MLR Models for Data Subsets

Matt Tyers

2024-08-30

## Overview of methods

### Variables

Multiple Linear Regression (MLR) was used to explore the effects of multiple design variables on stability, as quantified by stability scores calculated by inconsistency between four different measurements and their respective design values, according to the form below, for measurement of culvert . Scores were standardized by their respective standard deviation, to set all four sets of scores on an equivalent scale.

It should be noted that a previous analysis used the *absolute value of* the log ratio; this is no longer used. Strongly negative V scores may be interpreted as change in the negative direction with respect to measurements, positive V scores may be interpreted as change in the positive direction, and scores near zero may be interpreted as stable.

The following design variables were considered. These were variables identified as of High or Medium importance, and with no missing data. Note: this can be expanded to include variables with some missing data if desired, though this may have an unknown effect on model selection criteria.

## [1] "culvert\_shape" "lake\_outlet"   
## [3] "banks\_y\_n" "features\_y\_n"   
## [5] "reach\_3\_structure\_width" "reach3\_design\_cr"   
## [7] "reach\_3\_total\_length" "reach\_3\_gradient"   
## [9] "bank\_design\_type" "reach\_3\_width"   
## [11] "reach3\_design\_cr\_cat"

### Model Scoring

With many potential explanatory variables and a relatively limited sample size, overfitting (selecting too many variables) was a concern, since each culvert by itself had the potential to affect the overall model fit. To address this concern, the fit of a prospective model was evaluated according to its predictive power. This was calculated by means of leave-one-out cross validation (LOOCV), and scored using prediction root mean square error (RMSE), calculated according to the form:

in which represents the predicted value for instability score for culvert , predicted from a prospective model *without* using the data from culvert .

### Model-building Algorithm

Initially, I constructed an algorithm to fit models for every combination of variables, and do LOOCV scoring on all of them. This was time-consuming to run, probably prohibitively so if we want to consider more variables! I ended up writing an algorithm in which each variable was considered in turn, and the variable with the best predictive accuracy (lowest LOOCV RMSE) selected. Then, the remaining variables were considered and iteratively added, until there was no improvement to LOOCV RMSE. This agglomerative algorithm ended up finding the globally best predictive model in a fraction of the time.

### An Important Note on Interpretation

This dataset contains many explanatory variables that contain very similar information, or at least are strongly related to one another. For example, for instability related to Interior Channel Width, four design variables are much more important in themselves: Banks (y/n), Bank Design Type (1/2/3), Design CR (numeric), and Design CR (categorical). Banks (y/n) contains basically the same information as Bank Design Type (1/2/3), and when one of the two is selected, the other does not add appreciably to the model. Interestingly, there is a strong relationship between Banks and Design CR (you probably have better insights as to why!), so when a Banks variable is already in the model, the effect of a CR variable is washed out. It also begs the question: is the observed effect due to Banks, or due to CR? At my end, it’s impossible to say.

All this to say: the variables included in the final “best” model do not necessarily represent the most important variables by themselves! Most likely, they represent the most mutually-independent collection of variables that add meaningful predictive power.

### Plots to Follow

#### Multiple Regression

Two sets of plots are included for each of four V scores. The first sequence shows the results of the agglomerative MLR model-building algorithm, with the top row showing the LOOCV RMSE associated with all variables at each step (left to right), adding a new variable in each step. The first plot (top left) therefore illustrates which variables are important by themselves.

The plot at the top right shows the overall drop in LOOCV RMSE as variables are added. In two of the four cases, most of the drop in LOOCV RMSE happened with the first variable!

The next row illustrates the predictive power of each iteration of model building, with *LOOCV predicted* on the x-axis, as calculated by models fit *without* each observation, and the reported represents the of *prediction*. Finally, the plot at the bottom right of this set shows the fitted *V* on the x-axis and true *V* on the y-axis, along with the of model fit. The line overlayed with these plots is the y=x equality line, not a regression line.

#### Regression Tree

Below this is a set of plots from a regression tree for each variable. I have to confess, I know relatively little about the machinery behind regression trees. The R documentation for the rpart function just says that “it follows Breiman et. al (1984) quite closely.” Breiman et al is a textbook, and it’s one I haven’t read!

Two differences between MLR and regression trees that immediately come to mind are

* trees treat all numerical variables as categorical, and can perform better if there are nonlinear or threshold effects, but poorly if there are approximately linear effects
* nodes in a tree are conditional on the previous nodes, whereas MLR effects are global.

There is a set of tuning parameters for regression trees that I honestly haven’t messed with, that decide when to attempt a split and when not to. The first plot (top left) in this set shows trial values of minsplit, which is the minimum number of observations in a node to consider a split, vs LOOCV RMSE associated with trees computed at each value. A tree was then constructed with the best value of minsplit.

The next plot is Variable Importance. I don’t know how this is mathematically defined, but seems to be a measure of the overall importance of each variable at the top level. Often, variables with a high importance score are not used in the final tree, which is probably analogous to important variables not appearing in a final MLR model.

