Thera Bank Personal loans Analysis

Link dataset: https://www.kaggle.com/datasets/teertha/personal-loan-modeling

I. INTRODUCTION

A personal loan allows the purchase of goods and services. Basically, a loan means we buy goods on credit. A personal loan can be defined as the amount a person borrows from a bank or other financiers for their personal use. Personal loans are a type of consumer finance offered by lenders where banks and financial institutions help consumers deal with shortfalls. short-term personal financial shortfalls. These were large purchases of consumables and/or other large household expenses. These loans are secured by goods purchased as collateral or by someone involved in a supervisory or unsecured role.

This case is about a Thera Bank, whose management wanted the means to convert its passive customers into personal loan customers (while retaining them as depositors). They intend to launch a new marketing campaign, therefore they require information regarding the relationship between the factors in the data. Last year, the bank executed a responsible customer campaign that yielded a good rate of over 9%. This prompted the retail marketing department to join focused marketing initiatives. Better to aim for a higher success rate with less expenditure.

II. THEORETICAL BACKGROUND

1. Logistic Regression

The method of modeling the probability of a discrete result given an input variable is known as logistic regression. Logistic regression (or logit regression) is a technique for estimating the parameters of a logistic model in regression analysis (the coefficients in the linear combination). In binary logistic regression, there is a single binary dependent variable with two values labeled "0" and "1" that is coded by an indicator variable.

$$\text{Logistic function} = \frac{1}{1 + e^{x}}$$

2 Decision Tree

A decision tree is a form of prediction model in machine learning, which is a mapping from observations about a thing/phenomenon to

judgments about the object's goal value phenomenon. Each internal node represents a variable, and the line connecting it to its descendants reflects that variable's value. Given the values of the variables indicated by the path from the root node to that leaf node, each leaf node indicates the anticipated value of the target variable. Decision tree learning, or simply called decision trees for short, is a machine learning technique used in decision trees.
Decision trees can also be used to compute conditional probabilities in a descriptive manner. A decision tree is a combination of mathematics and computational techniques.
The data is given as records of the form:

$$(\boldsymbol{x}, \boldsymbol{y}) = (x_1, x_2, x_3, \ldots, x_k, y)$$

The dependent variable y is the variable that we need to understand, classify or generalize. x_1, x_2, x_3,..., are variables that will help us do that work.
Classification tree: if y is a categorical variable such as: gender (male or female), the outcome of a match (win or lose).

3 Random Forest

Random forest is a supervised learning method that may be applied to both classification and regression problems. Because random is random and forest is forest, I will build multiple decision trees using the Decision Tree algorithm in the Random Forest algorithm, but each decision tree will be unique (with random factor). The prediction results from the decision trees are then combined.
Random Forest works with missing data. When the Forest has more trees, we can avoid Overfitting with the data set. The Random Forest algorithm has a common application in the problem of finding potential customers and fraudulent customers.

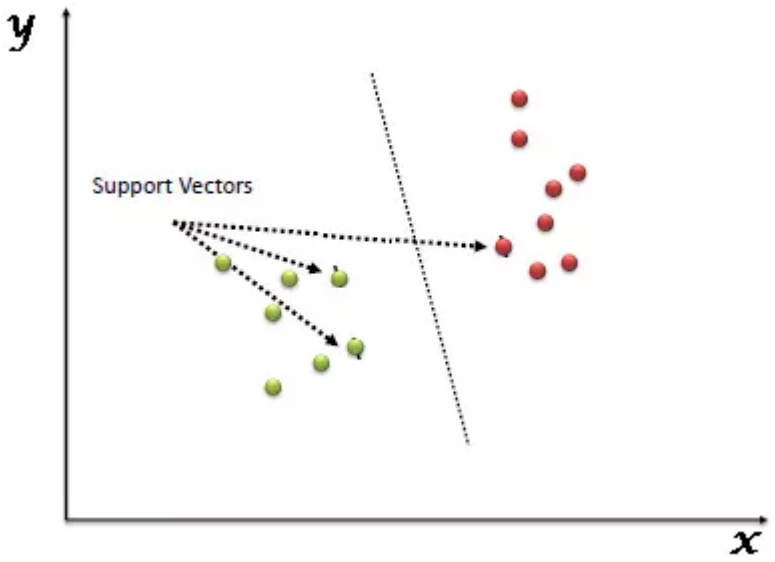
4. K-nearest neighbors (KNN)

K-Nearest Neighbor algorithm (KNN) is a supervised learning technique used to classify new observations by finding similarities between this new observation and data. whether available. KNN is a simple and intuitive model, but still highly effective because it is non-parametric. The model makes no assumptions about the data distribution. Moreover, it can be used directly for multi-class classification.

The KNN algorithm has many applications in the investment industry, including bankruptcy prediction, stock price prediction, corporate bond credit rating allocation, and custom bond and equity index generation.

5. Support Vector Machine(SVM)

SVM is a supervisory method that can be used for recursion or classification. However, it is primarily used for classification. We plot the data as points in n dimensions (where n is the number of features you have), with the value of each feature contributing to the overall association.The layers are then divided using a "hyper-plane" discovery. Simply put, a hyper-plane is a line that divides layers into two sections.



Support Vectors understand simply as objects on the observed coordinate graph, Support Vector Machine is a border to divide the two best classes.

III. DATA

1. Overview

This data is taken from on Kaggle. The dataset includes 5000 rows and 14 columns. The data includes customer demographics, customer relationship with the bank, and customer response to the latest personal loan campaign.

