AllLife Bank Personal Loan Campaign

Supervised Learning Classification Project

July 20th, 2023

Prepared by: Andres Herrejon Maya.

Contents / Agenda

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

Executive Summary

- Decision Tree Model Post-Prunned is the selected model with the best performance and no overfitting.
- A conversion of 8.80% for acceptance for Personal Loan is forecasted as a result of the application of this model.

Marketing Recommendations

- Target family of 3 for depositors and then for creditors. Has the highest conversion rate and the smallest population among the customers.
- Offer personal Loans through CD Accounts as a packaged combo showing financial earnings.

Business Problem Overview and Solution Approach

Business Problem Overview

AllLife Bank wants to explore ways of **converting its liability customers to personal loan customers** (while retaining them as depositors) and we need to **identify the potential customers who have a higher probability of purchasing the loan** based on the data of 5000 clients that we offered a personal loan from a previous campaign.

Solution Approach

To predict whether a liability customer will buy personal loans, to understand which customer attributes are most significant in driving purchases, and identify which segment of customers to target more, we are going to pick between **3 Decision Tree Models** the one that shows **the highest Recall Score as we want to minimize the false negatives and miss the business opportunity.** The models may differ in attributes or in size for **pruning methods may be applied.**

Data Preprocessing

Duplicate value check

All rows are unique as Customer ID has values from 1 to 5000 from a 5000 row data set.

Missing value treatment

No missing values found. Although anomalies were found on the Experience attribute where -3, -2 and -1 were found and treated as an error and replaced by their respective positive value.

Feature Engineering

We kept the first 2 digits of the ZIPCode attribute to keep only 7 groups of cities instead of handling 467 different ZIPCodes. The cities represented in each of those ZipCodes can be found on the appendix section.

Data Preprocessing

Outlier check

No treatment for outliers as the % of detected outliers are very low on present attributes:

Attribute	% of data detected as Outlier
Income	1.92 %
CCAvg	6.48 %
Mortgage	5.82 %

EDA Results

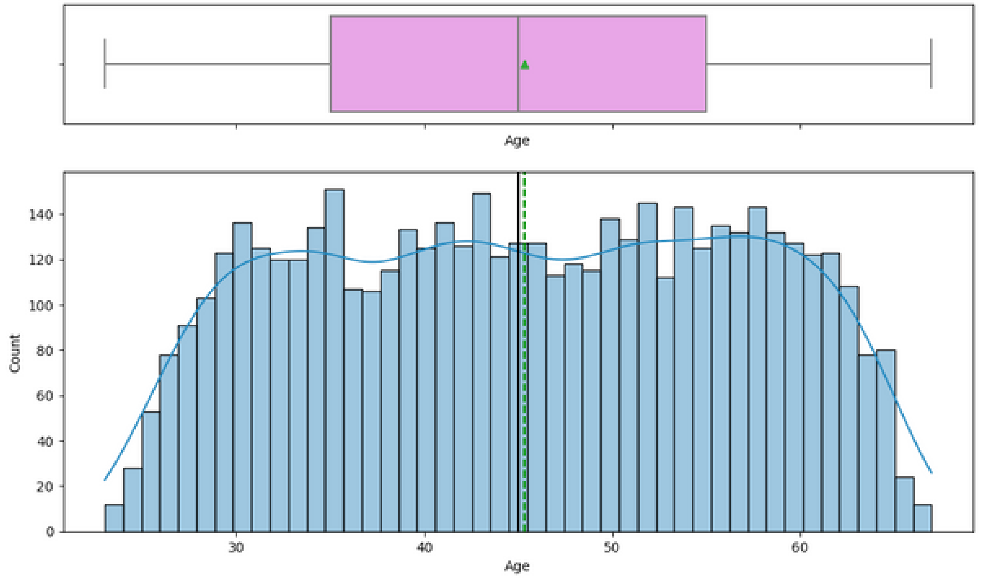
Age

Shows normal distribution and no outliers

Ages ranges from 23 to 65 years

On average, customers are on the 45 year old mark

50% of the customers fall in the range of 35 - 55 years old



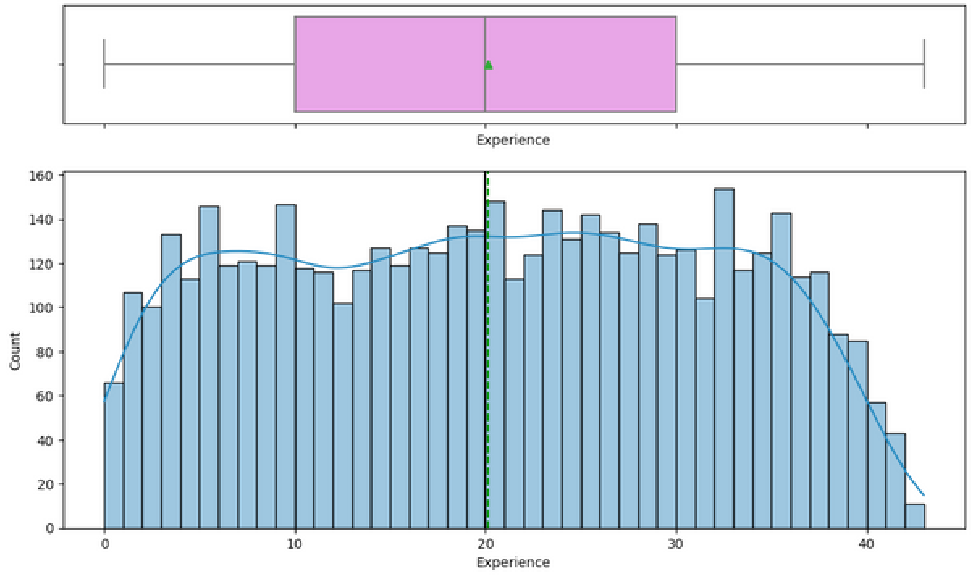
Experience

Shows normal distribution and no outliers as well.

Years of experience ranges from 0 to 43 years.

On average, customers have 20 years of experience.

50% of the customers fall in the range of 10 - 30 years of experience.



EDA Results

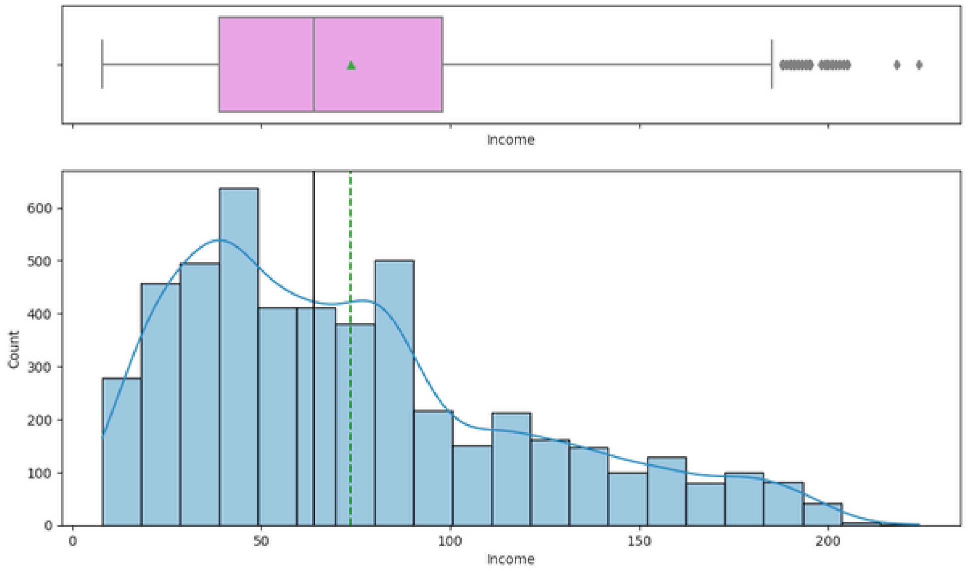
Income

Shows distribution skewed to the right and outliers are shown in the high income spectrum

Income ranges from $8K to $224K annually

On average, customers have an annual income of $73.77K

50% of the customers have an income in the range of $39 - $98K annually.



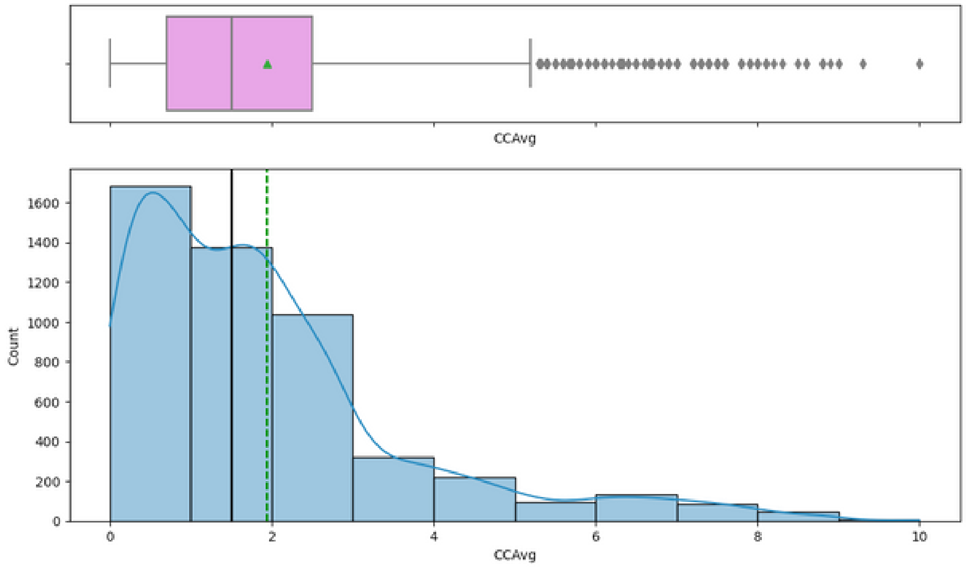
CCAvg

Shows also a distribution skewed to the right and outliers are shown in the high expenditure spectrum

Monthly expenses ranges from $0 to $10K

On average, customers have credit card payments of $1.93K

50% of the customers are paying between $700 and $2,500 monthly on credit cards.



EDA Results

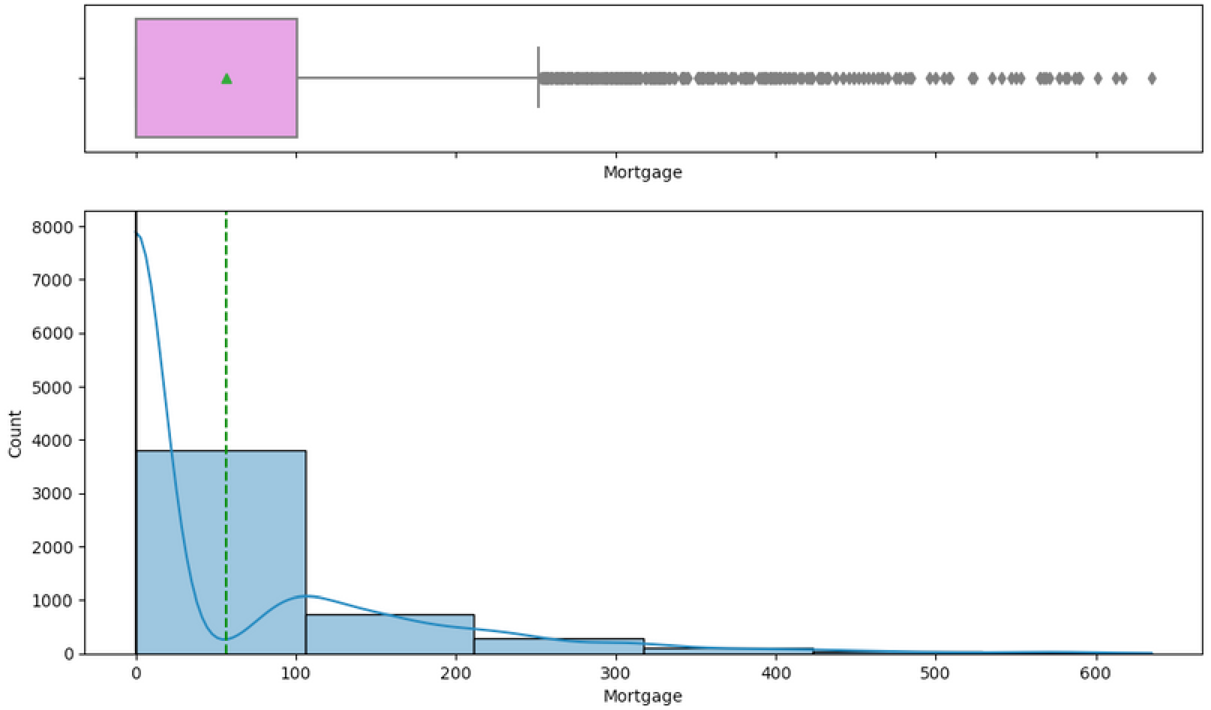
Mortgage

Shows distribution skewed to the right and outliers are shown in the high debt side

Mortgage ranges from $0 to $635K

On average, customers have a mortgage of $56.49K

75% of the customers fall in the mortgage range of $0 - $101K



EDA Results

Correlation Check

We find almost a perfect positive correlation between Experience and age. As a result, we are going to drop Experience from the data set when building the model.

