AllLife Bank

Personal Loan Campaign

AIML course - Decision Tree Modeling

Azin Faghihi
Role: Data Scientist
January 2025

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

Contents / Agenda

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

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

Executive Summary

- Our modeling parameter, personal_loan is an ***imbalanced class***.
- By conducting ***exploratory data analysis*** (univariate and bivariate) we explore relationships between all the variables (categorical & numerical)
    - We conclude that personal_loan is affected by income, education, family, credit card spending, and age.
- Due to the nature of the problem at hand, we chose **F1** as the score to maximize.
- We conduct ***pre- and post-pruning*** to try and reduce overfitting.
- We explore the relationships between the ***hyperparameters*** and model performance.
- The ***important features*** observed in our trained models match those of our observation in exploratory data analysis.

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

Business Problem Overview and Solution Approach

- ***Problem Overview:*** AllLife Banks is interested in growing its borrowers (asset customers) and the same time retain them as liability customers (depositors).
- ***Objective:*** Based on the data gathered from the last year's personal loan campaign, we would like to predict whether a liability customer will purchase personal loans, identify the key customer attributes influencing these purchases, and determine which customer segment to prioritize for targeting.
- ***Solution Approach:*** Using exploratory data analysis and machine learning modeling (decision tree), we identify the key features that influence whether a customer buys or does not buy a personal loan.

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

Data

- There are 5,000 (~5K) *rows* and 14 *columns* in the dataset.
- The *memory usage* is approximately 547.0 KB.
- There are no null values.
- There are no duplicated rows.

Memory Usage	547.0 KB
#	
Rows	5000
Columns	14
Null Values	0
Duplicated Rows	0

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

Data Dictionary

Column	Data Type	Description	# unique
Age	int64	Customer's age in completed years	45
Experience	int64	Number of years of professional experience	44
Income	int64	Annual income of the customer (in thousand dollars)	162
Family	category	The family size of the customer	4
CCAvg	float64	Average spending on credit cards per month (in thousand dollars)	108
Education	category	Education level (1: Undergrad; 2: Graduate; 3: Advanced/Professional)	3
Mortgage	int64	Value of house mortgage if any (in thousand dollars)	347
Personal_Loan	category	Did this customer accept the personal loan offered in the last campaign? (0: No, 1: Yes)	2
Securities_Account	category	Does the customer have securities account with the bank? (0: No, 1: Yes)	2
CD_Account	category	Does the customer have a certificate of deposit (CD) account with the bank? (0: No, 1: Yes)	2
Online	category	Do customers use internet banking facilities? (0: No, 1: Yes)	2
CreditCard	category	Does the customer use a credit card issued by any other Bank (excluding AllLife Bank)? (0: No, 1: Yes)	2
ZIPCode (SCF)	category	(sectional center facility) is the rightmost two digits of the ZIPCode	7

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

Data

- The following tables show the summary information of our variables.
  - ***Categorical***: unique counts, most common value and its corresponding frequency
  - ***Numerical***: mean, median, standard deviation, minimum, maximum, outlier counts, ...

Object/Categorical Column	unique	top	freq
Family	4	1	1472
Education	3	1	2096
Personal_Loan	2	0	4520
Securities_Account	2	0	4478
CD_Account	2	0	4698
Online	2	1	2984
CreditCard	2	0	3530
ZIPCode (SCF)	7	94	1472

Numerical Column	mean	std	min	25%	50%	75%	max	IQR	# Outliers (Upper)	# Outliers (Lower)	# Outliers	Outliers %
Age	45.3	11.5	23	35	45	55	67	20	0	0	0	0
Experience	20.1	11.4	0	10	20	30	43	20	0	0	0	0
Income	73.8	46	8	39	64	98	224	59	96	0	96	1.9
CCAvg	1.9	1.7	0	0.7	1.5	2.5	10	1.8	340	0	340	6.8
Mortgage	56.5	101.7	0	0	0	101	635	101	291	0	291	5.8

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

EDA Results

- The majority of our customers have neither **CD accounts** nor **securities accounts.**
- Most do not have other credit cards.
- More than half the customers have **graduate or advanced/professional degrees**.
- About 70% of the the customers **do not live alone**.
- More than half the users use **online** banking facilities.
- About 10% of the customers accepted the **personal loan** campaign last year.
- About 30% of the customers live in the SCF/**ZIP Code** (Sectional Center Facility) 94.
- The average **age** of the customers is 45 and on average they have 20 years of **experience**.
- Their average **income** is 74K and the average **mortgage** of those customers who have mortgages is 184K.
- The **average credit card spending** per month is 2K.

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

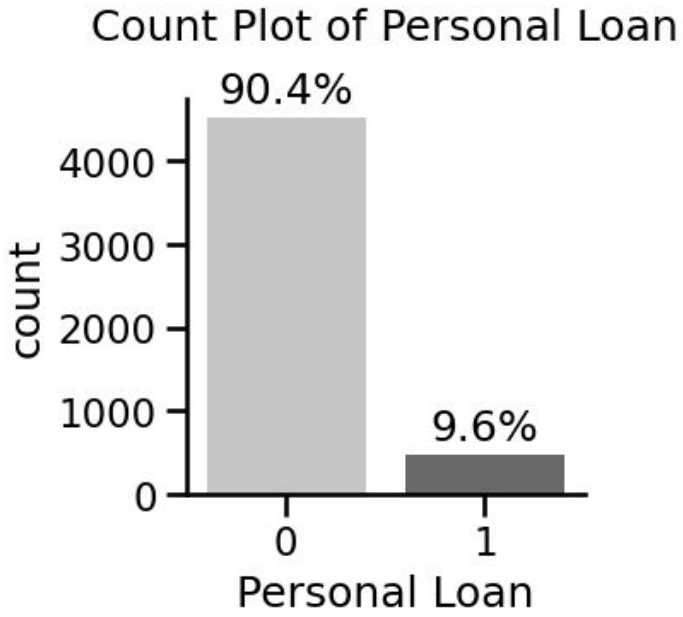
EDA Results

- **Age** and **experience** are both symmetrically distributed and are highly correlated.
- **Income**, **average spending on credit cards** and **mortgage** are also positively correlated.
- There's also a relation between **family** size/**educational** level/holding **CD accounts** on **personal loan** acceptance.
- The customers who buy the **personal loan** seem to have a higher **income/mortgage/credit card spending** on average.
- The median age/experience/income/mortgage of the people who live in SCF 96 seem to be lower than the other areas.
- The customers who have **CD accounts** on average have higher **mortgages**.

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

EDA - Univariate - *Personal Loan*

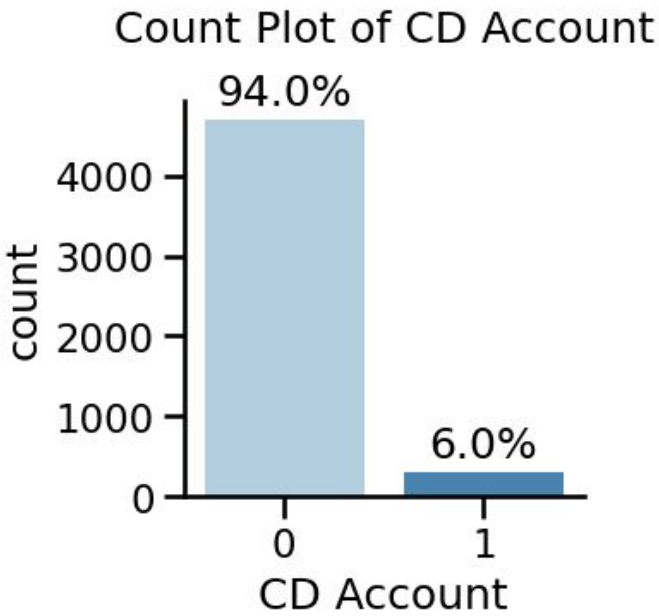
- The majority (~90%) of the customers ***did not accept*** the personal loan offered in the last campaign.



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

EDA - Univariate - *CD Account*

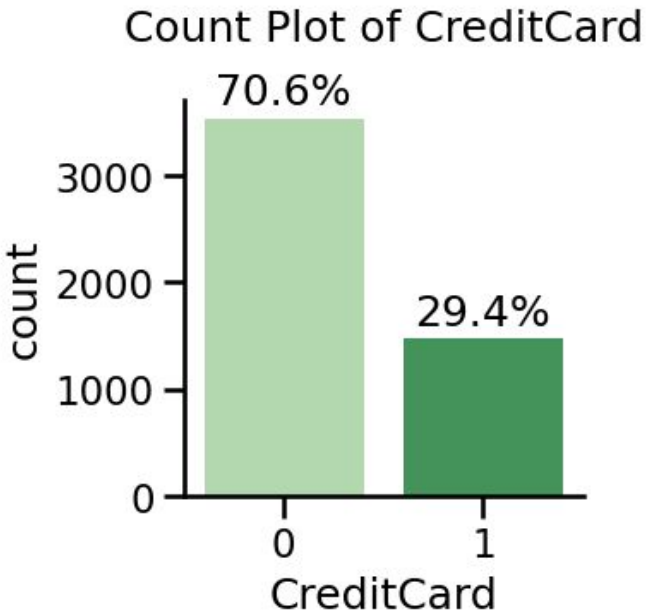
- The majority of the customers (94%) ***do not have*** a certificate of deposit account with the bank.



