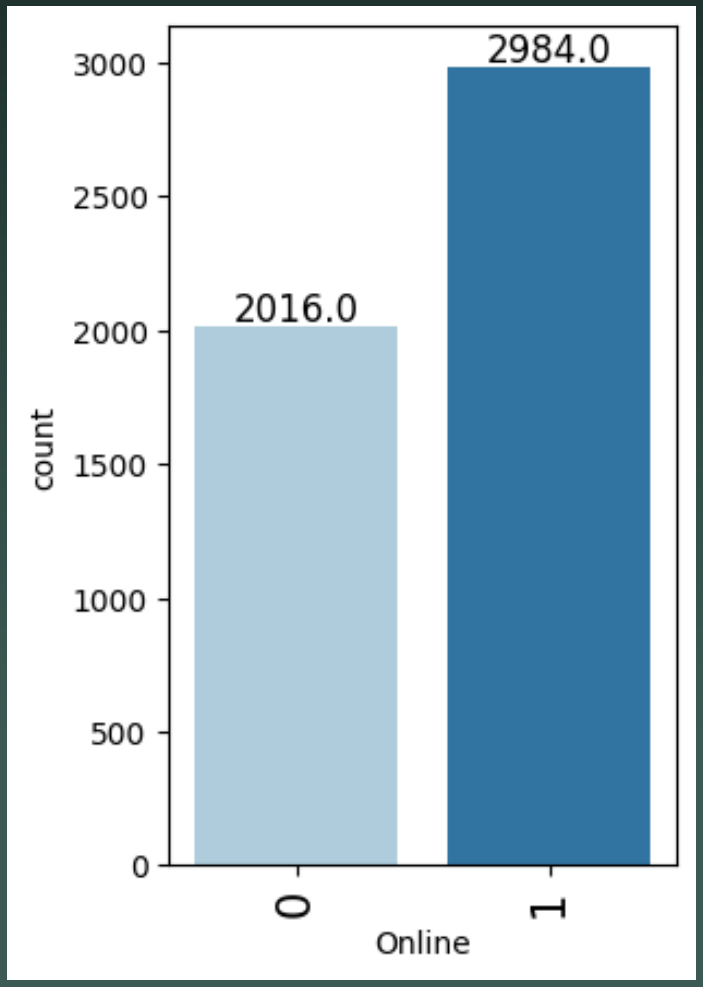
Appendix I

Bank Members Maintaining an AllLife CD Account



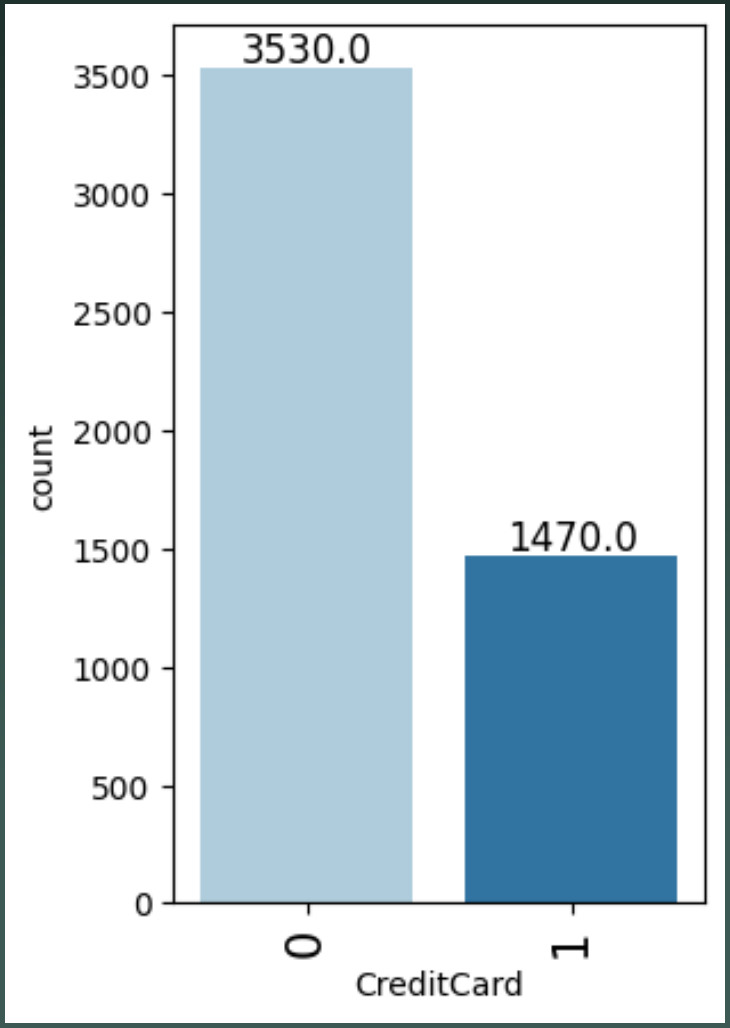
Appendix I

Members Utilizing an Online Bank Account



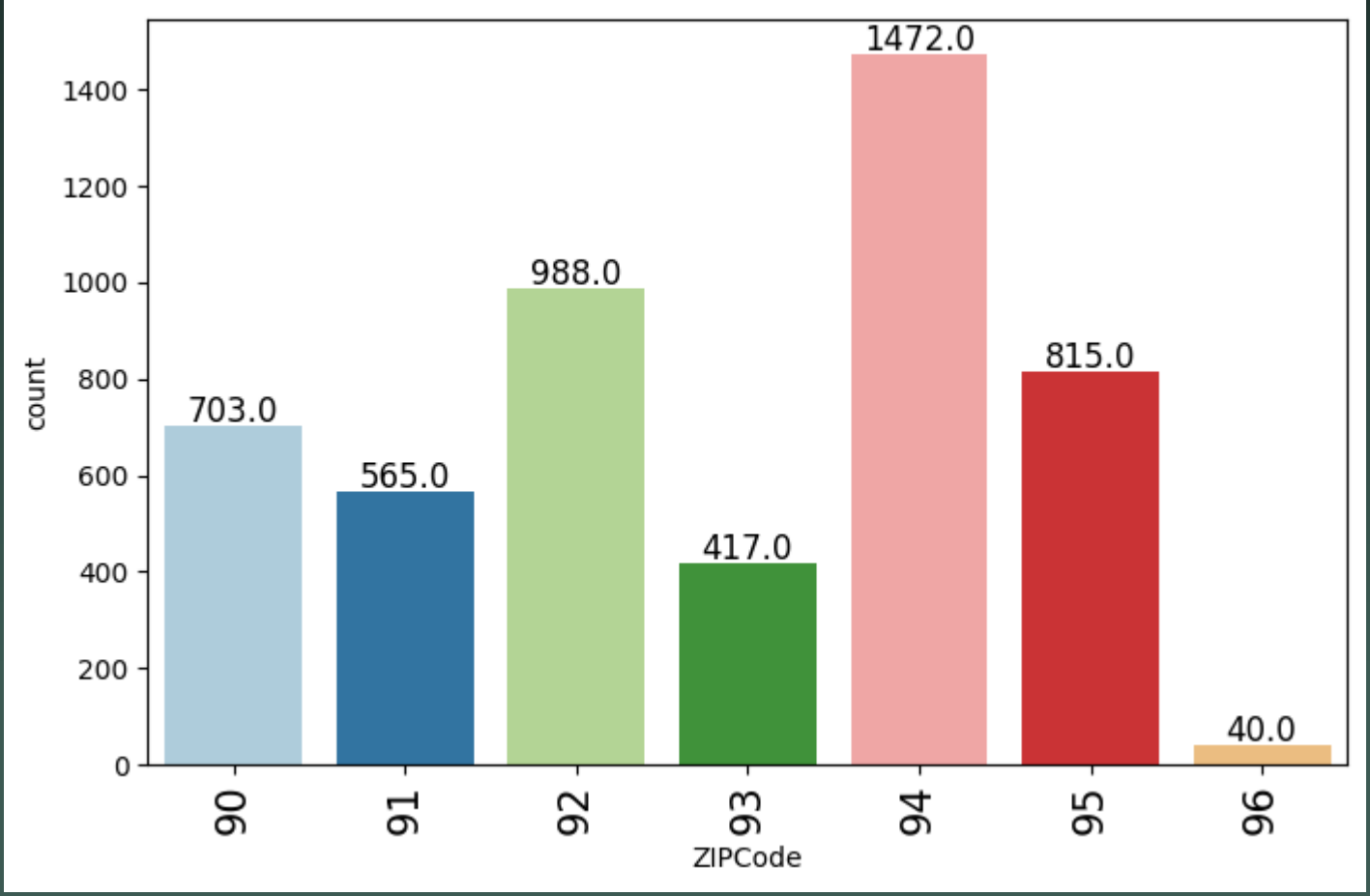
Appendix I

Members Possessing Non-Bank Credit Cards



Appendix I

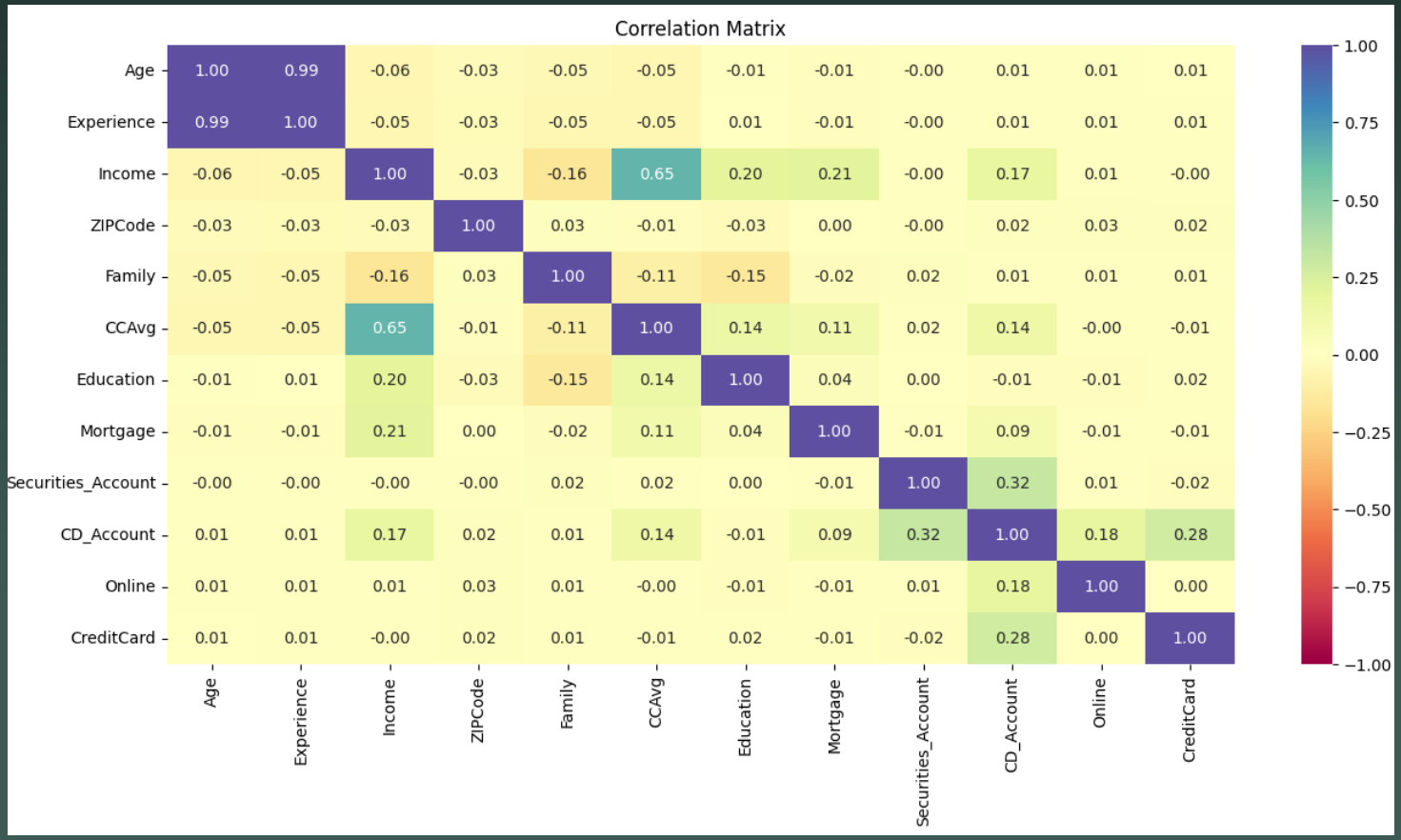
Member Home ZIP Codes = 7 Area Categories



Appendix II

Correlation Matrix

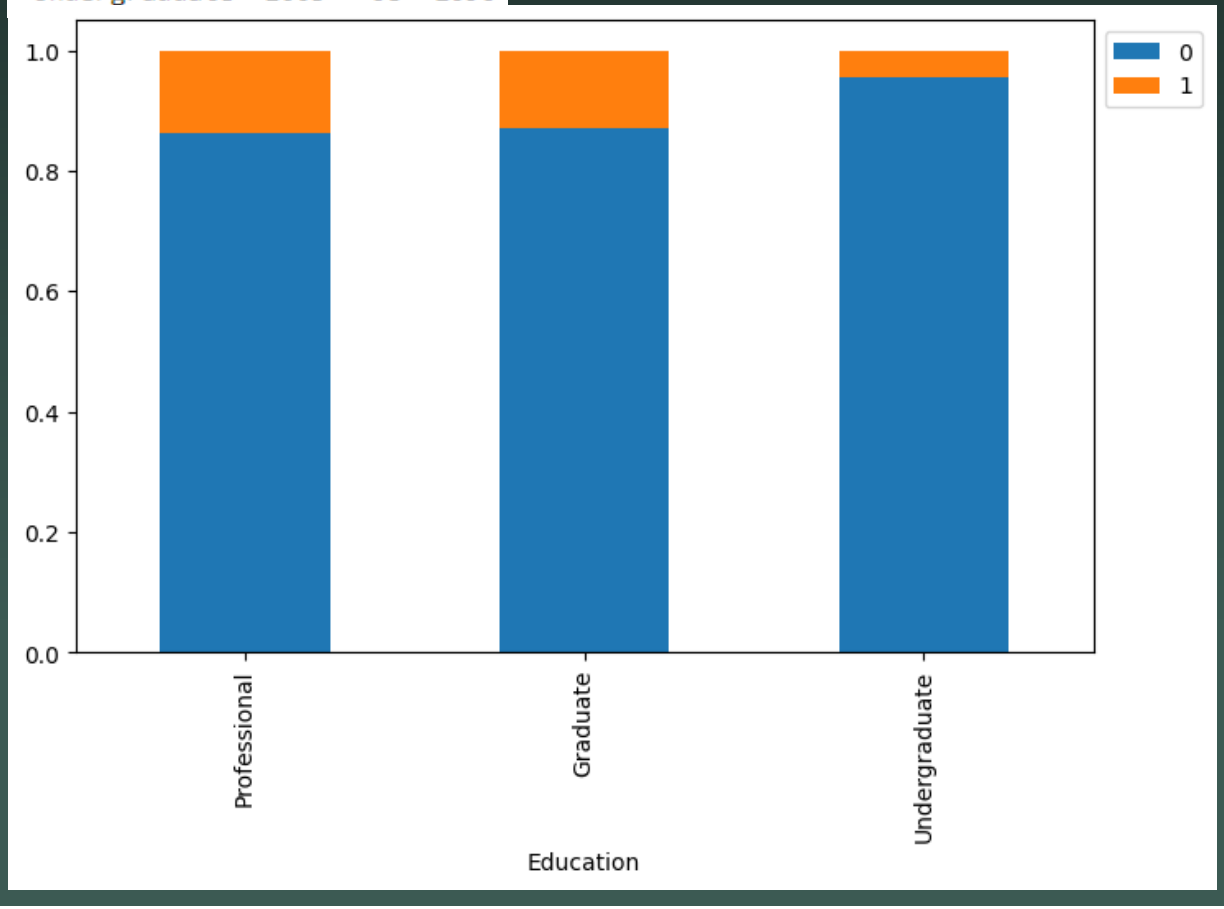