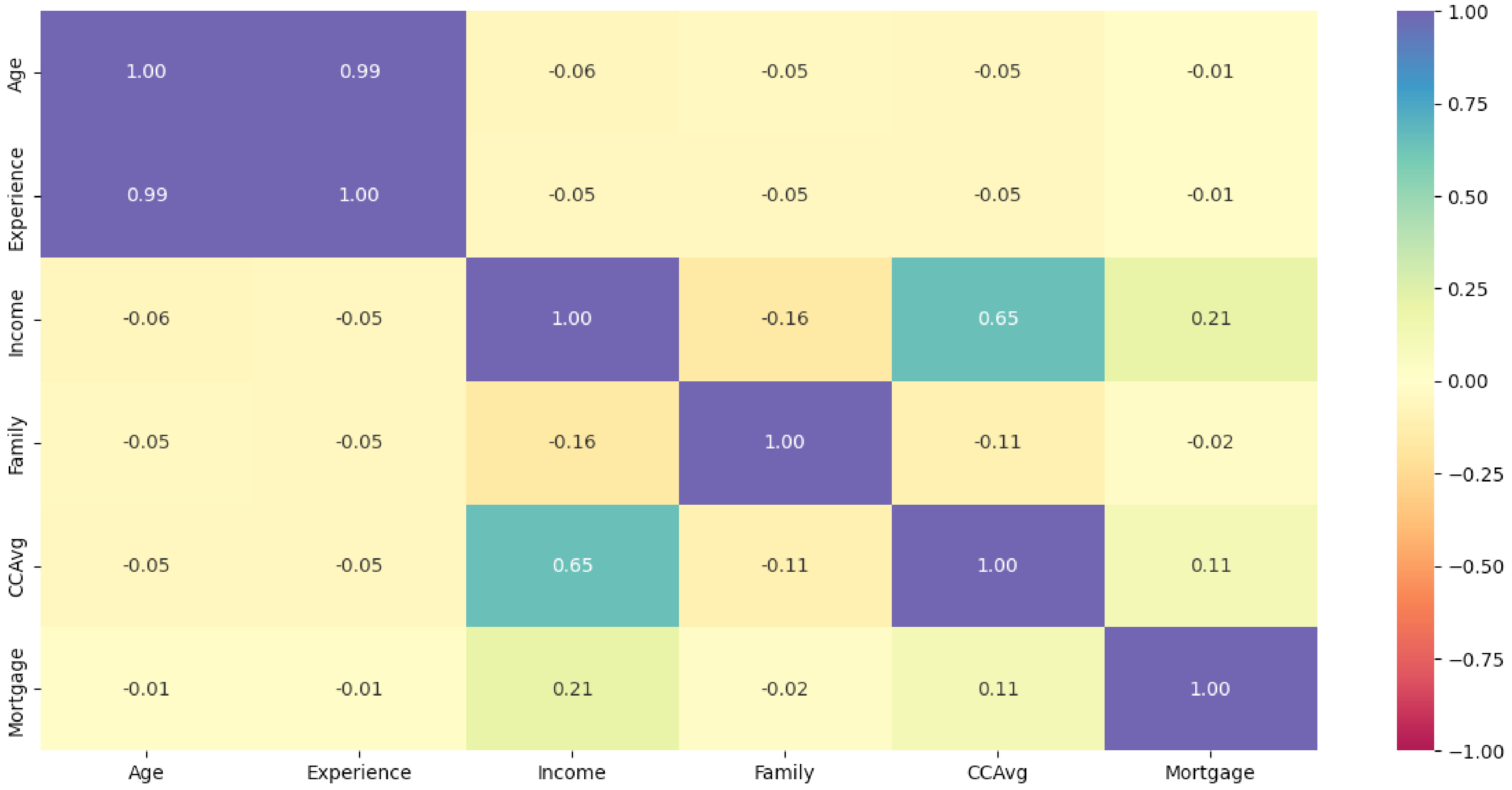
Next we have a strong positive correlation of .65 between CCAvg and Income as the higher the income higher credit line is authorized and customer can incur in more expenses

A positive correlation of .21 is found between Income and Mortgage. We can read this as the higher income, the client is able to get a higher mortgage.

A slightly positive correlation worth mentioning is between CCAvg and Mortgage of .11

Family poses negative correlations with Income and and CCAvg of -0.16 and -0.11 respectively.

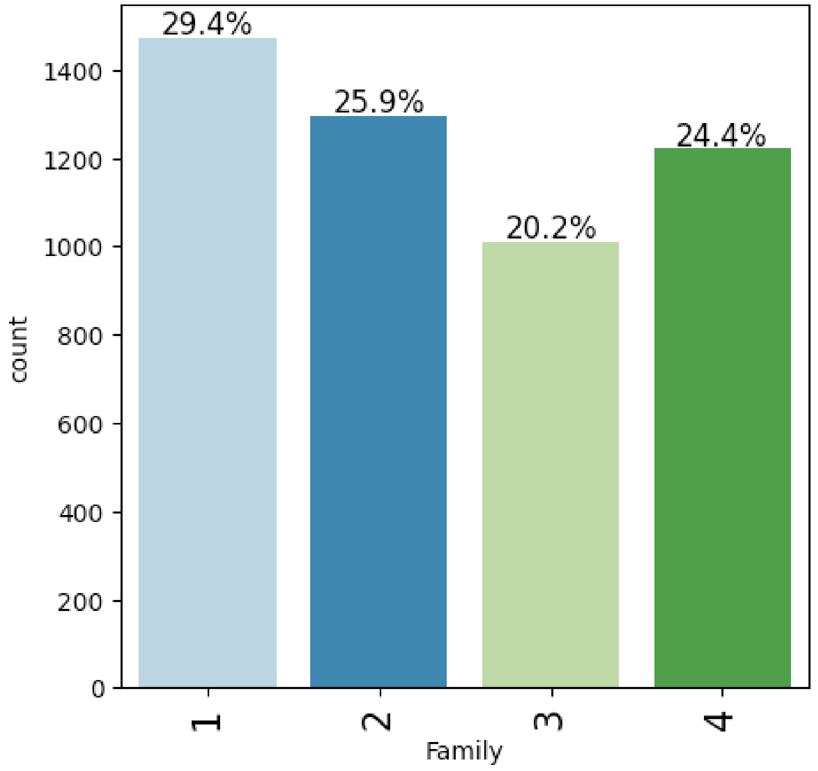
The rest of the correlations are nearly neutral.



EDA Results

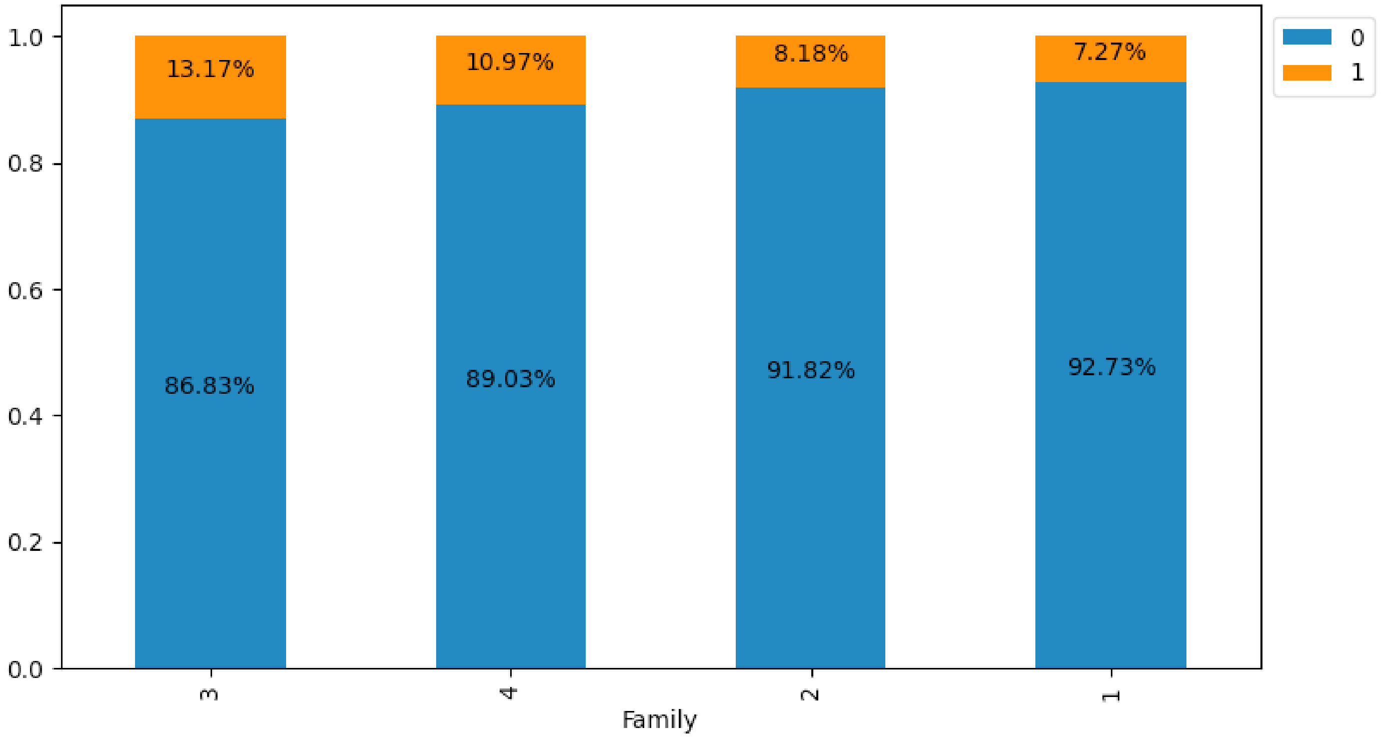
Family

- Family members ranges from 1 to 4
- almost 30% of the customers live in a household of 1
- On average, customers have a family of 2 members



Family and Personal Loan

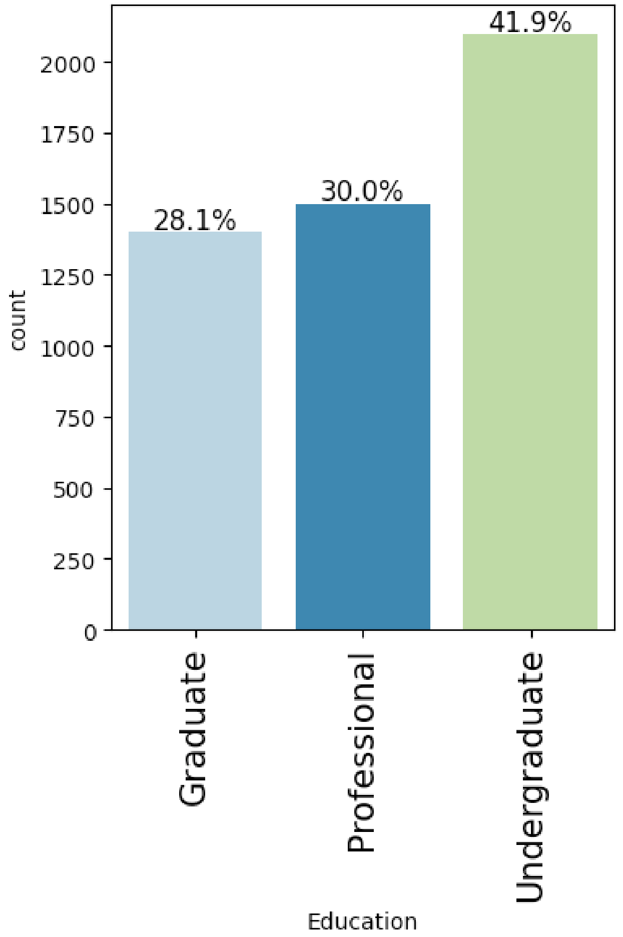
- Family of 3 Group has the top percentage (13.7%) of the customers that accepted the Personal Loan offer. Which make up only 20.2% of the sample (the smallest group of the data set)
- The largest group of Family members in the sample is 1 (29.4%) and corresponds to a 7.27% of offer acceptance (the lowest conversion rate observed)
- We can observe that as a general rule the bigger the number of family members, the probability to accept the Credit Card increases being the more prone to the 3 member group.



EDA Results

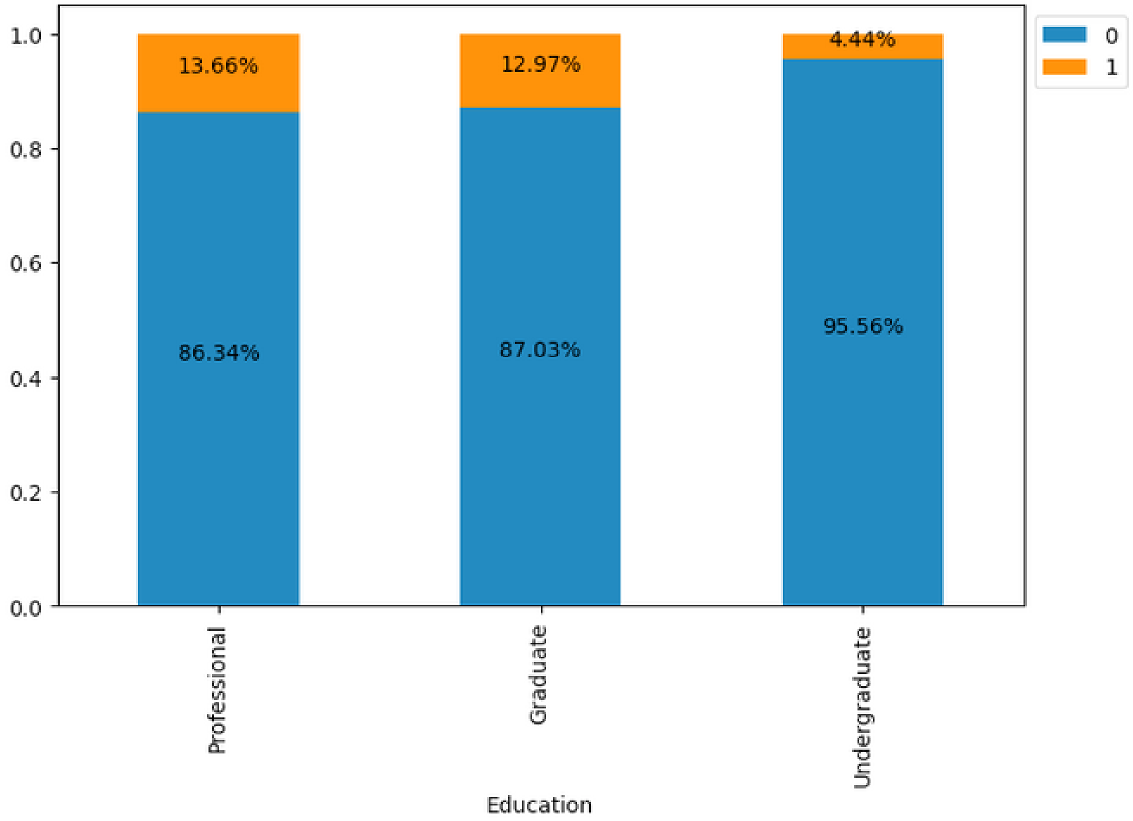
Education

- Customers that have completed academic studies (Professional + Graduate) have a near 60% - 40% proportion vs Undergraduate customers.
- 1/3 of the customers have advanced / Professional careers.



Education and Personal Loan

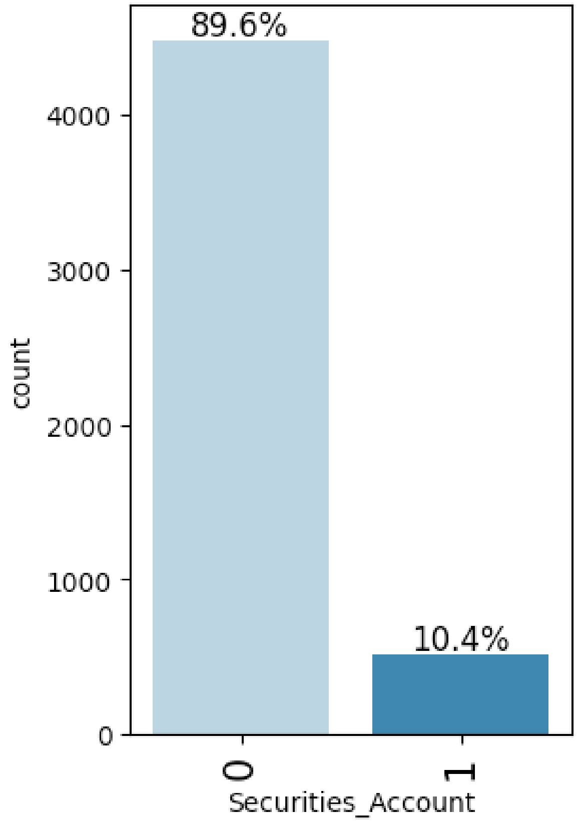
- We can observe a big difference in conversion rate between the Undergraduate group with a rate of 4.44% and the Professional and Graduate group with rates of 13.66% and 12.97% respectively that add up near 60% of the population in the data set.
- We observe that the advancement in career / academics is an attribute that increases the chances of conversion.



