

MINING

Hugh Chipman (Acadia University), Erika Nahm (DND),

What are “Network Data?”



- A naughty US Company
- Over 125,000 emails from 184 employees published.

From: hugh@enron.com
To: tim@enron.com
CC: sue@com.com, ken@enron.com
Date: 01:12 Dec 9, 2004 (ADT)
Subject: We're really in it now
Hi Tim, BlahBlahBlahBlah implicate BlahBlah criminal...

01:12 Dec 9, 2004

hugh@enron.com

tim@enron.com

- Thousands (millions) of such transactions available as our “data”.
- **Supervised Learning:** Classify graph nodes (here email addresses are senior/junior, 60 out of 184 are senior).
- Other goals possible - EDA, model entire graph, clustering, sampling questions.

Approach:

- Convert email headers from 125,000 into a “feature matrix” for 184 email addresses.
- Built 23 node-centric features:
 - **node-centric used here:** communication density, degree, time of day, ...
 - Others possible, (graph-based, subject co-variates like age)
- Usual classifiers + train/validate/test.
- Performance assessment via misclassification rates, lift curves, ROC curves.

Performance: Enron Data

	Misclass. (s.e.)	AUC' (s.e.)
Logistic regression (LR)	.25 (.02)	.50 (.13)
Decision trees (DT)	.31 (.02)	.35 (.11)
Random forest (RF)	.18 (.03)	.52 (.11)
Boosted trees (BT)	.22 (.02)	.45 (.14)

We want small misclass, large AUC'.

All methods better than random guessing

Rare target challenge...

- Another dataset
- 2.4 million transactions
- 21,533 nodes, but...
- **rare target:** only 652 nodes of “positive” class
- same 23 node-centric features
- find as many unknown positives as “quickly” as possible
 - produce **ranking**, and investigate “top- k ” unknown nodes
- how to evaluate performance in view of objective and class imbalance?

Rare target performance

- $T = P + N$ unknown nodes are ranked.
- $k \leq T$ is the threshold (operating value),
- top- k nodes are classified as positive,
- other $T - k$ are classified as negative.

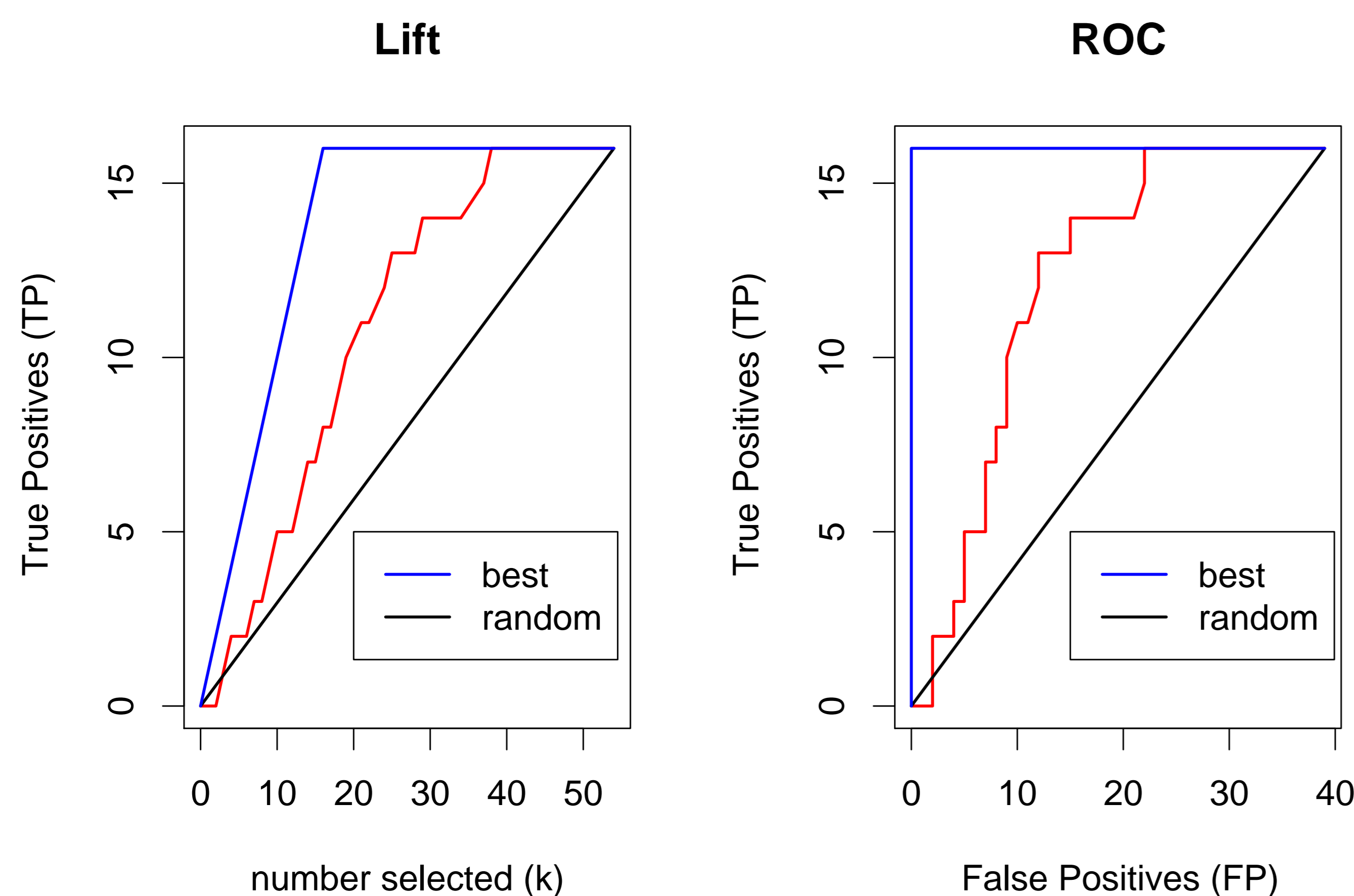
	predicted		
true	Pos	Neg	sum
Pos	TP	FN	P
Neg	FP	TN	N
sum	k	$T - k$	