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

EDA - Univariate - *Credit Card*

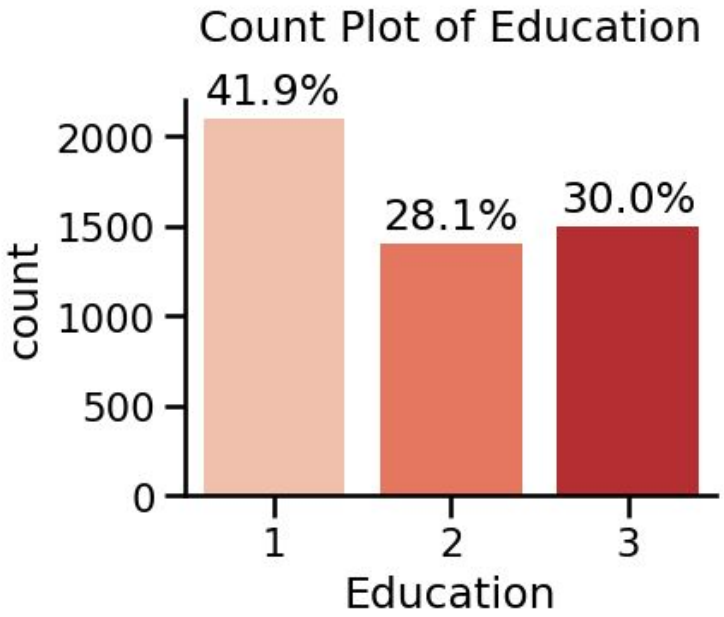
- The majority of the customers (~71%) ***do not have*** a credit card issued by any other bank.



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

EDA - Univariate - *Education*

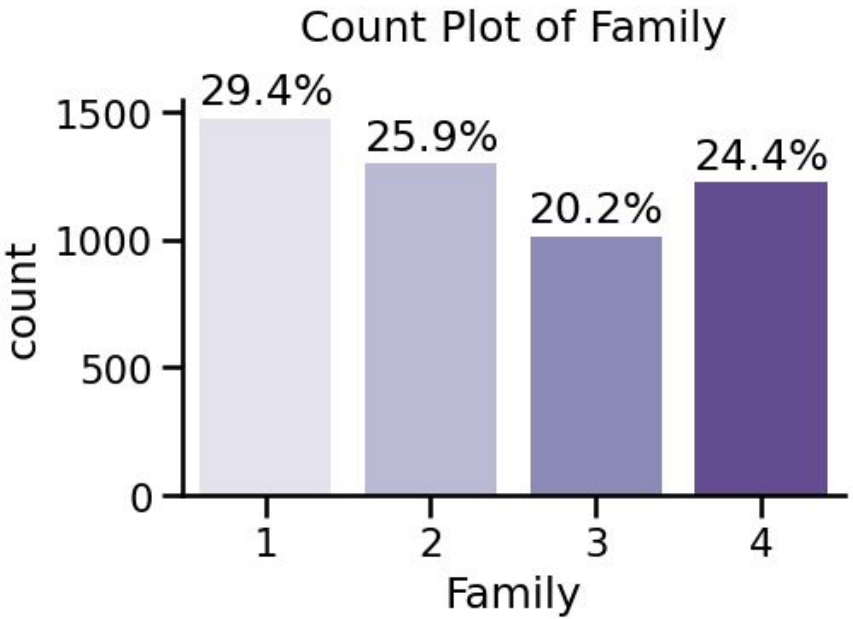
- The education level of most customers is ***undergraduate***.



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

EDA - Univariate - *Family*

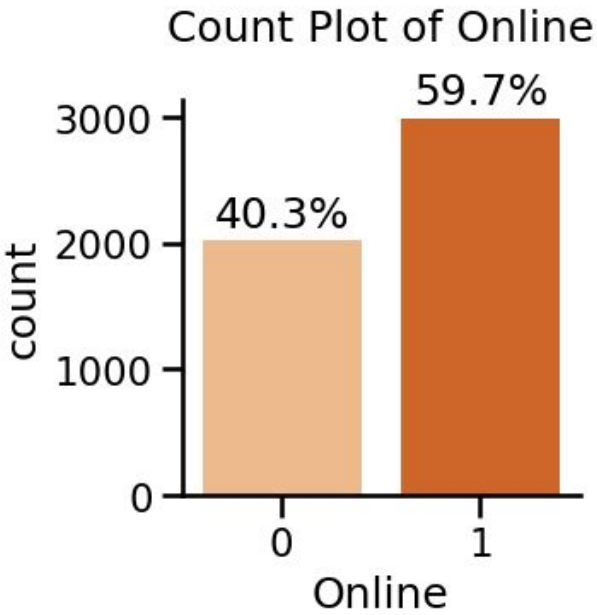
- The family size of most customers is ***1***.



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

EDA - Univariate - *online*

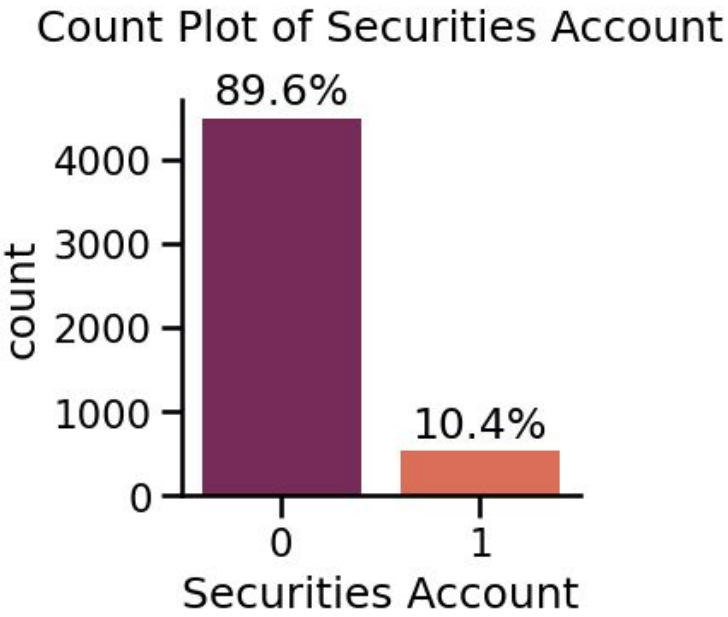
- More than half (~60 %) of the customers ***use*** internet banking facilities.



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

EDA - Univariate - *Securities Account*

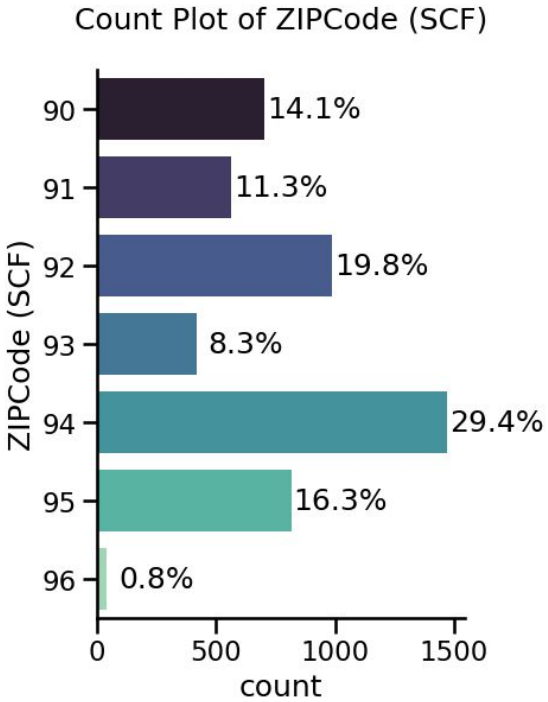
- The majority (~90 %) of the customers ***do not have*** securities account with the bank.



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

EDA - Univariate - *ZIP Code (SCF)*

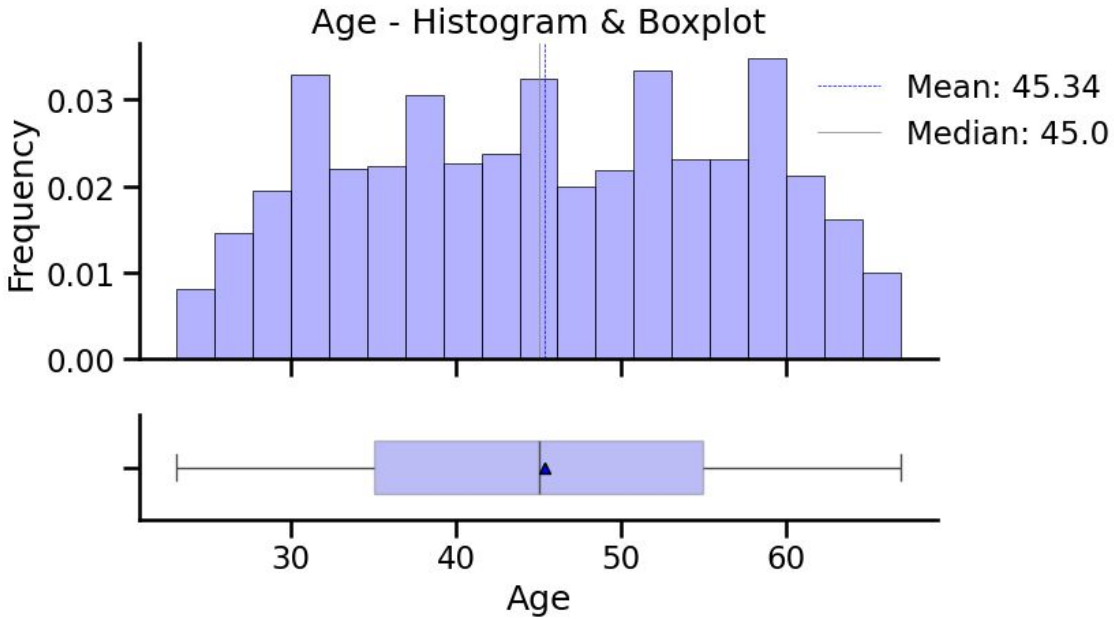
- The most common sectional center facility (home address) among the customers is 94, followed by 92.



