CS 6375.001 – MACHINE LEARNING

Titanic – Machine Learning from Disaster

Project Report
Fall 2016
November 28, 2016

Eswar Chowdary Ganta (exg151430)
Sai Charan Rao Vennamanani (sxv157130)
Dileep Gudena (dxg161730)
Divya Reddy Vudem (dxv151430)
Santhosh Kumar Kamishetty (sxk165130)

In this project, we aim at making a complete analysis of the TITANIC dataset to find what sorts of people were more likely to survive the shipwreck.

The Dataset - TITANIC

The TITANIC dataset is taken from an active Kaggle Competition and the link of which is given below:

https://www.kaggle.com/c/titanic/data?train.csv

- Number of attributes = 11 (including the class attribute)
- Number of instances = 891

The attributes are the following:

1. Survival – 0 if not survived and 1 if survived
2. Pclass – Passenger Class (1 - Upper, 2 - Middle and 3 – Lower class)
3. Name – Name of the passenger
4. Sex – Gender of the passenger
5. Age – Age of the passenger
6. Sibsp – Number of siblings/spouses aboard
7. Parch – Number of Parents/Children aboard
8. Ticket – Ticket Number
9. Fare – Ticket fare
10. Cabin – Cabin number
11. Embarked – Port of Embarkation (C, Q and S)

Techniques we have used:

- Decision Tree
- Support Vector Machine
- K-NN
- Random Forest
- Boosting

Experimental Methodology

- We employ the following procedure in our project –
  1. Pre-processing of the dataset
     - This step involves dealing with the NA values,
     - Selecting the attributes that influence the classification by observing the histograms and correlation plots,
     - Scaling the required attributes
  2. On the dataset

   - We perform each of the aforementioned techniques,
   - Also, vary the parameters and find the best one for the technique.
3. We evaluate the techniques using the following metrics –
   - Accuracy
   - Precision
   - Recall
   - F-measure
4. We plot the results that aid in comparing the performance of the classifiers.

Pre Processing

We now present the result of the work we've done so far.

- We removed the following attributes (after initial examination of the dataset) from the dataset as they don't impact the result significantly –
  - Passenger Number (This is just a serial number)
  - Name (Passenger name has nothing to do with his/her survival)
  - Ticket Number (Doesn't have relevance as it is just a booking ID to identify the survivor)
  - Ticket Fare (Just the cost of the ticket)
  - Cabin (Pclass already encodes the information about the cabin)
- We categorized the AGE attribute into three different intervals as follows:
  - Category_1 – 18 and below (Children)
  - Category_2 – Between 19 and 40
  - Category_3 – 41 and above

  We replaced the NAs in the age attribute with the average value of the age.

- The attribute SEX is labelled as follows:
  MALE – 1
  FEMALE – 2

Packages Used

Classifier	Package
Decision Tree	rpart, caret
SVM	e1071
Random Forest	randomForest
K-NN	class
Ada - Boosting	adabag

DECISION TREE

We have used Decision tree classifier. It is used to find what category of people have survived from the Titanic data. The parameters used in decision tree are MinSplit, MaxDepth, MinBucket, Cp. The following parameters have been changed in the experiments:

MinSplit - For a Split to be occurred, the minimal no. of observations needed that must exist in a node is the value of MinSplit.
MinBucket - The minimum number of observations in terminal node.
CP - Cp is Complexity Parameter.
MaxDepth - The Maximum depth of any node of the final tree.
We performed 10-Fold validation to obtain correct results.

The results of experiments are like the following:

Classifier	**Fold**	**CP**	**Minsplit**	**maxdepth**	**minbucket**	**Accuracy**	**Precision**	**Recall**	**FScore**
DT	10	0.001	2	1	1	81.005	0.789	0.797	0.793
DT	10	0.001	3	1	1	81.005	0.789	0.797	0.793
DT	10	0.001	4	1	1	81.005	0.789	0.797	0.793
DT	10	0.001	5	1	1	81.005	0.789	0.797	0.793
DT	10	0.001	10	1	1	81.005	0.789	0.797	0.793
DT	10	0.001	20	1	1	81.005	0.789	0.797	0.793
DT	10	0.001	20	2	1	81.005	0.789	0.797	0.793
DT	10	0.001	20	3	1	83.798	0.792	0.862	0.826

