CRM/ML/07 22.08.2021

“Credit Risk Modelling”

<u>Abstract</u>

The Credit Risk Model should accurately capture a customer’s payment behaviour. A customer which is current (good), will become delinquent and finally default. A good Model must predict the transition probabilities between these credit states, as accurately as possible.

A reliable & stable Model is one which discriminates well between Good/ Negligent/Default classes, remains stable and has acceptable out of sample performance.

The Customer behaviour is primarily driven by 3 major Factors.

1) Loan Variables
2) Customer Variables
3) Economic & Industrial Factors

There are generally more than 20 variables, affecting the customer payment behaviour.

Association between Predictors & Response variable. Fig1.1

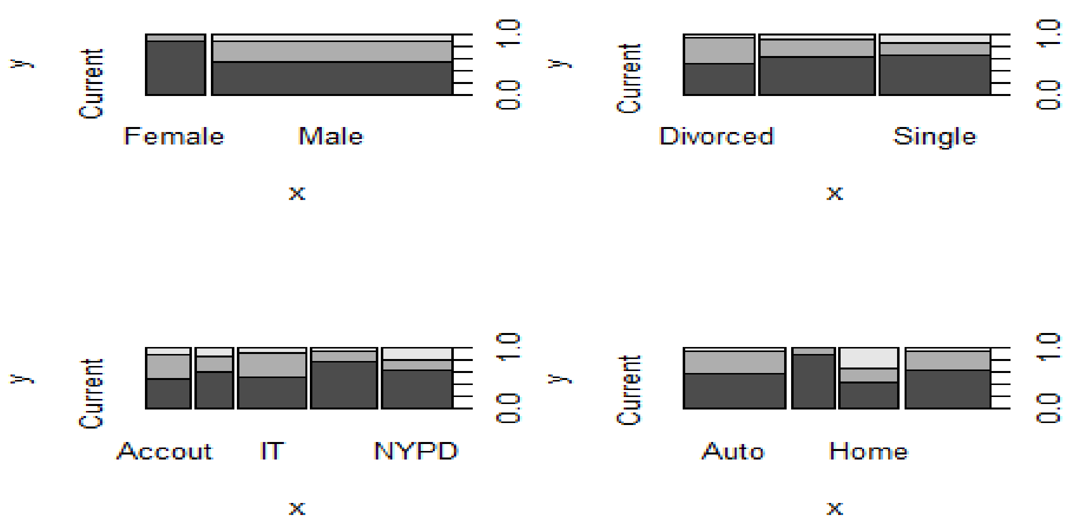


Fig 1.2

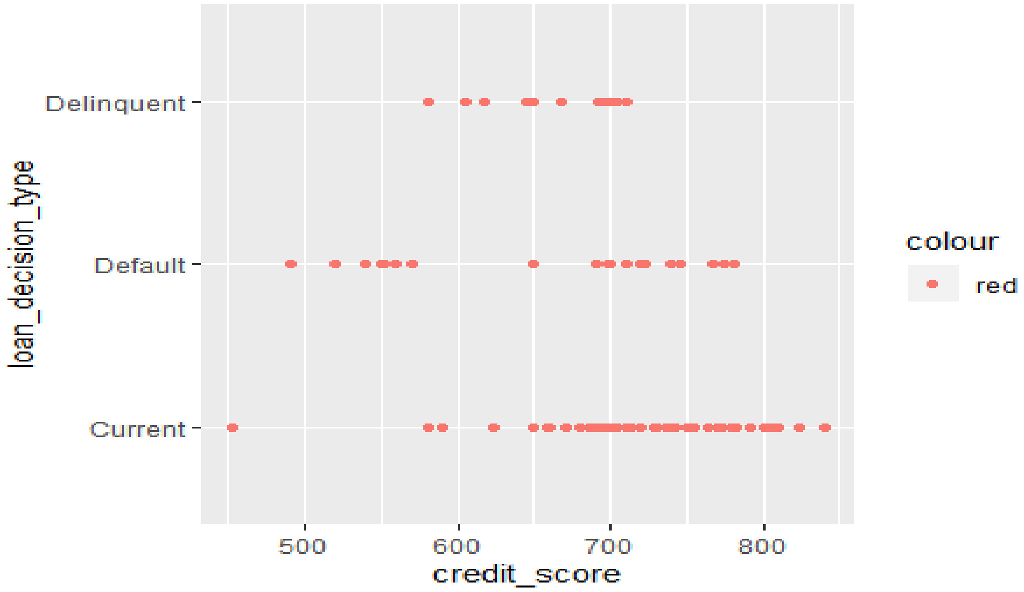
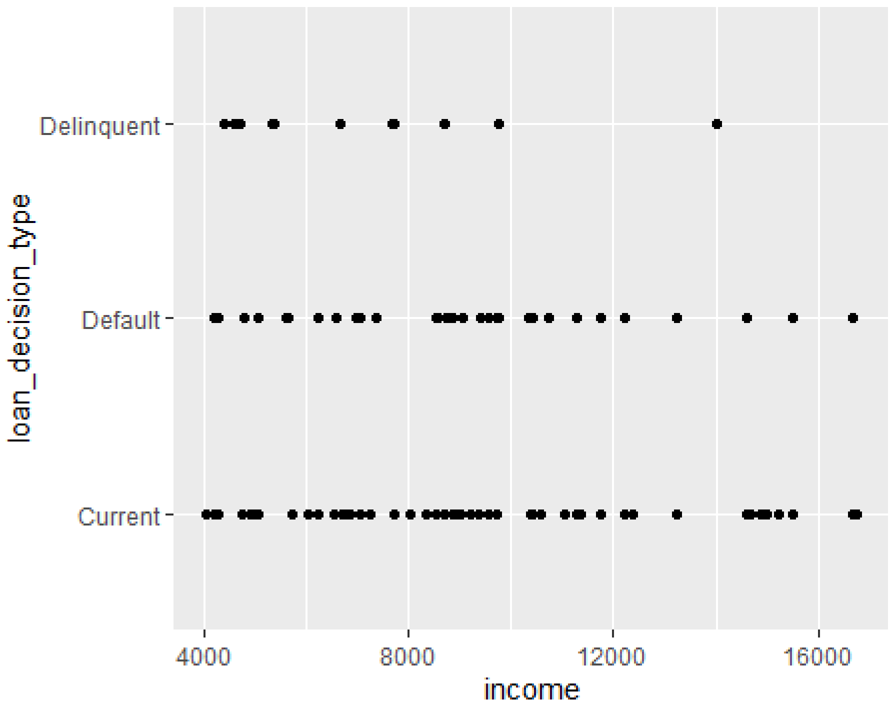
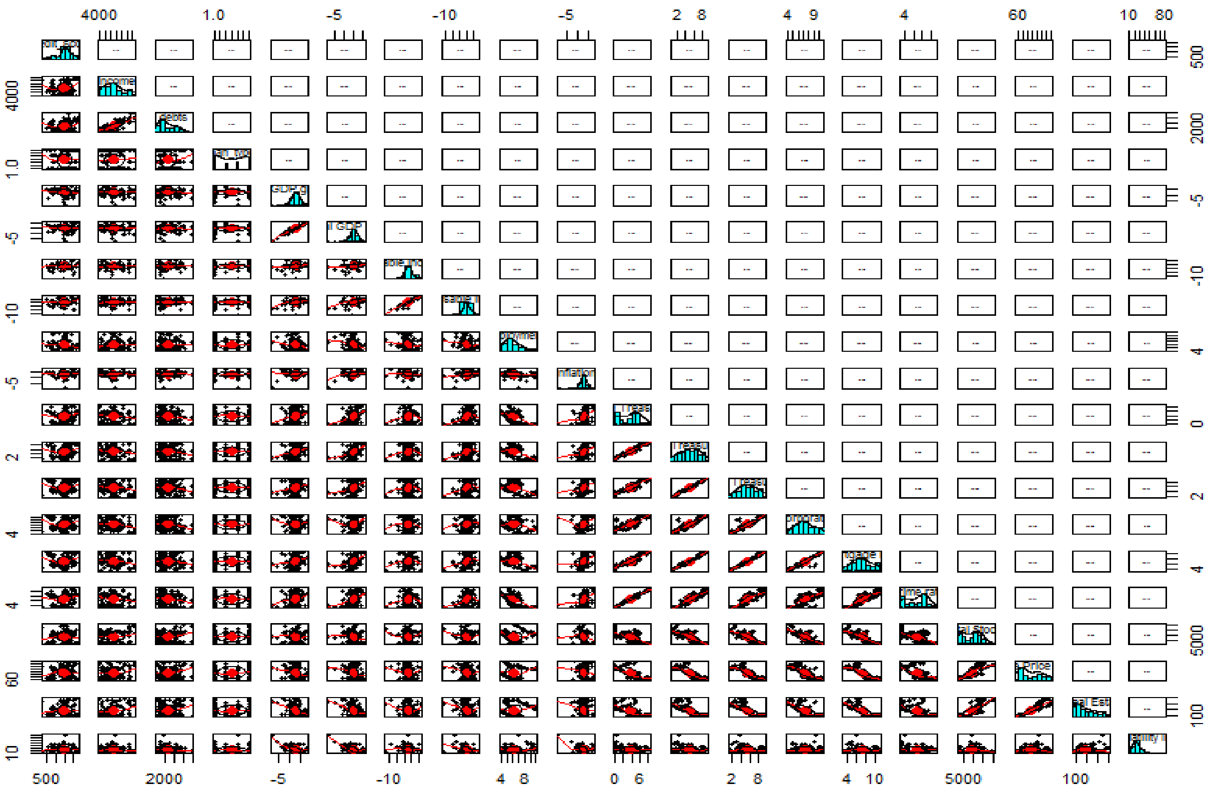

