

Q. Understanding the attributes - Find relationship between different attributes (Independent variables) and choose carefully which all attributes have to be a part of the analysis and why?

A. For the purpose of the analysis the following variables shall not be considered:

1. ID - Since ID does not have any relation to the user accepting personal loan.
2. ZIP Code – Since ZIP Code has a very high value of mode for a given value it's effect is negligible.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Age..in.years.	1	0.03	0.03	0.486	0.485915	
Experience..in.years.	1	0.00	0.00	0.019	0.891438	
Income..in.K.month.	1	109.78	109.78	2055.229	< 2e-16	***
Family.members	1	8.68	8.68	162.442	< 2e-16	***
CCAvg	1	1.33	1.33	24.932	6.14e-07	***
Education	1	22.93	22.93	429.216	< 2e-16	***
Mortgage	1	0.66	0.66	12.318	0.000453	***
Securities.Account	1	0.23	0.23	4.234	0.039673	*
CD.Account	1	20.45	20.45	382.927	< 2e-16	***
Online	1	0.68	0.68	12.685	0.000372	***
CreditCard	1	1.89	1.89	35.355	2.94e-09	***
Residuals	4970	265.48	0.05			


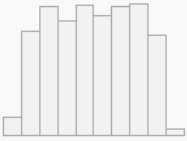
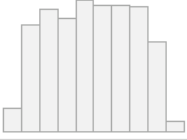
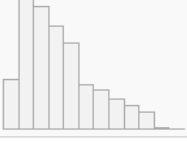

We can see from the above data that all the variables are important and contributing to the analysis.


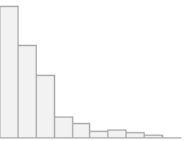


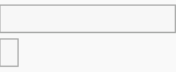

1. Income is an important factor with a very small p-value.
2. No. of family members is also very important.
3. CCAvg is also significant.
4. Education also contributes heavily.
5. Mortgage is a significant variable.
6. Securities Account is significant but not as much as the other variables.
7. CD Account is also an important factor.
8. Online is also important.
9. CreditCard is also an indicative variable.
10. Age and experience are not so significant.




Q. Exploratory Data analysis

Below we have looked at the different variables and their trends for the given dataset.

1. ID is insignificant variable and is useful only as an identifier.
2. Age has a normal distribution and is ranging from 23 to 67.
3. Experience is almost similar to age since both these factors generally go hand in hand.
4. Income which shows an approximately normal distribution is also an important contributing factor.
5. ZIP Code is highly skewed and does not significantly contribute to the analysis.
6. No. of family members is almost uniformly distributed and also contributes to the analysis.
7. CCAvg is heavily skewed to the right but is a significant contributor to our target variable.
8. Education I skewed to the left and is an important contributor to the target variable.
9. Mortgage is heavily skewed to the left and is an important factor.
10. Personal loan is the target variable and there are only 9.6% personal loan entries. Thus, the dataset is unbalanced.
11. Securities account, online account, CD account and CreditCard are categorical variables that are contributing to the target variable.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	ID [numeric]	Mean (sd) : 2502.4 (1441.8) min < med < max: 1 < 2502.5 < 5000 IQR (CV) : 2495.5 (0.6)	4982 distinct values		4982 (100%)	0 (0%)
2	Age (in years) [numeric]	Mean (sd) : 45.3 (11.5) min < med < max: 23 < 45 < 67 IQR (CV) : 20 (0.3)	45 distinct values		4982 (100%)	0 (0%)
3	Experience (in years) [numeric]	Mean (sd) : 20.1 (11.5) min < med < max: -3 < 20 < 43 IQR (CV) : 20 (0.6)	47 distinct values		4982 (100%)	0 (0%)
4	Income (in K/month) [numeric]	Mean (sd) : 73.7 (46) min < med < max: 8 < 64 < 224 IQR (CV) : 59 (0.6)	162 distinct values		4982 (100%)	0 (0%)
5	ZIP Code [numeric]	Mean (sd) : 93152.5 (2123) min < med < max: 9307 < 93437 < 96651 IQR (CV) : 2697 (0)	467 distinct values		4982 (100%)	0 (0%)

