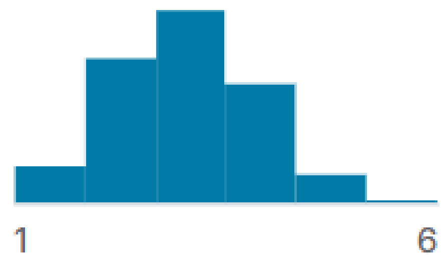
Task AES: Automated Essay Scoring

Federico Rausa – mat . 919795

Dataset

Variables:

- Full text: essay written by the student as unique string
- Score: evaluation of the essay given by the teacher (int from 1 to 6), who should follow general criteria

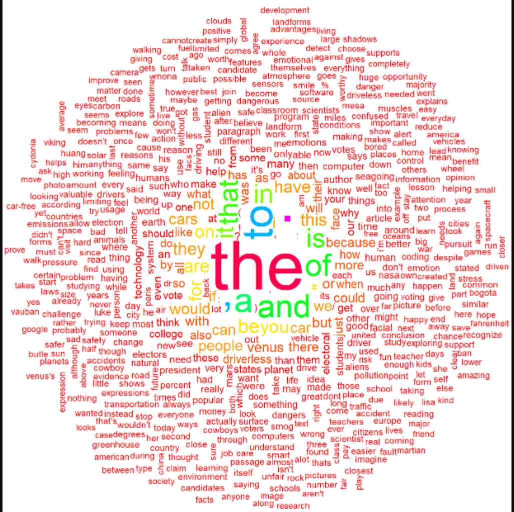
full_text	score
17307 unique values	
Many people have car where they live. The thing they don't know is that when you use a car alot of t...	3
I am a scientist at NASA that is discussing the "face" on mars. I will be explaining how the "face" ...	3
People always wish they had the same technology that they have seen in movies, or the best new piece...	4
We all heard about	4

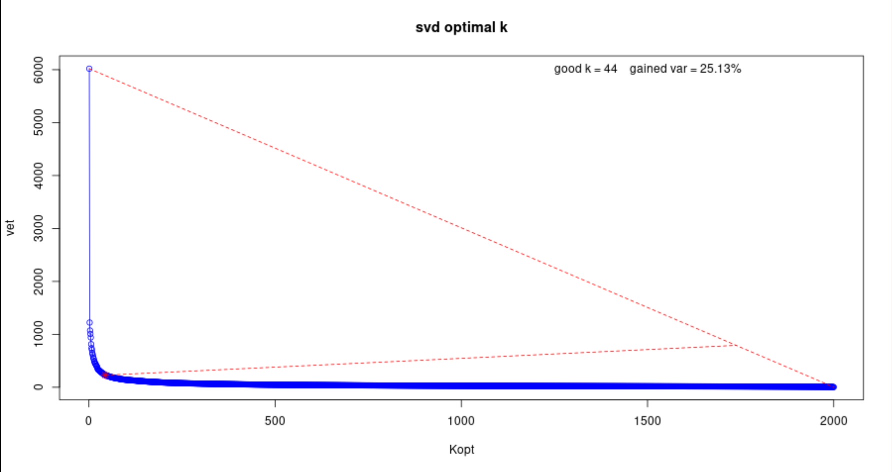
Full Text preprocessing:

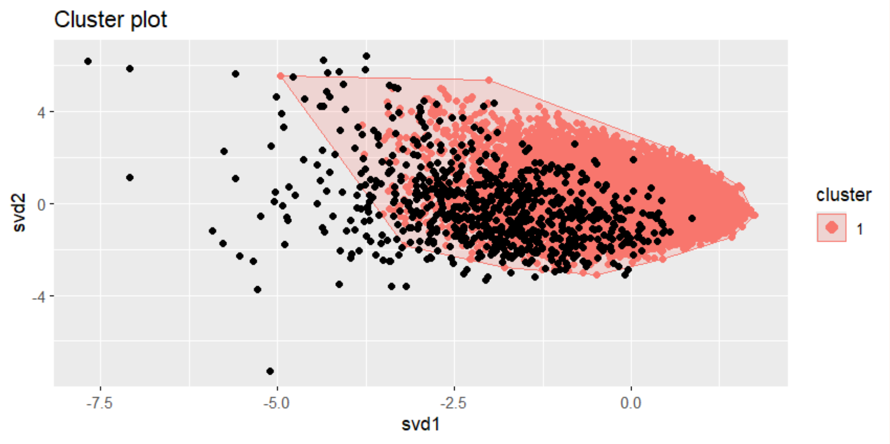
1. DTM + trimming

2. SVD and choice of #components
   (25 for 20% explained var)

3. Noise Removal with DBSCAN
   (#noisePts=~760, #clusters=1, eps=5, minPts=5)







Data splitting for supervised learning

- Splitting with Stratified Sampling (follow score distribution)
- Validation set: ~ 4000 obs (for hyper-params optimization)
- Train-Test set: ~ 12000 obs (for bootstrap 50% comparisons)