	Columns	Description
1	ID	Customer ID is unique for each customer which is assigned by the Bank.
2	Age	Age of a particular Customer in completed years.
3	Experience	professional experience in terms of years.
4	Income	Annual income of the customer (\$ 000).
5	ZIP Code	Home Address ZIP code of the customer.
6	Family	Size of the customer's Family.
7	CCAvg	Average spending on credit cards per month by customers ($000)
8	Education	Education Level of the customers. 1: Undergrad; 2: Graduate; 3: Advanced/Professional.
9	Mortgage	If any customer has any house mortgage then the value of house mortgage. (\$ 000).
10	Securities Account	Does the customer have a certificate of deposit (CD) account with the bank?
11	CD Account	Does the customer use internet banking facilities?
12	Online	Does the customer use a credit card issued by UniversalBank?
13	Credit Card	Did this customer accept the personal loan offered in the last campaign? (Target Attribute)
14	Personal Loan	Does the customer have a securities account with the bank?

There are no empty or (NaN) values in the dataset. The dataset has a combination of categorical numeric attributes, but all categorical data are represented numerically.

2. Data processing

Drop columns operation

- The ID column in our database has a unique number for every client.
- The ZIP Code column in our database has a ZIP Code for the city of the clients.
- So there is no relation between the ID column or ZIP Code column and any other variable.

- So it will be useful when dropping them to prevent occurence of misleading information.

```
df.drop(['id', 'zip_code'], axis=1, inplace=True)
```

The dataset after drop columns.

	age	experience	Income	family	ccavg	education	mortgage	personal_loan	securities_account	cd_account	online	CreditCard
0	25	1	49	4	1.6	1	0	0	1	0	0	0
1	45	19	34	3	1.5	1	0	0	1	0	0	0
2	39	15	11	1	1.0	1	0	0	0	0	0	0
3	35	9	100	1	2.7	2	0	0	0	0	0	0
4	35	8	45	4	1.0	2	0	0	0	0	0	1

```
df.describe()
```

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	personal_loan	securities_account	cd_account	Online	CreditCard
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.00000	5000.000000	5000.000000
mean	45.338400	20.104600	73.774200	2.396400	1.937938	1.881000	56.498800	0.096000	0.104400	0.06040	0.596800	0.294000
std	11.463166	11.467954	46.033729	1.147663	1.747659	0.839869	101.713802	0.294621	0.305809	0.23825	0.490589	0.455637
min	23.000000	-3.000000	8.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
25%	35.000000	10.000000	39.000000	1.000000	0.700000	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
50%	45.000000	20.000000	64.000000	2.000000	1.500000	2.000000	0.000000	0.000000	0.000000	0.00000	1.000000	0.000000
75%	55.000000	30.000000	98.000000	3.000000	2.500000	3.000000	101.000000	0.000000	0.000000	0.00000	1.000000	1.000000
max	67.000000	43.000000	224.000000	4.000000	10.000000	3.000000	635.000000	1.000000	1.000000	1.00000	1.000000	1.000000

Dataset describe:

- We can see in the describe cell above:
  + The min value of the Experience column is -3, but we know the Experience values must be positive (>0). So, we will change any negative Experience value by the mean.

```
df['Experience'][df['Experience'] < 0] = df['Experience'].mean()
df.describe()
```

The describe database after changing any negative value in the Experience column.

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	personal_loan	securities_account	cd_account	Online	CreditCard
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5000.00000	5000.000000	5000.000000
mean	45.338400	20.328688	73.774200	2.396400	1.937938	1.881000	56.498800	0.096000	0.104400	0.06040	0.596800	0.294000
std	11.463166	11.253009	46.033729	1.147663	1.747659	0.839869	101.713802	0.294621	0.305809	0.23825	0.490589	0.455637
min	23.000000	0.000000	8.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
25%	35.000000	11.000000	39.000000	1.000000	0.700000	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000
50%	45.000000	20.104600	64.000000	2.000000	1.500000	2.000000	0.000000	0.000000	0.000000	0.00000	1.000000	0.000000
75%	55.000000	30.000000	98.000000	3.000000	2.500000	3.000000	101.000000	0.000000	0.000000	0.00000	1.000000	1.000000
max	67.000000	43.000000	224.000000	4.000000	10.000000	3.000000	635.000000	1.000000	1.000000	1.00000	1.000000	1.000000

+ We will convert the CCAvg from monthly average to annual average like income column.

