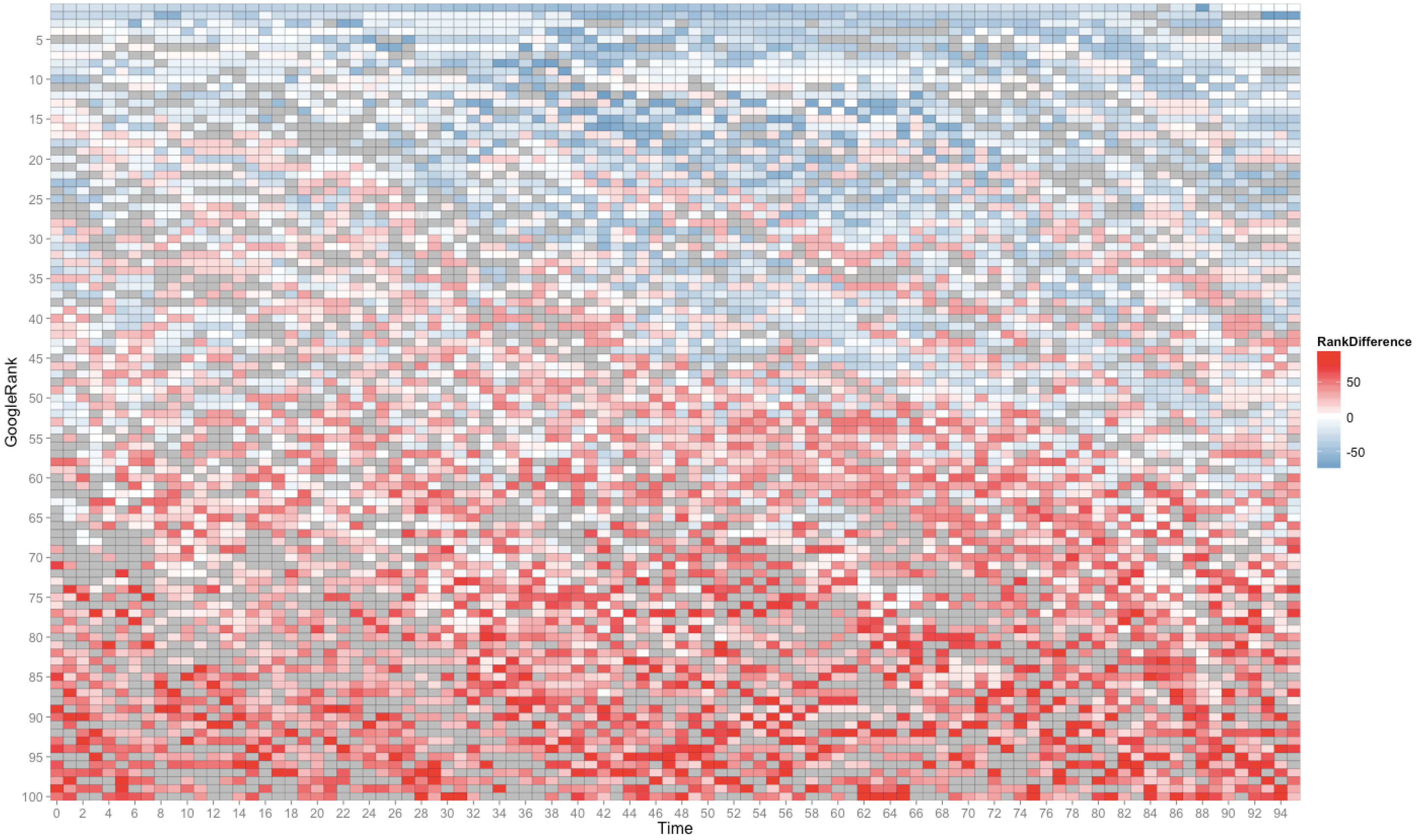
THE UTILITY PROBLEM OF WEB CONTENT POPULARITY PREDICTION

Nuno Moniz [1] and Luís Torgo [1,2]
[1] INESC TEC / University of Porto
[1,2] Dalhousie University

ACM Hypertext 2018
Baltimore, Maryland, USA

11th July, 2018

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
LIAAD
INESCTEC
LABORATÓRIO ASSOCIADO
U.PORTO


RankDifference
50
0
-50
GoogleRank
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Time
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
62
64
66
68
70
72
74
76
78
80
82
84
86
88
90
92
94

Introduction

- Profusion of user-generated web content
- Imbalanced degree of attention received
- Great interest in anticipating such attention
- 10+ years of related work

Web Content Popularity Prediction

- Formalized as classification, regression, forecasting, etc.

