Kaggle Seasons #06

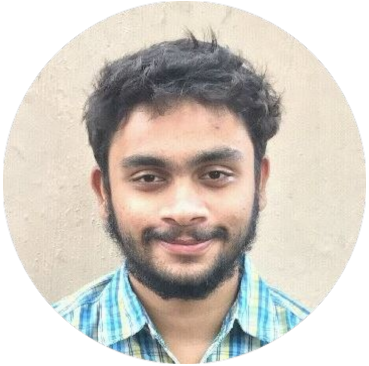
QUEEN MARY MACHINE
LEARNING SOCIETY

Meet the Crew

Karl

Max

Vish



What is Kaggle?

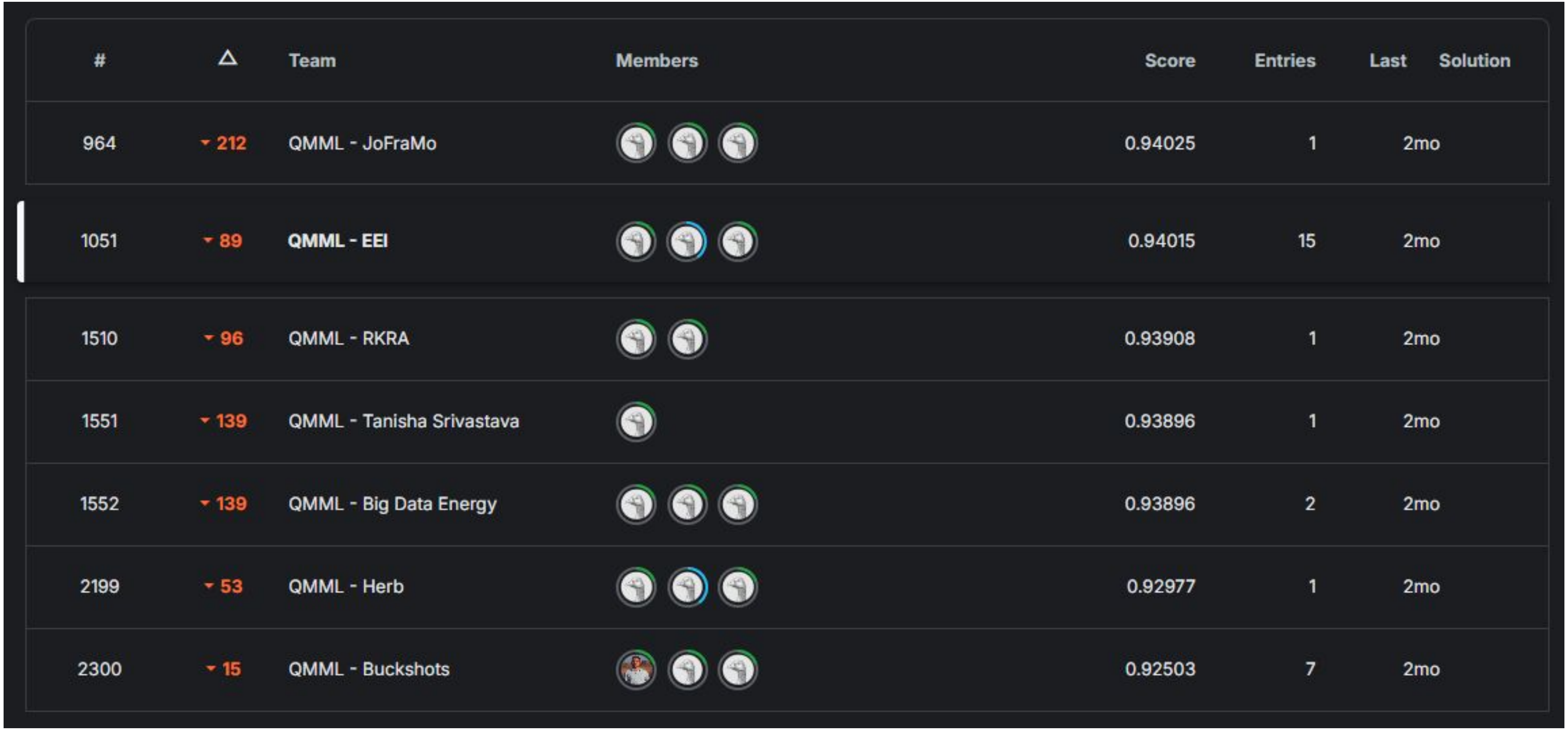
- *The* platform for everything AI/ML/Data Science
- Go-to platform for ML/Data Science competitions
- Teams competing from all around the world
- Boosts Employability (amazing for CV!)
- Opportunity to:
  - Meet talented people
  - Discover cutting edge methods
  - Develop real-world skills from real-life industry projects