	1	2	3	4	5	6	N
scoreMainDist	7.23	27.29	36.29	22.68	5.6	0.9	17307
scoreValidationDist	6.97	28.04	37.01	22.61	4.82	0.55	4002
scoreTrainTestDist	7.15	27.86	37.02	22.61	4.84	0.53	12551
scoreNoiseDist	10.08	13.79	20.29	24.27	22.55	9.02	754

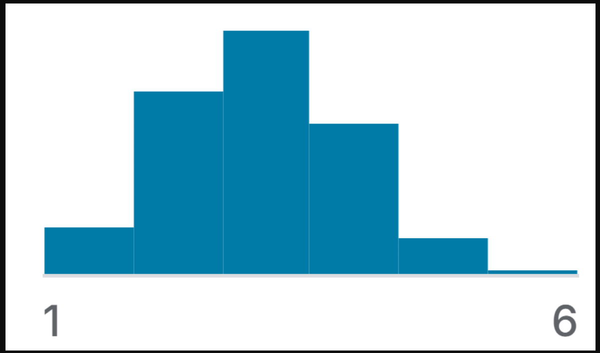
Evaluation metrics used:

- Quadratic Weigthed Cohen's Kappa (QWK) :

higther penalties for errors far from the diagonal of the Confusion Matrix
[-1, 1]

$$QWK = 1 - \frac{\sum W_{ij} O_{ij}}{\sum W_{ij} E_{ij}} \quad \text{with weigths } W_{ij} = \frac{(i-j)^2}{(K-1)^2}$$

- Accuracy [0,1]
- MAE [0, +Inf)
- F1 score on extreme classes (1 and 6) [0,1]



Different supervised models for different task

- Regression Models
(minimize MAE):
Lasso, KNN, SVD, GBM, XGB

- Classification Methods
(maximize Accuracy):
Decision Tree, SVD, GBM, XGB

- Ranking Methods
(maximaze QWK):
Cumulative Logit, XGB with Pairwise Loss, ranking with binary groups

Hyper – parameters optimization on Validation Set

With Grid-Search:

- Lasso (L1 regularization)
- KNN (K)

With Bayesian Optimization:

- SVD (kernel, C, gamma)
- GBM (learning rate, #trees for boosting, bag fraction for each tree)
- XGB (learning rate, #trees for boosting, regularization,
- gamma/minimal error reduction for a new split)

Ranking models

- Ordinal logistic regression:

$$P(y_i \leq j) \; = \; \frac{exp(\alpha_j + \beta x_i)}{1 + exp(\alpha_j + \beta x_i)}$$

$$P(y_i = j) \; = \; P(y_i \leq j) \; - \; P(y_i \leq j - 1)$$

- XGB for the minimization of the Pairwise Logistic Loss:

$$pairwiseLoss = L(\overline{\boldsymbol{x}}, \overline{y}, f) \; = \sum_{(i,j) \in Pairs} loss_{ij}$$

$$loss_{ij} = -\left(z_{ij} * ln\left(p_{ij}\right) + \left(1 - z_{ij}\right) * ln\left(1 - p_{ij}\right)\right)$$

$$\Delta_{\mathrm{ij}} = f(x_i) - f\left(x_j\right)$$

$$p_{ij} = \frac{1}{\left(1 + exp\left(-\Delta_{ij}\right)\right)}$$

$$z_{ij} = \begin{cases} 1 & y_i > y_j \\ 0 & y_i < y_j \end{cases}$$

Ensemble strategy for ranking

Group of binary models for CDF prediction:

- K-1 binary models to predict K-1 cdf dummies
- Same hyper-parameters to follow

j=1	j=2	j=3	j=4	j=5	j=6
0	0	1	0	0	0
0	1	0	0	0	0
0	0	0	1	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	1	0	0	0	0

