greatlearning
Power Ahead

Personal Loan Analysis

Contents

Step 1: Reading and Understanding the Data

Step 2: Performing EDA - Exploratory Data Analysis

Step 3: Data Cleaning and Preparation

Step 4: Data Visualization

- 4.1 Univariate Analysis for all features having continuous data
- 4.2 Bivariate Analysis for all features having continuous data
- 4.3 Univariate Analysis for all features having categorical data
- 4.4 Bivariate Analysis for all features having categorical data

Step 5: Data Preprocessing

- 5.1 Missing Value Treatment (Imputing Missing Values)
- 5.2 Outlier Detection and Treatment
- 5.3 Scaling/Normalization of Features

Step 6: Model Building, Predicting and Evaluation, Insights & Classification model

Step 7: Logistic Regression

Step 8: Decision Tree

Step 9: Conclusion & Recommendation

Business Problem Overview and Solution Approach

Objective & Business Goal:

Data Analysis to identify the potential customers who have a higher probability of purchasing the loan.

The classification goal is to predict the likelihood of a liability customer buying personal loans which means we must build a model which will be used to predict which customer will be most likely to accept the offer for personal loan, based on the specific relationship with the bank across various features given in the dataset. Here I will be using the Supervised Learning methods to predict which model is best for this problem amongst Logistic Regression.

Data Overview:

The file Bank.xls contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 accepted the personal loan that was offered to them in the earlier campaign.

Exploratory Data Analysis

No.of columns: 14

No.of rows: 5000

List of rows and description:

Nonimal Variables :

ID, ZIP Code

Categorical variables :

Family, Education

Interval Variables :

Age, Experience, Income, CCAvg, Mortgage

Binary Categorical Variable :

CD Account, Security Account, Online, Credit Card, Personal Loan

Datatype of given data set:

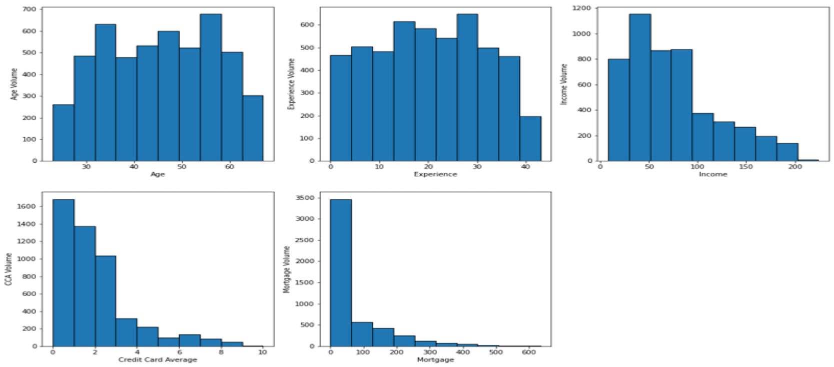
All the data-types are either int64 or float64.

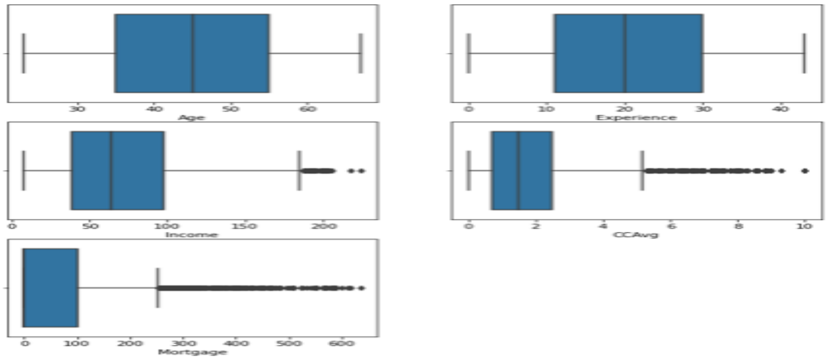
There is no null or missing values in the given dataset.

Coeffients:

Negative value in the Experience field (-3.0).