EDA Results

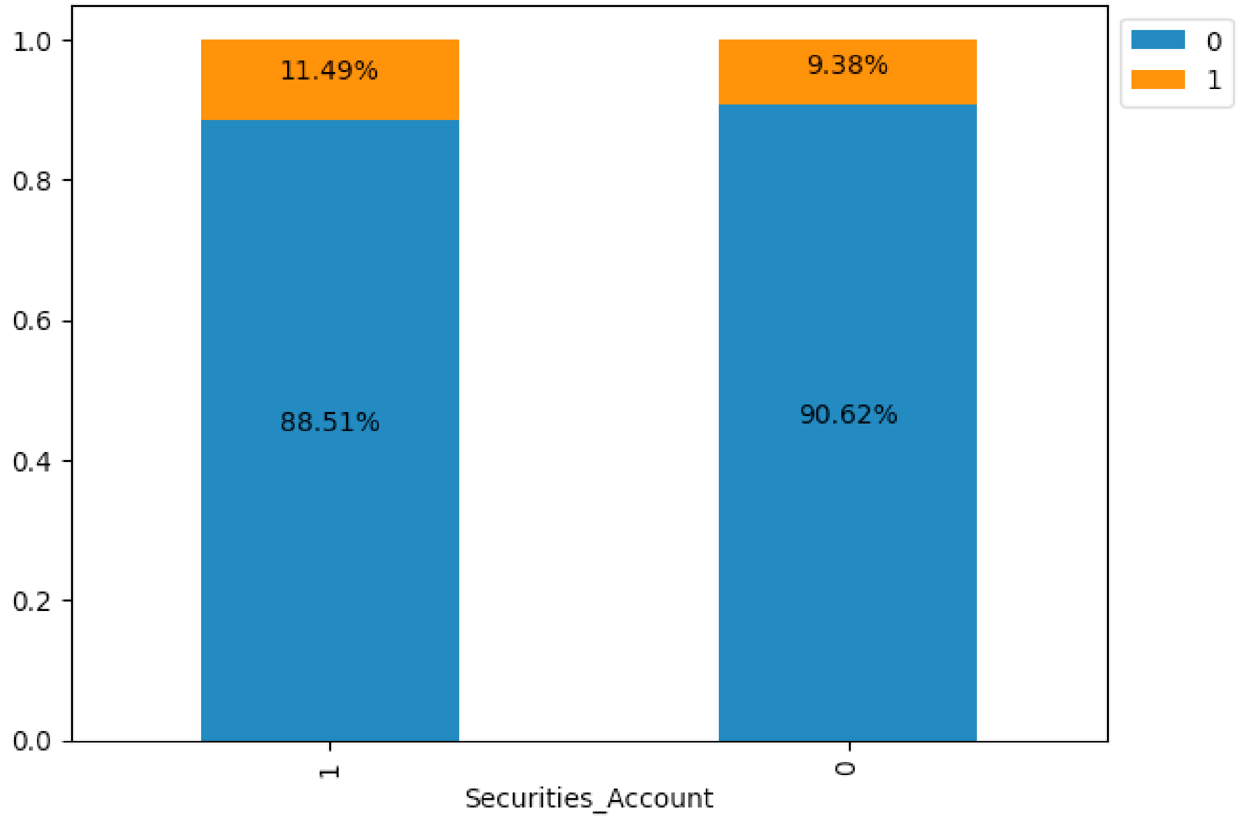
Securities_Account

- Only 10.4% of the customers are investing their money through a portfolio of some kind.



Securities_Account and Personal Loan

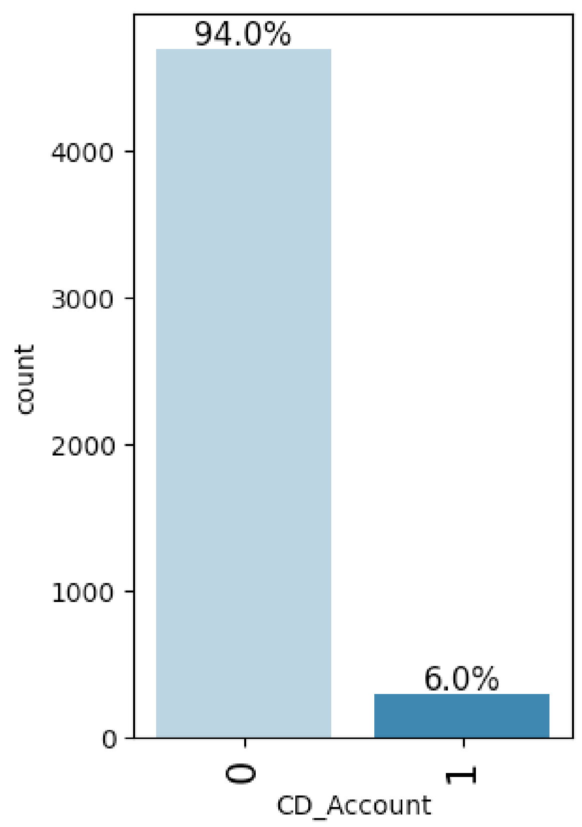
- Although liquidity might be a problem which solution may come with some financial cost result of a forced sell of assets the conversion rate is only 2% higher in those customers who have a Securities Account over those who don't.



EDA Results

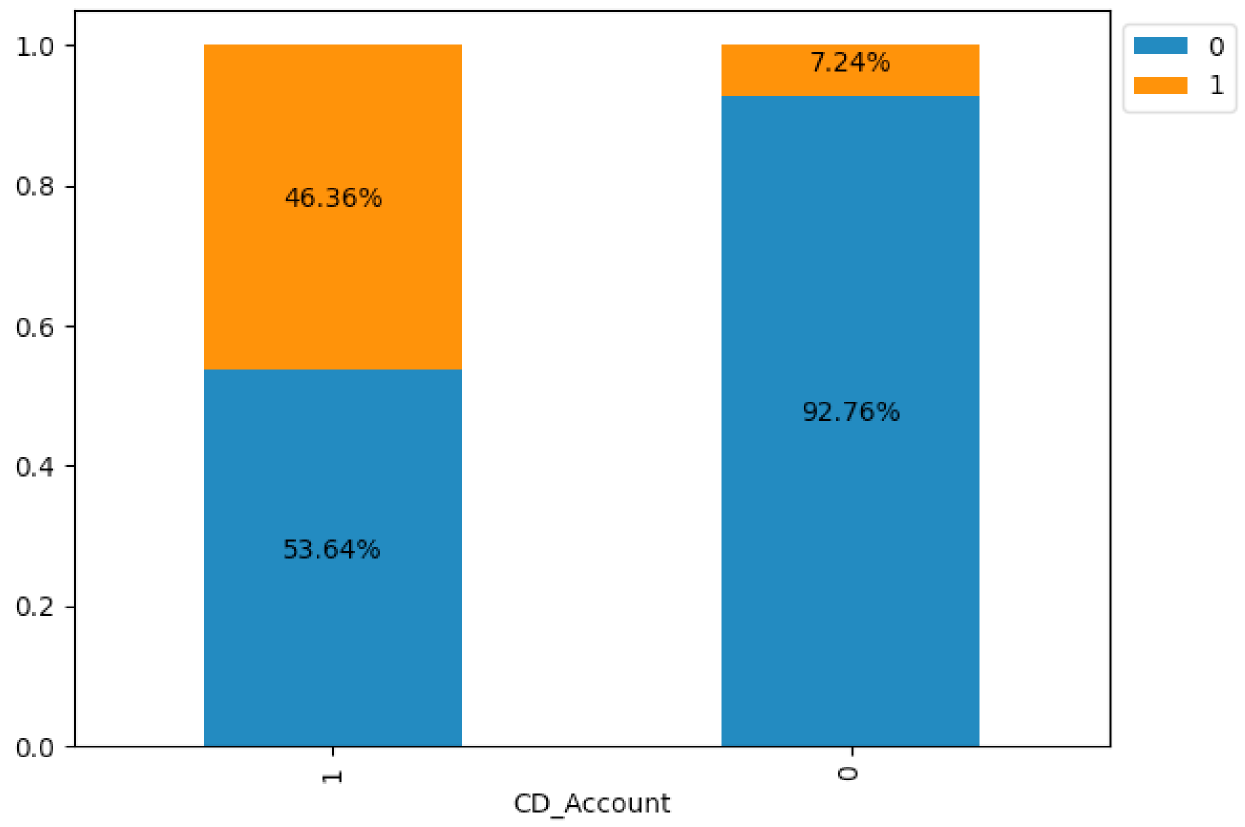
CD_Account

- In the case of Certificates of Deposit, only 6% of the customers have the potential problem of have a liquidity problem for having money invested for a period of time.



CD_Account and Personal Loan

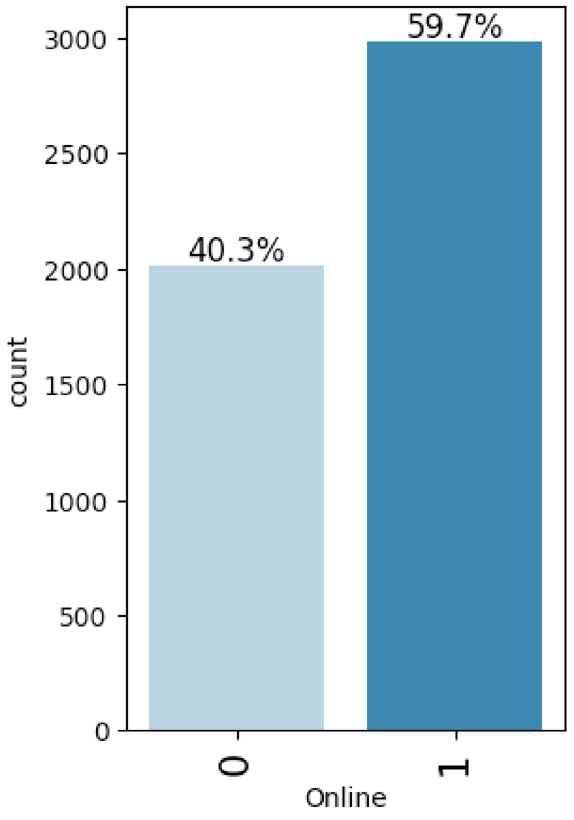
- The highest conversion rate observed is among the group of customers that have a CD_Account. This could be related with the ability to finance their expenses while their money is giving returns through investments that may apply penalties if withdrawn early.



EDA Results

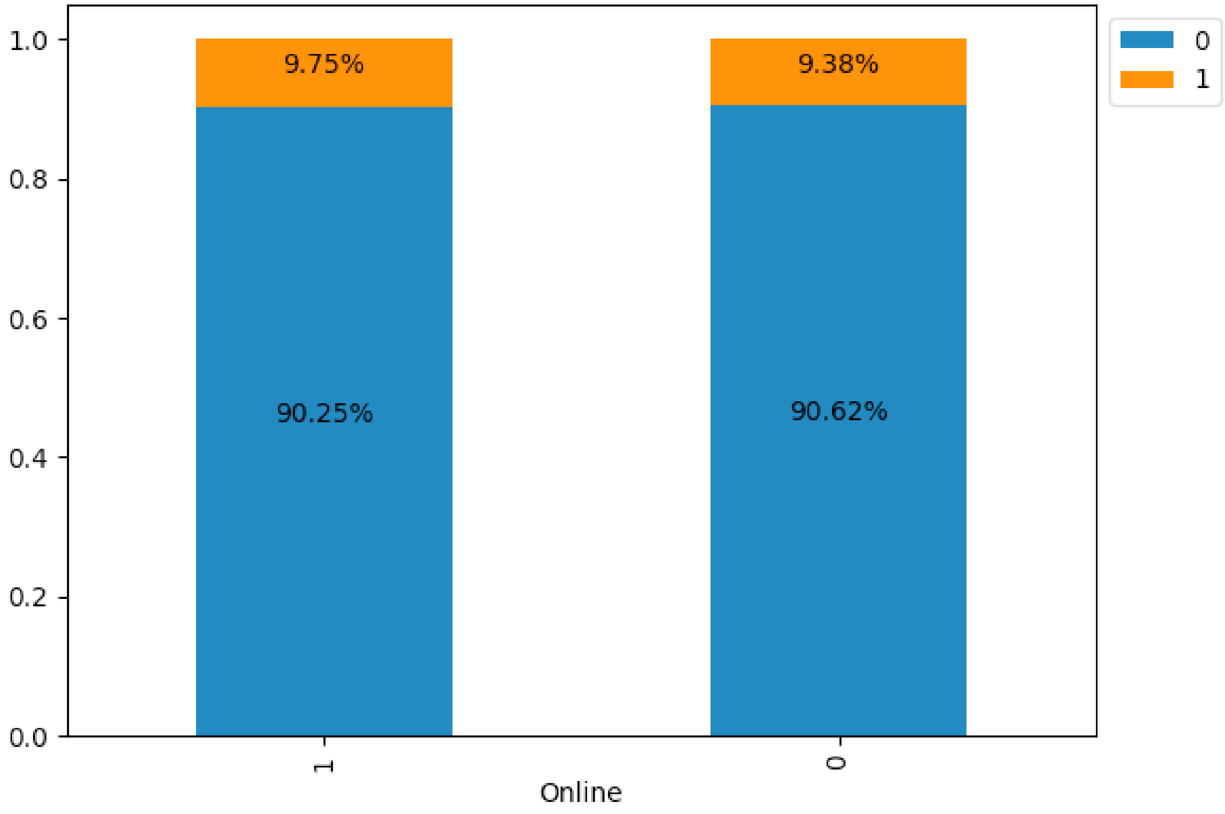
Online

- Almost 60% of the customers have online banking



Online and Personal Loan

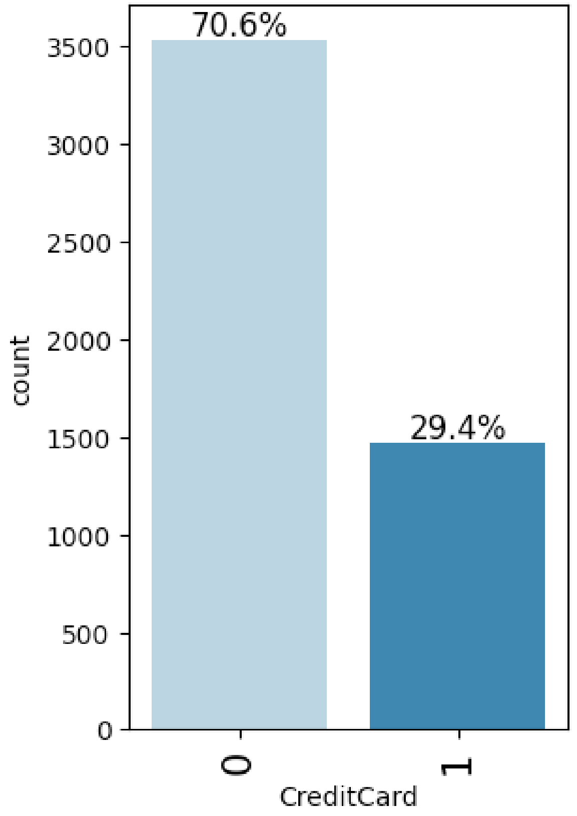
- Similar conversion rates are shown across customers with online baking and no online banking with 9.75% and 9.38% respectively.
- No relevant difference on this attribute on conversion rate observed.