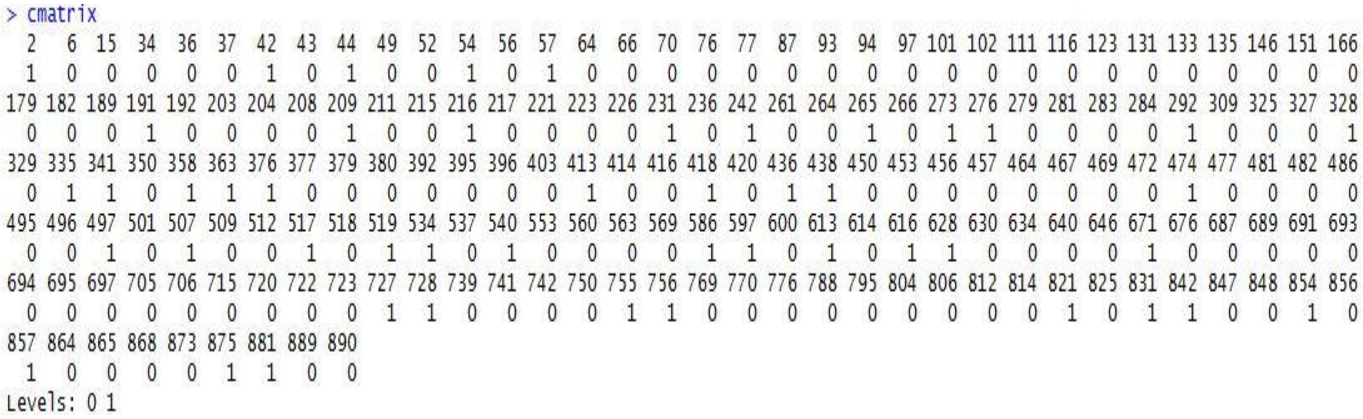
DT	10	0.001	20	3	2	83.798	0.792	0.862	0.826
DT	10	0.001	20	3	3	83.798	0.792	0.862	0.826
DT	10	0.001	20	5	1	84.357	0.803	0.860	0.826
DT	10	0.001	20	5	3	84.357	0.803	0.860	0.826
DT	10	0.001	20	5	5	84.357	0.803	0.866	0.826
DT	10	0.001	15	2	2	81.005	0.789	0.797	0.793
DT	10	0.001	15	3	3	83.798	0.792	0.862	0.826
DT	10	0.00001	20	5	1	84.357	0.803	0.860	0.826
DT	10	0.001	20	7	5	83.798	0.802	0.845	0.823
DT	10	0.00001	20	7	5	83.798	0.802	0.845	0.823
DT	10	0.001	25	5	5	84.357	0.802	0.845	0.823
DT	10	0.001	25	7	5	83.798	0.802	0.845	0.823

After the experiments are done, we found out that the measure values (accuracy, precision, recall, fscore) are greater for the tree with minsplit=20, maxdepth=5, minbucket=5 and cp=0.001.

The following are the values obtained:

CLASSIFIER	**DECISION TREE**
No. of Folds	10
Accuracy	84.357%
Precision	0.803
Recall	0.866
F-Score	0.826

Using the above parameter values, The predicted class labels are as follows:



```
> cmatrix
  2   6  15  34  36  37  42  43  44  49  52  54  56  57  64  66  70  76  77  87  93  94  97 101 102 111 116 123 131 133 135 146 151 166 
  1   0   0   0   0   0   1   0   1   0   0   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 
179 182 189 191 192 203 204 208 209 211 215 216 217 221 223 226 231 236 242 261 264 265 266 273 276 279 281 283 284 292 309 325 327 328 
  0   0   0   1   0   0   0   0   1   0   0   1   0   0   0   0   1   0   1   0   0   1   0   1   1   0   0   0   0   1   0   0   0   1 
329 335 341 350 358 363 376 377 379 380 392 395 396 403 413 414 416 418 420 436 438 450 453 456 457 464 467 469 472 474 477 481 482 486 
  0   1   1   0   1   1   1   0   0   0   0   0   0   0   1   0   0   1   0   1   1   0   0   0   0   0   0   0   0   1   0   0   0   0 
495 496 497 501 507 509 512 517 518 519 534 537 540 553 560 563 569 586 597 600 613 614 616 628 630 634 640 646 671 676 687 689 691 693 
  0   0   1   0   1   0   0   1   0   1   1   0   1   0   0   0   0   1   1   0   1   0   1   1   0   0   0   0   1   0   0   0   0   0 
694 695 697 705 706 715 720 722 723 727 728 739 741 742 750 755 756 769 770 776 788 795 804 806 812 814 821 825 831 842 847 848 854 856 
  0   0   0   0   0   0   0   0   0   1   1   0   0   0   0   1   1   0   0   0   0   0   0   0   0   0   1   0   1   1   0   0   1   0 
857 864 865 868 873 875 881 889 890 
  1   0   0   0   0   1   1   0   0 
Levels: 0 1
```

We note from the above table that 44 persons of the test sample were among those survived the wreck and are predicted correctly by the classifier.

After the observation from the instances obtained the following are the results obtained:

Pclass	Middle class people survived the most.
Sex	Majority of the people survived are females.
Age	The survived peoples age range is between 18 and 40
SibSp	Majority of the survived peoples have a sibling.
Parch	Majority of the survivals do not have parents.
Embarked	Majority of the survivors were from the Southampton Port of Entry.

From the above results obtained the majority of the people survived from the Titanic wreck are the ***Middle Class female adults with a sibling***.

Support Vector Machine

From the results obtained in SVM for a n fold cross validation, based on the accuracy, precision, recall and F-Measure values, the ***radial*** kernel is giving the best results for the SVM model that we created. The parameters that we have taken into consideration are ***cost, kernel, gamma*** and ***tolerance***.

For a linear kernel, the measure of accuracy, precision, recall or F-measure are almost similar irrespective of the value change in cost and tolerance. We cannot place a gamma value for a linear kernel, because the gamma value comes into picture only if the kernel is not a linear kernel.

