

## 1. Background and Objectives

This dataset consists of part of passenger's information in Titanic which sank during her maiden voyage in 1912 because of colliding with an iceberg. The whole dataset consists of 1309 records and 12 variables which means there are 1309 passengers with some information about their identities could be found in the dataset such as age, gender and class of cabin...However, it is not a complete dataset due to many null values. Thus, there is still missing information about their ages or whether they were survived at last. Those null values will be the target in this task.

Thus, there are mainly three objectives. Firstly, ANN model will be applied **to predict the missing age** in the dataset during the data preparation stage. However, the final target of this task will be prediction on passenger's survival. Therefore, the dataset with filled nulls in first objective will be a training dataset for **predict passenger's survival** later. Four DM techniques including **Classification tree, Artificial neural network, Logistic regression and K-nearest neighbor** will be compared to obtain the best prediction performance. Finally, the model with best performance will be refined to **improve the performance**.

## 2. Data preparation

Before feeding dataset into each model, there are some procedures like data cleansing, data transformation and variable selection needed to be done. Data preparation is very important that not only can it ensure validation of dataset into the model (i.e. bugs may occur if data type is incorrect), it can also improve the performance of the prediction ability of the model. In this task, R and Excel will be the tools for data preparation.

### 2.1 Dataset description

<u>Column</u>	<u>Name</u>	<u>Variable type</u>	<u>Description</u>
1	PassengerId	Nominal	Id of passenger
2	*Survived	Binary	Survival (0 = No; 1 = Yes)
3	Pclass	Ordinal	Class of the cabin (1 = 1st; 2 = 2nd; 3 = 3rd)
4	Name	Nominal	Passenger name
5	Sex	Binary	Passenger gender
6	Age	Continuous	Passenger age
7	SibSp	Integer-valued	No. of Siblings or Spouses Aboard
8	Parch	Integer-valued	No. of Parents or Children Aboard
9	Ticket	Nominal	Ticket No. of passenger
10	Fare	Continuous	Passenger fare in British pound
11	Cabin	Nominal	Cabin No. of passenger Port of Embarkation
12	Embarked	Nominal	(C = Cherbourg; Q = Queenstown; S = Southampton)

Data dimension: 1309 x 12

Final target variable: Survived

```
> summary(d)
```

PassengerId	Survived	Pclass	Name	Sex	Age
Min. : 1	Min. :0.0000	Min. :1.000	Connolly, Miss. Kate	female:466	Min. : 0.17
1st Qu.: 328	1st Qu.:0.0000	1st Qu.:2.000	Kelly, Mr. James	male :843	1st Qu.:21.00
Median : 655	Median :0.0000	Median :3.000	Abbing, Mr. Anthony		Median :28.00
Mean : 655	Mean :0.3838	Mean :2.295	Abbott, Master. Eugene Joseph		Mean :29.88
3rd Qu.: 982	3rd Qu.:1.0000	3rd Qu.:3.000	Abbott, Mr. Rossmore Edward		3rd Qu.:39.00
Max. :1309	Max. :1.0000	Max. :3.000	Abbott, Mrs. Stanton (Rosa Hunt)		Max. :80.00
	NA's :418		(Other) :1301		NA's :263

SibSp	Parch	Ticket	Fare	Cabin	Embarked
Min. :0.0000	Min. :0.000	CA. 2343: 11	Min. : 0.000	:1014	: 2
1st Qu.:0.0000	1st Qu.:0.000	1601 : 8	1st Qu.: 7.896	C23 C25 C27 : 6	C:270
Median :0.0000	Median :0.000	CA 2144 : 8	Median : 14.454	B57 B59 B63 B66: 5	Q:123
Mean :0.4989	Mean :0.385	3101295 : 7	Mean : 33.280	G6 : 5	S:914
3rd Qu.:1.0000	3rd Qu.:0.000	347077 : 7	3rd Qu.: 31.275	B96 B98 : 4	
Max. :8.0000	Max. :9.000	347082 : 7	Max. :512.329	C22 C26 : 4	
		(Other) :1261		(Other) : 271	

The dimension of dataset is 1309 x 12 including 4 numeric variables: Age, SibSp, Parch, Fare and the rest are categorical variables. The target variable is in column 2 "Survived". From the summarized data from R output, there are columns with index function that contains too many levels like PassengerId, Name and Ticket. Besides, some null values could be found in the dataset like Survived, Age, Cabin and Embarked. These columns should be processed before feeding into DM models.