EDA Results

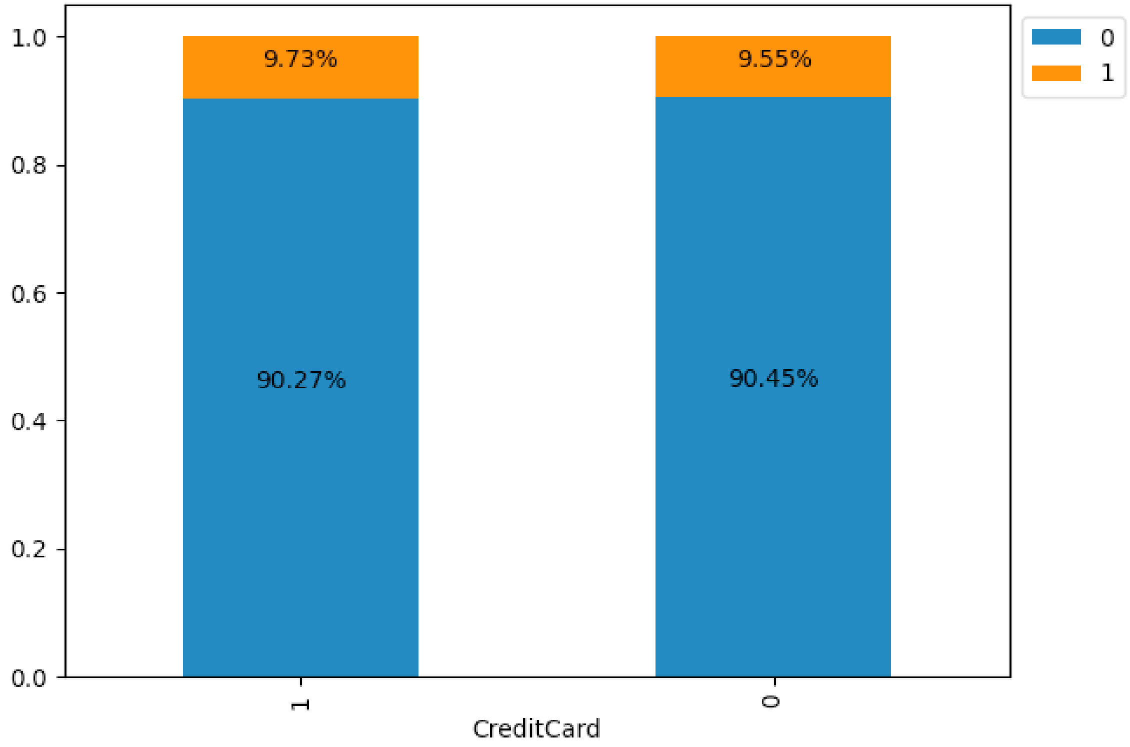
CreditCard

- Almost 30% of customers already have a credit card with another bank.



Credit Card and Personal Loan

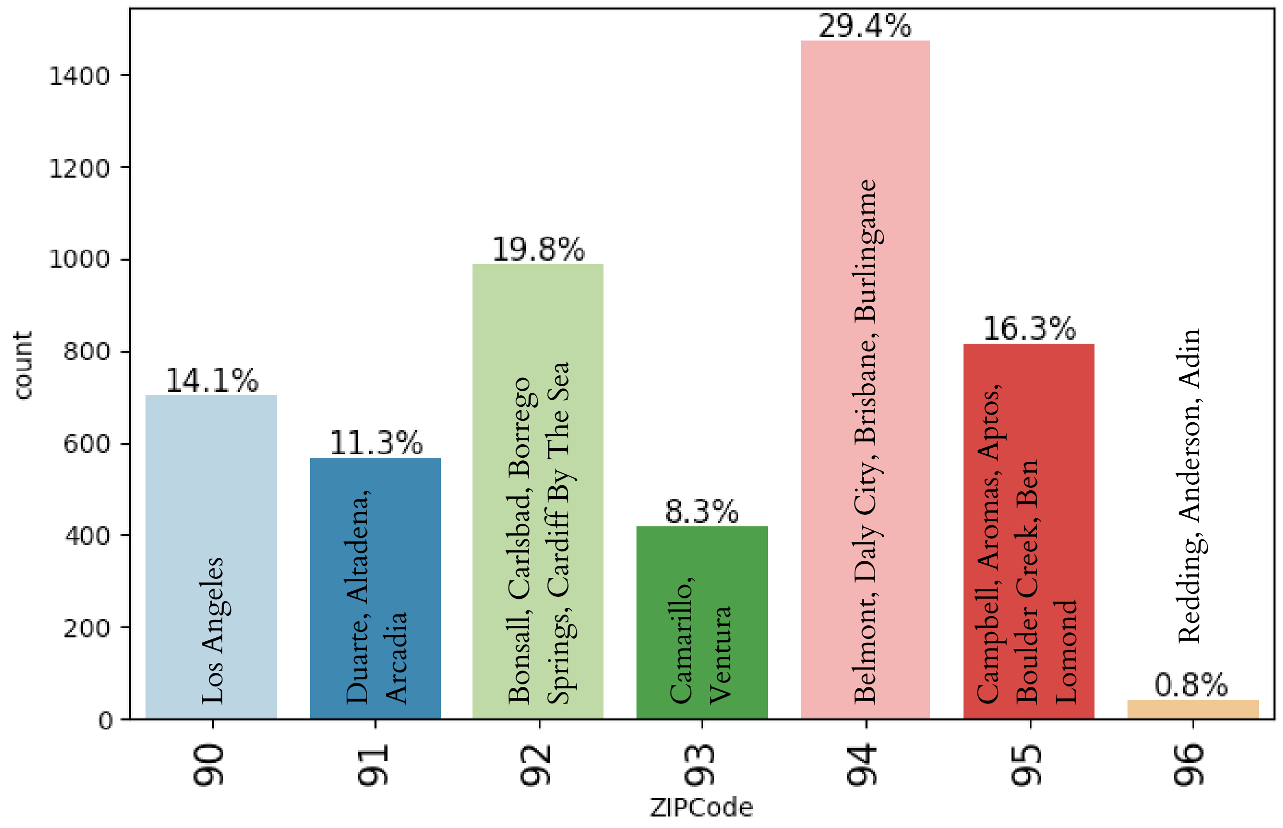
- Similar conversion rates are shown across customers with an existing Credit Card from another bank and no credit card with 9.73% and 9.55% respectively.
- No relevant difference on this attribute on conversion rate observed weather this is the first credit card customer would have or the second or more.



EDA Results

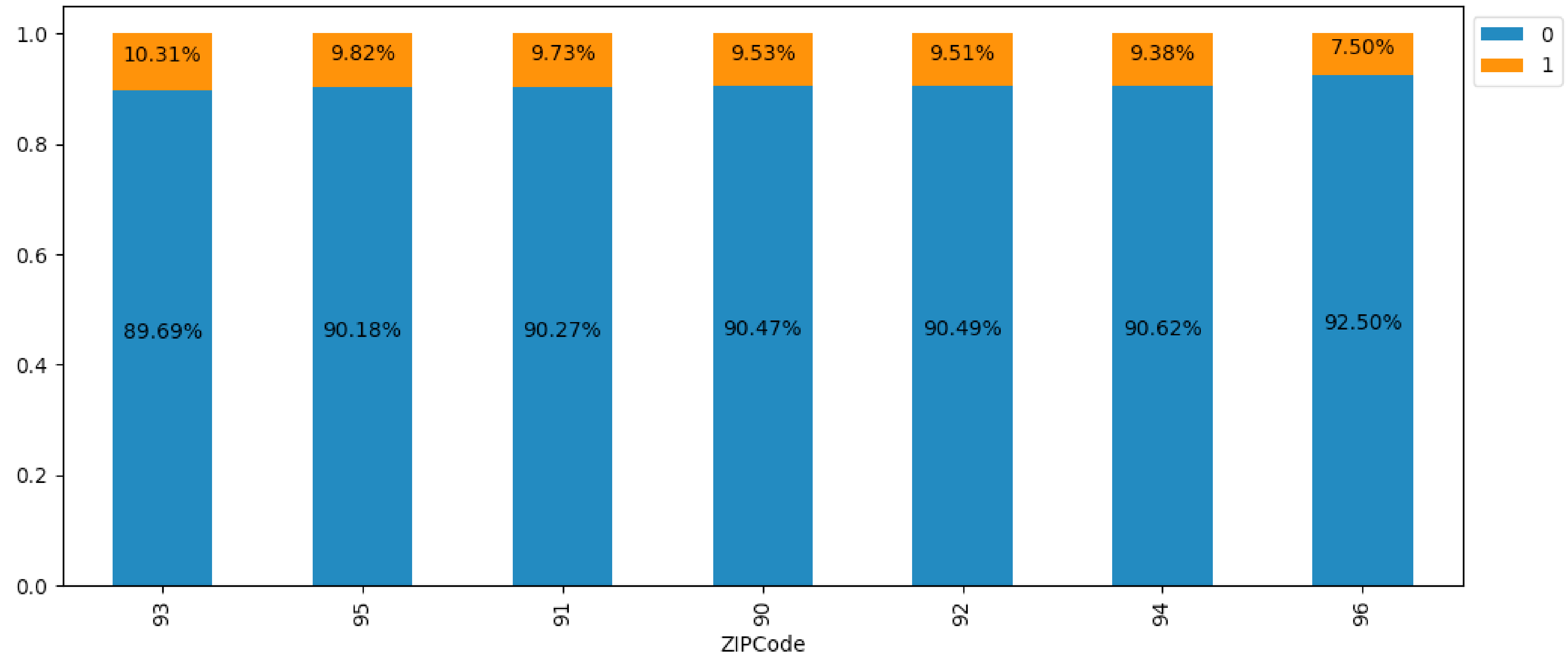
ZipCode

- The distribution of the 7 clusters resulted from keeping the first 2 digits of the ZipCode and showing the main cities corresponding to those clusters
- The biggest cluster is 94.



ZipCode and Personal Loan

- The largest cluster is for prefix 94 and shows a Personal Loan acceptance of 9.38% which is over the mentioned healthy conversion rate of 9%
- Most of them follow this tendency being cluster 96 the one with lowest success rate of 7.5% but also this cluster is the smallest cluster representing only 0.8% of the data set.

