AllLife Bank

PGP-AIML: Machine Learning Project

5/24/2024

By Anmol Verma

Contents / Agenda

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

Executive Summary

Objective: To build a predictive model using which the bank can identify customers who will be interested in taking a personal loan. For this purpose, we determine the key factors that impact a customer's interest in taking a personal loan

Important Factors: Exploratory data analyses revealed income, family size and education as the most important factors impacting interest in purchase loans. Specifically, the likelihood for taking personal loans was high for the following types of customers

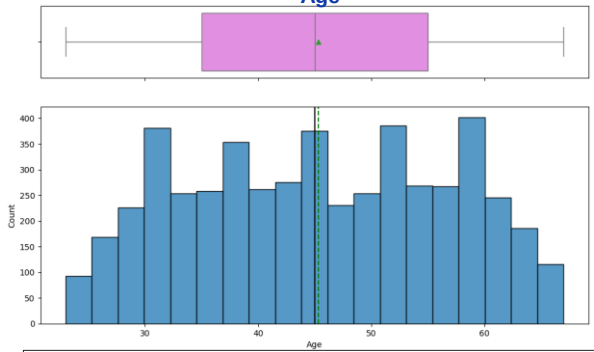
- Higher income customers (specifically, customers with income over $116K)
- Those with a family of more than 2 (larger families tend to have more expenses and thus, possibly exhibit greater demand for credit)
- Customers who are more educated (with education levels greater than under-graduate)
- ~50% of customers who have Certificate of deposit (CD_Account) with the bank also took a personal loan– the bank should target this customer segment as well

Business Problem Overview and Solution Approach

- AllLife Bank is a US bank with a growing customer base. The majority of its customers are liability customers (depositors). The bank has a small borrower base.
- The bank is interested in expanding its borrower base so it can earn more through interest on loans. The management is particularly interested in 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 health conversion rate of > 9% success. The retail marketing department is now interested in devising campaigns for improved targeting to increase this success ratio.
- We analyse the data provided by AllLife Bank and build a machine learning model to help the marketing department identify potential customers who are more likely to take personal loans

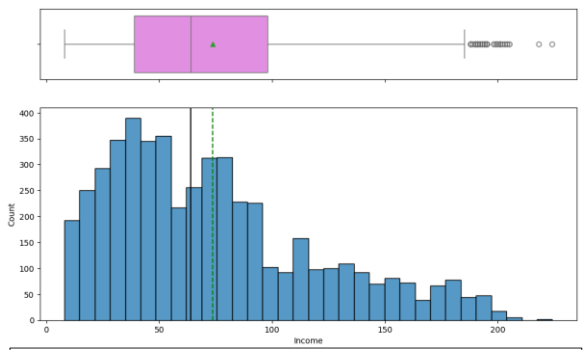
EDA Results: Univariate Analyses

Age

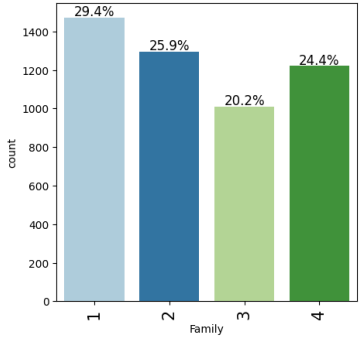


- The mean age is ~45Y and the distribution is normal
- Age and experience are naturally correlated and have a similar distribution

Income

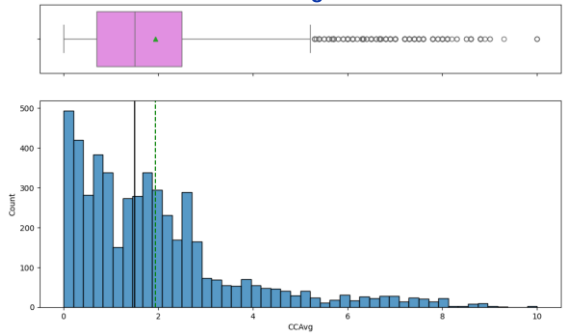


The mean income is 73.7K and the distribution is right skewed with outliers. 50% of the customers have income USD <64K



