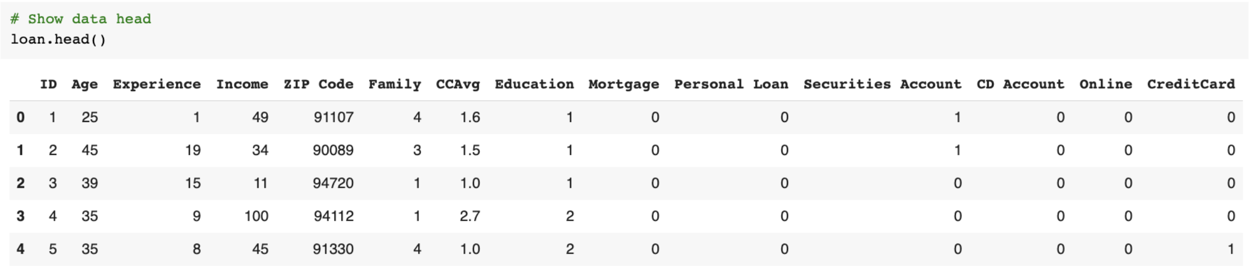
Loan prediction model

Group 10.

Adenuga, Temitope Oluwadamilola
Sahiti Priyanka Boyanapalli
Manisha Erukulla
Patala, Rohit Saurya
Ojha, Dhwani Bakulkumar

Thera Banks management wants to explore ways of converting its liability customers to personal loan customers while retaining them as depositors. This increases profitability and growth for the business, but personal loans are risky. A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

The data being used includes 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 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

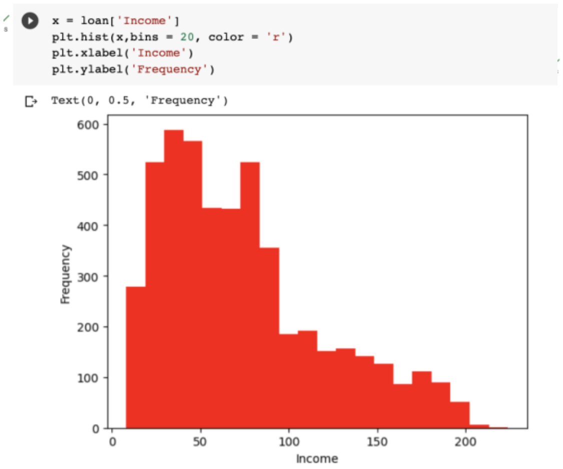


```
# Show data head
loan.head()
```

	ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
0	1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
1	2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	0	1

Summary Statistics: In this section we are studying the distribution of our data, looking for any correlations between variables, as well as the histograms of variables to understand its structure. The images below show the distribution of variables in the dataset.

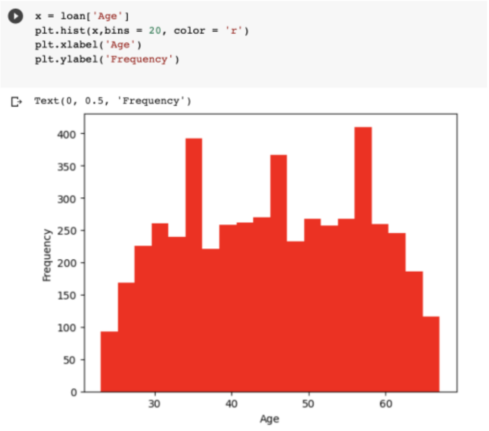
HISTOGRAM OF INCOME DISTRIBUTION



```
x = loan['Income']
plt.hist(x,bins = 20, color = 'r')
plt.xlabel('Income')
plt.ylabel('Frequency')
```

```
Text(0, 0.5, 'Frequency')
```

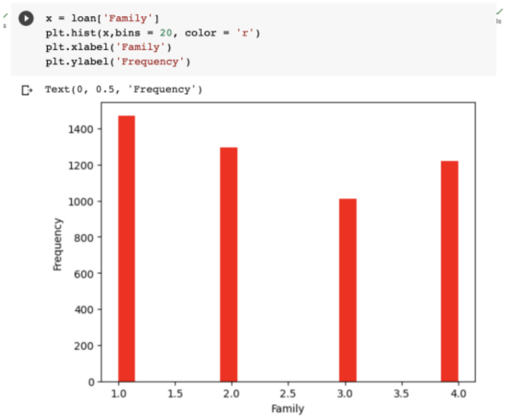
HISTOGRAM OF AGE DISTRIBUTION



```
x = loan['Age']
plt.hist(x,bins = 20, color = 'r')
plt.xlabel('Age')
plt.ylabel('Frequency')
```

```
Text(0, 0.5, 'Frequency')
```