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

EDA - Univariate - *Age*

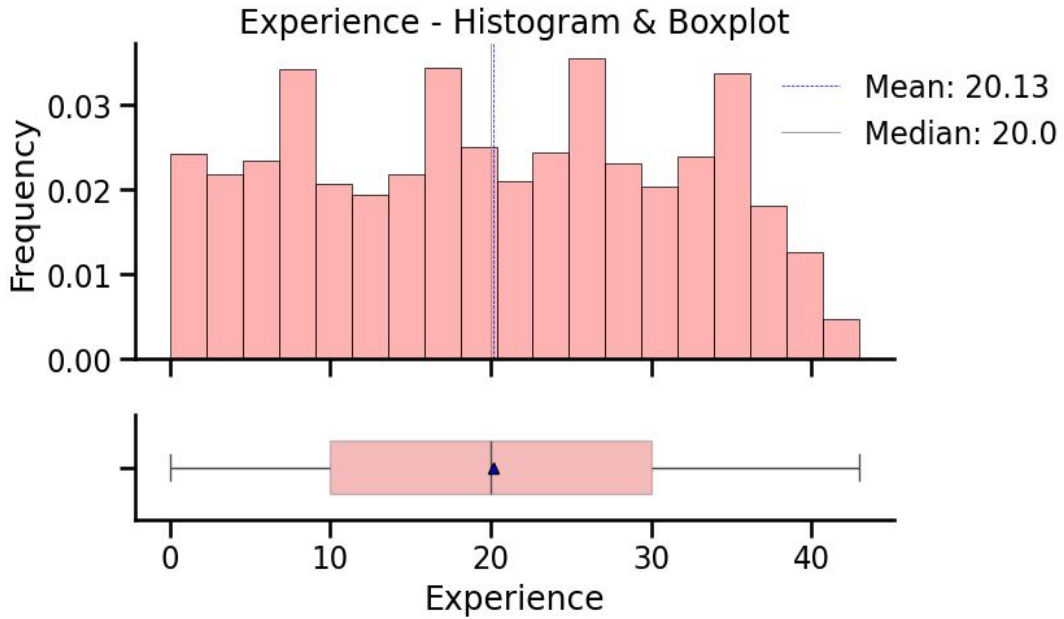
- The age of the customers is almost symmetrically distributed around the mean of 45 years. The youngest customer is 23 years old, and the oldest customer is 67 years old.



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

EDA - Univariate - *Experience*

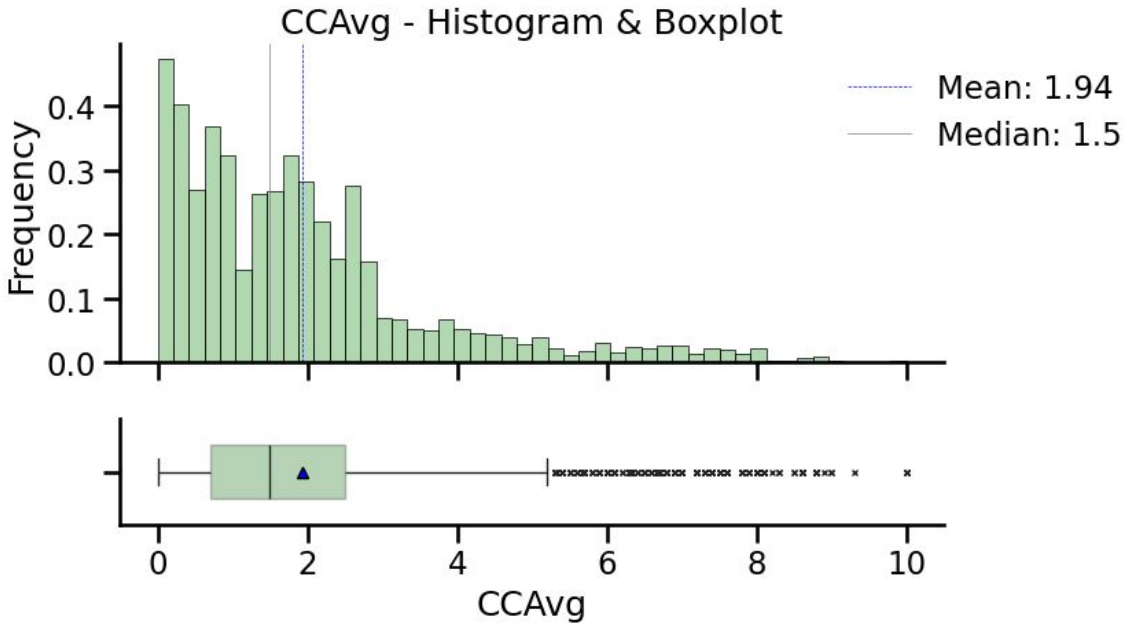
- The number of years of professional experience of the customers is almost symmetrically distributed around the mean of 20 years. The most experienced customer has 43 years of experience.



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

EDA - Univariate - *CC Avg*

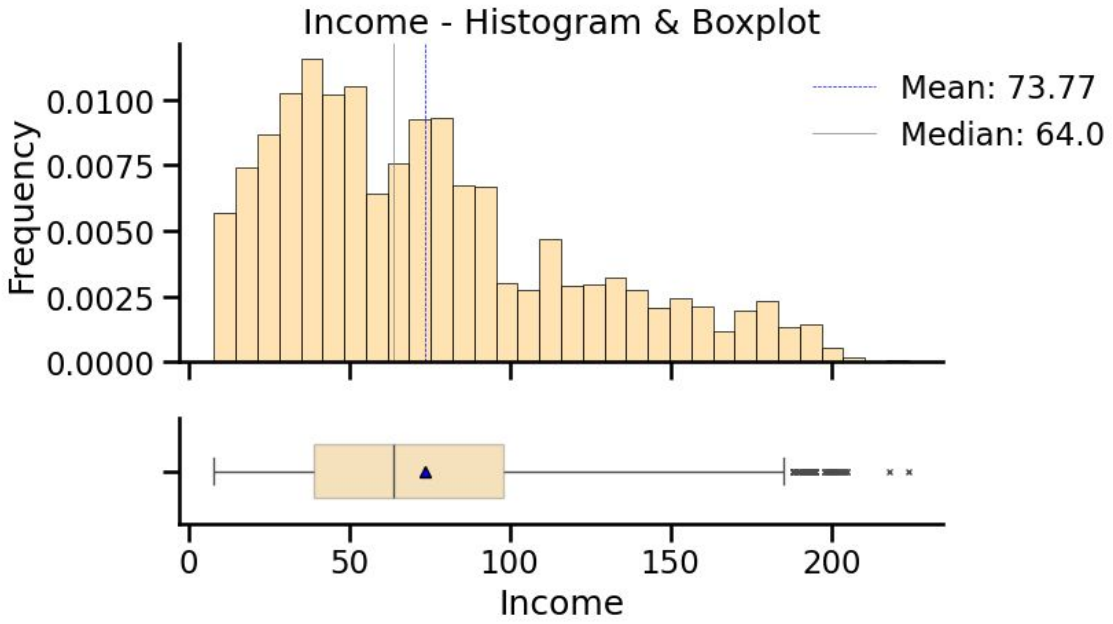
- The average spending on credit cards per month is highly positively skewed (right-skewed). The median is around $1,500. Around 6% of these values are outliers.



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

EDA - Univariate - *Income*

- The annual income of the customers is highly positively skewed (right-skewed) with the mean of ~ 74K and median of 64K. The minimum income is 8K and the maximum income is 224K. Around 2% of these values are outliers.

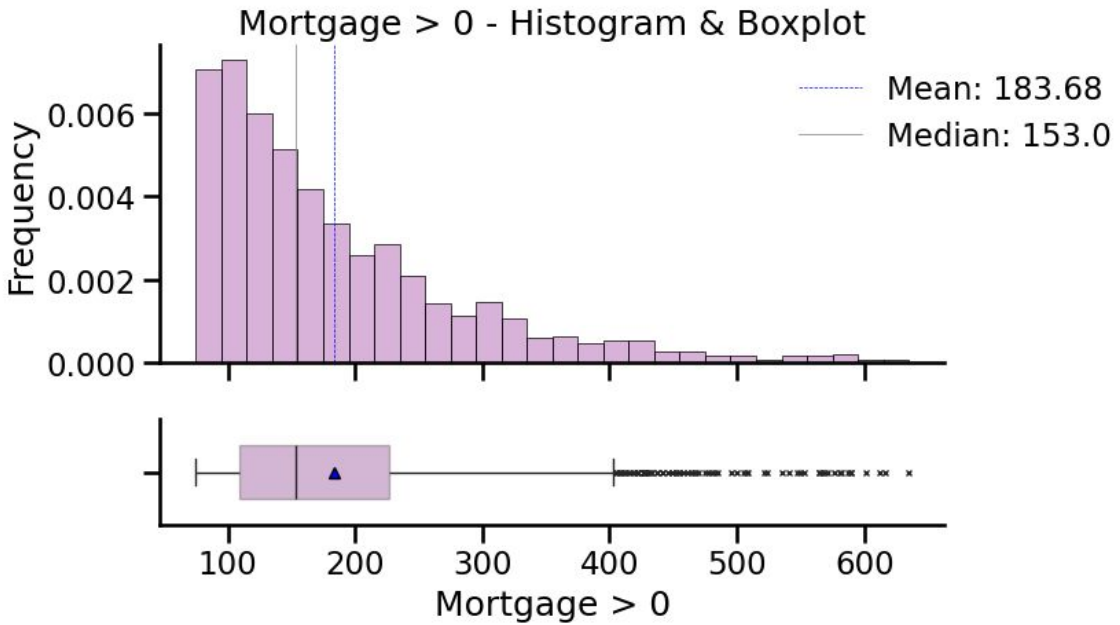


Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA - Univariate - *Mortgage*

- Around 70% of the customers do not have mortgage. Amongst those who do, it is highly positively skewed with an average of ~ 184K and median of 153K. Around 5% of these values are outliers.