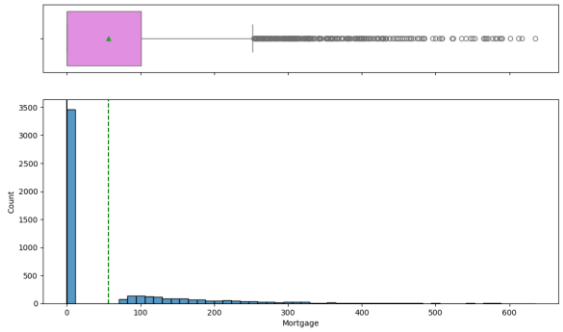
~29% of customers are singles and another 26% have a family size of 2

CCAvg.

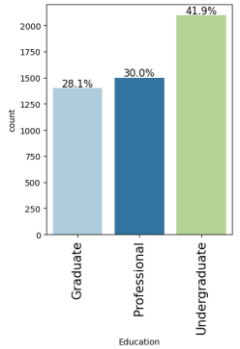


Mean spending on credit cards per month is 1.9K per month and the distribution is right skewed with outliers.

Mortgage

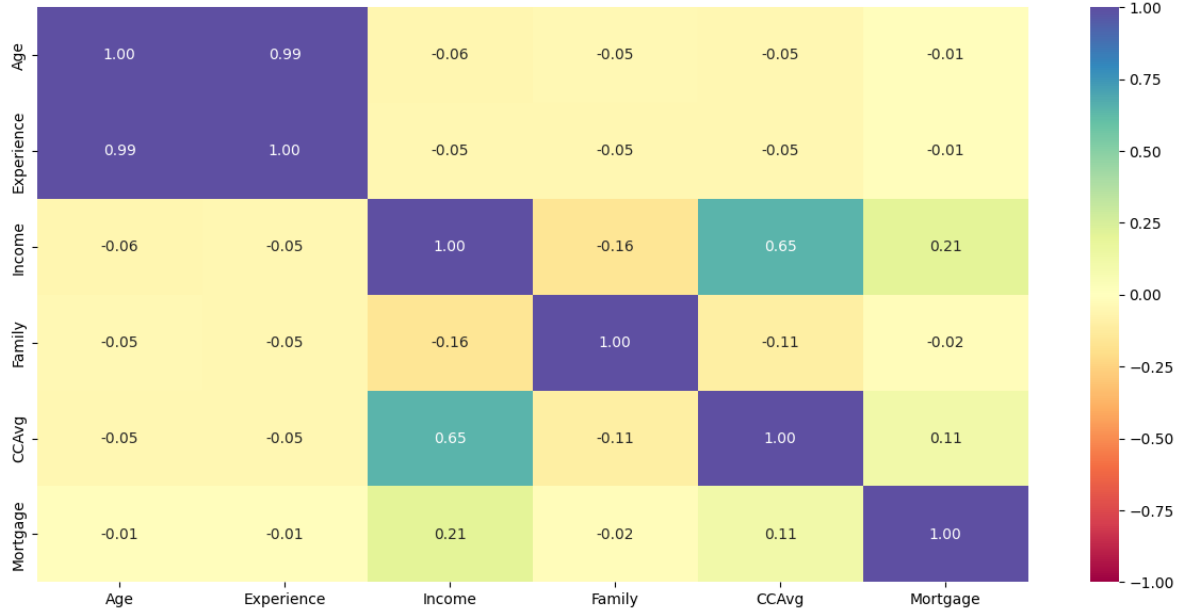


Mean value of house mortgage is 56.5K. Many customers don't have a mortgage but the data also has many high-value outliers



