



HYBRID SYSTEM FOR ENHANCED PROFIT PREDICTION

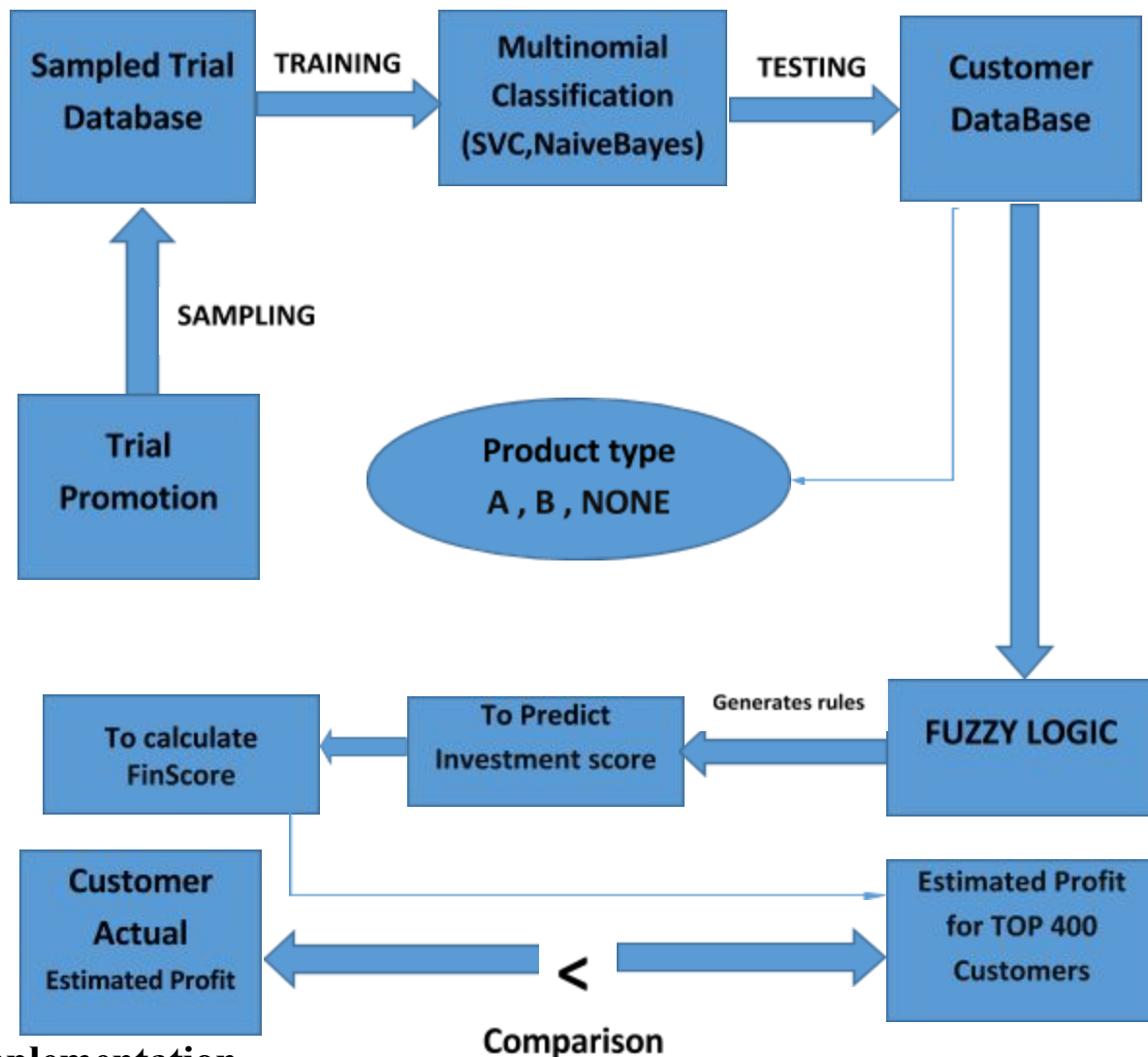
Naitik Shukla
Vignesh Srinivasan
Navneet Goswami

Abstract	6
Sampling and Classification	7
Fuzzification of Customer Inverse Potential Score	10
Fuzzy System Structure:	11
RULES	11
Comparison Result Between Actual(Given) and Predicted FinScore:	13
Variables	14
How To Run Through The Code For Fuzzy Logic:	22

Abstract

The project demands a forecasting environment to run an email campaign such that sales of product A and B can be achieved with maximum profit from the topmost probable customers. A hybrid model can be a possible solution, to replicate better results, from a trial campaign done earlier.

