MATH5743M: Statistical Learning - Lecture 8

Dr Seppo Virtanen, s.virtanen@leeds.ac.uk

Semester 2: 14 Mar 2022

Outline

Ensemble learning or a combination of models

- averaging reduces noise
- different models may focus on different aspects
- combine strengths of the different models.

Wisdom of the crowd or collective wisdom

Figure 1: Francis Galton noted how collective intelligence (average guess 1,197lb) gave an accurate estimate for the weight of a bull (1,198lb).

Wisdom of the crowd or collective wisdom

Crowds acting together can be capable of intelligence that goes beyond what any individual can achieve on their own.

Condorcet's theorem

Trial by jury, where twelve citizens sit in judgement of the accused and decide on his/her guilt or innocence. But why do we have twelve jury members, rather than just a single individual?

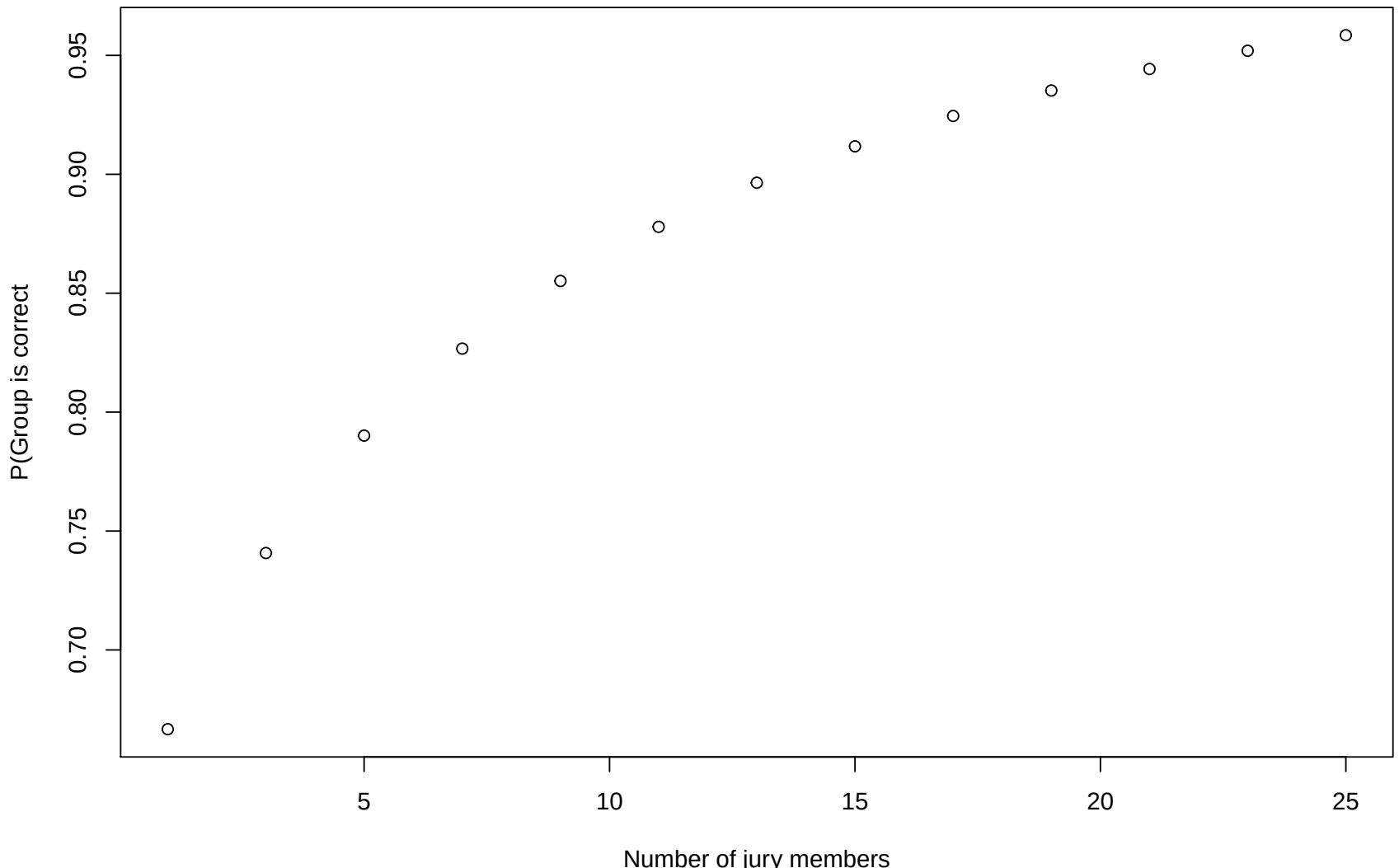
Consider N individuals asked to decide on the truth of some proposition. Assume that each individual has a probability of being correct p, and that $p > 1/2$ (even a single individual is more likely to be correct than wrong).