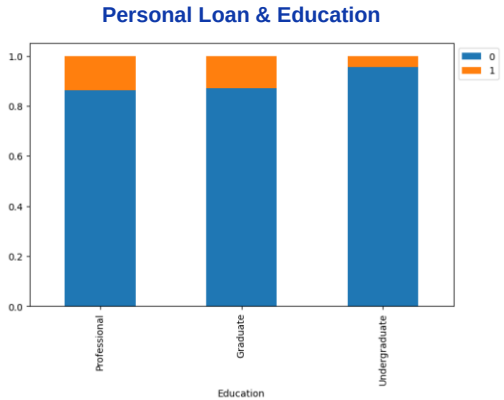
42% have an undergraduate degree

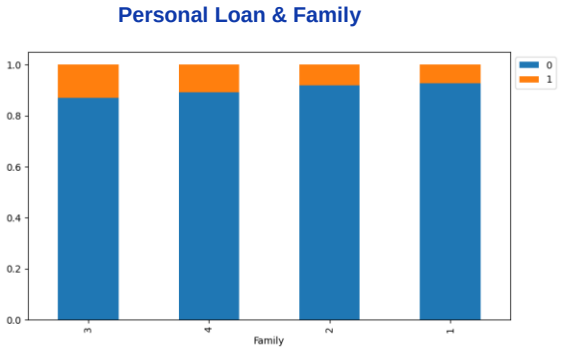
EDA Results: Bivariate Analyses (I)

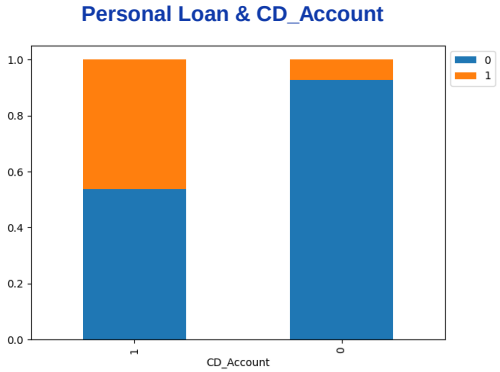


- There is 65% correlation between CCavg & Income. This makes sense as higher income customers tend to have higher credit card average bills.
- Understandably so, there is 99% experience between age and experience – thus, it makes sense to drop one of these as they will provide similar information to the model
- Counter-intuitively, family and income have negative correlation, however, the correlation is not so strong as to allow us to make a definitive conclusion.

EDA Results: Bivariate Analyses (II)