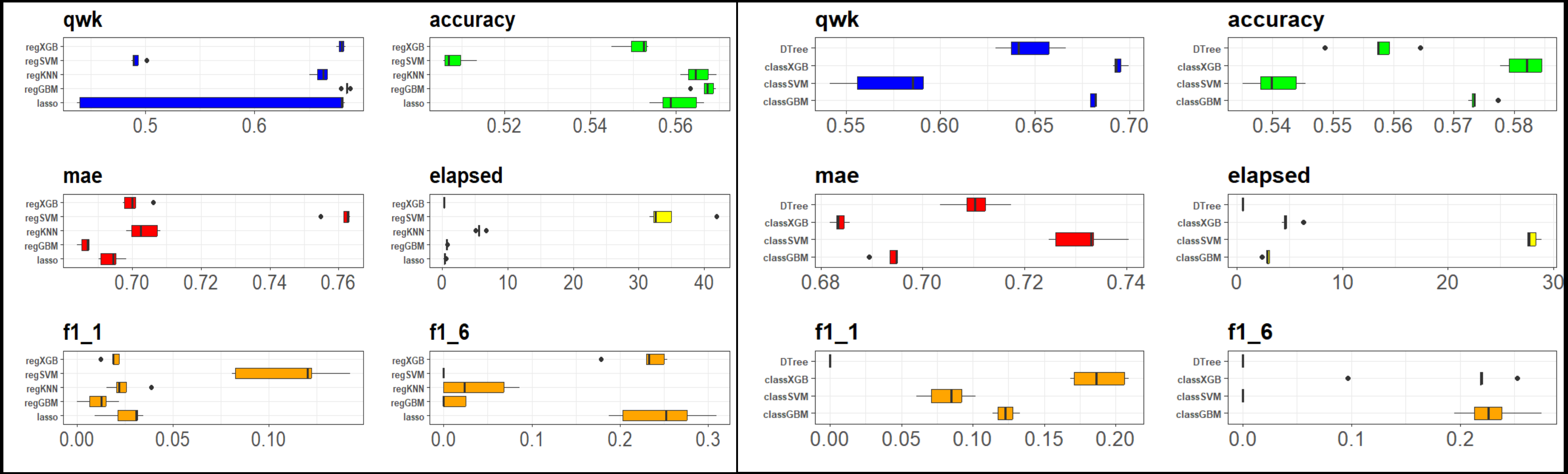
j<=1	j<=2	j<=3	j<=4	j<=5	j<=6
0	0	1	1	1	1
0	1	1	1	1	1
0	0	0	1	1	1
0	0	1	1	1	1
0	0	0	1	1	1
0	1	1	1	1	1