## 2.2Data processing

### Null value

For column "Embarked", it's null value percentage is very low(0.15%) so those 2 records will be label as the majority group "S". For column "Cabin", it is not considered to be put into the model as it contains too many null values(80%). For column "Age", it contains 263 null values which occupies a medium proportion of dataset. Null values from age column will be predicted by ANN model first.

### Dataset transformation

For those columns with index function, they will be removed from dataset because they have no meaning for prediction. However, there is a worth keeping information in column Name. From this column, a passenger's title can be extracted for knowing more about the identity of that passenger. Thus, title is extracted from name for later use. As there are still too many levels for column title, it is reshaped to reduce the levels that only 5 levels remained. This part of job is mostly done by Excel.

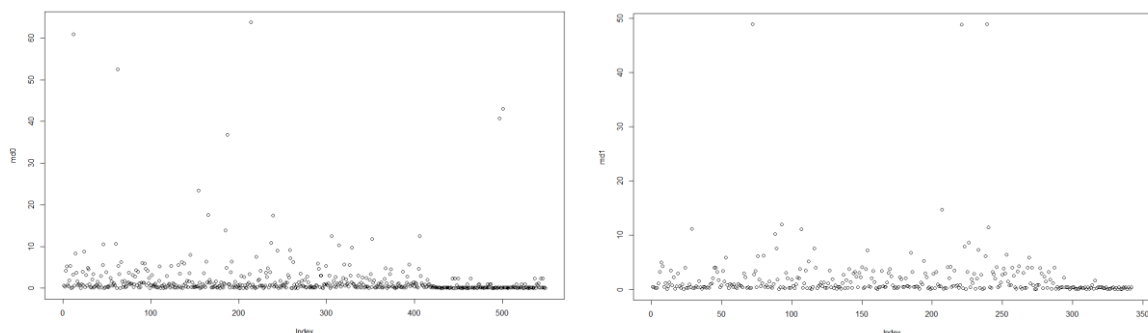
The screenshot shows an Excel spreadsheet with the 'Title' column extracted from the 'Name' column. The 'Title' column has 5 levels: Master, Miss, Mr, Mrs, and Others. The 'Summary' table shows the distribution of these titles.

Survived	Pclass	Sex	Age	Title
Min. :0.0000	Min. :1.000	female:466	Min. : 0.17	Master: 61
1st Qu.:0.0000	1st Qu.:2.000	male :843	1st Qu.:21.00	Miss :260
Median :0.0000	Median :3.000		Median :28.00	Mr :757
Mean :0.3838	Mean :2.295		Mean :29.88	Mrs :197
3rd Qu.:1.0000	3rd Qu.:3.000		3rd Qu.:39.00	Others: 34
Max. :1.0000	Max. :3.000		Max. :80.00	
NA's :418			NA's :263	

## 2.3 Outlier detection

As some statistical techniques are sensitive to outliers, such as logistic regression. Outlier detection will be processed before applying statistical model. Numeric variables like Age and Fare will undergo outlier detection and the outlier will be removed from the dataset. As only logistic regression needs the process in this task, so this process output will be saved as separate training dataset. Other models which are not sensitive to outliers will not use this output dataset.

In this dataset, continuous variables "Age" and "Fare" are undergo outlier detection. 27 outlier records are discovered as shown in the diagram below. These records will be removed from dataset if the statistical model is sensitive to outliers.



## 2.4 Selected dataset

After data transformation, the data dimension becomes 864 x 9 if outliers are considered or 891 x 9 if outliers are not considered. Column PassengerId, Name, Ticket, and Cabin are removed. Column title is created by extracting Name. Therefore only 9 variables will be feed into prediction models.

