IASD M2 at Paris Dauphine

Become a Kaggle Master

1: Introduction and Course overview

Eric Benhamou

Agenda

Part I: general concepts

1. Introduction to Kaggle (concept and API)
2. Competition, metrics
3. Validation
4. Hyper parameters tuning
5. Model ensemble with blending and stacking

Part II: Competitions

5. Predict Financial markets
6. Analyze News
7. Design your portfolio

Grading

- Two training Kaggle: 20% x 2 = 40% + Final project 60%

- For each Kaggle:
  - public presentation
  - Slide
  - Google Collab notebook

- Score:
  - 80% kaggle score (baseline=10, best 20)
  - 20% presentation

Calendar

1. Tues Jan 24: Lecture 1
2. Thur Jan 26: Lecture 2 + Intro Competition 1
3. Tues Jan 31: Lecture 3
4. Thur Feb 02: Competition 1 Presentation
5. Tues Mar 7 : Lecture 4
6. Thur Mar 9: Competition 2 Presentation
7. Mon Mar 13: Lecture 5
8. Fri Mar 17: Competition 3 Presentation

Origin

- Data science competitions have long been around and they have experienced growing success over time, starting from a niche community of passionate competitors, drawing more and more attention, and reaching a much larger audience

- Starting in 1970s, ICPC, the International Collegiate Programming Contest coded in Fortran
- In 2000s, KDD Cup (first competition took place in 1007)
- Netflix competition from 2006 to 2009 with a 10% improvement

kaggle

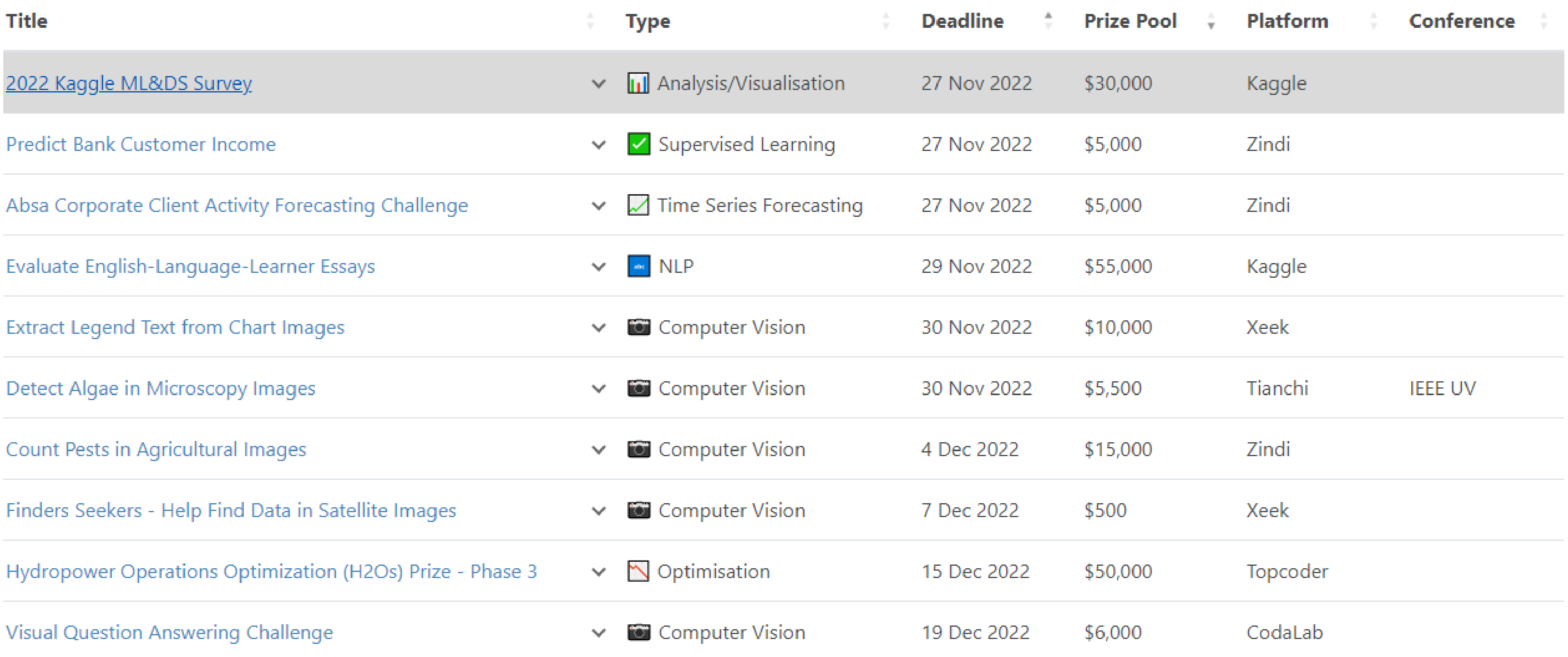
- Kaggle started in Feb 2010, with Anthony Goldbloom
  - Geoffrey Hinton, the “godfather” of deep learning, won a Kaggle competition hosted by Merck in 2012
  - *Tianqi Chen* launched XGBoost for the *Higgs Boson Machine Learning Challenge*
  - Jeremy Howard won some competition became Kaggle CTO and then created www.fast.ai
  - *Jeremy Achin* and *Thomas de Godoy* created DataRobot which develops AutoML
  - François Chollet created Keras to be able to have