Positive Correlation: Experience & Age , Credit Card Monthly Spend & Income
Slightly Positive Correlation: CD & Securities Account , Credit Card & CD account
All other feature correlates are neutral or negligible



Appendix II

Loan Interest by Education

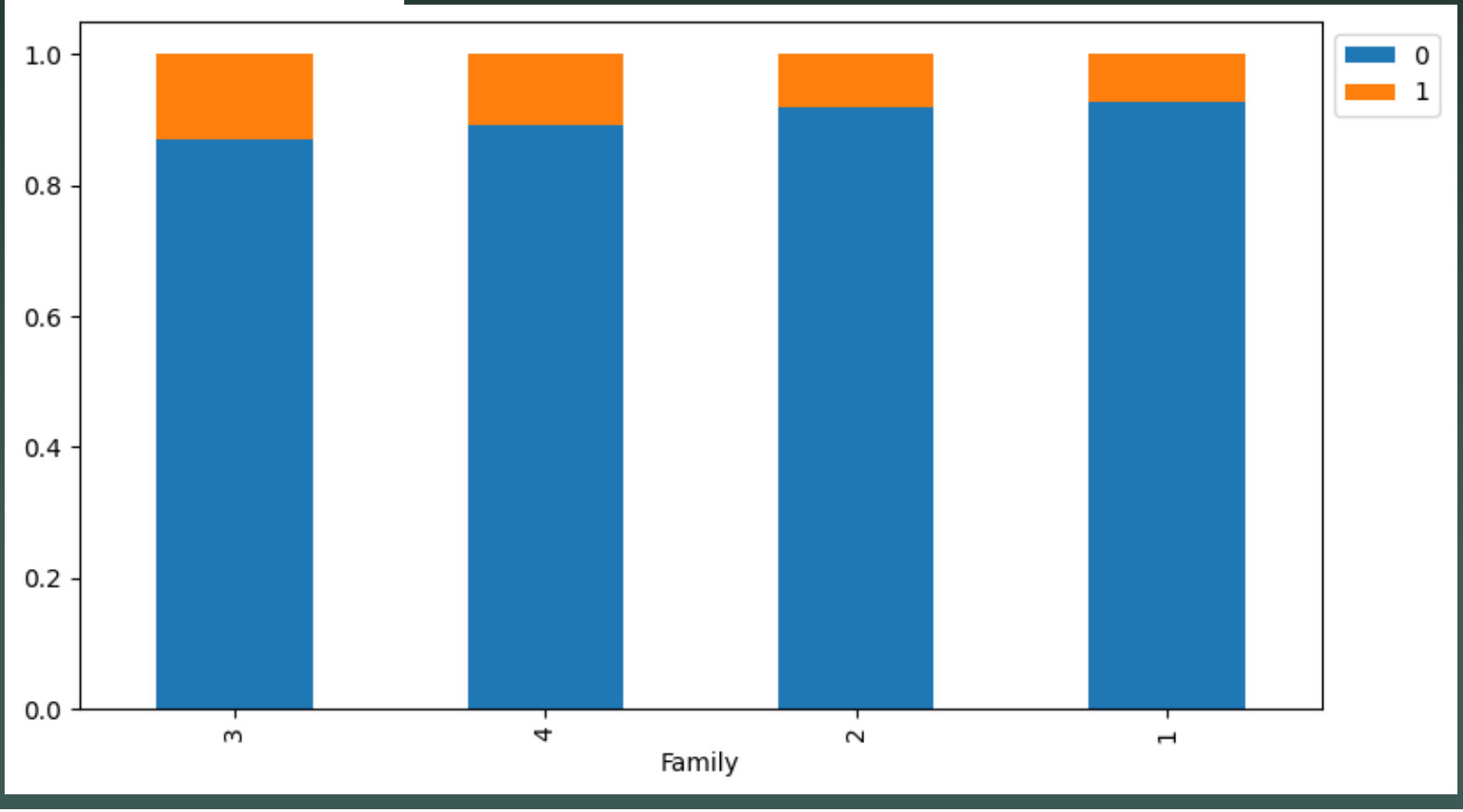
Personal_Loan	0	1	All
Education			
All	4520	480	5000
Professional	1296	205	1501
Graduate	1221	182	1403
Undergraduate	2003	93	2096