No Mortgage	proportion
True	69%
False	31%

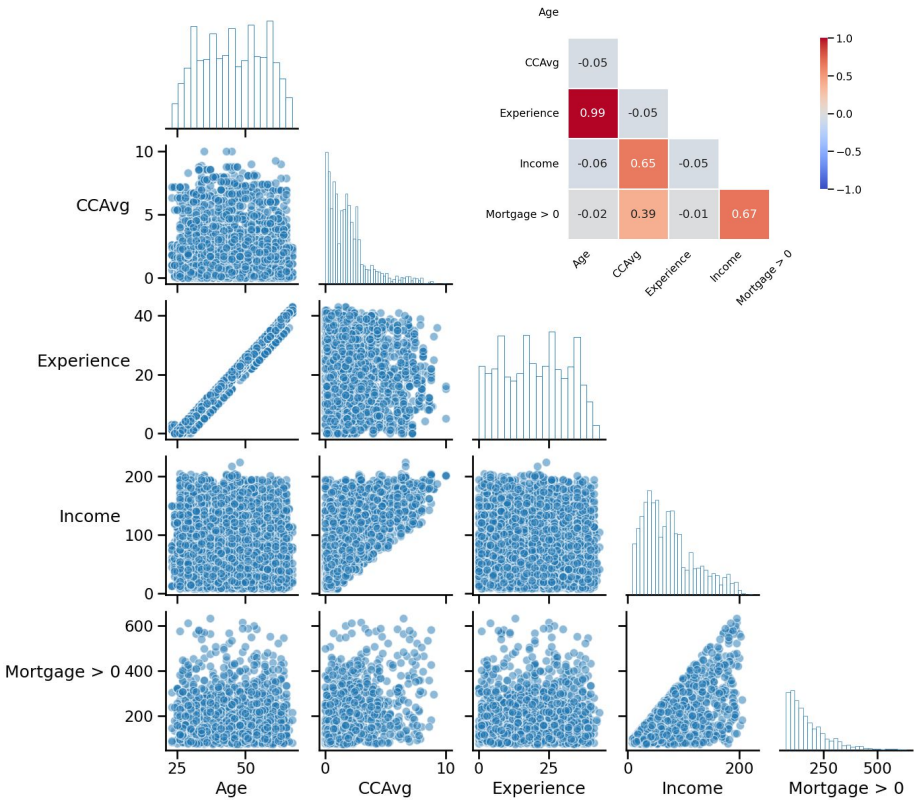


Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA - Bivariate - *Numerical Variables*

- ***Age & Experience*** are highly correlated. As seen in the pairplot, there's a linear relationship between them. Hence, the Experience column should be dropped before modeling.

Variable 1	Variable 2	Correlation
Age	Experience	0.99
Income	Mortgage > 0	0.67
CCAvg	Income	0.65
CCAvg	Mortgage > 0	0.39

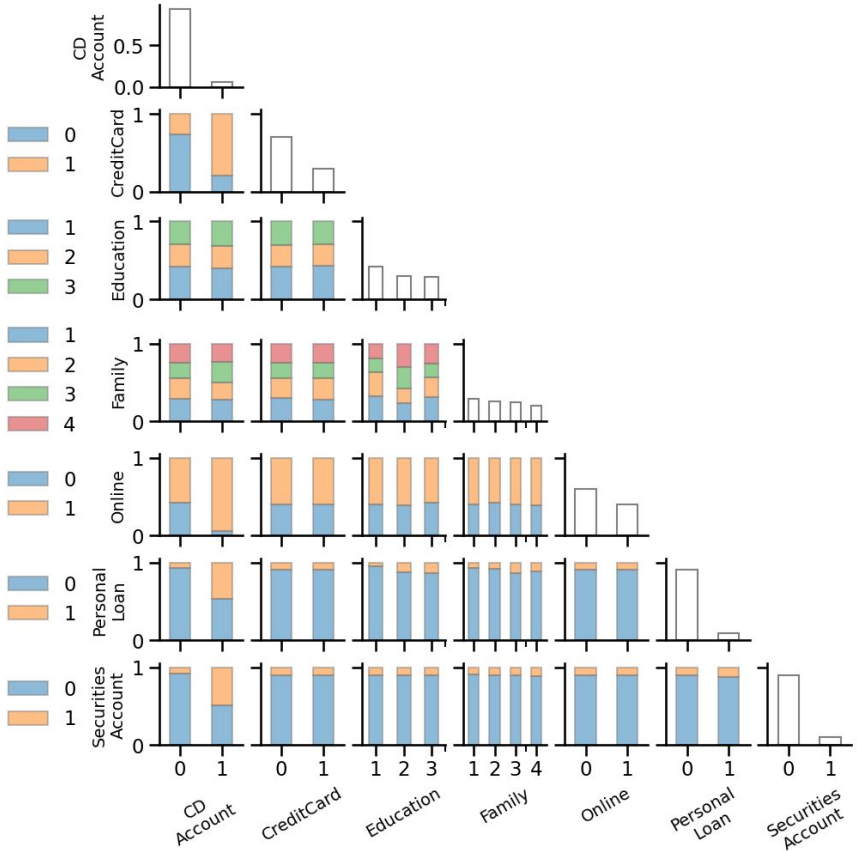


Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

EDA - Bivariate - *Categorical Variables*

- By conducting χ^2 test of independence we observe that the following variables have effects on each other:

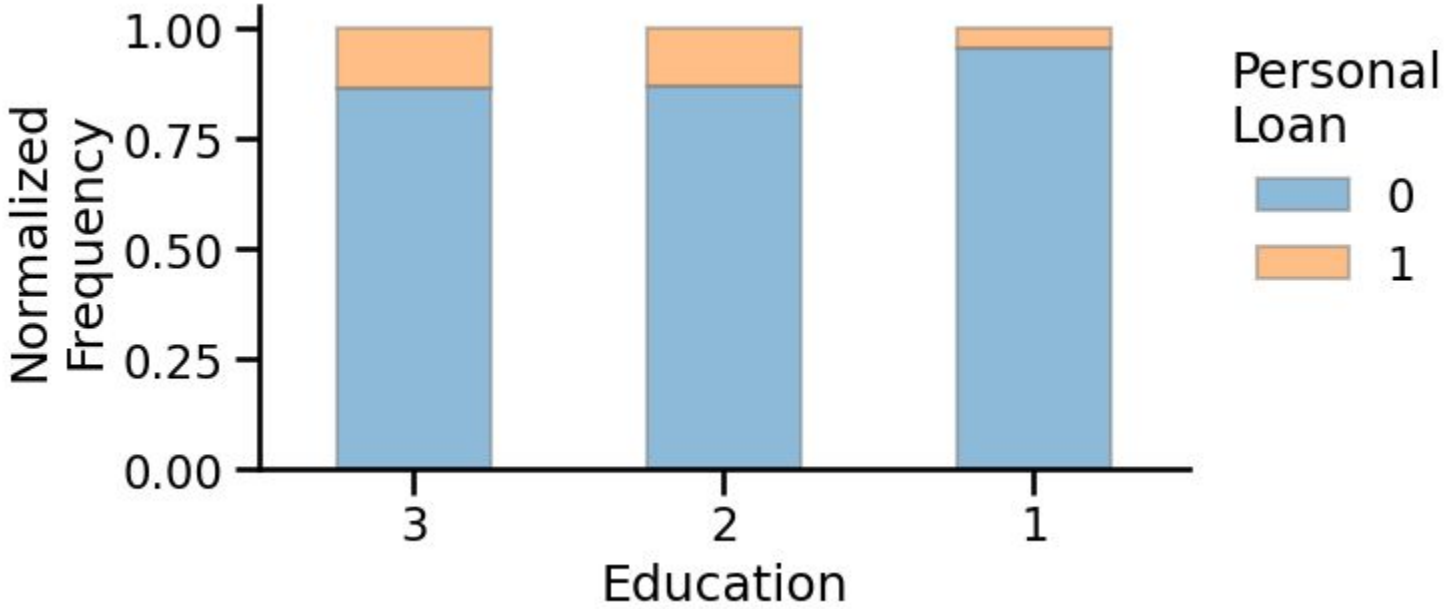
Category 1	Category 2	p-value
CD_Account	Securities_Account	2.3e-110
CD_Account	Personal_Loan	7.4e-110
CD_Account	CreditCard	7.3e-86
CD_Account	Online	3.5e-35
Education	Family	7.3e-34
Education	Personal_Loan	7.0e-25
Family	Personal_Loan	1.6e-06
CD_Account	Family	1.8e-02



