Benchmark for Ordinal Regression in pen & paper RPG game design

Jolanta Śliwa

Tabletop RPGs

- Takes place in a fantasy world
- A group of people play as heroes
- Gameplay based on Game Master narration

Monster

- Players fight together with monsters
- Encounters are important part of the game
- **Level** represents how hard it is to defeat the creature
- The fight should be challenging, but the victory has to be possible
- Monsters have a set of statistics, e.g. strength

Ordinal regression

Type of regression in which the goal is to predict a variable which is discrete and ordered.

Dependent variable: **LEVEL** (integer between -1 and 21)

Blue Dragon

AC 42; **HP** 370

Fort +32; **Ref** +30, **Will** +33

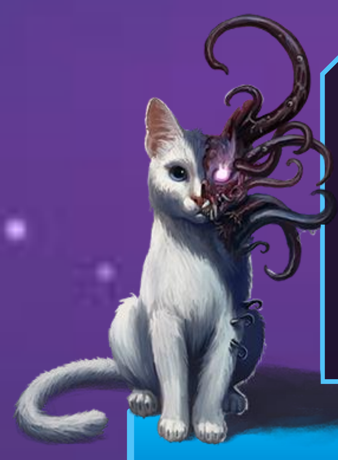
Perception: +31; darkvision, scent

Str +7, **Dex** +4, **Con** +6, **Int** +7, **Wis** +5, **Cha** +7

Melee - jaws +35 [+30/+25]

Damage 3d10+15 piercing + 2d12 electricity

Spells DC 43; *hallucinatory terrain* (lvl 4)...



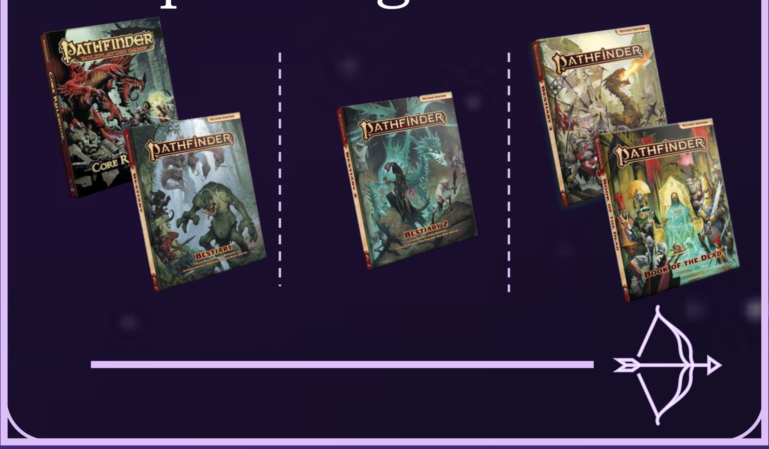
New monster design

1. Choose your monster's characteristics
2. But how to know the level?

Answer: Playtest, guessing or **AI/ML**

Split

Chronological split and expanding window



Metrics

Regression: RMSE, MAE, $RMSE^M$, MAE^M (macroaveraged)