Appendix II

Personal Loan Interest by Family

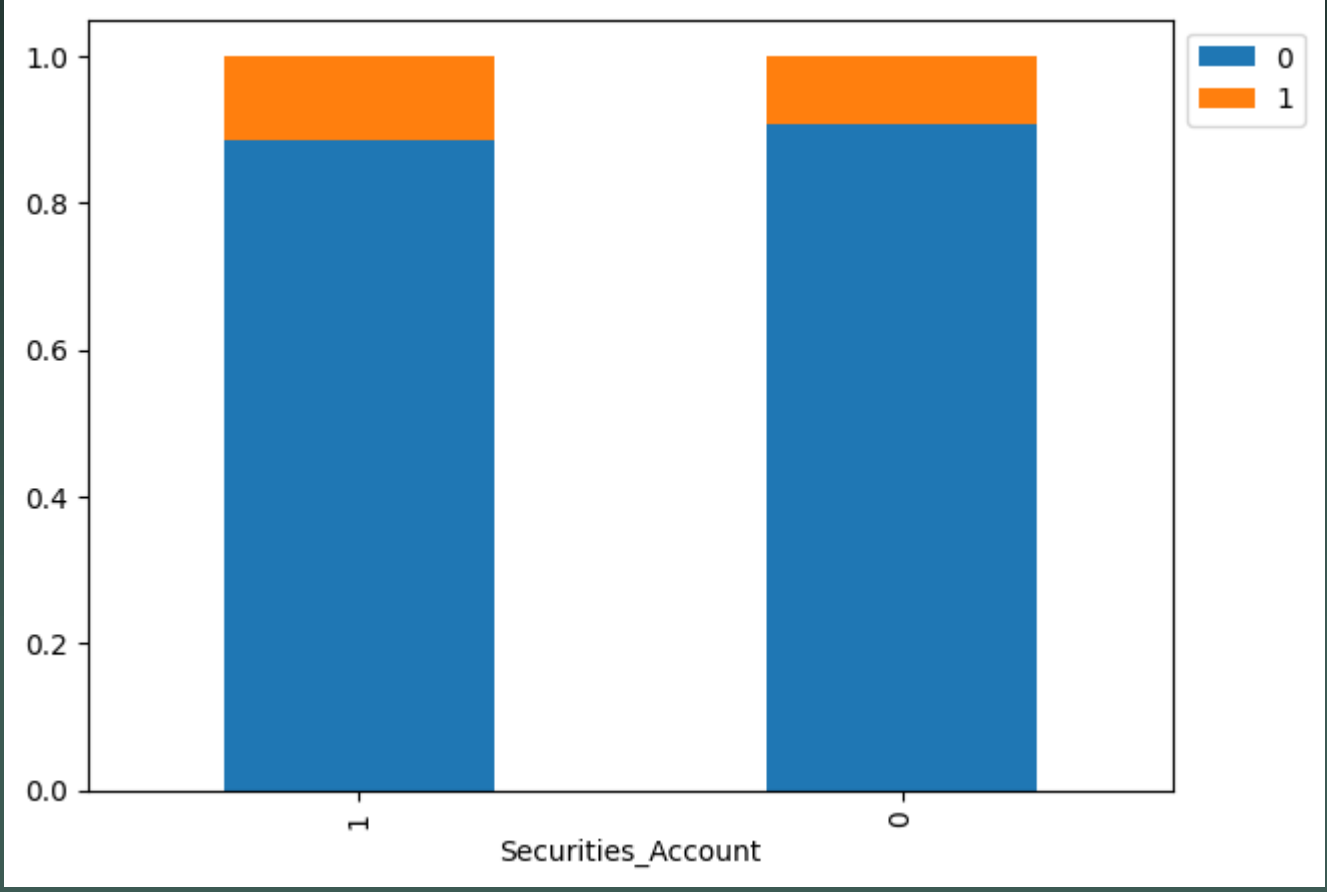
Personal_Loan	0	1	All
Family			
All	4520	480	5000
4	1088	134	1222
3	877	133	1010
1	1365	107	1472
2	1190	106	1296



Appendix II

Personal Loan vs Security Account

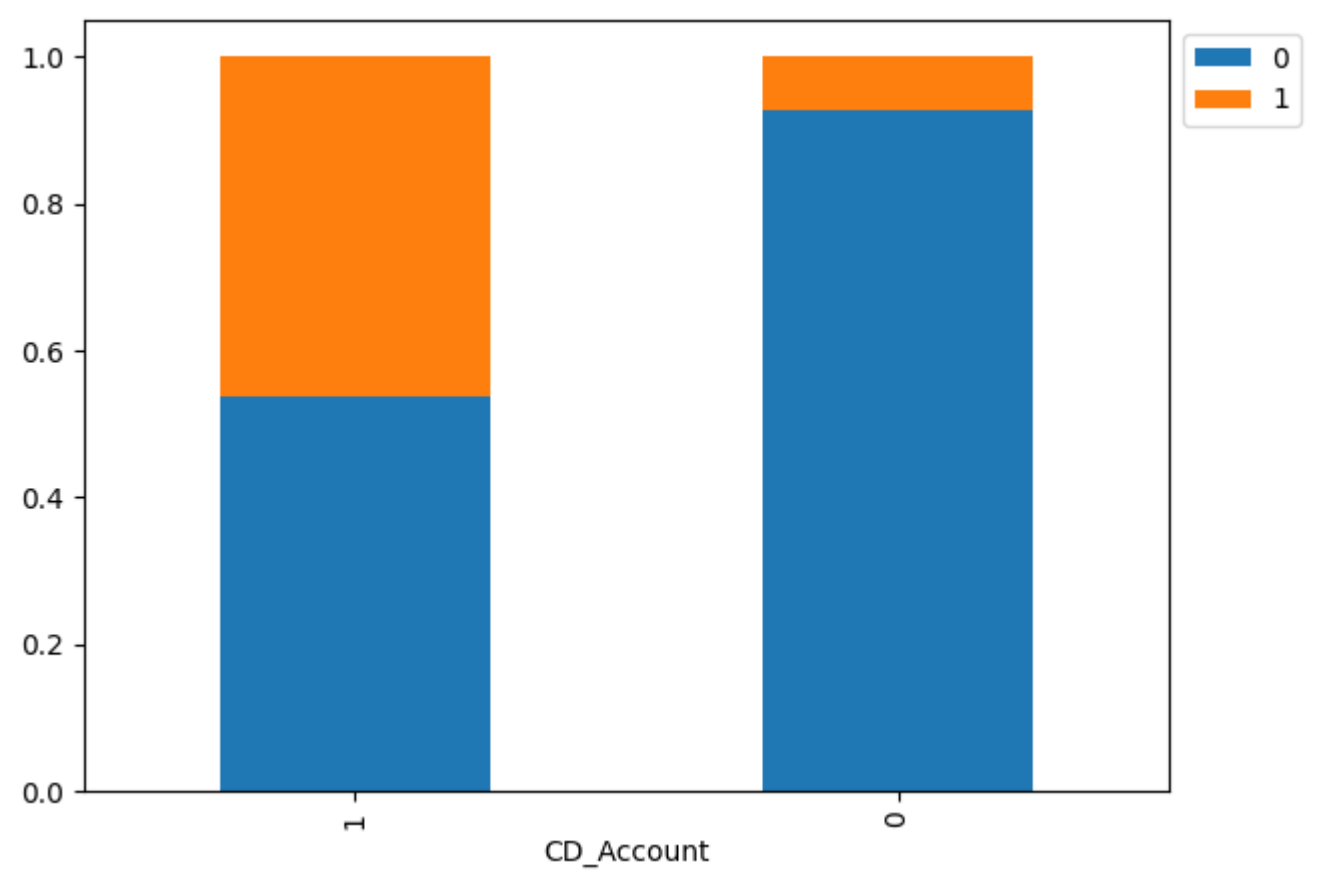
Personal_Loan	0	1	All
Securities_Account			
All	4520	480	5000
0	4058	420	4478
1	462	60	522