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

EDA - Bivariate - *Personal Loan vs Education*

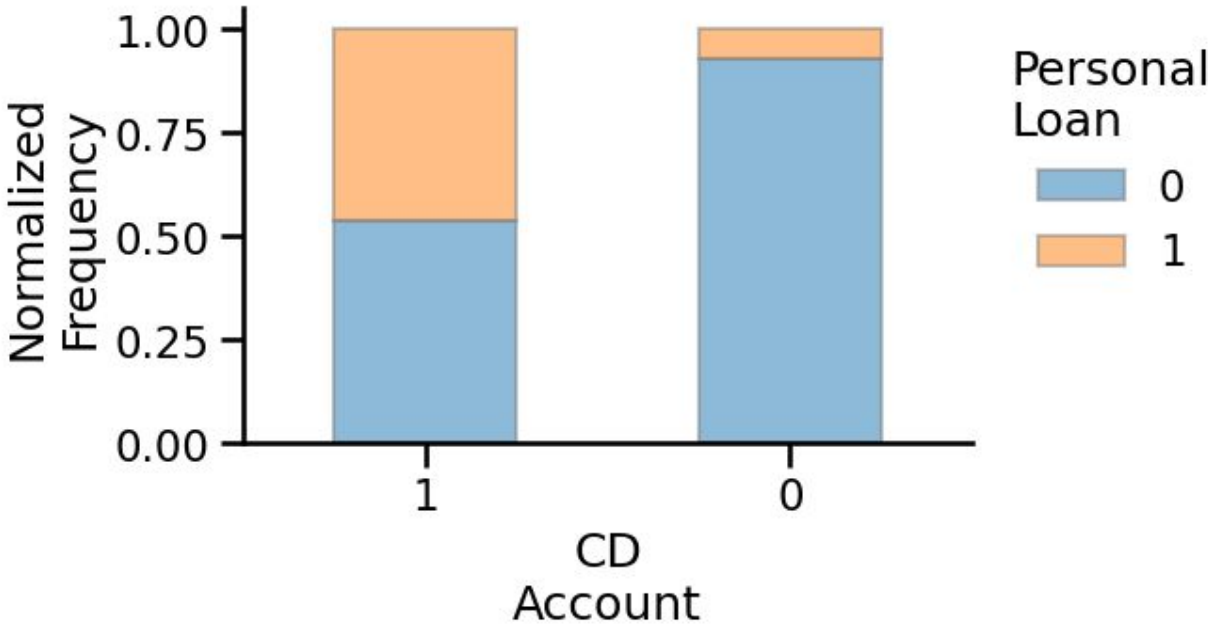
- There seems to be more tendency among those customers who are educationally advanced/professionals to accept the personal loan campaign.



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

EDA - Bivariate - *Personal Loan vs CD Account*

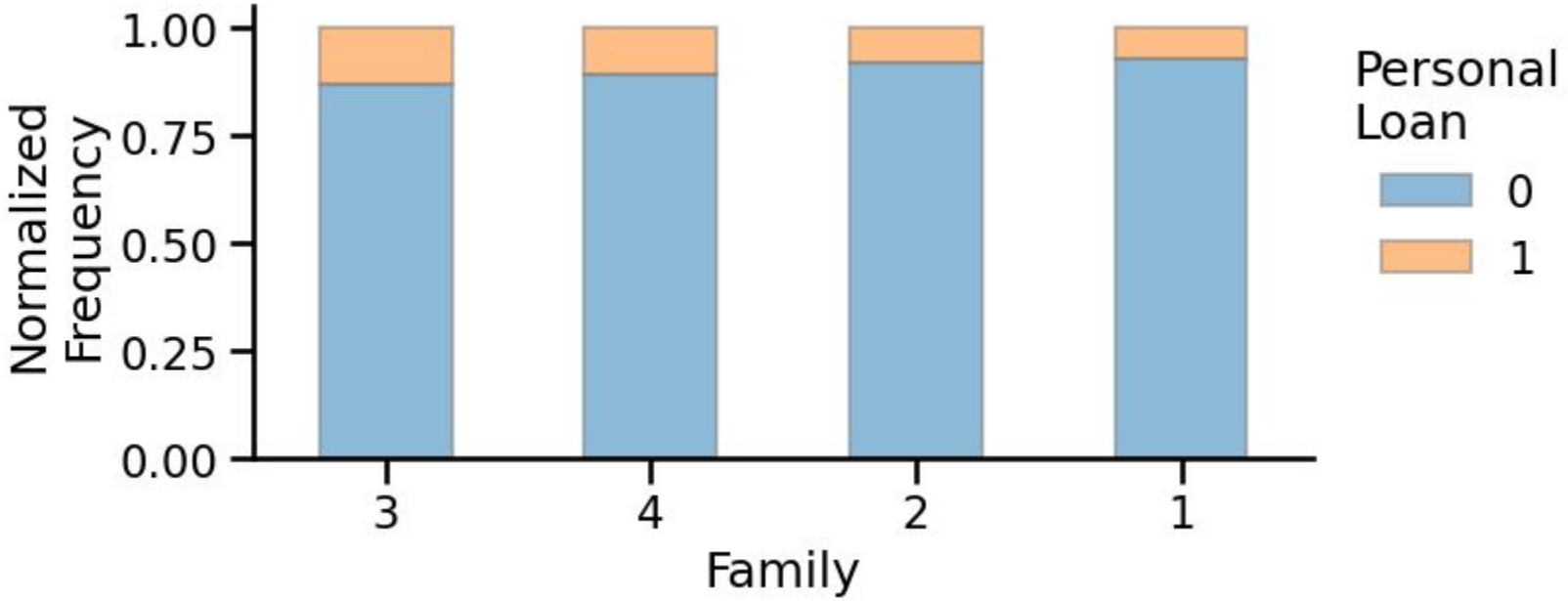
- We observe that the proportion of customers who have CD Accounts who buy the personal loan is higher than those do not.



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

EDA - Bivariate - *Personal Loan vs Family*

- We observe that the bigger families might be more interested in purchasing the personal loan.



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

EDA - Bivariate - *Categorical-Numerical Variables*

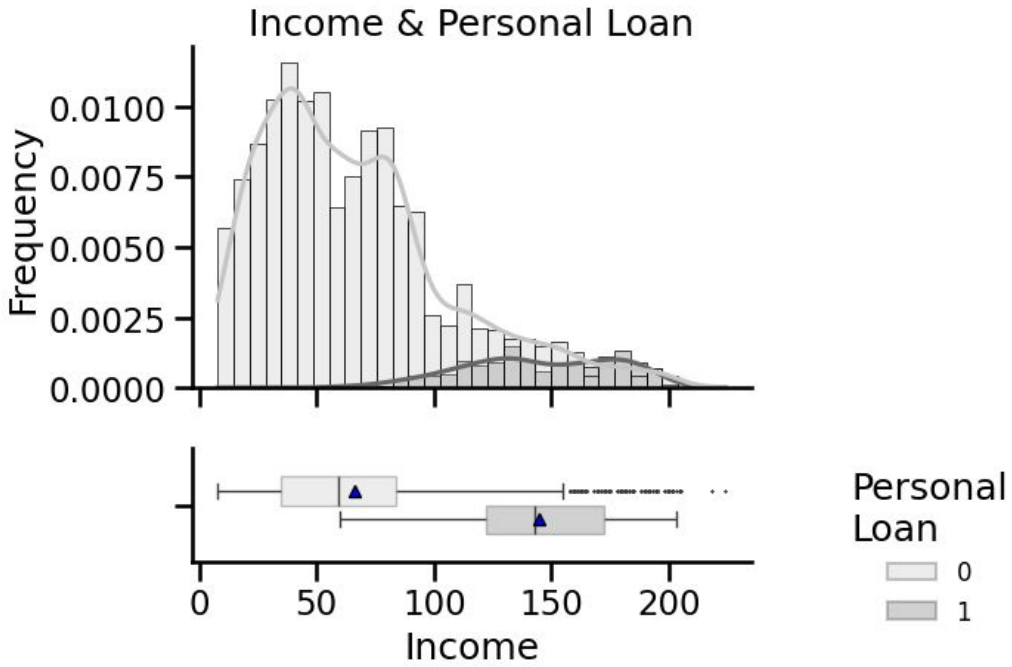
- By conducting **One-Way ANOVA F-test** we observe that the following variables have effects on each other:
    - In the following slides, we take a look at some of these relationships.

Category	Numerical	p-value
Personal_Loan	Income	0.0e+00
Personal_Loan	CCAvg	3.8e-159
Education	Income	8.0e-54
Family	Income	3.1e-39
CD_Account	Income	1.2e-33
Education	CCAvg	7.1e-28
Personal_Loan	Mortgage	5.7e-24
CD_Account	CCAvg	3.1e-22
Family	CCAvg	2.0e-20
CD_Account	Mortgage	2.5e-10
Family	Age	3.6e-05
Family	Experience	3.9e-05
Education	Age	5.1e-03
Education	Mortgage	7.6e-03
Family	Mortgage	2.5e-02

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

EDA - Bivariate - *Personal Loan vs Income*

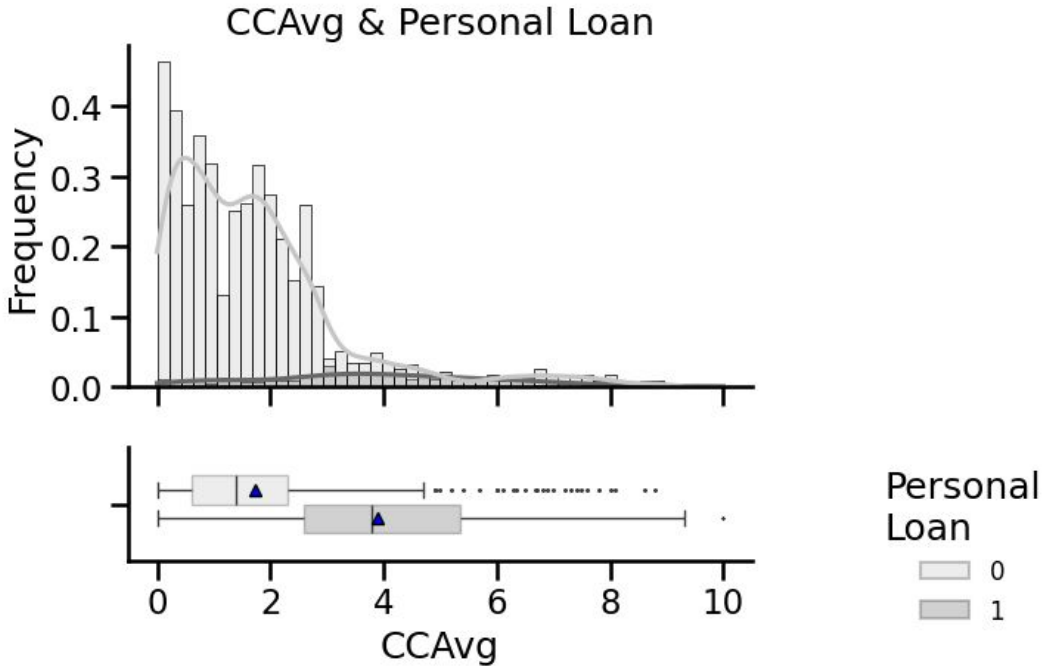
- We observe that the average income of customers who accepted the personal loan offer is significantly ***higher*** than that of customers who did not.



