# **MTech KE Unit 7 Hybrid Systems**

Bharat Nagaraju (A0178258N)

Vigneshram Andiappan Selvaraj (A0178215A)

Contents

[**MTech KE Unit 7 Hybrid Systems** 1](#_Toc20777371)

[Workshop 1A & 1B 3](#_Toc20777372)

[Problem Statement 3](#_Toc20777373)

[Observations 3](#_Toc20777374)

[Workshop 1A 4](#_Toc20777375)

[Regression model 4](#_Toc20777376)

[Optimization Model – Genetic Algorithm Model 5](#_Toc20777377)

[Results 6](#_Toc20777378)

[Validation 6](#_Toc20777379)

[Workshop 1B 7](#_Toc20777380)

[Design 7](#_Toc20777381)

[Predictive Model – Genetic Algorithm Model 1 7](#_Toc20777382)

[Result 8](#_Toc20777383)

[Optimization Model – Genetic Algorithm Model 2 9](#_Toc20777384)

[Workshop 2B 10](#_Toc20777385)

[Problem Statement 10](#_Toc20777386)

[Objective 10](#_Toc20777387)

[Data Preprocessing 11](#_Toc20777388)

[Modelling 11](#_Toc20777389)

[Testing & Ensembles 15](#_Toc20777390)

[Results 15](#_Toc20777391)

# 

# Workshop 1A & 1B

### Problem Statement

Milton advertises new products via 5 websites by placing banners on these websites. The websites can display banners from a total of 6 designs. At any point in time, only 1 banner can be displayed on a website.

As each website comes with a cost for listing the banners, Milton would like to optimize the cost of advertising. The table of charges is as shown in Table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cost | Website 1 | Website 2 | Website 3 | Website 4 | Website 6 |
| $/ hr | 15 | 10 | 8 | 8 | 12 |

**Table 1 Advertisement Banner Cost**

In addition to the charges, Milton have listed the following criteria for the business model.

1. A budget of $300 divided across all 5 websites
2. Banner can be displayed from 0.00 - 24.00 and the duration must be for less than 24 hrs.
3. Banner display must be continuous throughout the day

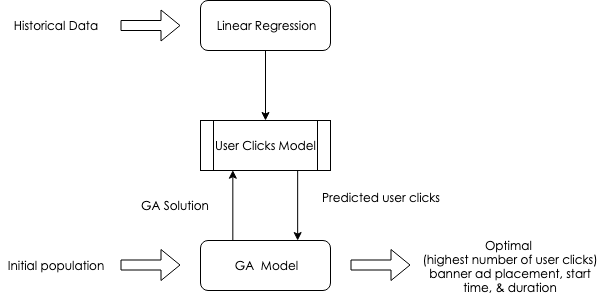
### Observations

The dataset consists of 1,000 observations and 17 columns. Each observation contains information about the cost, the total number of users click from a permutation of 5 websites with 6 banners and display start and end time.

We are required to develop a hybrid intelligent system that consists of a predictive model and optimization model. The initial approach is to create a predictive model that could provide user clicks based on the above listed parameters. We would then pass the output of the predicted user click as the fitness function for the optimization model, to obtain the highest number of user clicks

# Workshop 1A

The overall solution of the hybrid intelligent system can be seen below.



**Figure 1 Workshop 1A – Hybrid system design**

### Regression model

A linear regression model was fitted using all the observations and the resulting model was used to predict the number of user clicks given the website, the ad banner, the start and the end time.