$$P(y_i \leq j) \; > P(y_i \leq j-1)$$

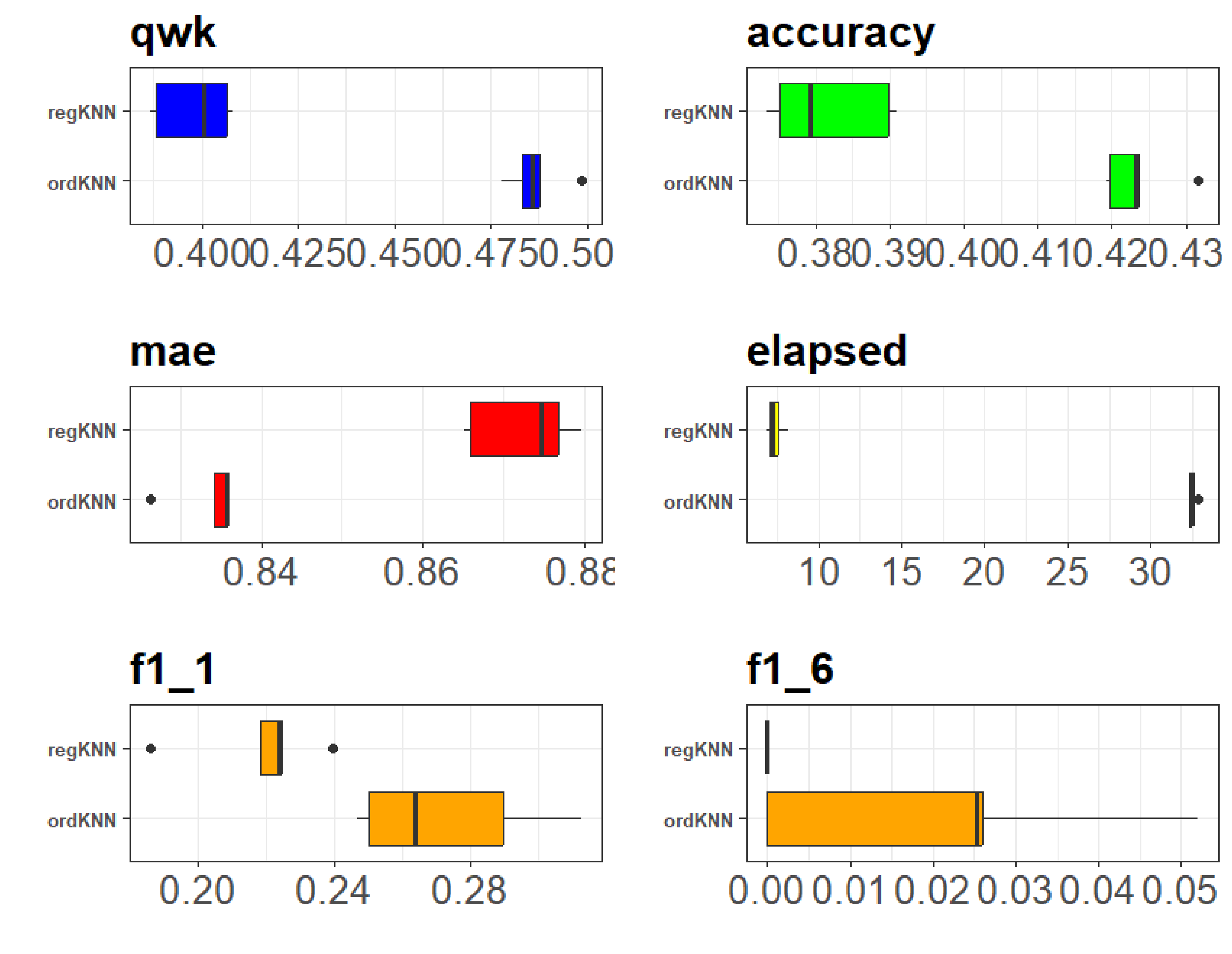
$$y_{ij} = \begin{cases} 1 & y_i \geq j \\ 0 & y_i < j \end{cases}$$

