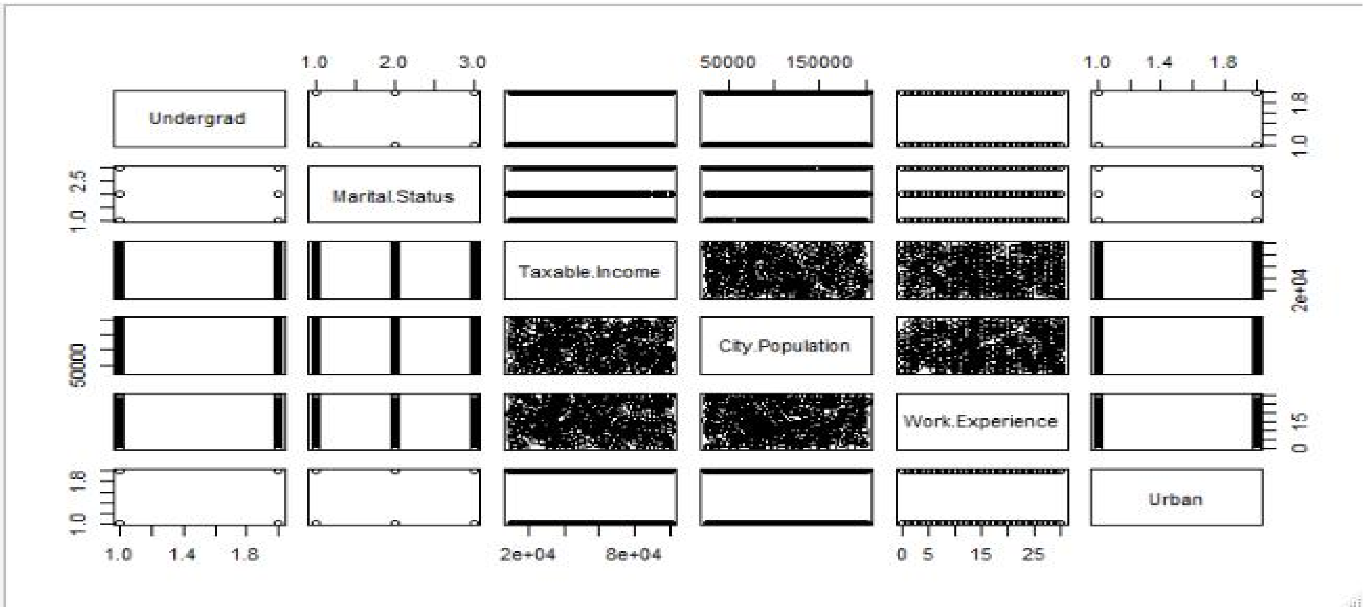
Random Forest

Example-Fraud Check Dataset

```
'data.frame':	600 obs. of  6 variables:
 $ Undergrad      : Factor w/ 2 levels "NO","YES": 1 2 1 2 1 1 1 2 1 2 ...
 $ Marital.Status : Factor w/ 3 levels "Divorced","Married",..: 3 1 2 3 2 1 1
3 3 1 ...
 $ Taxable.Income : int  68833 33700 36925 50190 81002 33329 83357 62774 8351
9 98152 ...
 $ City.Population: int  50047 134075 160205 193264 27533 116382 80890 131253
102481 155482 ...
 $ Work.Experience: int  10 18 30 15 28 0 8 3 12 4 ...
 $ Urban          : Factor w/ 2 levels "NO","YES": 2 2 2 2 1 1 2 2 2 2 ...
```