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

EDA - Bivariate - *Personal Loan vs CC Avg*

- We observe that the mean of the average spending on credit cards (per month) of customers who accepted the personal loan offer is ***higher*** than that of customers who did not.



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

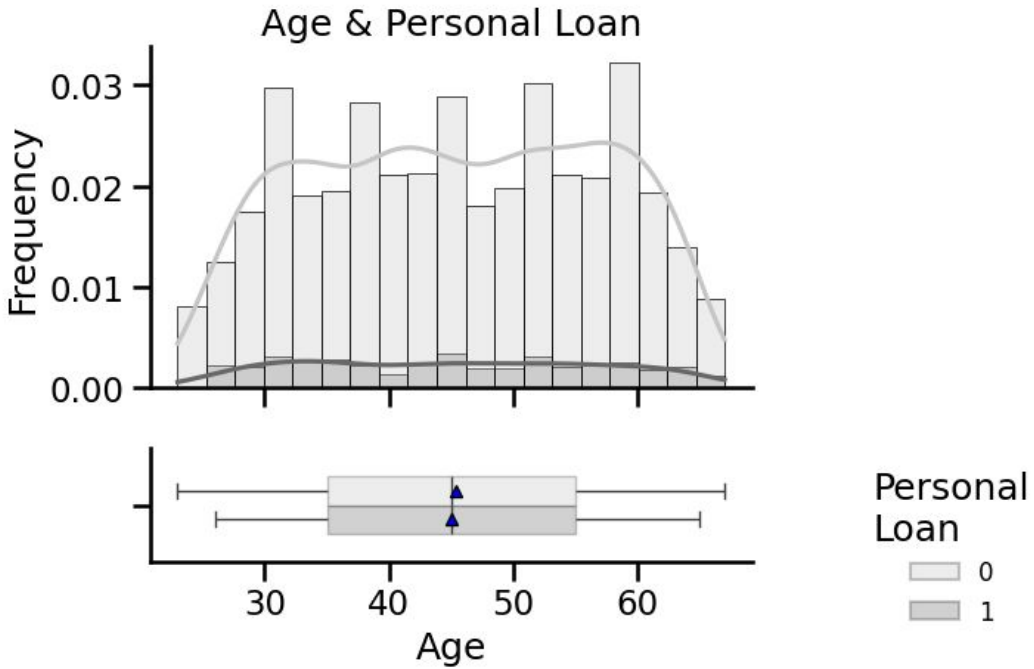
EDA - Bivariate - *Personal Loan vs Mortgage*

- We observe that among those customers who have a mortgage, those with ***higher*** mortgages tend to accept the personal loan campaign.

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

EDA - Bivariate - *Personal Loan vs Age*

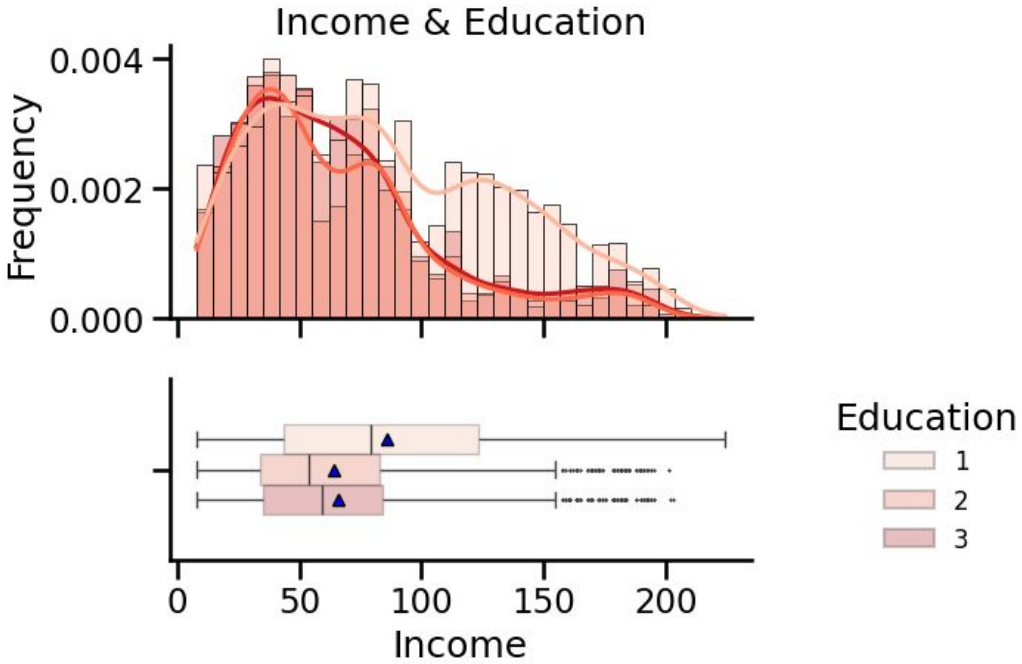
- It looks like the Age ***does not have an effect*** on buying a personal loan.



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

EDA - Bivariate - *Income vs Education*

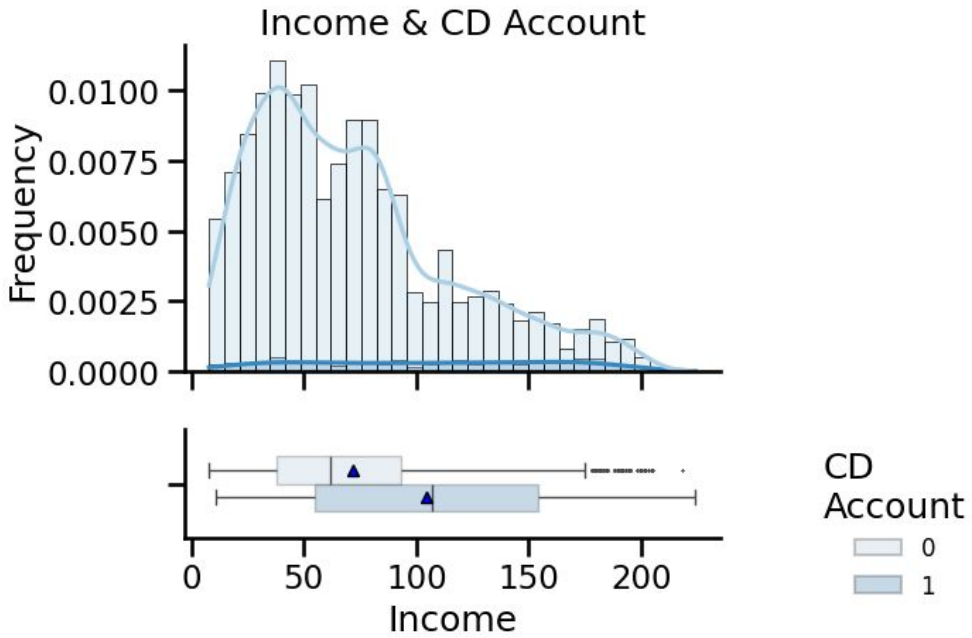
- We observe that the median income of undergraduate customers is higher than those of higher levels of education; however there are more high income outliers among the higher education customers.



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

EDA - Bivariate - *Income vs CD Account*

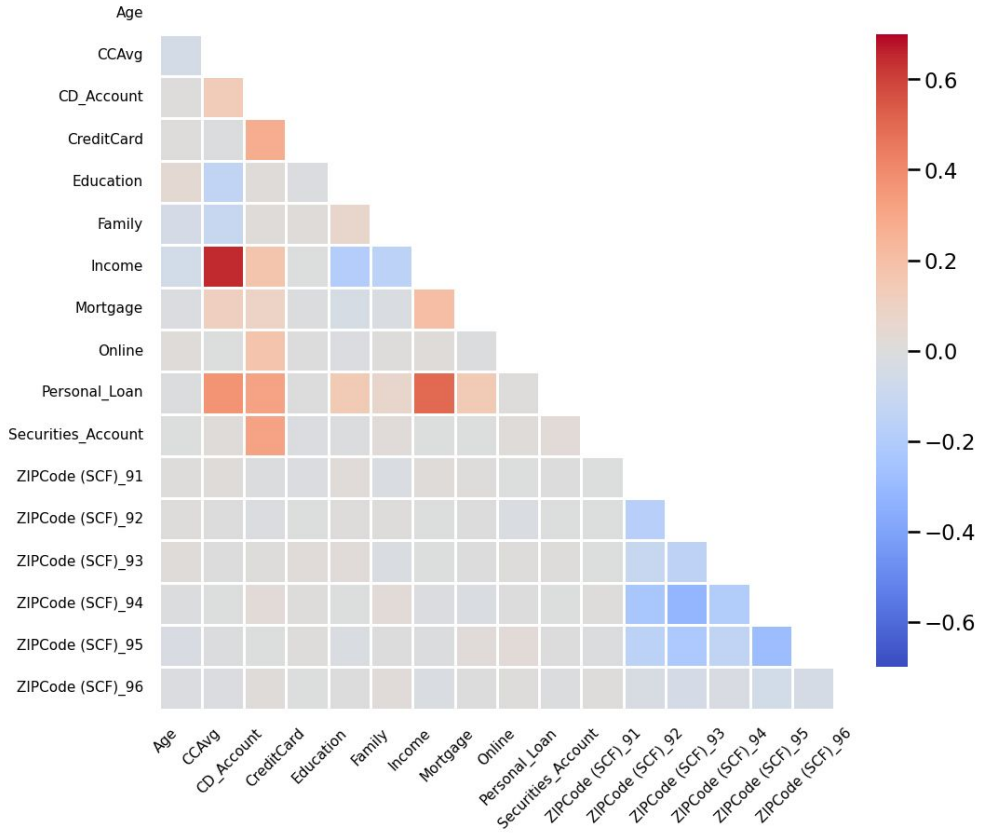
- We observe that the median income of those customers who have CD accounts is higher than that of customers who do not.