If these N individuals vote, and the vote of individual i is X_i, with $X_i = 1$ implying the correct decision and $X_i = 0$ the wrong decision, then the probability that the group will make the correct decision is:

$$\begin{aligned} P(\text{Group is correct}) &= P(\sum_{i=1}^{N} X_i > N/2) \\ &= P(Z > N/2),\ Z = \sum_{i=1}^{N} X_i \sim \mathcal{B}(N, p) \end{aligned} \tag{1}$$

Condorcet's theorem

```
N = seq(1, 25, 2); p = 2/3
p_group_correct = 1 - pbinom(N/2, N, p)
plot(N, p_group_correct, xlab="Number of jury members", yla
```



Condorcet's theorem

With a jury of 15 the group will get its decision correct over 90% of the time, even though each individual is only right 66% of the time.

If we imagine that instead of different jurors, we had a collection of different *classifiers*, this shows how an *ensemble or combination* of many bad/basic classifiers can produce very accurate results.

The Netflix prize

Many teams of researchers entered the competition and produced very complicated algorithms that improved on Netflix's own, but for a long time no one was able to beat the magic figure of 10% (necessary criterion for the prize).

In 2009, ensemble of previous methods was able to reach this improvement.

Ensemble of statistical models: Each statistical model was able to spot patterns in the data that other models had missed; as a result the combined prediction from all of the models put together was able to perform much better than any single model.

The elements of collective intelligence

1. Diversity. Models are different.
2. Independence. Models are not dependent on each other.
3. Decentralisation. Models do not 'see' each other.
4. Aggregation. (Weighted) average as a final prediction.