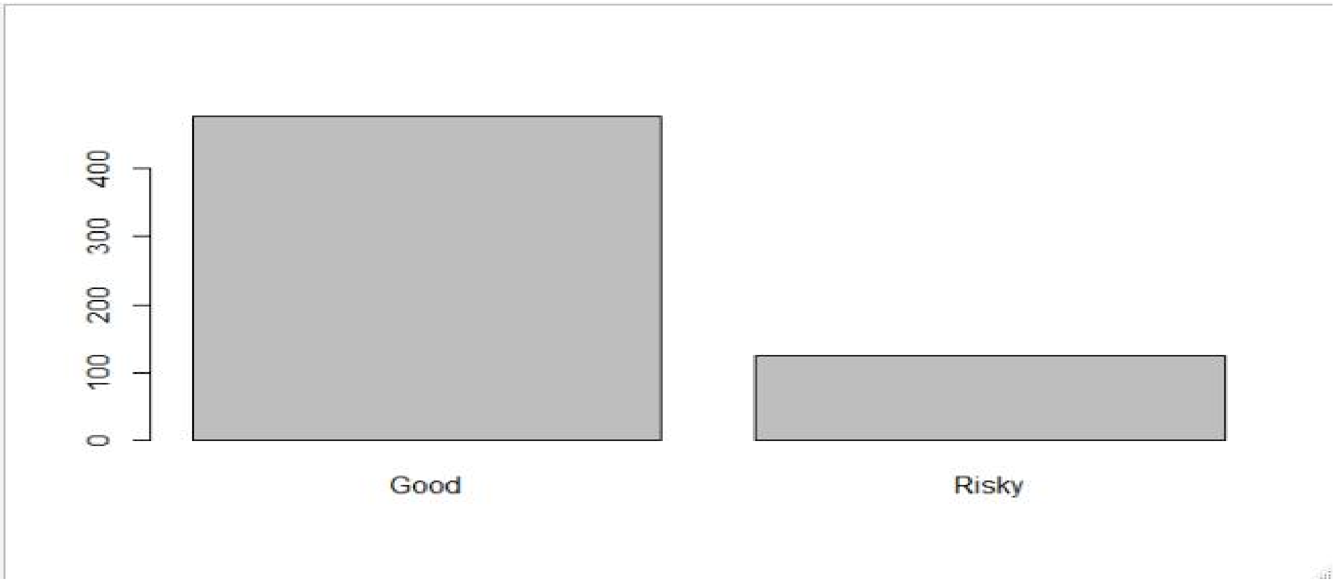
In the above data frame 3 variables are factors and rest all are numeric and target variable is Taxable.Income

Now we create another variable type, which is factor and contain desired results Good or Risky.



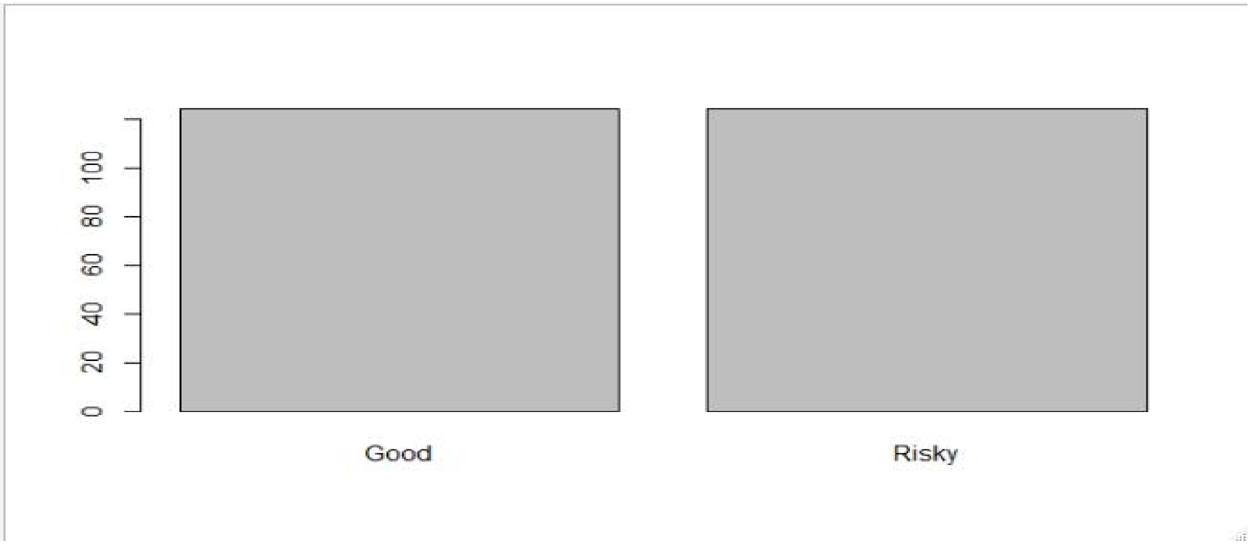
From the pairs plot, none of variable is correlated with our target variable Taxable.Income and uniform distributed scatter plots between all the numeric variable.

Treatment With Imbalanced Data →



```
 Good Risky
  476   124
```

From above plot, our target variable is imbalanced, so we will make ratio equal as 1.