Expt #	**Fold**	**Parameter1 Cost**	**Parameter2 kernel**	**Parameter3 gamma**	**Parameter4 tolerance**	**Average Accuracy(in %)**	**Average Precision**	**Average Recall**	**F-measure**
1	10	10	linear	-	0.01	78.69	0.767	0.776	0.771
2	10	12	linear	-	0.05	78.74	0.767	0.776	0.772
3	10	18	linear	-	0.05	78.55	0.769	0.776	0.772
4	**10**	**20**	linear	-	**0.01**	**78.73**	**0.769**	**0.780**	**0.775**
5	10	15	linear	-	0.01	78.59	0.776	0.779	0.773

In polynomial kernel, if we are either increasing the cost or the value of gamma, the time required for processing the algorithm is huge comparatively. We are getting the best results in polynomial for cost 10, gamma 0.25 and tolerance 0.01. If we closely observe the values for polynomial kernel, the measures like accuracy, F-measure, precision and recall are better when the gamma values are less comparatively.

Expt #	Fold	Parameter1 Cost	Parameter2 kernel	Parameter3 gamma	Parameter4 tolerance	Average Accuracy(in %)	Average Precision	Average Recall	F-measure
1	10	10	polynomial	0.25	0.01	81.36	0.788	0.813	0.800
2	10	12	polynomial	0.33	0.05	80.96	0.783	0.807	0.795
3	10	18	polynomial	0.5	0.05	79.14	0.762	0.792	0.776
4	**10**	**15**	polynomial	0.67	**0.01**	**80.23**	**0.779**	**0.797**	**0.788**
5	10	10	polynomial	0.75	0.01	79.40	0.768	0.790	0.779
6	10	20	polynomial	0.33	0.05	81.01	0.784	0.805	0.794
7	10	12	polynomial	0.67	0.01	78.64	0.763	0.784	0.773
8	10	15	polynomial	0.25	0.05	79.66	0.769	0.788	0.778
9	10	18	polynomial	0.5	0.01	79.99	0.774	0.796	0.784
10	10	20	polynomial	0.5	0.01	78.40	0.760	0.774	0.767

Radial kernel is producing consistent and better measures of precision, recall, accuracy and F-measure. With the decrease in gamma value and increase in cost value, the measures are increasing. If we closely observe, the tolerance value effects the measures of the experiment. Lower the value of tolerance, better the values of accuracy, precision, recall and F-measures.

Expt #	Fold	Parameter1 Cost	Parameter2 kernel	Parameter3 gamma	Parameter4 tolerance	Average Accuracy(in %)	Average Precision	Average Recall	F-measure
1	10	10	radial	1	0.001	81.25	0.783	0.815	0.798
2	10	15	radial	0.5	0.01	81.31	0.782	0.816	0.799
3	10	18	radial	0.25	0.05	80.59	0.775	0.805	0.790
4	**10**	**12**	**radial**	**0.33**	**0.01**	**80.31**	**0.772**	**0.805**	**0.788**
5	10	20	radial	0.45	0.001	80.52	0.774	0.805	0.789
6	10	12	radial	0.67	0.01	81.19	0.780	0.817	0.798
7	10	15	radial	0.25	0.05	81.10	0.788	0.814	0.801
8	10	10	radial	0.33	0.01	80.91	0.778	0.813	0.795
9	10	18	radial	0.09	0.01	81.22	0.782	0.813	0.795
10	10	20	radial	0.25	0.05	80.18	0.769	0.806	0.787

Out of all the observations that we have observed, we can closely relate the data as:

Lower the values of gamma, tolerance and higher the value of cost we are obtaining the best results in the respective kernels.

Test Results:

svm.pred	Pclass	Sex	Age	SibSp	Parch	Embarked
1	3	2	1	0	0	4
1	2	2	3	0	0	4
0	2	1	2	0	0	4

0	2	1	2	0	0	4
0	1	1	2	0	0	4
1	3	2	2	0	0	3
1	1	2	2	1	0	2
1	1	1	2	1	0	2
0	3	2	2	1	0	4
0	3	1	2	1	1	2
0	3	1	2	1	0	2
1	2	2	1	0	0	4
0	3	2	2	0	0	4
0	3	1	2	0	0	4
0	3	2	2	1	0	4
0	1	1	2	0	1	2
0	2	1	2	1	0	2
0	3	1	2	0	0	4
1	1	2	2	0	2	4
0	3	2	2	0	0	4
0	2	1	1	0	0	4
0	3	1	2	0	0	4
0	3	1	3	0	2	4
1	3	1	1	0	2	4
0	2	1	2	0	0	4
0	3	1	2	0	0	4
0	3	1	2	0	0	2
0	2	1	2	0	0	4
0	3	1	2	0	0	4
1	1	1	3	2	0	3
0	1	1	2	1	1	4
0	3	1	2	0	0	3
0	1	1	3	1	1	4
0	1	1	2	0	0	4
1	1	2	2	0	0	4
0	3	1	2	0	0	4
1	1	2	2	1	0	2
1	2	2	2	1	1	4
1	2	2	2	1	0	4
0	3	1	2	0	0	4
0	3	1	2	0	0	4
1	1	2	2	0	0	2
0	3	1	3	0	0	4