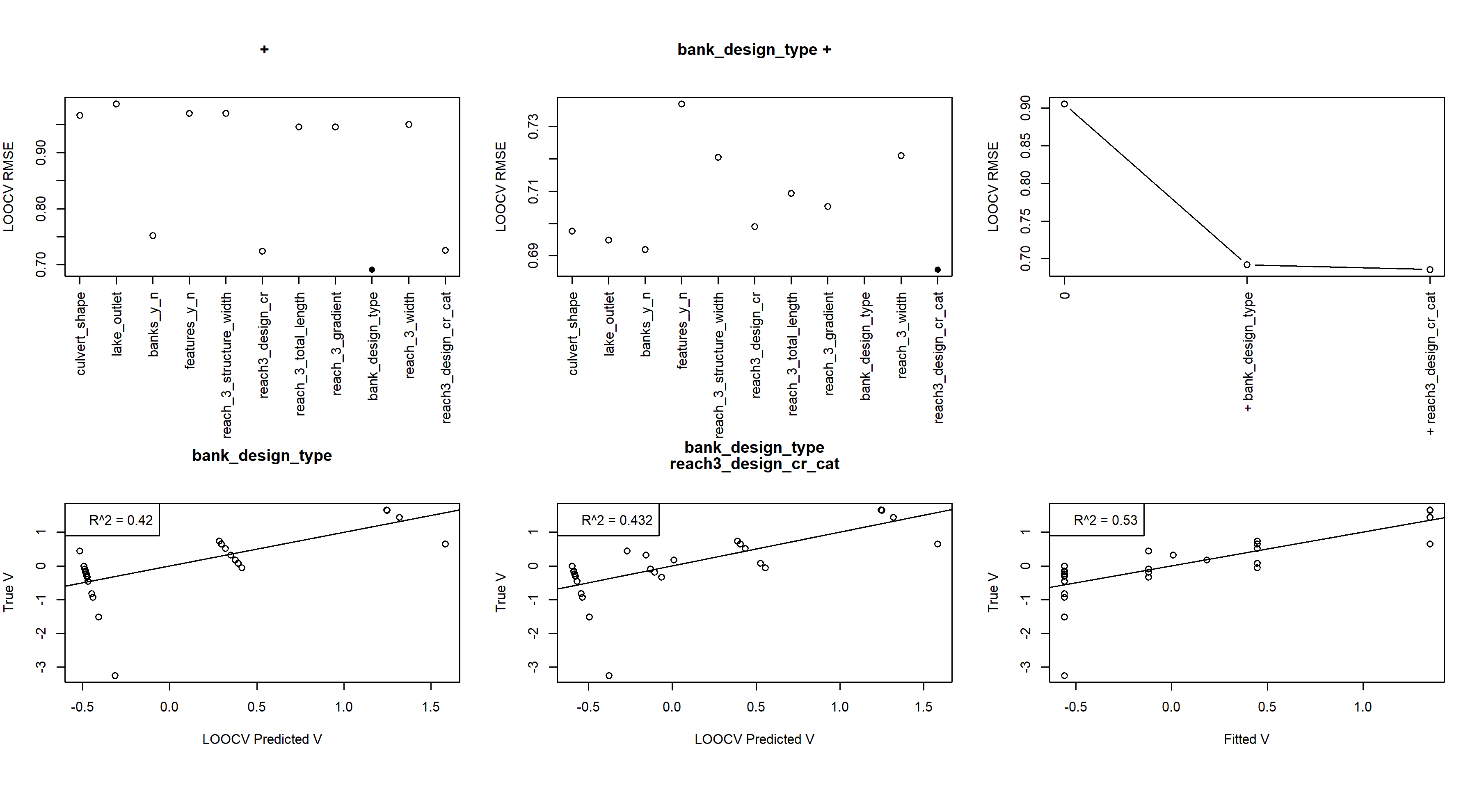
Finally, true *V* vs LOOCV predicted and Fitted *V*, and their respective values of . These can be compared to their MLR counterparts, the final two plots of the MLR set. Again, the overlayed line is the y=x line, not a regression line.

## Interior Channel Width - Multiple Regression

### Reach 3 gradient <= 1.5

Top model:

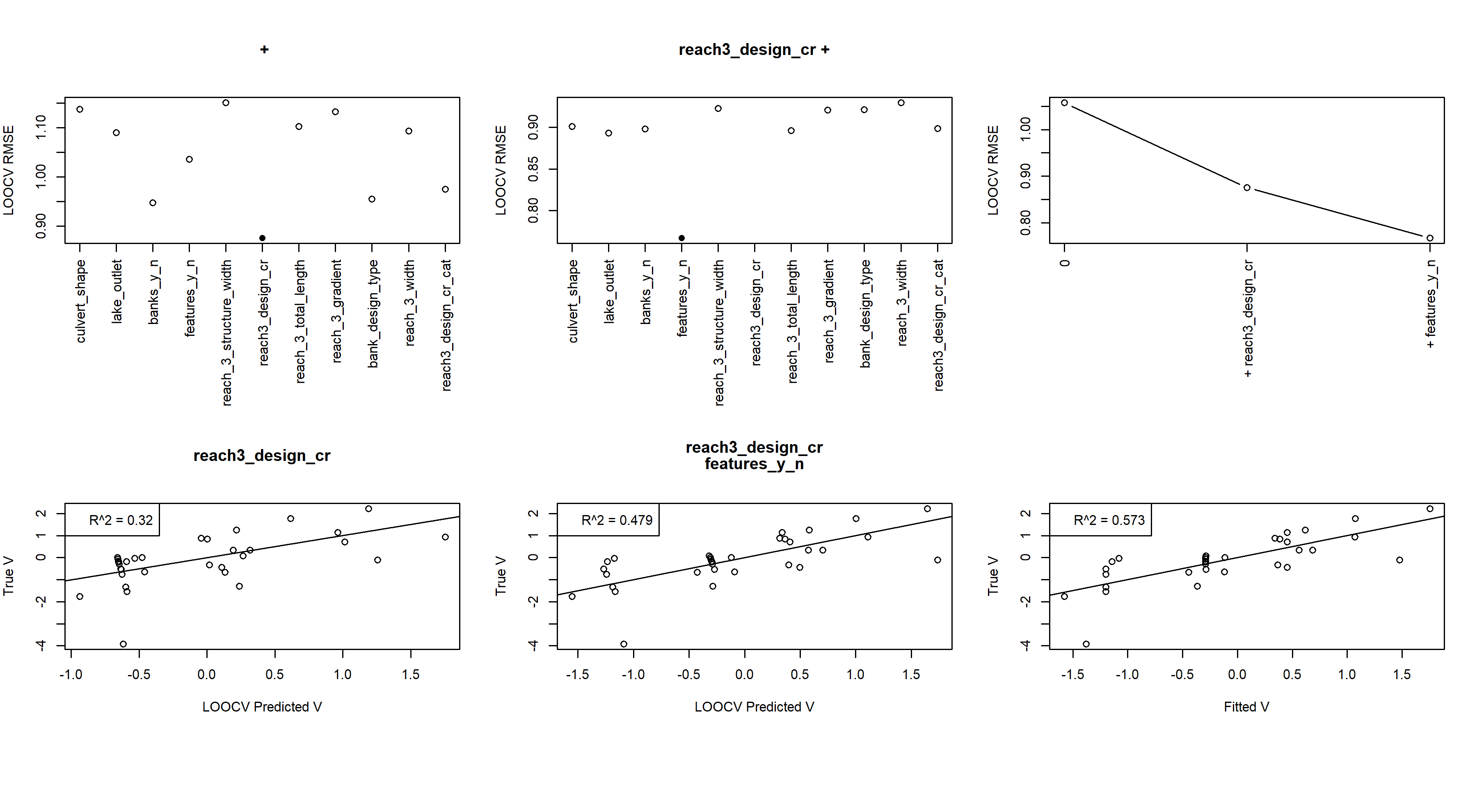
V Interior Channel Width ~ bank\_design\_type + reach3\_design\_cr\_cat