6	Family members [numeric]	Mean (sd) : 2.4 (1.1) min < med < max: 1 < 2 < 4 IQR (CV) : 2 (0.5)	1: 1464 (29.4%) 2: 1292 (25.9%) 3: 1009 (20.2%) 4: 1217 (24.4%)		4982 (100%)	0 (0%)
7	CCAvg [numeric]	Mean (sd) : 1.9 (1.7) min < med < max: 0 < 1.5 < 10 IQR (CV) : 1.8 (0.9)	108 distinct values		4982 (100%)	0 (0%)
8	Education [numeric]	Mean (sd) : 1.9 (0.8) min < med < max: 1 < 2 < 3 IQR (CV) : 2 (0.4)	1: 2088 (41.9%) 2: 1399 (28.1%) 3: 1495 (30.0%)		4982 (100%)	0 (0%)
9	Mortgage [numeric]	Mean (sd) : 56.5 (101.8) min < med < max: 0 < 0 < 635 IQR (CV) : 101 (1.8)	347 distinct values		4982 (100%)	0 (0%)
10	Personal Loan [numeric]	Min : 0 Mean : 0.1 Max : 1	0: 4504 (90.4%) 1: 478 (9.6%)		4982 (100%)	0 (0%)
11	Securities Account [numeric]	Min : 0 Mean : 0.1 Max : 1	0: 4463 (89.6%) 1: 519 (10.4%)		4982 (100%)	0 (0%)

12	CD Account [numeric]	Min : 0 Mean : 0.1 Max : 1	0: 4682 (94.0%) 1: 300 (6.0%)		4982 (100%)	0 (0%)
13	Online [numeric]	Min : 0 Mean : 0.6 Max : 1	0: 2013 (40.4%) 1: 2969 (59.6%)		4982 (100%)	0 (0%)
14	CreditCard [numeric]	Min : 0 Mean : 0.3 Max : 1	0: 3517 (70.6%) 1: 1465 (29.4%)		4982 (100%)	0 (0%)

Q. Splitting data in Train and Test dataset

A. We have split the data into a training dataset and a testing dataset to develop the model and then we apply the model to our hold out sample. The dataset has been split thus:

Training set

```
table(trainingset$Personal.Loan)
  0    1
3152 336
> prop.table(table(trainingset$Personal.Loan))
      0      1
0.90366972 0.09633028
```

The training set has 3488 records. It also has a success ratio of 9.6%, which is very close to the actual dataset.

Testing set

```
table(testingset$Personal.Loan)
  0    1
1352 142
> prop.table(table(testingset$Personal.Loan))
      0      1
0.90495315 0.09504685
```