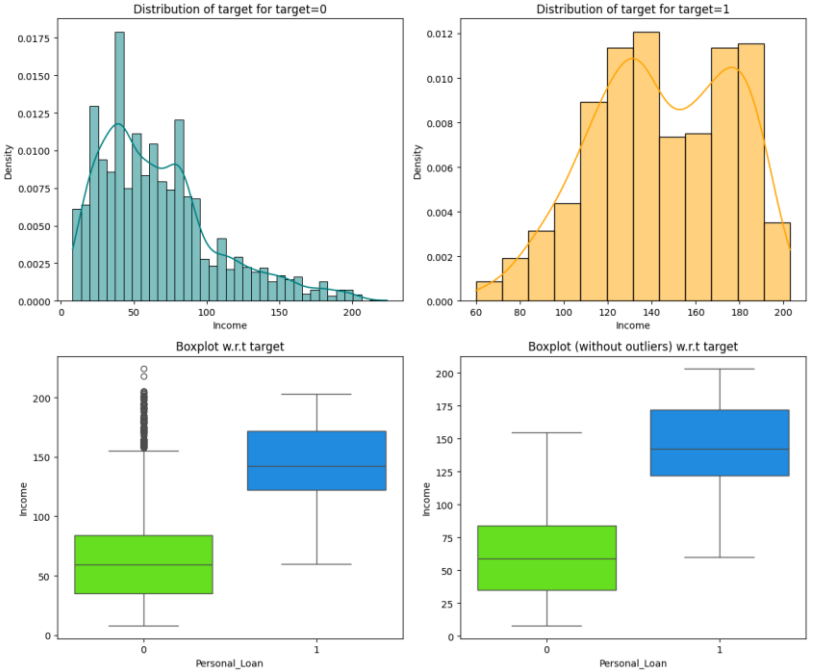




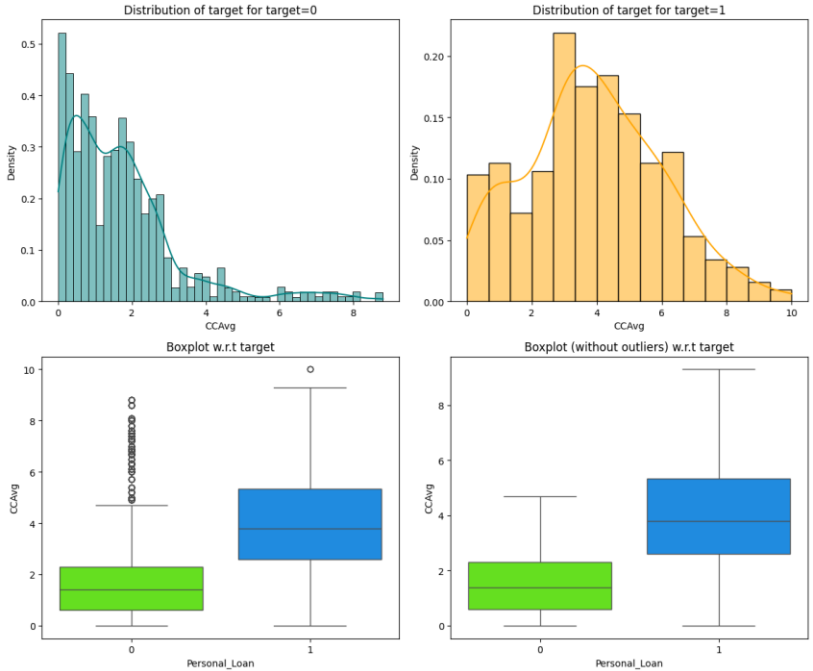
Main observations and targeting strategies

- Majority of the customers that accepted the personal loan offered in the last campaign were either graduates or advanced/professional graduates. **Customers with a higher education** are a sensible category to target for personal loans
- Majority of the customers that accepted the personal loan offered in the last campaign were family size of 3 or 4. This suggests that family size may be a factor in predicting a customer's interest in purchasing a personal loan, with **larger families having a higher propensity to take out a loan**.
- **Customers with a CD account are much more likely to purchase a personal loan** compared to those without a CD account and thus may be another segment to target for personal loans
- **Factors that mattered less:** Customers that have online accounts, securities accounts or credit card balances or those from a specific zipcode did not show meaningfully more propensity to purchase personal loans. Similarly, age and experience also mattered less as factors. Our current hypothesis is that these are not important factors in determining a customer's interest in personal loans

EDA Results: Bivariate Analyses (III)