Delinquent
Default
Current
loan_decision_type
500
600
700
800
credit_score
colour
red

Fig 1.3


Delinquent
Default
Current
loan_decision_type
4000
8000
12000
16000
income

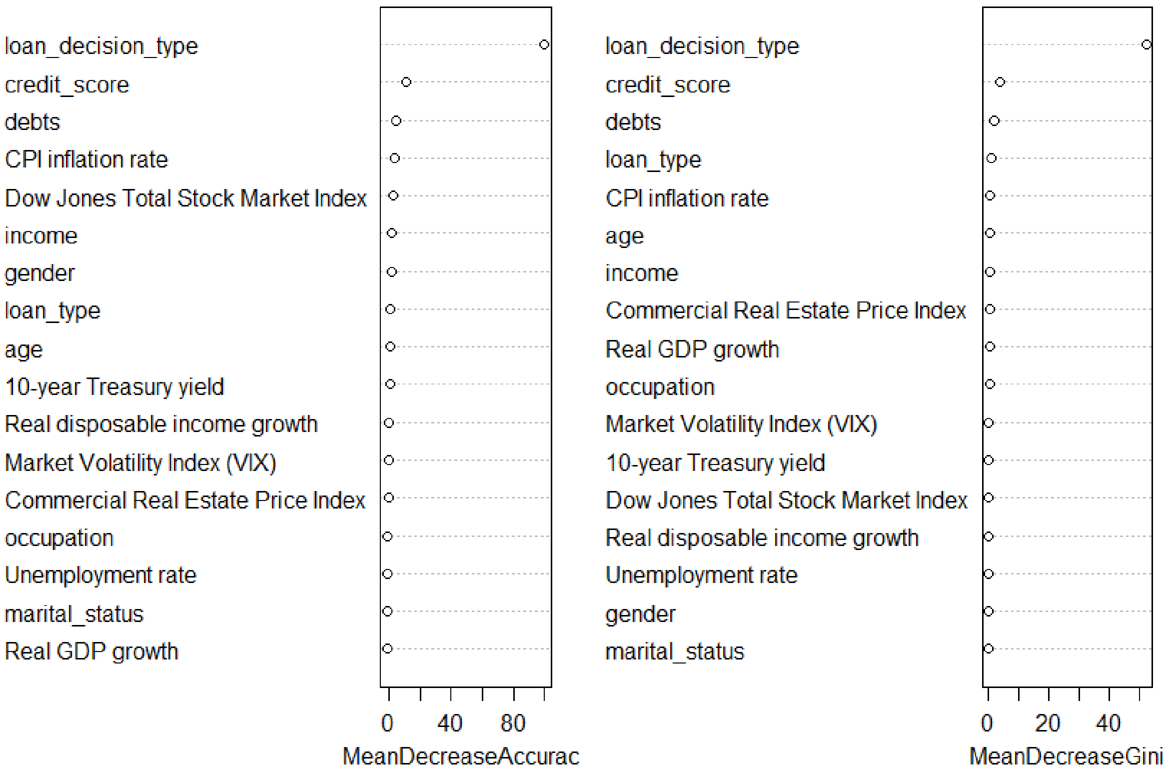
Correlation between Variables: Fig 1.4



Predictor Variable Importance:

Predictors	MeanDecreaseGini
gender	0.05584458
age	0.37814901
marital_status	0.01288889
occupation	0.16184203
credit_score	3.48790686
income	0.28514822
debts	1.62657855
loan_type	0.66605989
Real GDP growth	0.16953758
Real disposable income growth	0.08912802
Unemployment rate	0.05736890
CPI inflation rate	0.50735048
10-year Treasury yield	0.11321638
Dow Jones Total Stock Market Index	0.10972037
Commercial Real Estate Price Index	0.20104675
Market Volatility Index (VIX)	0.13340554

Mod.ccar3



Data Processing & Modeling Process:

Data set was refined after cleaning, preprocessing and partitioned in 2 sets,Training set and Test set.

The Models were developed on Training Data s set and were tested on Test set. Model accuracy was assessed using confusion matrix,

```
str(customer_loan_refined)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':      1114 obs. of  9
variables:
 $ gender             : num  1 1 2 1 1 1 2 2 2 1 ...
 $ age                : num  36 36 34 48 32 44 60 60 60 48 ...
 $ marital_status     : num  2 2 2 2 3 3 3 3 3 2 ...
 $ occupation         : num  5 5 3 1 2 1 4 4 4 1 ...
 $ credit_score       : num  710 720 720 670 720 540 840 824 824
 $ income             : num  9371 9371 9010 6538 8679 ...
 $ debts              : num  2000 3014 1000 2099 1000 ...
 $ loan_type          : num  4 1 2 3 3 4 4 1 2 4 ...
 $ loan_decision_status: Factor w/ 2 levels "0","1": 2 2 2 2 2 1 2 2
```

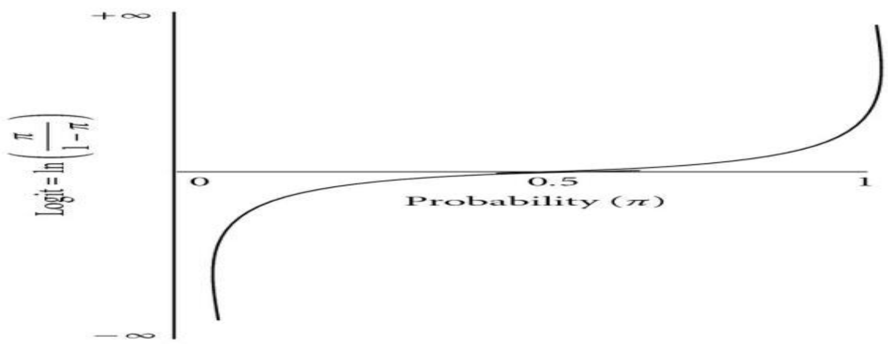
There are 2 broad methods used for Credit Risk Modelling.

1) **Statistical Methods:** The Models generally used are: LOGIT, LDA, MDA, Baye’s Classifier.
Statistical Models have certain assumptions about the data, the relationship between Predictors & Response variables. The Model has predetermined Functional form and the data is used to estimate the parameters. The Model is finally subjected to statistical validations, before being put in production.

These Models do not work well when the number of predictors > 20. Model becomes unstable and its performance drops, out of sample.

The class imbalance problem and high multicollinearity among predictors, affects the performance of these models.

Variable selection: Methods such as forward, backward, and stepwise selection and IC based measures (Akaike information criterion (AIC)) and Bayesian information criterion (BIC)) are available; however none they give incorrect estimates of the standard errors and *P*-values, when number of predictors are >20.



.**Logistic Model: Fig1.1**

As evident from S curve.

The gradient of S curve is 0 at 0.5 and is asymptote at extreme ranges.

