Were used five datasets (attrition, cars evaluation, churn, german credit, HPC) to evaluate the codings and compare them.

Seventy-five percent of the data per training set were used to start the analyses, and 10-fold-cross-validation were used to optimize the models.

When the outcome variable showed two levels, the area under the roc curve was maximized, and for the other datasets the multinomial log likelihood was optimized

For each dataset, the models were fit to the data in their original form and to the dummy (unordered) variables.

For the ordered datasets was added a model using ordinal dummy varibials (i.e., polynomial contrast)

10 different models (Single CART trees, Bagged CART trees, Single C5.0 trees and single C5.0 rulesets, Single conditional inference trees, Boosted CART trees, Boosted C5.0 trees, Boosted C5.0 rules, Random forests using CART trees, Random forests using conditional (unbiased) inference trees) were used, each adapted to the data set, and predefined software parameters were used.

For each individual model, performance metrics were estimated using resampling and total time to train and optimize the model.

From the comparison of factor vs ordinary dummy and factor vs unorderd dummy it shows that the difference in performance is calculated as follows:

$$\%difference = \frac{factor-dummy}{facotr} \text{ x } 100$$

For the attrition, churn and german credit datasets, each has two classes, the roc curve shows that no difference is present between the simple factor and dummy variable codifca in the area under the roc curve. This results appear even when comparing factor varibailes with dummy varibailes generated by polynomial constrast for ordinal predictors.

When ensemble methods are used (stochastic gradient boosting and bagged car trees) there is a 2% to 4% drop in the curve when using factor rather than dummy variables

The second metric used was overall accuracy. We used two different models and they show two differences results. The two models shows difference in codings when comparing dummy and factor.

Example churn shows no differences from the roc curve metric.

Regarding to the car evaluation dataset, it shows that the factor encodings are better than dummy varibials, this may be due to the fact that it consists of 4 classes and all the predictors are categorical. For this reason, it's possible that shows the maximum effect compared to other datasets.

No coding differences are present in the case of dummy varibials generated by contrast polynomial

Also in the car evalutation dataset, there are no differences between factor coding compared to contrast polynomial, but compared to the unordered dummy variable , factor coding are higher. This indicates the trend in the data follows the polynomial model.

Regarding performance, the difference are rare. In the car dataset since the variables are all categorical and this is a good indicator to use factors.

Few differences were observed and it is difficult to predict when these differences will happen .

Also the training time of the models was calculated and it was observed that the speed-up of factors is higher than dummy variables (2.5 slower). It is possible to conclude that factor-based models are trained more efficiently. This maybe happened because of the extended number of predictors due to the generation of dummy variables , which needs more computational time (exception are models using conditional inference trees)

The way qualitative predictors are coded is related to summary measures , for example tree-based models calculate variabiles importance scores that measure how much a predictor influences the results.

The tree model measures the effect of a split on improving model performance. The predictors that are used in the split and these improvements are aggregated and can be used as the importance score. For example if a split involves all the values of the predictor example Saturday versus other days it is likely that the score for the whole variable are very large compared to the single variable (Saturday or non-Saturday)

Conclusion: use preditions without converting them to dummy variables