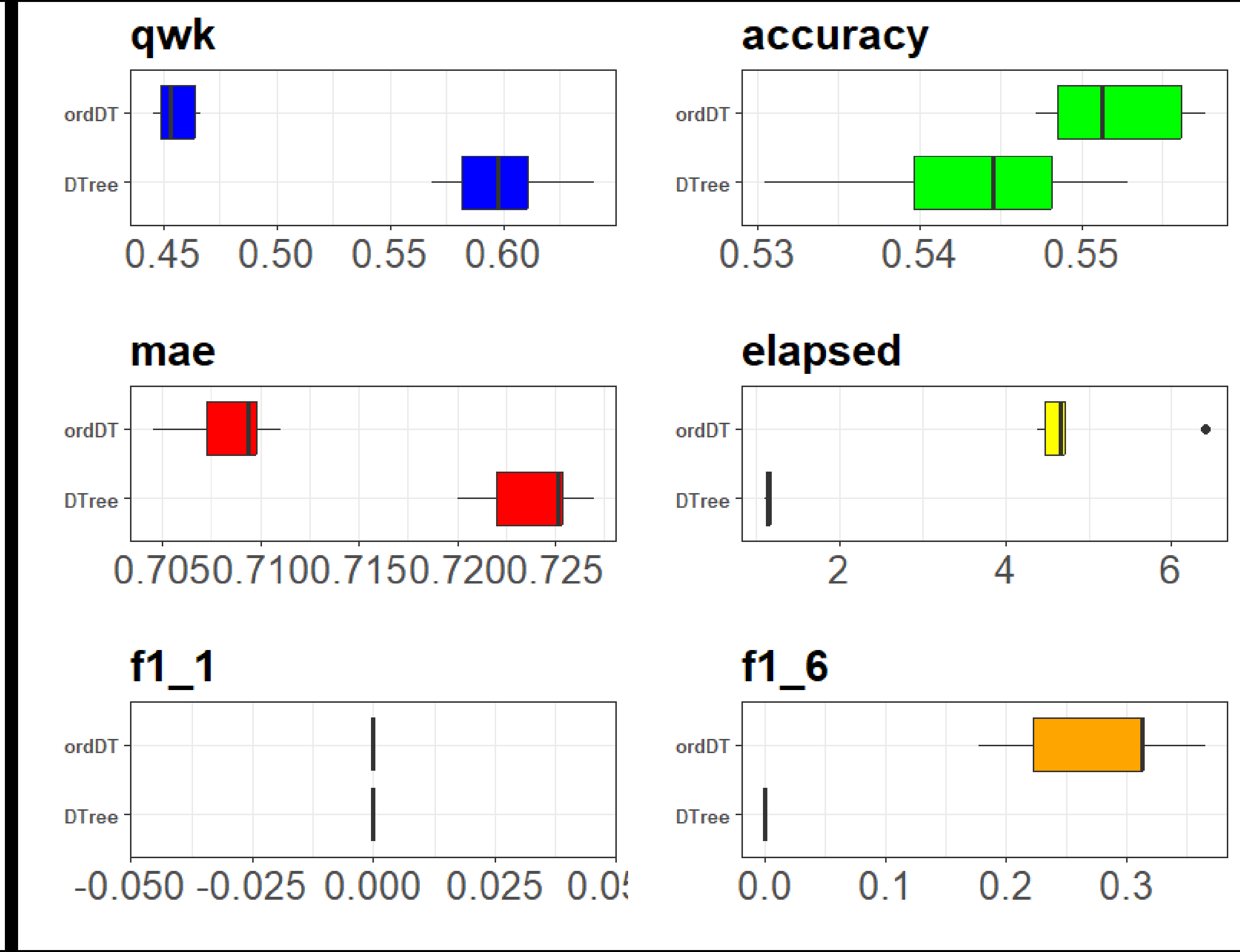
$$P(y_i = j) = \big(1 - P(y_i \geq j)\big) - \big(1 - P(y_i \geq j-1)\big) = P(y_i \geq j-1) \; - \; P(y_i \geq j)$$