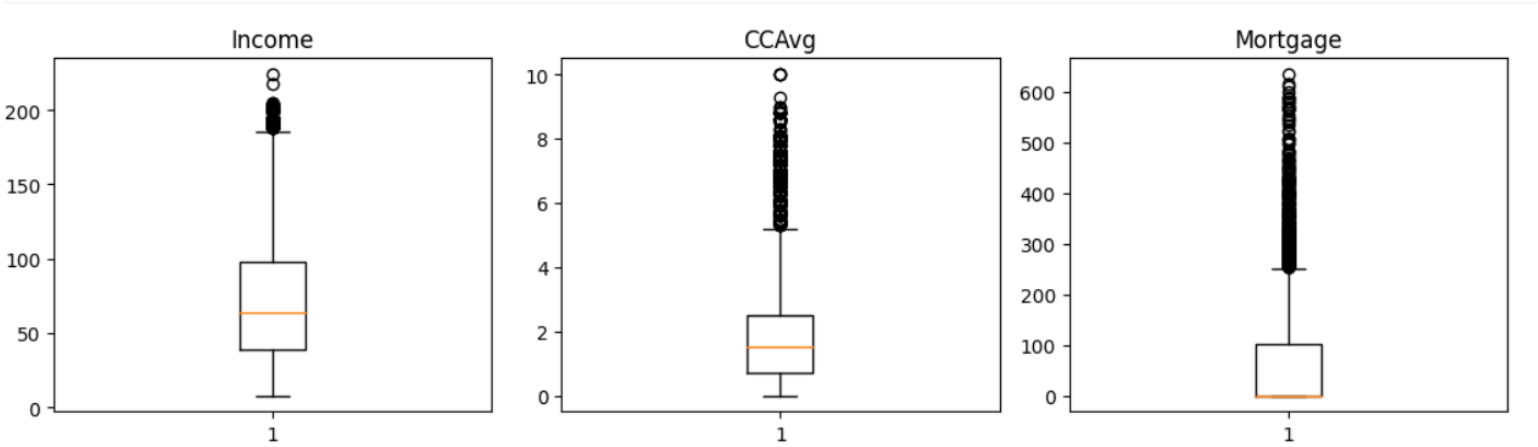
Income can be a significant predictor of a customer's interest in purchasing a personal loan as there is a noticeable shift in the distribution of income between the two groups, with those interested in personal loans showing higher incomes. Customers with income higher than $100K are more likely to take personal loans and could be a good segment to target.



Customers with higher average credit card spending (CCAvg) are more likely to purchase a personal loan as there is a noticeable shift in the distribution of CCAvg between the two groups, with those interested in personal loans showing higher spending.

Data Preprocessing (I/II)

- **Missing Values:** The data has no missing values
- **Outliers:** There are a lot of outliers in income, CCAvg (average spending on credit card per month) and Mortgage (value of house mortgage) however, we've chosen to not treat them and leave them as is because we want to preserve data integrity and prevent AllLife Bank from losing critical customer data.



Data Preprocessing (II/II)

Great Learning
POWER AHEAD

Feature Engineering

- **ZipCode:** We convert the code to string type, determine the number of unique two-digit prefixes of the ZIP codes (7) and, then update the ZipCode column to contain only these two-digit prefixes. We then convert the Zip code column to a categorical data type. We do this operation because in this case, the specific geographical region indicated by the first two digits of the ZIP code is more relevant than the full ZIP code. In this way, we convert the 467 unique zip codes into 7 groupings.
- **Converting data type of categorical features to category:** We perform this operation on categorical variables like Education, Personal Loan, Securities Account, CD_Account, Online, Credit Card and ZipCode. We do this to appropriately highlight their semantic meaning and enhance the effectiveness of machine learning algorithms.

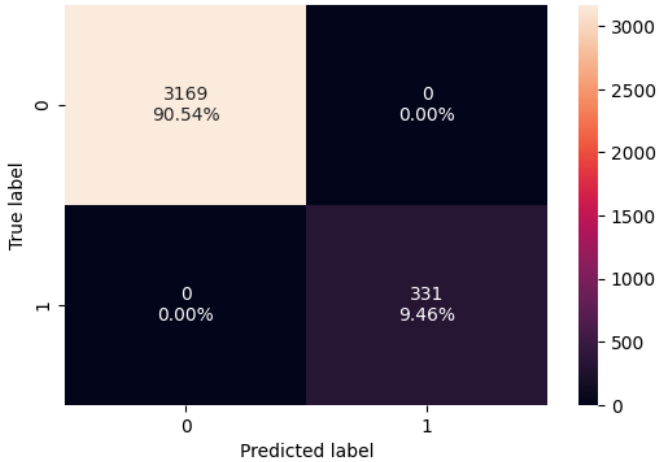
Data preprocessing for modeling: Upon checking for anomalous values, we discover that *experience* has three negative values. We replace these with absolute values as experience cannot be negative.