The formula to predict the total user count is:

|  |
| --- |
| User Click ~ Start timew + End timew + Ad Bannerw  for website w (w=1 to 5) |

The result of the regression model is shown in Table 2. As we can see in the statistics Table 3, the R square is which means that our model explains 98% data variance and we can conclude that the model fits well.

|  |  |
| --- | --- |
| **Table 2 Coefficients of Regression Model**  Coefficients | |
| Intercept | 0 |
| W1 Start Time | -8236.5291 |
| W1 End Time | 8768.010269 |
| W1 Ad Banner | -833.7449276 |
| W2 Start Time | -12412.38392 |
| W2 End Time | 11812.79329 |
| W2 Ad Banner | -614.9915423 |
| W3 Start Time | -11385.53664 |
| W3 End Time | 10402.62474 |
| W3 Ad Banner | 739.2322702 |
| W4 Start Time | -10160.44336 |
| W4 End Time | 9900.014686 |
| W4 Ad Banner | -147.5864926 |
| W5 Start Time | -8389.118608 |
| W5 End Time | 9106.052736 |
| W5 Ad Banner | -2.658177274 |

|  |  |
| --- | --- |
| ***Table 3 Regression Statistics*** | |
| Multiple R | 0.955160771 |
| R Square | 0.912332099 |
| Adjusted R Square | 0.910995698 |
| Standard Error | 24989.70173 |
| Observations | 1000 |

### Optimization Model – Genetic Algorithm Model

The Genetic Algorithm (GA) model was then fitted to find the optimal value of the start and end time of a banner to a website. We were using arbitrary number as our initial population value and used Excel to fit the model. The formula from linear regression model was then used to predict the number of clicks.

**Fitness Function**

Maximum number of predicted user clicks.

**Constraints**

The following set of constraints are used in the GA model:

* Total cost ≤ $300
* Ad Banner ID = Integer
* Ad Banner ID = All Different
* Ad Banner ID ≥ 1
* Ad Banner ID ≤ 6
* End time ≥ Start time
* 0 ≤ Start time ≤ 24
* 0 ≤ End time ≤ 24

**Solver Parameters**

Solving Method = “Evolutionary”

Mutation Rate = 0.075

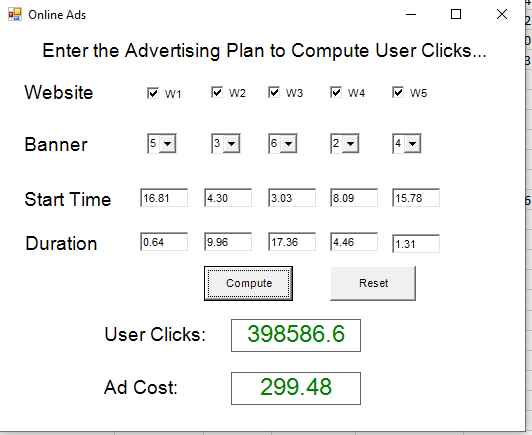
Population Size = 100

Maximum Time without Improvement = 30

### Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Website** | **Banner** | **Start Time** | **EndTime** | **TimeTaken** |
| W1 | 1 | 16.81316548 | 17.4550327 | 0.641867223 |
| W2 | 3 | 4.296145281 | 14.25887944 | 9.962734161 |
| W3 | 6 | 3.032069166 | 20.39237075 | 17.36030159 |
| W4 | 5 | 8.091066065 | 12.5548582 | 4.463792132 |
| W5 | 4 | 15.78079464 | 17.09675074 | 1.315956109 |
|  | 2 | 9.21144E+29 | 8.69382E+29 | -5.1762E+28 |

### Validation

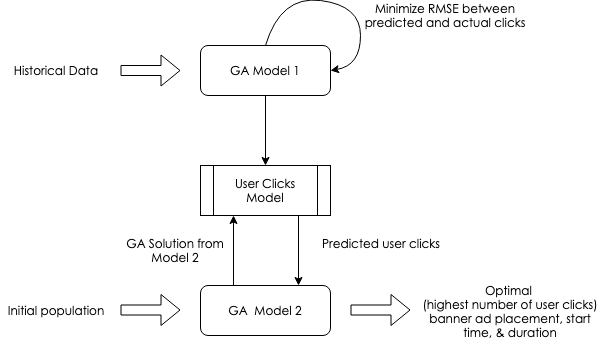


# Workshop 1B

Based on experience, the number of user clicks is dependent on the time of day in which the advertisements are displayed. The day is divided into three time periods and the number of users clicks per hour is stable. The user clicks achievable is also modelled using the duration of the placement and the clicks per hour multiplied by a scale factor. Using this knowledge, modifications to our existing system from 1A are required.