Bootstrap final comparisons: classification and regression

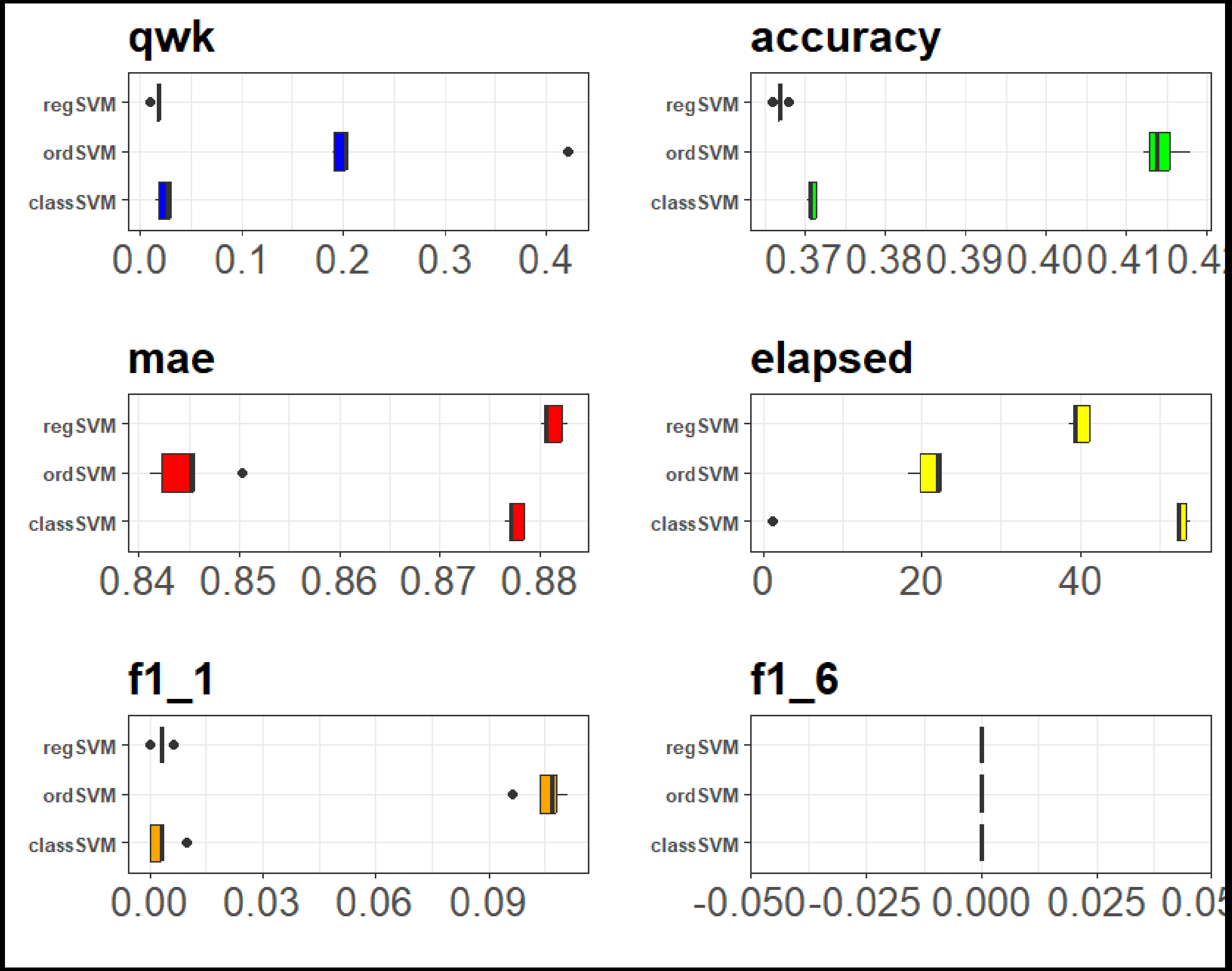


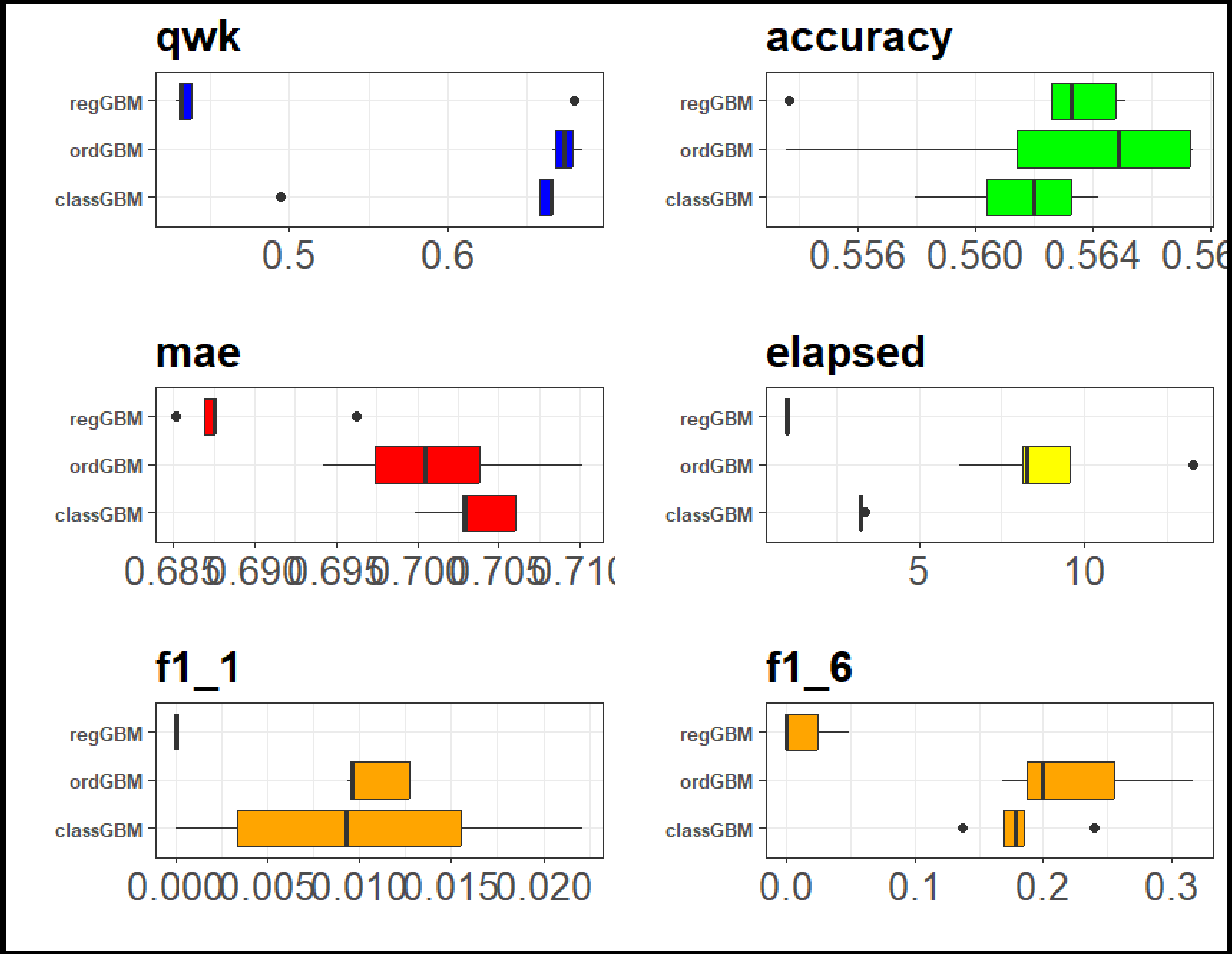
Bootstrap final comparisons: improvements of ranking groups (KNN , DT)