Model Building

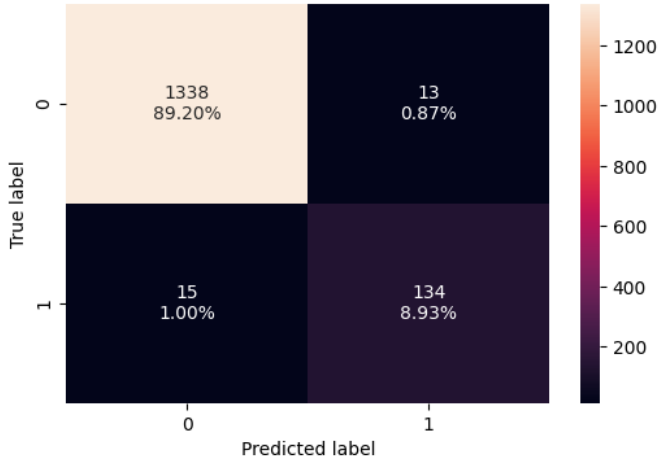
Objective: To predict whether an AllLife bank depositor will buy a personal loan or not

- **Creating functions:** We create functions to compute different metrics and confusion matrix to check the performance of a classification model built using sklearn. We do this so don't have to use the same code repeatedly
- **Building decision tree:** We use Gini Impurity as the criterion to measure the quality of a split. Gini Impurity measures how often a randomly chosen element would be incorrectly classified if it was labelled according to distribution of labels in the dataset.
- **Model evaluation criteria:** Model can make wrong predictions in two cases: a) we offer personal loans to customers who don't want them (wasted effort) b) we don't offer personal loans to customers who want them (wasted opportunity)
  Which case is more important?
  Since this exercise is focused on enhancing revenue prospects for the bank, we would like to offer personal loans to as many prospects as possible. Case (a) is acceptable but Case (b) is not. Thus, our objective is to **reduce false negatives** and we will thus, **maximize recall**

Model Performance summary (training & testing data)



	Accuracy	Recall	Precision	F1
0	1.0	1.0	1.0	1.0



	Accuracy	Recall	Precision	F1
0	0.981333	0.899329	0.911565	0.905405

Most significant predictors for buying a personal loan as indicated by the decision tree:

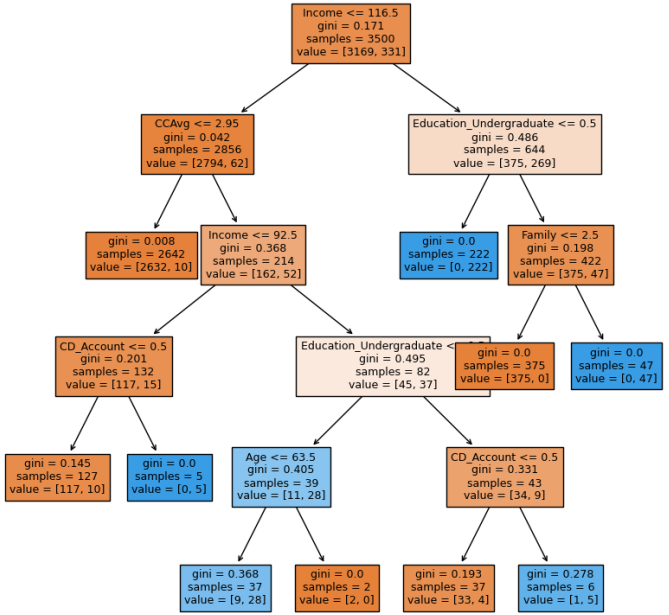
- Education Undergraduate
- Income
- Family
- Credit Card Spending (CCAvg)