### Design

The design of the hybrid intelligent system can be seen below.



**Figure 2 Workshop 1B – Hybrid system architecture**

### Predictive Model – Genetic Algorithm Model 1

We fitted a GA model as our predictive model to find the optimal 2 cut-off points (CP 1 and CP 2), the number of user clicks achievable per hour for each duration (click 1, click 2, and click 3), and the scale factor of the advertisement banners.

**Fitness Function**

Minimum root-mean-square error of actual and projected user clicks. Projected user clicks were calculated using the following formula:

|  |
| --- |
| ((Duration 1 \* Click 1) + (Duration 2 \* Click 2) + (Duration 3 \* Click 3)) \* Scale Factor |

The duration for each period were derived from start time, end time, and cut-off points using the following formula:

|  |
| --- |
| Duration 1 = IF[min(cp1, ET) - max(0, ST) < 0, 0, min(cp1, ET) - max(0, ST)]  Duration 2 = IF[min(cp2, ET) - max(cp1, ST) < 0, 0, min(cp2, ET) - max(cp1, ST)]  Duration 3 = IF[max(ET, cp2) - max(cp2, ST) < 0, 0, max(ET, cp2) - max(cp2, ST)]  where:  cp1 to cp3 = cutpoint 1 to 3  ET = end time  ST = start time |

**Constraints**

The following set of constraints are used in the GA model:

* 0 ≤ CP 1 ≤ 24
* CP 1 ≤ CP 2
* 0 ≤ Scale factors ≤ 10
* 0 ≤ click 1-3 ≤ 20000
* Click 1-3 = integer

**Solver Parameters**

Solving Method = “Evolutionary”

Mutation Rate = 0.075

Population Size = 100

Maximum Time without Improvement = 30

### Result

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Website** | **Click 1** | **Click 2** | **Click 3** | **Scale Factors** | **CutPoints** |
| Website 1 | 6447 | 3573 | 6339 | 1.36964758 | 4.868991696 |
| Website 2 | 5998 | 6791 | 7971 | 1.27516615 | 8.373441004 |
| Website 3 | 8295 | 11591 | 5443 | 1.59195594 | 2.77634E+29 |
| Website 4 | 5008 | 9608 | 5977 | 1.328197651 | 5.51105E+29 |
| Website 5 | 5520 | 3059 | 7039 | 1.517526051 | 2.22224E+29 |

**Table 5 Cut-off points and user clicks achievable per hour for each period**

### Optimization Model – Genetic Algorithm Model 2

The Genetic Algorithm (GA) model was then fitted to find the optimal value of the start and end time of a banner to a website. We were using arbitrary number as our initial population value and used Excel to fit the model. The cut-off points and user click achievable per hour from GA 1 were used to calculate the projected total user clicks.

**Fitness Function**

Maximum number of predicted user clicks. The predicted user clicks are calculated using the same formula as we used in GA 1:

|  |
| --- |
| ((Duration 1 \* Click 1) + (Duration 2 \* Click 2) + (Duration 3 \* Click 3)) \* Scale Factor |

**Constraints**

The following set of constraints are used in the GA model:

* Total cost ≤ $300
* Ad Banner ID = Integer
* Ad Banner ID = All Different
* Ad Banner ID ≥ 1
* Ad Banner ID ≤ 6
* End time ≥ Start time
* 0 ≤ Start time ≤ 24
* 0 ≤ End time ≤ 24

**Solver Parameters**

Solving Method = “Evolutionary”

# Workshop 2B

### Problem Statement

Sentosa Bank has two new investment products –A & B. They conduct a trial mailing -1000 customers are selected randomly and offered both products.

