NTNU | Norwegian University of Science and Technology

Project Presentation

TDT4173 Machine Learning
Fall 2022

Ruslan Khalitov
Sverre Herland

Outline

- Problem overview
- Plaace introduction and data set description
- Evaluation
- Virtual teams
- What should be delivered?
- Grading
- Computing resources

Purpose

Learn how to solve real-world problems using machine learning models.

Learning goals:

- Data analysis
- Feature engineering
- Validation
- Model selection
- Hyperparameters tuning
- Solving ML competitions

The task

- Accurately solve a prediction problem
- Real-world data
- You will compete with Virtual Teams (VTs) we've created
- Better score → More VTs defeated → higher grade you receive

Kaggle

- **The competition is hosted on Kaggle InClass**
- **Kaggle is the largest platform for data science competitions**
- **Why Kaggle?**
  - You can check solutions **automatically**
  - Allows private InClass competitions for students
  - User-friendly interface
  - World-known platform

Kaggle

- **Invitation Link**

https://www.kaggle.com/t/3affe88e40c44dde87d1ff836ded9e92


plaace
NTNU
Project in TDT4173
Fall 2022

plaace

Plaace delivers a data-driven platform to match retail tenants and retail properties.

By using external data sources and analytics, the platform allows tenants to establish themselves in the best possible locations, and property owners can choose the concepts that are best suited to their property.

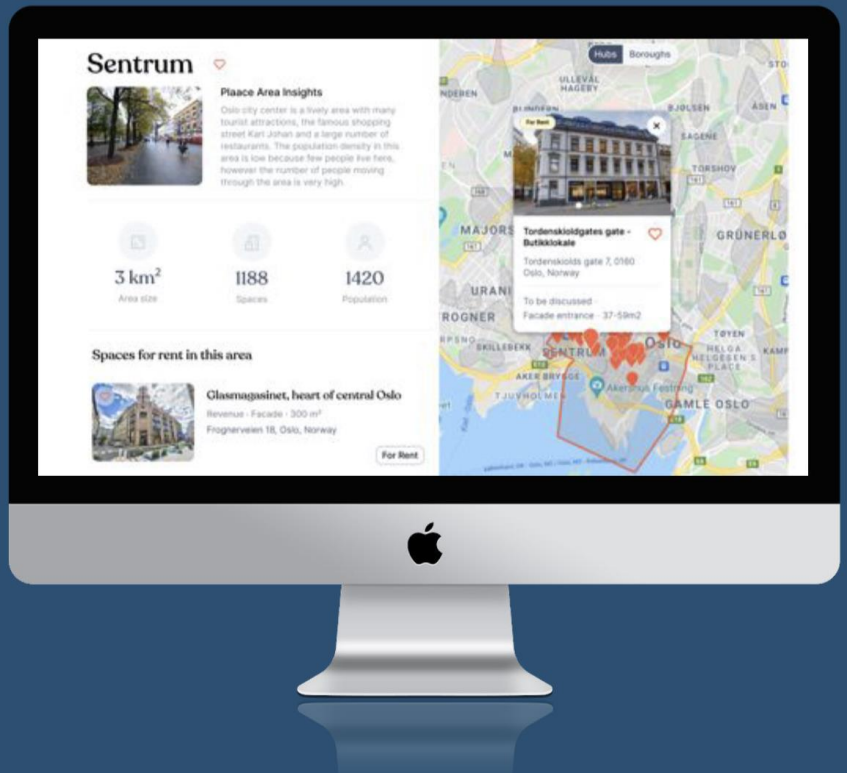
Meet Plaace - A platform for retail- and property owners

plaace

Matching and collaboration

- See available and off-market spaces for retail & hospitality
- Showcase your business profile
- Create lists, score and collaborate with stakeholders

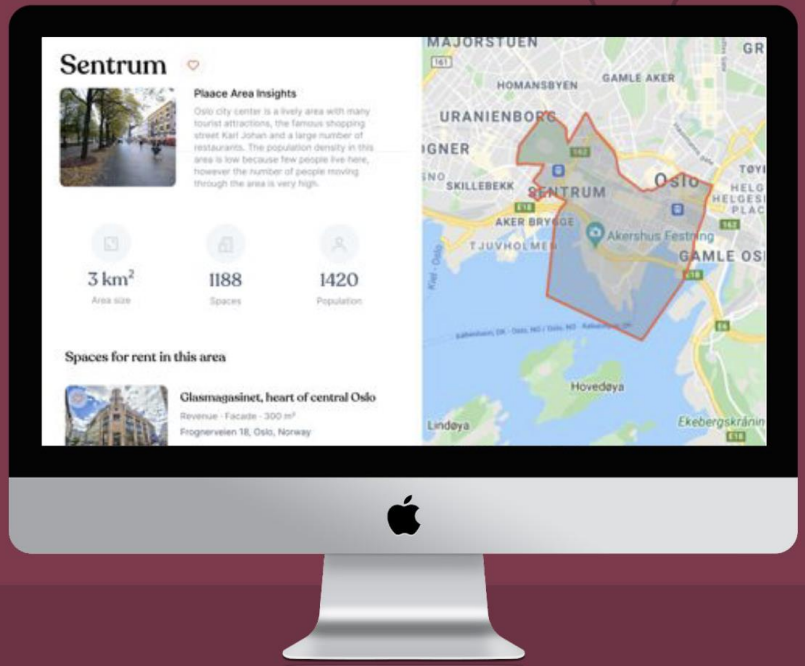
- Advertise a portfolio of retail space both on and off market
- Find and assess relevant tenants fitting your criteria
- Receive and handle rental requests