0	3	1	3	0	0	4
1	3	2	2	0	0	3
0	3	1	1	5	2	4
0	3	2	2	0	0	4
0	3	1	2	0	0	2
0	1	1	2	0	0	4
0	3	2	2	1	0	4
0	3	1	2	0	0	2
1	2	2	3	0	0	4
1	2	2	2	1	2	4
1	2	2	2	0	0	2
0	1	1	2	0	0	4
0	3	1	2	0	0	4
0	2	1	2	0	0	4
0	3	1	2	0	0	4
1	3	2	2	0	0	3
0	1	1	2	0	0	4
1	3	1	1	1	1	2
0	3	1	2	0	0	2
0	3	1	2	0	0	4
0	2	1	3	0	0	4
0	2	1	2	1	0	4
0	3	1	3	0	0	4
1	1	1	3	1	0	2
0	1	1	2	0	0	4
0	1	1	2	0	0	4
0	3	1	2	0	0	4
0	3	1	2	0	0	4
0	2	1	2	1	1	4
0	3	1	2	1	0	4
1	3	2	1	0	0	3
0	1	1	3	0	0	4
0	1	1	2	0	0	2
0	3	1	3	0	0	4
1	2	2	3	0	0	4
0	1	1	3	1	0	4
1	1	2	2	0	0	4
0	1	1	2	0	0	2
0	3	1	2	0	0	3
0	3	1	2	0	0	4

0	3	1	2	0	0	2
0	3	1	1	0	0	4
0	1	1	2	0	0	4
0	3	1	2	0	0	2
1	2	1	1	0	2	2
0	3	1	2	0	0	3
0	3	1	2	0	0	4
0	3	1	3	0	0	4
1	1	2	2	1	0	2
1	1	2	3	0	1	2
0	3	1	2	0	0	3

- From the results of testing data, svm model has predicted 27 survivals out of 94 instances.
- Ratio of number of males survived to number of females survived is 6:21
- Age 0-18:19-40:41+ ratio of survival is 6:15:6

The above 2 attributes show significant differences that can be seen from the results. The statistics show that the ***females*** and the ***middle-aged*** people had more chances of survival, compared to the other categories. Even, the training data set shows similar significant results.

KNN

KNN Classifier:

We have considered KNN classifier for our analysis where we considered k nearest training examples in the feature space for classification of people who have survived in the Titanic using 10 cross fold validation.

KNN classifier Parameters:

K value: Number of neighbors considered in Feature Space

In KNN classifier, we have tabulated the results (Accuracy, Precision, Recall and F-Score) as below changing k parameter.

No.	Fold	K value	Accuracy	Precision	Recall	F-Score
1	10	5	79.9214	0.79839	0.76611	0.78172
2	10	10	80.5472	0.80741	0.77450	0.79036

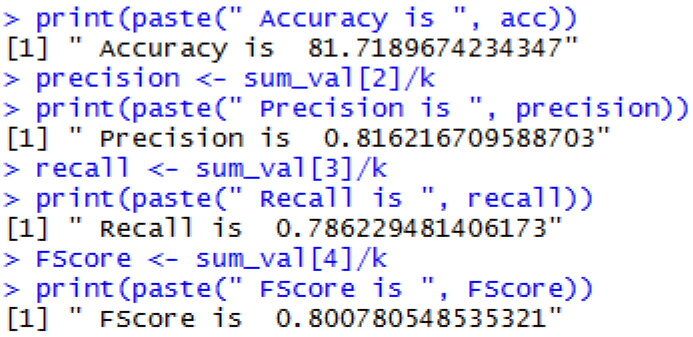
3	10	15	81.7189	0.81621	0.78622	0.80078
4	10	20	79.3315	0.80191	0.75908	0.77948
5	10	25	79.2417	0.80079	0.75836	0.77860
6	10	30	78.24686	0.79191	0.74390	0.76694
7	10	35	77.22693	0.77367	0.734161	0.75330
8	10	40	76.87979	0.77179	0.72920	0.74980
9	10	45	76.8940	0.77988	0.73241	0.75513
10	10	50	75.8946	0.76128	0.71507	0.73735

From the above observations, maximum accuracy is obtained for k value as 15 and below is the summary of the result set considered.

Classifier	**KNN**
Number of Folds in Cross Validation	10
Accuracy	81.7189
Precision	0.81621
Recall	0.78622
F-Score	0.80078

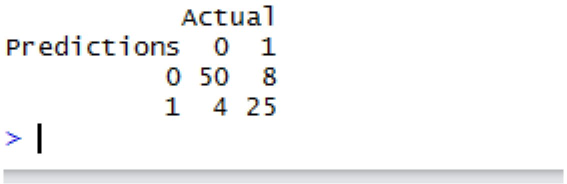
We now examine the class labels of the prediction to analyze the result of the classifier. The following R – snapshot gives us the picture of the class labels of prediction on the test set.