Univariate Analysis for the continuous variables:

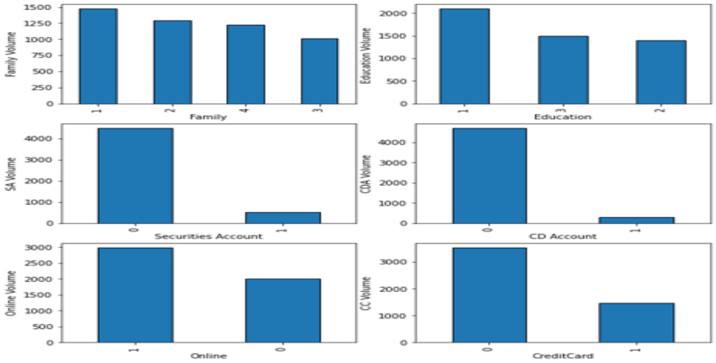




- Age and Experience has normal distribution.
- Income, Credit Card Average and Mortgage are highly skewness

- Age has normally distributed, 35 to 55 age
- Experience has normal distribution, 11 to 30 years
- Remaining field has positive skewness

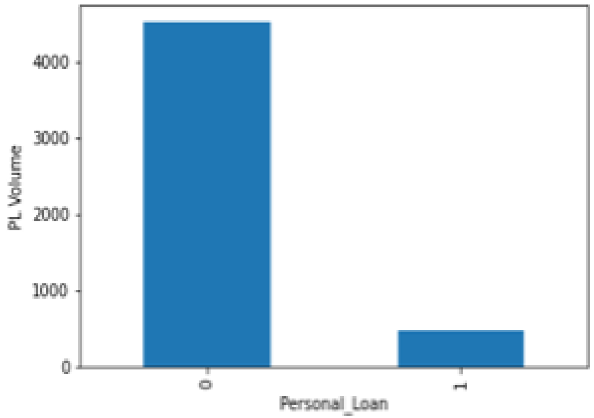
Univariate Analysis of the discrete variables:



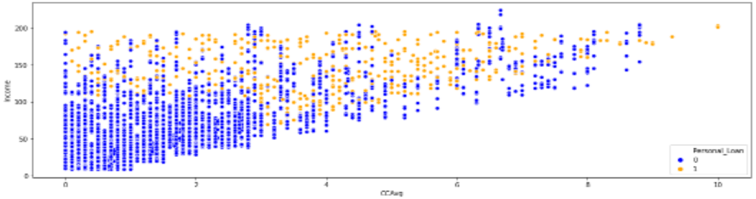
- Family and Education has normal distribution
- Variate in the Securities Account and CD Account

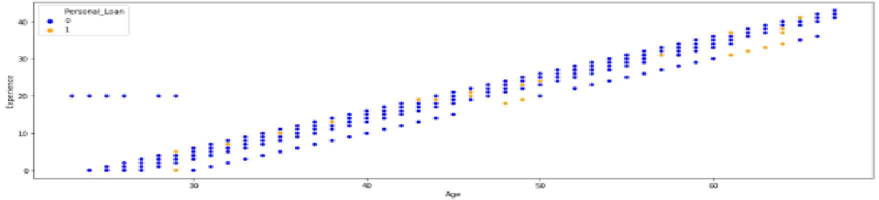
Checking dependent variable:

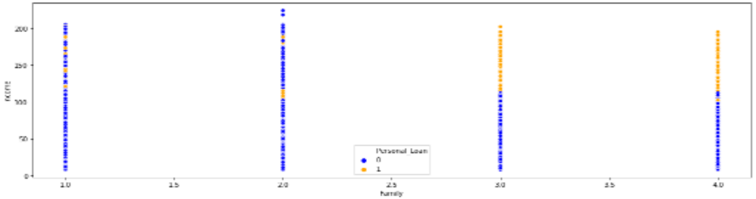
Personal Loan 0 – 4520
Personal Loan 1 -- 480