kaggle

Last Semester Review

- Participated in 1 external hackathon at Google
- Organized our own Christmas Hackathon with 5+ teams competing
- Held several social events
- Competed in several monthly Kaggle competitions (10 teams)
    - November competition with 7 teams (shown next slide)

Last Semester Review



#	△	Team	Members	Score	Entries	Last	Solution
964	▾ 212	QMML - JoFraMo		0.94025	1	2mo	
1051	▾ 89	**QMML - EEI**		0.94015	15	2mo	
1510	▾ 96	QMML - RKRA		0.93908	1	2mo	
1551	▾ 139	QMML - Tanisha Srivastava		0.93896	1	2mo	
1552	▾ 139	QMML - Big Data Energy		0.93896	2	2mo	
2199	▾ 53	QMML - Herb		0.92977	1	2mo	
2300	▾ 15	QMML - Buckshots		0.92503	7	2mo	

Plans for this semester

- Lectures and interactive sessions with Kaggle grandmasters
- More external hackathons for all skill ranges
- Industry-focused competitions
- More social events
- More collaborations with societies and companies

What now?

1. Fill out the survey form (shown next slide)
2. Go to your group (survey based results)
3. Talk to everyone in the group
4. Break into teams of 3-4

5 Minutes for Sign-up!

SLACK

FORM

Intro to Gradient Boosting

QUEEN MARY MACHINE
LEARNING SOCIETY

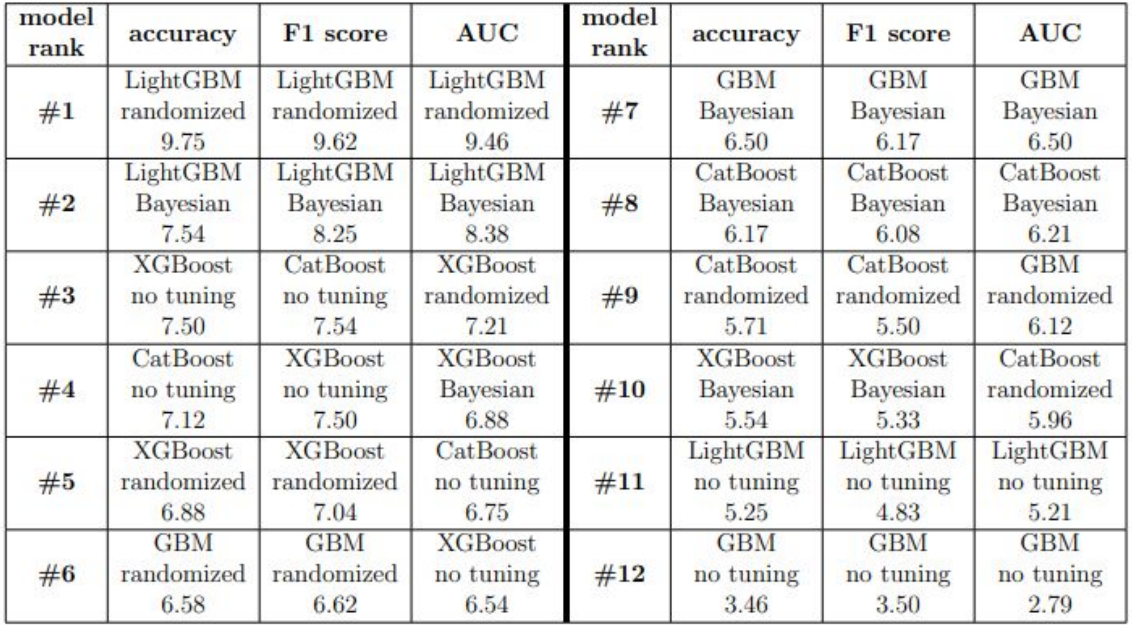
What is gradient boosting? (Continued)