Appendix II

Personal Loan vs CD Account

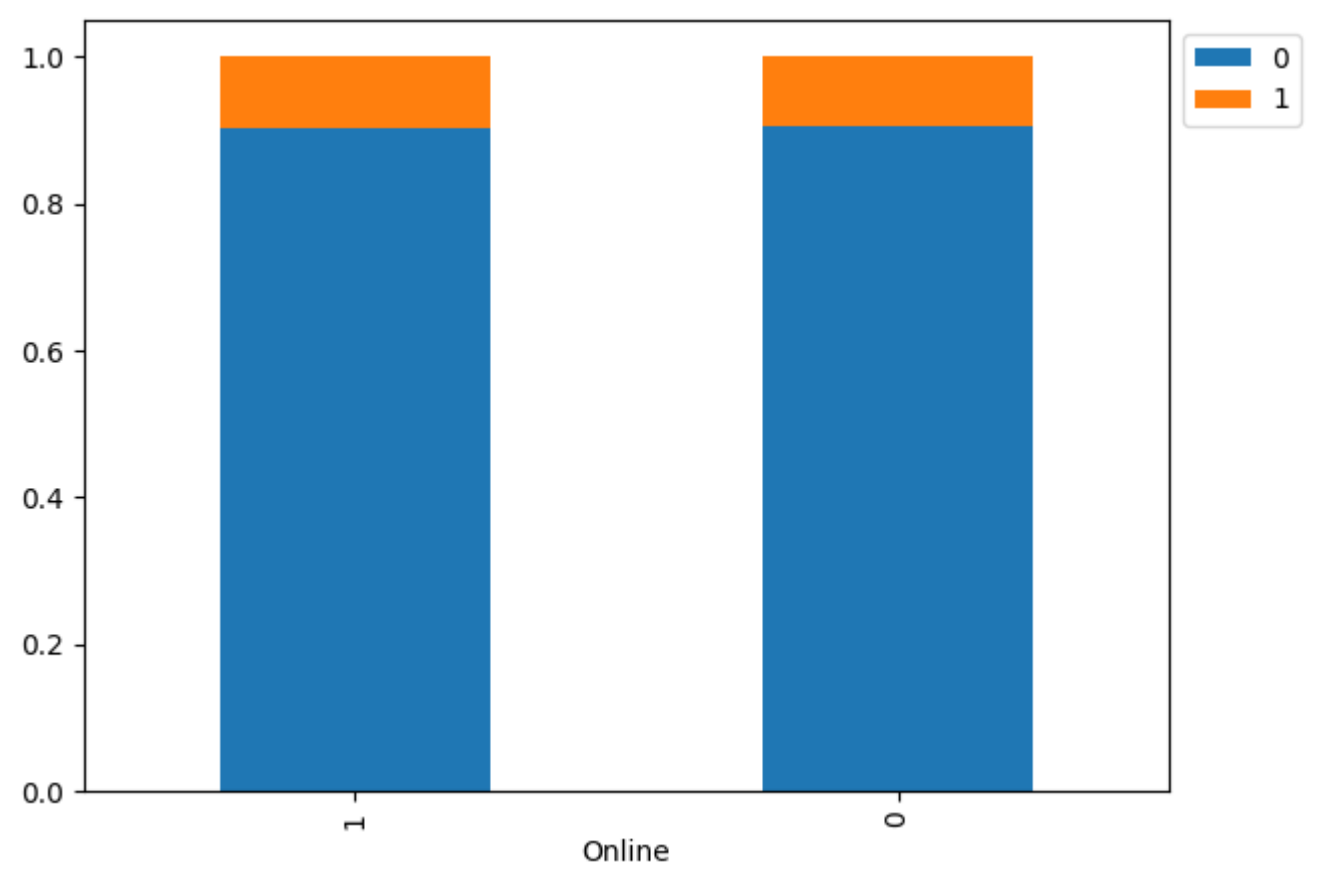
Personal_Loan	0	1	All
CD_Account			
All	4520	480	5000
0	4358	340	4698
1	162	140	302



Appendix II

Personal Loan vs Online Banking

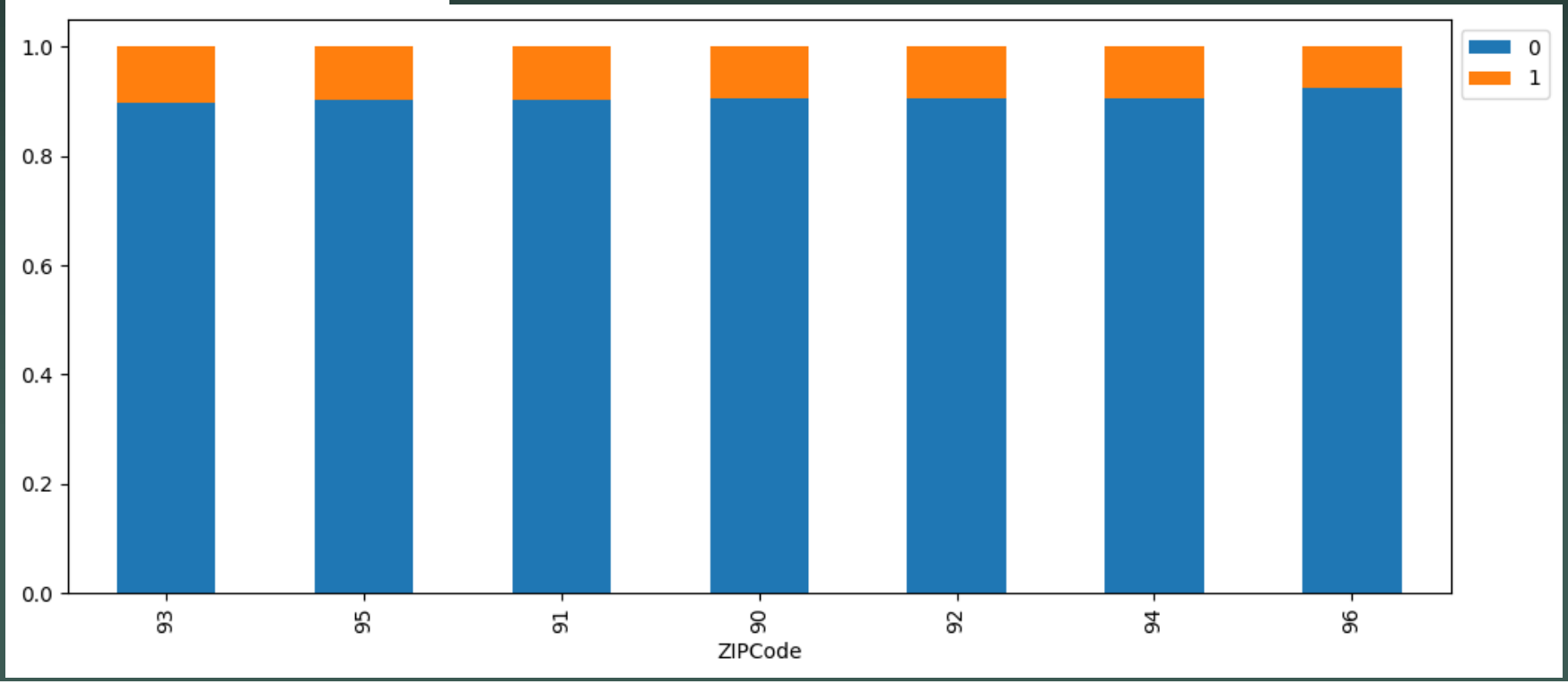
Personal_Loan	0	1	All
Online			
All	4520	480	5000
1	2693	291	2984
0	1827	189	2016