Outcome of processed dataset:

```
> summary(d)
Survived Pclass      Sex      Age      SibSp      Parch
0:549    1:216   female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
1:342    2:184   male  :577   1st Qu.:22.00   1st Qu.:0.000   1st Qu.:0.0000
          3:491           Median :29.69   Median :0.000   Median :0.0000
                      Mean   :29.76   Mean   :0.523   Mean   :0.3816
                      3rd Qu.:36.00   3rd Qu.:1.000   3rd Qu.:0.0000
                      Max.   :80.00   Max.   :8.000   Max.   :6.0000

      Fare      Embarked  Title
Min.   : 0.00   C:168   Master: 40
1st Qu.: 7.91   Q: 77   Miss  :182
Median :14.45   S:646   Mr    :517
Mean   :32.20           Mrs   :125
3rd Qu.:31.00           Others: 27
Max.   :512.33
```

```
d      891 obs. of 9 variables
Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 2 2 ...
Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 1 3 3 2 3 ...
Sex : int 2 1 1 1 2 2 2 1 1 1 ...
Age : num 22 38 26 35 35 54 2 27 14 4 ...
SibSp : int 1 1 0 1 0 0 3 0 1 1 ...
Parch : int 0 0 0 0 0 0 1 2 0 1 ...
Fare : num 7.25 71.28 7.92 53.1 8.05 ...
Embarked: int 3 1 3 3 3 3 3 3 1 3 ...
Title : int 3 4 2 4 3 3 1 4 4 2 ...
```

### 3. Analysing techniques

Classification DM techniques will be applied to predict whether a passenger is survived from the catastrophe or not as it is a classification problem and the target variable is a binary output. Classification tree, Artificial neural network, Logistic regression and K-nearest neighbour will be applied to compare the performance of the model based on this dataset. The chosen four techniques are all supervised learning techniques because the dataset contains the target label, the models are trained and will recognize the pattern with known information about passenger's survival.

In each techniques, there will be a slight difference in data transformation. For example, some may need scaling while others may need outlier detection. Therefore data transformation will be done case by case depending on which model is used. However, most of them will apply resampling method of k-fold cross validation in order to get a better assessment of the testing dataset result.

#### 3.1 Artificial neural network (Logistic output)

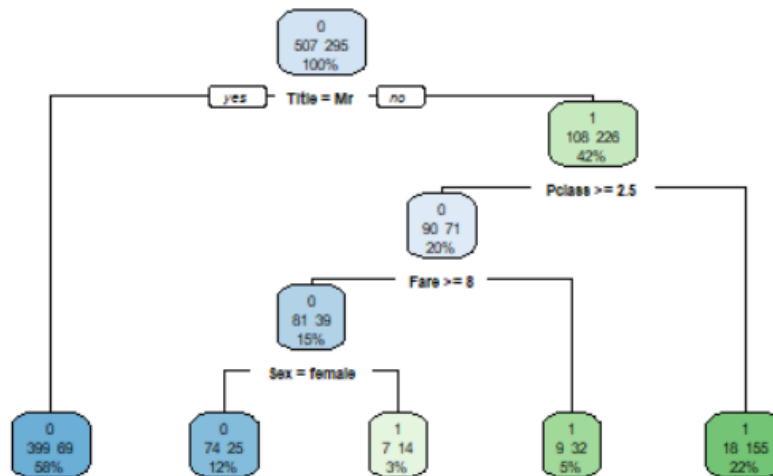
```
> summary(titanictrain.nn)
a 8-3-1 network with 31 weights
options were - entropy fitting
b->h1 i1->h1 i2->h1 i3->h1 i4->h1 i5->h1 i6->h1 i7->h1 i8->h1
92.46 -55.83 -13.61 -2.25 -7.57 -22.45 2.35 18.29 -2.25
b->h2 i1->h2 i2->h2 i3->h2 i4->h2 i5->h2 i6->h2 i7->h2 i8->h2
-140.24 18.39 75.25 0.03 0.94 6.60 -0.05 7.93 -7.19
b->h3 i1->h3 i2->h3 i3->h3 i4->h3 i5->h3 i6->h3 i7->h3 i8->h3
17.08 7.37 25.13 -18.22 -46.49 21.67 0.58 8.07 7.71
b->o h1->o h2->o h3->o
1.27 1.98 -3.53 4.52
```

Training classification table: 0 1 0 483 125 1 14 188	Training error rate $(125+14)/(139+188+483)$ = 0.1716049
Testing classification table: 0 1 0 48 8 1 4 21	Testing error rate $(4+8)/(4+8+48+21)$ =0.1481481

Data transformation in this model includes Scaling of continuous variable "Age" and "Fare", as well as recoding character vector to numeric vector like "Sex", "Embarked" and "Title". This step can ensure equal importance of variable in the input layer. It has logistic output so the target variable is transformed into factor object. K fold cross validation is applied in resembing process. The value of K is set to be 10 this time.