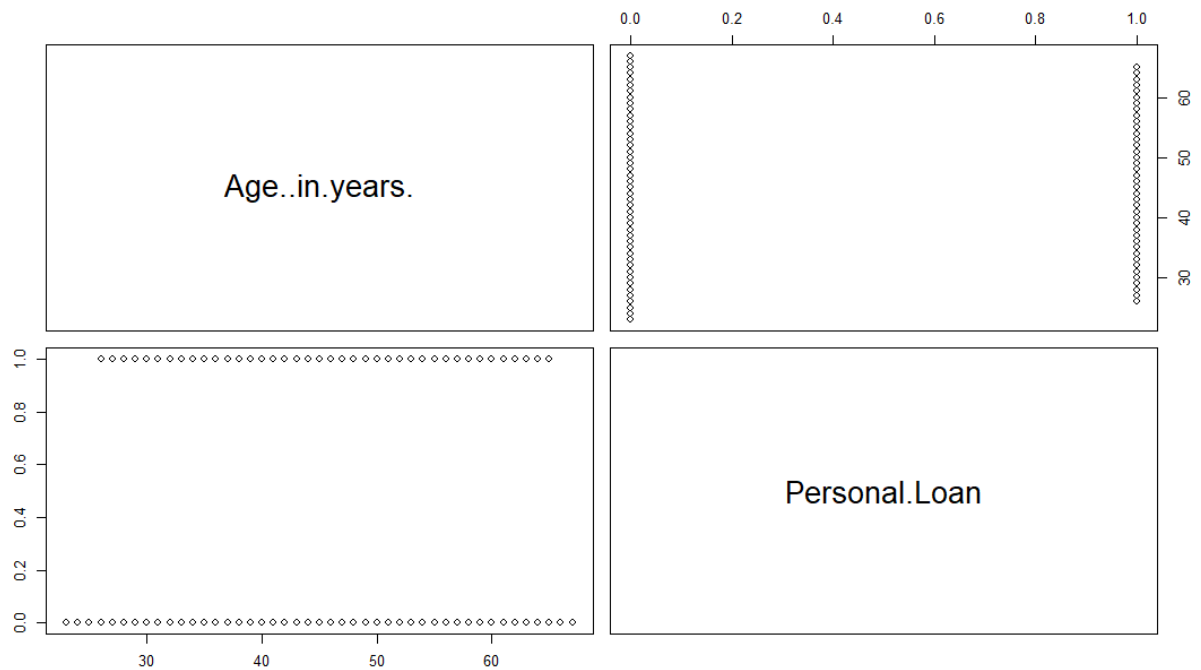
The testing set has 1494 records. It also has a success ratio of 9.5%, which is very close to the actual dataset.

Hence, we can use both these datasets for our training and validation and also for the model performance measures.

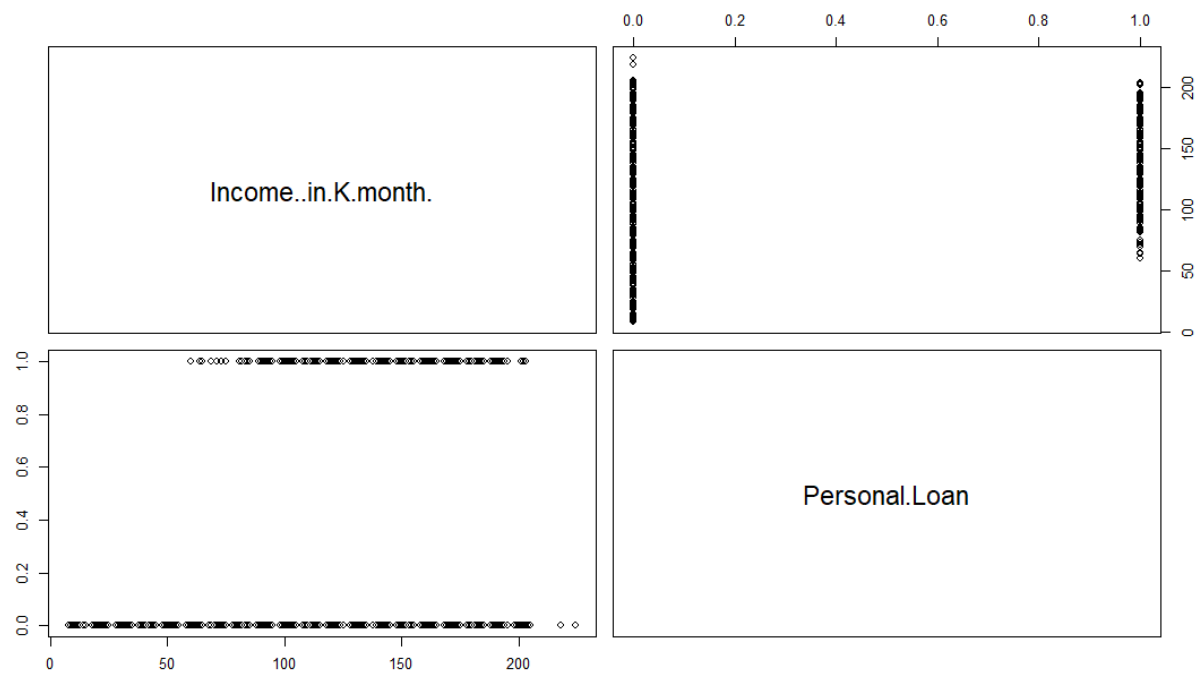
Q. Some Charts and Graphs to show case the relationship between Independent and Dependent Variables