Architecture & Plan:



Implementation

1. Sampling and Classification

The trial campaign conducted on 1000 customers provides result which are biased towards a category and hence an overall sampling is done to generate valid cases for deducing and training a Naive Bayes and SVM Classifier Model, having three categories, which are A, B & None. The model is trained with the total set of 1000 and tested on the 20% of the same dataset (Ratio 80:20).

When checked on raw data provided for Distribution wrt each Target class, we got below information, which clearly shows skewness of data by class:

```
> print(prop.table(table(data$status))) Whole Data
      A      B      None
0.19472914 0.05710102 0.74816984
> print(prop.table(table(trial_train1$status))) Train Data
      A      B      None
0.20475320 0.06032907 0.73491773
> print(prop.table(table(trial_test$status))) Test Data
      A      B      None
0.15441176 0.04411765 0.80147059
> |
Class wise Distribution
```

So it was in need to perform some changes on raw data so that this biased data towards 'None' class can be reduced but still maintains original distribution.

We have done Oversampling on raw data only for 'Training data' of model as can be seen below. Testdata was not been touched for any modifications.

Oversampling performed on the imbalanced data Using SMOTE

```
> balancedData <- SMOTE(status~., trial_train1, perc.over = 2300, k = 7, perc.under = 100, learner = NULL)
> print(prop.table(table(balancedData$status)))
```

After Smote 1(for B)

```
      A      B      None
0.1044487 0.5106383 0.3849130
```

```
> trial_train <- SMOTE(status~., balancedData, perc.over = 200, k = 15, perc.under = 200, learner = NULL )
> print(prop.table(table(trial_train$status)))
```

After Smote 1(for A)

```
      A      B      None
0.4285714 0.3218695 0.2495591
```

Class wise Distribution after SMOTE

Below are the comparison result of both data on unaltered TestData , so as to better guess for which model giving better accuracy.

Naive Bayes and SVM Model Compare for TestSet after UpSampling On TrainSet

SVM

Confusion Matrix and Statistics

	Reference		
Prediction	A	B	None
A	19	0	42
B	0	3	14
None	2	3	53

Overall Statistics

Accuracy : 0.5515
95% CI : (0.4639, 0.6368)
No Information Rate : 0.8015
P-Value [Acc > NIR] : 1

Kappa : 0.2312
McNemar's Test P-Value : NA

Naive Bayes

Confusion Matrix and Statistics

	Reference		
Prediction	A	B	None
A	21	0	0
B	0	6	0
None	0	0	109

Overall Statistics

Accuracy : 1
95% CI : (0.9732, 1.0000)
No Information Rate : 0.8015
P-Value [Acc > NIR] : 8.486e-14

Kappa : 1
McNemar's Test P-Value : NA

By looking into above Confusion Matrix we can clearly say that , Naive Bayes outperform SVC in this small TestDataset (unchanged). So we went on and perform prediction of NaiveBayes trained model on complete 1000 records given for TrialAd.

Below is the result for same:

Naive Bayes for Full Data

Confusion Matrix and Statistics

Prediction	Reference		
	A	B	None
A	133	0	0
B	0	39	0
None	0	0	511

Overall Statistics

Accuracy : 1
95% CI : (0.9946, 1)
No Information Rate : 0.7482
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 1
McNemar's Test P-Value : NA

2. Fuzzification of Customer Inverse Potential Score

To come up with similar results obtained from the previous trial campaign, specific rules and patterns will be required to derive the best set of customers who can effectively improve the expected total profit. Hence fuzzy logic does the job in achieving that goal.

Generate Investment Potential scores & Max Profit:

The main purpose of fuzzy logic is used to predict the investment potential score of the customers by generating optimized rules such that it satisfies the business cases. The overall investment potential of a customer highly depends on customer's account activity rather than the personal attributes of the customer. Taking such cases into consideration the fuzzy logic model is trained to predict the potential investment score ranging from (0 to 10) along with supporting crisp rules that defines the criteria and helps to predict the nature of linguistic attribute such as Customer Investment Potential score.