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

EDA - Bivariate - *All the variables*

- We can convert the categorical variables into dummy variables and construct the correlation matrix for all the columns
- This confirms our previous observations.



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

Data Preprocessing

- There were ***no duplicates*** in the original data.
- There were ***no missing values*** in the data.
- ***Anonymous Values***: The value of the *Experience* field was negative for 52 instances. These values were *converted to positive* numbers since the years of experience is always *positive* (or equal to zero).
- ***Feature Engineering***: SCF (sectional center facility) is set to be the rightmost two digits of the ZIPCode.
- Data preprocessing for modeling:
  - The *ID column is dropped* because it does not convey any useful information.
  - The *Experience* column is dropped since it's highly correlated with the *Age*.

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

Data Preprocessing

- Outlier Check:
  - Here are the outliers for the numerical columns.
  - For the modeling section, we are not treating the outliers.

Numerical Column	Outliers %
Age	0
Experience	0
Income	1.9
CCAvg	6.8
Mortgage	5.8

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

Model Building

Model
Overfitting
Pruning

We care using **decision tree** algorithm from scikit-learn to train our model for predicting the customers' decision to purchase/not purchase the personal loan.

- We split the data into **train and test sets**. We will train the model on the train set and would evaluate our model by predicting the corresponding results in the test set.

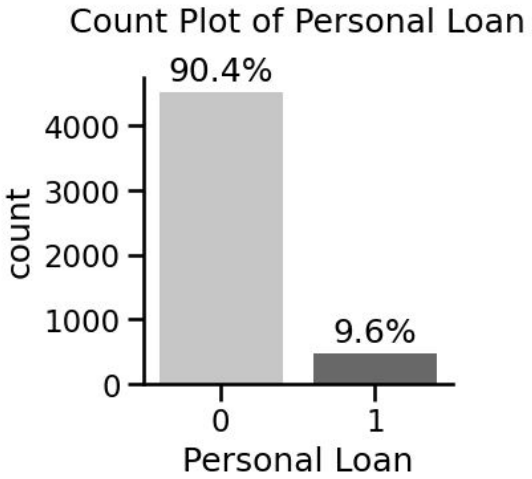
We will use **pruning** techniques to try and reduce **overfitting**.

- **Pre-pruning**: Hyperparameter tuning given a grid of hyperparameters to improve the performance and robustness of the model.
- **Post-pruning:** The nodes with the smallest effective alpha are pruned first.

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

Model Building
Classes

- The proportion of the customers who purchase the personal loan is ~10%. The model could become biased toward the (other) dominant class.
- To resolve this problem we can set the class weight to be be balanced which automatically adjusts the weights to be inversely proportional to the frequencies in the input data.



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

Model Performance Summary

Model Evaluation Criterion

Our model can make wrong predictions in two ways:

- ***False Positive***: Predicting that a customer will purchase the personal loan when they will not.
  - If increased, it raises the cost of targeting uninterested customers.
  - If minimized, it improves the precision.
- ***False Negative***: Predicting that a customer will not be interested in purchasing the personal loan when they would be.
  - If increased, we miss opportunities for selling personal loans.
  - If minimized, it improves the recall.

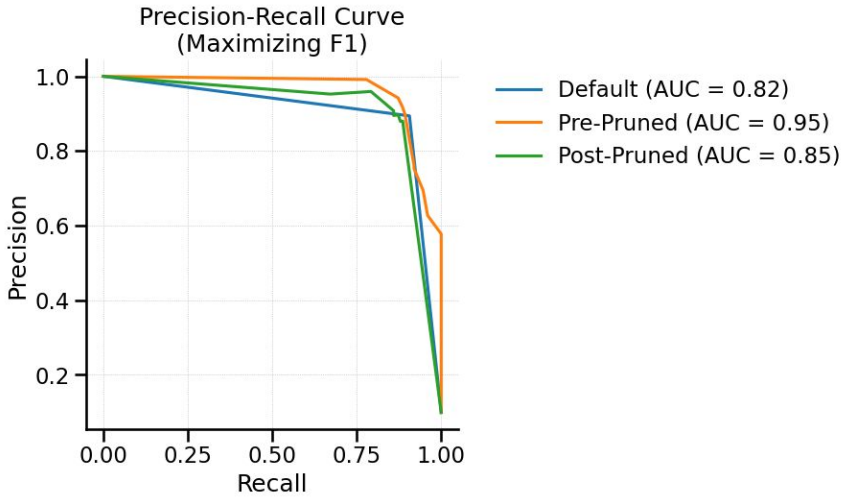
Depending on the cost that is more important to us, we can either minimize recall or precision. By minimizing F1, we could minimize both recall and precision.

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

Model Performance Summary

Final Decision Tree Model & its parameters

We select the pre-pruned model optimized for maximizing F1 using Gini criterion. The selected model has a depth of 6. The F1 score for the test data is .9.



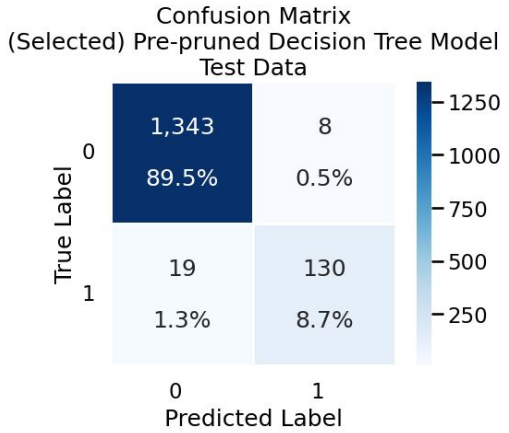
Decision Tree Model	Data	Accuracy	Recall	Precision	F1	Max Depth	Max Leaf Nodes	Min Samples Split	Criterion	alpha	Class Weight	Impurity
(Selected) Pre-pruned	Train	0.99	0.92	0.96	0.94	6	50	50	gini	0.0	None	0.017346
(Selected) Pre-pruned	Test	0.98	0.87	0.94	0.91	6	50	50	gini	0.0	None	0.017346

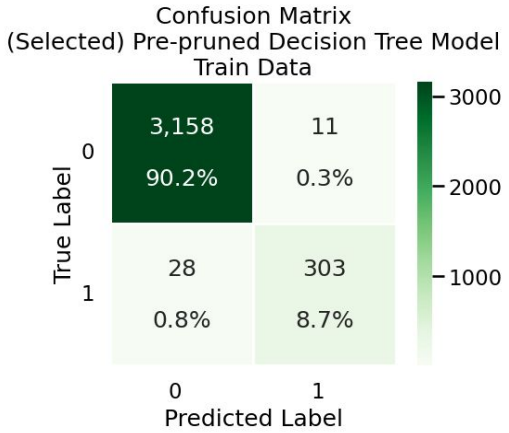
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

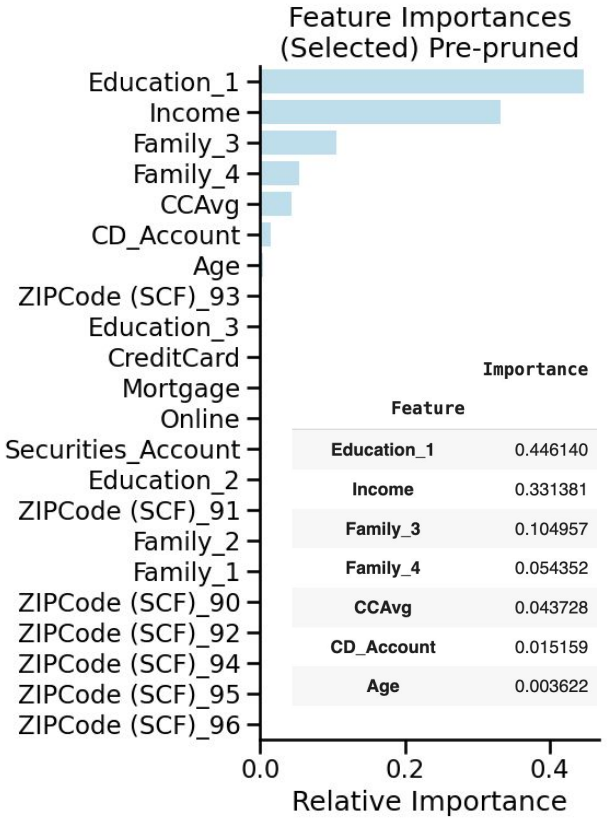
Model Performance Summary

Important Features

- Here are the important features concluded from our selected model with the highest F1
- The confusion matrices show the true positives, true negatives, false positives and false negatives for train and test data