Meet Plaace - A platform for retail- and property owners

plaace

Location Intelligence



- Movement
- Demography
- Competitors
- Spend

Telia vipps Statistisk sentralbyrå Statistics Norway Prognosesenteret

Meet Plaace - A platform for retail- and property owners

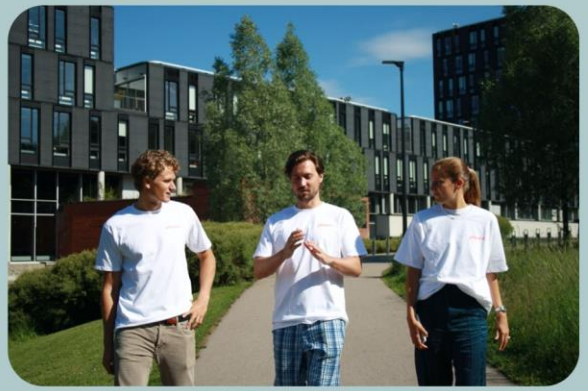
plaace

Bringing it together with ML driven recommendations

Opportunities in Plaace

plaace

- Specialization project and master's thesis
- Internship
- Part time work
- Full time position



Task and data details

Task: Predict revenue of all types of retail stores in Norway.

The challenge requires extensive analysis of the data foundation, creation of a wide variety of informative input features, and development of supervised regression ML models.

Overview of data sets

plaace

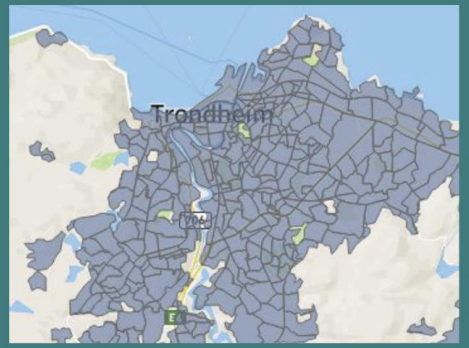
1. Spatial data
   - Official geographical units in Norway

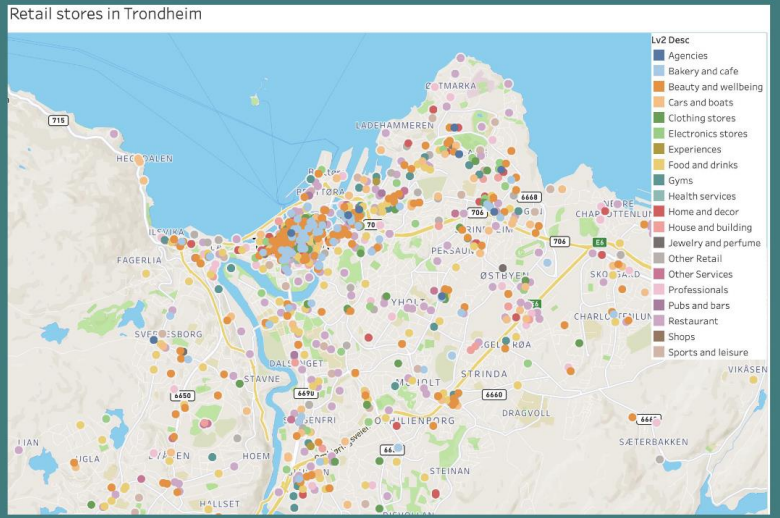
2. Demography data:
   - Age distributions
   - Household compositions
   - Income households

3. Mobility data:
   - Public transportation stops

4. Retail company data:
   - Exact location, revenue, #employees, etc.
   - Retail hierarchy

(5. Additional data?)
   - ...





1. Spatial data

plaace

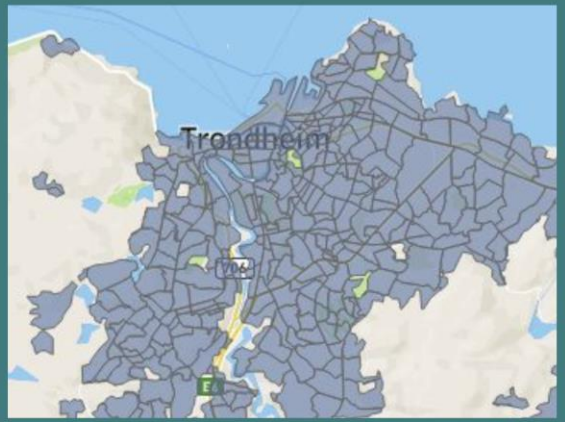
Statistisk sentralbyrå
Statistics Norway

Foundation for mapping other data, density measures, etc.

Grunnkrets:

grunnkrets_ID	year	grunnkrets_name	district_name	municipality_name	geometry	area_km2

- Smallest official geographical boundary
- Approx. 13 000 entries after negligible terrain has been removed



2. Demography data

plaace

Statistisk sentralbyrå
Statistics Norway

Data describing the population in the different areas

Age distribution:

grunnkrets_ID	year	age_0	age_1	...	age_90

- Number of people of different ages for every *grunnkrets* in Norway

Household composition:

grunnkrets_ID	year	single_parent_children_0_5	...	couple_children_18_or_above

- Number of people within 8 household categories for every *grunnkrets*

Income households:

grunnkrets_ID*	year	singles	...	couple_without_children

- Median income for different household types

* data is originally on the level of *districts*, and is mapped directly to the underlying grunnkretses