The Sensitivity computations becomes difficult and unreliable

Fitting a data set to a Statistical Models and finetuning it to meet its assumption is a **waste** of computational effort.

2) **Machine Learning:** Models used in this category are Decision tree, Random forest, Ensemble techniques, Neural Nets, Support Vectors, etc.

They do not assume any functional relationship, nor they require any statistical assumptions about dataset (Linearity, Normality, stable covariance structure, etc.).

Different ML algorithms are applied on the data set and the one with least test set error rate is selected, using 10 fold cross validation method.

Some of these Models have inbuilt test sets and work on bootstrapped samples. The cross validation of error across these samples ensures robust prediction. There are well developed methods for Variable Importance Ranking and Selection.

The Model sensitivity, the Sensitivity index, considering the extremes values of Predictor range and Importance ranking can be done.

Model Performance.

Model 1:Logistics Model:

```
Call:

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.85264  -0.07926   0.03195   0.11459   1.66751

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.579e+01  7.434e+00  -3.469 0.000522 ***
gender          1.105e+00  1.959e+00   0.564 0.572714
age             1.801e-01  8.608e-02   2.093 0.036359 *
marital_status  2.082e+00  9.578e-01   2.174 0.029687 *
occupation      3.202e-01  4.965e-01   0.645 0.518914
credit_score    2.476e-02  7.882e-03   3.141 0.001685 **
income          1.376e-03  5.201e-04   2.646 0.008138 **
debts          -4.712e-03  1.386e-03  -3.400 0.000673 ***
loan_type      -4.365e-01  4.046e-01  -1.079 0.280692
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 106.998  on 90  degrees of freedom
Residual deviance:  32.271  on 82  degrees of freedom
AIC: 50.271
```

```
Number of Fisher Scoring iterations: 8
```

```
tablelm1
          actual
predicted  0  1
        0  5  2
        1  2 14

> accuracylm1
[1] 0.826087
```

Model 2: LDA Model

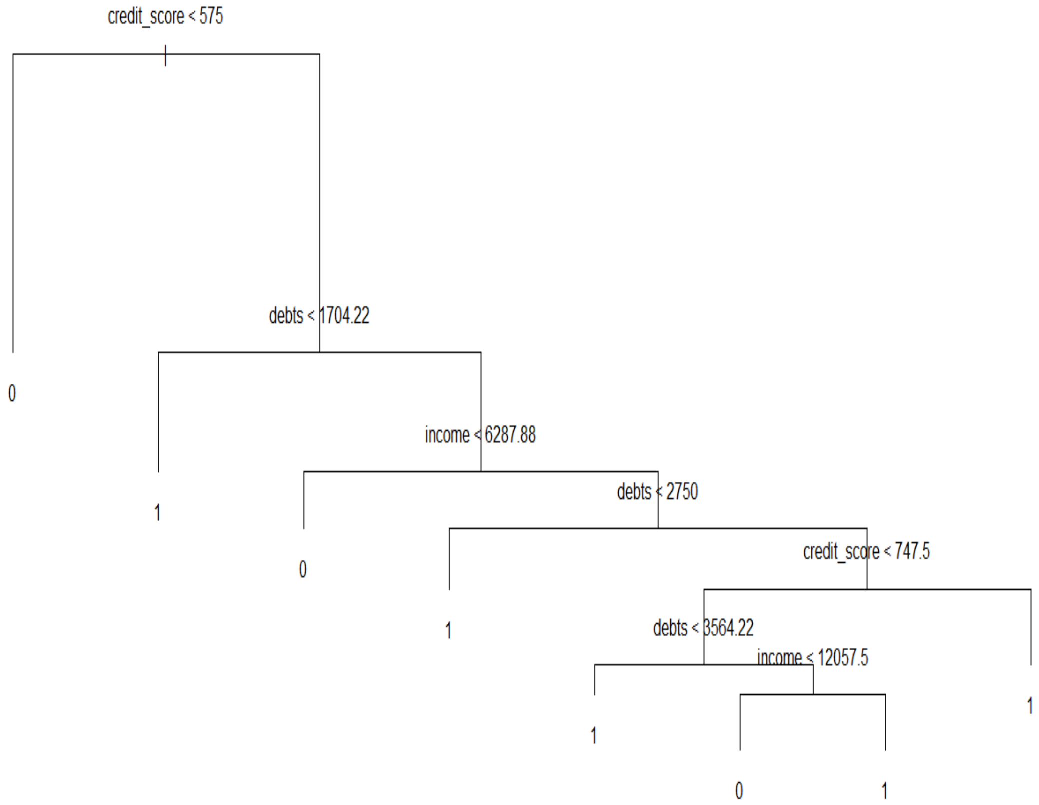
```
Call:

Prior probabilities of groups:
        0         1
0.2747253 0.7252747

Group means:
    gender      age marital_status occupation credit_score   income
debts
0 1.120000 37.56000       1.840000    3.08000     640.4000 8864.687
3477.042
1 1.272727 38.54545       2.136364    3.30303     725.0909 9148.658
2360.390
  loan_type
0  2.400000
1  2.439394

Coefficients of linear discriminants:
                         LD1
gender         -0.0757520341
age             0.0383488855
marital_status  0.4310195815
occupation      0.0729713994
credit_score    0.0098128319
income          0.0003413658
debts          -0.0012400317
loan_type      -0.0782903565

tablelda
          actual
predicted  0  1
        0  6  2
        1  1 14
accuracy_lda
[1] 0.8695652
```