A. We can see that a number of factors hugely impact the people who have taken a personal loan. The graphs below show how there is a strong relationship between many variables such as income, family members, education etc. and how they contribute to people taking loans.

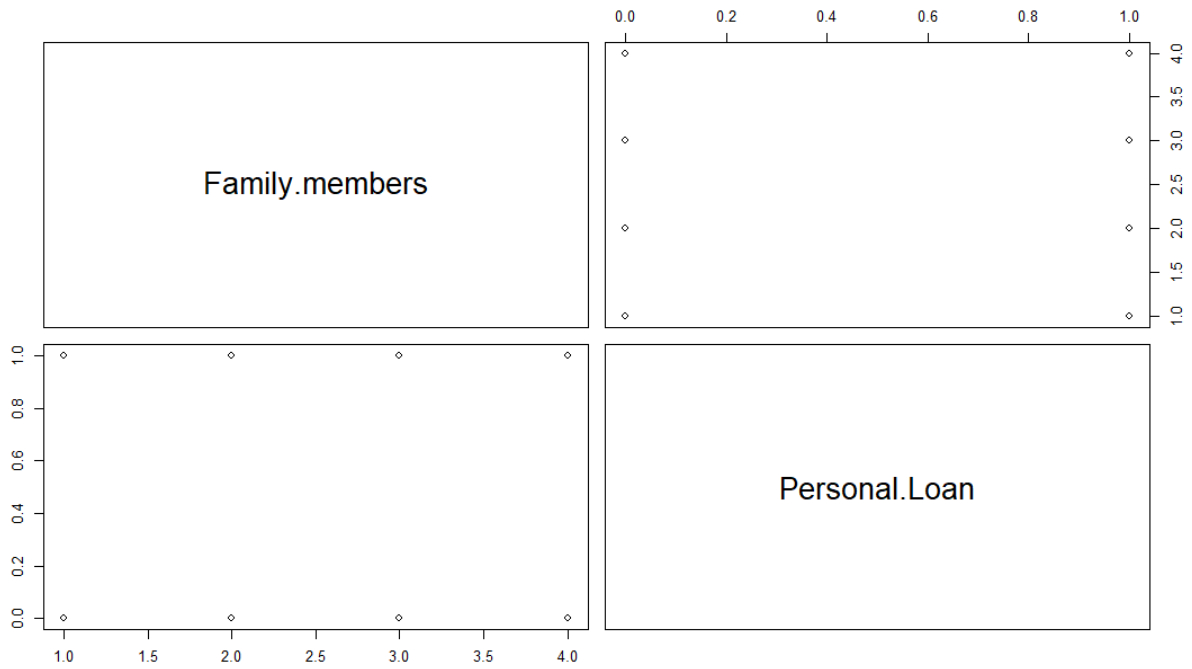
The following graph shows the relationship between age and personal loan.