3. Mobility data

plaace

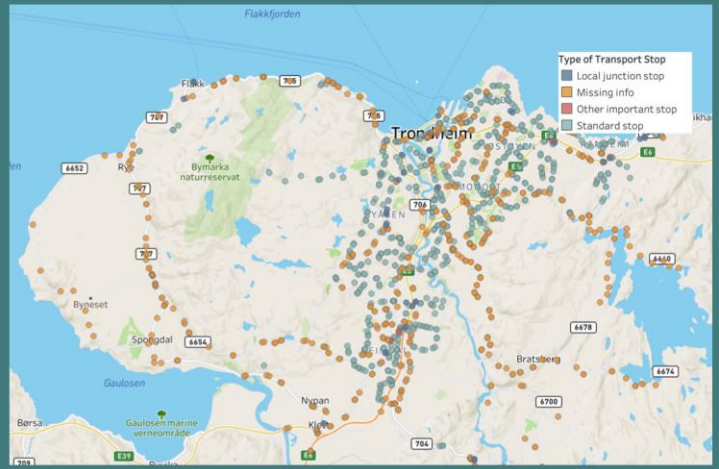
Static proxy for movement

Public transportation stops:

busstop_id	importance_lv	lat/lon

Statens vegvesen

- All 68k bus stops in Norway with exact location
- 6 importance levels, indication of volume



4. Retail company data

plaace

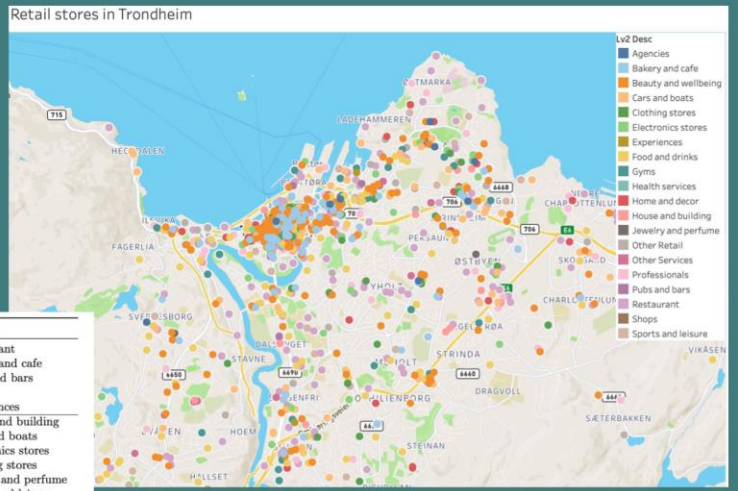
Company specific data for all retailers in Norway

Prognosesenteret

Retailers:

store_id	year	lat/lon	store_name	grunnkrets_id	hierarchy_id	chain	mall	**revenue**	+++

- Data describing existing retail stores
- The revenue variable is the target for predictions
- Approx 21 000 stores with revenue, and 27 000 without revenue per year



Plaace hierarchy:

lv1	lv1_desc	...	lv4	lv4_desc

- Retail hierarchy created by Plaace
- 109 distinct at lv4

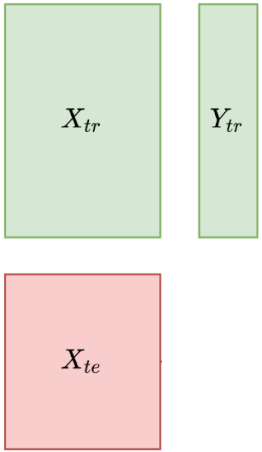
lv1	lv1_desc	lv2	lv2_desc
1	Dining and Experiences	1.1	Restaurant
1	Dining and Experiences	1.2	Bakery and cafe
1	Dining and Experiences	1.3	Pubs and bars
1	Dining and Experiences	1.4	Shops
1	Dining and Experiences	1.5	Experiences
2	Retail	2.1	House and building
2	Retail	2.2	Cars and boats
2	Retail	2.3	Electronics stores
2	Retail	2.4	Clothing stores
2	Retail	2.5	Jewelry and perfume
2	Retail	2.6	Sports and leisure
2	Retail	2.7	Food and drinks
2	Retail	2.8	Other Retail
2	Retail	2.9	Home and decor
3	Services	3.1	Health services
3	Services	3.2	Beauty and wellbeing
3	Services	3.3	Professionals
3	Services	3.4	Other Services
3	Services	3.5	Agencies
3	Services	3.6	Gyms

Evaluation metric

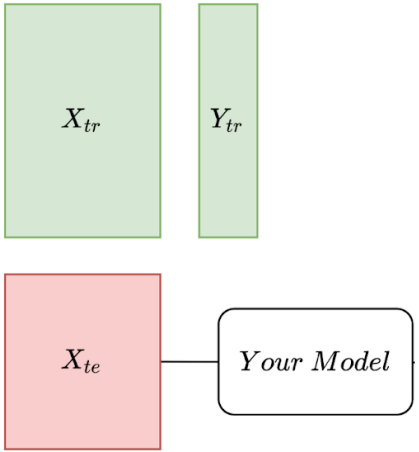
$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(x_i) - \log(y_i))^2}$$