The 8- 3 -1 ANN model being trained has 8 input layers, 3 hidden layers and 1 output layer with 31 parameters in the model. There are several trials to test how many hidden layers would give the best result. It shows the best result based on error rate assessment when hidden layer is 3. From the classification table, outcome of this model is quite good that training error rate is 17 % and the testing error rate 15%.

### 3.2 Classification tree



n= 802

node), split, n, loss, yval, (yprob)  
\* denotes terminal node

- 1) root 802 295 0 (0.6321696 0.3678304)
- 2) Title=Mr 468 69 0 (0.8525641 0.1474359) \*
- 3) Title=Master, Miss, Mrs, Others 334 108 1 (0.3233533 0.6766467)
- 6) Pclass>=2.5 161 71 0 (0.5590062 0.4409938)
- 12) Fare>=8.0396 120 39 0 (0.6750000 0.3250000)
- 24) Sex=female 99 25 0 (0.7474747 0.2525253) \*
- 25) Sex=male 21 7 1 (0.3333333 0.6666667) \*
- 13) Fare< 8.0396 41 9 1 (0.2195122 0.7804878) \*
- 7) Pclass< 2.5 173 18 1 (0.1040462 0.8959538) \*

<b>Training classification table:</b> c   0   1 1 434 83 2 51 234	<b>Training error rate</b> $(51+83)/(51+83+234+434)$ = 0.1670823
<b>Testing classification table:</b> pr1 0 1 0 56 5 1 8 20	<b>Testing error rate</b> $(5+8)/(5+8+56+20)$ = 0.1460674

Classification tree is a robust technique that it does not require any distributional assumptions and it is not sensitive to outliers. Thus, seldom data transformation job needed to be done. Random forest which is a resembling method is adopted in order to get a more accurate assessment.

From R output diagrams, there are 6 decision rules obtained from classification tree. Those rules reveal the key factors that determine passenger's survival are title, pclass, fare and sex. The rule with best support is when Title not equal to "Mr", this rule occupy 58% of whole dataset. Confidence is 85% which has high accuracy in this rule. Those key factor information may be useful in later refinement of model training.

From the classification table, outcome of this model is quite the same as ANN output that training error rate is 16.7 % and the testing error rate 14.6%.

### 3.3 Logistical regression

```
Call:
glm(formula = Survived ~ ., family = binomial, data = x1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0992  -0.5225  -0.3709   0.5324   2.5510

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  20.323164  547.817748   0.037  0.970407
Pclass      -0.830439   0.189720  -4.377  1.20e-05 ***
Sexmale     -15.591949  547.817176  -0.028  0.977294
Age         -0.035564   0.010705  -3.322  0.000893 ***
SibSp       -0.784685   0.153513  -5.112  3.20e-07 ***
Parch       -0.397727   0.148785  -2.673  0.007514 **
Fare        0.025789   0.006953   3.709  0.000208 ***
EmbarkedQ    0.299066   0.427470   0.700  0.484165
EmbarkedS   -0.343207   0.282922  -1.213  0.225099
TitleMiss   -16.368631  547.817468  -0.030  0.976163
TitleMr     -3.605216   0.605181  -5.957  2.57e-09 ***
TitleMrs    -15.489854  547.817519  -0.028  0.977442
TitleOthers -3.538253   0.850666  -4.159  3.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1045.18  on 784  degrees of freedom
Residual deviance:  611.53  on 772  degrees of freedom
AIC: 637.53

Number of Fisher Scoring iterations: 14
```

Training classification table: pred3 0 1 0 426 74 1 58 227	Training error rate $(58+74)/(74+58+426+227)$ = 0.1681529
Testing classification table: pred 0 1 0 42 11 1 5 21	Testing error rate $(5+11)/(5+11+42+21)$ = 0.2025316

Logistic regression is a parametric method that it requires distributional assumption and is sensitive to outliers so dataset without outliers will be input into this model. K fold cross validation is also applied in resembling process like when training ANN.