Model Performance Improvement : Pre pruning (I)

The decision tree model with default parameters has an overfitting problem and is not able to generalize well. We used **grid search** to get an optimal model and did pre-pruning using the following hyperparameters

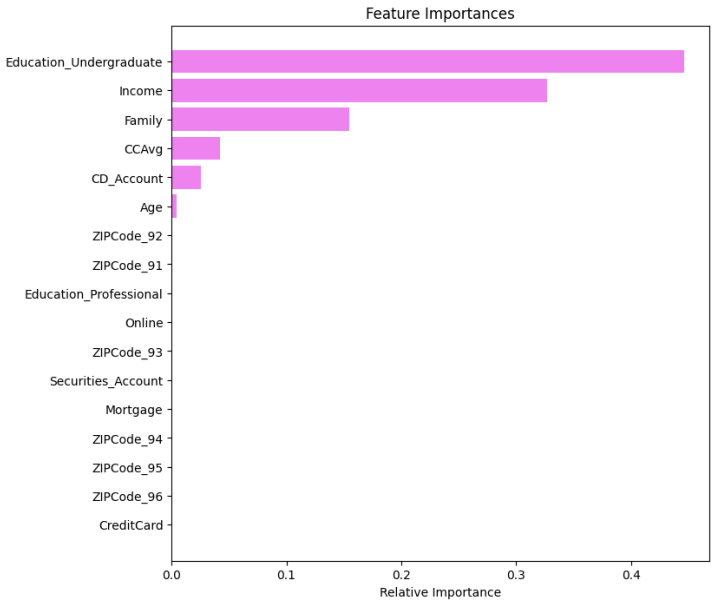
- Max_depth
- Min_samples_leaf
- Max_leaf_nodes

As can be seen from the pre-pruned visualization on the right, model performance improves using pre-pruning.



Model Performance Improvement : Pre pruning (II)

Feature importance after pre-pruning



Comparing training and testing performance (Pre-pruning)

Training data

	Accuracy	Recall	Precision	F1
0	0.990286	0.927492	0.968454	0.947531

Testing Data

	Accuracy	Recall	Precision	F1
0	0.98	0.865772	0.928058	0.895833

- Model performance on training data has reduced but ~93% recall is still good.

Model Performance Improvement : Pre pruning (III)

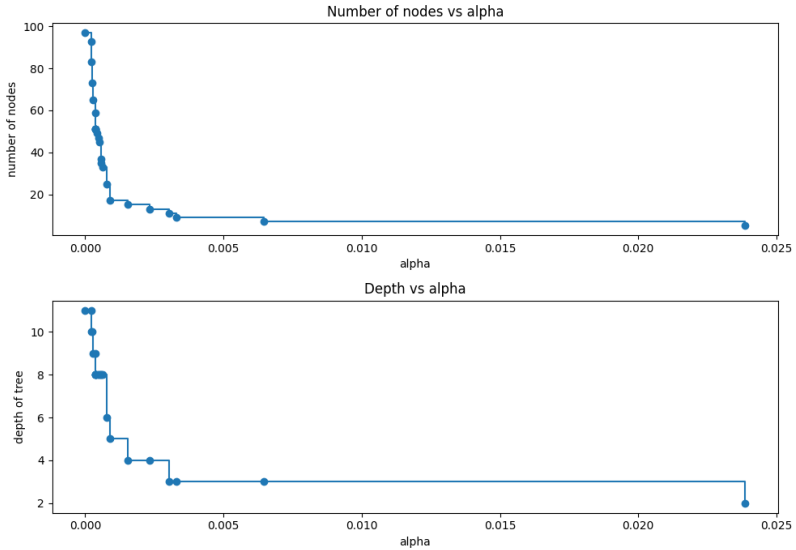
Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)
Accuracy	1.0	0.990286
Recall	1.0	0.927492
Precision	1.0	0.968454
F1	1.0	0.947531