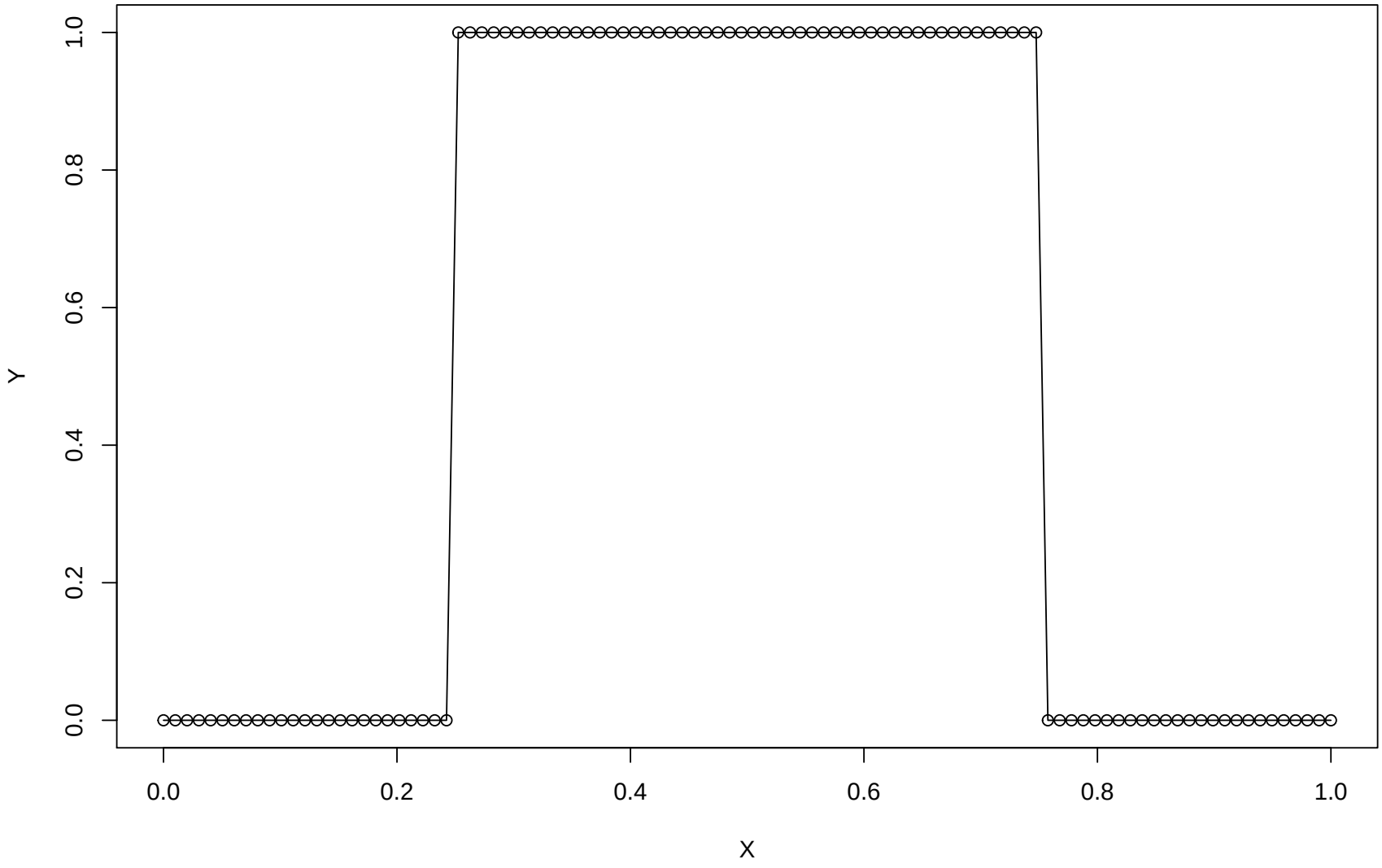
Bootstrap Aggregating (BAGGING)

Bootstrapping: We can generate a new data set by sampling from indices uniformly *with* replacement.

```
data = faithful
bootstrap_indices = sample(length(faithful),
                           length(faithful), replace=TRUE)
data_bootstrap = data[bootstrap_indices, ]
```

Intuition: Summing over many logistic regression models trained on different sampled subsets of the data can outperform a single model.