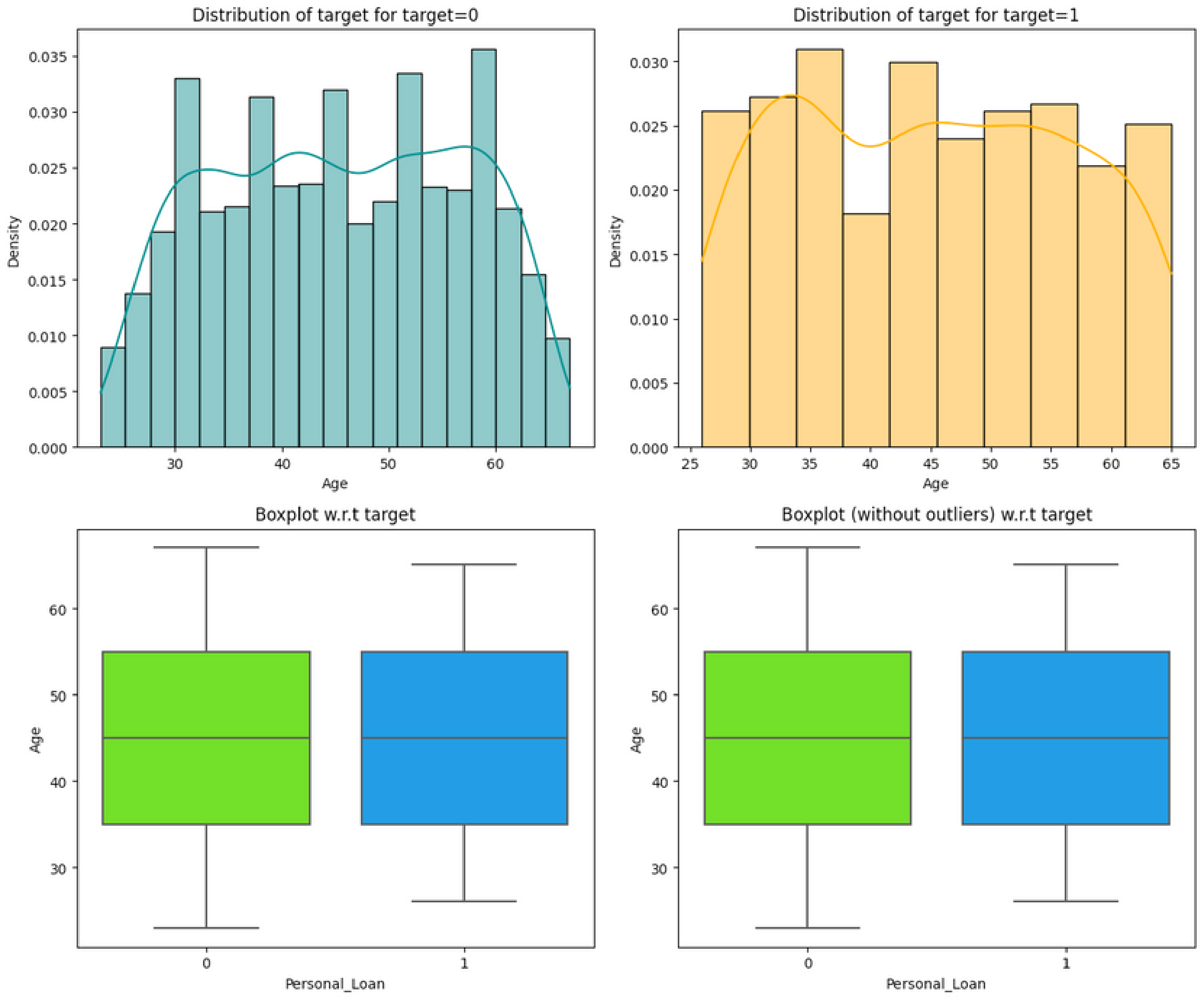


EDA Results

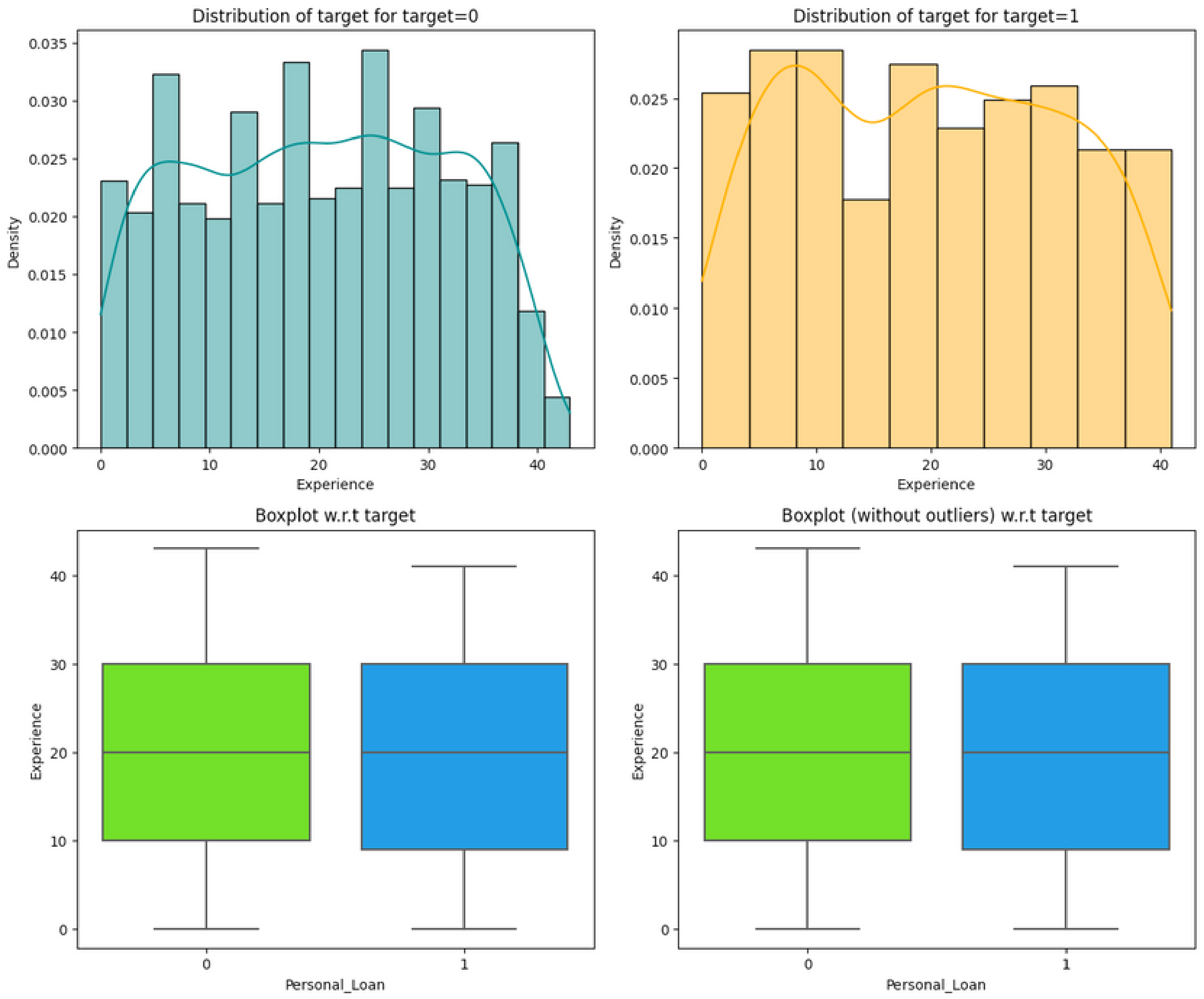
Age and Experience impact on Personal Loan purchase interest

- We can observe alike distributions on Age and Experience impact on purchase interest of a Personal Loan.
- This is because they have almost a perfect positive correlation (logically the older a customer gets the more experienced) which is the final reason on why Experience attribute is going to be dropped before building our model.

Age and Personal Loan



Experience and Personal Loan

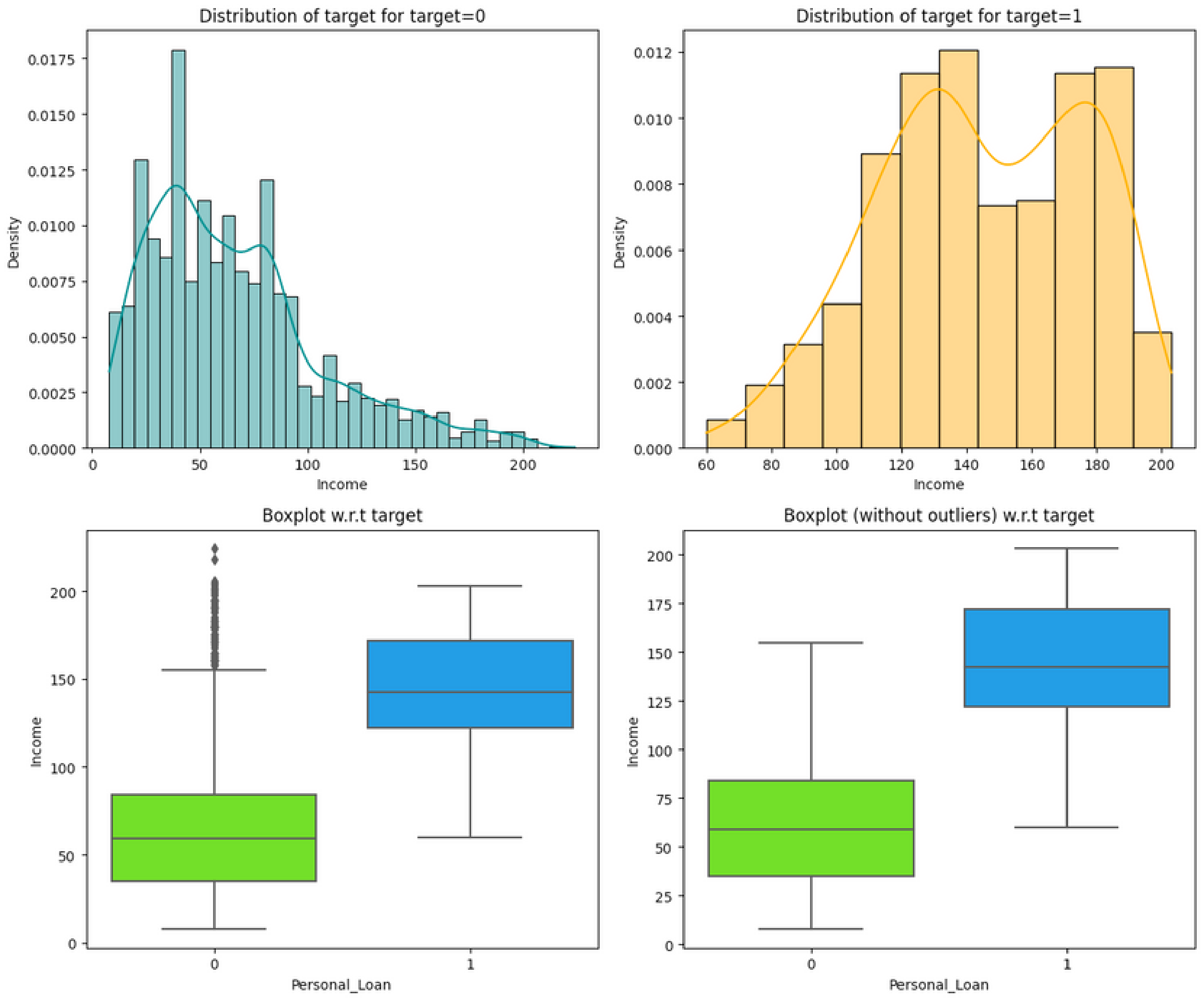


EDA Results

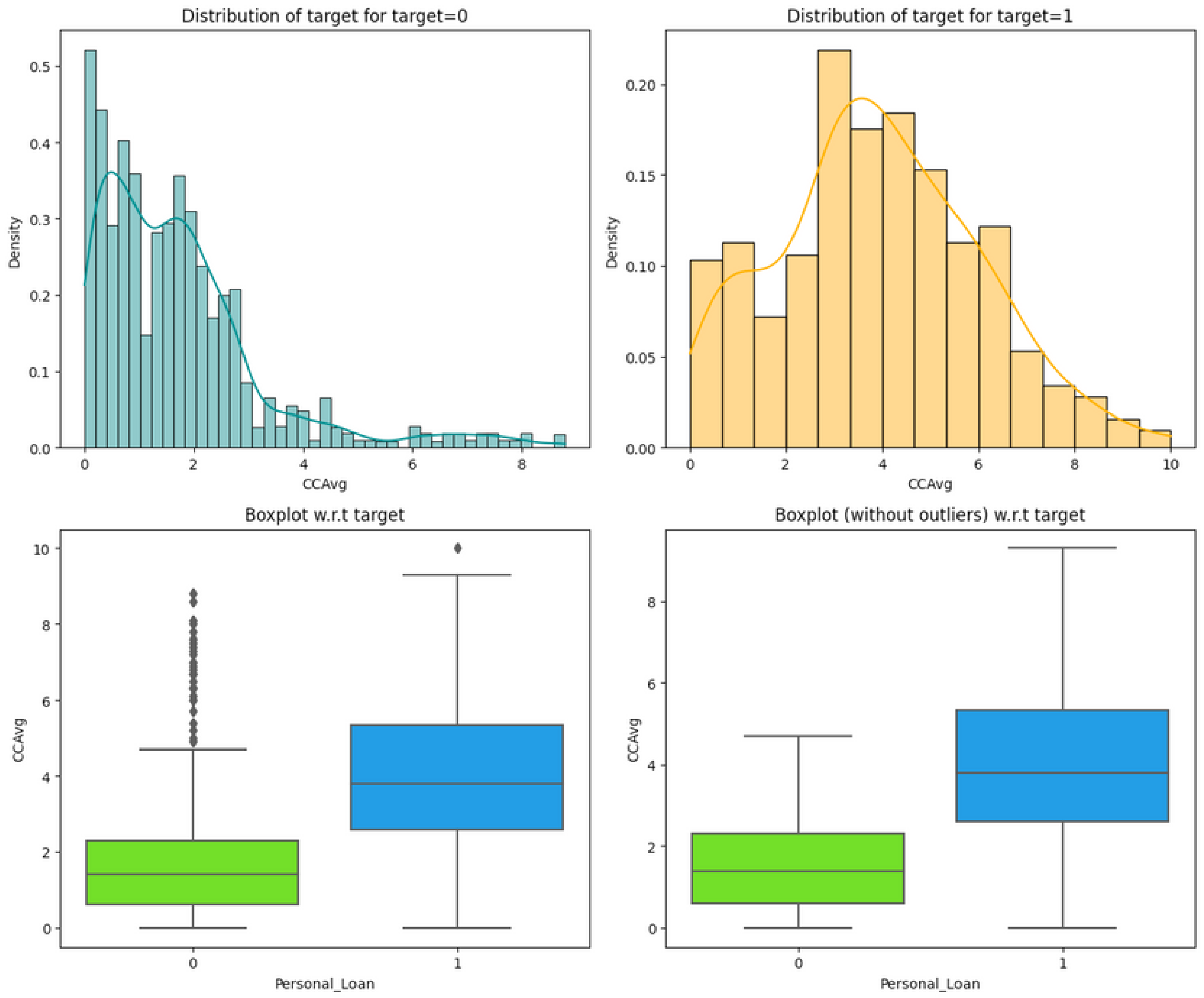
Income and CCAvg impact on Personal Loan purchase interest

- We can observe that the customers that rejected the Personal Loan have the highest density in the low income side and in the low CCAvg payments side. eople that rejected the offer and have a large income and large CCAvg are treated as outliers
- On the portion of the customers that accepted the offer, we can observe that are the customers that have high income but low CCAvg expenditure.

Income and Personal Loan



CCAvg and Personal Loan



Data Preprocessing

Data preprocessing for modeling

Dropped Experience attribute as it is very correlated to age and has the same distribution.

Education attribute values were replaced as follows:
1: Undergraduate; 2: Graduate; 3: Professional

Education, Personal_Loan, Securities_Account, CD_Account, Online, CreditCard and ZIPCode columns were converted to category data type.

Only Education and Zip Code have values that need to apply dummies, the rest are boolean type treated as categorical.

Data Preprocessing

Data preprocessing for modeling (cont.)

Set the independent variable "X" to all columns except for Personal Loan and Experience.

Set the dependent variable "y" to independent.

We split the data in a 70% for training and 30% for testing in our model building.
Resulting in 3500 rows for training and 1500 for testing.
9.45% of positive responses for Personal_Loan on training set
9.93% of positive responses for Personal_Loan on testing set

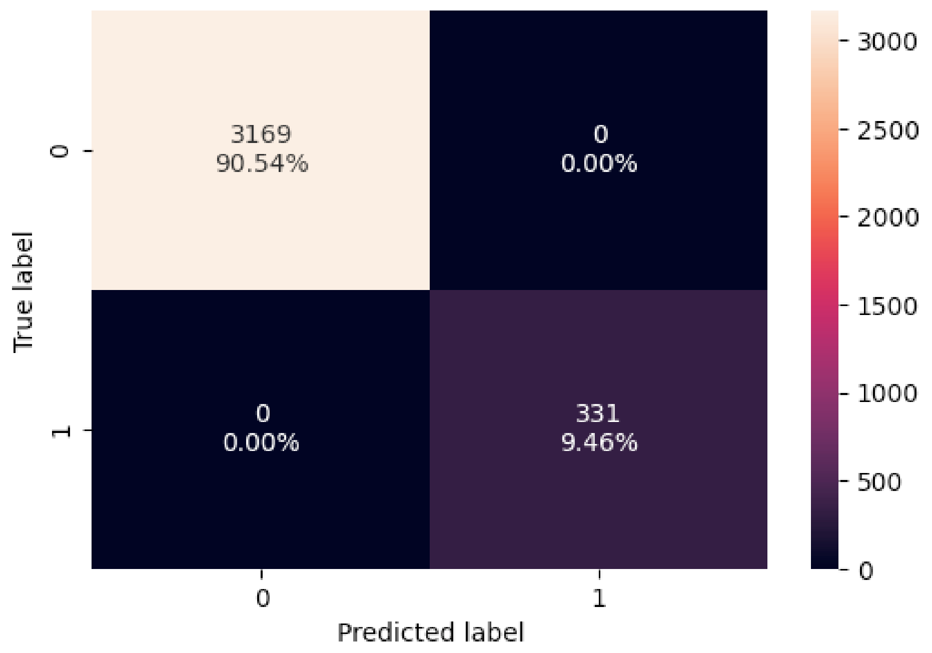
Model Building

Complete Decision Tree Model

- We build a full complete decision tree with the only criterion parameter set to 'gini' so we can explore the complete decision tree.