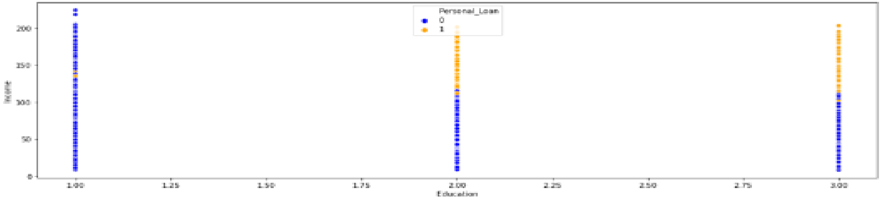


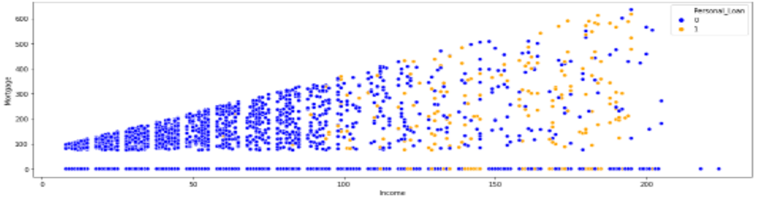
Comparison charts using different variables by dependent variable:

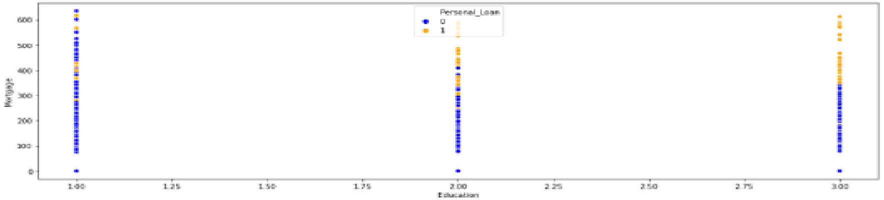






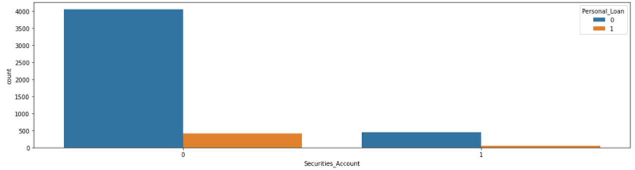


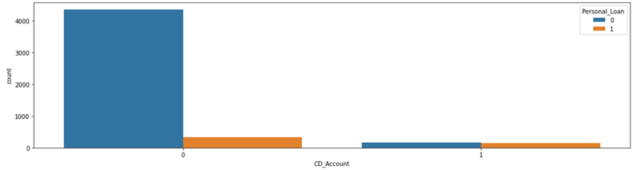


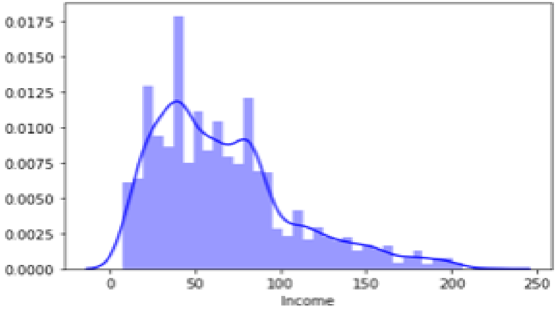


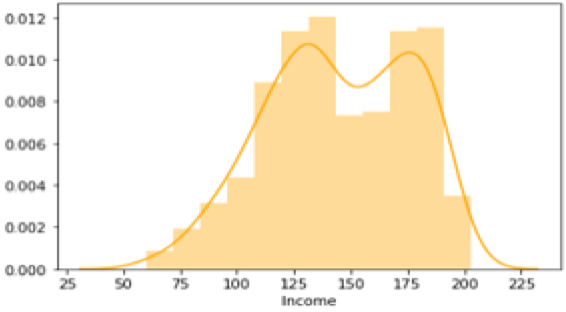
- Based on CCAvg, Family & Income increases, personal loan also got increasing