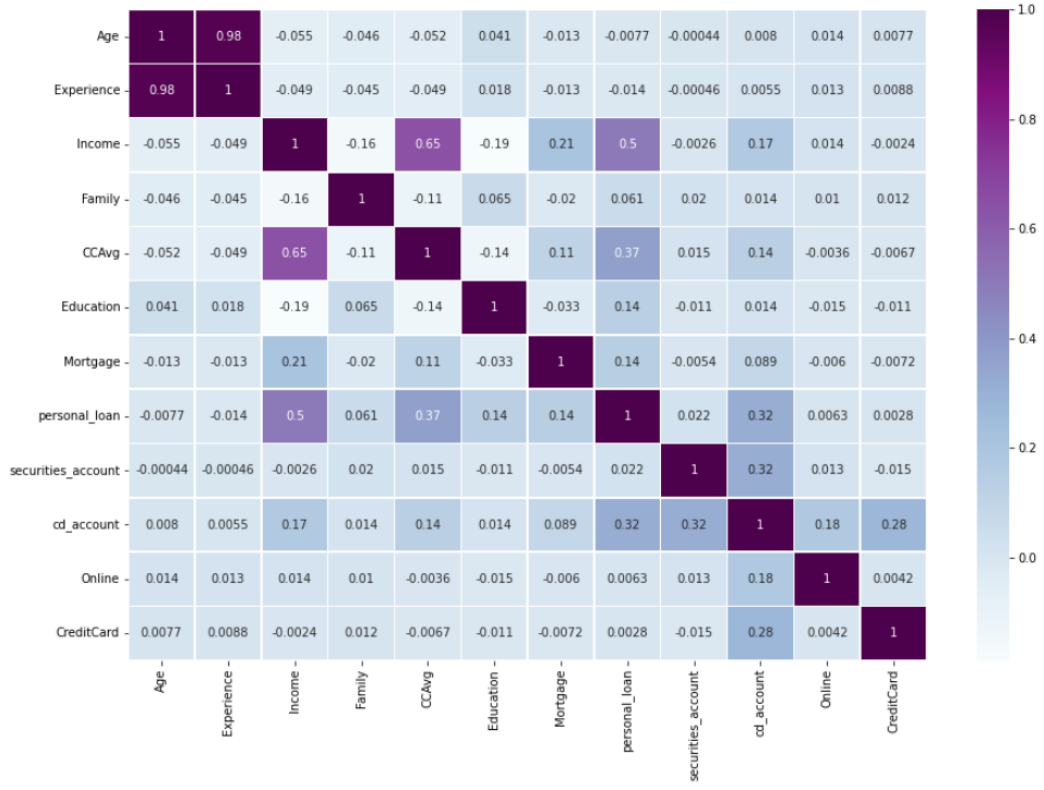
```
1  df['CCAvg'] = df['CCAvg']*12
2  df.head()
```

The dataset after convert:

	Age	Experience	Income	Family	CCAvg	Education	Mortgage	personal_loan	securities_account	cd_account	Online	CreditCard
0	25	1.0	49	4	19.2	1	0	0	1	0	0	0
1	45	19.0	34	3	18.0	1	0	0	1	0	0	0
2	39	15.0	11	1	12.0	1	0	0	0	0	0	0
3	35	9.0	100	1	32.4	2	0	0	0	0	0	0
4	35	8.0	45	4	12.0	2	0	0	0	0	0	1

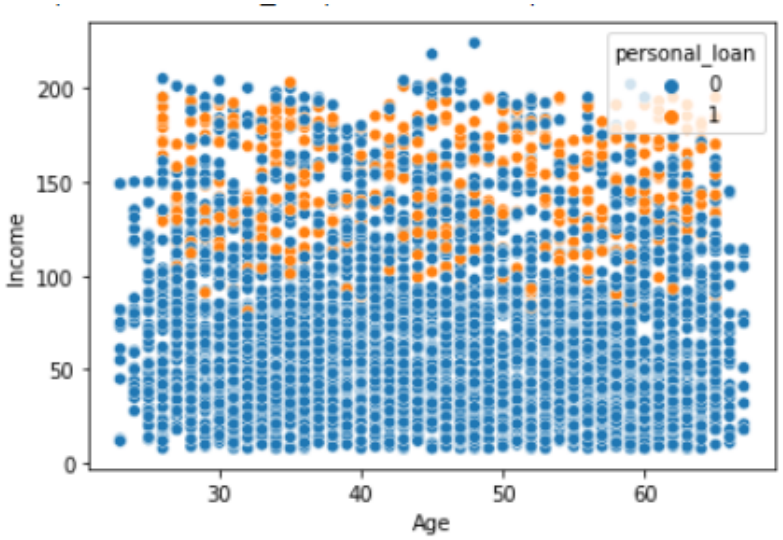
3. Data analysis

Plot correlation between columns.



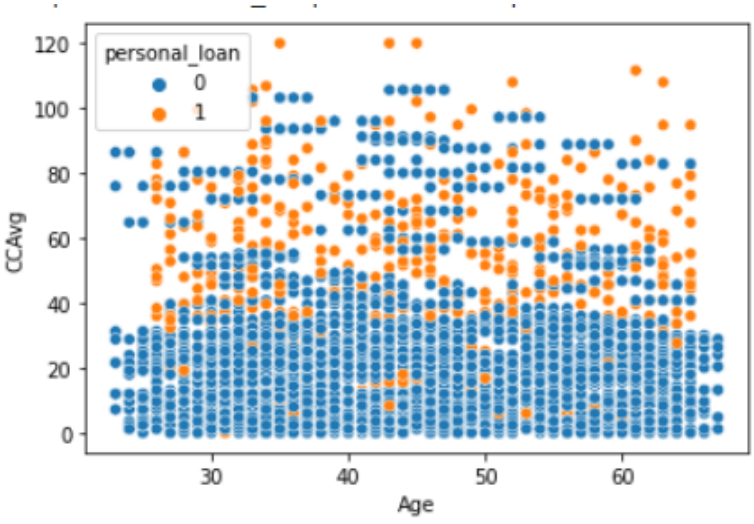
- We can see:
+ ‘Age’ and ‘Experience’ are highly correlated with each other, it almost 1.
+ ‘Income’ and ‘CCAvg’ correlated with each other.
+ ‘cd_account’ has correlation with ‘credit_card’, ‘securities_account’, ‘Online’, ‘CCAvg’, and ‘Income’.
+ ‘personal_loan’ has correlation with ‘Income’, ‘CCAvg’, ‘cd_account’, ‘Mortgage’, and ‘Education’.
+ ‘Mortgage’ has moderate correlation with ‘Income’.
+ ‘Income’ has correlation with ‘CCAvg’, ‘personal_loan’, ‘cd_account’, and ‘Mortgage’.