Classification: Accuracy, Accuracy@1

OR: Somer's D

Models

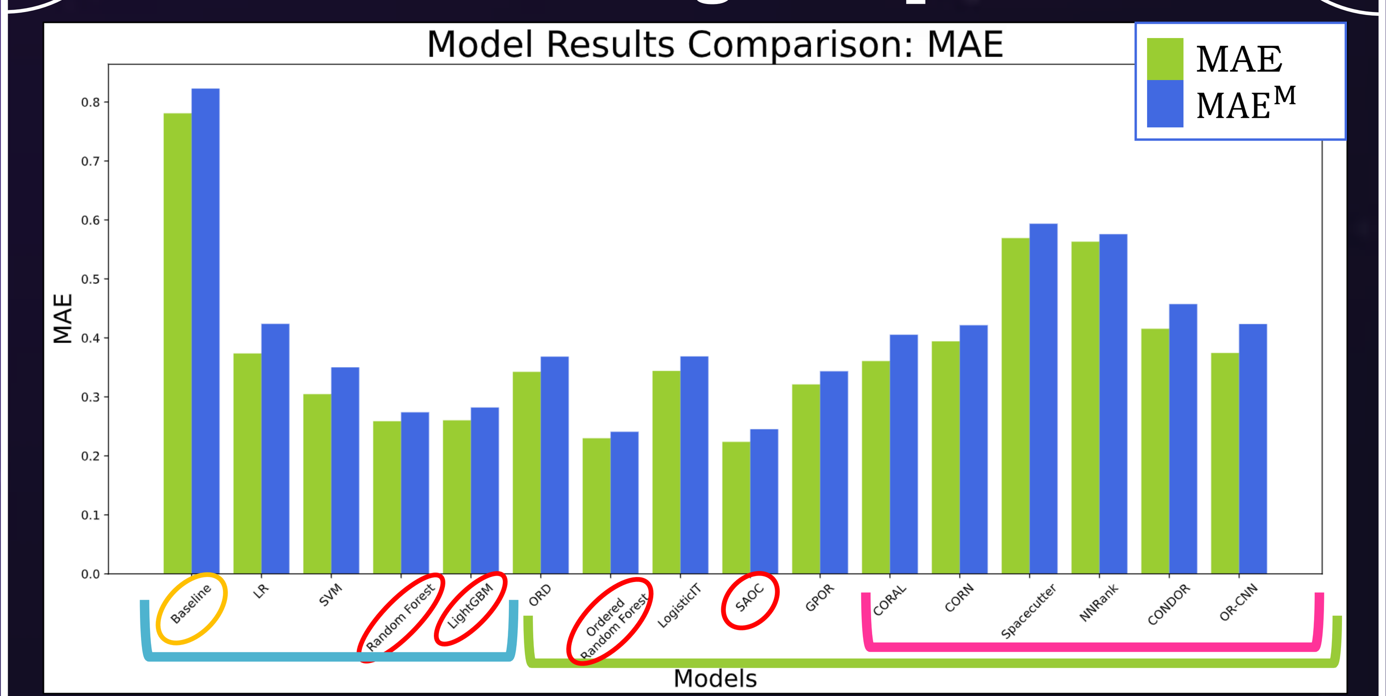
- Human inspired baseline
- Classical regression models (e.g. RF, LightGBM) with rounding
- Dedicated models based on RF, logistic regression and NN

Rounding

- Classic rounding (0.5)
- Single threshold tuning
- Threshold per level tuning:
  - TPE
  - Shortest path problem