RMSLE

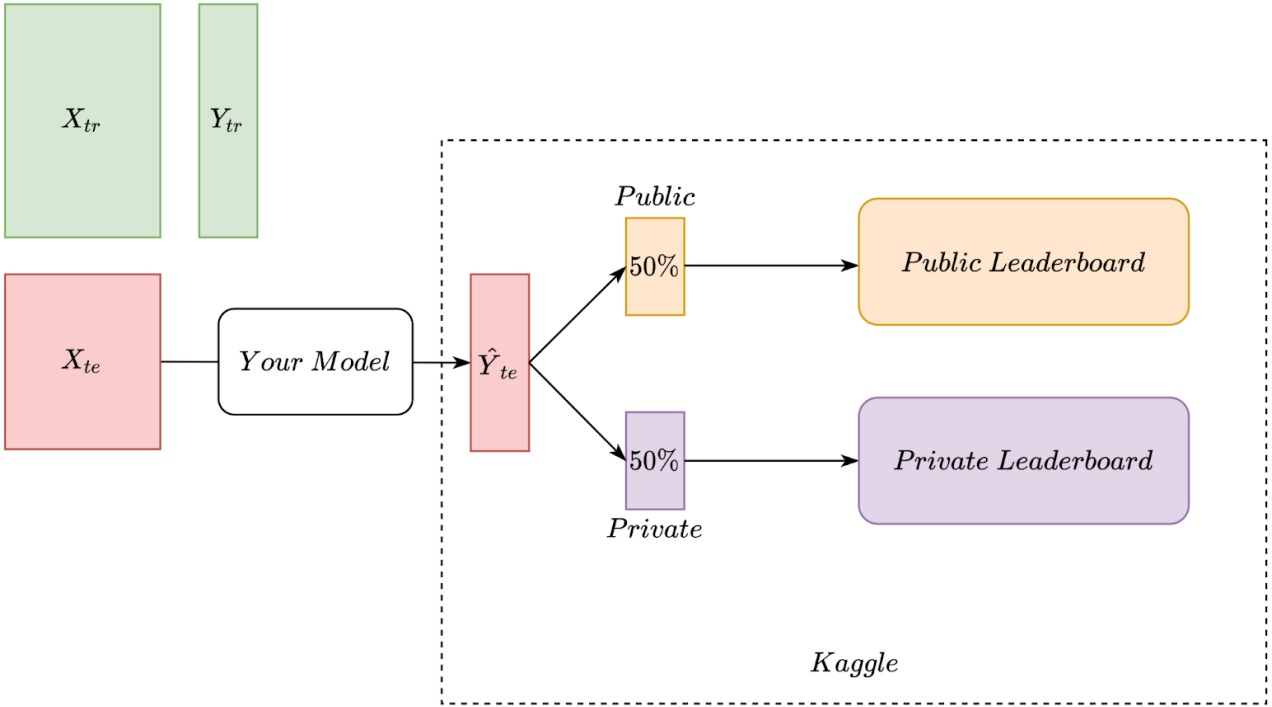
Evaluation. Public/Private



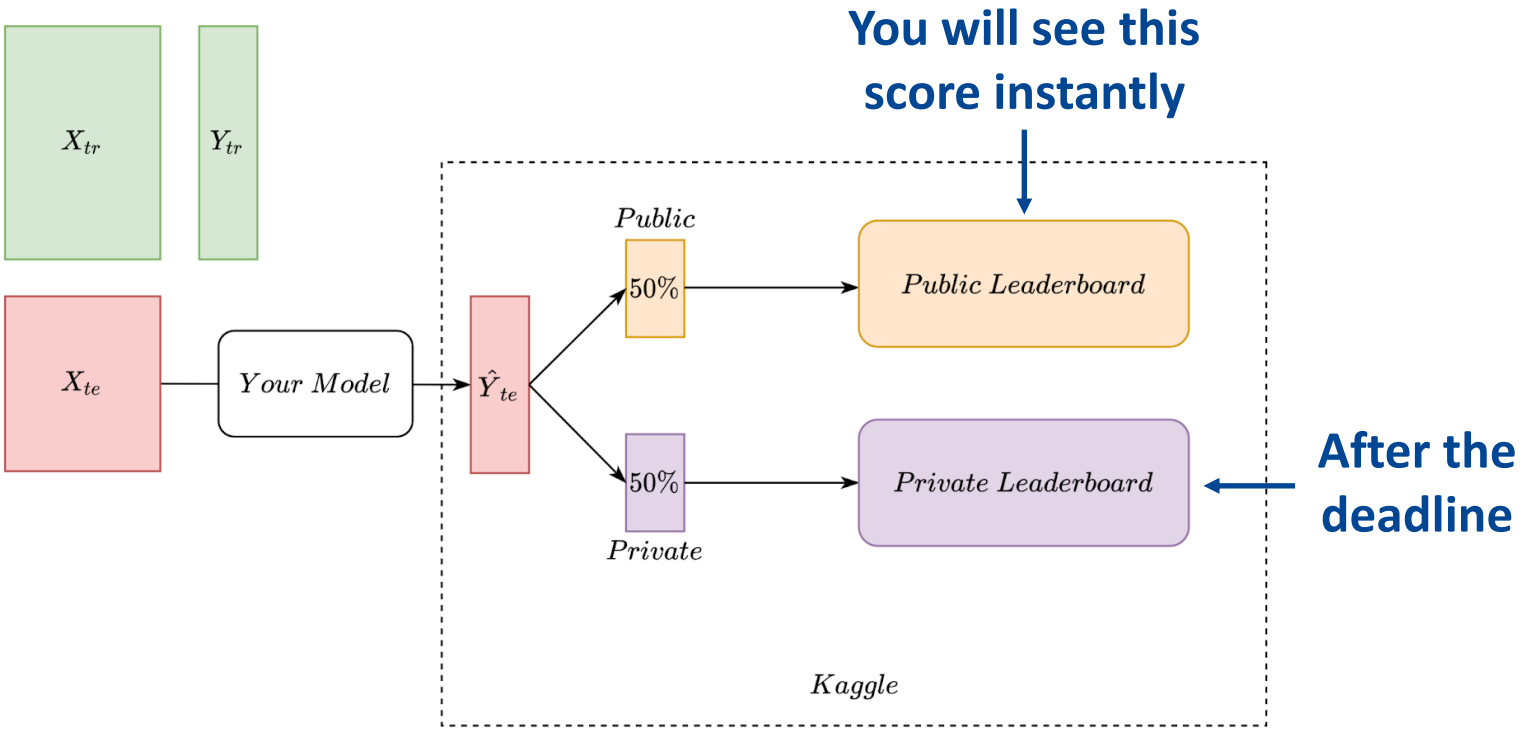
Evaluation. Public/Private



Evaluation. Public/Private



Evaluation. Public/Private

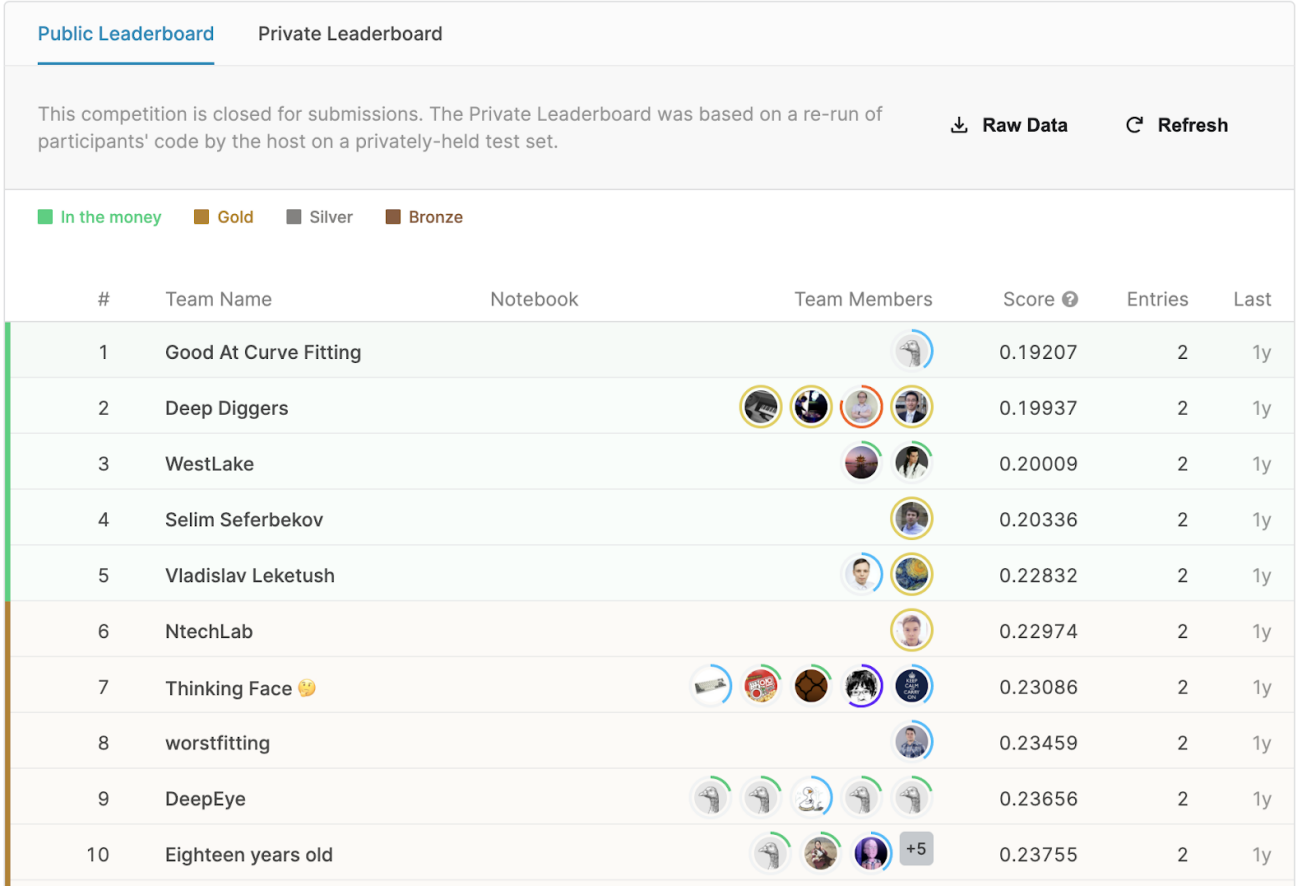


Evaluation. Public/Private

- You don’t know the exact Public/Private split
- You can see your **estimated score** on the test set instantly
- You can submit **max 5 times** per day
- **The final score of your solution is your Private score**
- Before the competition end you have to select 2 solutions you think are best:
  - Tip: select different models, so at least one of them will score high

Evaluation. Public/Private

- You have extra information from the public leaderboard (50%)
- Mind the distribution difference
- **The best public score does not necessarily correspond to the best private score**
- **Do not overfit on the public leaderboard score**