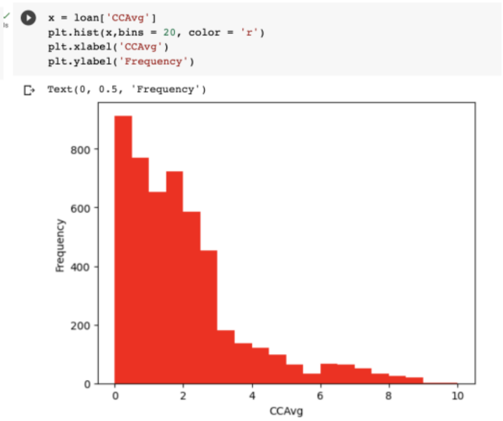
HISTOGRAM OF FAMILY SIZE DISTRIBUTION



```
x = loan['Family']
plt.hist(x,bins = 20, color = 'r')
plt.xlabel('Family')
plt.ylabel('Frequency')
```

```
Text(0, 0.5, 'Frequency')
```

HISTOGRAM OF AVERAGE CREDIT CARD SPEND DISTRIBUTION



```
x = loan['CCAvg']
plt.hist(x,bins = 20, color = 'r')
plt.xlabel('CCAvg')
plt.ylabel('Frequency')
```

```
Text(0, 0.5, 'Frequency')
```

Classification is one of the most commonly used techniques in machine learning, with the aim of classifying an input into one of several predefined categories. Classification models are perfect for this business problem because it allows us to drill down on customers with a higher chance of accepting the offer. Decision trees, logistic regression, k-means clustering, support vector machines (SVMs), and Naive Bayes are popular algorithms used in classification models. In this report, we will explore the implementation and performance of these algorithms on our dataset.

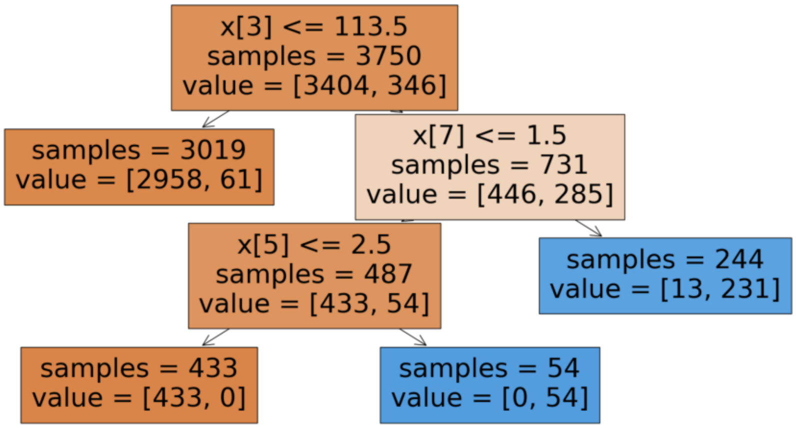
Data Preprocessing: In this step, we prepare the data for modeling by cleaning, transforming, and scaling it. The data in the banking dataset is already clean and well-formatted, but we did run code to ensure any duplicate values were removed and dropped columns with missing data. We also need to split the dataset into training and testing sets to evaluate the performance of the models

Model Building: We will now build five classification models using the following algorithms: decision trees, logistic regression, k-means clustering, SVMs, and Naive Bayes.

Decision Trees

Decision trees are a popular algorithm for classification tasks. They work by recursively partitioning the data into subsets based on the values of the input features. At each step, the algorithm selects the feature that provides the most information gain, which is a measure of how much the feature reduces uncertainty about the class label. We built a decision tree model using the scikit-learn library in Python. The accuracy of the model was 98.05% meaning the decision tree algorithm correctly identified which customers would accept the personal loan offer.