Results for KNN(K=15):



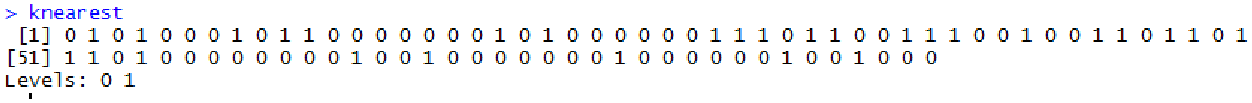
```
> print(paste(" Accuracy is ", acc))
[1] " Accuracy is  81.7189674234347"
> precision <- sum_val[2]/k
> print(paste(" Precision is ", precision))
[1] " Precision is  0.816216709588703"
> recall <- sum_val[3]/k
> print(paste(" Recall is ", recall))
[1] " Recall is  0.786229481406173"
> FScore <- sum_val[4]/k
> print(paste(" FScore is ", FScore))
[1] " FScore is  0.800780548535321"
```

Prediction Table for KNN(k=15):



```
             Actual
Predictions   0  1
           0 50  8
           1  4 25
> |
```

Prediction of Class Labels:



```
> knearest
 [1] 0 1 0 1 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 1 1 0 1 1 0 0 1 1 1 0 0 1 0 0 1 1 0 1 1 0 1
[51] 1 1 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0
Levels: 0 1
```

From the above class label analysis, 29 persons have been survived and after careful analysis of these person's data, we can come to the below conclusions.

PClass	**Lower class** passengers were the majority among the survived.
Sex	Most of the people who survived are **females** and the ratio of male and females who survived is 1:5
Age	**Adults** between 18 and 40 years of age were the major survivors.
SibSp	Majority of the survivors have no siblings or spouses aboard.
Parch	Majority of the survivors have no parents or children with them aboard.
Embarked	Majority of the survivors were from the Southampton POE

From the above results, we can conclude that most of the people who survived are ***Lower Class Female Adults*.**

Random Forest

Random Forest Classifier is implemented on the Titanic dataset in *R* to figure out what category of people survived the titanic ship wreck. The classifier takes in various parameters like ntree, proximity and importance, etc., & experiments were made to select appropriate set of parameters. The following are the parameters that are varied in the experiments.

Ntree – This is the number of trees to grow in the process of classification

Proximity – This argument takes either TRUE or FALSE and when set to TRUE, proximity among the instances is considered during classification.

Importance – This argument also takes either TRUE or FALSE and when set to TRUE, the importance of the predictors is also taken into account while building the model and this in turn affect the prediction on the test data.

We performed 10-fold cross validation to obtain reliable results and also due to the fact that averaging cancels out the effect of noise on the results.

The results of various experiments are tabulated as follows:

Classifier	Fold	ntree	Importance	Proximity	Accuracy	Precision	Recall	F-measure
RF	10	300	TRUE	TRUE	81.067	0.818	0.780	0.799
RF	10	400	TRUE	TRUE	80.952	0.818	0.797	0.797
RF	10	450	TRUE	TRUE	80.847	0.816	0.796	0.796
RF	10	500	TRUE	TRUE	80.965	0.816	0.798	0.798
RF	10	550	TRUE	TRUE	81.181	0.820	0.799	0.799
RF	10	600	TRUE	TRUE	81.069	0.818	0.798	0.798
RF	10	700	TRUE	TRUE	81.088	0.818	0.799	0.799

Classifier	Fold	ntree	Importance	Proximity	Accuracy	Precision	Recall	F-measure
RF	**10**	**300**	**FALSE**	**TRUE**	**81.520**	**0.822**	**0.783**	**0.802**
RF	10	400	FALSE	TRUE	81.086	0.816	0.781	0.798
RF	10	450	FALSE	TRUE	80.859	0.812	0.779	0.795
RF	10	500	FALSE	TRUE	80.864	0.812	0.780	0.796
RF	10	550	FALSE	TRUE	80.648	0.812	0.775	0.793
RF	10	600	FALSE	TRUE	81.196	0.820	0.780	0.799
RF	10	700	FALSE	TRUE	80.979	0.814	0.780	0.797