Model 3: Naïve Baye's classifier.

```
Call:

A-priori probabilities:
Y
        0         1 
0.2747253 0.7252747 

Conditional probabilities:
   gender
Y       [,1]      [,2]
  0 1.120000 0.3316625
  1 1.272727 0.4487746

   age
Y       [,1]     [,2]
  0 37.56000 11.96551
  1 38.54545 15.23981

   marital_status
Y       [,1]      [,2]
  0 1.840000 0.8504901
  1 2.136364 0.7623115

   occupation
Y      [,1]     [,2]
  0 3.08000 1.411855
  1 3.30303 1.424723

   credit_score
Y       [,1]     [,2]
  0 640.4000 96.41922
  1 725.0909 59.53686

   income
Y       [,1]     [,2]
  0 8864.687 3131.956
  1 9148.658 3727.252

   debts
Y       [,1]     [,2]
  0 3477.042 1612.375
  1 2360.390 1604.257

   loan_type
Y       [,1]     [,2]
  0 2.400000 1.384437
  1 2.439394 1.204208
tablebayes

bayes.pred  0  1
         0  2  2
         1  5 14
0.7391304
```

Model 4: Classification Tree:

```
  Classification tree:

Variables actually used in tree construction:
  "credit_score" "debts"        "income"

Number of terminal nodes:  8
Residual mean deviance:  0.2672 = 22.18 / 83
Misclassification error rate: 0.05495 = 5 / 91
```



tabletree1

```
          actual
predicted  0  1
        0  6  1
        1  1 15
```

accuracytree

```
[1] 0.8695652
```

Model 5: Random Forest:

```
Call:
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 13.19%
Confusion matrix:
   0  1 class.error
0 14 11  0.44000000
1  1 65  0.01515152

rf.tree$importance
               MeanDecreaseGini
gender                0.7412788
age                   3.8926255
marital_status        1.9822257
occupation            1.8166870
credit_score         11.2709188
income                4.8311361
debts                 8.4794197
loan_type             1.9432035


tablerf
         actual
predicted  0  1
        0  3  1
        1  4 15
accuracyrf
[1] 0.7826087
```

Key Findings:

1) When the data is clean, with minimal noise and a clear distinction between classes is possible , Baye's Classifier &LDA seems to outperform all the other models.

2) When data set matrix (nxp), becomes very high, Statistical model fails because Matrix generally do not have full rank and their inverse does not exist.

3) When data is noisy, the class distinction cannot be made using a Straight line or a Hyperplane, Machine Learning Models exhibit better performance. Models Like Decision tree, RF, work in presence of outliers and missing values.

4) When the no. of predictors increases and the decision boundary exists in large dimensional feature space, Support Vector Machines is the best Model to be used.

5) Machine learning Models exhibits very high accuracy on training set, but performance drops drastically on test set. This overfitting problem must be dealt carefully.

The Research Team,

ALBEDO ENERGY
##