### Reach 3 gradient > 1.5

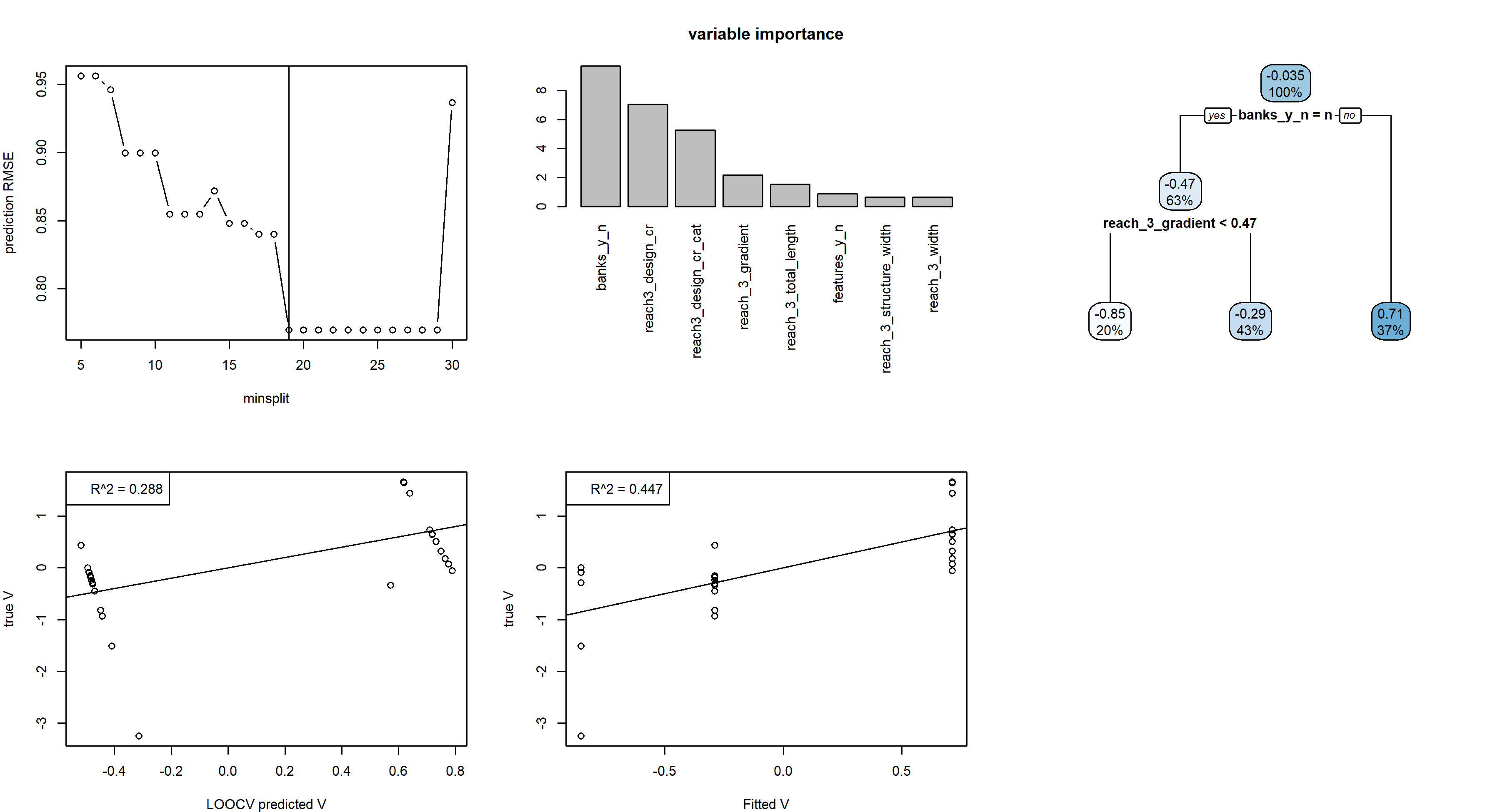
Top model:

V Interior Channel Width ~ reach3\_design\_cr + features\_y\_n

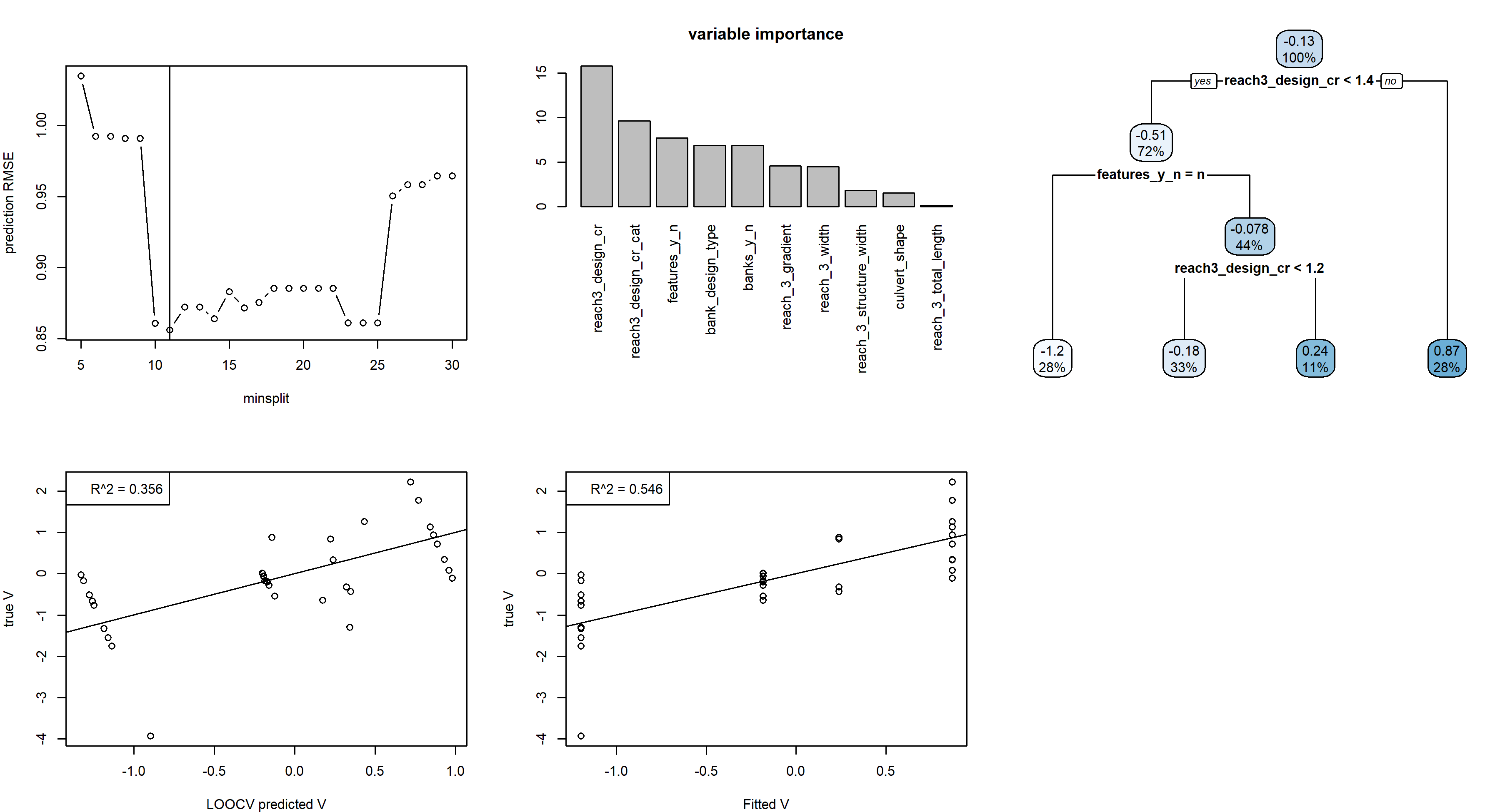


## Interior Channel Width - Regression Tree

### Reach 3 gradient <= 1.5



### Reach 3 gradient > 1.5

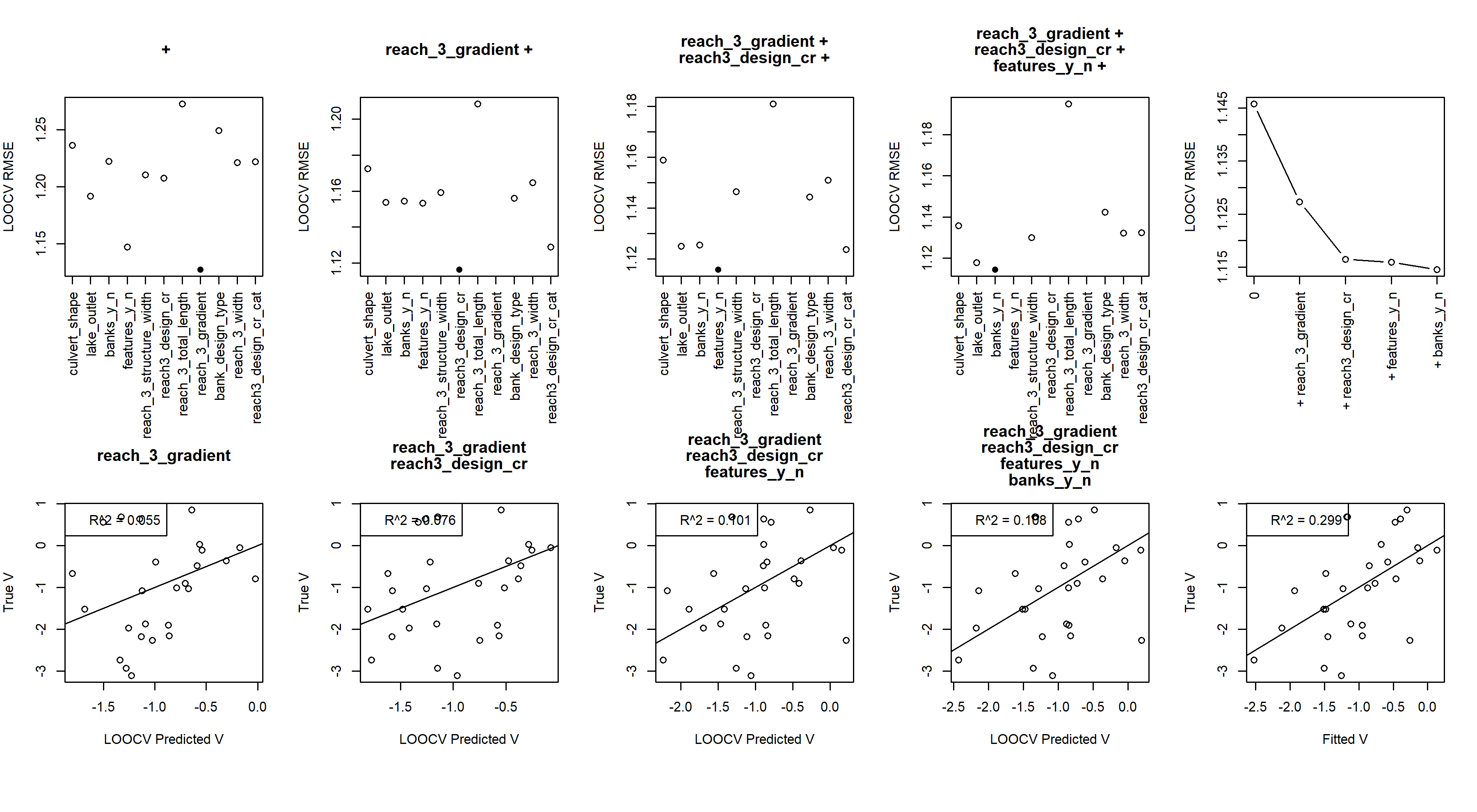


## Interior Gradient - Multiple Regression

### Reach 3 gradient <= 1.5

Top model:

V Interior Gradient ~ reach\_3\_gradient + reach3\_design\_cr + features\_y\_n + banks\_y\_n