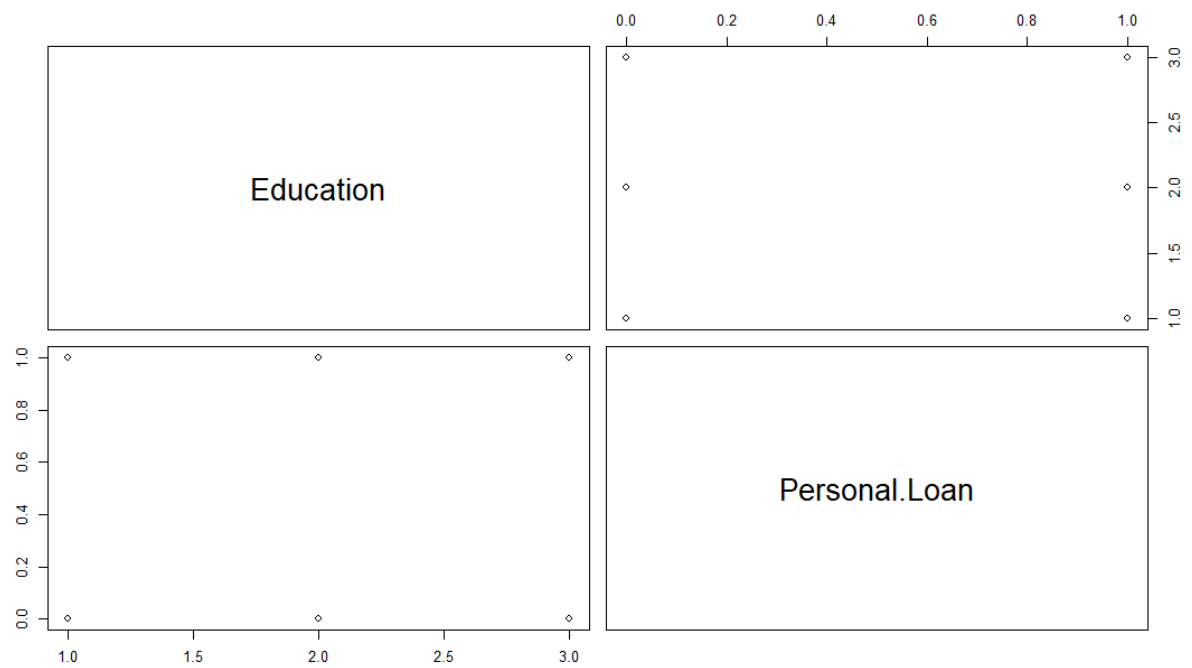
The following graph shows the relation between income and personal loan.



The following graph shows the relation between family members and personal loan.



The graph below shows the relation between education and personal loan.



Q. Model Development (Any one of the below techniques to be used)

- Random Forest
- CART

A. We shall use the random forest to build our given model.

```
randomForest(formula = as.factor(trainingset$Personal.Loan) ~
= trainingset[, c(-1, -5)], ntree = 201, mtry = 3, nodesize = 70, imp
ortance = TRUE)
Type of random forest: classification
Number of trees: 201
No. of variables tried at each split: 3

OOB estimate of error rate: 2.04%
Confusion matrix:
      0      1 class.error
0 3147      5 0.001586294
1   66 270 0.196428571
```

The given model was successful in building a good random forest. We have achieved an OOB rate of 2.04%.

Q. Model Performance Measures

A. To see the success of the model we shall see the following measures.

1. OOB rate: For the given model we have achieved an OOB rate of 2.04%, which shows that the model is successful in predicting the consumer behaviour.
2. GINI: For the given model we have obtained a GINI of 0.883 which shows that the model is highly successful and also very robust.

```
> gini
[1] 0.8828758
```