A fuzzy hybrid system is built to provide the basis from which decision can be made or patterns discerned, as the process involves fuzzification ie customer attributes or details,

which are numerical , are first limited to a range using membership function. The membership function used in our model is trapezoidal.

Fuzzy System Structure:

The system structure identifies the fuzzy logic inference flow from the input variables to the output variables. The fuzzification in the input interfaces translates analog inputs into fuzzy values. The fuzzy inference takes place in rule blocks which contain the linguistic control rules. The output of these rule blocks are linguistic variables. The defuzzification in the output interfaces translates them into analog variables.

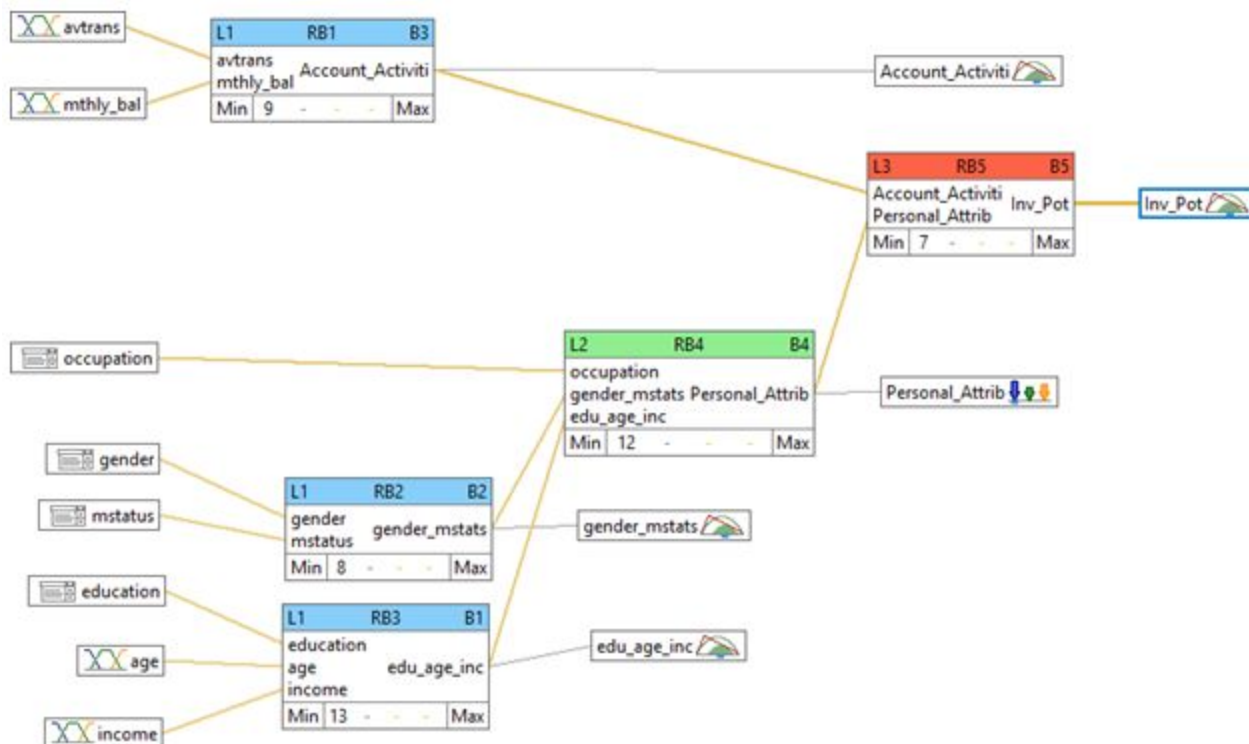


Figure 1: Structure of the Fuzzy Logic System

RULES:

Further the rules on the customer attributes (age,Income) are generated for calculating the investment potential score . Basic rules blocks are included below as screenshots

