

Indian Liver Patients Capstone Project

Adele Taylor

08/01/2021

Introduction

The Indian Liver Patients data set (Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.) contains 416 liver patient records and 167 non liver patient records collected from North East of Andhra Pradesh, India. Dataset contains age of the patient, gender of the patient, results of blood tests including eight measures of interest and a field used to split the data into two sets (patient with liver disease, or no disease). The aim of this project is simply to create an ensemble model which predicts whether a patient has or does not have liver disease based on the blood test results.

The dataset was divided into training and validation sets, a selection of models trained and tested initially on a random subset of the training set, before best performing models combined into an ensemble and checked against the validation set.

All the code for the analyses mentioned in this report is included in the `ataylor_india_liver_patients.R` file.

Analysis

This is a small dataset (579 observations of 11 variables) and we will try a straightforward approach of six simple algorithms from the `caret` package. First we will check their feasibility, again partitioning the training set into training and test sets in ratio 9:1.

Models chosen were a general linear model (“glm”), k-nearest neighbours (“knn”) random forest (“rf”), random partition (“rpart”), linear discrimination analysis (“lda”) and quadratic discrimination analysis (“qda”). The resulting root mean square errors (RMSEs) are shown in the table below:

```
## Warning in set.seed(4, sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```


[illegible]

```

## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

```



```

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

```



```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

	x
glm	0.4748
knn	0.4285
rf	0.4375
rpart	0.4599
lda	0.4385
qda	0.4262

These are all pretty good low RMSEs, so we shall incorporate all models into the final ensemble, as shown in the R code below.

```
final_model <- function(validation_set, training_set){
  predictions <- sapply(models, function(model){
    fit <- train(Disease ~ ., model=model, data=training_set)
    return((predict(fit, validation_set)))
  })
}
```

Results

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
```

```
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

[illegible]

[illegible]

[illegible]

```

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

```

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to do
## classification? If so, use a 2 level factor as your outcome column.
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

```
## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```


[illegible]


```
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?

## Warning in randomForest.default(x, y, mtry = param$mtry, ...): The response has
## five or fewer unique values. Are you sure you want to do regression?
```

Using this final model on the validation set gave a final RMSE of 0.3745.

Conclusion

Due to time constraints, this was a pretty bare bones approach. A more nuanced approach would look at specific sets of the inputs to see if still reasonable results can be obtained without using the full blood panel (so if only some results are known). It would also be worth testing whether there are differences between the effectiveness of the algorithm based on the gender of the patient as there are 441 men and only 142 women included in the dataset.