- Gradient boosting is a form of generalised form of boosting
- Boosting is an ensemble learning approach which starts from weak models and iteratively improves them with the help of other weak models
- Gradient boosting achieves iterative improvement by training new models on the residuals of previous models
- Instead of fitting models directly on residuals, which can often be noisy, gradient boosting instead fits on more noise-resistant, gradient-based estimators of residuals, hence the name (more details next week!)

What is gradient boosting? (Continued)

- Gradient boosting works most effectively with decision trees as the initial weak learners, as they are less prone to overfitting
- Therefore, almost all state-of-the-art gradient boosting algorithms are based on decision trees
- Notable examples are CatBoost, XGBoost, and LightGBM (again, more details next week!)
- Gradient boosting generally outperforms traditional supervised learning algorithms such as linear/logistic regression, random forest, etc on structured data

What is gradient boosting? (Continued)



model rank	accuracy	F1 score	AUC	model rank	accuracy	F1 score	AUC
#1	LightGBM randomized 9.75	LightGBM randomized 9.62	LightGBM randomized 9.46	#7	GBM Bayesian 6.50	GBM Bayesian 6.17	GBM Bayesian 6.50
#2	LightGBM Bayesian 7.54	LightGBM Bayesian 8.25	LightGBM Bayesian 8.38	#8	CatBoost Bayesian 6.17	CatBoost Bayesian 6.08	CatBoost Bayesian 6.21
#3	XGBoost no tuning 7.50	CatBoost no tuning 7.54	XGBoost randomized 7.21	#9	CatBoost randomized 5.71	CatBoost randomized 5.50	GBM randomized 6.12
#4	CatBoost no tuning 7.12	XGBoost no tuning 7.50	XGBoost Bayesian 6.88	#10	XGBoost Bayesian 5.54	XGBoost Bayesian 5.33	CatBoost randomized 5.96
#5	XGBoost randomized 6.88	XGBoost randomized 7.04	CatBoost no tuning 6.75	#11	LightGBM no tuning 5.25	LightGBM no tuning 4.83	LightGBM no tuning 5.21
#6	GBM randomized 6.58	GBM randomized 6.62	XGBoost no tuning 6.54	#12	GBM no tuning 3.46	GBM no tuning 3.50	GBM no tuning 2.79

Table 6: Final rankings of 12 models for accuracy, F1 score and AUC

[1] Florek, P. & Zagdański, A. Benchmarking state-of-the-art gradientboosting algorithms for classification. Preprint at https://doi.org/10.48550/arXiv.2305.17094 (2023)

How important is gradient boosting in ML?

kaggle

How important is gradient boosting in ML?

VERY. Next question.

kaggle

How important is gradient boosting in ML?

- Gradient boosting is the current state of the art for supervised learning on structured data
- The vast majority of Kaggle competition winners and top performance utilise gradient boosting somewhere in the loop
- However, gradient boosting is less optimal for unstructured data, which requires more adaptive approaches such as NNs
- Gradient boosting also isn’t as effective on high-dimensional datasets due its lack of parallelisation and computational cost

Thanks for listening!

QUEEN MARY MACHINE
LEARNING SOCIETY