Appendix II

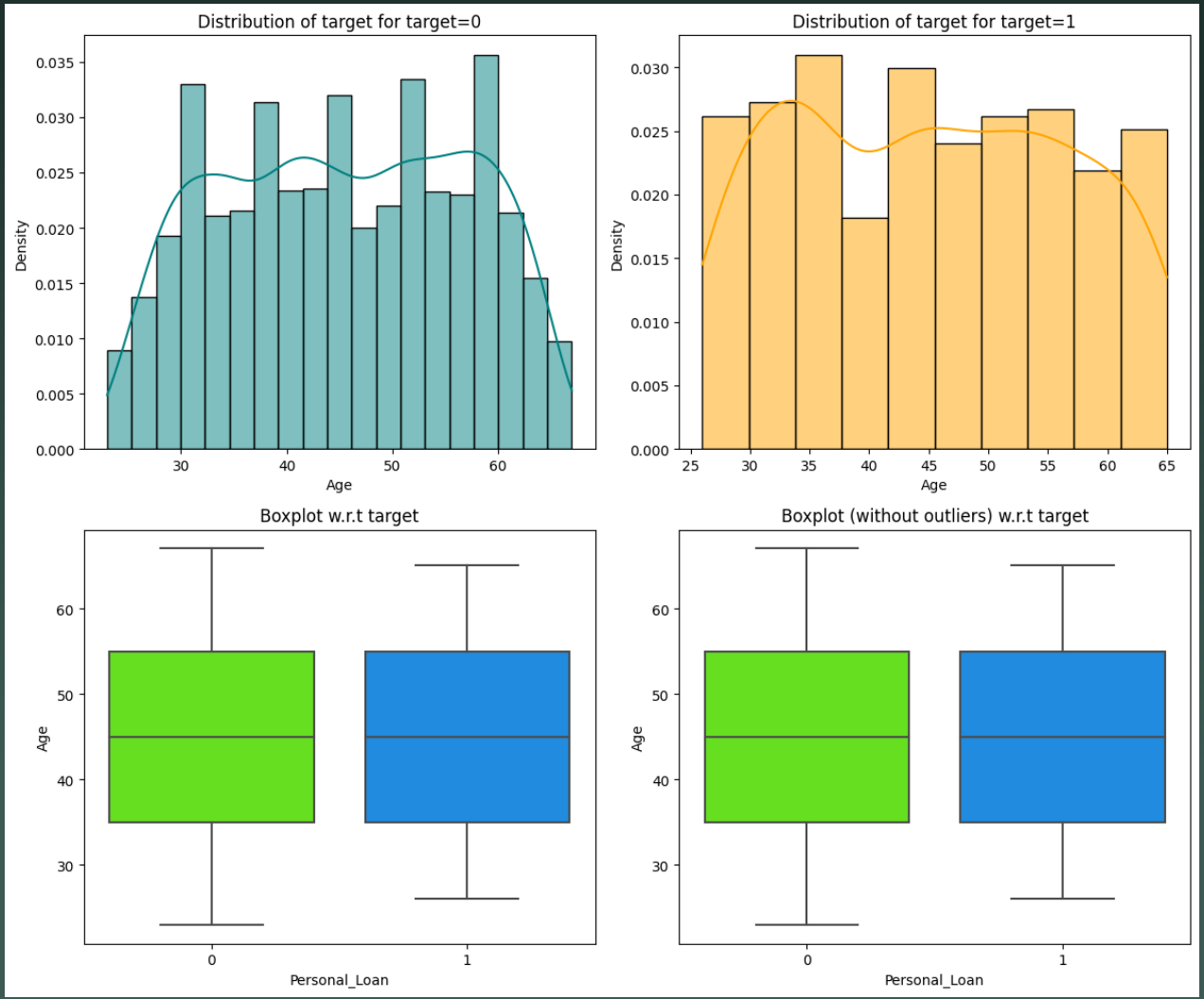
Personal Loan vs Zip Code Category

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



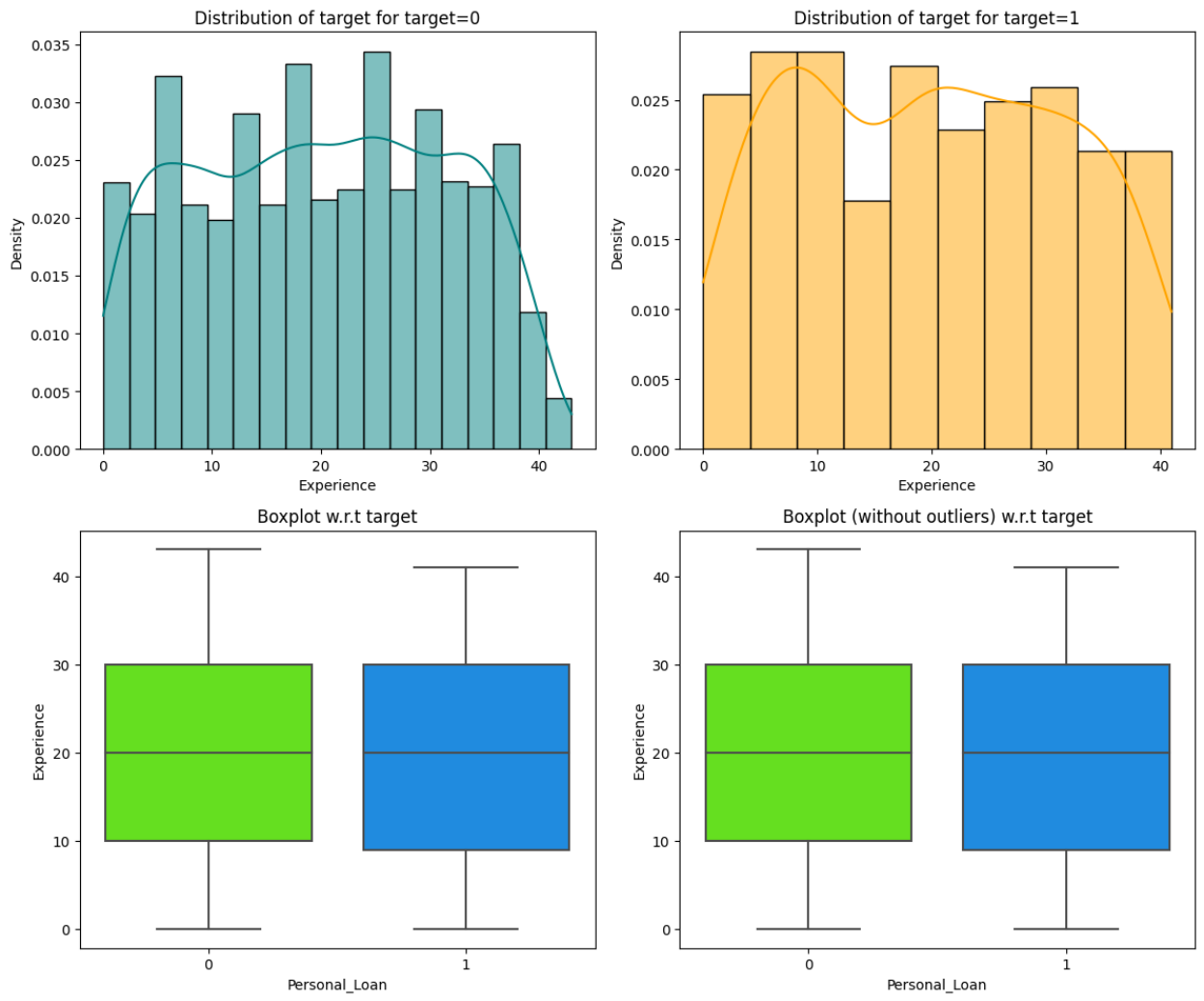
Appendix II

Personal Loan vs Age



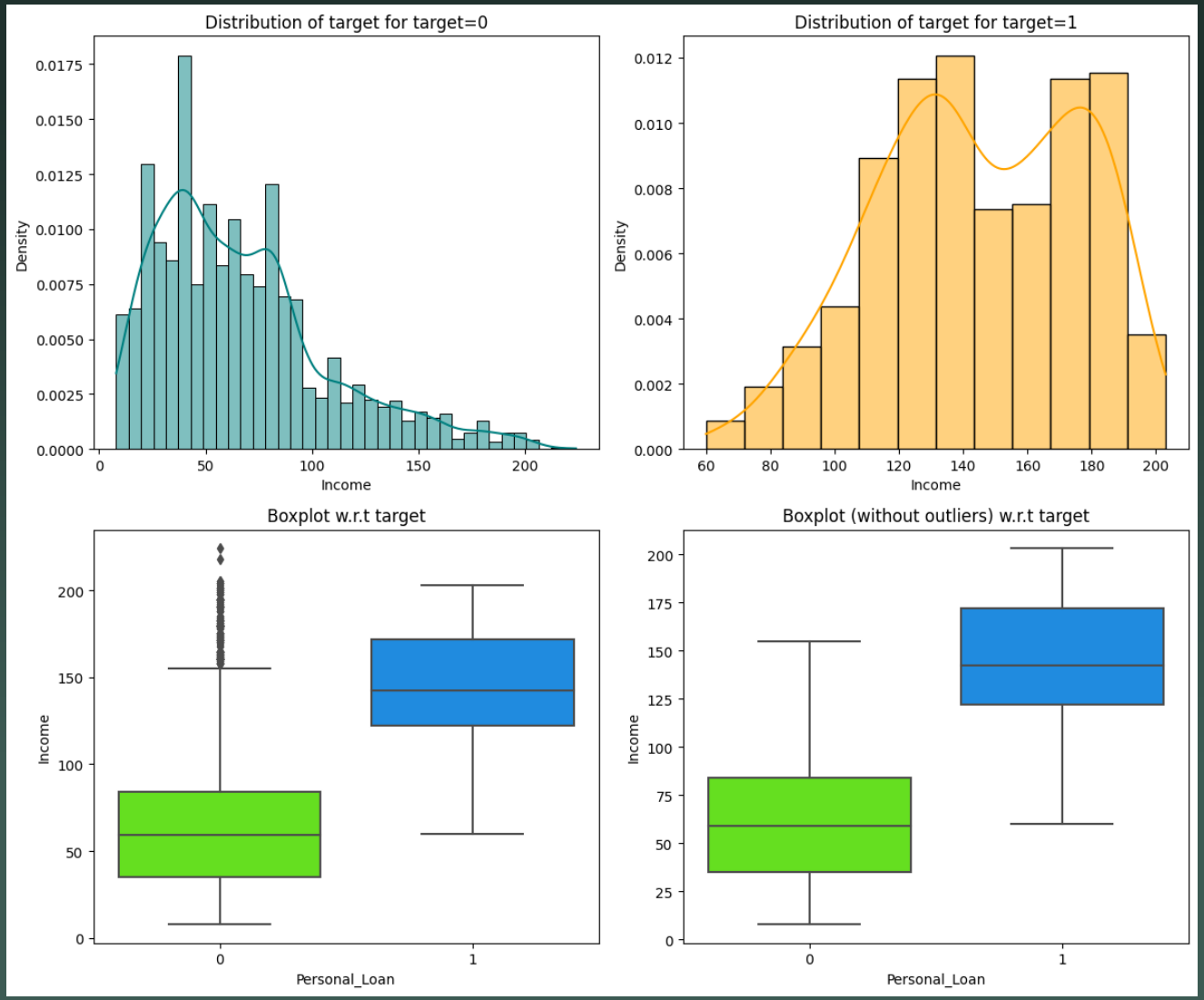
Appendix II

Personal Loan vs Experience



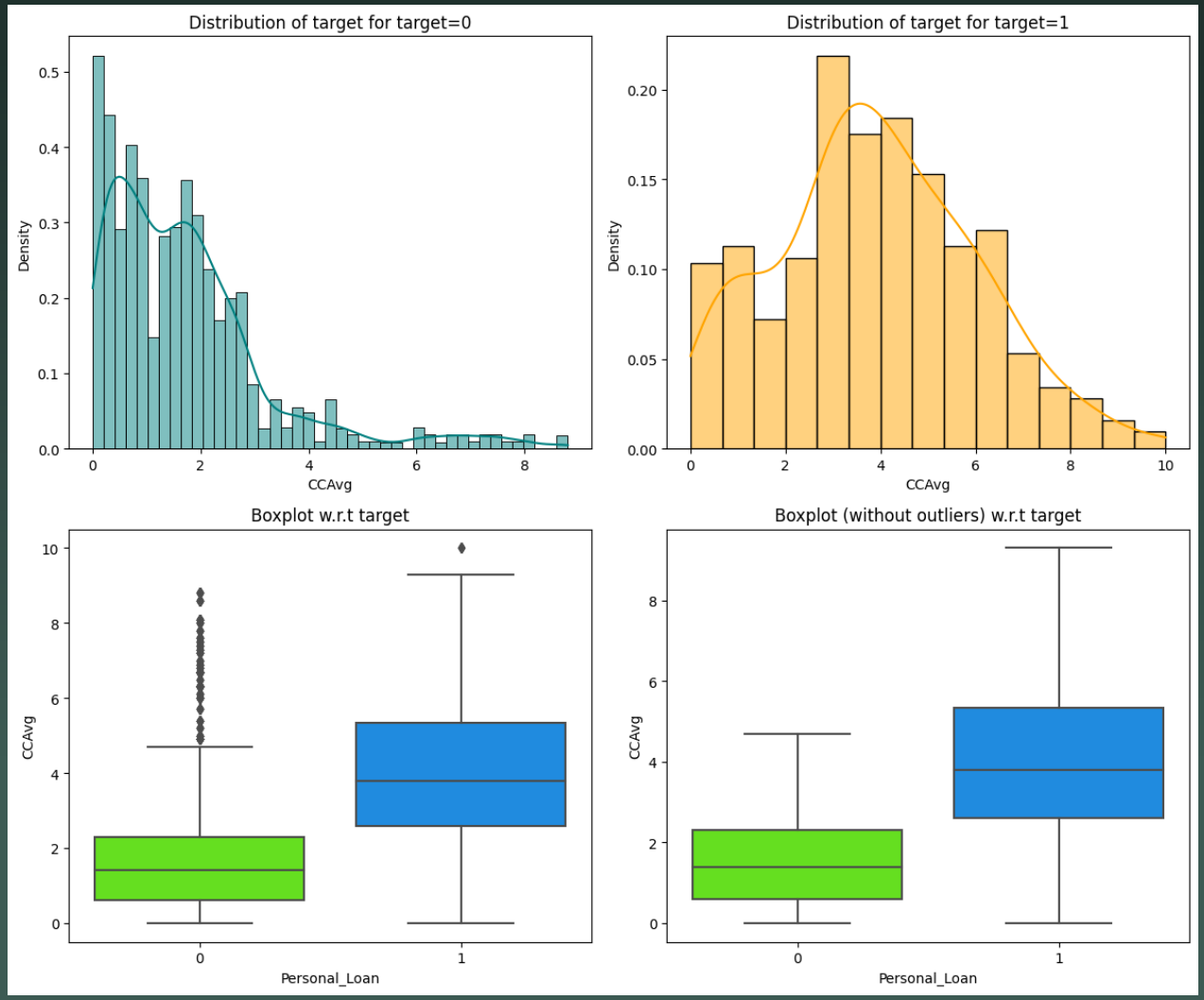
Appendix II

Personal Loan vs Income



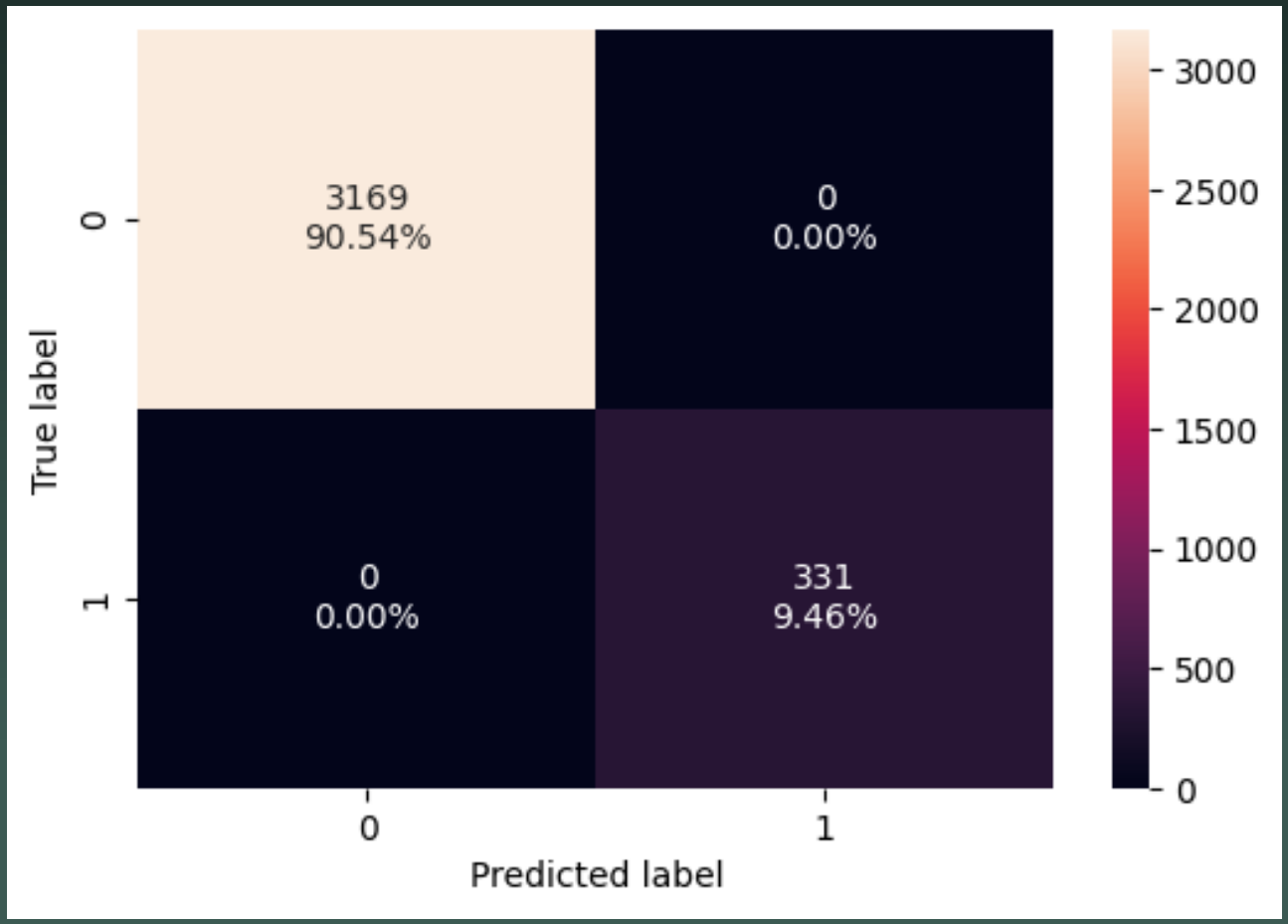
Appendix II

Personal Loan vs Monthly Credit Card Spend



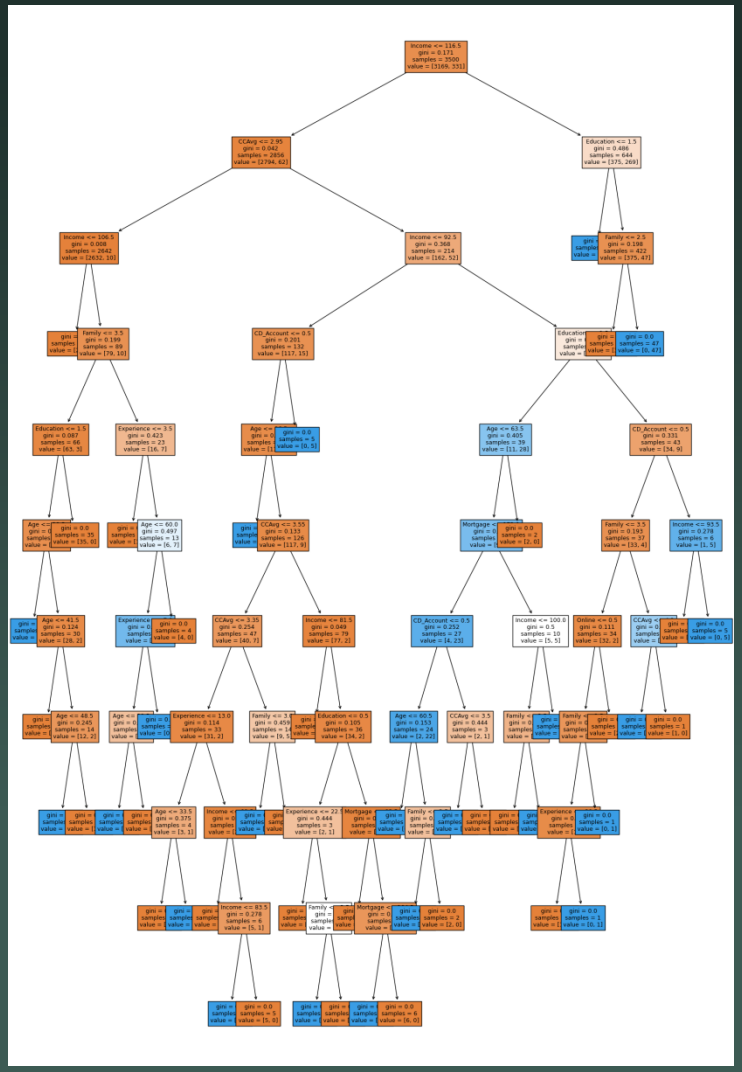
Appendix III

Checking Model Performance of Training Data



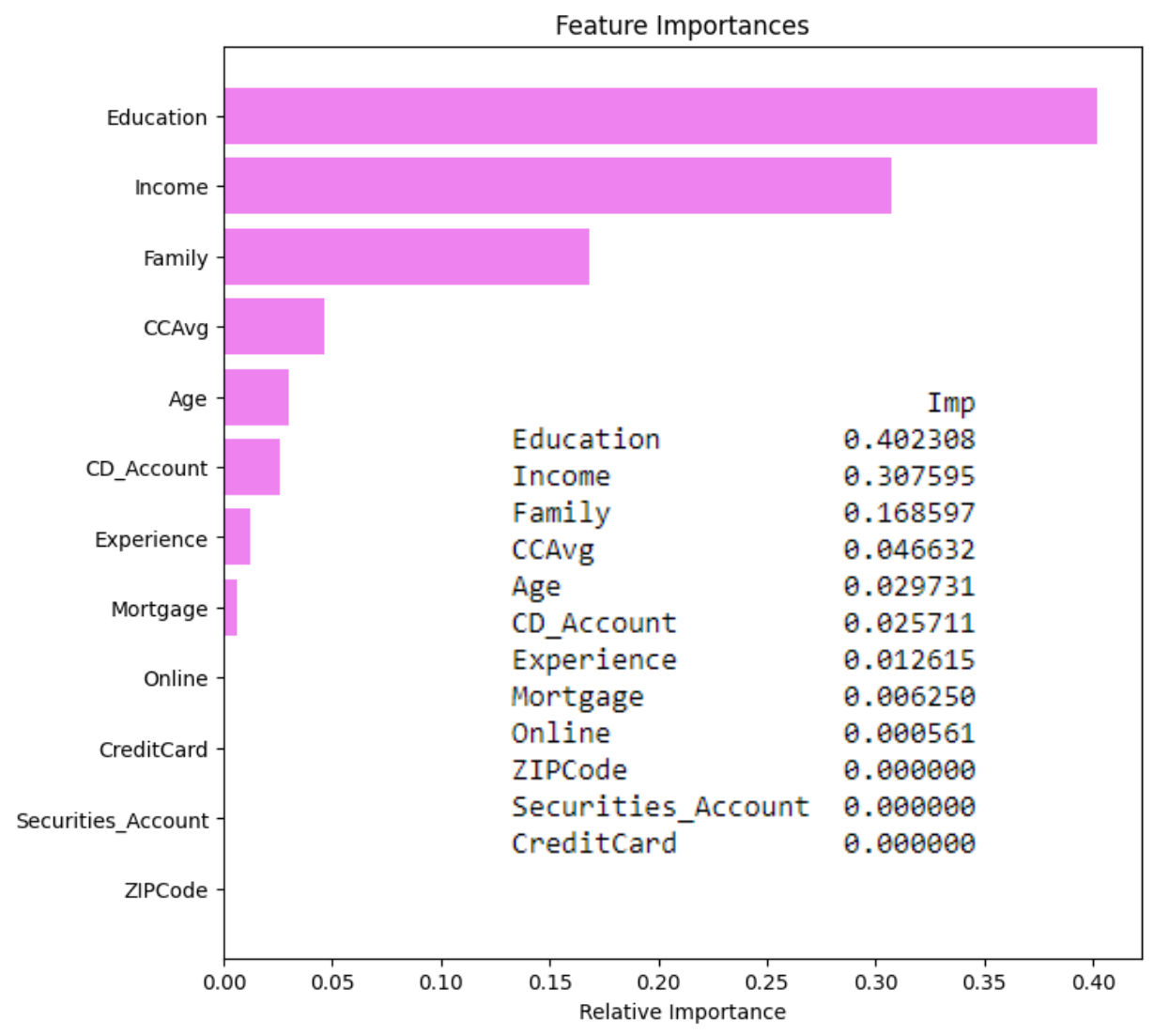
Appendix III

Visualizing Decision Tree



Appendix III

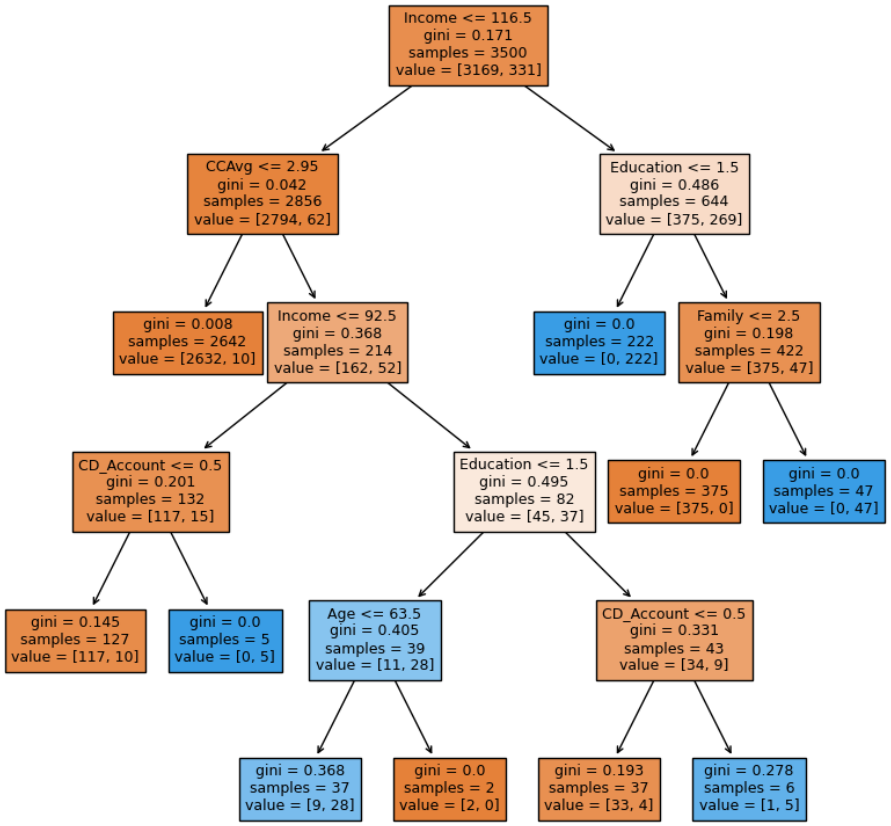