- Graphical summaries: Lift/ROC curves

NETWORK DATA

François Théberge (University of Ottawa)

Lift, ROC curves

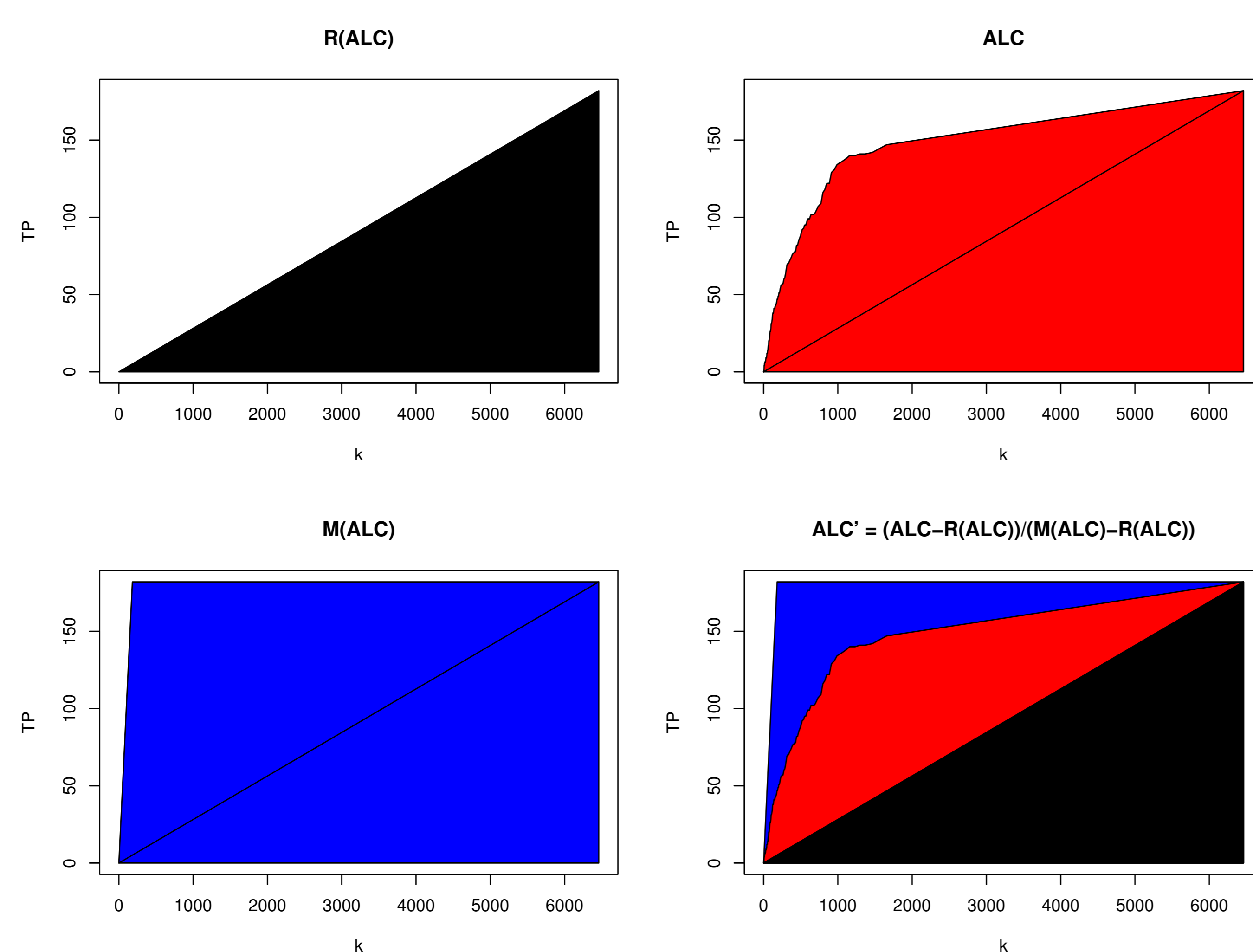


Common performance measures include:

- AUC (Area under ROC)
- ALC (Area under lift curve)

Result: $AUC' = ALC'$ (suitably normalized)

- $ALC \Rightarrow ALC'$ normalization below:

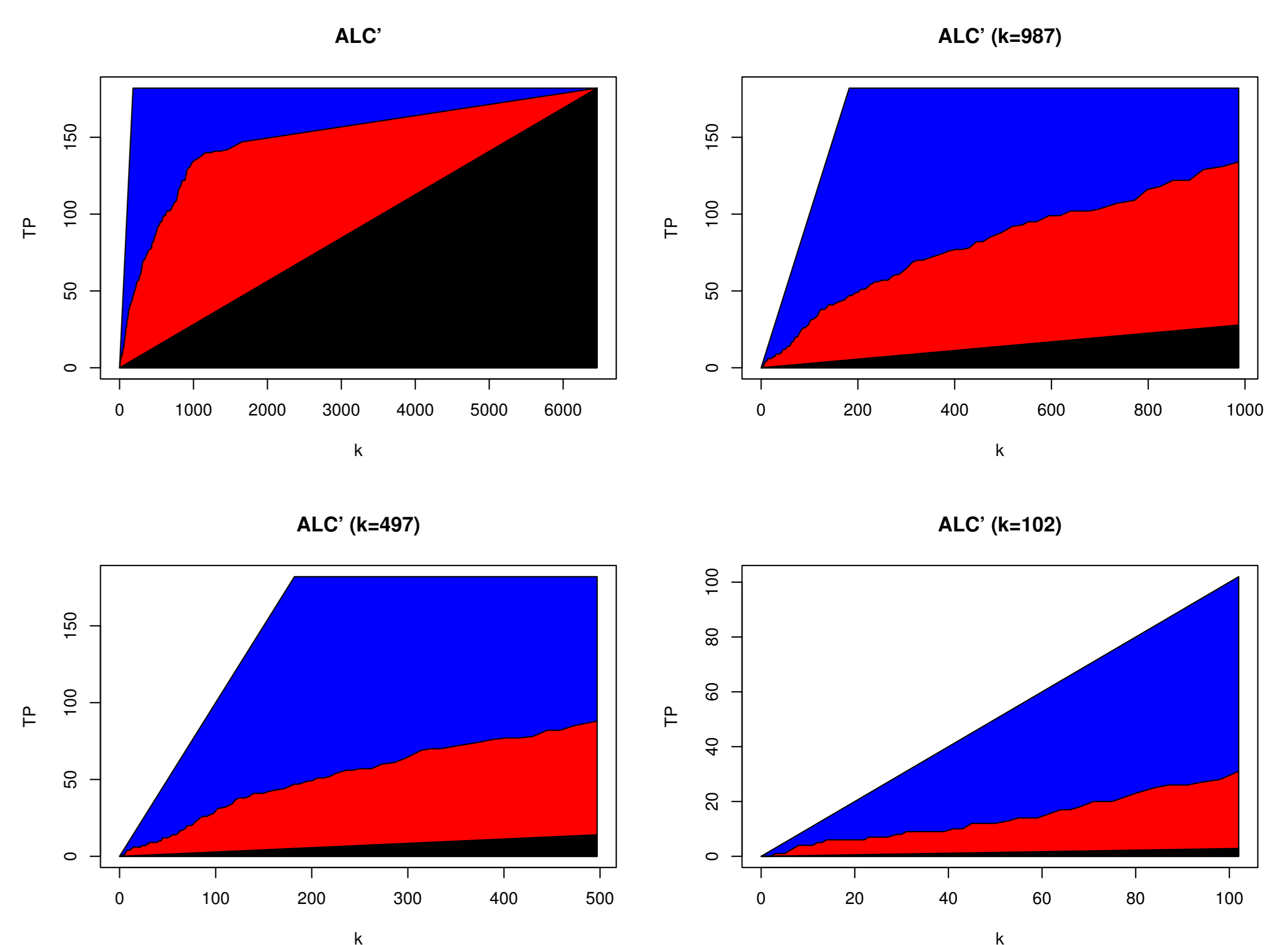


In lower right plot,

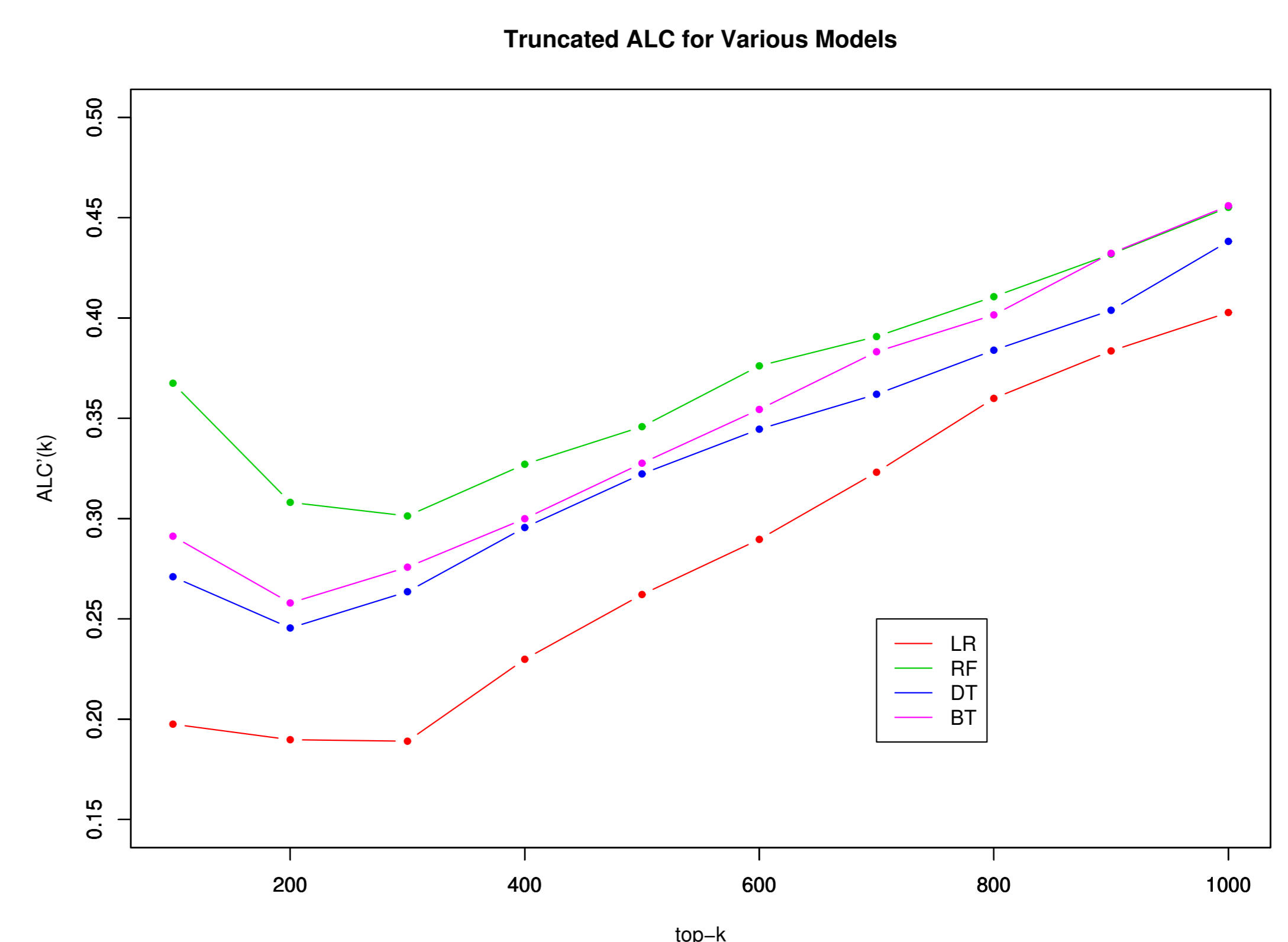
$$ALC' = \frac{\text{red}}{\text{red} + \text{blue}} = \frac{\text{improvement over random}}{\text{best possible}}$$

Improvement: truncated ALC

- AUC truncation problematic since horizontal axis (FP) is random.
- ALC'_k : normalize & truncate ALC, selecting top k cases.



ALC'_k results: rare target ex.



Significance? Derive $\text{Var}(ALC'_k)$

- Here, $\text{se}(ALC'_k) \leq 0.023$
 \Rightarrow all quite significant.

Aside: train/test split is tricky