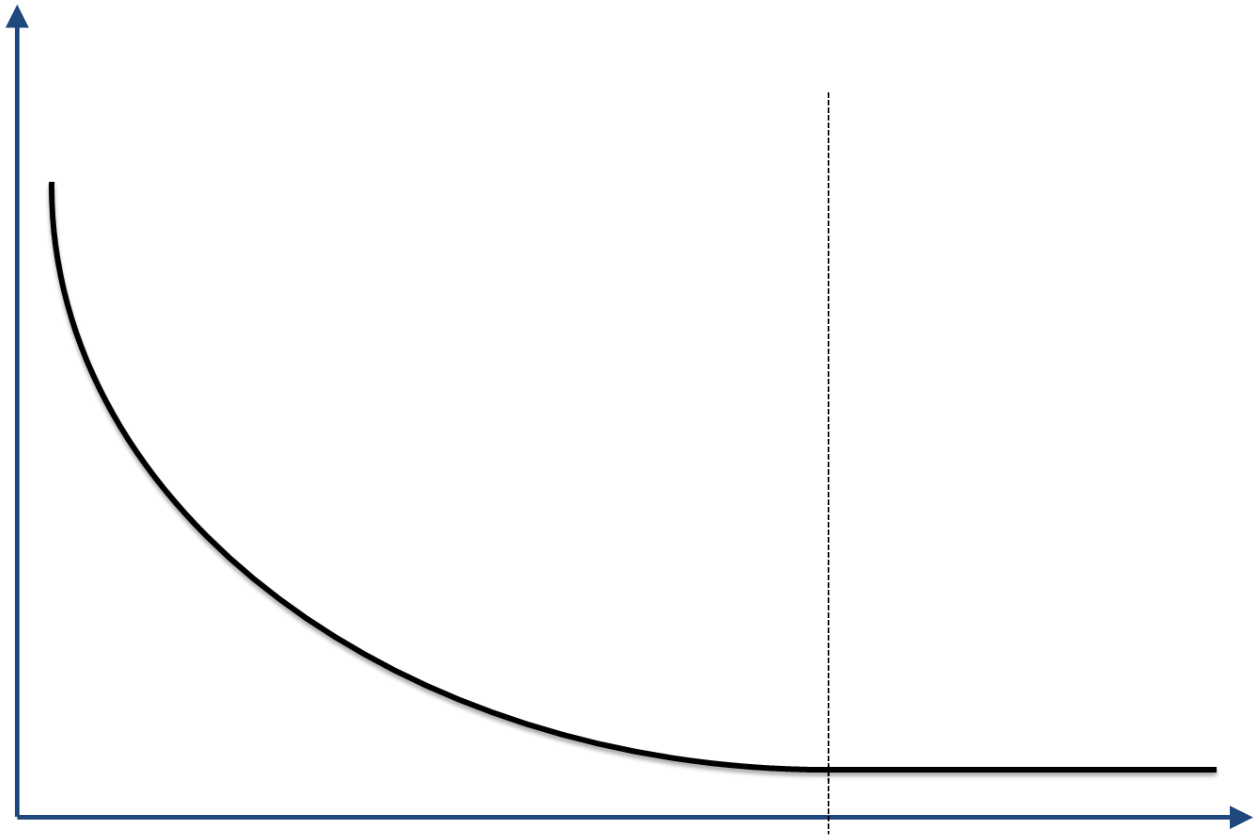
- Uses historical data to build prediction models in order to anticipate future popularity