BAGGING



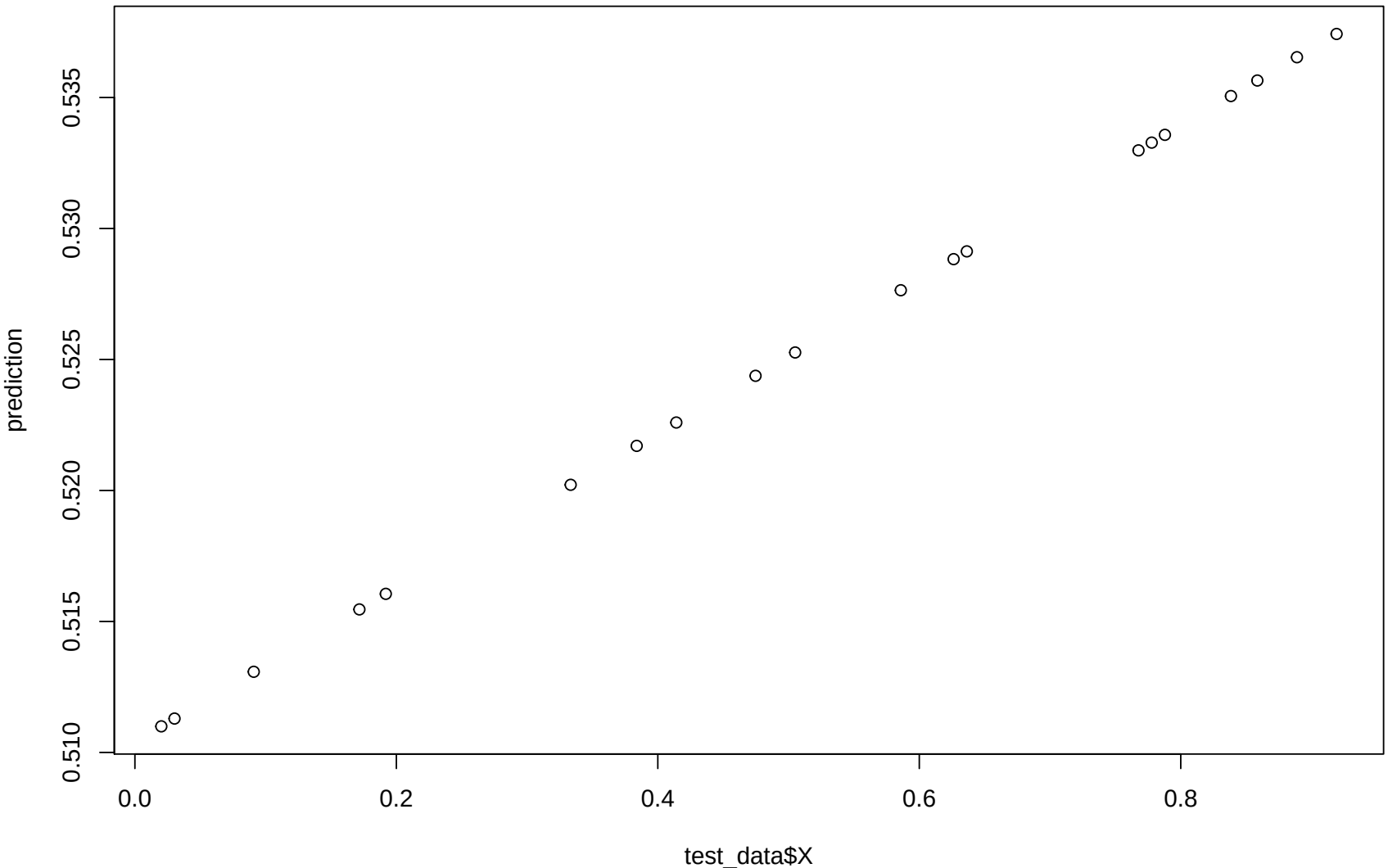
BAGGING

```
#Make 1000 bootstrap samples
#Record prediction onto test data and aggregate
bsPrediction = rep(0, dim(test_data)[1])
for (i in 1:1000){
  bs_idx = sample(dim(train_data)[1], 5, replace=TRUE)
  bs_data = train_data[bs_idx,]
  bsModel = suppressWarnings(
    glm(Y ~ X, data=bs_data, family=binomial))
  bsPrediction = bsPrediction +
    predict(bsModel, newdata=test_data,
            type="response")/1000
}
```

BAGGING

Simple logistic regression performs poorly.