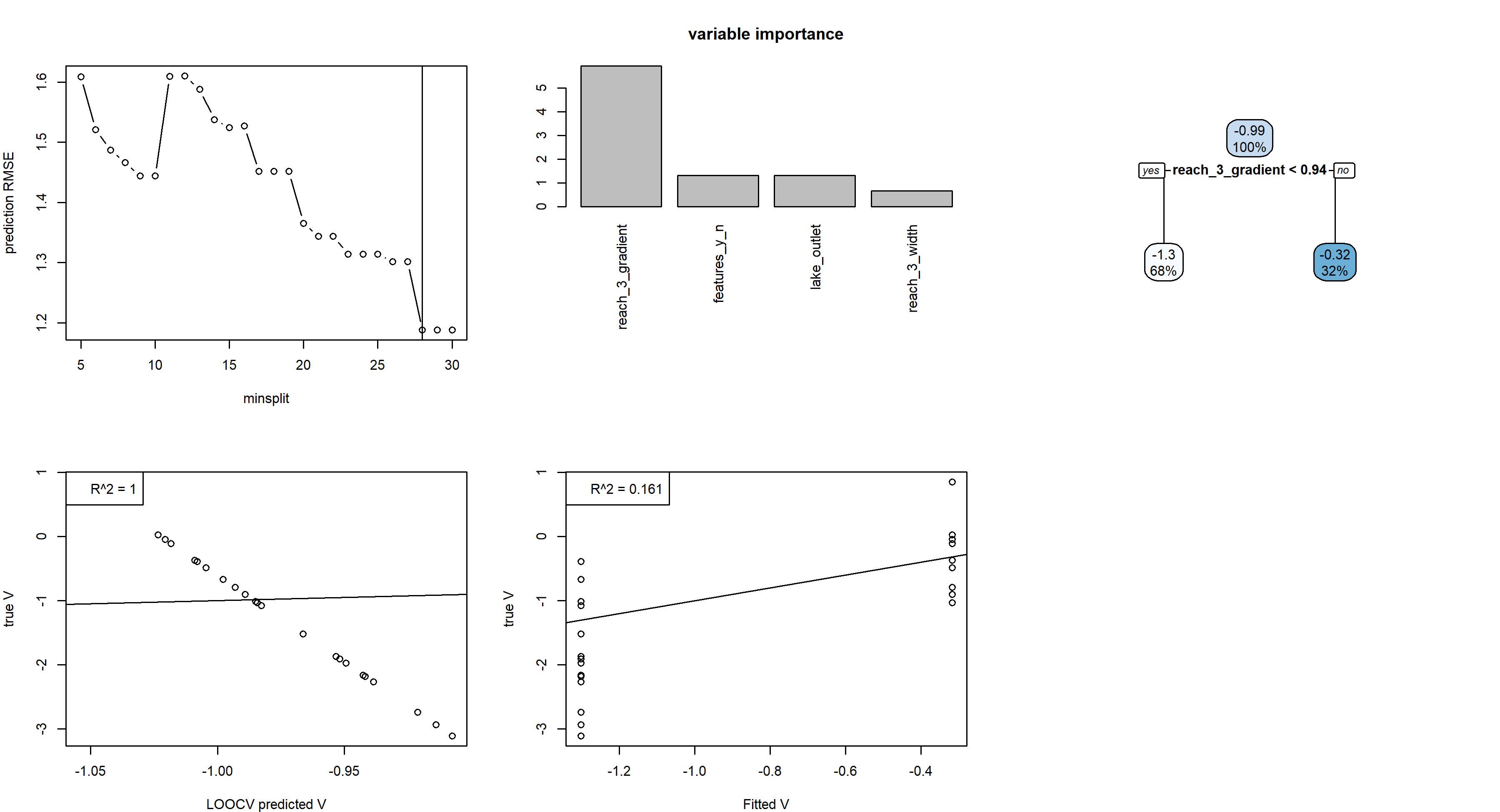
### Reach 3 gradient > 1.5

Top model:

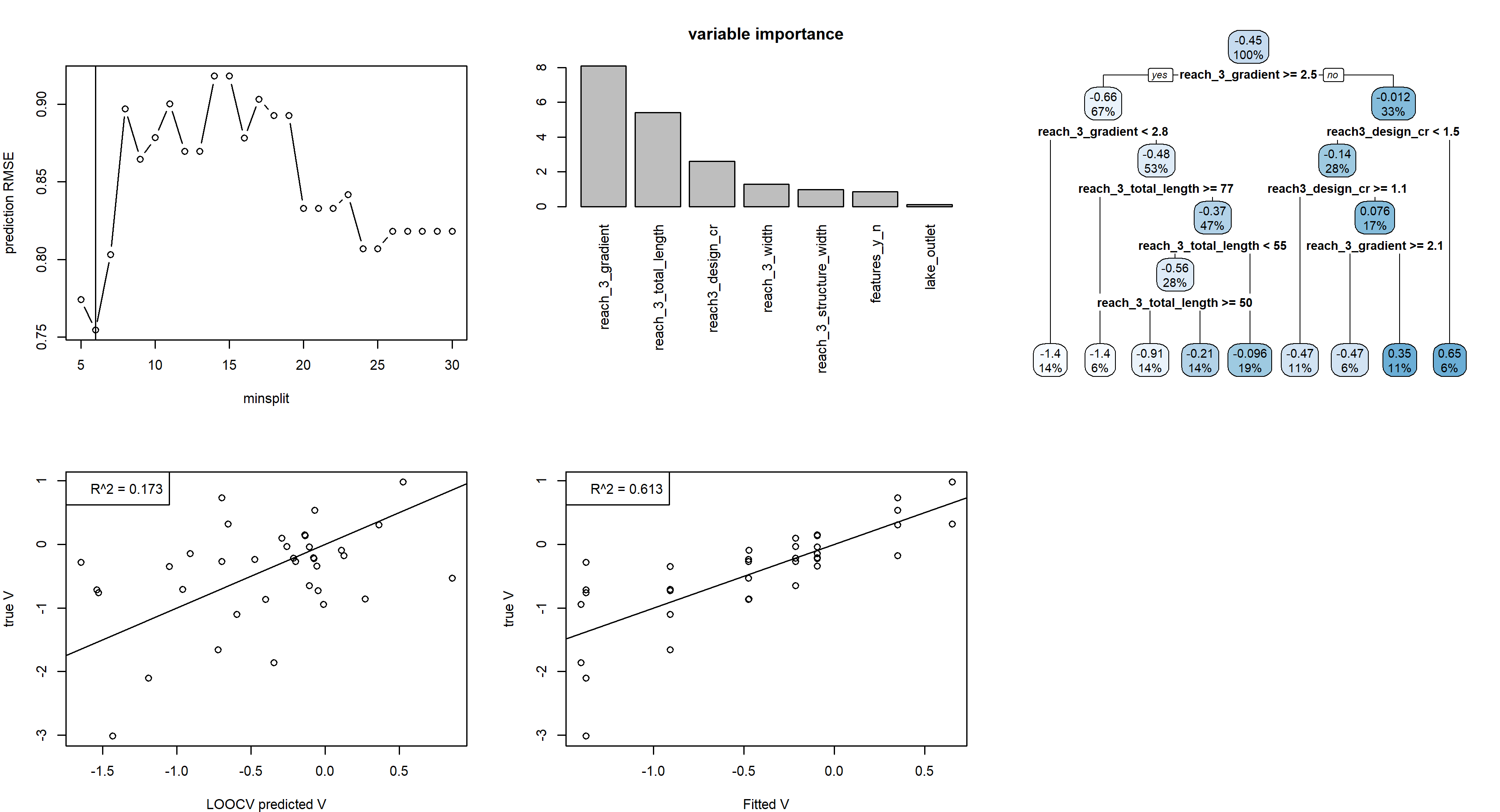
V Interior Gradient ~

## Interior Gradient - Regression Tree

### Reach 3 gradient <= 1.5



### Reach 3 gradient > 1.5

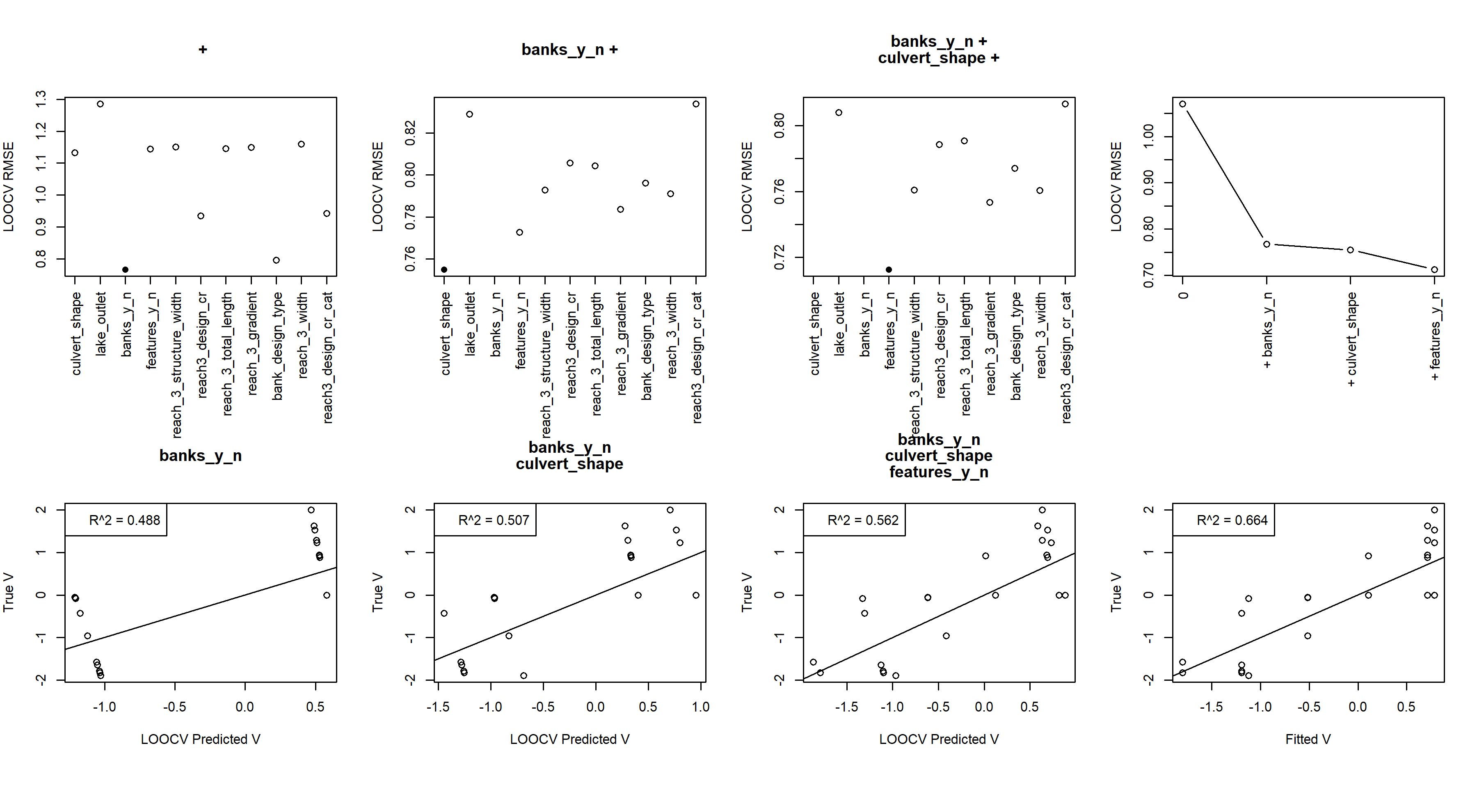