Standard Assumption: Uniform Domain Preferences

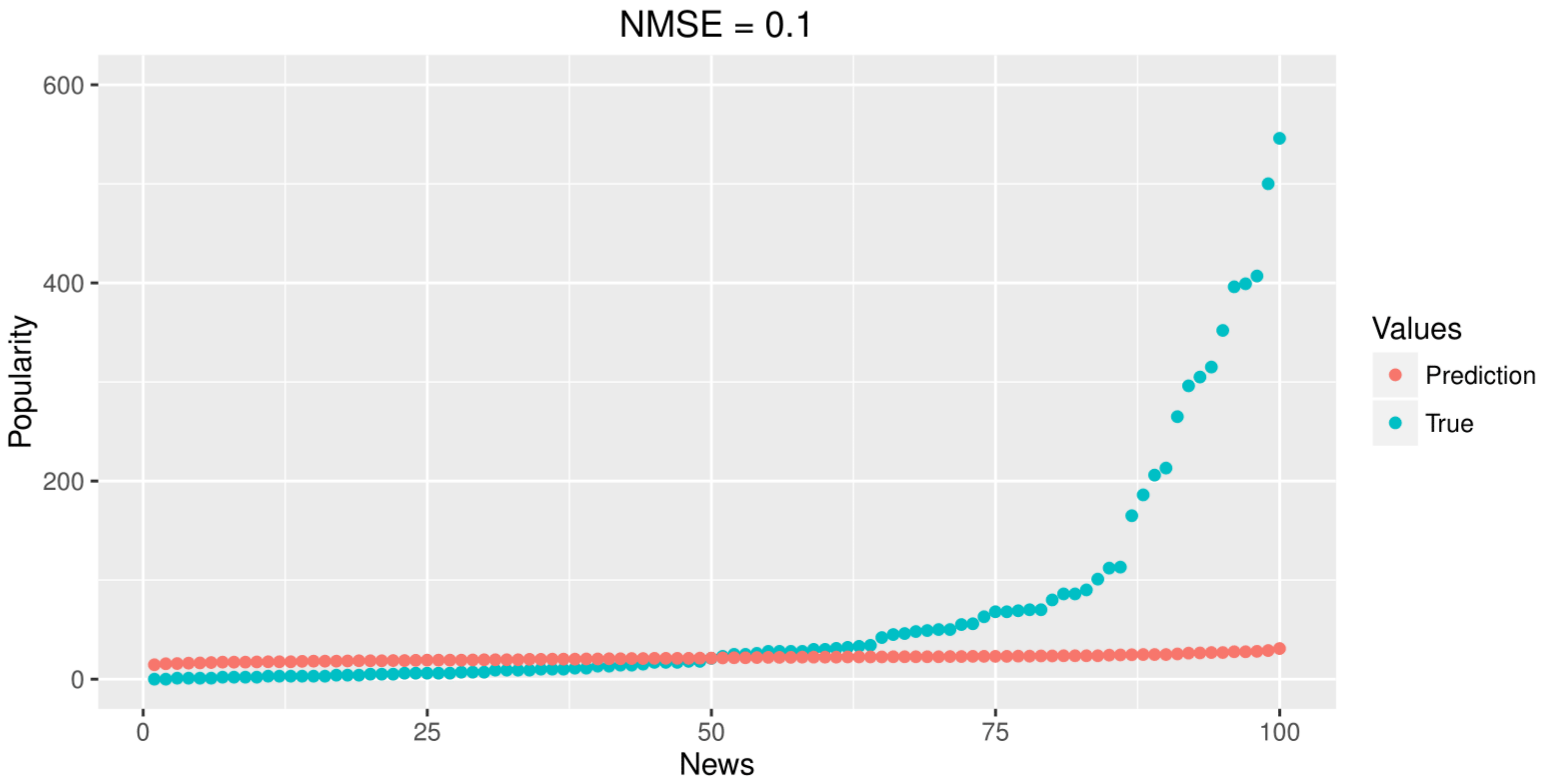
(in other words, every item is equally important)

What seems to be the problem?