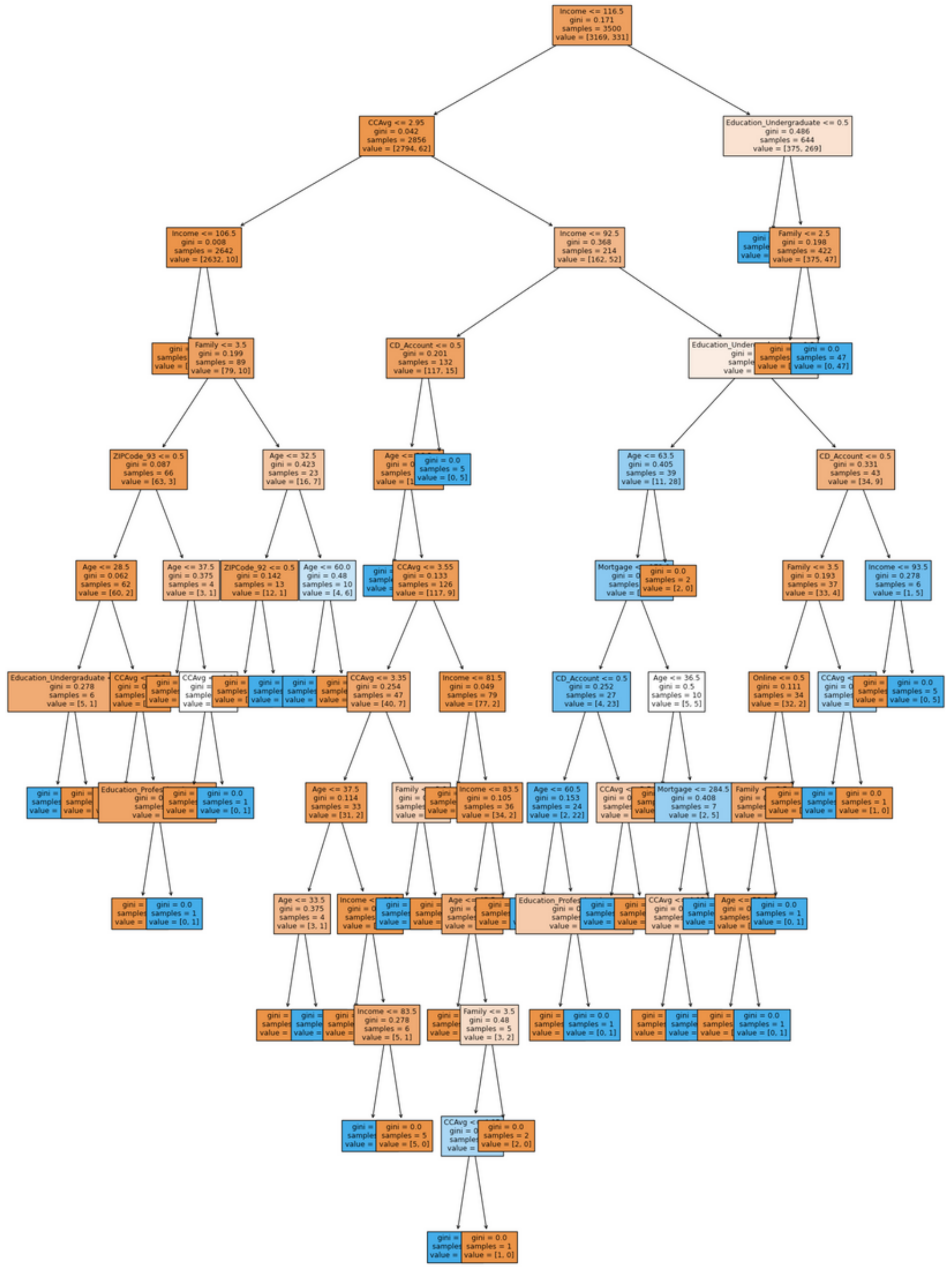
Training

- We can observe that we have 0.00% False Negatives and 0.00% False Positives.
- A predicted 9.46 conversion of True Positives is shown.
- We observe a full grown tree.
- This model is showing overfitting on the performance indicators.



Performance

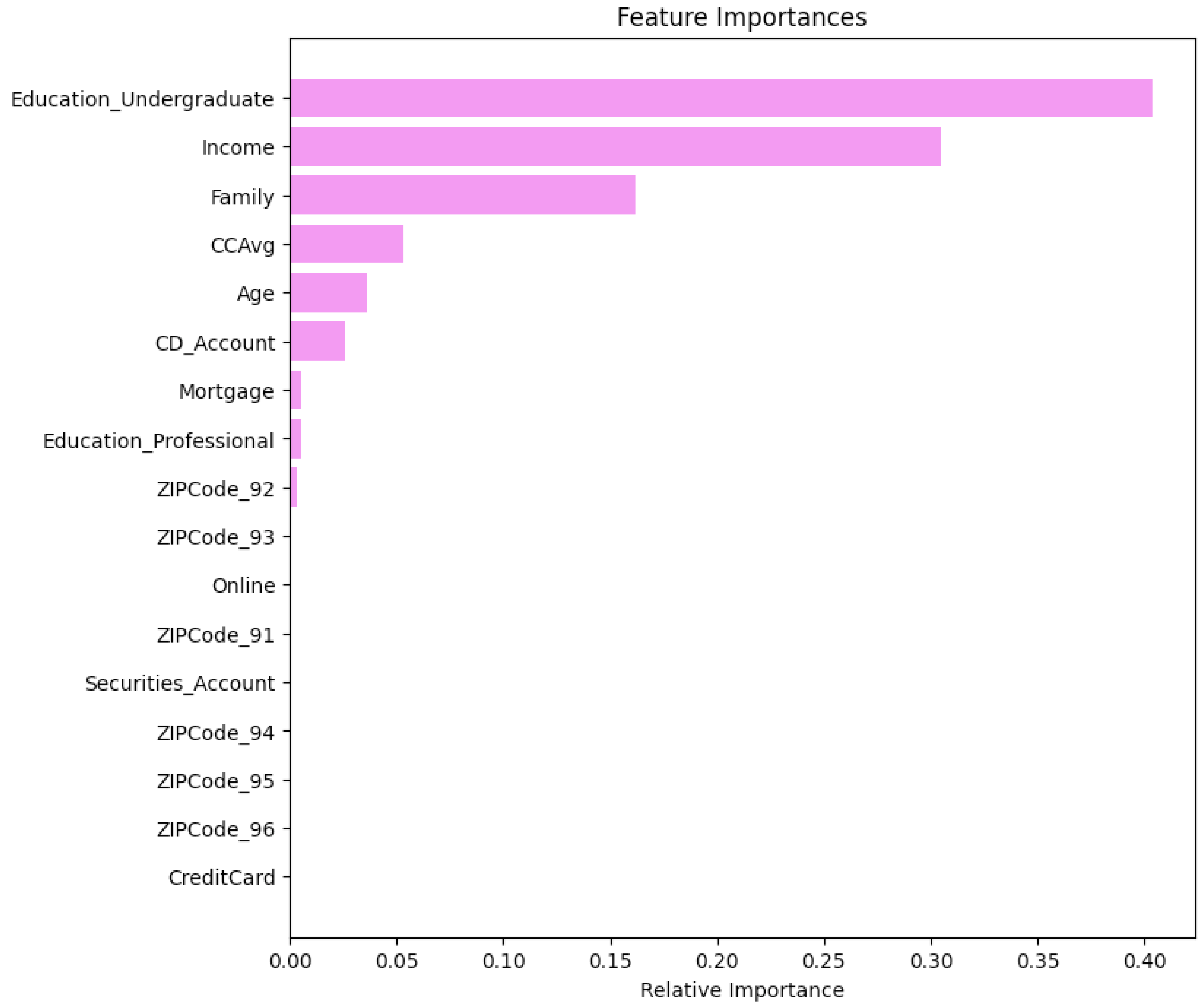
Accuracy	Recall	Precision	F1
1.00	1.00	1.00	1.00



Model Building

Complete Decision Tree Model

- We observe on the Feature importances that Education_Undergraduate, Income, Family, CCAvg, Age, CD_Account, Mortgage, Education_Professional and ZIPCode_92 are the ones with the most Relative Importance.

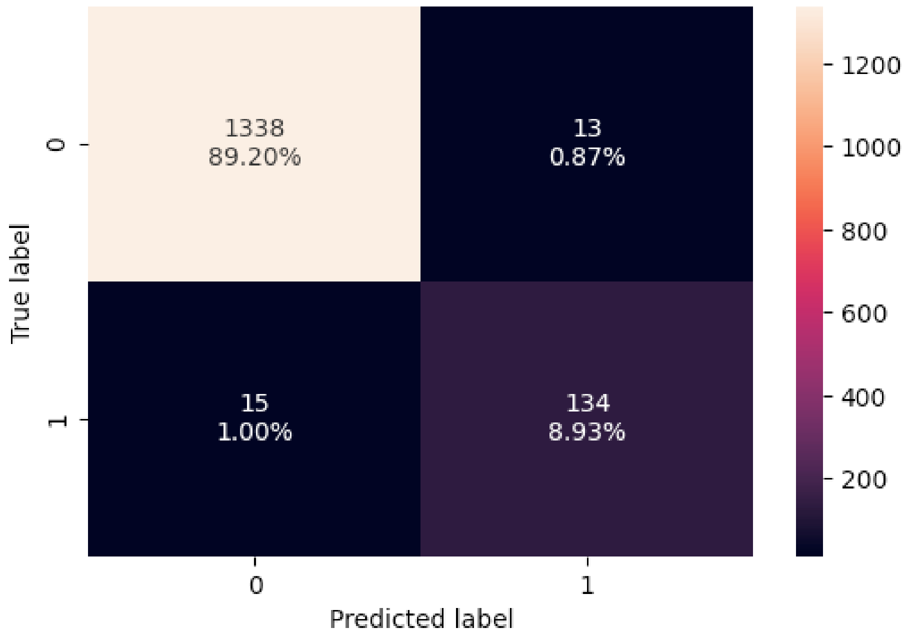


Model Building

Complete Decision Tree Model

Testing

- We can observe that we have 1.00% False Negatives and 0.87% False Positives
- A predicted 8.93 conversion of True Positives is shown.
- On performance we are getting a good recall score and decent levels of Precision, F1 and Accuracy



Performance

Accuracy	Recall	Precision	F1
0.981333	0.899329	0.911565	0.905405

Model Performance Improvement

Pre-Prunning

- We are going to find the best combination of parameters based on recall score from this grid:
- Max depth: 6, 15
- Min samples leaf: 1, 2, 5, 7, 10
- Max leaf nodes: 2, 3, 5, 10

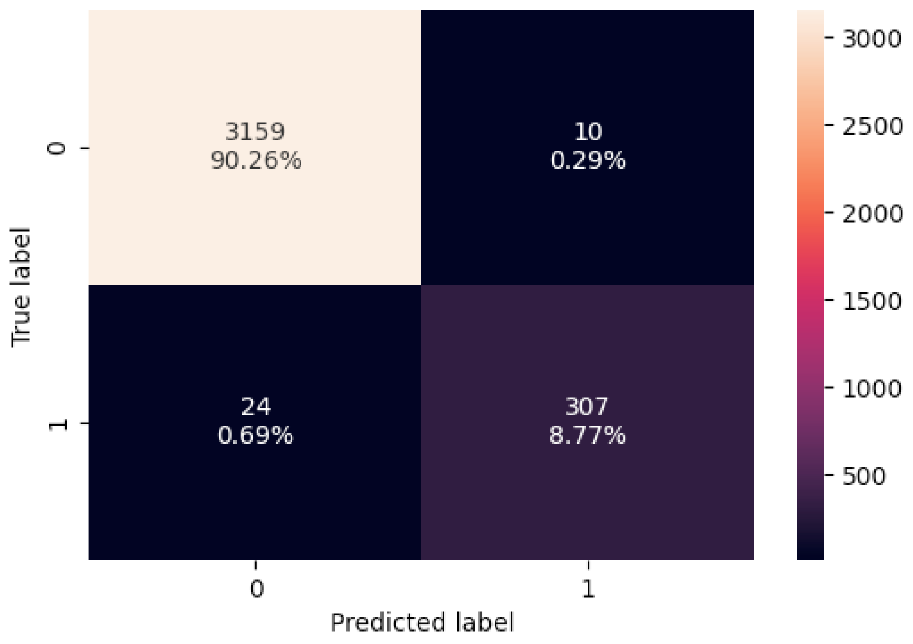
- The parameters chosen based on the best recall score is:
- Max depth: 6, Max leaf nodes: 10.

Model Performance Improvement

Pre-Prunning

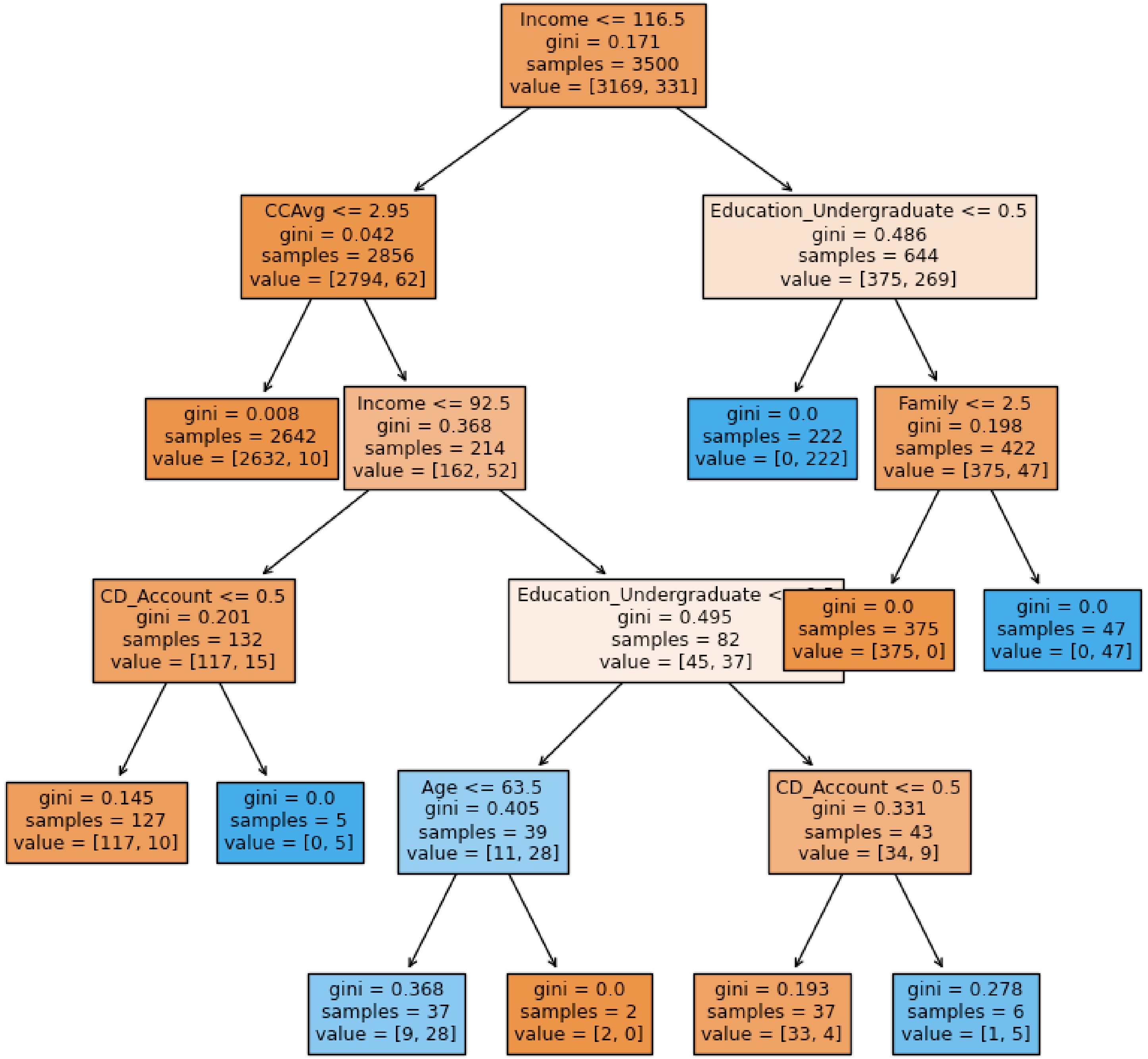
Training

- We can observe that we have 0.69% False Negatives and 0.29% False Positives.
- A predicted 8.77% conversion of True Positives is shown.
- We observe a full grown tree with the pre-pruinning parameters mentioned.
- On performance we observed no more overfitting and a very good recall score.