- Below comparison chart, based on income vs personal loan

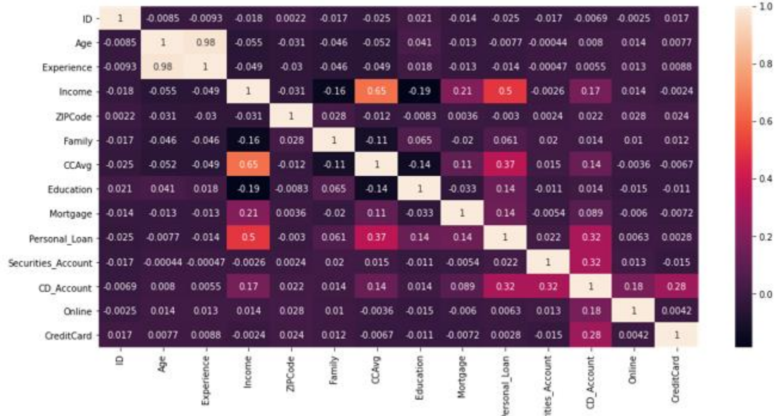






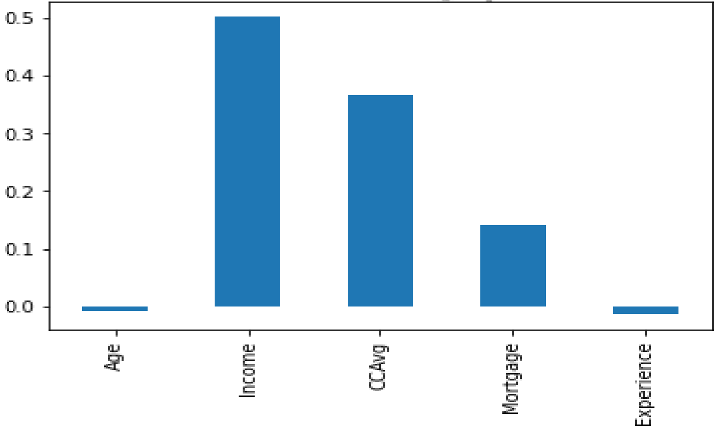


Bivariate Analysis:



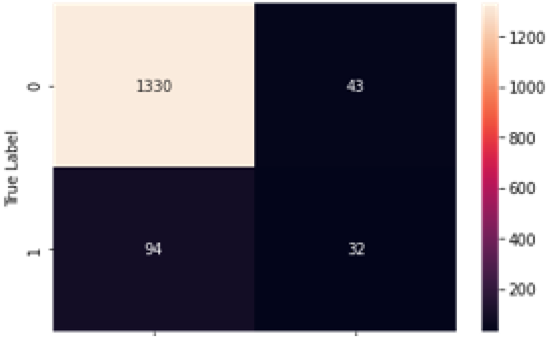
	ID	Age	Experience	Income	ZIPCode	Family	CCAvg	Education	Mortgage	Personal_Loan	rities_Account	CD_Account	Online	CreditCard
ID	1	-0.0085	-0.0093	-0.018	0.0022	-0.017	-0.025	0.021	-0.014	-0.025	-0.017	-0.0069	-0.0025	0.017
Age	-0.0085	1	0.98	-0.055	-0.031	-0.046	-0.052	0.041	-0.013	-0.0077	-0.00044	0.008	0.014	0.0077
Experience	-0.0093	0.98	1	-0.049	-0.03	-0.046	-0.049	0.018	-0.013	-0.014	-0.00047	0.0055	0.013	0.0088
Income	-0.018	-0.055	-0.049	1	-0.031	-0.16	0.65	-0.19	0.21	0.5	-0.0026	0.17	0.014	-0.0024
ZIPCode	0.0022	-0.031	-0.03	-0.031	1	0.028	-0.012	-0.0083	0.0036	-0.003	0.0024	0.022	0.028	0.024
Family	-0.017	-0.046	-0.046	-0.16	0.028	1	-0.11	0.065	-0.02	0.061	0.02	0.014	0.01	0.012
CCAvg	-0.025	-0.052	-0.049	0.65	-0.012	-0.11	1	-0.14	0.11	0.37	0.015	0.14	-0.0036	-0.0067
Education	0.021	0.041	0.018	-0.19	-0.0083	0.065	-0.14	1	-0.033	0.14	-0.011	0.014	-0.015	-0.011
Mortgage	-0.014	-0.013	-0.013	0.21	0.0036	-0.02	0.11	-0.033	1	0.14	-0.0054	0.089	-0.006	-0.0072
Personal_Loan	-0.025	-0.0077	-0.014	0.5	-0.003	0.061	0.37	0.14	0.14	1	0.022	0.32	0.0063	0.0028
Securities_Account	-0.017	-0.00044	-0.00047	-0.0026	0.0024	0.02	0.015	-0.011	-0.0054	0.022	1	0.32	0.013	-0.015
CD_Account	-0.0069	0.008	0.0055	0.17	0.022	0.014	0.14	0.014	0.089	0.32	0.32	1	0.18	0.28
Online	-0.0025	0.014	0.013	0.014	0.028	0.01	-0.0036	-0.015	-0.006	0.0063	0.013	0.18	1	0.0042
CreditCard	0.017	0.0077	0.0088	-0.0024	0.024	0.012	-0.0067	-0.011	-0.0072	0.0028	-0.015	0.28	0.0042	1