They plan a second mailing campaign in which the trial promotion results are used to help select 400 customers likely to buy one of the new products

A trained bank officer will visit each selected customer to try to sell them one or other product. To save costs, some staff are trained to sell product A and others product B.

### Objective

Build a hybrid system to select 400 customers that maximize the expected campaign profit.

Expected profit for campaign = Σcustomers Expected profit for customeri

Expected profit\* for customeri= customer investment score \* 0.6 if product purchased = A

= customer investment score if product purchased = B

= 0 if no product purchased

Investment potential is related to the customer’s account activity as well as their personal attributes

1. Account activity is considered more important than personal attributes.
2. Account activity is measured by examining the customers average monthly transactions and average monthly balance. A customer with high values for both has more investment potential.
3. Personal factors relating to investment potential
   1. Gender -males have more potential than females, this is less true for unmarried women
   2. Income –higher is better
   3. Age -investment potential peaks around middle-age
   4. Occupation -retirees have low potential, professionals (doctors, lawyers etc.) have the highest
   5. Education –a higher level is better. Education is more important for middle-aged customers. For older customers income is more important than education-level.

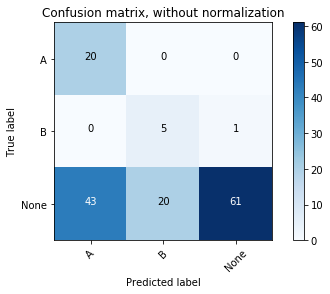
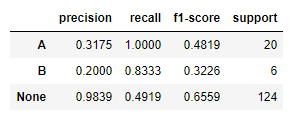
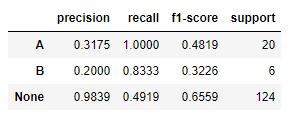
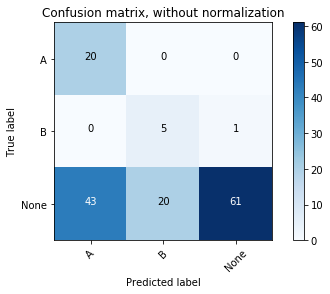
### Data Preprocessing

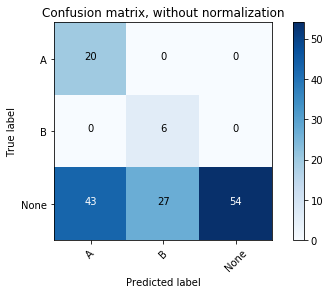
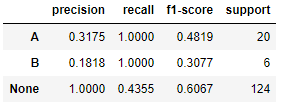
1. Perform - Perform one-hot encoding for all the categorical features and split the data into training and test data
2. Save the dataset for modelling
3. Insert synthetic data - Run the SMOTE algorithm to balance the output classes as mentioned [here](https://imbalanced-learn.org/en/stable/generated/imblearn.over_sampling.SMOTE.html)
4. Above step is required to balance the dataset



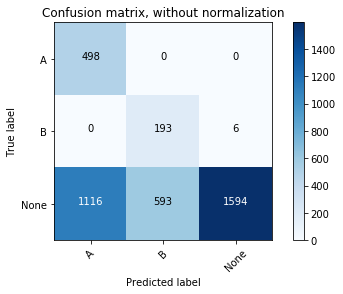
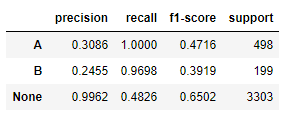
### Modelling

Modelling using Boosters

1. Data balancing– New rows are added for all classes using equal classes strategy and various model boosters such as AdaBoost, Gradient Boosting, XG Boost
2. AdaBoost: Confusion Matrix is below  
    
3. Gradient Boosting: Confusion Matrix is below  
   
4. XG Boost: Confusion Matrix is below

Test Model on Test Dataset:



Modelling for Binary Classification

1. Data Normalisation: Class labels are generated for binary classification and various models are built and their performances are compared. For Binary Classification – Class labels are 1 and 0. 1 if “None” else 0.
2. Neural network:

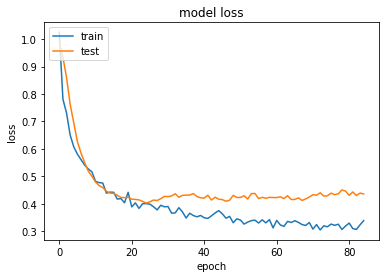
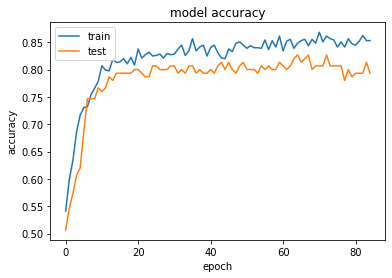
|  |  |
| --- | --- |
| Training Accuracy | Test Accuracy |
| 0.8388235294117647 | 0.8133333333333334 |

1. Random Forest Classifier:

|  |  |
| --- | --- |
| Training Accuracy | Test Accuracy |
| 0.9776470588235294 | 0.82 |

1. Multi Layer Neural Network using Keras

|  |  |
| --- | --- |
| Training Accuracy | Test Accuracy |
| 0.86 | 79.33 |



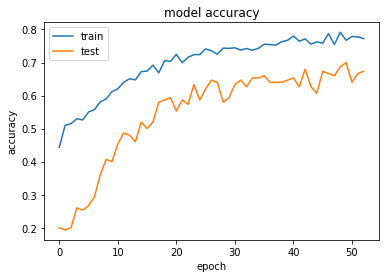
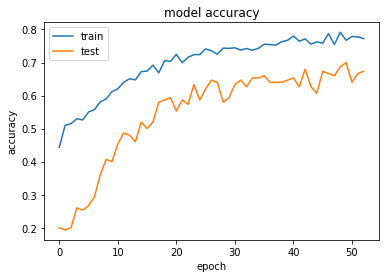


Modelling for MultiClass Classification

1. Data balancing is performed to add data to imbalanced classes using equal classes strategy
2. SVM, Random Forest Classifier, Neural Network provides average performance

|  |  |  |
| --- | --- | --- |
| Model | Training Accuracy | Test Acccuracy |
| SVM | 0.7522032141005702 | 0.42 |
| Neural Network | 0.8522550544323484 | 0.6333333333333333 |
| Random Forest Classifier | 0.7864178330741317 | 0.56 |

1. Multi Layer Neural Network using Keras

Training Accuracy: 0.8900984966303784

Test Accuracy: 0.6733333333333333

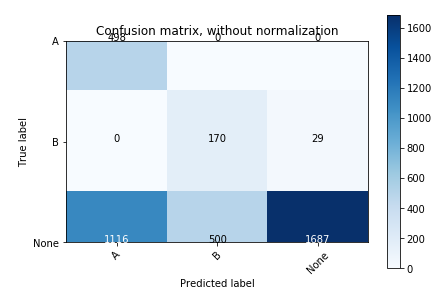
The model accuracy is still low



### Testing & Ensembles

1. Data is normalized
2. Perform the predictions using the model
3. Merge status predicted by classifier and investment score generated by Fuzzy Rule



### Results

Comparison of Different Models

A screenshot of a cell phone

Description automatically generated

AUC for Type - A

A close up of a map

Description automatically generated

AUC for Type - B

A close up of a map

Description automatically generated

Fuzzy Rules:

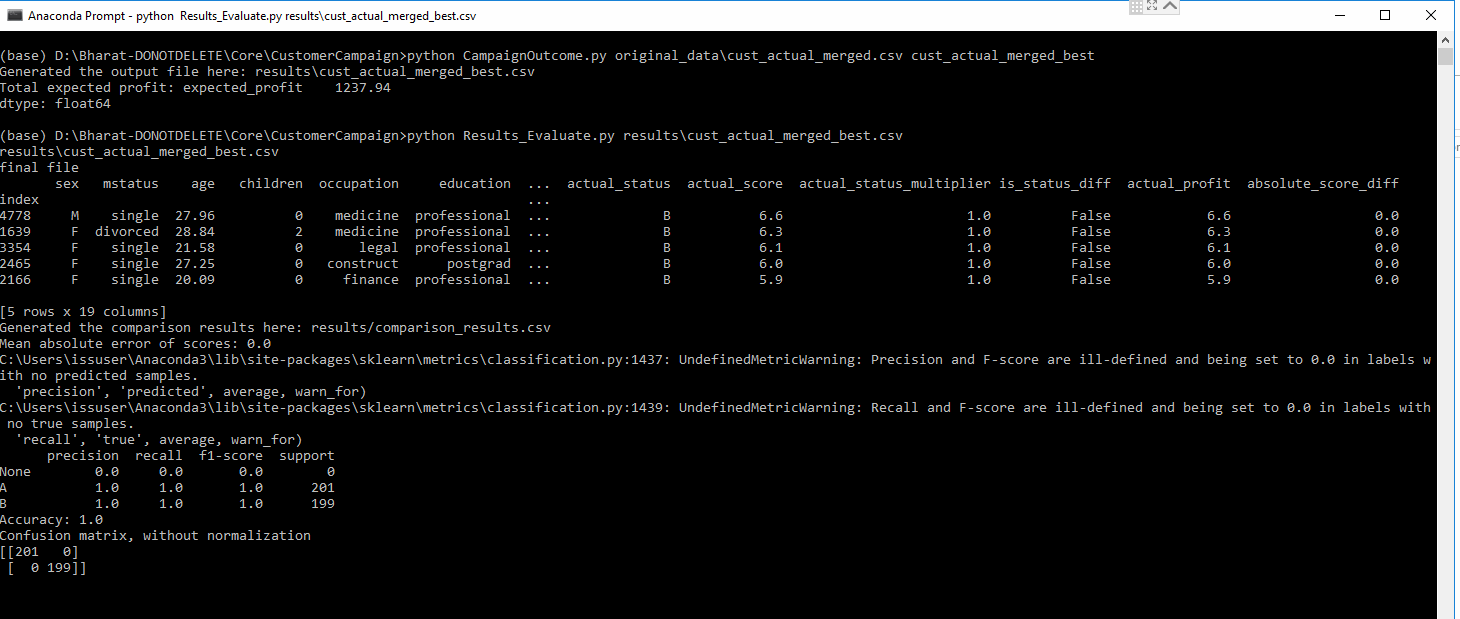


|  |  |
| --- | --- |
| Income -> Investment Score | Age -> Investment Score |
| Age -> Income\_Education Weighting with range (-0.5, 0.5) | Account Activity Fuzzy Rules |

<**python CampaignOutcome.py "original\_data\cust\_actual\_merged.csv\" cust\_actual\_merged\_best**> is used to sort the input data with the status and the score columns, generated from the classification model and expert system respectively and select the top 400 customers to be targeted for the campaign.

The file is generated in the `results` folder with the list of 400 customers

<**python Results\_Evaluate.py results\cust\_actual\_merged\_best.csv**> is used to compare the status and the scores from the input file and generate a confusion matrix and mean absolute error respectively. It also generates a CSV file inside <**results**> folder that has the difference for each record in the input.



**Actual Profit: 1237.94**

**Expected Profit: 1238**

From the results it can be seen that neural network has the best performance in terms of coverage of the data. Though K NN has the best AUC, it is not performing good in terms of the Recall and Precision. This possess a serious challenge as the important class misclassification result in bad campaign results.

Hence considering all these it can be seen that Neural network has performed well in terms of all the metrics. Also, looking at the final result, the results achieved is more or less perfect after normalising, balancing the data and combining with fuzzy rules.