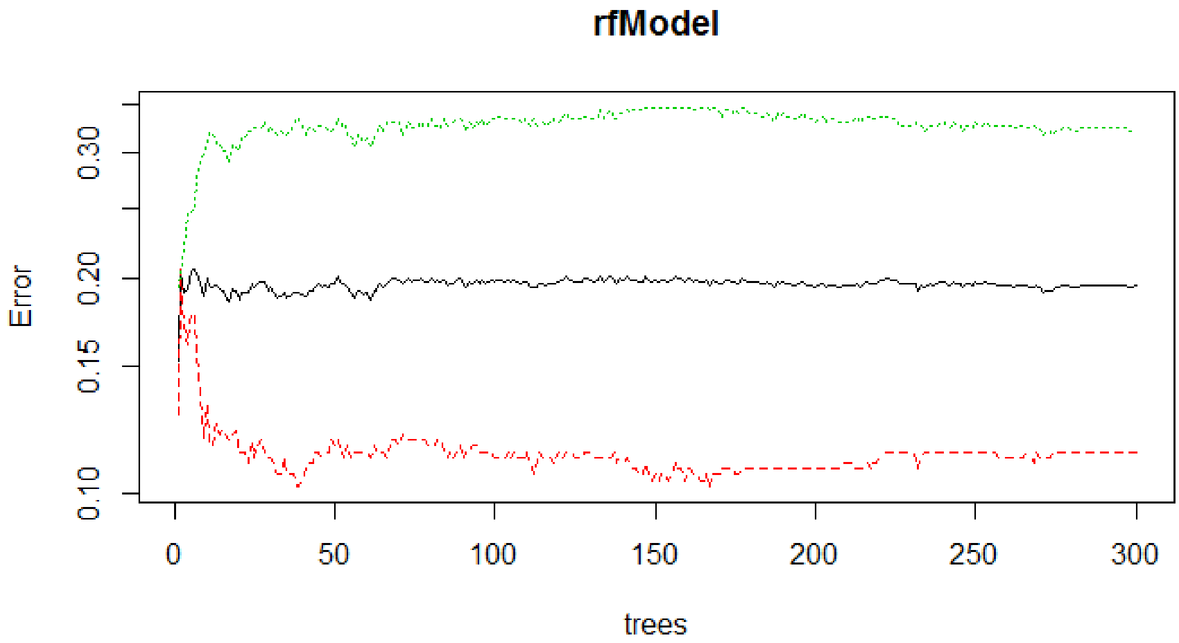
Classifier	Fold	ntree	Importance	Proximity	Accuracy	Precision	Recall	F-measure
RF	10	300	TRUE	FALSE	81.067	0.818	0.780	0.799
RF	10	400	TRUE	FALSE	80.952	0.818	0.776	0.797
RF	10	450	TRUE	FALSE	80.847	0.816	0.777	0.796
RF	10	500	TRUE	FALSE	80.965	0.816	0.780	0.798
RF	10	550	TRUE	FALSE	81.180	0.820	0.780	0.799
RF	10	600	TRUE	FALSE	81.069	0.818	0.779	0.798
RF	10	700	TRUE	FALSE	81.008	0.818	0.780	0.799
Classifier	**Fold**	**ntree**	**Importance**	**Proximity**	**Accuracy**	**Precision**	**Recall**	**F-measure**
RF	10	300	FALSE	FALSE	81.520	0.822	0.783	0.802
RF	10	400	FALSE	FALSE	81.086	0.816	0.781	0.798
RF	10	450	FALSE	FALSE	80.859	0.812	0.779	0.795
RF	10	500	FALSE	FALSE	80.864	0.812	0.780	0.795
RF	10	550	FALSE	FALSE	80.648	0.812	0.775	0.793
RF	10	600	FALSE	FALSE	81.196	0.820	0.780	0.799
RF	10	700	FALSE	FALSE	80.979	0.814	0.780	0.767

From the obtained results, we chose the number of trees to be 300, the importance is set to FALSE and the proximity among the instances is considered. Apart from accuracy, other metrics such as precision, recall and f- measure were also considered to maintain fairness in evaluation.

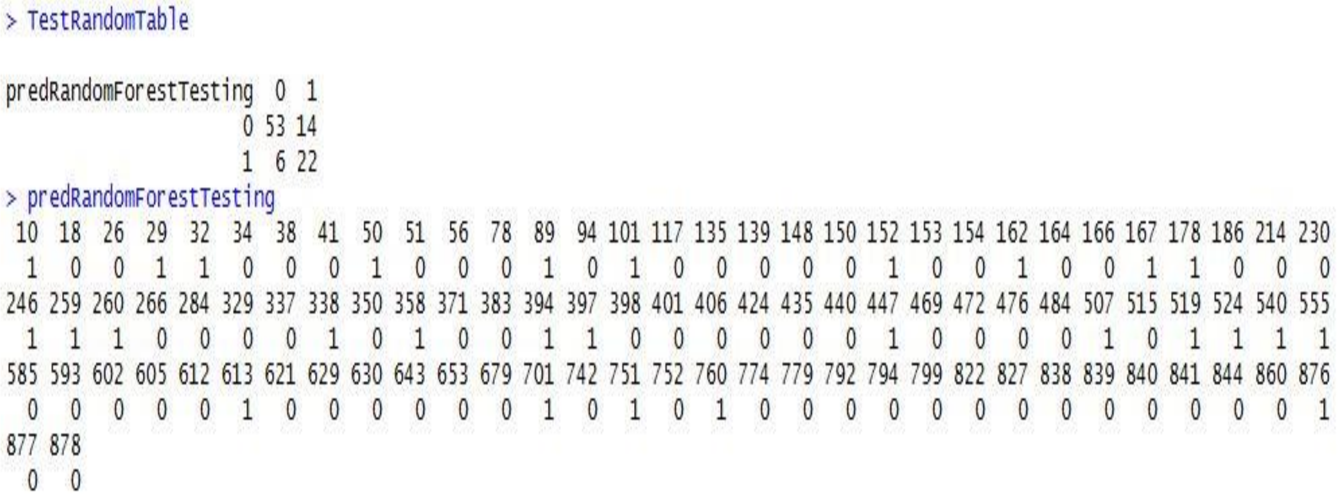
The following table summarizes the **performance of the Random Forest Classifier** on the Titanic Dataset:

Classifier	**Random Forest**
Number of Folds in Cross Validation	10
Accuracy	81.520 %
Precision	0.822
Recall	0.783
F - measure	0.802

The plot of test MSE of the classification is as below:



We now examine the class labels of the prediction to analyze the result of the classifier. The following R – snapshot gives us the picture of the class labels of prediction on the test set.