- Popularity follows a heavy-tailed distribution



Evaluation/Optimization



Utility

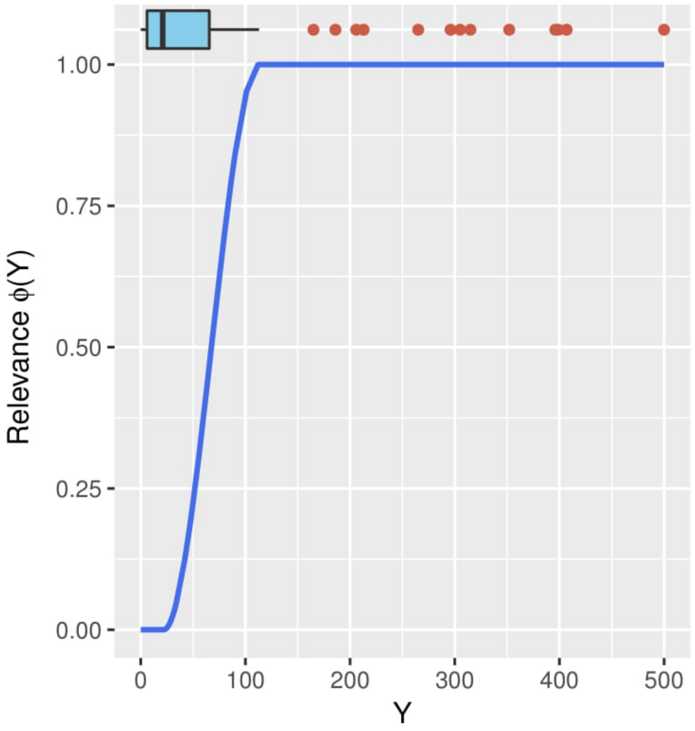
- Standard Learning Tasks

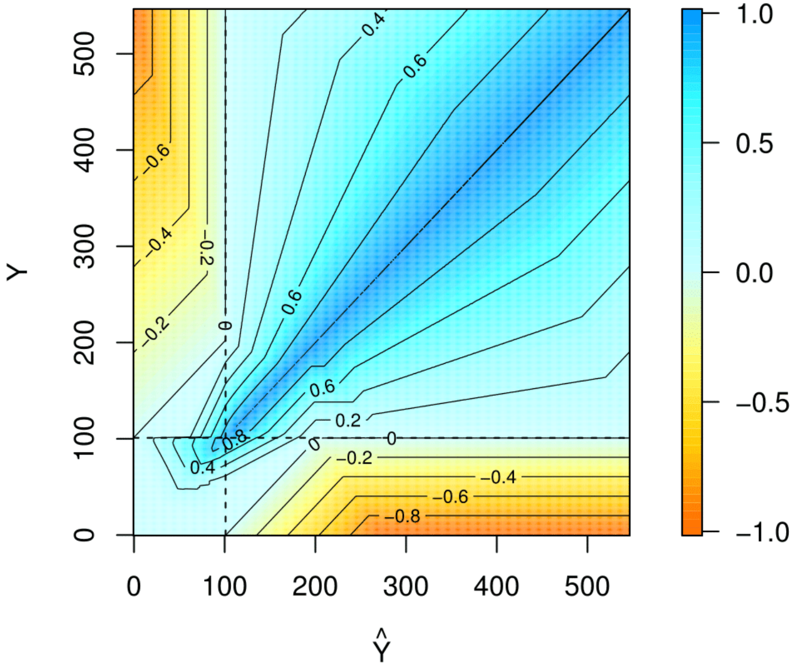
Equally accurate predictions have the same utility
More accurate predictions have greater utility
Less accurate predictions have lesser utility

- Utility-based Learning Tasks

Equally accurate prediction may have different utility
More accurate predictions may have lesser utility
Less accurate predictions may have greater utility

Utility-based Evaluation





Evaluation Metrics: Utility-based F-Score, Precision and Recall

Experimental Evaluation

- Task: Imbalanced Regression
- Methodology: Monte Carlo simulations

- Assumptions
  - Highly skewed distribution of target variable
  - Non-uniform domain preferences