Results

Chronological split



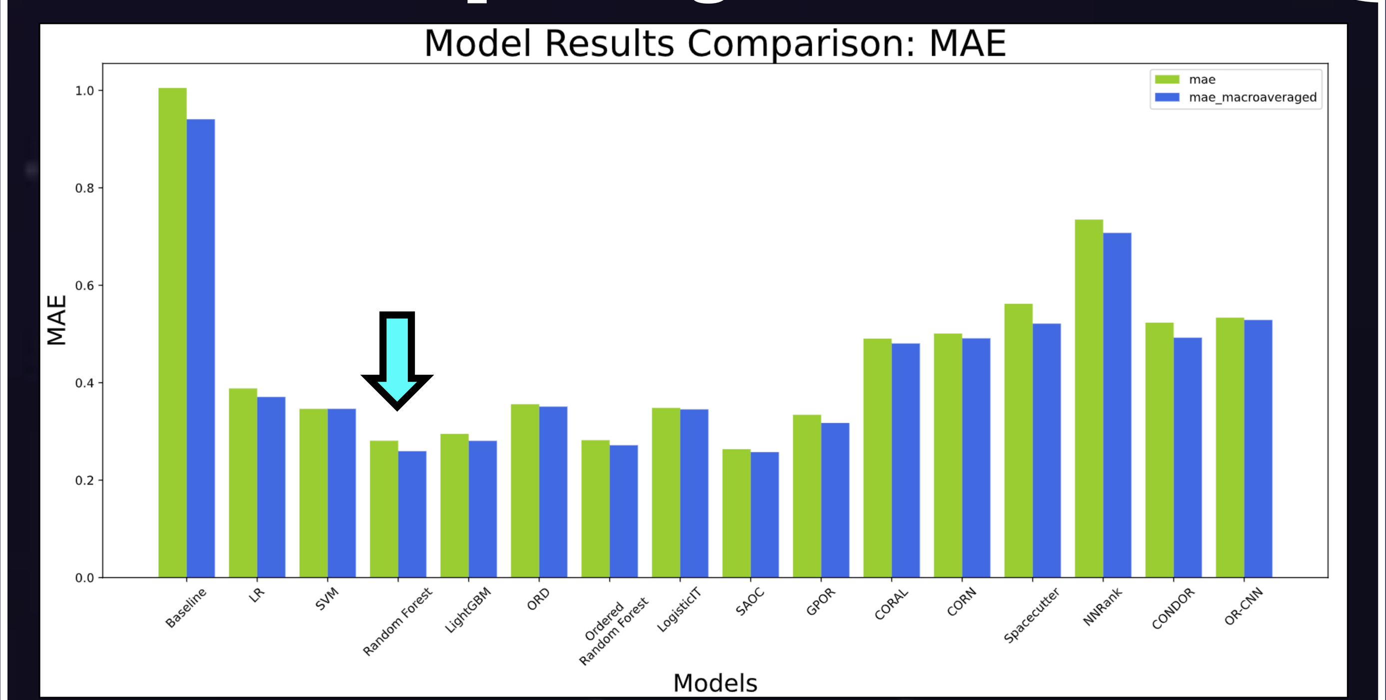
- Best results for all tree-based models
- All models outperform human-inspired baseline
- Ordinal models do not have better performance than regression + rounding
- Surprisingly bad results for NN-based methods
- As expected, results for macroaveraged metrics are worse (more realistic)

Accuracy (chronological)

Model	Accuracy	Accuracy@1
Baseline	46%	86%
LR	67%	97%
SVM	73%	98%
RF	77%	98%
LightGBM	77%	98%
ORD	70%	97%
ORF	81%	97%
LogisticIT	70%	97%
SAOC	81%	97%
GPOR	72%	97%
CORAL	68%	97%
CORN	65%	97%
Spaccutter	51%	93%
NNRank	50%	95%
CONDOR	64%	95%
OR-CNN	67%	97%

- All models work great with most errors around 1 level

Expanding window



- Opposite to what was expected results for macroaveraged metrics are better (lower)
- Why? -> see below

Rounding comparison

Model	Round 0.5	Global R_1	TPE R_1	Graph R_1	Global R_2	TPE R_2	Graph R_2
Baseline	0.861	0.861	0.853	0.823	0.861	0.864	0.823
LR	0.425	0.425	0.433	0.427	0.425	0.428	0.424
SVM	0.354	0.361	0.365	0.350	0.361	0.356	0.363
RF	0.277	0.295	0.298	0.274	0.295	0.289	0.276
LightGBM	0.282	0.492	0.366	0.483	0.329	0.295	0.324

- Tested on 2 sets: $R_1 = \{0.05, 0.1, \ldots, 0.95\}$ and $R_2 = \{0.25, 0.3, \ldots, 0.75\}$
- The best rounding strategy depends on the model

Playtests

Windows results
expanding window (RF)

Nr	MAE	MAE^M	RMSE	$RMSE^M$	Accuracy	Accuracy@1
1	0.56	0.55	1.11	1.04	59%	91%
2	0.49	0.22	0.77	0.50	57%	94%
3	0.25	0.28	0.50	0.53	75%	100%
4	0.32	0.35	0.60	0.63	69%	98%
5	0.15	0.10	0.38	0.31	85%	100%
6	0.28	0.34	0.68	0.81	77%	97%
7	0.39	0.33	0.66	0.66	63%	99%
8	0.16	0.19	0.42	0.46	85%	99%
9	0.16	0.15	0.40	0.39	84%	100%
10	0.25	0.24	0.54	0.53	76%	98%
11	0.21	0.22	0.50	0.53	81%	98%
12	0.29	0.29	0.64	0.62	75%	97%
13	0.13	0.10	0.38	0.34	88%	99%

- Generally more data == better
- Fluctuationsdue to data distribution shifts