Plot scatter personal_loan between Age and Income.



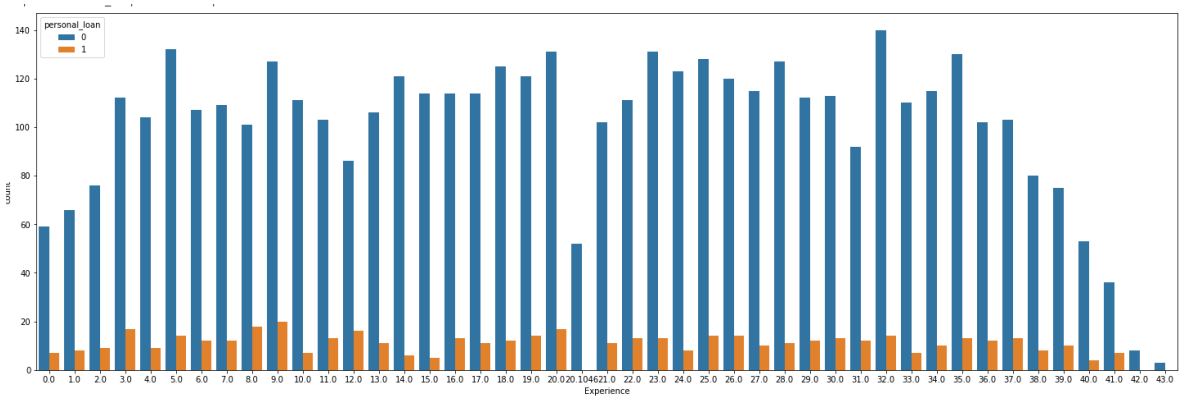
We can see, client with income more than 100k are more likely to get a loan.

Plot scatter personal_loan between Age and CCAvg.

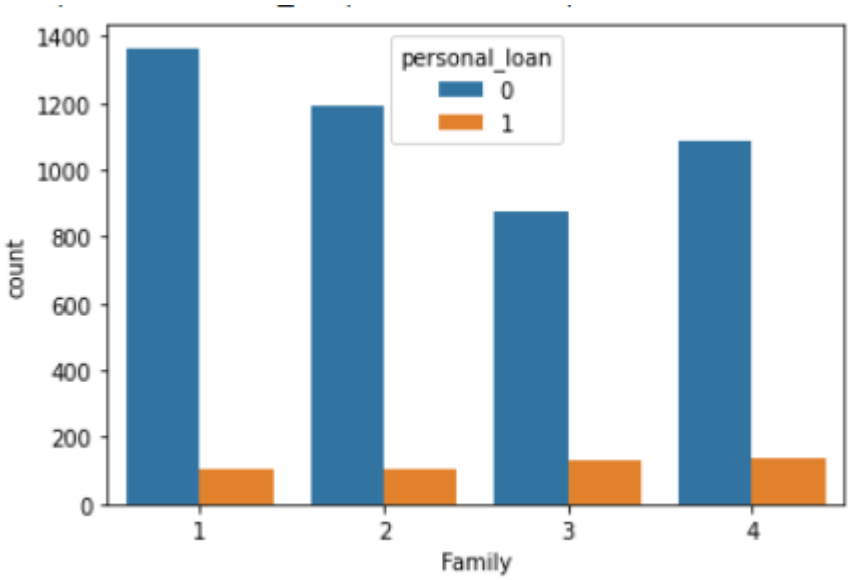


Clients with an annual CV average(CCAvg) more than 30 are more likely to get loans.

Plot count personal_loan by Experience.