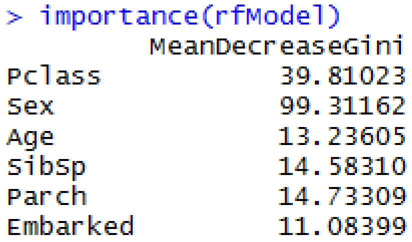


```
> TestRandomTable

predRandomForestTesting  0  1
                      0 53 14
                      1  6 22
> predRandomForestTesting
 10  18  26  29  32  34  38  41  50  51  56  78  89  94 101 117 135 139 148 150 152 153 154 162 164 166 167 178 186 214 230 
  1   0   0   1   1   0   0   0   1   0   0   0   1   0   1   0   0   0   0   0   1   0   0   1   0   0   1   1   0   0   0 
246 259 260 266 284 329 337 338 350 358 371 383 394 397 398 401 406 424 435 440 447 469 472 476 484 507 515 519 524 540 555 
  1   1   1   0   0   0   0   1   0   1   0   0   1   1   0   0   0   0   0   0   1   0   0   0   0   1   0   1   1   1   1 
585 593 602 605 612 613 621 629 630 643 653 679 701 742 751 752 760 774 779 792 794 799 822 827 838 839 840 841 844 860 876 
  0   0   0   0   0   1   0   0   0   0   0   0   1   0   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   1 
877 878 
  0   0 
```

We note from the above table that 22 persons of the test sample were among those survived the wreck and are predicted correctly by the classifier.

The importance of the attributes in predicting the class in RF model is as follows:



```
> importance(rfModel)
         MeanDecreaseGini
Pclass           39.81023
Sex              99.31162
Age              13.23605
SibSp            14.58310
Parch            14.73309
Embarked         11.08399
```

Upon careful observation and analysis of the survived instances, we come across the following facts. The results of the analysis are presented in a table indicating what category of persons survived w.r.t each attribute.

PClass	**Upper class** passengers were the majority among the survived.
Sex	**Female** passengers were mostly survived.
Age	**Adults** between 18 and 40 years of age were the major survivors.
Sibsp	*Majority of the survivors have no siblings or spouses aboard.
Parch	*Majority of the survivors have no parents or children with them aboard.
Embarked	Majority of the survivors were from the Southampton POE

*People might not necessarily travel with their families (say siblings, spouses, parents, children, etc.,) but some children might travel just with a nanny or people can travel with their friends, etc.,

From the results available based on the Random Forest Classifier, we concluded that the majority of the survivors of the Titanic ship wreck were ***Upper Class Female Adults***.

Boosting

Boosting Classifier is implemented on the Titanic dataset in R to figure out what category of people survived the titanic ship wreck. The parameters considered in this classifier are minsplit, mfinal and maxdepth, etc... The task is to select most appropriate set of parameters influencing the classifier. The following are the parameters that are varied in the experiments.

minsplit – Minimum number of splits that must be performed.

Mfinal - an integer, the number of iterations for which boosting is run or the number of trees to use. Defaults to mfinal=100 iterations

max depth – maximum depth of the individual regression estimators. The maximum depth limits the number of nodes in the tree. Tune this parameter for best performance; the best value depends on the interaction of the input variables.

We performed 10-fold cross validation to obtain reliable results and also due to the fact that averaging cancels out the effect of noise on the results.

The results of various experiments are as follows:

Classifier	Fold	minsplit	Mfinal	Maxdepth	Accuracy	Precision	Recall	F-measure
BOOSTING	10	2	10	3	80.552	0.802	0.773	0.787
BOOSTING	10	3	10	3	80.692	0.806	0.777	0.791
BOOSTING	10	4	10	3	81.033	0.812	0.784	0.797
BOOSTING	10	5	10	3	80.600	0.804	0.777	0.790
BOOSTING	10	6	10	3	80.181	0.803	0.771	0.786
BOOSTING	10	7	10	3	80.378	0.803	0.780	0.791
BOOSTING	10	8	10	3	79.845	0.800	0.770	0.784

Classifier	Fold	minsplit	Mfinal	Maxdepth	Accuracy	Precision	Recall	F-measure
BOOSTING	10	2	11	3	81.200	0.817	0.776	0.795
BOOSTING	10	3	11	3	79.691	0.794	0.766	0.779
BOOSTING	10	4	11	3	80.024	0.800	0.771	0.785
BOOSTING	10	5	11	3	80.751	0.812	0.779	0.795
BOOSTING	10	6	11	3	80.951	0.813	0.780	0.796
BOOSTING	10	7	11	3	80.275	0.800	0.778	0.789
BOOSTING	10	8	11	3	81.167	0.816	0.783	0.799