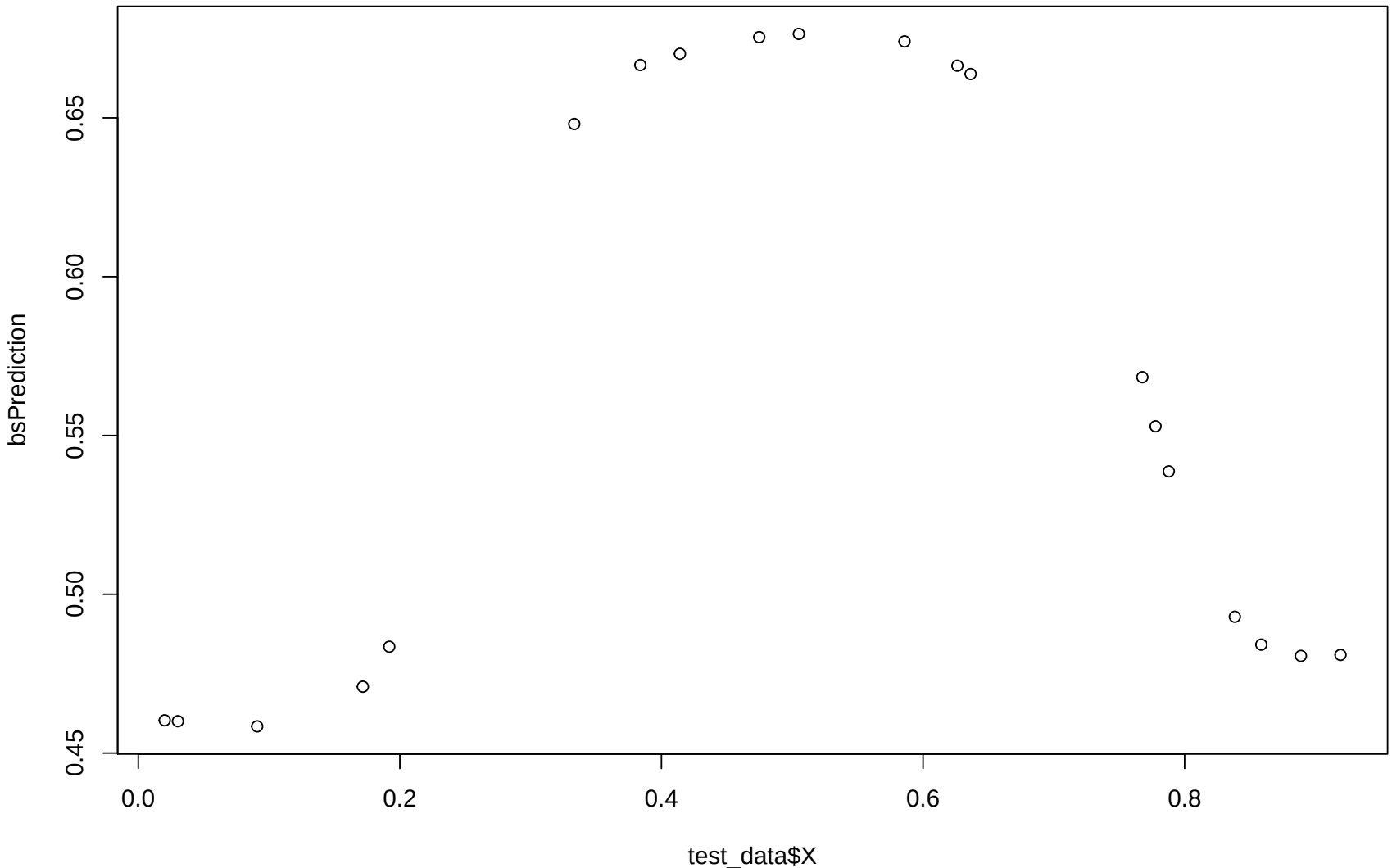
```
## [1] -14.11583
```



BAGGING

Ensemble works better.

```
## [1] -11.44711
```



BAGGING

In this case a single logistic regression cannot capture the fact that the probability is non-monotonic in X.

By learning from different sampled subsets of the data, the BAGGED logistic regressions capture different features of the underlying relationship, in this case the rise in probability at $X = 0.25$ and the decrease at $X = 0.75$.

Discussion

If we want a collective or ensemble to perform well, we should try to increase the diversity and independence of the different models within it, and the information they can draw on.

One way we can do that is by decreasing the overlap in the data each model sees the training set.