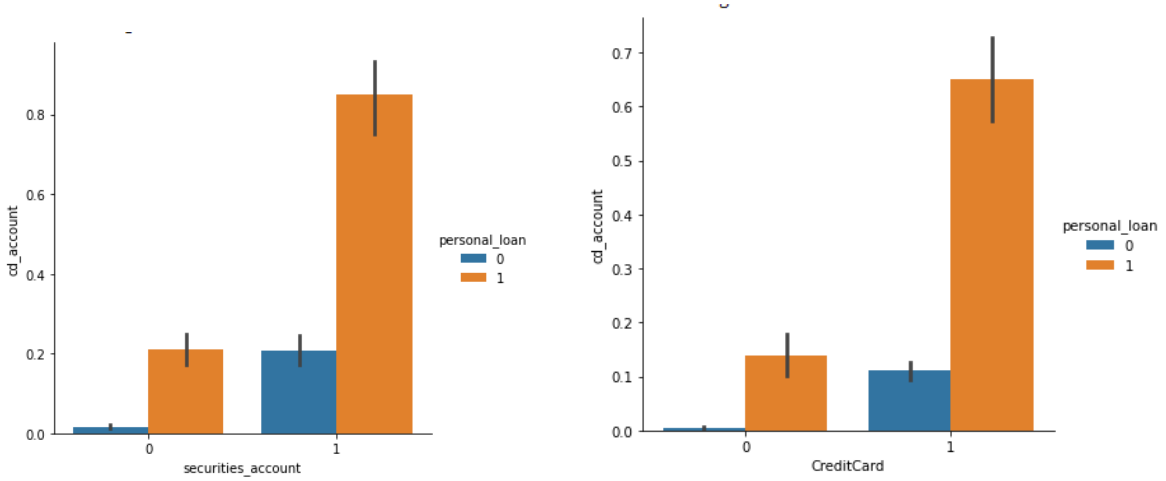


Plot count personal_loan by family



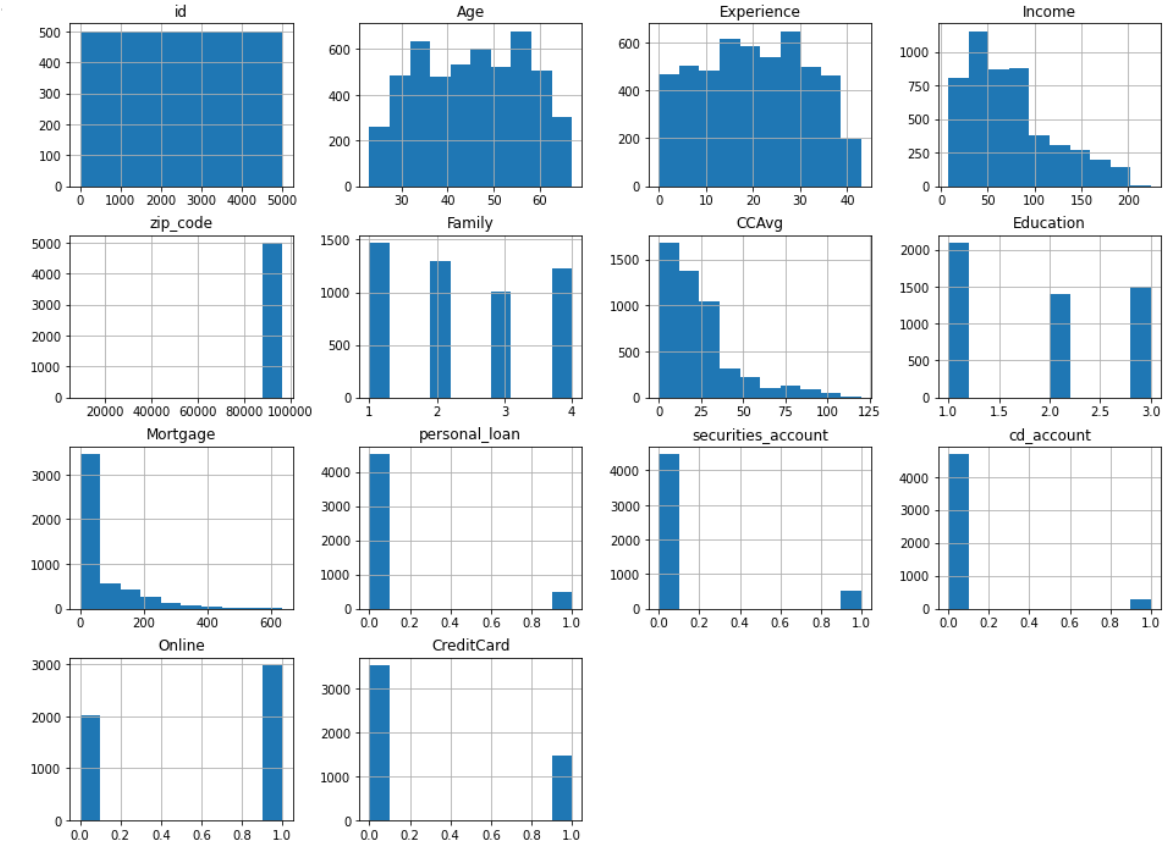
We can see in the previous two graphs the Family and Experience has low affect in the personal loan attribute. But, people with from 2 to 4 members are likely to take personal loans.

Count Plot cd_account with securities_account and CreditCard



From two graphs, we can see: It seems that customers with multiple bank accounts have higher creditworthiness, in addition, they are more likely to get a loan.

Plot hist map



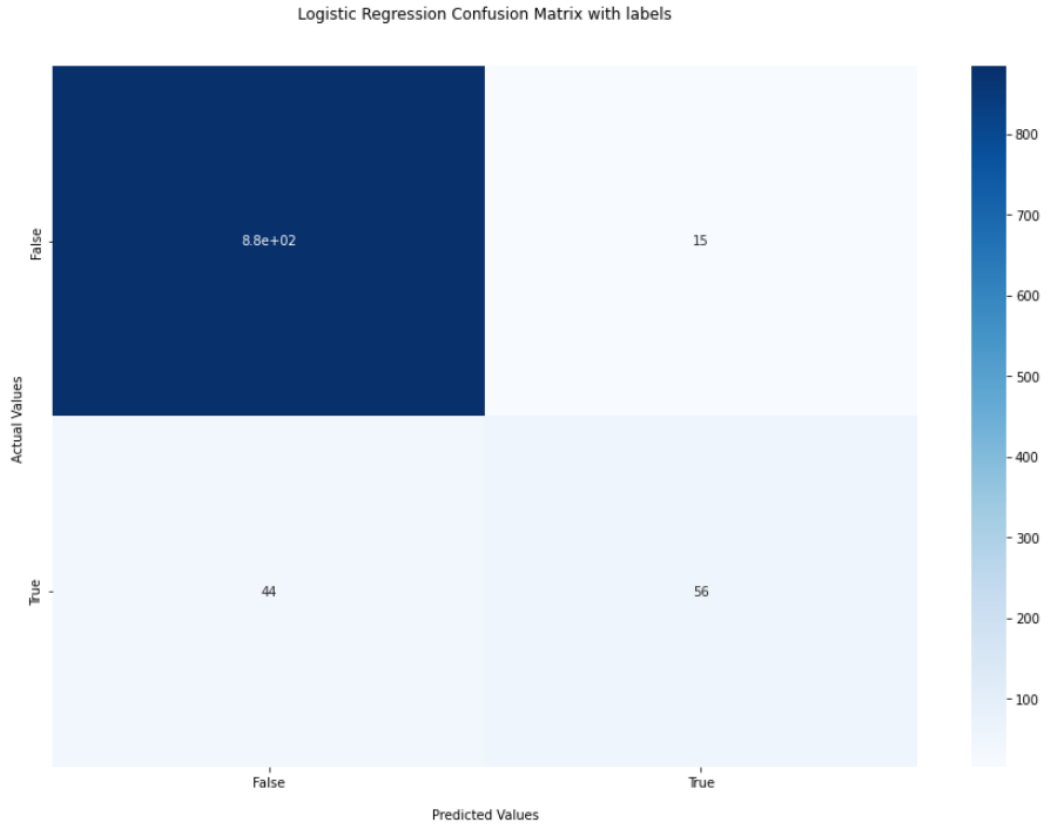
From hist map, we can see:

- The ‘Age’ column has an even distribution. Here we can see that the mean values between the ages are almost the same(mean and mid mean is almost same). The majority of customers are between the ages of 25 and 65 years.
- The ‘Experience’ column is normally distributed. The mean value here is also close to the mid mean value.
- The distribution of ‘Age’ and ‘Experience’ looks suspiciously similar. They might be correlated.
- ‘Income’ has a right skewed distribution.
- CCAvg also has a right skewed distribution. Most customers spend an average of 12k to 100k per year.
- ‘Mortgage’ distribution is also right skewed. The majority of individuals have mortgage loans under 40k.

IV. FINDINGS AND DISCUSSION

In this part, we will use accuracy score, confusion matrix, precision, f1-score to evaluate model quality in this problem. In addition, I will use the AUC-ROC indexes to compare models with each other. Then we will choose the best model for this problem.

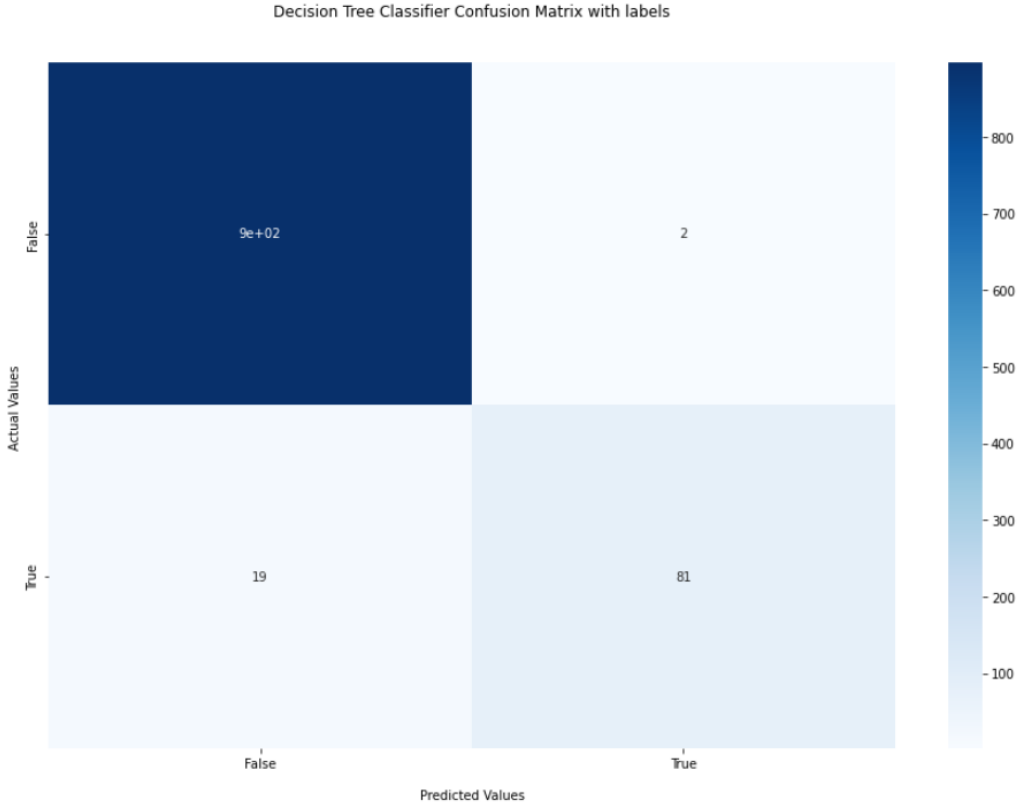
1. Logistic Regression



	precision	recall	f1-score	support
0	0.95	0.98	0.97	900
1	0.79	0.56	0.65	100
accuracy			0.94	1000
macro avg	0.87	0.77	0.81	1000
weighted avg	0.94	0.94	0.94	1000

The accuracy of Logistic Regression is 94%. F1 score is 65%. Recall is 56%, which means that, out of all the customers, they would actually buy the loan.