- Income shows the highest correlation with CCAvg (0.65)
- Age and Experience are very highly correlated(0.98) with each other.
- Age and Experience are highly correlated, and the correlation is almost 1.
- 'Income' and 'CCAvg' is moderately correlated.
- We can see in above heat map there is association of 'CD Account' with 'Credit Card', 'Securities Account', 'Online', 'CCAvg' and 'Income'.
- 'Income' influences 'CCAvg', 'Personal Loan', 'CD Account' and 'Mortgage'.



- Income & CCAvg shows the highest correlation

Logistic Regression



Accuracy on train set: 0.91

Accuracy on test set: 0.91

Recall score: 0.25

ROC AUC score: 0.61

Precision score: 0.43

```
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      1373
           1       0.43      0.25      0.32       126

    accuracy                           0.91      1499
   macro avg       0.68      0.61      0.63      1499
weighted avg       0.89      0.91      0.90      1499
```

Insights:

- True Positive (observed=1,predicted=1): Model predicted that 32 customers shall take Personal loan and they customer took it
- False Positive (observed=0,predicted=1): Model Predicted 43 Personal loan will take, and the customer did not take it, but bank didn't loose any money
- True Negative (observed=0,predicted=0): Model Predicted 1330 Personal loan will not take, and the customer did not take it
- False Negative (observed=1,predicted=0): Model Predicted 94 Personal loan will not take, and the customer took it - This is where model should have done better

Build Decision Tree Model:

Confusion matrix:

Personal Loan 0: 0.905429
Personal Loan 1: 0.094571

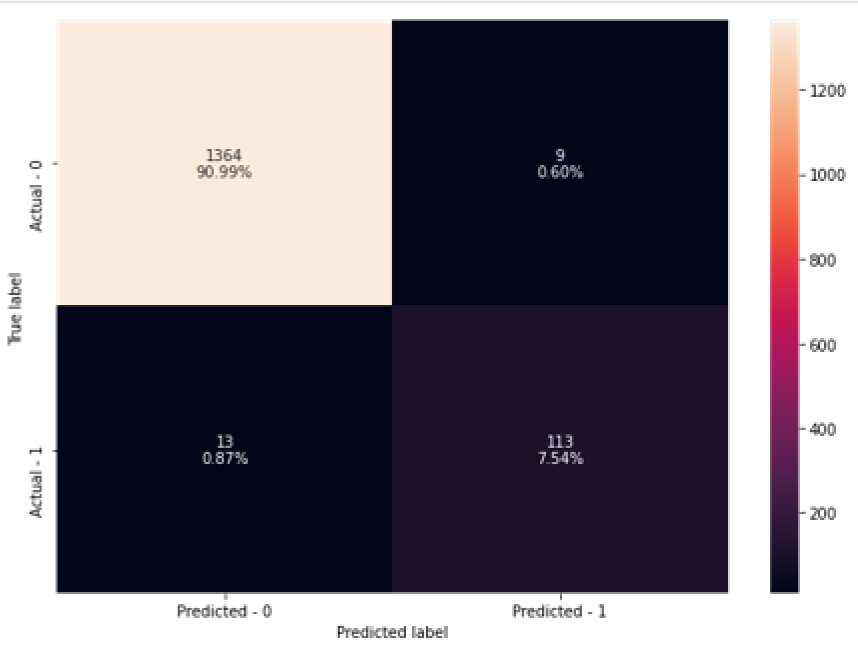
Data has 90.99% of Positive

Recall Score:

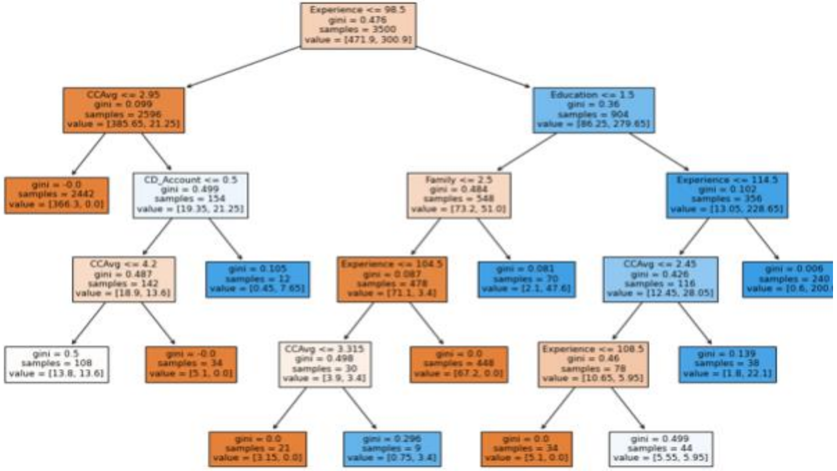
Recall on training set : 1.0

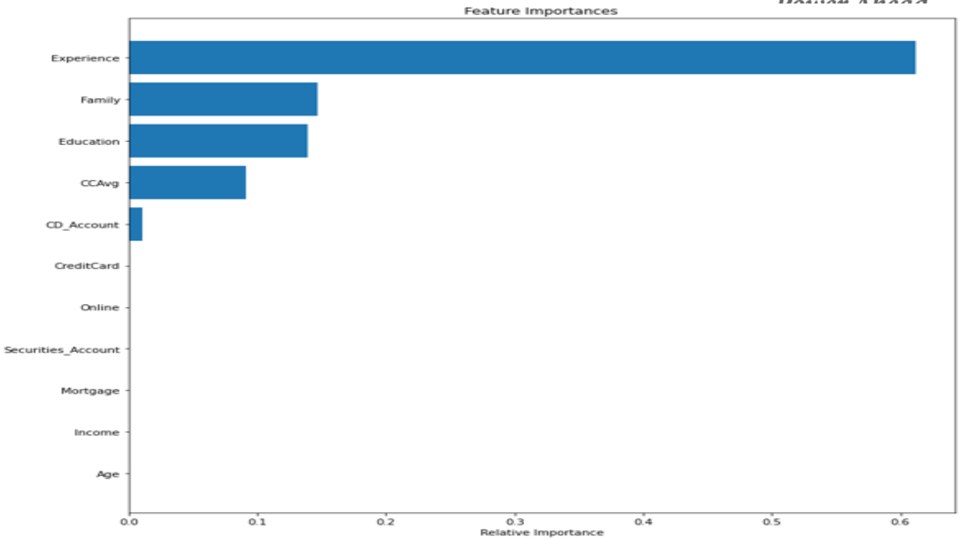
Recall on test set : 0.8968253968253969

- There is no huge disparity in performance of model on training set and test set