Other competitions platforms

- **US:**
  - **DrivenData** https://www.drivendata.org/competitions/
  - **Numerai** https://numer.ai/
  - **CrowdANALYTIX** https://www.crowdanalytix.com/
- **France**: https://challengedata.ens.fr/ and https://codalab.lisn.upsaclay.fr/
- **China**: https://tianchi.aliyun.com/competition/
- **India**: https://datahack.analyticsvidhya.com/
- **Japan**: Signate : https://signate.jp/competitions/
- **Africa**: Zindi https://zindi.africa/competitions
- **Switzerland**: https://www.aicrowd.com/
- Russia: https://ods.ai/competitions

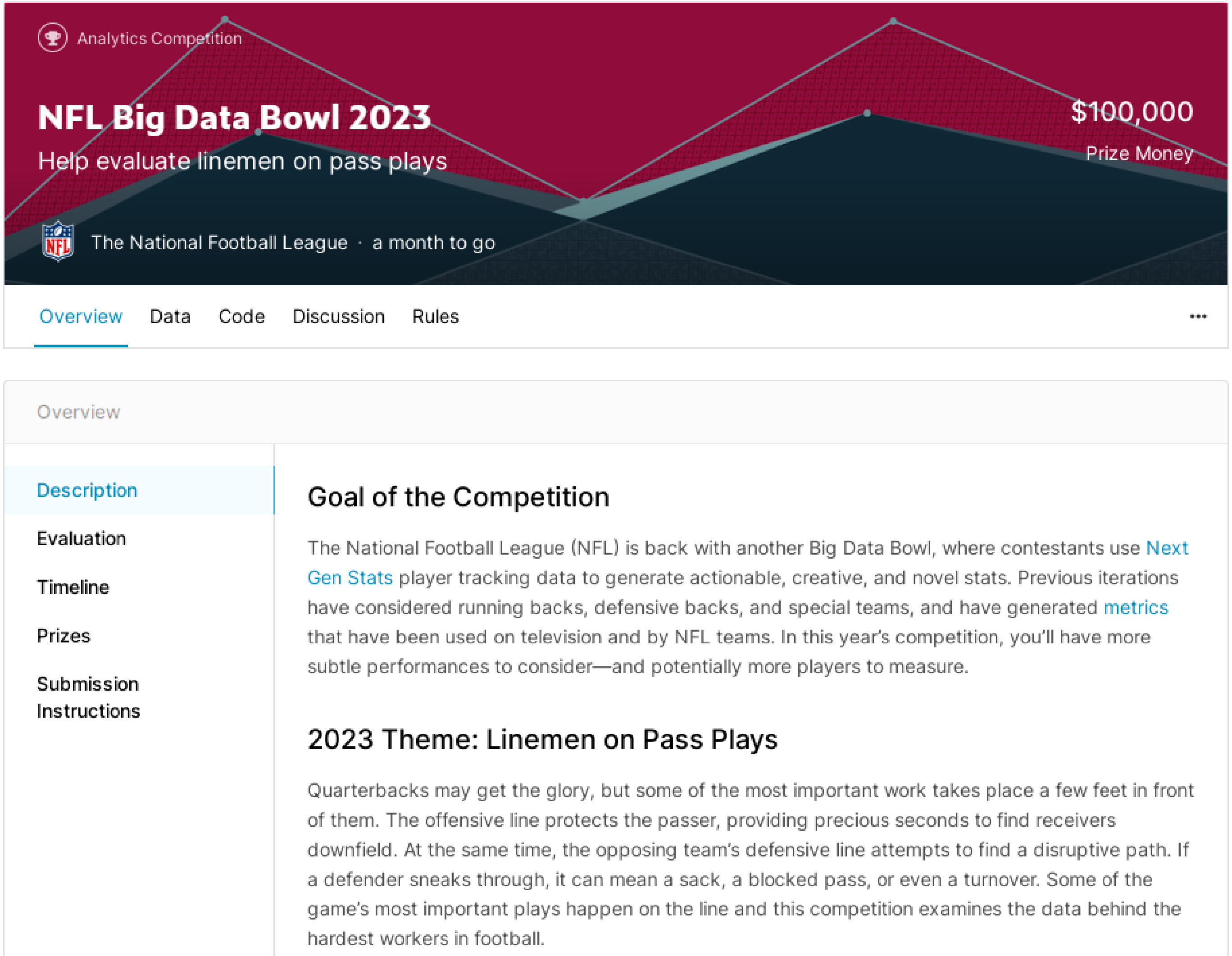
Dashbord of ML Competitions

- https://mlcontests.com/



Title	Type	Deadline	Prize Pool	Platform	Conference
2022 Kaggle ML&DS Survey	📊 Analysis/Visualisation	27 Nov 2022	$30,000	Kaggle	
Predict Bank Customer Income	✅ Supervised Learning	27 Nov 2022	$5,000	Zindi	
Absa Corporate Client Activity Forecasting Challenge	📈 Time Series Forecasting	27 Nov 2022	$5,000	Zindi	
Evaluate English-Language-Learner Essays	🔤 NLP	29 Nov 2022	$55,000	Kaggle	
Extract Legend Text from Chart Images	📷 Computer Vision	30 Nov 2022	$10,000	Xeek	
Detect Algae in Microscopy Images	📷 Computer Vision	30 Nov 2022	$5,500	Tianchi	IEEE UV
Count Pests in Agricultural Images	📷 Computer Vision	4 Dec 2022	$15,000	Zindi	
Finders Seekers - Help Find Data in Satellite Images	📷 Computer Vision	7 Dec 2022	$500	Xeek	
Hydropower Operations Optimization (H2Os) Prize - Phase 3	📉 Optimisation	15 Dec 2022	$50,000	Topcoder	
Visual Question Answering Challenge	📷 Computer Vision	19 Dec 2022	$6,000	CodaLab	

How a Kaggle competition works?



Goal of the Competition

The National Football League (NFL) is back with another Big Data Bowl, where contestants use Next Gen Stats player tracking data to generate actionable, creative, and novel stats. Previous iterations have considered running backs, defensive backs, and special teams, and have generated metrics that have been used on television and by NFL teams. In this year's competition, you'll have more subtle performances to consider—and potentially more players to measure.

2023 Theme: Linemen on Pass Plays

Quarterbacks may get the glory, but some of the most important work takes place a few feet in front of them. The offensive line protects the passer, providing precious seconds to find receivers downfield. At the same time, the opposing team's defensive line attempts to find a disruptive path. If a defender sneaks through, it can mean a sack, a blocked pass, or even a turnover. Some of the game's most important plays happen on the line and this competition examines the data behind the hardest workers in football.