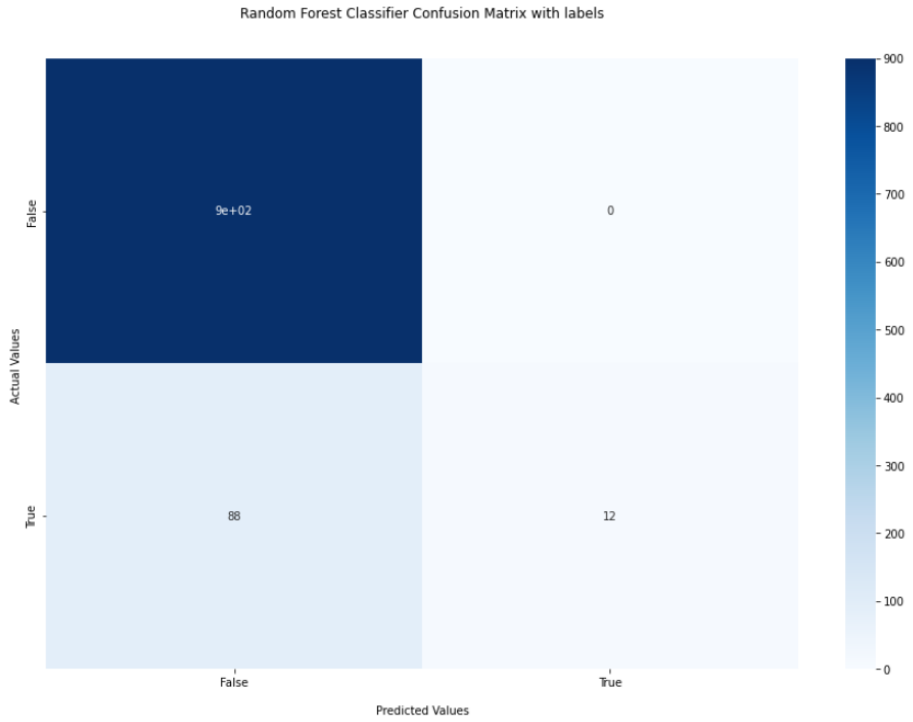
2. Decision Tree Classifier



	precision	recall	f1-score	support
0	0.98	1.00	0.99	900
1	0.98	0.81	0.89	100
accuracy			0.98	1000
macro avg	0.98	0.90	0.94	1000
weighted avg	0.98	0.98	0.98	1000

Same, accuracy of the Decision Tree Classifier is 98%. It is higher than Logistic Regression.F1 score and Recall are also higher with 89% and 81%.

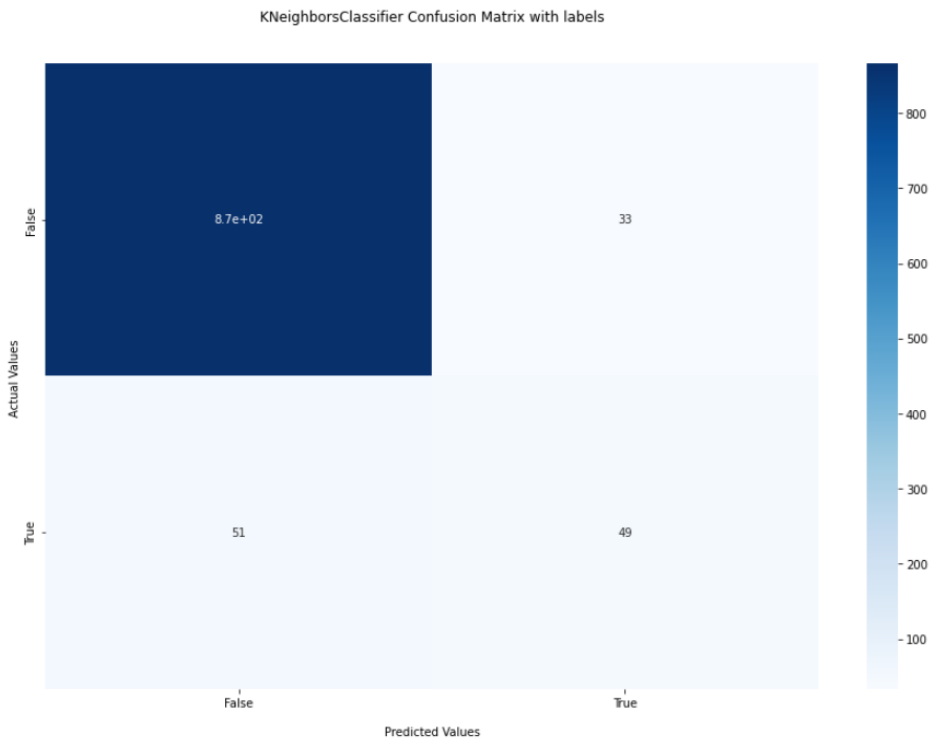
3. Random Forest Classifier



	precision	recall	f1-score	support
0	0.91	1.00	0.95	900
1	1.00	0.12	0.21	100
accuracy			0.91	1000
macro avg	0.96	0.56	0.58	1000
weighted avg	0.92	0.91	0.88	1000

The accuracy is 91%. It is less than Decision Tree and Logistic Regression. Recall and f1 score are low.