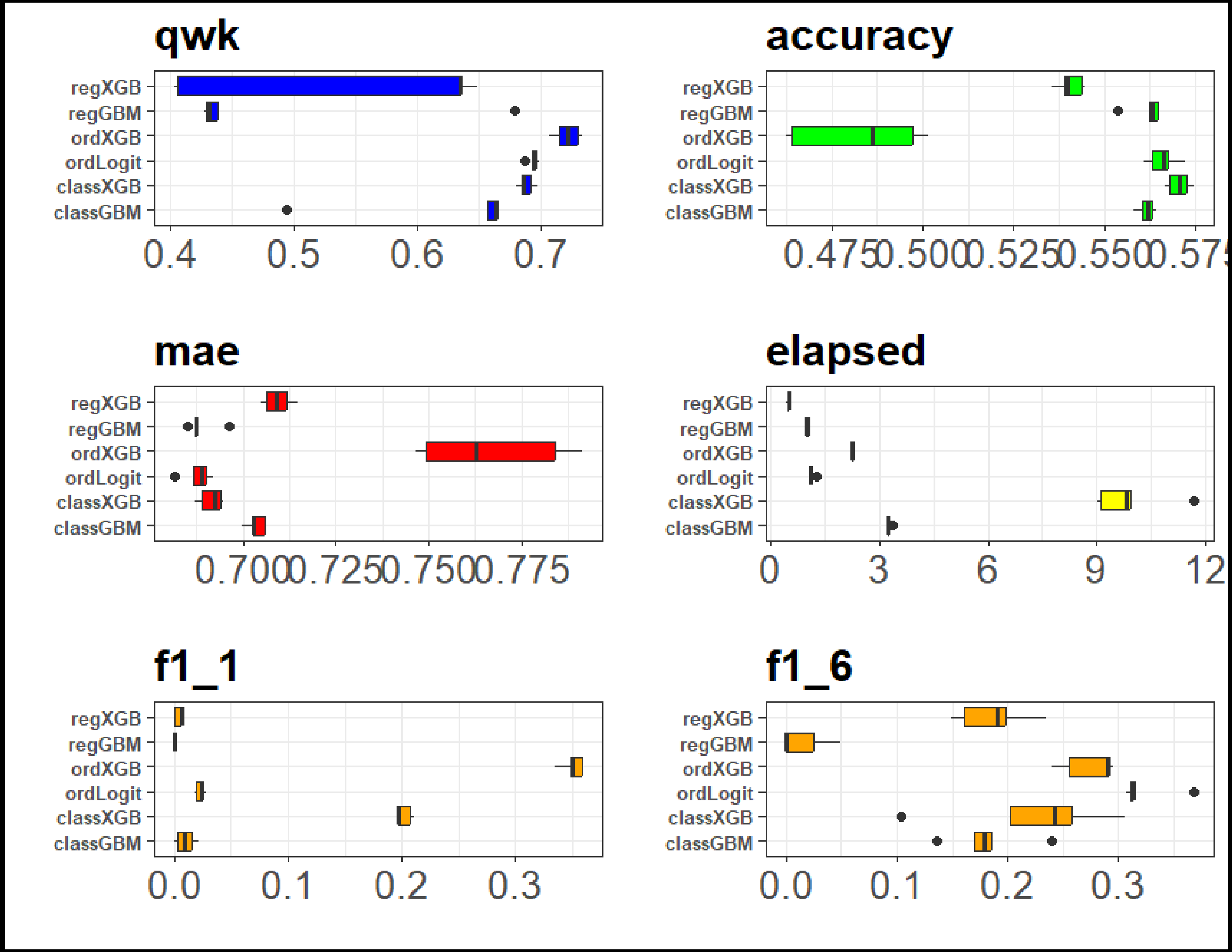


Bootstrap final comparisons : improvements of ranking groups (SVD , GBM)