Test performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)
Accuracy	0.981333	0.980000
Recall	0.899329	0.865772
Precision	0.911565	0.928058
F1	0.905405	0.895833

Cost Complexity Pruning



- We tried cost complexity pruning to overcome the overfitting problem
- However, best model obtained with cost-complexity pruning is similar to the initial default model
- The pre-pruned tree thus gives the best generalized performance with strong precision and recall

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
Decision Tree	1	0.981	1	0.899	1	0.911
Decision Tree - Pre-Pruning	0.99	**0.98**	0.927	**0.865**	0.968	**0.928**

APPENDIX

Data Background

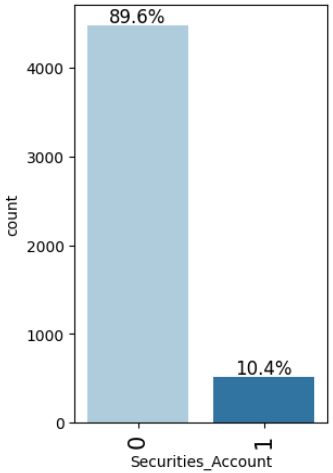
	count	mean	std	min	25%	50%	75%	max
ID	5000.0	2500.500000	1443.520003	1.0	1250.75	2500.5	3750.25	5000.0
Age	5000.0	45.338400	11.463166	23.0	35.00	45.0	55.00	67.0
Experience	5000.0	20.104600	11.467954	-3.0	10.00	20.0	30.00	43.0
Income	5000.0	73.774200	46.033729	8.0	39.00	64.0	98.00	224.0
ZIPCode	5000.0	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

Select Observations

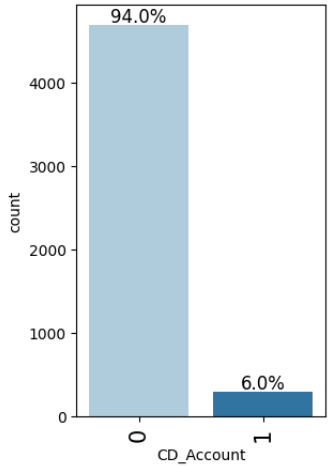
- The data has 5000 customers with many attributes including age, experience, income, family, CC Avg, education, mortgage etc
- We drop Customer ID as it will not add value to the analyses
- Average age is 45 years, ranging from 23 to 67 years. The average experience is 20 years. Age and experience both are symmetrically distributed about mean and median
- Average Income is $73K; Income has a wide range from $8K to 224K
- Average family size is 2.4

EDA Results: Univariate Analyses

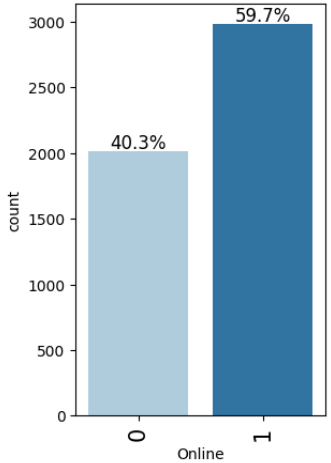
Many banking products are under-penetrated among the customer base



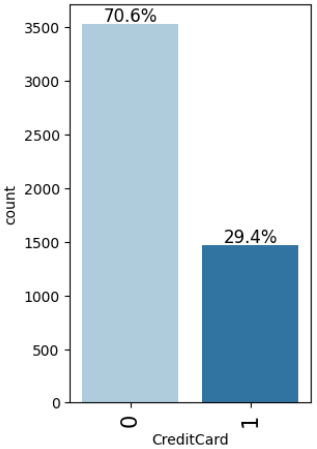
90% customers don't have a securities account with the bank



94% customers don't have a certificate of deposit account with the bank



40% of customers don't have internet banking facilities



71% of customers don't have a credit card issued by another bank and only 29% do

Great Learning

Happy Learning !