Classifier	Fold	minsplit	Mfinal	Maxdepth	Accuracy	Precision	Recall	F-measure
BOOSTING	10	2	11	4	81.097	0.811	0.780	0.795
BOOSTING	10	3	11	4	79.134	0.787	0.765	0.776
BOOSTING	10	4	11	4	80.227	0.803	0.774	0.788
BOOSTING	10	5	11	4	81.169	0.808	0.788	0.798
BOOSTING	**10**	6	11	4	82.103	0.824	0.794	0.808
BOOSTING	10	7	11	4	80.340	0.808	0.774	0.791
BOOSTING	10	8	11	4	81.362	0.814	0.788	0.801

Classifier	Fold	minsplit	Mfinal	Maxdepth	Accuracy	Precision	Recall	F-measure
BOOSTING	10	2	12	4	81.711	0.824	0.786	0.804
BOOSTING	10	3	12	4	81.027	0.810	0.780	0.795
BOOSTING	10	4	12	4	82.336	0.824	0.799	0.811
BOOSTING	10	2	13	5	80.8135	0.809	0.778	0.793
BOOSTING	10	3	13	5	79.689	0.792	0.774	0.783
BOOSTING	10	4	13	5	80.575	0.809	0.779	0.794
BOOSTING	10	5	13	5	80.665	0.807	0.774	0.790

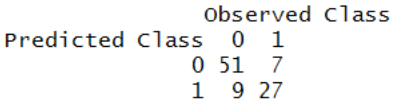
From the obtained results, we chose minsplit to be 4, the mfinal is set to 12 and the maxdepth as 4. Apart from accuracy, other metrics such as precision, recall and f- measure were also considered to maintain fairness in evaluation.

The following table summarizes the **performance of the Boosting Classifier** on the Titanic Dataset:

Classifier	**Boosting**
Number of Folds in Cross Validation	10
Accuracy	82.336
Precision	0.824
Recall	0.799
F – measure	0.811

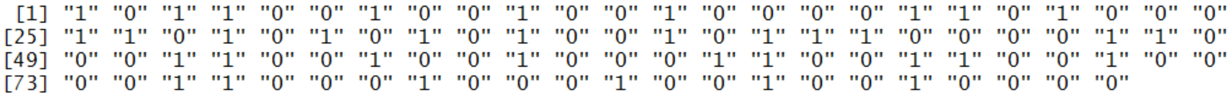
We now examine the class labels of the prediction to analyze the result of the classifier. The following R – snapshot gives us the picture of the class labels of prediction on the test set.

- ➔ Predicted



```
               Observed Class
Predicted Class  0  1
              0 51  7
              1  9 27
```

Predicted Class Labels



```
 [1] "1" "0" "1" "1" "0" "0" "1" "0" "0" "1" "0" "0" "1" "0" "0" "0" "0" "1" "1" "0" "1" "0" "0" "0"
[25] "1" "1" "0" "1" "0" "1" "0" "1" "0" "1" "0" "0" "1" "0" "1" "1" "1" "0" "0" "0" "0" "1" "1" "0"
[49] "0" "0" "1" "1" "0" "0" "1" "0" "0" "1" "0" "0" "0" "1" "1" "0" "0" "1" "1" "0" "0" "1" "0" "0"
[73] "0" "0" "1" "1" "0" "0" "0" "1" "0" "0" "0" "1" "0" "0" "1" "0" "0" "1" "0" "0" "0" "0"
```

We note from the above table that 20 persons of the test sample were among those survived the wreck and are predicted correctly by the classifier.

Upon careful observation and analysis of the survived instances, we come across the following facts. The results of the analysis are presented in a table indicating what category of persons survived w.r.t each attribute.

PClass	**Middle class** passengers were the majority among the survived.
Sex	**Female** passengers were mostly survived.
Age	**Adults** between 18 and 40 years of age were the major survivors.
Sibsp	Majority of the survivors have no siblings or spouses aboard.
Parch	Majority of the survivors have no parents or children with them aboard.
Embarked	Majority of the survivors were from the Southampton POE

From the results available based on the Boosting Classifier, we concluded that the majority of the survivors of the Titanic ship wreck were ***Middle Class Female Adults***.

Conclusion

Below are the best accuracies obtained for different classifiers after logging values with different parameters.

Classifier	**Parameters Considered**	**Accuracy (in %)**	**Precision**	**Recall**	**F-Score**
Decision Tree	minsplit=20, maxdepth=5, minbucket=5, cp=0.001	84.357	0.803	0.866	0.826
SVM	Cost=10, gamma=0.25, kernel=polynomial	81.360	0.788	0.813	0.800
K-NN	K=15	81.718	0.816	0.786	0.800
Random Forest	Ntree=300, importance=false, proximity=false	81.520	0.822	0.783	0.802
Boosting	Mfinal=12, minsplit=4, maxdepth=4	82.336	0.824	0.799	0.811

From the above observations, we can see that Decision Tree classifier has the maximum accuracy but decision tree is not consistent and small change in the data lead to large variation in accuracy. Hence instead of decision tree, we can take both ***Random Forest and Boosting*** classifiers which gained highest accuracy with consistency.

The statistics from all the classifiers show that the ***females*** and the ***middle-aged*** people had more chances of survival, compared to the other categories. Even, the training data set shows similar significant results. The training data has 303 entries who survived. 214 ***females*** survived and 188 ***middle-aged*** people survived in the training set.