Feature Importance



	Imp
Education	0.402308
Income	0.307595
Family	0.168597
CCAvg	0.046632
Age	0.029731
CD_Account	0.025711
Experience	0.012615
Mortgage	0.006250
Online	0.000561
ZIPCode	0.000000
Securities_Account	0.000000
CreditCard	0.000000

Appendix III

Tuning Performance of the Decision Tree

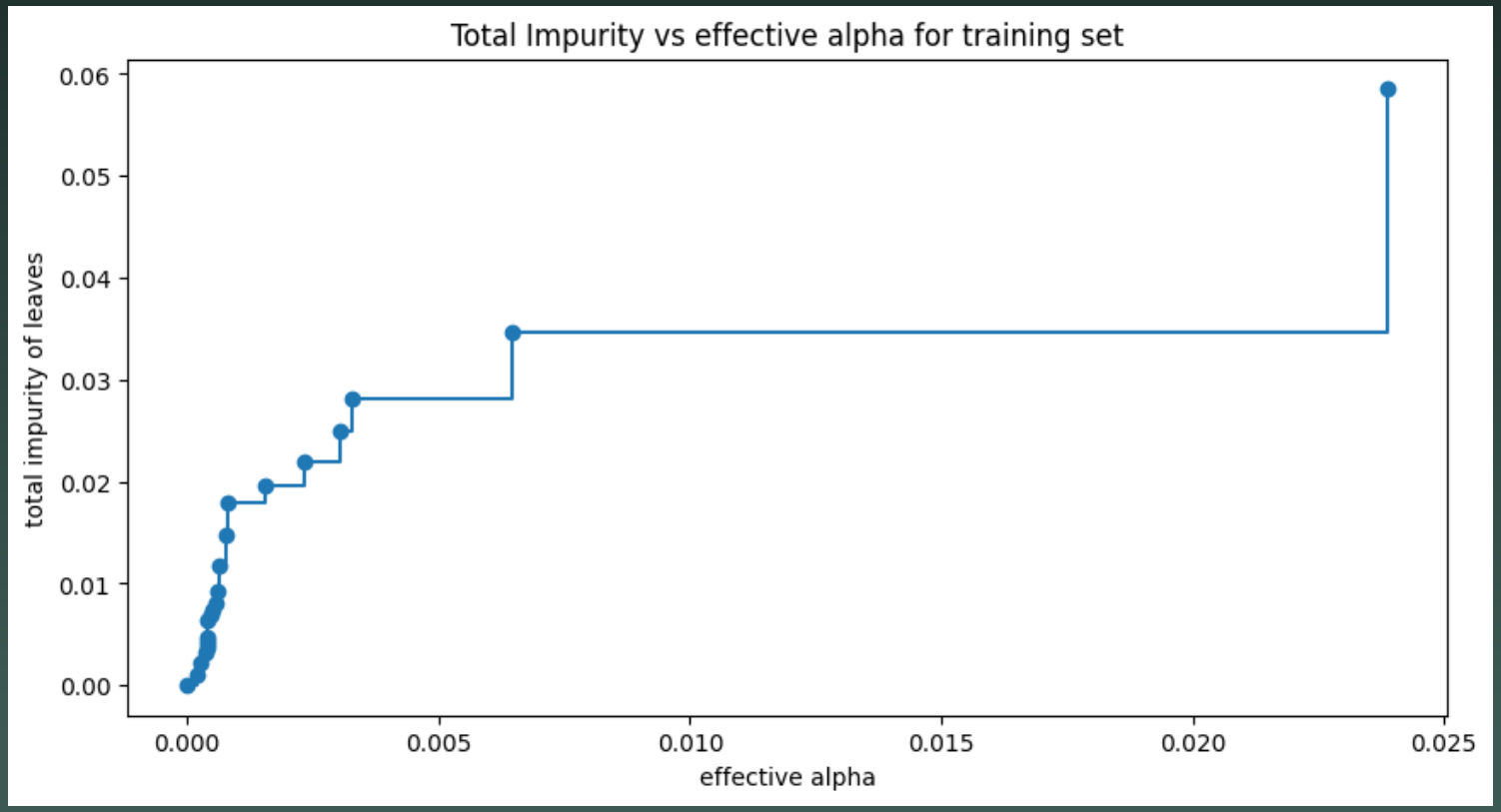



```
|--- Income <= 116.50
|   |--- CCAvg <= 2.95
|   |   |--- weights: [2632.00, 10.00] class: 0
|   |--- CCAvg >  2.95
|   |   |--- Income <= 92.50
|   |   |   |--- CD_Account <= 0.50
|   |   |   |   |--- weights: [117.00, 10.00] class: 0
|   |   |   |--- CD_Account >  0.50
|   |   |   |   |--- weights: [0.00, 5.00] class: 1
|   |   |--- Income >  92.50
|   |   |   |--- Education <= 1.50
|   |   |   |   |--- Age <= 63.50
|   |   |   |   |   |--- weights: [9.00, 28.00] class: 1
|   |   |   |   |--- Age >  63.50
|   |   |   |   |   |--- weights: [2.00, 0.00] class: 0
|   |   |   |--- Education >  1.50
|   |   |   |   |--- CD_Account <= 0.50
|   |   |   |   |   |--- weights: [33.00, 4.00] class: 0
|   |   |   |   |--- CD_Account >  0.50
|   |   |   |   |   |--- weights: [1.00, 5.00] class: 1
|--- Income >  116.50
|   |--- Education <= 1.50
|   |   |--- weights: [0.00, 222.00] class: 1
|   |--- Education >  1.50
|   |   |--- Family <= 2.50
|   |   |   |--- weights: [375.00, 0.00] class: 0
|   |   |--- Family >  2.50
|   |   |   |--- weights: [0.00, 47.00] class: 1
```

