Exploratory Model: Random Forests.

The Random Forest algorithm, was the first algorithm to implement to the project.

“Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual tree”(Wikipedia).

For various input cases, 6 fold cross validation was implemented and the results for each case were gathered. At this point, it would be useful for the reader, to show the exact labels for each activity:

|  |  |  |
| --- | --- | --- |
| Label Number | Activity | Sum of Activities in Dataset |
| 1 | Sleeping | 14 |
| 2 | Toileting | 87 |
| 3 | Showering | 14 |
| 4 | Breakfast | 81 |
| 5 | Grooming | 49 |
| 6 | Spare\_Time/TV | 78 |
| 7 | Leaving | 14 |
| 8 | Lunch | 61 |
| 9 | Snack | 11 |

Case 1. Input: Location, Type, Place:

model1 OOB estimate of error rate: 48.17%

Accuracy:51.83%

Confusion matrix:

1 2 3 4 5 6 7 8 9 class.error

1 2 5 0 0 2 3 0 2 0 0.8571429

2 6 38 0 0 20 14 1 8 0 0.5632184

3 0 0 0 5 8 1 0 0 0 1.0000000

4 1 1 0 35 9 1 0 34 0 0.5679012

5 0 1 0 0 37 9 1 1 0 0.2448980

6 1 4 0 0 6 51 12 4 0 0.3461538

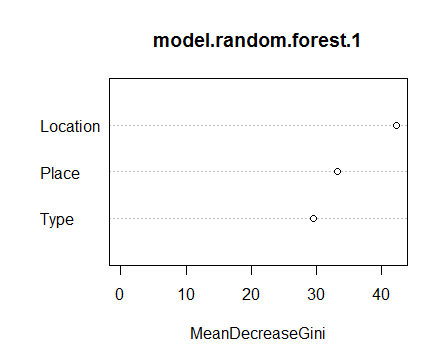
7 0 0 0 0 2 12 0 0 0 1.0000000

8 0 0 0 8 1 2 1 49 0 0.1967213

9 0 1 0 0 1 4 0 5 0 1.0000000

The first implementation of the random forest algorithm used as inputs only the attributes “Location” ,”Place” and “Type”. This approach returned an accuracy of 51.83%.Top scoring activities were “Grooming”(37/49) and “Lunch” (49/61). On the other hand, the worse results were for “Showering”, ”Leaving”, and “Snack” (not a single activity was predicted).

The picture below shows the variable Importance plot for this model. It seems that attribute “Location” is by far the most important” decision maker” for the model. That’s something really valuable for the exploratory attribute procedure, which will eventually lead to the best possible results.



Case 2 Input: Location, Type, Place,hours

model2 OOB estimate of error rate: 37.16%

Accuracy:62.84%

Confusion matrix:

1 2 3 4 5 6 7 8 9 class.error

1 5 3 0 0 2 3 0 1 0 0.6428571

2 7 40 0 0 20 14 1 3 2 0.5402299

3 0 0 0 5 8 1 0 0 0 1.0000000

4 1 1 2 67 9 1 0 0 0 0.1728395

5 0 7 1 3 27 9 1 0 1 0.4489796

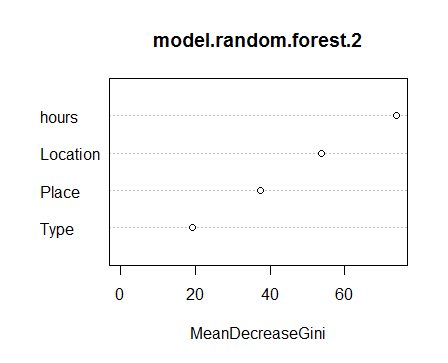
6 1 4 0 1 6 54 9 1 2 0.3076923

7 0 0 0 0 2 6 6 0 0 0.5714286

8 0 0 0 3 1 2 1 54 0 0.1147541

9 0 1 0 0 1 3 1 1 4 0.6363636

In this case, attribute “hours” joins the input team. “Hours” is a time attribute, including only the “hh” instance from the start time (00,01,02,03,..21, 22, 23). This entrance, made the results far better. Accuracy was improved and only one activity (Showering) presented 100% error. Best scores here were “Breakfast” (67/81) and “Lunch”(54/61). In the plot below, it is easily noticeable that attribute “hours” was the most important for this classificator.



Case 3 Input: Location, Type, Place, hours, end.hours

model3 OOB estimate of error rate: 36.43%

Accuracy: 63,57%

Confusion matrix:

1 2 3 4 5 6 7 8 9 class.error

1 5 4 0 0 2 3 0 0 0 0.6428571

2 7 42 0 0 20 14 1 1 2 0.5172414

3 0 0 0 5 8 1 0 0 0 1.0000000

4 1 1 3 66 9 1 0 0 0 0.1851852

5 1 7 1 5 24 9 1 0 1 0.5102041

6 0 8 0 1 4 60 3 0 2 0.2307692

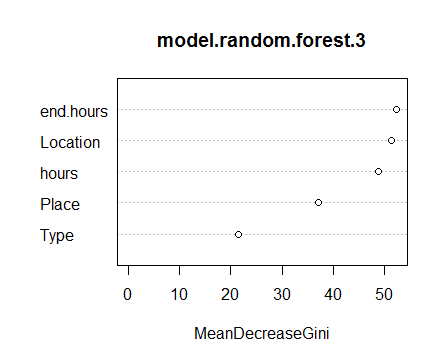
7 0 0 0 0 2 6 6 0 0 0.5714286

8 0 0 0 3 1 2 1 54 0 0.1147541

9 0 2 0 0 1 3 1 1 3 0.7272727

In this trial, attribute “end. hours” is joined. This attribute is the same as “hours” but includes the “hh” instance from the “end.time” attribute.

Accuracy raised about 0.27%. Best results showed up for “Shower” (54/61) and “Breakfast(66/81). Significantly better were the results for “Spare\_Time/TV”(60/78) while result got worse for “Snack”. The worst results were (once again) for “Showering” with a class error of 100%. In the graph below, it is obvious that the latest attribute (end.hours), was highly predictive for this model.



Input Case 4 Input: Location, Type, Place, Start Time, End Time,Duration