## Bank Length - Multiple Regression

### Reach 3 gradient <= 1.5

Top model:

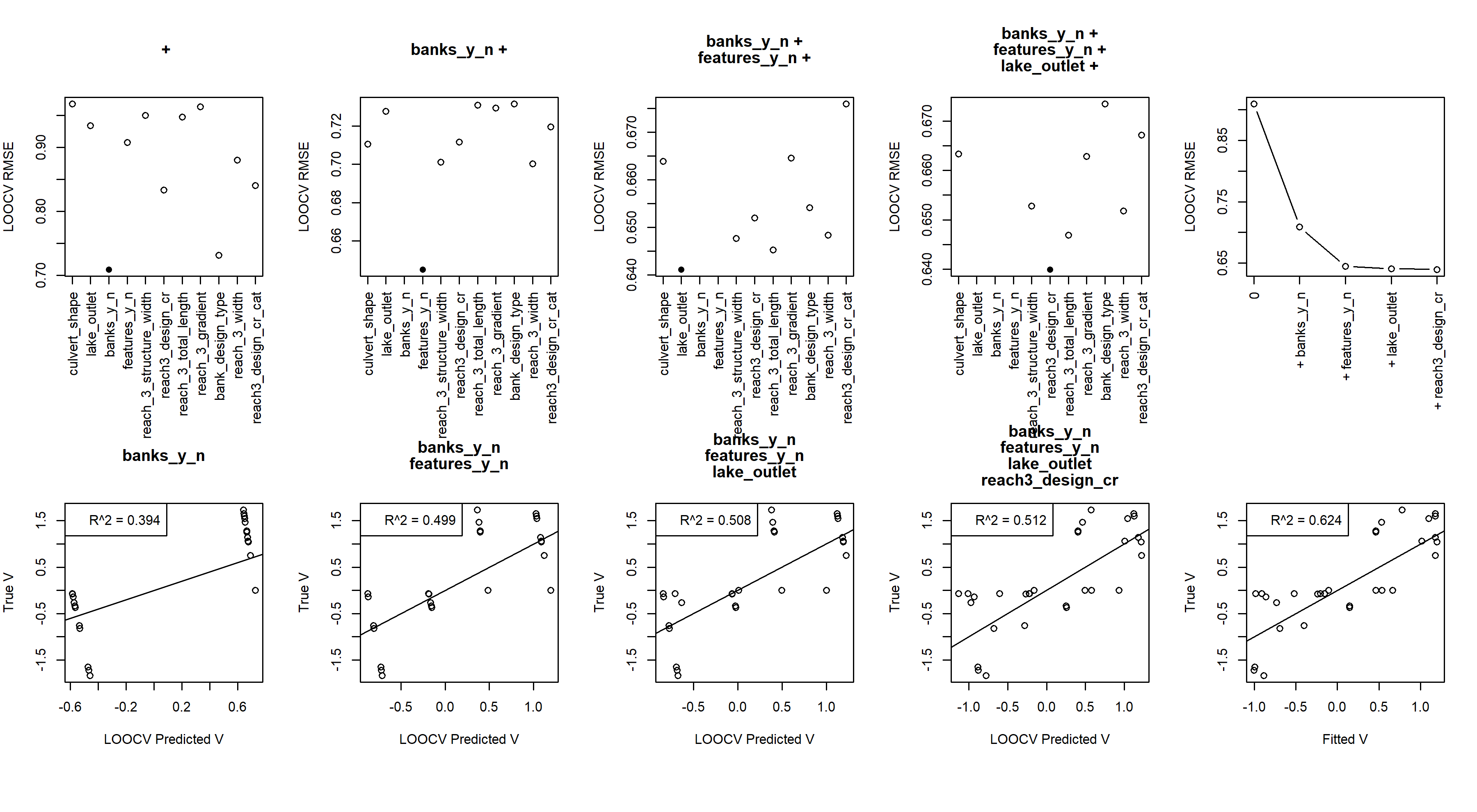
V Bank Length ~ banks\_y\_n + culvert\_shape + features\_y\_n