Decision tree logic with X[3] – Experience as the beginning node.

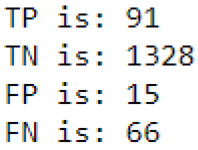


Logistic Regression

Logistic regression is a statistical technique used to analyze data that has a binary outcome, meaning that the dependent variable can take on only two possible values, such as 0 or 1. It is a type of regression analysis that models the probability of a certain outcome based on one or more predictor variables. In logistic regression, the dependent variable is modeled as a function of one or more independent variables using a logistic function.

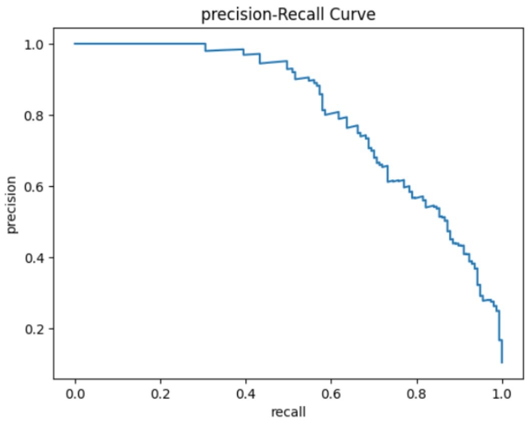
We built a logistic regression model using the scikit-learn library in Python and used Personal Loan as the variable where 0= Not Approved and 1= Approved.

We created a confusion matrix (by importing it from sklearn.metrics) which is used to evaluate the performance of the classification model and got the following values- TP = True Positive; TN = True Negative; FP=False Positive; FN=False Negative

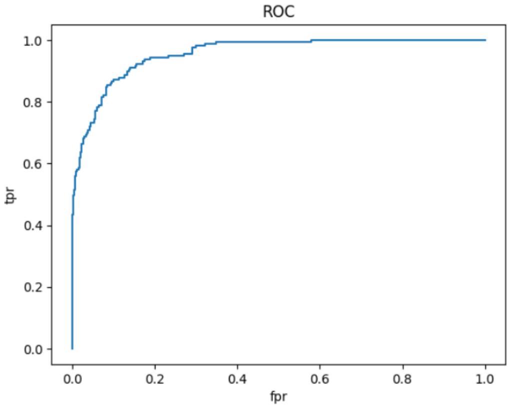


```
TP is: 91
TN is: 1328
FP is: 15
FN is: 66
```

Precision- Recall Curve (Recall score vs precision score).



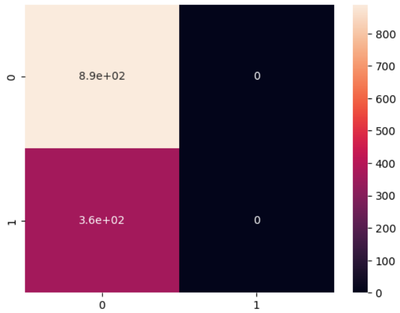
The ROC curve (False positive rate vs True positive rate)



SVMs

SVMs are a powerful algorithm for classification tasks. They work by finding the hyperplane that maximally separates the instances of different classes. The hyperplane is chosen in such a way that the margin between the closest instances of different classes is maximized. We built an SVM model using the scikit-learn library in Python.

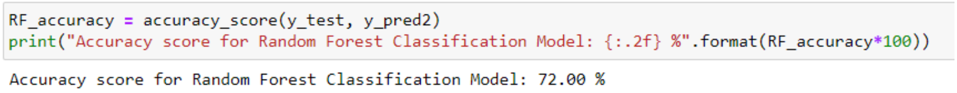
This is the heatmap that we got after running the SVM model and the accuracy score for the model as 71.12%.



Random Forest

Random forest is a popular machine learning algorithm used for classification, regression, and other tasks. It is an ensemble learning technique that combines multiple decision trees to create a more robust and accurate predictive model.