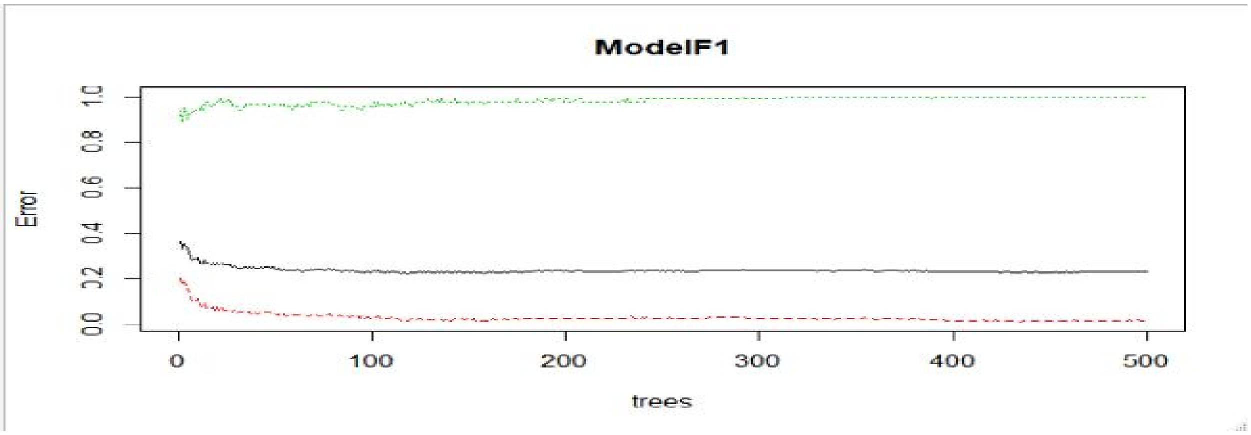
Now our data is equal in ratio.

Model-1 ➔

Confusion Matrix

```
        Predicted
Actual   Good Risky
  Good    144     3
  Risky    32     1
```