- Data
  - News from Google News and Yahoo! News
  - Popularity from Facebook, Google+, LinkedIn
  - 20-minute slices
  - Available in UCI

Methods

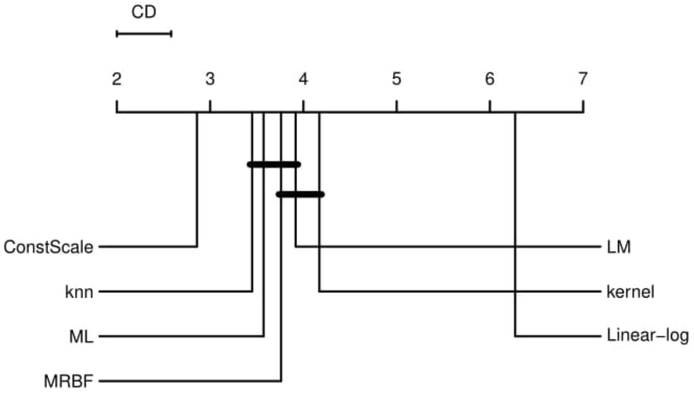
- Constant Scaling (Szabo and Huberman)
- Log-Linear (Szabo and Huberman)
- Multivariate linear regression (Pinto et al.)
- MLR with Radial Basis Functions (Pinto et al.)
- Linear with Sentiment Features (Asur and Huberman)

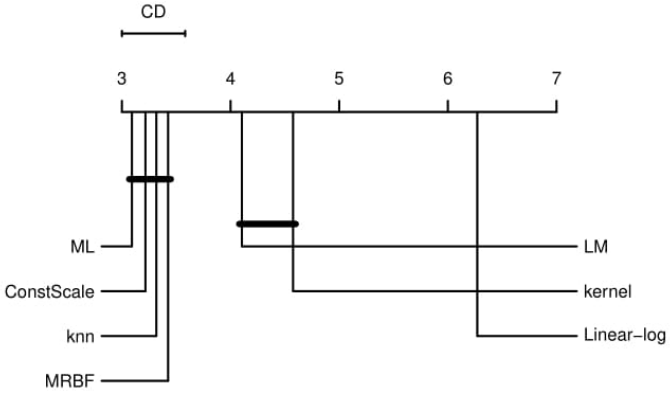
Proposals based on Imbalanced Domain Learning:

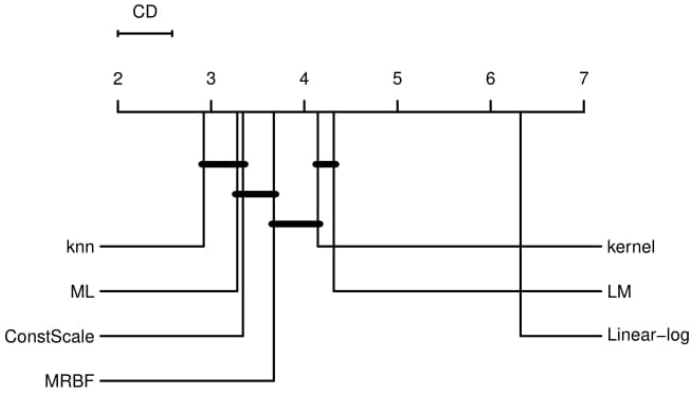
- Kernel-based Approach
- KNN-based Approach