We imported the RandomForestClassifier from sklearn-kit and got an accuuracy of **72%**.



```
RF_accuracy = accuracy_score(y_test, y_pred2)
print("Accuracy score for Random Forest Classification Model: {:.2f} %".format(RF_accuracy*100))
```

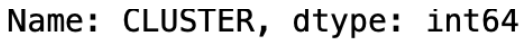
```
Accuracy score for Random Forest Classification Model: 72.00 %
```

Unsupervised Learning – Kmeans Clustering

Despite it not being a strict requirement, we were quite invested in the idea of exploring some Unsupervised Learning for our Project. Employing a method like Clustering points and brings together records that share an association with multiple fields. For our use case, these records the model groups up essentially give us customer archetypes among the Thera Bank userbase. Furthermore, we wanted to look specifically at customers linked to being Personal Loan buyers, as these are the people Thera Bank is most interesting in knowing. By looking at any clusters associated with Personal Loan purchases, we will find the other demographics and characteristics these buyers are associated with. This information is invaluable in helping Thera Bank understand their customers and can better educate their decisions, whether it is in their operations or their marketing.

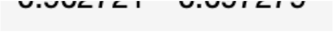
We used KMeans Clustering on the dataset. After a lengthy pre-processing period, we tested different cluster count values to find the one that was clearly distinct from the others and had a clear association to Personal Loan purchasing. We took the average values of all the records in each cluster to get the association with each characteristic.

We started with 3 clusters and found that they were relatively even in record distribution.



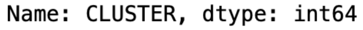
```
Name: CLUSTER, dtype: int64
```

Personal Loan association was even around 9% for each cluster, which is about the same as the purchase rate in the original dataset (9.6%).



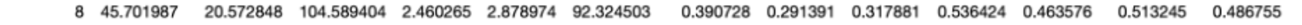
This wasn't of much use to us as the characteristics associated with these clusters don't belong to people especially inclined to buy Personal Loans. In addition, there were no clear differences among the other attributes between each cluster.

We tested more values, and it was with 8 clusters that we managed to find the largest distinction between customer with Personal Loan purchases and those without.

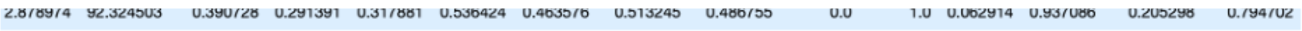


```
Name: CLUSTER, dtype: int64
```

Cluster 8, populated by 302 customers, had the strongest association with Personal Loan purchases at 0.46.

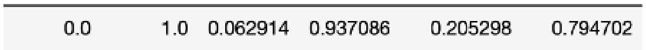


```
8   45.701987   20.572848   104.589404   2.460265   2.878974   92.324503   0.390728   0.291391   0.317881   0.536424   0.463576   0.513245   0.486755
```

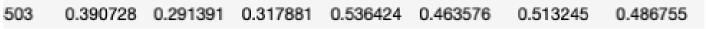


```
2.878974   92.324503   0.390728   0.291391   0.317881   0.536424   0.463576   0.513245   0.486755   0.0   1.0   0.062914   0.937086   0.205298   0.794702
```

There was a steady increase in Personal Loan purchase rate as we added more clusters, and this stopped at k=8. So we took a closer look at this cluster for other values that varied from the remaining clusters.



```
0.0   1.0   0.062914   0.937086   0.205298   0.794702
```



```
503   0.390728   0.291391   0.317881   0.536424   0.463576   0.513245   0.486755
```

We found that the Annual Income for customers in this cluster was notably higher compared to the other clusters, at $104,589.40 per year. Credit card spending per month was higher as well, at $2,878.97 per month. This cluster also consisted of customers with mortgages worth more than average, at $92,324.50. These customers had a very strong association with being Online Banking users and being owners of Thera Bank credit cards at values of 93.71% and 79.47% respectively.

The association with Personal loan purchasing was 46.35%. Compared to the average success rate, this is quite an increase from 9.6%.

What does this mean for Thera Bank? Our conclusion is that if they were to pick two customers among their userbase, one at random and one fitting the characteristics outlined by cluster 8, the customer is nearly five times more likely to purchase a personal loan.

Thera Bank can use this information when choosing potential customers and targeting existing ones through their marketing.

Conclusion.

The models used to classify our business problem showed very satisfactory accuracy scores but we decided to select the decision tree model because of the added benefit of knowing the logic of the model and being able to visualize it. Although the Random Forest model had an accuracy of 98.16% we believe the benefit of understanding the logic is a bit more valuable in achieving the business objective.