### Reach 3 gradient > 1.5

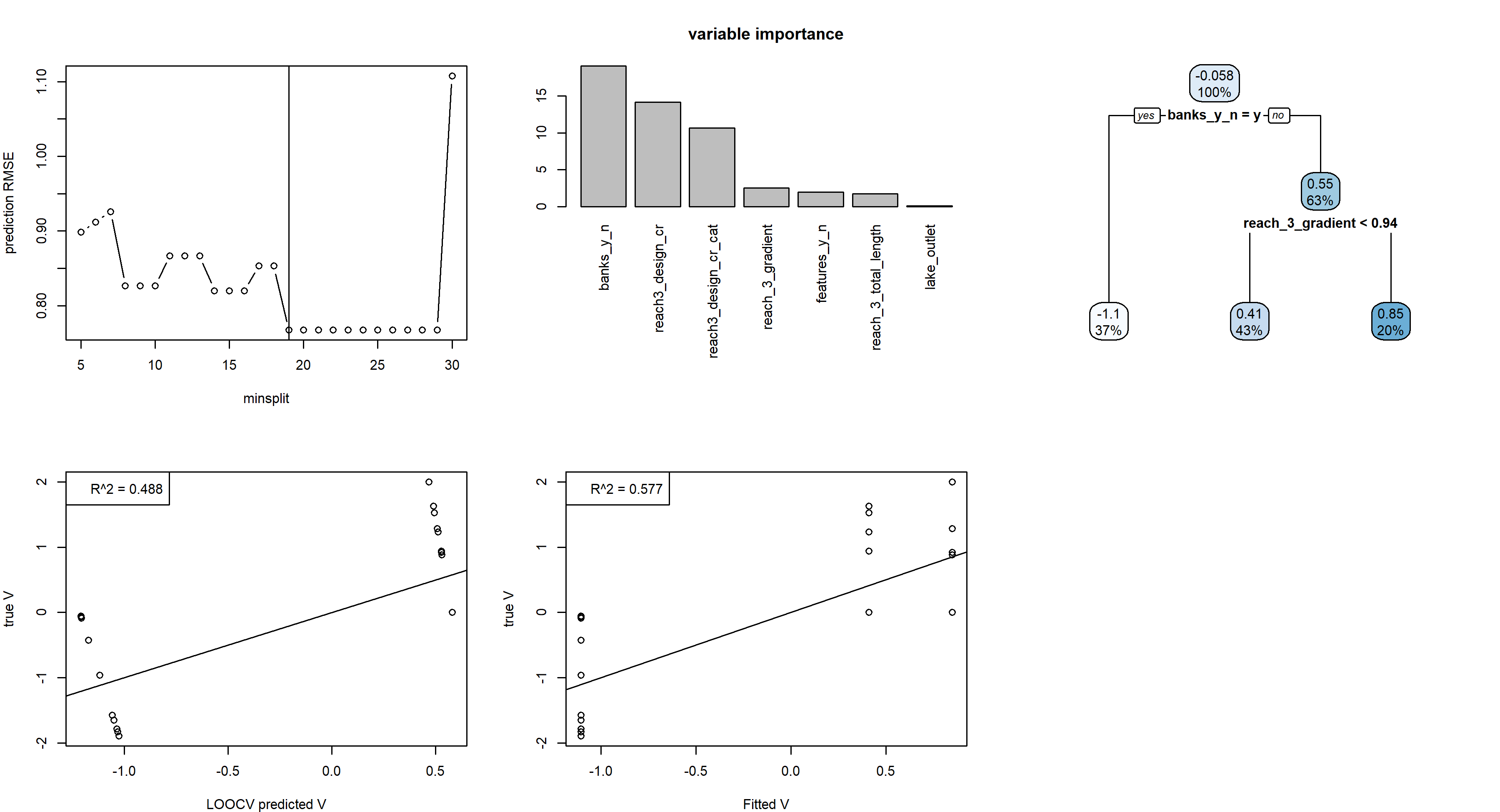
Top model:

V Bank Length ~ banks\_y\_n + features\_y\_n + lake\_outlet + reach3\_design\_cr



## Bank Length - Regression Tree

### Reach 3 gradient <= 1.5



### Reach 3 gradient > 1.5