Results – Predictive Analytics

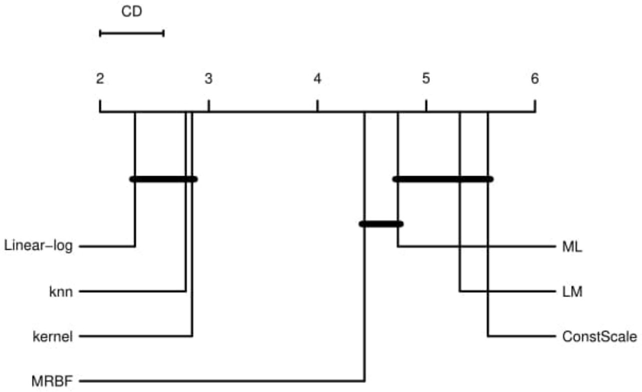
RMSE

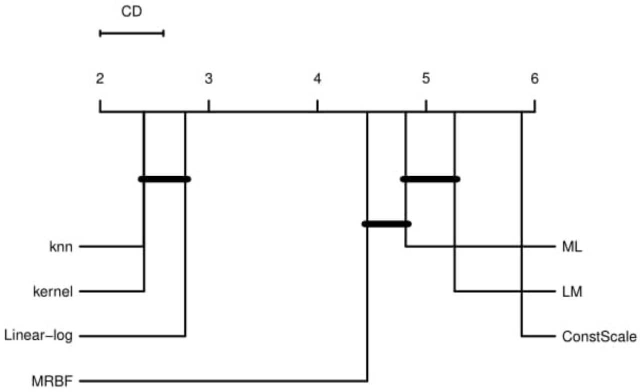


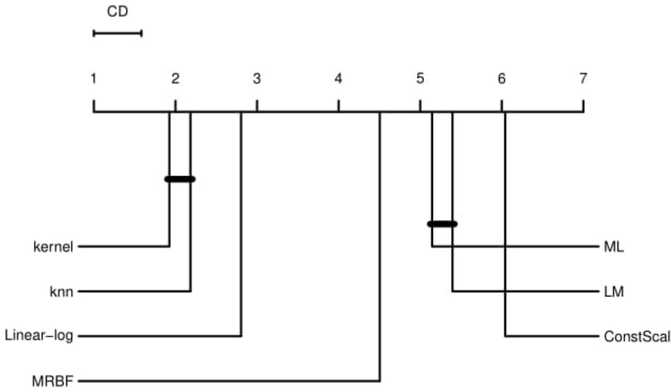




Utility-based F1-Score

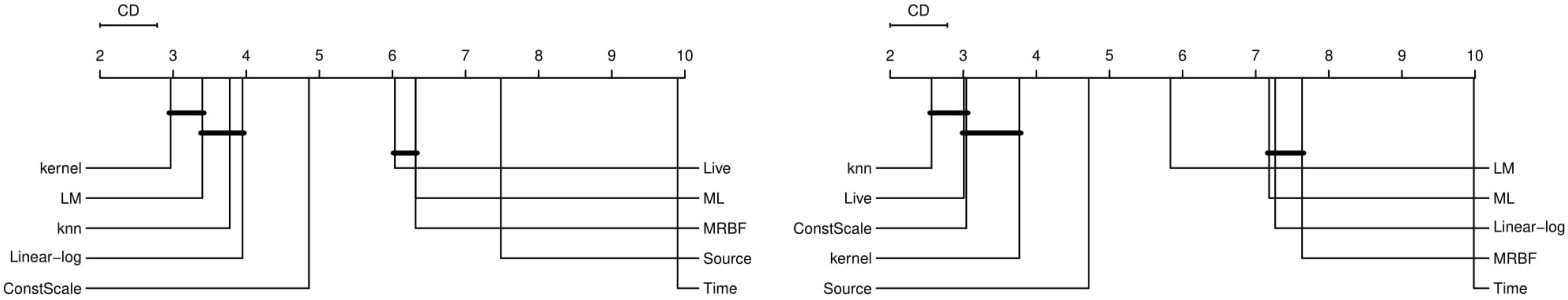






Results – Rankings

NDCG@10 (Google News & Yahoo! News)



Conclusions

- Previous proposals with best results in predicting average behaviour are within the worst in predicting the popularity of highly relevant cases;

- Simple local-based methods are capable of providing a better ability at predicting items with high levels of popularity;

- Utility-based formalization enables better results in ranking proposals concerning timely suggestions of highly popular content.

Thank you! Questions?

Nuno Moniz
nmmoniz@inesctec.pt

Luís Torgo
ltorgo@dal.ca

Code + Data + Presentation @ github:nunompmoniz