Feature	Importance
Education_1	0.446140
Income	0.331381
Family_3	0.104957
Family_4	0.054352
CCAvg	0.043728
CD_Account	0.015159
Age	0.003622

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

Model Performance Summary

Performance Metrics

- Here we have tabulated summary of key performance metrics for training and test data of all the models optimized for maximizing F1

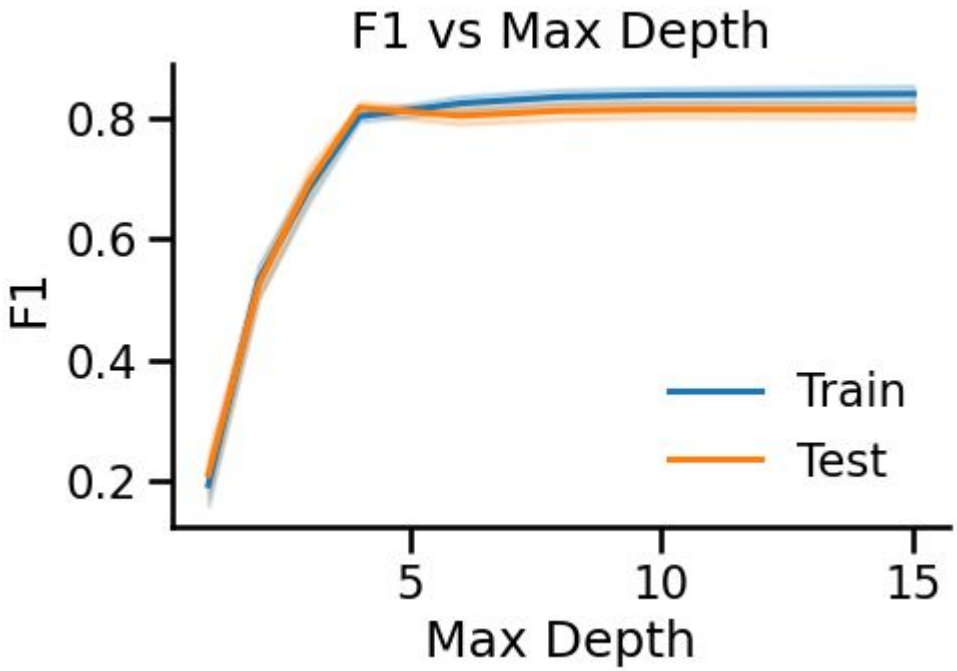
Maximizing Score	Model	Data	Accuracy	Recall	Precision	F1
F1	Default	Train	1.00	1.00	1.00	1.00
	Default	Test	0.98	0.91	0.89	0.90
	Pre-Pruned	Train	0.99	0.92	0.96	0.94
	Pre-Pruned	Test	0.98	0.87	0.94	0.91
	Post-Pruned	Train	1.00	1.00	0.95	0.97
	Post-Pruned	Test	0.98	0.89	0.88	0.88

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

Model Performance Improvement

Pre-pruning & Max Depth

By conducting pre-pruning and turning the model with the max depth, we see the that the F1 improves by increasing the depth at first and then stops improving. Same is true for accuracy, recall and precision.



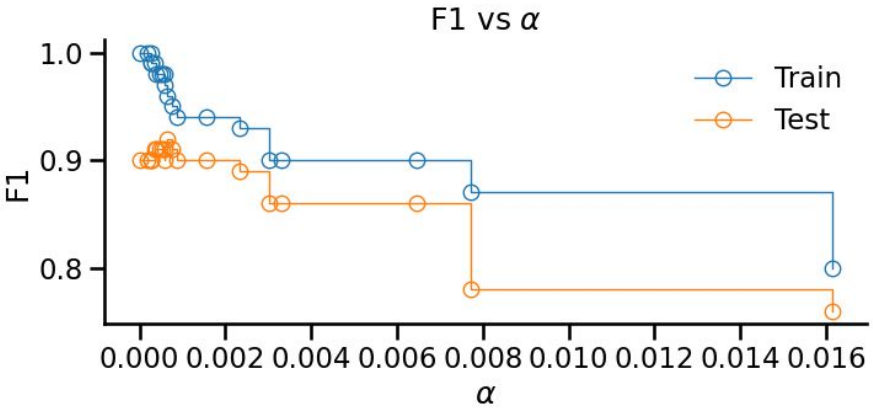
Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

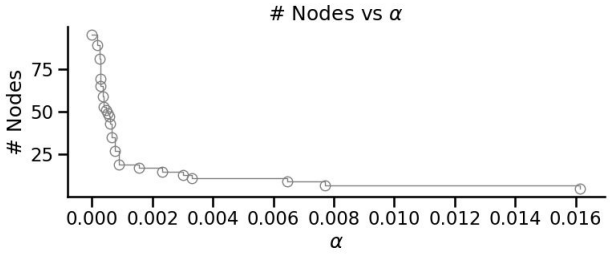
Model Performance Improvement

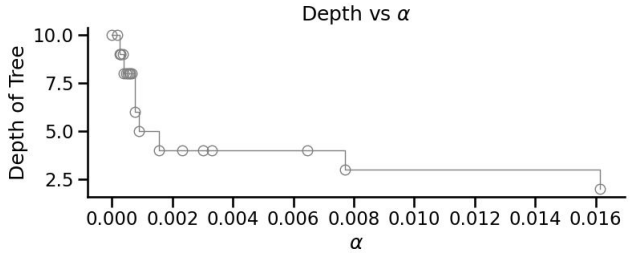
Post-pruning & α

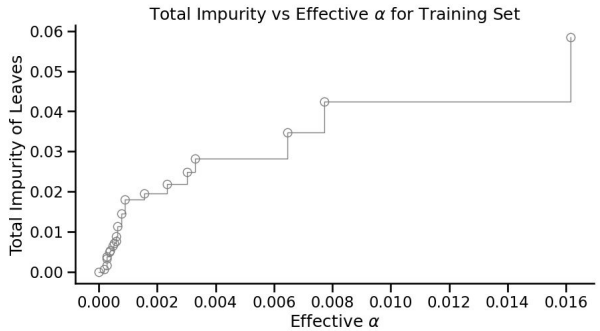
Tree depth & Nodes & f1

We observe that the number of nodes and the depth of the tree and f1 increase with decreasing α, however, the impurity increases.









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

APPENDIX

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

Data Background and Contents

Original Data Dictionary

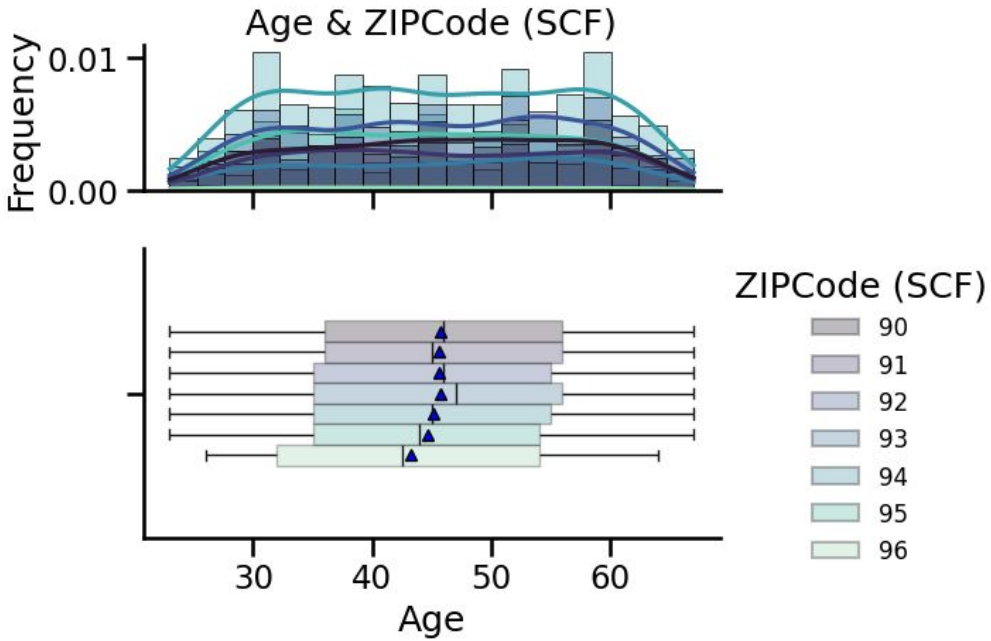
Column	Data Type	Description	# unique
ID	int64	Customer ID	5000
Age	int64	Customer's age in completed years	45
Experience	int64	Number of years of professional experience	47
Income	int64	Annual income of the customer (in thousand dol...	162
ZIPCode	int64	Home address ZIP code	467
Family	int64	The family size of the customer	4
CCAvg	float64	Average spending on credit cards per month (in...	108
Education	int64	Education level (1: Undergrad; 2: Graduate; 3:...	3
Mortgage	int64	Value of house mortgage if any (in thousand do...	347
Personal_Loan	int64	Did this customer accept the personal loan off...	2
Securities_Account	int64	Does the customer have securities account with...	2
CD_Account	int64	Does the customer have a certificate of deposi...	2
Online	int64	Do customers use internet banking facilities? ...	2
CreditCard	int64	Does the customer use a credit card issued by ...	2

Memory Usage	547.0 KB
#	
Rows	5000
Columns	14
Null Values	0
Duplicated Rows	0
int64	13
float64	1

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

EDA - Bivariate - *ZIP Code (SCF) vs Age*

- The median age of customers who live in SCF 96 is lower.



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

Great Learning

Happy Learning !