RULEBLOCK RB1

```
AND : MIN;          // Use 'min' for 'and'
ACT : MIN;          // Use 'min' activation method
ACCU : MAX;         // Use 'max' accumulation method
RULE 1 : IF avtrans IS High AND avbal IS High THEN accactivity IS High;
RULE 2 : IF avtrans IS Medium AND avbal IS Medium THEN accactivity IS Medium;
RULE 3 : IF avtrans IS Low AND avbal IS Low THEN accactivity IS Low;
RULE 4 : IF avtrans IS High AND avbal IS Medium THEN accactivity IS Medium;
RULE 5 : IF avtrans IS Medium AND avbal IS High THEN accactivity IS High;
RULE 6 : IF avtrans IS Low OR avbal IS Low THEN accactivity IS Low;
```

END_RULEBLOCK

RULEBLOCK RB2

```
AND : MIN;          // Use 'min' for 'and'
ACT : MIN;          // Use 'min' activation method
ACCU : MAX;         // Use 'max' accumulation method
RULE 1 : IF Gender IS Male AND mstats IS married THEN gender_mstats IS Medium;
RULE 2 : IF Gender IS Male AND mstats IS single OR mstats IS widow OR mstats IS divorced THEN gender_mstats IS High;
RULE 3 : IF Gender IS Female AND mstats IS single OR mstats IS widow THEN gender_mstats IS High;
RULE 4 : IF Gender IS Female AND mstats IS married THEN gender_mstats IS Low;
RULE 5 : IF Gender IS Female AND mstats IS divorced THEN gender_mstats IS Medium;
```

END_RULEBLOCK

RULEBLOCK RB3

```
AND : MIN;          // Use 'min' for 'and'
ACT : MIN;          // Use 'min' activation method
ACCU : MAX;         // Use 'max' accumulation method
RULE 1 : IF Age IS middle AND Education IS Postgrad OR Education IS Professional AND Income IS High THEN edu_age_inc IS High;
RULE 2 : IF Age IS middle AND Education IS Secondary AND Income IS High THEN edu_age_inc IS Medium;
RULE 3 : IF Age IS middle AND Education IS Tertiary THEN edu_age_inc IS Low;
RULE 4 : IF Age IS old AND Income IS High THEN edu_age_inc IS High;
RULE 5 : IF Age IS old AND Income IS Medium THEN edu_age_inc IS Medium;
RULE 6 : IF Age IS old AND Income IS Low THEN edu_age_inc IS Low;
RULE 7 : IF Age IS young AND Income IS High THEN edu_age_inc IS High;
RULE 8 : IF Age IS young AND Income IS Medium THEN edu_age_inc IS Medium;
RULE 9 : IF Age IS young AND Income IS Low THEN edu_age_inc IS Low;
```

END_RULEBLOCK

RULEBLOCK RB4

```
AND : MIN;
ACT : MIN;
ACCU : MAX;
RULE 1 : IF Occupation IS Retired THEN personal_attr IS Low;
RULE 2 : IF Occupation IS Legal OR Occupation IS IT OR Occupation IS Medicine OR Occupation IS Finance OR Occupation IS Government THEN
personal_attr IS High;
RULE 3 : IF edu_age_inc IS Medium OR edu_age_inc IS High AND gender_mstats IS High THEN personal_attr IS High;
RULE 4 : IF edu_age_inc IS Low AND gender_mstats IS High THEN personal_attr IS Low;
RULE 5 : IF edu_age_inc IS High AND gender_mstats IS Medium THEN personal_attr IS High ;
RULE 6 : IF edu_age_inc IS Medium AND gender_mstats IS Medium THEN personal_attr IS Medium;
RULE 7 : IF edu_age_inc IS Low AND gender_mstats IS Medium THEN personal_attr IS Low;
RULE 8 : IF edu_age_inc IS Medium AND gender_mstats IS Low THEN personal_attr IS Low;
RULE 9 : IF Occupation IS Construct OR Occupation IS Manuf OR edu_age_inc IS High THEN personal_attr IS Medium;
```

END_RULEBLOCK


```

RULEBLOCK RB5
  AND : MIN;      // Use 'min' for 'and'
  ACT : MIN;      // Use 'min' activation method
  ACCU : MAX;     // Use 'max' accumulation method
  RULE 1 : IF accactivity IS High AND personal_attr IS High THEN Potential IS High;
  RULE 2 : IF accactivity IS High AND personal_attr IS Medium THEN Potential IS High;
  RULE 3 : IF accactivity IS Medium AND personal_attr IS High OR personal_attr IS Medium THEN Potential IS Medium;
  RULE 4 : IF accactivity IS Low AND personal_attr IS Low OR personal_attr IS Medium THEN Potential IS Low;
  RULE 5 : IF accactivity IS Low AND personal_attr IS High THEN Potential IS Medium;
END RULEBLOCK