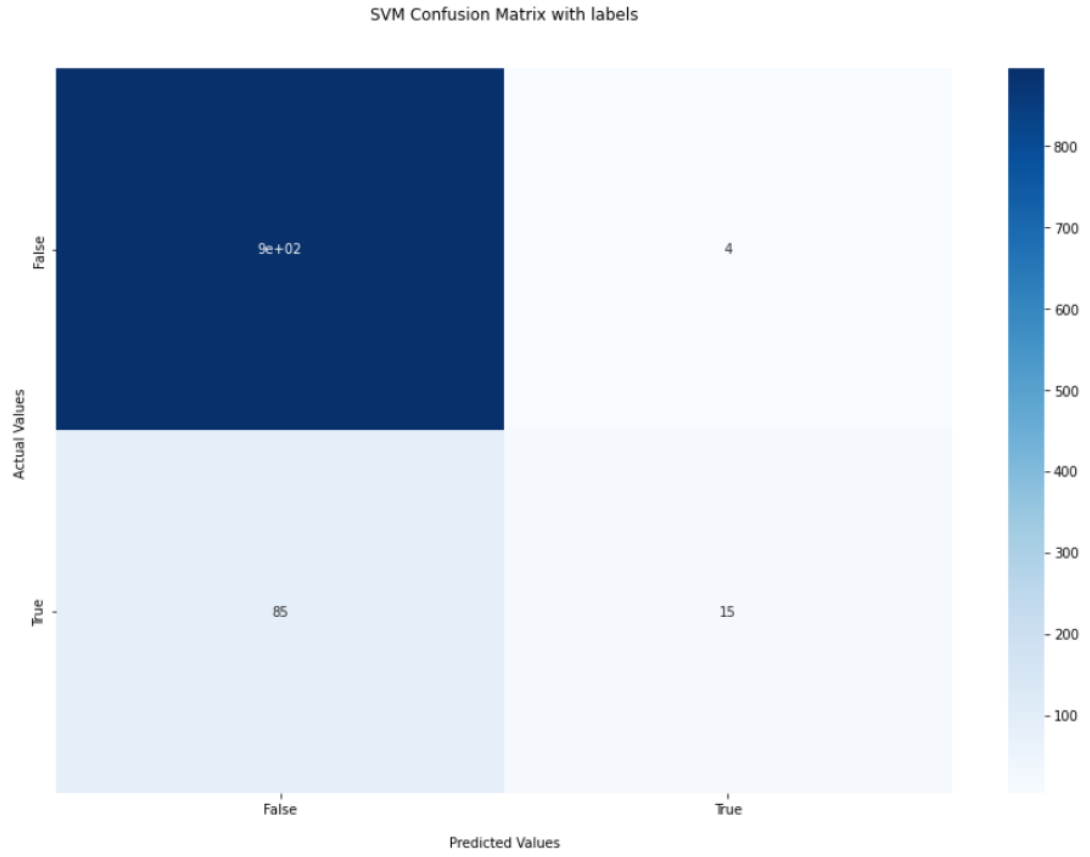
4. KNN



	precision	recall	f1-score	support
0	0.94	0.96	0.95	900
1	0.60	0.49	0.54	100
accuracy			0.92	1000
macro avg	0.77	0.73	0.75	1000
weighted avg	0.91	0.92	0.91	1000

The accuracy of KNN is higher than Random forest. But it is also less than Decision Tree. F1 score and recall are low.

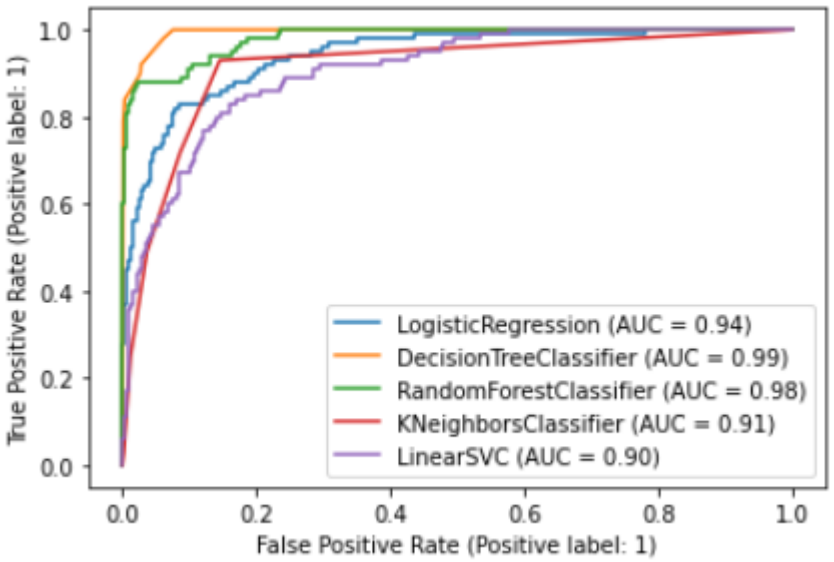
5. SVM



	precision	recall	f1-score	support
0	0.91	1.00	0.95	900
1	0.79	0.15	0.25	100
accuracy			0.91	1000
macro avg	0.85	0.57	0.60	1000
weighted avg	0.90	0.91	0.88	1000

SVM has accuracy 91%. Beside, recall and f1 score are also low too.

6. Compare models performance



We can see, Decision Tree is the model that has ROC-AUC highest at 99%.

V. Conclusion

As we are building models for prediction who will take personal loan. We can see, Decision Tree Classifier has accuracy and AUC of up to 98%. It means models predict a very high positive rate. So, it will help reduce the chance of failure. So Decision Tree is the best fit model to evaluate whether to lend to users or not. This will help the bank reduce risks of lending and bad debt.