Timeline and submissions

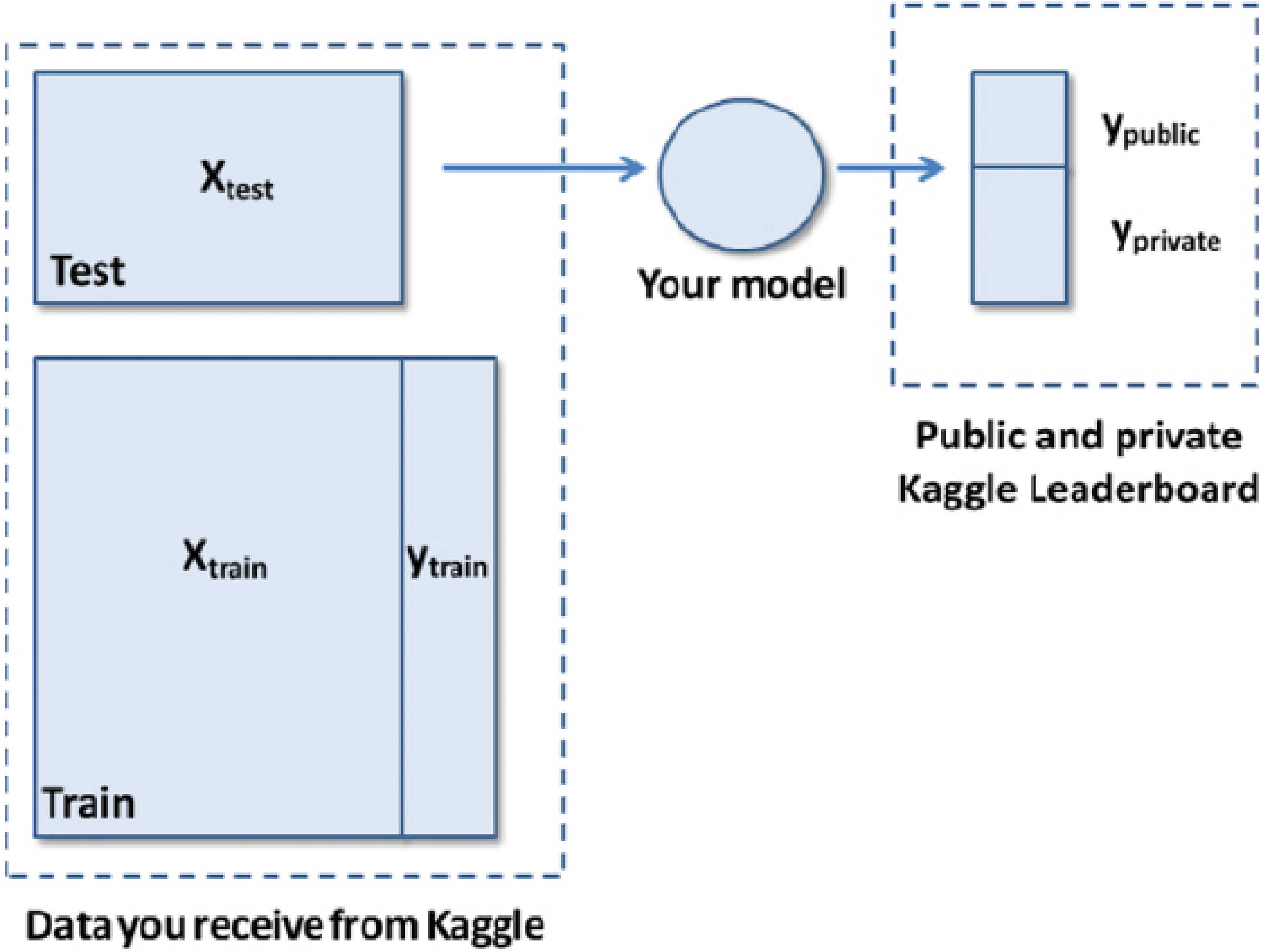
- **Deadline**:
  - Start
  - End
  - Team merger deadline

- **Rules**:
  - eligibility for a prize?
  - external data ?
  - number of submissions per day?
  - number of final solutions?

Kaggle API

- Allow to download data and use on your computer or cloud
- https://www.kaggle.com/docs/api

Data


X_{test}
Test
X_{train}
y_{train}
Train
Data you receive from Kaggle
Your model
y_{public}
$y_{private}$
Public and private
Kaggle Leaderboard

Types of competitions and examples

- Featured
- Masters (private and invite-only competitions=
- Annuals
- Research (*Facebook Recruiting Competition, Google Landmark Recognition, etc..)*
- Recruitment
- Getting Started (*Digit Recognizer, Titanic, House Prices)*
- Playground
- Analytics
- Community

What can go wrong in competition?

- Leakage from the data
- Probing from the leaderboard (the scoring system)
- Overfitting and consequent leaderboard shake-up
- Private sharing

Improving results

- The system works the best if the **task is well defined** and **the data is of good quality**. In the long run, the performance of solutions improves by small gains until **it reaches an asymptote**.
- The process can be sped up by **allowing a certain amount of sharing among participants** (as happens on Kaggle by means of discussions, and sharing Kaggle
- Notebooks and extra data provided by the datasets found in the Datasets section).

- **Competitive pressure** in a competition suffices to **produce always-improving solutions**. When the competitive pressure is paired with some degree of sharing among participants, the improvement happens at an even faster rate – hence why Kaggle introduced many incentives for sharing.

Computational resources

- Kaggle notebook
- https://kaggledays.com/

Rules

- https://www.kaggle.com/docs/notebooks
- 12 hours execution time for CPU/GPU, 9 hours for TPU
- 20 gigabytes of auto-saved disk space (/kaggle/working)

Resources

- CPU specifications:
  - 4 CPU cores
  - 16 gigabytes of RAM
- GPU specifications:
  - 2 CPU cores
  - 13 gigabytes of RAM
- TPU specifications
  - 4 CPU cores
  - 16 gigabytes of RAM