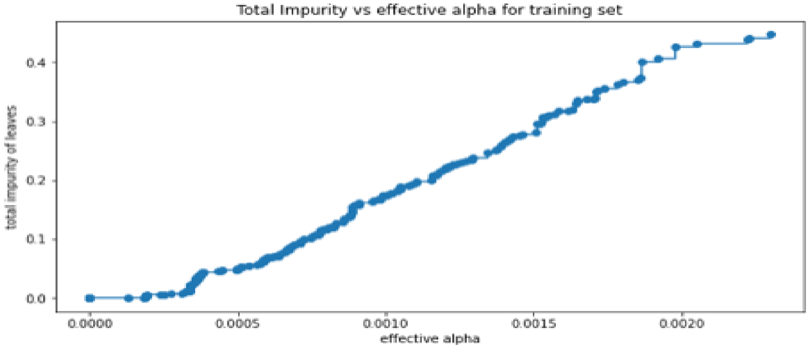
Visualizing the Decision Tree:

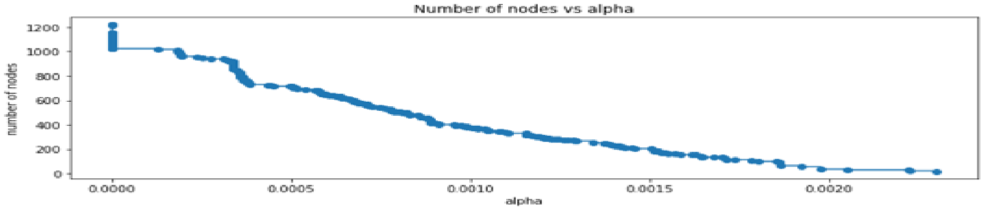


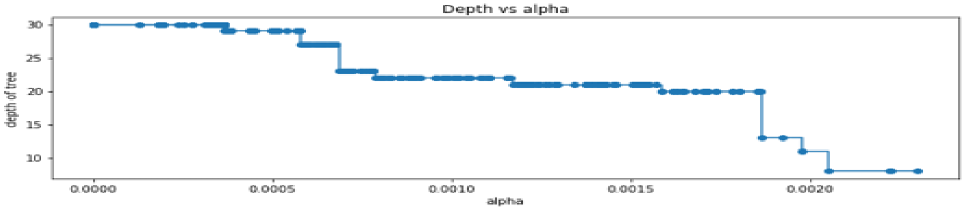


- Experience & Family are at the top two important features to predict

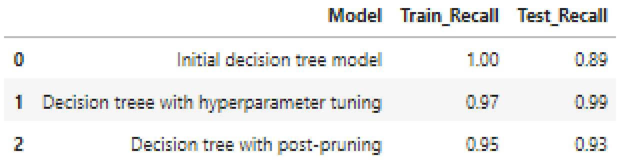
Cost Complexity Pruning:







Comparing all the decision tree models:



	Model	Train_Recall	Test_Recall
0	Initial decision tree model	1.00	0.89
1	Decision treee with hyperparameter tuning	0.97	0.99
2	Decision tree with post-pruning	0.95	0.93

- Decision tree model with hyperparameter tuning has given the best recall score on data.

Conclusion and Recommendations:

•I have analyzed the "Personal Loan" using different techniques and used Decision Tree Classifier to build a predictive model for the same.

•The model built can be used to predict which feature is going to contribute to Personal loan generation.

•Visualized different trees and their confusion matrix to get a better understanding of the model. Easy interpretation is one of the key benefits of Decision Trees.

•Verified the fact that how much less data preparation is needed for Decision Trees and such a simple model gave good results even with outliers and imbalanced classes which shows the robustness of Decision Trees.

•Experience, Family, Education, CCAvg and CC_Account are the most important variable in predicting the customers that will contribute to the revenue.

•The aim of the Bank is to convert liability customers into loan customers.

•It seems like 'Logistic Regression' algorithm & Decision tree model with hyperparameter tuning' have the highest accuracy and we can choose that as our final model

Thank you