There are 12 parameters in the logistic regression, but some parameters are with high p values. These parameters may be taken away in further refinement as they are not significant factors in the prediction. Whereas the key factors obtained by logistic regression is "Pclass", "Age", "SibSp", "Parch", "Fare" and when "Title" is Mr or Others. This finding is quite consistent with Classification tree's result.

From the classification table, the training error rate is 17% and the testing error rate is 20%. The performance of logistic regression is lower than that of ANN or classification tree.

### 3.4 K-nearest neighbour

<pre>&gt; titanic_knn&lt;-k_nn(z1,z2,c1,d2[,1],v=20 k= 1 error rate= 0.4044944 k= 2 error rate= 0.3258427 k= 3 error rate= 0.3258427 k= 4 error rate= 0.3707865 k= 5 error rate= 0.2696629 k= 6 error rate= 0.258427 k= 7 error rate= 0.258427 k= 8 error rate= 0.2359551 k= 9 error rate= 0.247191 k= 10 error rate= 0.247191 k= 11 error rate= 0.2359551 k= 12 error rate= 0.258427 k= 13 error rate= 0.258427 k= 14 error rate= 0.247191 k= 15 error rate= 0.258427 k= 16 error rate= 0.258427 k= 17 error rate= 0.247191 k= 18 error rate= 0.2696629 k= 19 error rate= 0.258427 k= 20 error rate= 0.258427 best k= 8 error rate= 0.2359551</pre>	Testing classification table: <pre>0 1 0 53 10 1 11 15</pre>
	Testing error rate $(10+11)/(10+11+53+15)$ $= 0.2359551$

K-nearest neighbour is a distance-based method, so it requires scaling of dataset for both categorical variables and continuous variables. It is also robust to outliers, so a normal cleansed dataset is enough to train the model. This model is a lazy learning technique so rather than applying a trained model, it make prediction using the training dataset when a testing record is available. As for this reason, training error rate is not able to be obtained so only testing error rate could be output. Also, cross validation is skipped for this lazy learner. In order to obtain the best k values, an improved knn function is applied and the best value of k is 8 in this case.

From testing dataset classification table, the error rate is 23.6% and it is the highest among four models in this dataset.

### 3.5 Comparison between four models in titanic dataset

Data Mining Technique	Training error rate (%)	Testing error rate (%)
Artificial neural network	17	15
Classification tree	16.7	14.6
Logistical regression	17	20
K-nearest neighbour	NA	23.6

From the comparison table, K-nearest neighbour has the lowest performance in testing error rate. For the other three models, they have similar performance in training dataset as they all have training error rate around 17%. However, Artificial neural network and Classification tree stand out when encountering testing dataset. They both have testing error around 15% while logistic regression's testing error is 20%.

In conclusion, Artificial neural network and Classification tree perform the best in this dataset.

#### 4. Refinement of Random Forest

In order to optimize the performance, a second trial will be conducted. As the best model for titanic dataset among four models is Random forest. Therefore classification tree will be the only focus in the following refinement. Dataset will be reshaped and variables will be selected again in a bid to improve the performance. Also, Lift chart will be used for assessment apart from error rate.

##### 4.1 Reshape dataset

From logistic regression and classification tree output in the previous stage, they both indicate column SibSp is not a significant factor, so it will be removed from dataset.