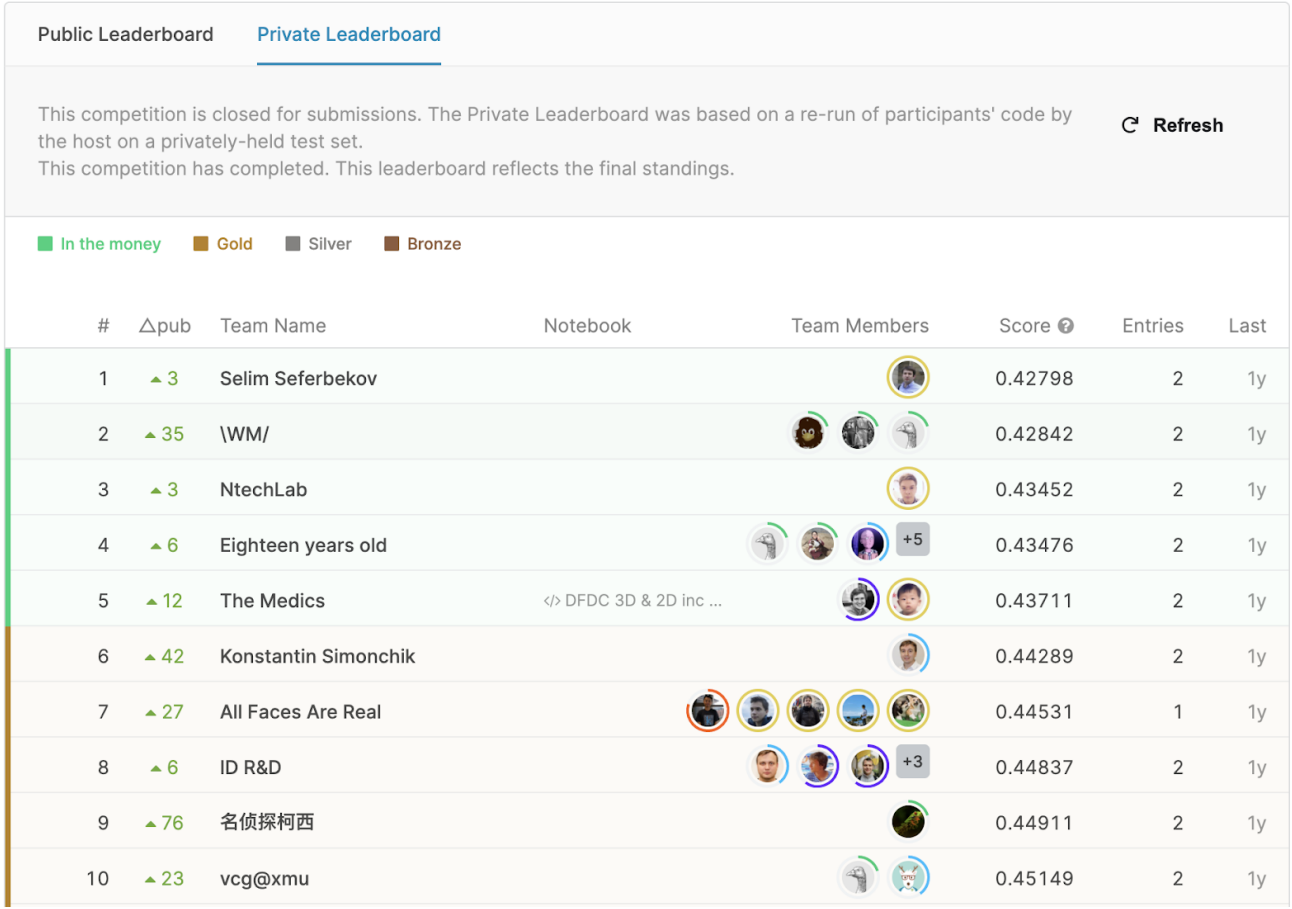
Public Leaderboard | Private Leaderboard

This competition is closed for submissions. The Private Leaderboard was based on a re-run of participants' code by the host on a privately-held test set.

Raw Data | Refresh

In the money | Gold | Silver | Bronze

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Good At Curve Fitting			0.19207	2	1y
2	Deep Diggers			0.19937	2	1y
3	WestLake			0.20009	2	1y
4	Selim Seferbekov			0.20336	2	1y
5	Vladislav Leketush			0.22832	2	1y
6	NtechLab			0.22974	2	1y
7	Thinking Face 🤔			0.23086	2	1y
8	worstfitting			0.23459	2	1y
9	DeepEye			0.23656	2	1y
10	Eighteen years old		+5	0.23755	2	1y