Performance

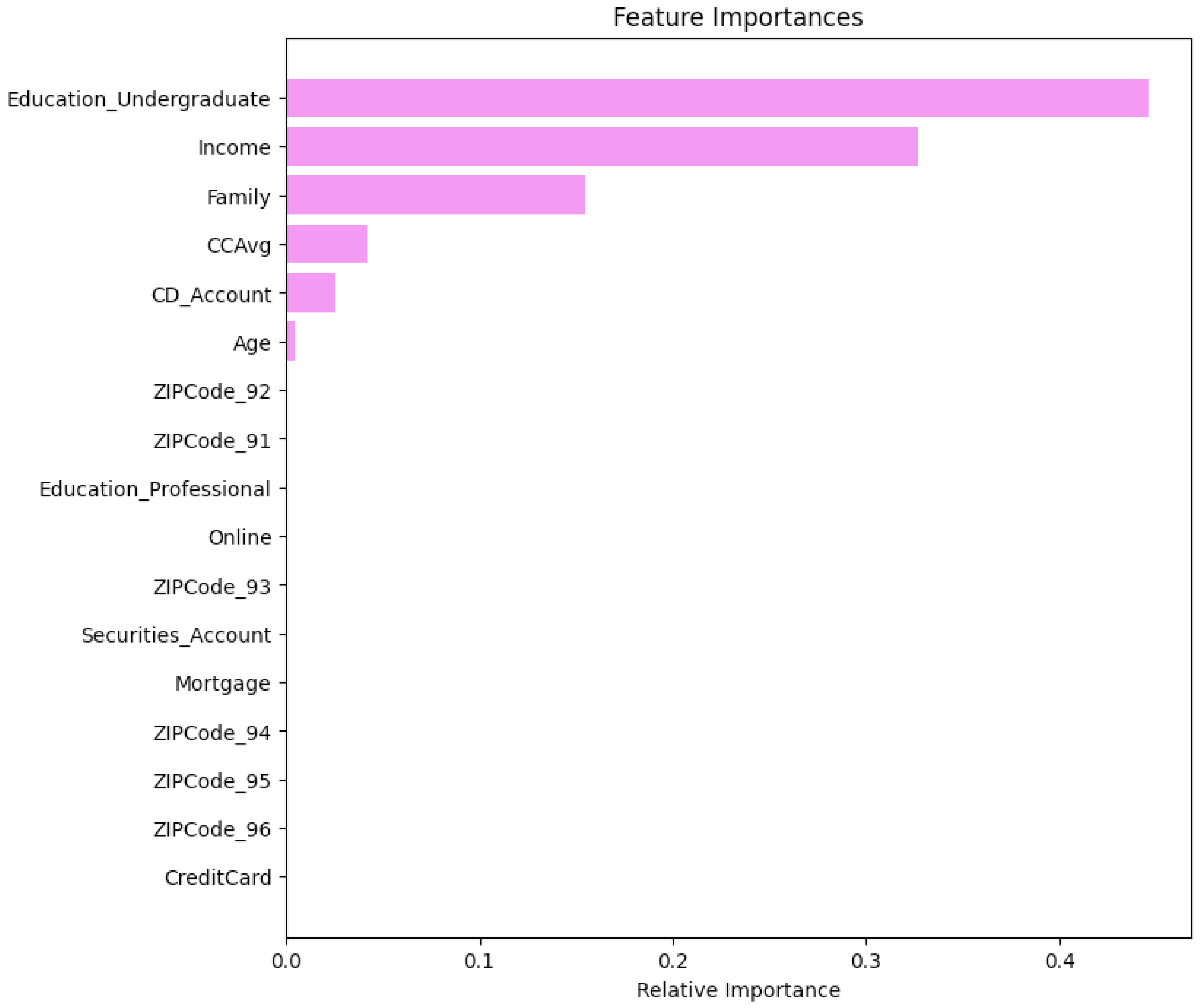
Accuracy	Recall	Precision	F1
0.990286	0.927492	0.968454	0.947531



Model Building

Pre-Prunning

- We observe a reduction on the Feature importances from the complete tree and only Education_Undergraduate, Income, Family, CCAvg, CD_Account, Age, are the ones with the most Relative Importance on this model.

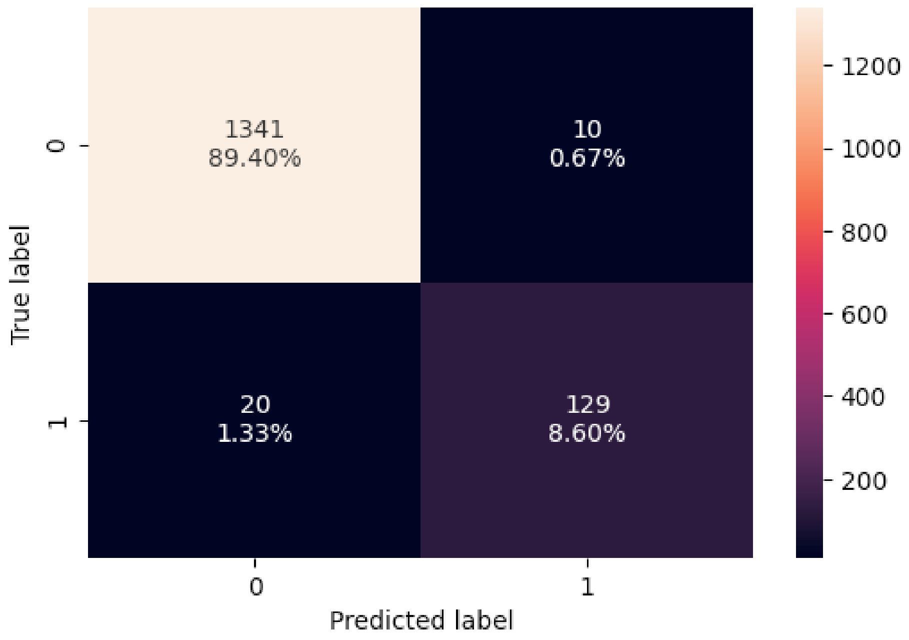


Model Building

Pre-Prunning

Testing

- We can observe that we have 1.33% False Negatives and 0.67% False Positives
- A predicted 8.60% conversion of True Positives is shown.
- On performance we observe a recall score relatively close to the score shown on the training data. but most performance indicators are low.



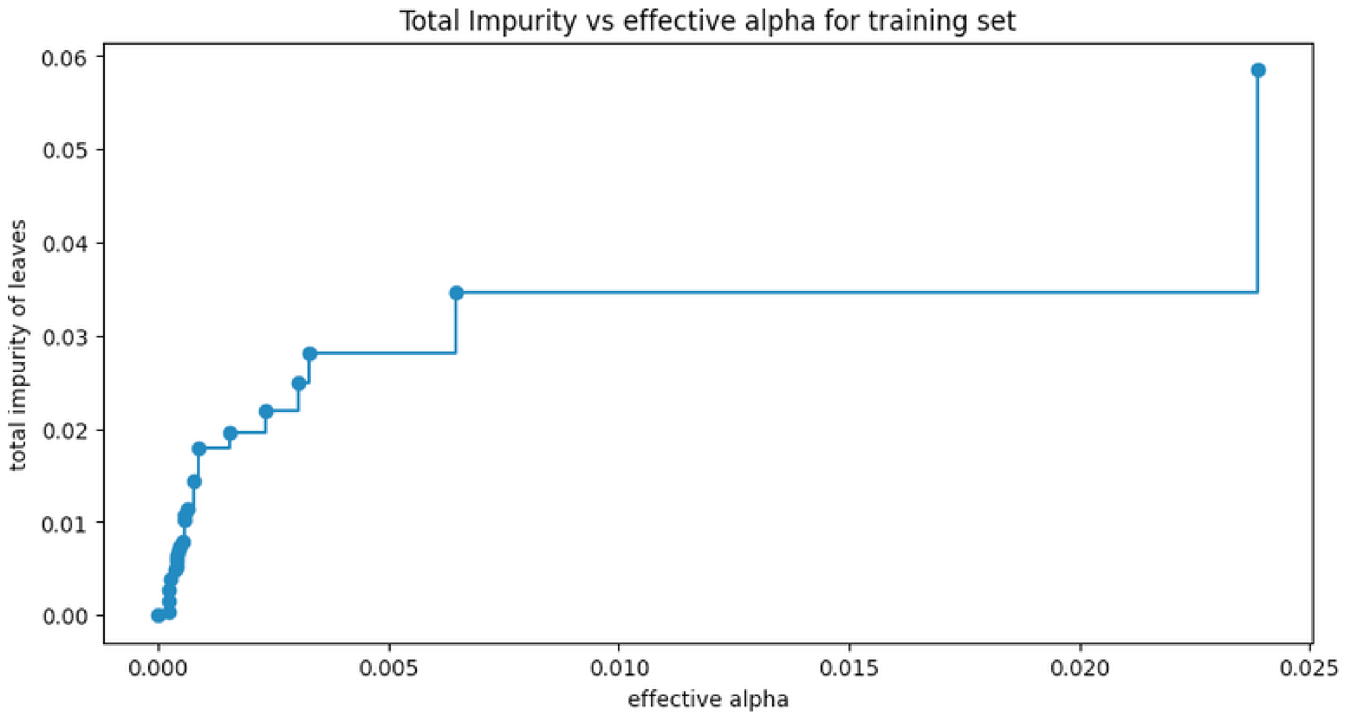
Performance

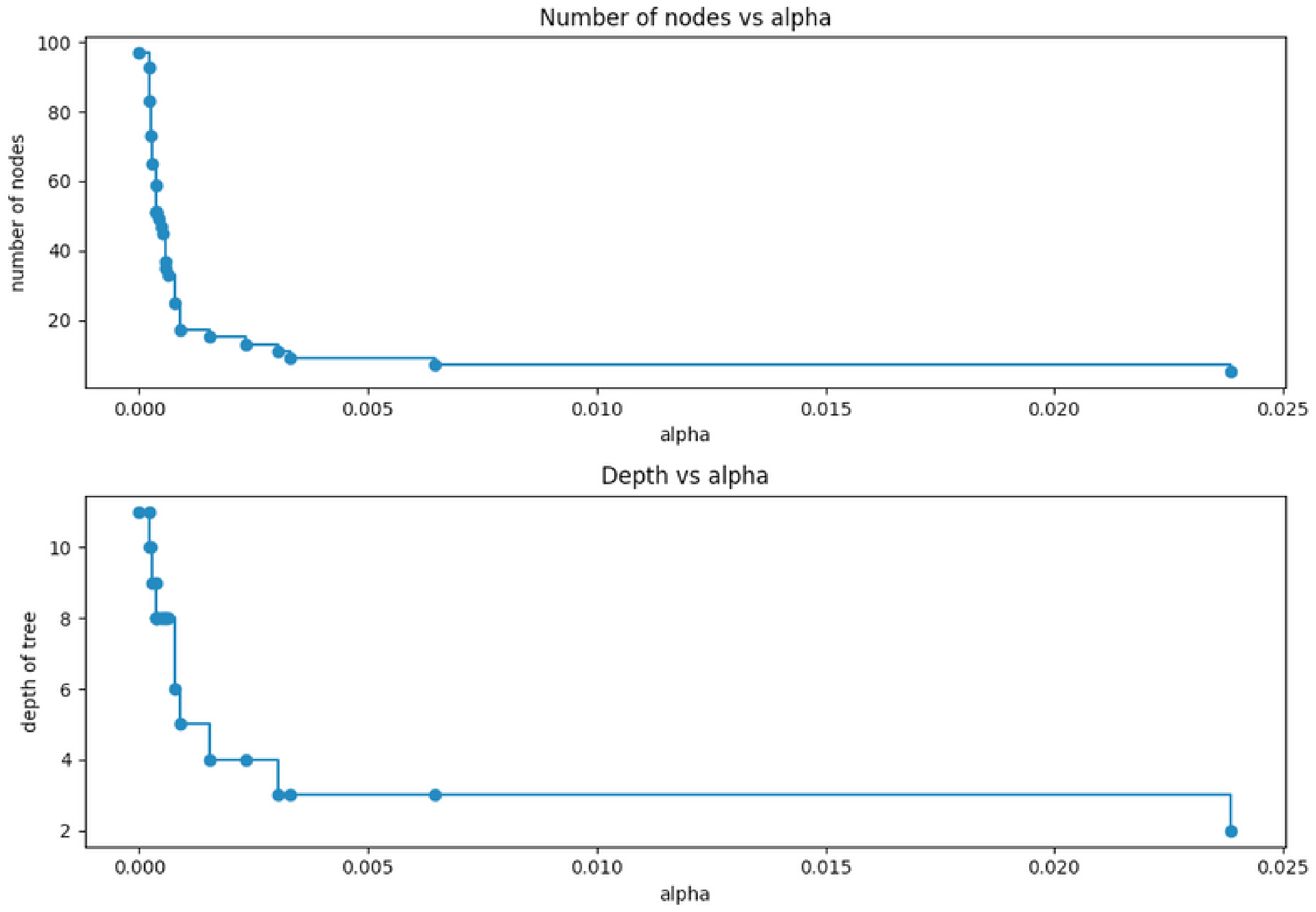
Accuracy	Recall	Precision	F1
0.98	0.865772	0.928058	0.895833

Model Performance Improvement

Cost-Complexity

- We are going to find the best ccp_alpha that balances the fit with the training data and the complexity of the tree.
- We can observe that the impurity trade off is best just after the 0.005 CCP_alpha value. After that the jump of impurity increase is to steep.
- Also in the number of nodes we see that no relevant decrease in nodes happen just after 0.005 ccp_alpha value and no significant depth decrease ofter that value.