The continuous variable Age will be transformed into ordinal variables by banding. Age under 18 is group 1, Age between 18 to 30 is group2, Age between 30 to 40 is group3 and Age above 40 is group 4. This banding method takes reference from column statistics as below

Transformation of column age by banding:

Age	Age Group Banding	Age Group frequency count
Min. : 0.42	Group1 : below 18	<b>AgeGroup</b> <b>1:154</b>
1st Qu.:22.00	Group2 : 18 - 30	<b>2:584</b>
Median :29.69	Group2 : 30 - 40	<b>3:240</b>
Mean :29.76	Group6 : above 40	<b>4:331</b>
3rd Qu.:36.00		
Max. :80.00		

Summary of reshaped dataset:

Survived	Pclass	Sex	AgeGroup	Parch
0:549	1:216	female:314	1:113	Min. :0.0000
1:342	2:184	male :577	2:440	1st Qu.:0.0000
	3:491		3:175	Median :0.0000
			4:163	Mean :0.3816
				3rd Qu.:0.0000
				Max. :6.0000
Fare	Embarked	Title		
Min. : 0.00	C:168	Master: 40		
1st Qu.: 7.91	Q: 77	Miss :182		
Median : 14.45	S:646	Mr :517		
Mean : 32.20		Mrs :125		
3rd Qu.: 31.00		others: 27		
Max. :512.33				

Now, the dataset dimension is reduced to 891 x 8, and there are only two continuous variables in the dataset including Parch and Fare. The rest of variables are categorical variables. Those categorical variables with a reasonable amount of levels.

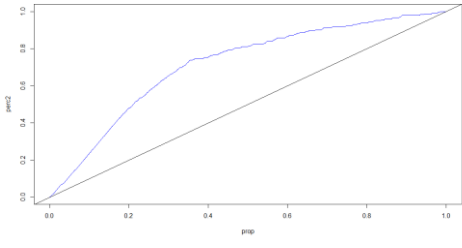
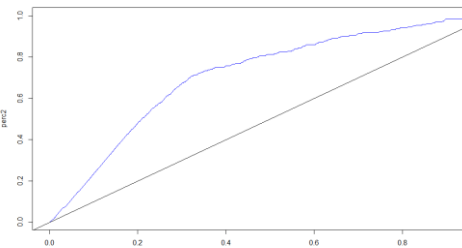


## 4.2 Comparison

Classification table of refined model training

Training classification table: c 0 1 1 452 94 2 33 223	Training error rate $(94+33)/(94+33+452+223)$ = 0.1583541
Testing classification table: pr1 0 1 0 62 7 1 2 18	Testing error rate $(2+7)/(2+7+62+18)$ = 0.1011236

Comparison of random forest performance before and after refinement

	Before refinement	After refinement
Training error rate	16.7%	15.8%
Testing error rate	14.6%	10%
Lift chart		

From the comparison table, the training error rate has improved a bit from 16.7% to 15.8 %. As there is only 1 % improvement so it is hard to observe to change by the Lift chart. However, testing error rate do improve a lot by 14.6 % error rate before and drop to 10 % after refinement.

In conclusion, methods of **refinement can be traced back to data preparation stage and model trials**. Data preparation like processing null values, variable selection data transformation can have impact on the final prediction performance. Besides, some models may not be the best to give final prediction, but they can contribute in better prepare dataset. For example, they can show key factors like logistic regression or classification tree. On the other hand, models can also help to predict null values in data preparation stage. All these can make the data preparation more sophisticated, and hence improve the final prediction model performance indirectly. **Thus, making use of data mining techniques in different stages can utilize data mining power and get the optimum result in the end.**

## 5. Reference

Data source: <https://www.kaggle.com/c/titanic>

Document list

Folder	File Name	Format	Description
Data preparation	Raw data from Kaggle	file	Dataset downloaded from Kaggle.com
	Titanic_whole_dataset	CSV	Combined and transformed data by Excel
	FillNa_age	R	Process null values of Age column
	fill_na	CSV	Output of "FillNa_age.R" It is cleansed dataset which will be input to DM models
	Dect_outlier	R	Detect outliers of dataset
	fill_naNout	CSV	Output of "Dect_outlier R" It is cleansed dataset which will be input to logistic regression
Model training	ann_final	R	Artificial neuron network program
	ctree_final	R	Classification tree program
	knn_final	R	K nearest neighbour program
	lreg_final	R	Logistic regression program
Refinement	fill_naNreshape	CSV	Dataset in fill_na.csv is reshaped by Excel
	refinement_ctree	R	Refined random forest
Final report	Final report	doc	Written report