```

Meanwhile the fuzzy inference engine performs an approximate reasoning by associating the input variables ie the customer attributes with the fuzzy rules generated

The Defuzzification technique used is Center Of Gravity which incorporates the membership values and the effective widths of the membership function in calculating a crisp output ie. (fuzzy degree of membership in the qualifying linguistic set.). It effectively calculates the optimal width automatically by the genetic algorithm through training on the previous trial campaign dataset.

Below find the attached code for Fuzzy system

3. Comparison Result Between Actual(Given) and Predicted FinScore:

Case1:

When Decreasing Sort FinScore based on Given Score and comparing result with same Index Prediction Score:

Actual	Predicted
Top400	Top400
1237.94	1153.277

Case2:

When Decreasing Sort FinScore based on Predicted Score and comparing result with same Index Original(given) Score:

Predicted	Given
Top400	Top400
1248.297	1175.72

Excel with results and formula (custdatabase-fuzzyResults.xlsx) are attached in project folder.

By sorting down for top 400 customers we can predict the expected profit. The above results shows the comparisons between the original score and the predicted one. This Expected Profit for each Customer are calculated based on whether the Product purchased is 'A' or 'B' or 'None'

$$\begin{aligned}\text{Expected Profit} = & 0.6 * \text{Predicted Score Using Fuzzy (For Product 'A')} \\ & \text{Predicted Score Using Fuzzy(For Product 'B')} \\ & 0 \text{ (For 'None')}$$

Variables

This part contains the definition of all linguistic variables and of all membership functions.

Linguistic variables are used to translate real values into linguistic values. The possible values of a linguistic variable are not numbers but so called 'linguistic terms'.

The following tables list all variables of the system as well as the respective fuzzification or defuzzification method. Also the properties of all base variables and the term names are listed.

.1 Inputs









#	Variable Name	Type	Unit	Min	Max
1	age		years	0	100
2	avtrans			0	10000
3	education		-	0	3
4	gender		-	0	1
5	income		SGD	0	20000
6	mstatus		-	0	3
7	Avbal(mthly_bal)			500	89000
8	occupation		-	0	8

Table 2: Variables of Group "Inputs"

Fuzzification Methods



Trapezoid



Categorical

.2 Outputs








#	Variable Name	Type	Unit	Min	Max
9	Accactivity			0	1
10	edu_age_inc			0	1
11	gender_mstats			0	1
12	Potential			1	11
13	Personal_Attrib			0	1

Table 3: Variables of Group "Outputs"