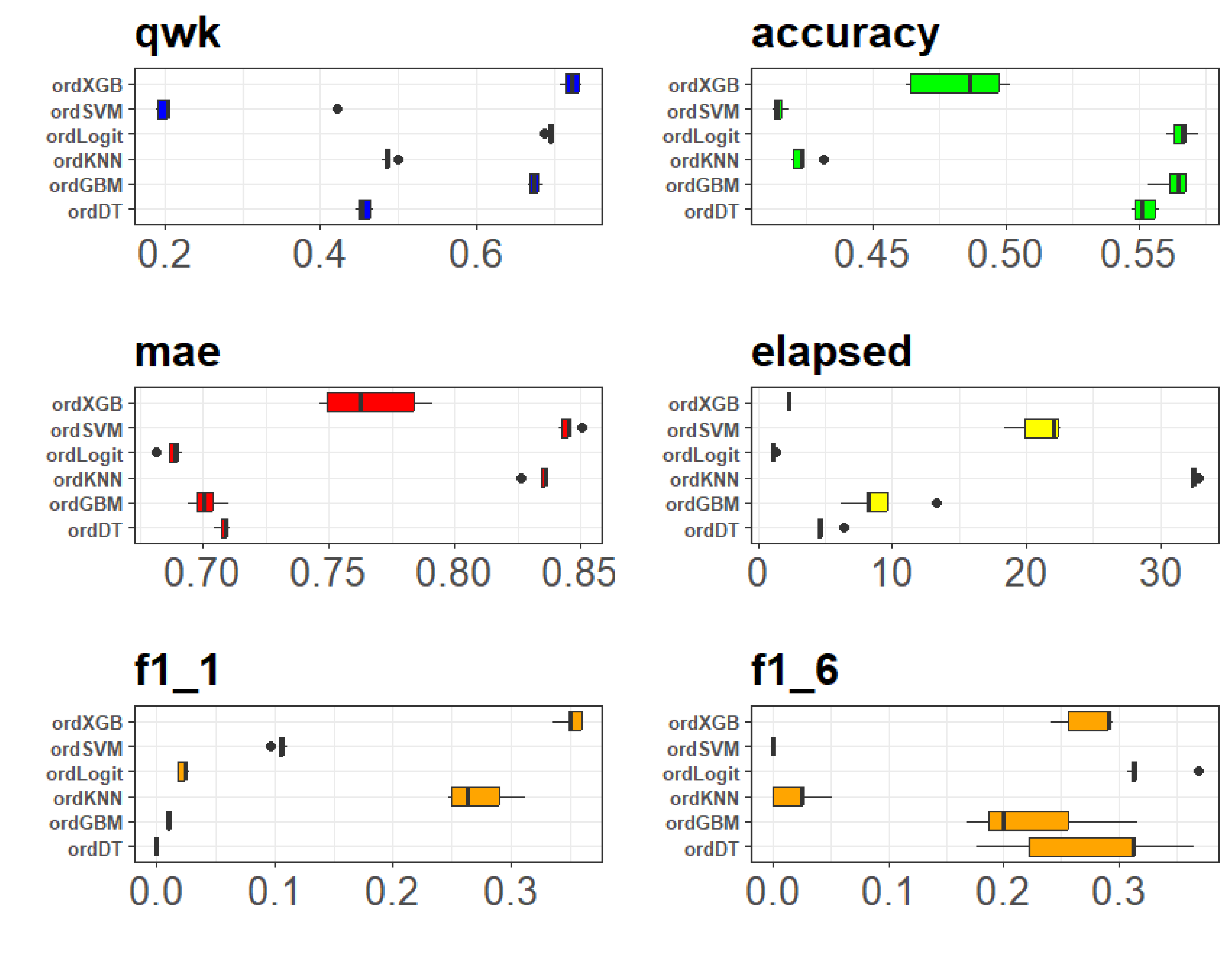




Bootstrap final comparisons: best models



Bootstrap final comparisons: ranking models



THANK YOU