Accuracy ➔0.8055

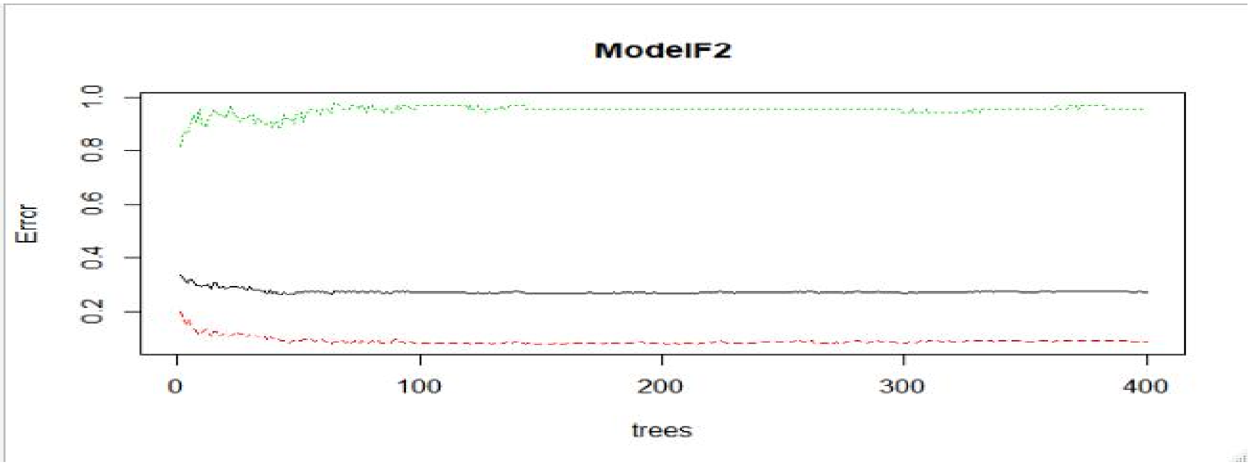


Model-2 ➔

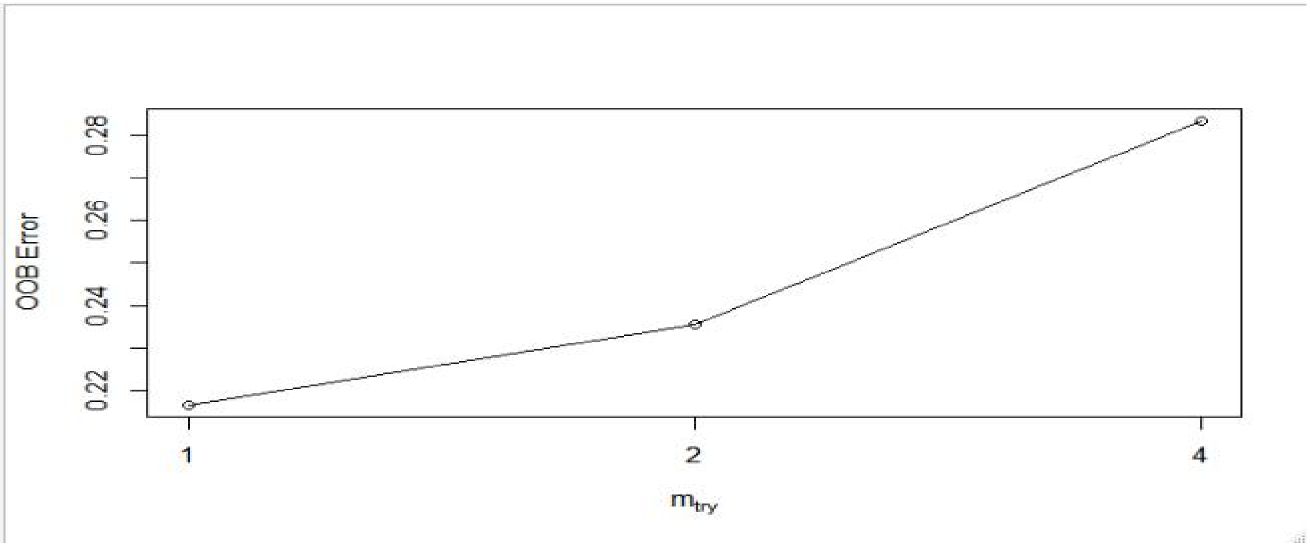
Confusion Matrix

```
        Predicted
Actual   Good Risky
  Good    131    16
  Risky    32     1
```

Accuracy ➔0.7333



Turing the Random Forest

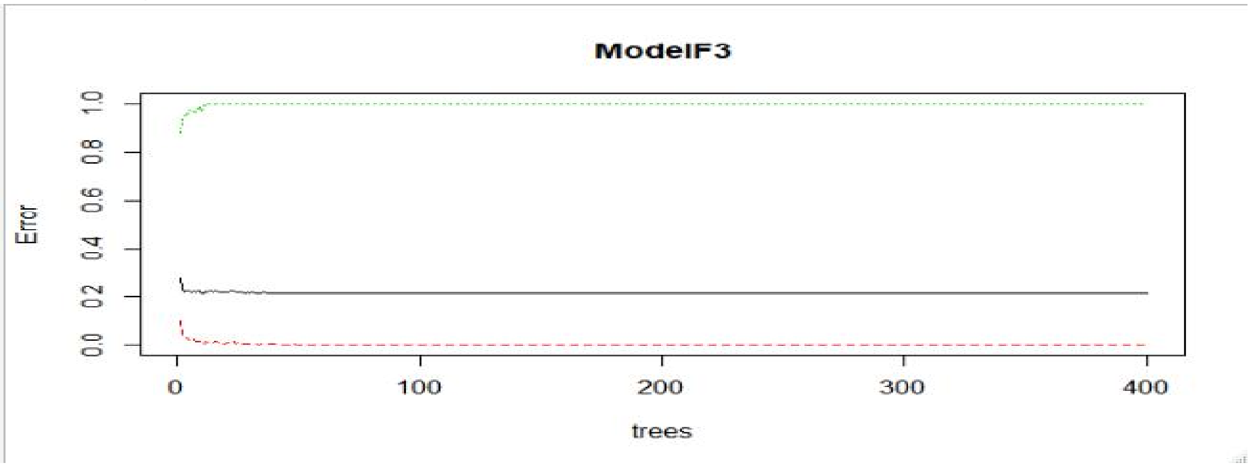


Model-3 ➔

Confusion Matrix

```
Predicted
Actual   Good Risky
  Good   147     0
  Risky   33     0
```

Accuracy ➔0.8166



From above information we can infer that Model-3 is good model with accuracy 81.66%.