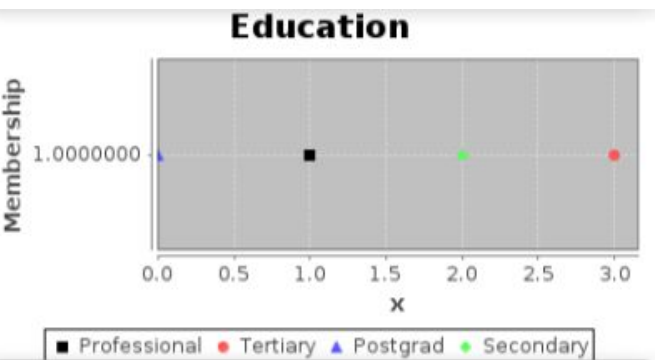
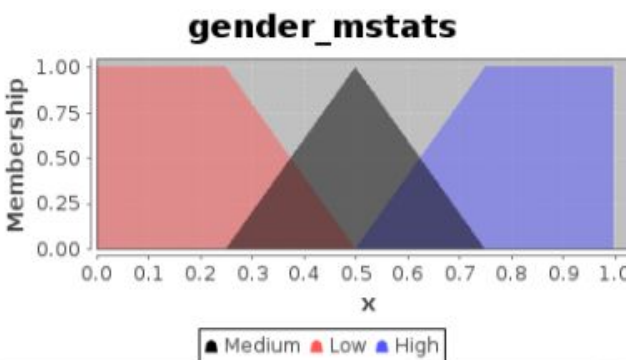
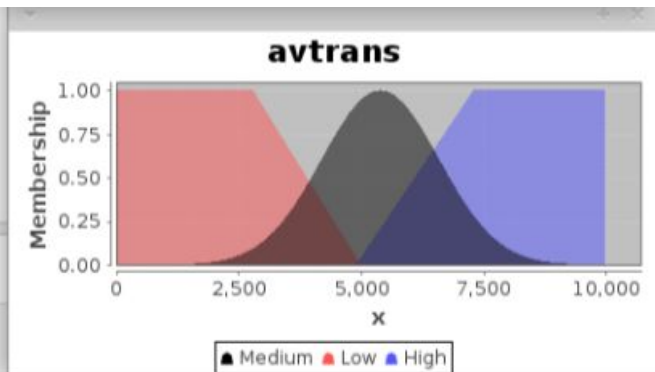
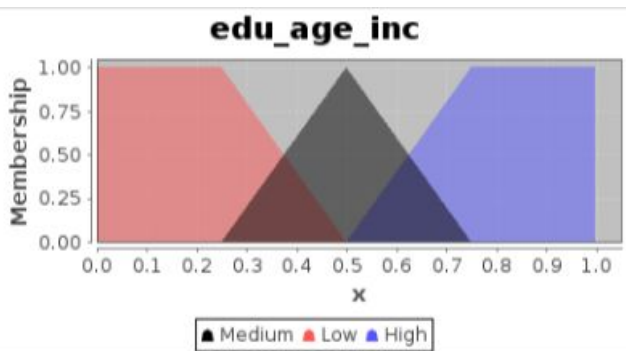
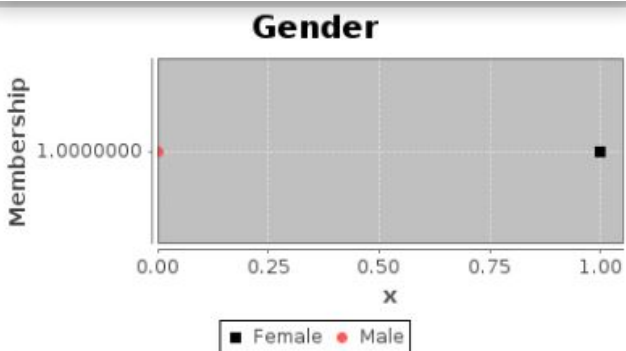
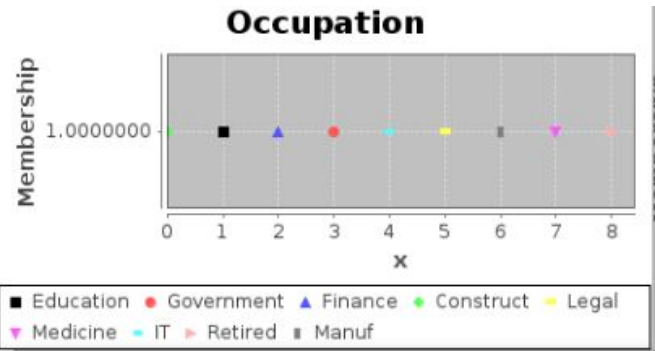
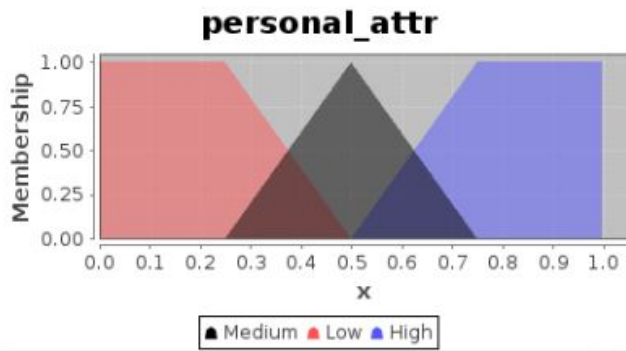
Defuzzification Methods