Appendix III

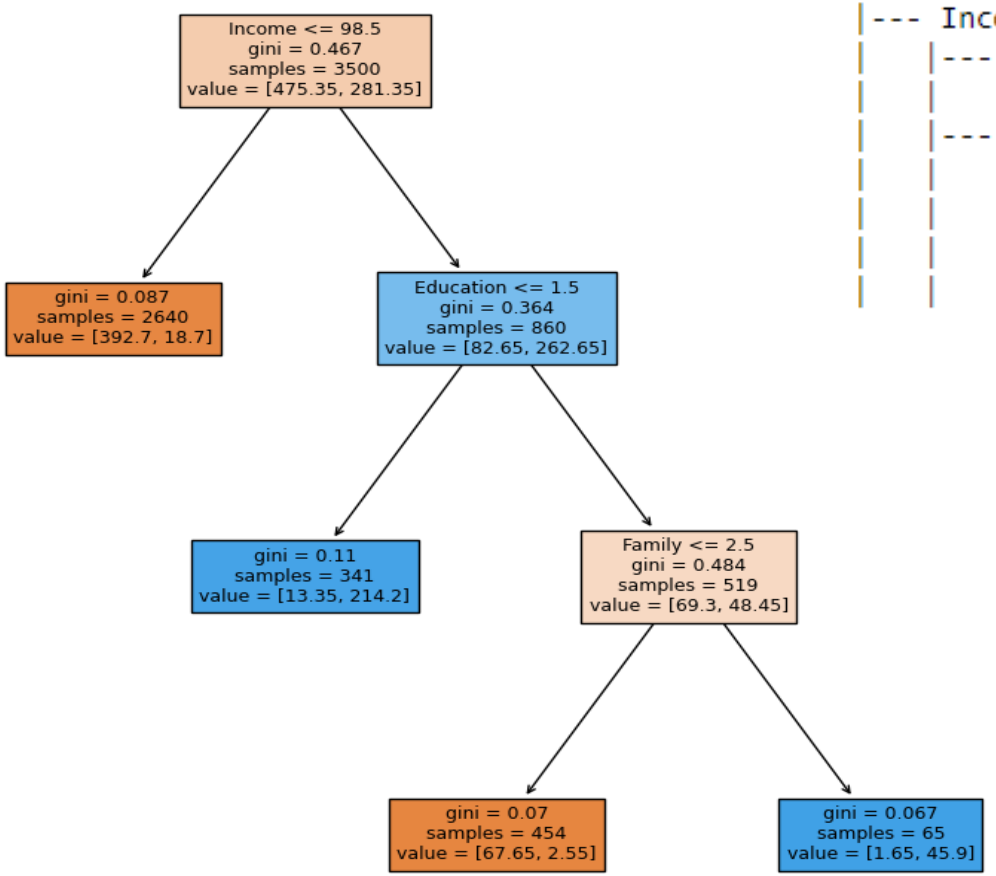
Cost-Complexity Pruning Alpha Values | Impurities

	ccp_alphas	impurities
0	0.000000	0.000000
1	0.000186	0.001114
2	0.000268	0.002188
3	0.000359	0.003263
4	0.000381	0.003644
5	0.000381	0.004025
6	0.000381	0.004406
7	0.000381	0.004787
8	0.000409	0.006423
9	0.000476	0.006900
10	0.000508	0.007407
11	0.000582	0.007989
12	0.000593	0.009175
13	0.000641	0.011740
14	0.000769	0.014817
15	0.000792	0.017985
16	0.001552	0.019536
17	0.002333	0.021869
18	0.003024	0.024893
19	0.003294	0.028187
20	0.006473	0.034659
21	0.023866	0.058525
22	0.056365	0.171255



Appendix III

Model Tuning Visualization

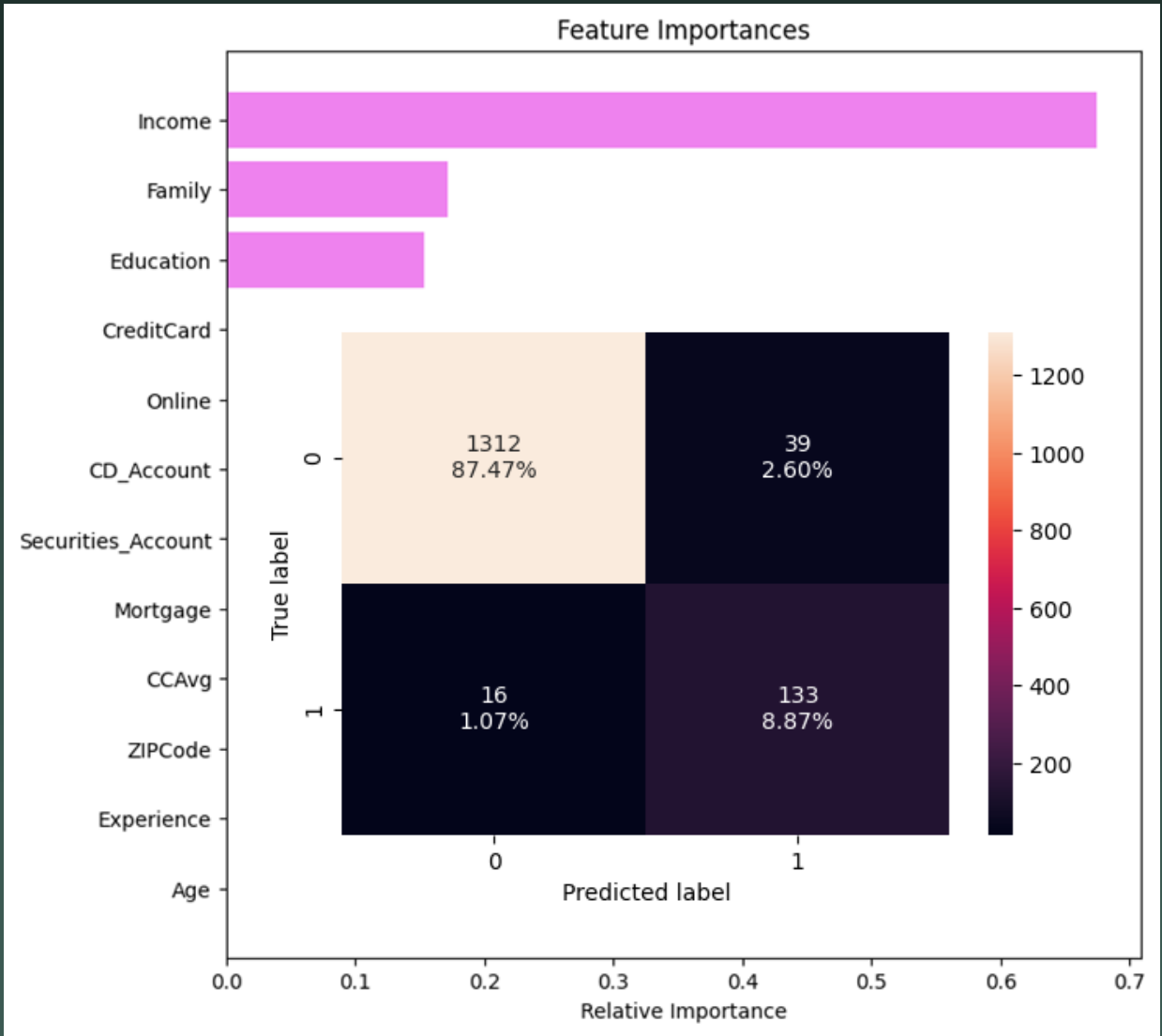



```
|--- Income <= 98.50
|   |--- weights: [392.70, 18.70] class: 0
|--- Income >  98.50
|   |--- Education <= 1.50
|   |   |--- weights: [13.35, 214.20] class: 1
|   |--- Education >  1.50
|   |   |--- Family <= 2.50
|   |   |   |--- weights: [67.65, 2.55] class: 0
|   |   |--- Family >  2.50
|   |   |   |--- weights: [1.65, 45.90] class: 1
```

	Imp
Income	0.674921
Family	0.171953
Education	0.153126
Age	0.000000
Experience	0.000000
ZIPCode	0.000000
CCAvg	0.000000
Mortgage	0.000000
Securities_Account	0.000000
CD_Account	0.000000
Online	0.000000
CreditCard	0.000000

Appendix III

Feature Importance & Performance Check For Test Data



Appendix III

Training Performance Comparison

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.980667	1.0	0.964286
Recall	0.892617	1.0	0.924471
Precision	0.910959	1.0	0.753695
F1	0.901695	1.0	0.830393