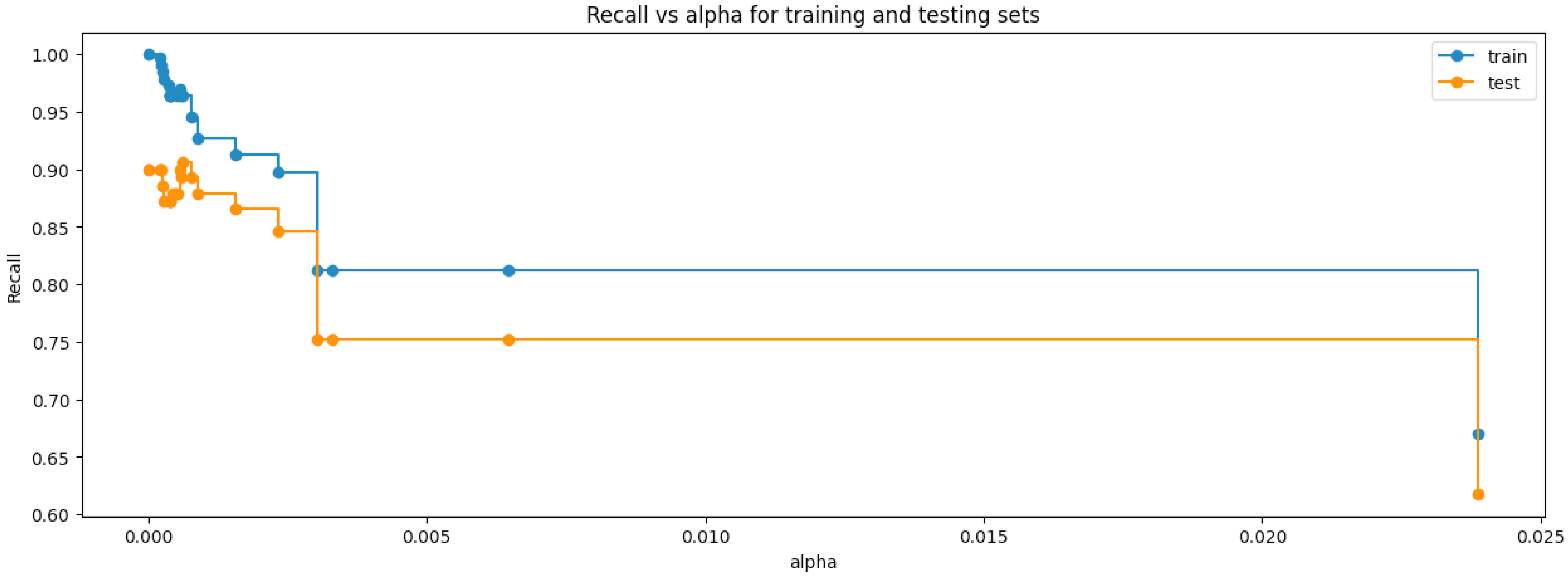




Model Performance Improvement

Cost-Complexity

- We find that across ccp_alpha values from 0.004 to 0.025 there's no significant increase in recall score until ccp_alpha gets really small. We have a value just after 0.005
- We get the best ccp_alpha of: 0.0006209286209286216 and we build our model for post-prunning with the class_weight : 0: 0.15, 1: 0.85

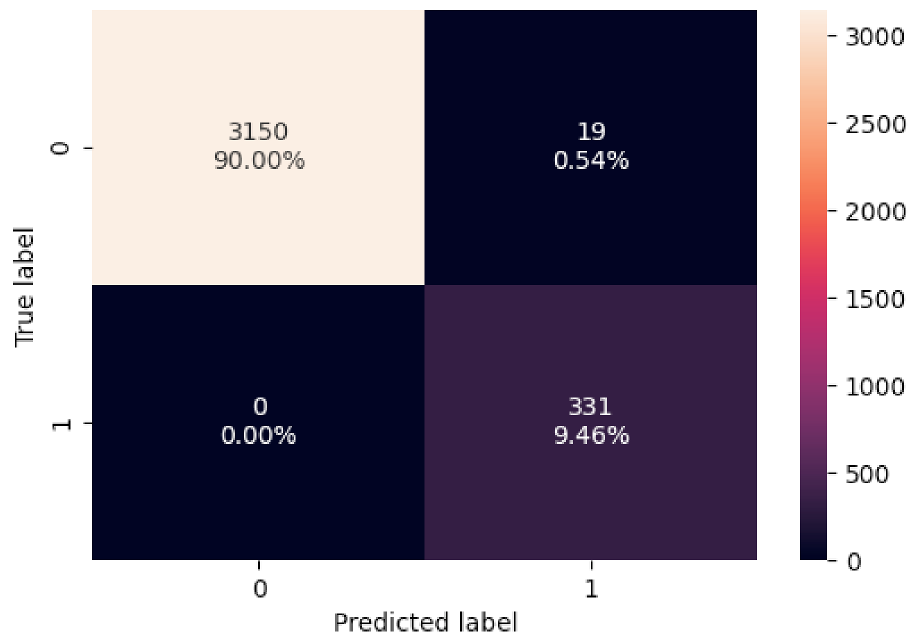


Model Performance Improvement

Post-Prunning

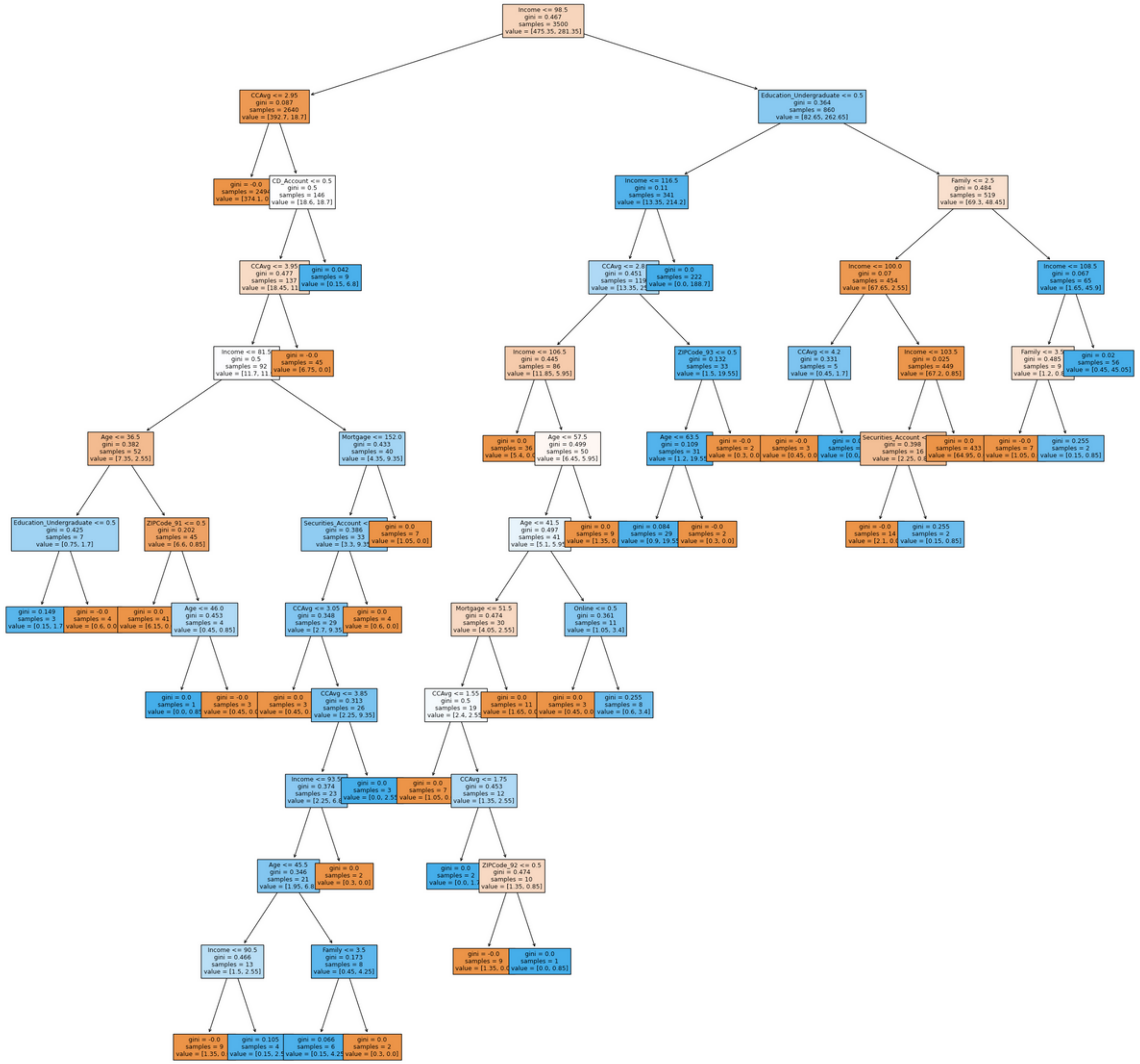
Training

- We can observe that we have 0.00% False Negatives and 0.54% False Positives.
- A predicted 9.46% conversion of True Positives is shown.
- We observe the tree with the post-prunning parameters mentioned.
- On performance we observe that we are close to overfitting.



Performance

Accuracy	Recall	Precision	F1
0.994571	1.0	0.945714	0.9721

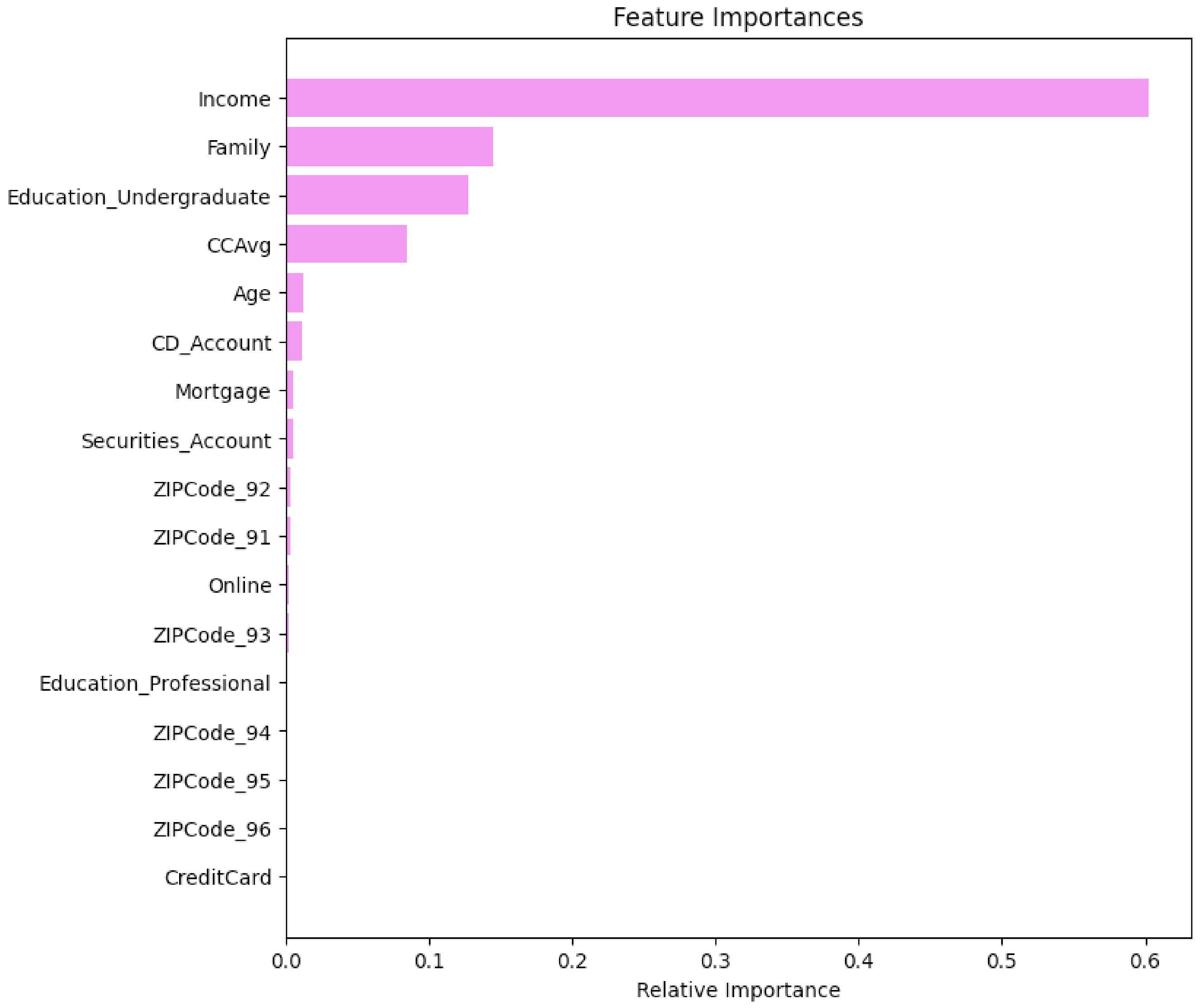


Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

Model Building

Post-Prunning

- We observe an increase on the Feature importances from the pre-prunned tree.

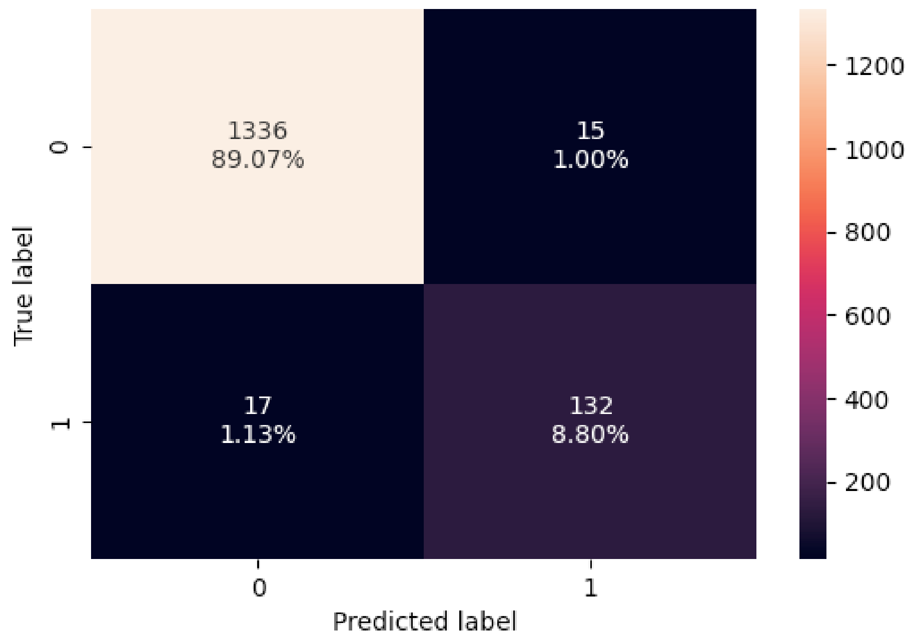


Model Building

Post-Prunning

Testing

- We can observe that we have 1.13% False Negatives and 1.00% False Positives
- A predicted 8.80% conversion of True Positives is shown.
- On performance we observe a good recall score.



Performance

Accuracy	Recall	Precision	F1
0.978667	0.885906	0.897959	0.891892

Model Performance Summary

Model evaluation criterion

As we want to minimize the number of false negatives as we want top avoid opportunity cost for the bank

Overview of the final decision tree model and its parameters

Th final decision tree is the post-pruned model as doesn't show any overfitting on training and a good recall score and good conversion on testing.

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

ccp_alpha of: 0.0006209286209286216 and we build our model for post-prunning with the class_weight : 0: 0.15, 1: 0.85

Model Performance Summary

Summary of key performance metrics for training and test data of all the models in tabular format for comparison

Training performance comparison			
	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	1	0.990286	0.994571
Recall	1	0.927492	1
Precision	1	0.968454	0.945714
F1	1	0.947531	0.9721

Testing performance comparison:			
	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.981333	0.98	0.978667
Recall	0.899329	0.865772	0.885906
Precision	0.911565	0.928058	0.897959
F1	0.905405	0.895833	0.891892

APPENDIX

Data Background and Contents

Data background

The dataset contains 5000 customer data from AllLife Bank from a previous Personal Loan campaign with a conversion of over 9%.

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)

Happy Learning !