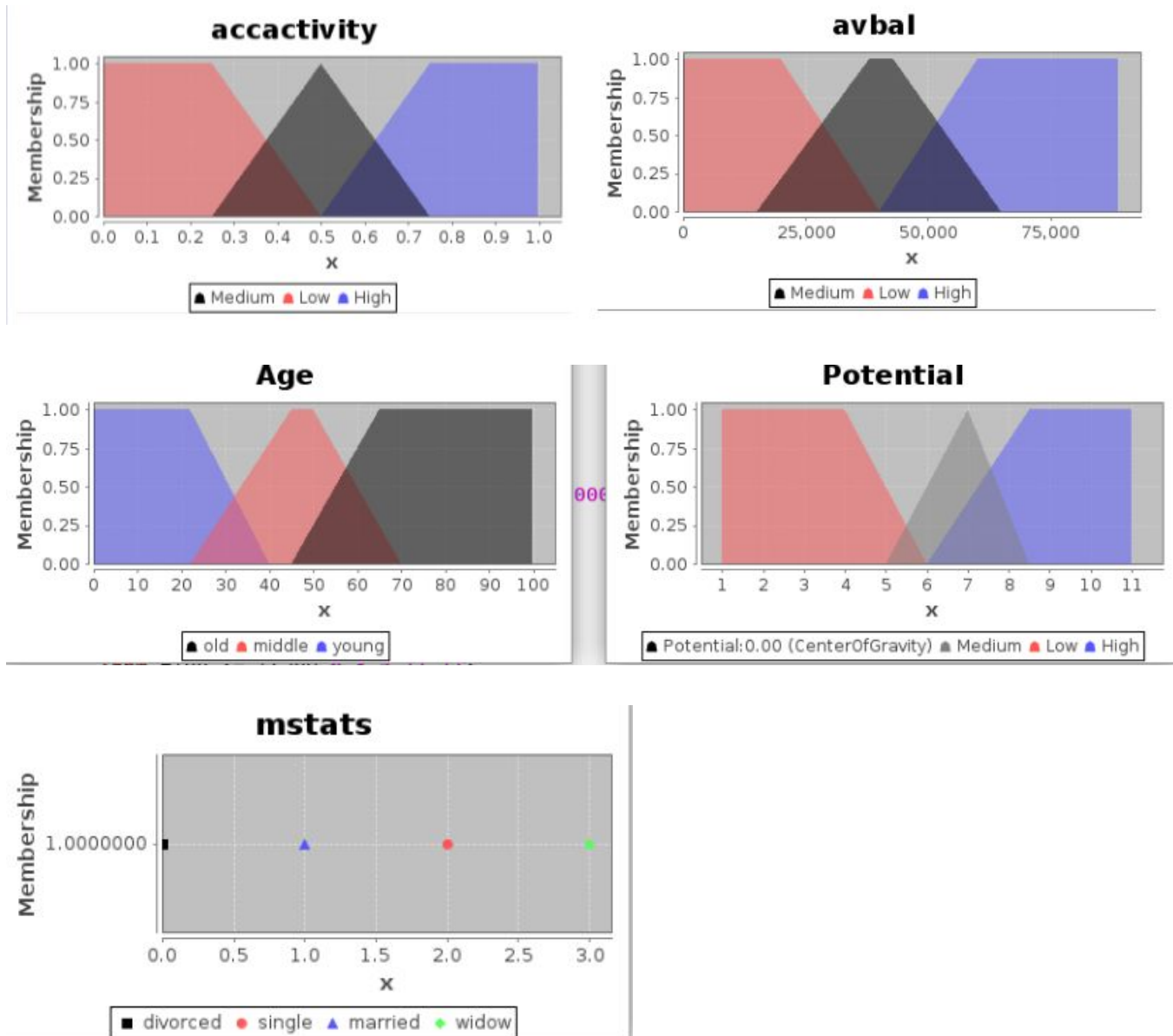


CoG

Input Variables Details:

Below are the graphic visualization for the Fuzzy Set Space defined for each variables and following with more details about each variable.