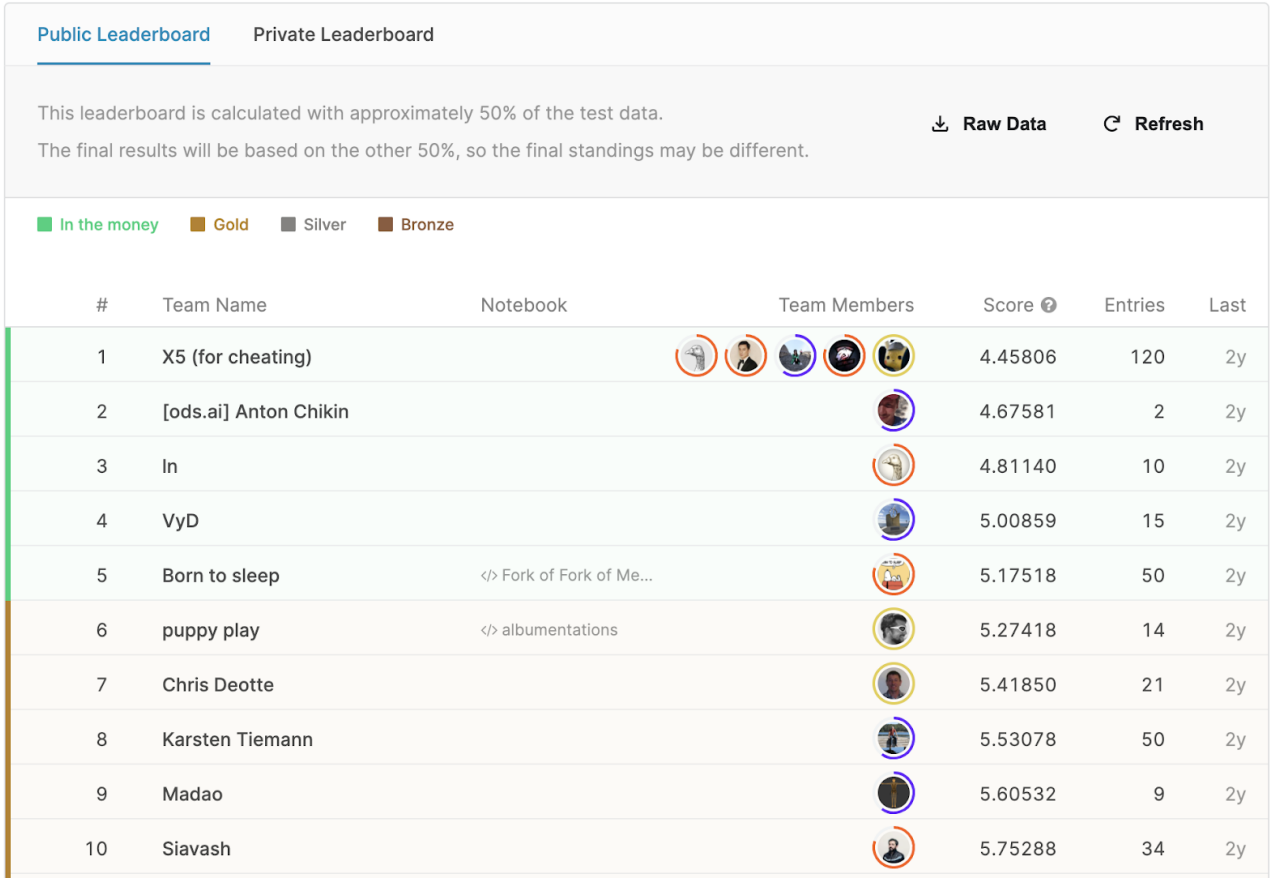
Public Leaderboard | Private Leaderboard

This competition is closed for submissions. The Private Leaderboard was based on a re-run of participants' code by the host on a privately-held test set.
This competition has completed. This leaderboard reflects the final standings.

Refresh

In the money | Gold | Silver | Bronze

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲ 3	Selim Seferbekov			0.42798	2	1y
2	▲ 35	\WM/			0.42842	2	1y
3	▲ 3	NtechLab			0.43452	2	1y
4	▲ 6	Eighteen years old		+5	0.43476	2	1y
5	▲ 12	The Medics	</> DFDC 3D & 2D inc ...		0.43711	2	1y
6	▲ 42	Konstantin Simonchik			0.44289	2	1y
7	▲ 27	All Faces Are Real			0.44531	1	1y
8	▲ 6	ID R&D		+3	0.44837	2	1y
9	▲ 76	名侦探柯西			0.44911	2	1y
10	▲ 23	vcg@xmu			0.45149	2	1y



Public Leaderboard | Private Leaderboard

This leaderboard is calculated with approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

Raw Data | Refresh

In the money | Gold | Silver | Bronze

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	X5 (for cheating)			4.45806	120	2y
2	[ods.ai] Anton Chikin			4.67581	2	2y
3	In			4.81140	10	2y
4	VyD			5.00859	15	2y
5	Born to sleep	</> Fork of Fork of Me...		5.17518	50	2y
6	puppy play	</> albumentations		5.27418	14	2y
7	Chris Deotte			5.41850	21	2y
8	Karsten Tiemann			5.53078	50	2y
9	Madao			5.60532	9	2y
10	Siavash			5.75288	34	2y

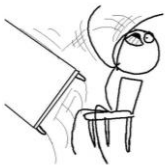
Public Leaderboard | Private Leaderboard

This competition's private leaderboard was evaluated by a private metric. This competition has closed and is no longer accepting submissions.
This competition has completed. This leaderboard reflects the final standings.

Refresh

In the money | Gold | Silver | Bronze

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲ 438	Tom	GAN dogs starter B...		55.42142	25	2y
2	▲ 372	[kaggler-ja] yabea	[Sub] RaLS BigGAN...		56.47019	39	2y
3	▲ 407	lli			70.41224	32	2y
4	▲ 362	bestfitting	i96_1e-4_6e_4_bs3...		77.80956	33	2y
5	▲ 367	[ods.ai] Dmitry Vorobiev	doggies BigGAN su...		82.30464	78	2y
6	▲ 431	Theo Viel	ProCGan 30		82.78395	15	2y
7	▲ 209	Doge			83.77130	59	2y
8	▲ 441	tkato	submit_ac_2		89.17595	38	2y
9	▲ 460	Mark Peng	small-stylegan-v6-...		89.64117	43	2y
10	▲ 457	Johannl	SN-DCGAN		95.45683	43	2y



Outline

- Problem overview
- Plaace introduction and data set description
- Evaluation
- Virtual teams
- What should be delivered?
- Grading
- Computing resources