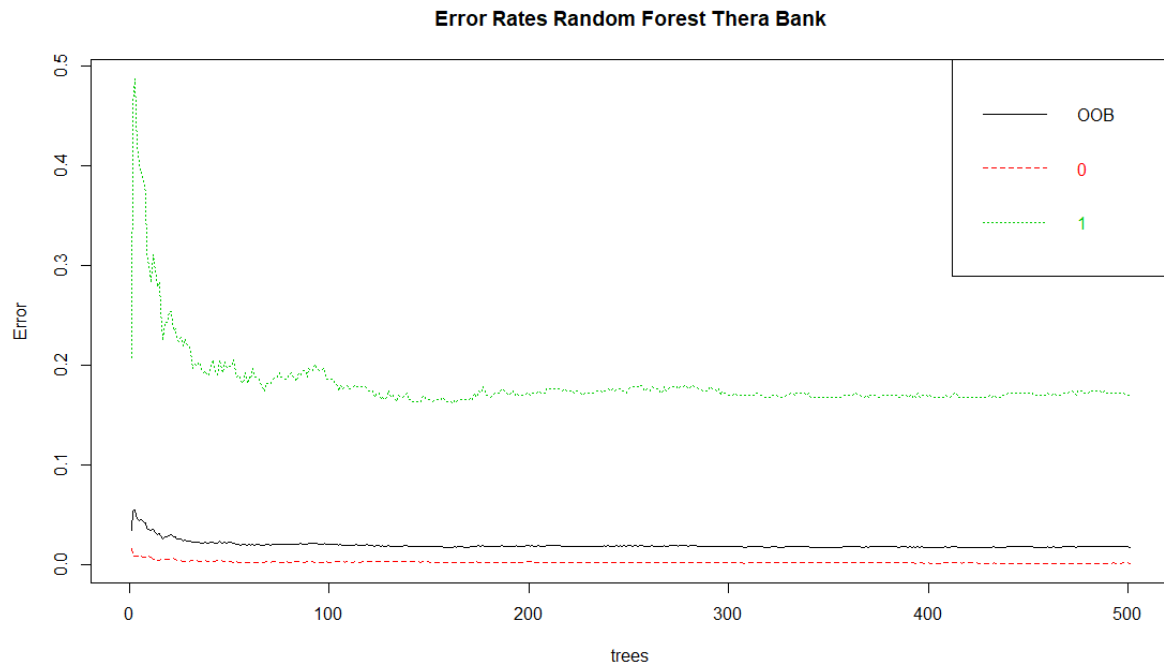
3. AUC: The area under curve is another measure that shows the success of the model. For our model we have obtained an AUC of 0.996. This again shows that the model is highly robust.

```
auc
[1] 0.9956641
```

Hence, we can conclude that the model is highly robust and successful.

Q. Validation of Model

A. Once we have formed the random forest model, we can check the success of the model by tuning it to a useful form. Below we have plotted a graph of error vs no. of trees. We can see that beyond the 200th tree there is no considerable reduction in the error and hence we can build a model with 201 trees. We will now tune the model to see the most suitable number of mtry variables for the trees.



```
mtry = 3  OOB error = 0%
Searching left ...
mtry = 2  OOB error = 0.04%
-Inf 1e-04
Searching right ...
mtry = 4  OOB error = 0%
```

Hence, we can see that for the given model the optimum no of trees is 3,4. Since the no. of variables under consideration is not too high, we have a wide range for optimum no. of variables.

Now we see whether the random forest shows the same contribution of variables as was seen earlier in the exploratory data analysis.

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Income..in.K.month.	56.22	47.61	58.30	225.88
Education	52.78	41.22	52.11	134.14
Family.members	40.11	28.93	39.95	67.76
CCAvg	17.57	16.80	20.19	97.77
CD.Account	10.46	10.31	13.95	37.80
CreditCard	4.30	2.39	5.90	1.32
Age..in.years.	5.03	3.88	5.79	4.57
Mortgage	8.62	-6.77	5.66	15.11
Experience..in.years.	4.49	1.68	4.86	3.80
Securities.Account	2.96	1.60	4.60	0.77
Online	0.10	1.22	1.04	0.52

Here we can see again that the contribution of all the variables to the consumer choice is in line with our expectations and the maximum contribution is from Income, education, family members etc, as can be seen in the output above.

Q. Model Performance on Hold Out Sample

A. Finally we apply our model to the hold out sample and see the behaviour of the model.

The model output for the hold out is as follows:

	0	1
0	1350	23
1	2	119

Here the rows represent the predicted class and the columns represent the actual.

As we can see the model produces an error rate of 1.6% for this holdout sample. Hence, we can conclude with confidence that the model is successful.