1. Terms of Variable "age"

Term "young"

- Type: Trapezoid(standard)
- Points: 0,0,22,40

Term "middle"

- Type: Trapezoid(standard)
- Points: 22,45,50,70

Term “old”

- Type: Trapezoid(standard)
- Points: 45 65 100 100

2. Terms of Variable "Income"

Term "Low"

- a. Type: Trapezoid(standard)
- b. Points: 0, 0 ,5000, 10000

Term “Medium”

- Type: Trapezoid(standard)
- Points: 4500 ,9325 ,10150 ,15000

Term “High”

- Type: Trapezoid(standard)
- Points: 9150 15000 20000 20000

3. Terms of Variable "avtrans"

Term "Low"

- a. Type: Trapezoid(standard)
- b. Points: 0 ,0 ,2800 ,5000

Term “Medium”

- Type:Gauss(standard)
- Mean,Std: 5400 1200

Term “High”

- Type: Trapezoid(standard)
- Points: 4900, 7300, 10000 ,10000

4. Terms of Variable "mstats"

Term "Divorced"

- a. Type: Point
- b. Points: 0

Term “Married”

- Type: Point
- Points: 1

Term “Single”

- Type: Point
- Points: 2

Term “widow”

- Type: Point
- Points: 3

5. Terms of Variable "personal_attr"

Term "Low"

- a. Type: Trapezoid(standard)
- b. Points: 0,0,0.25,0.50

Term “medium”

- Type: Trapezoid(standard)
- Points: 0.25, 0.50, 0.50, 0.75

Term “High”

- Type: Trapezoid(standard)
- Points: 0.5,0.75,1,1

6. Terms of Variable "gender_mstats"

Term "Low"

- a. Type: Trapezoid(standard)
- b. Points:0,0,0.25,0.50

Term “Medium”

- Type: Trapezoid(standard)
- Points: 0.25,0.50, 0.50,0.75

Term “High”

- Type: Trapezoid(standard)
- Points: 0.5, 0.75, 1,1

7. Terms of Variable "edu_age_inc"

Term "Low"

- a. Type: Trapezoid(standard)
- b. Points: 0, 0, 0.25, 0.50

Term "Medium"

- Type: Trapezoid(standard)
- Points: 0.25, 0.50, 0.50, 0.75

Term "High"

- Type: Trapezoid(standard)
- Points: 0.5, 0.75, 1, 1

8. Terms of Variable "accactivity"

Term "Low"

- a. Type: Trapezoid(standard)
- b. Points: 0, 0, 0.25, 0.50

Term "Medium"

- Type: Trapezoid(standard)
- Points: 0.25, 0.50, 0.50, 0.75

Term "High"

- Type: Trapezoid(standard)
- Points: 0.5, 0.75, 1, 1

9. Terms of Variable "Education"

Term "Postgrad"

- a. Type: Point
- b. Points: 0

Term "Professional"

- Type: Point
- Points: 1

Term "Secondary"

- Type: Point
- Points: 2

Term “Tertiary”

- Type: Point
- Points: 3

10. Terms of Variable "Occupation"

Term "Construct"

- a. Type: Point
- b. Points: 0

Term “Education”

- Type:Point
- Points: 1

Term “Finance”

- Type: Point
- Points: 2

Term “Government”

- Type: Point
- Points: 3

Term “IT”

- Type: Point
- Points: 4

Term “Legal”

- Type: Point
- Points: 5

Term “Manuf”

- Type: Point
- Points: 6

Term “Medicine”

- Type: Point
- Points: 7

Term “Retired”

- Type: Point
- Points: 8

11. Terms of Variable "Gender"

Term "Male"

- a. Type: Point
- b. Points: 0

Term “Female”

- Type: Point
- Points: 1

12. Terms of Variable "Potential"

Term "Low"

- a. Type: Trapezoid(standard)
- b. Points: 0.5, 0.5 ,3 ,5

Term “Medium”

- Type: Trapezoid(standard)
- Points: 3, 5 ,5, 7

Term “High”

- Type: Trapezoid(standard)
- Points: 5.5, 7 ,10, 10

How To Run Through The Code For Fuzzy Logic:

- Download jFuzzyLogic.jar from the Folder
- Using The above jar Type Java -jar jFuzzyLogic.jar *.fcl (You have to give proper name for FCL file)
- FCL file can be downloaded from the folder
- Open the fuzzy folder which is uploaded using some JAVA IDE
- Change the path of fcl file in 87 line and path for at the line 39 which points to 4000 sets of customer database (**Remember to convert all the catagorical variable to Numeric value**)
- Run the fuzzy.java file