Outline

- Problem overview
- Plaace introduction and data set description
- Evaluation
- Virtual Teams
- What should be delivered?
- Grading
- Computing resources

Disclaimer:

all the character images in this presentation were generated using the Stable Diffusion Neural Network on https://beta.dreamstudio.ai with the membership license

Some slides have sound effects, please make sure your speakers are muted if you're on a lecture

Press start to continue

NTNU| Norwegian University of Science and Technology

The Story

Three brave students
Applied to the TDT4173 Machine Learning course
As a consequence they were summoned to the mysterious Kaggle platform

To compete in a tournament

whose outcome will decide the fate of the …

COURSE GRADE

This tournament is known as

Machine Learning Kombat

Choose Your Fighter

Choose Your Fighter

NTNU | Norwegian University of Science and Technology

ML Kombat

- 10 virtual teams sorted by difficulty level
- Similar to MK towers
- The more virtual teams you defeat the more base points you will get

Johnny Cage

Liu Kang

NTNU| Norwegian University of Science and Technology

Kenshi

NTNU| Norwegian University of Science and Technology

Kitana

NTNU| Norwegian University of Science and Technology

Raiden

NTNU | Norwegian University of Science and Technology

Sub-Zero

NTNU| Norwegian University of Science and Technology

Scorpion

Goro

Triborg

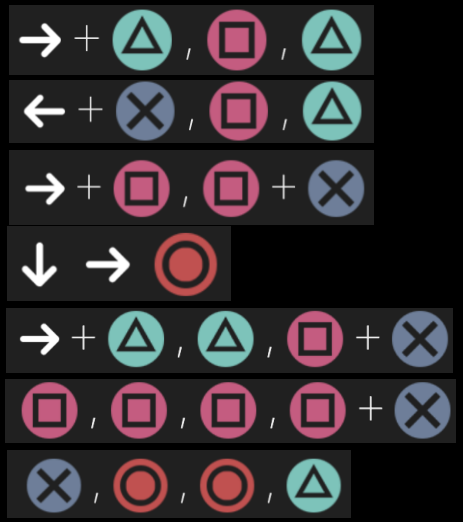
NTNU | Norwegian University of Science and Technology

Shao Kahn

The VTs are not fixed yet

Frozen date is 20/10

Useful combinations



- Deep data analysis and EDA
- Accurate preprocessing
- Good feature engineering
- Strong ML models
- Optimized Hyperparameters
- Stacking/Mixing
- Model results analysis

More details are in the next lectures

Outline

- Problem overview
- Plaace introduction and data set description
- Evaluation
- Virtual teams
- What should be delivered?
- Grading
- Computing resources

Each group

- **Submits 4 things:**
  - Select two predictions on kaggle before the deadline
  - Two **short** Jupyter notebooks:
    - Includes only necessary steps to reproduce your selected predictions
    - naming: short_notebook_{kaggle_submission_id}.ipynb
  - A **long** Jupyter notebook:
    - Contains all attempts in your group work (EDA, all models, algorithms, feature engineering, results interpretation)

- **Submissions to Kaggle, notebooks to Blackboard**

What can I use?

- **Any** languages, tools, platforms, AutoML (offline), libraries, file formats during development.
- But you should use Jupyter Notebook for your delivery
- You **can not use external data**, other than we provided
- Writing massive data in code is not allowed

Late submissions

- **Deadline: 13/11/2022 23:59 Oslo Time**
- Strict.

- Up to this time you should submit both submissions to Kaggle and notebooks to Blackboard

- Late submission deadline: 16/11/2022 23:59 Oslo Time **(-5 points**)
- Even later = the whole group fails the course

Grading

- **Project points = base points + possible deductions**

Letter	Points
A	89-100
B	77-88
C	65-76
D	53-64
E	41-52
F	0-40

Base points are proportional to the number of VTs defeated:

- max 100 (all Vts)
- min 41 (defeat 1 VT)

We use the best (out of 2 submitted) to evaluate your solution

Grading

Possible deductions:

- Pass Individual assignment in the second chance (-5)
- Late submission (-5)
- No exploratory data analysis (EDA) (-3)
- Only one predictor is used (-3)
- No feature engineering (-3)
- No model interpretation (-3)

All deductions are binary (full or no deduction)

Computing resources

NTNU computing resources (TAs: Ivar and Ørjan)

Google Cloud Credits (TAs: Mathias and Torstein)

1. NTNU Through command prompt

1. Open the command prompt
2. Connect to a NTNU network either directly or through vpn
3. Type inn ssh username@tdt4173-xx.idi.ntnu.no
    1. Username is your NTNU username, e.g., ivarrefs
    2. xx is the computer your group get assigned, e.g., 01
4. Type inn your NTNU password
5. You are now in, and should be able to run code from your command prompt. Several of the packages you need for the assignment is already installed, but you might need to install more which can be done through sudo pip install.

2. NTNU Through VS code

1. Download VS code, and connect to a NTNU network
2. Go to extensions and download “remote ssh”

3. Go to “remote explorer”
4. Choose “SSH targets” and click the “+” sign (add ssh)
5. Type in ssh username@tdt4173-xx.idi.ntnu.no
6. Try to open the VM and choose Linux as machine
7. You are now good to go.
    1. You can click on new file and write your python code here and save it.
    2. To upload the data, I have just opened a file with and copied the data from the csv file to this and saved it. Probably better ways of doing this, google your way to it!
8. This video should explain all you need to know to be able to connect to the VMs through VS-code: https://www.youtube.com/watch?v=ar_ZjFu0FP4&t=114s&ab_channel=MannyinTheCloud