

SC4000/CZ4041/CE4041: Machine Learning

Course Project Description

Kelly KE
School of Computer Science and Engineering,
NTU, Singapore

Detailed Project Description

- This is a group-based course project
- Each group consists of at most 5 members (≤ 5)
- Individual “group” is allowed, but not recommended
- Each group can choose one of the Kaggle competitions listed on the next slide as the course project
- Assessments:
 - Project report (30%)
 - Presentation video (10%)

Course Project Candidates – Kaggle Competitions

- Zillow Prize: Zillow's Home Value Prediction (Zestimate)
url: <https://www.kaggle.com/c/zillow-prize-1>
- Sberbank Russian Housing Market
url: <https://www.kaggle.com/c/sberbank-russian-housing-market/>
- Google Smartphone Decimeter Challenge 2022
url: <https://www.kaggle.com/competitions/smartphone-decimeter-2022>
- Store Item Demand Forecasting Challenge
url: <https://www.kaggle.com/c/demand-forecasting-kernels-only/>
- Nomad2018 Predicting Transparent Conductors
url: <https://www.kaggle.com/c/nomad2018-predict-transparent-conductors/>
- HuBMAP – Hacking the Kidney
url: <https://www.kaggle.com/competitions/hubmap-kidney-segmentation>
- PetFinder.my – Pawpularity Contest
url: <https://www.kaggle.com/competitions/petfinder-pawpularity-score>
- Elo Merchant Category Recommendation
url: <https://www.kaggle.com/c/elo-merchant-category-recommendation/>
- Northeastern SMILE Lab – Recognizing Faces in the Wild
url: <https://www.kaggle.com/c/recognizing-faces-in-the-wild/>

Programming Languages

- Programming Languages:
 - Any programming language can be used, e.g., Python, C/C++, Java, R, etc.
 - Any open-source ML toolbox can be used
- Note: directly using the source codes released by participants of Kaggle competitions are NOT allowed (penalty will be made if found)

Key Dates

- Send information on group members via email:
 - by 17th Feb. 2023 (Friday of Week 6)
 - Phase I: find group members by yourself (a forum on NTULearn course site has been created for help)
 - Phase II: will help those who are not able to form a group
- Submit required files via NTULearn:
 - by 11:59pm, 21st Apr. 2023 (Friday of Week 14)

FEBRUARY						
S	M	T	W	T	F	S
			1	2	3	4
5	5	6	7	8	9	10
6	12	13	14	15	16	17
7	19	20	21	22	23	24
	26	27	28			

APRIL						
S	M	T	W	T	F	S
						1
12	2	3	4	5	6	7
13	9	10	11	12	13	14
	16	17	18	19	20	21
	23	24	25	26	27	28
	30					

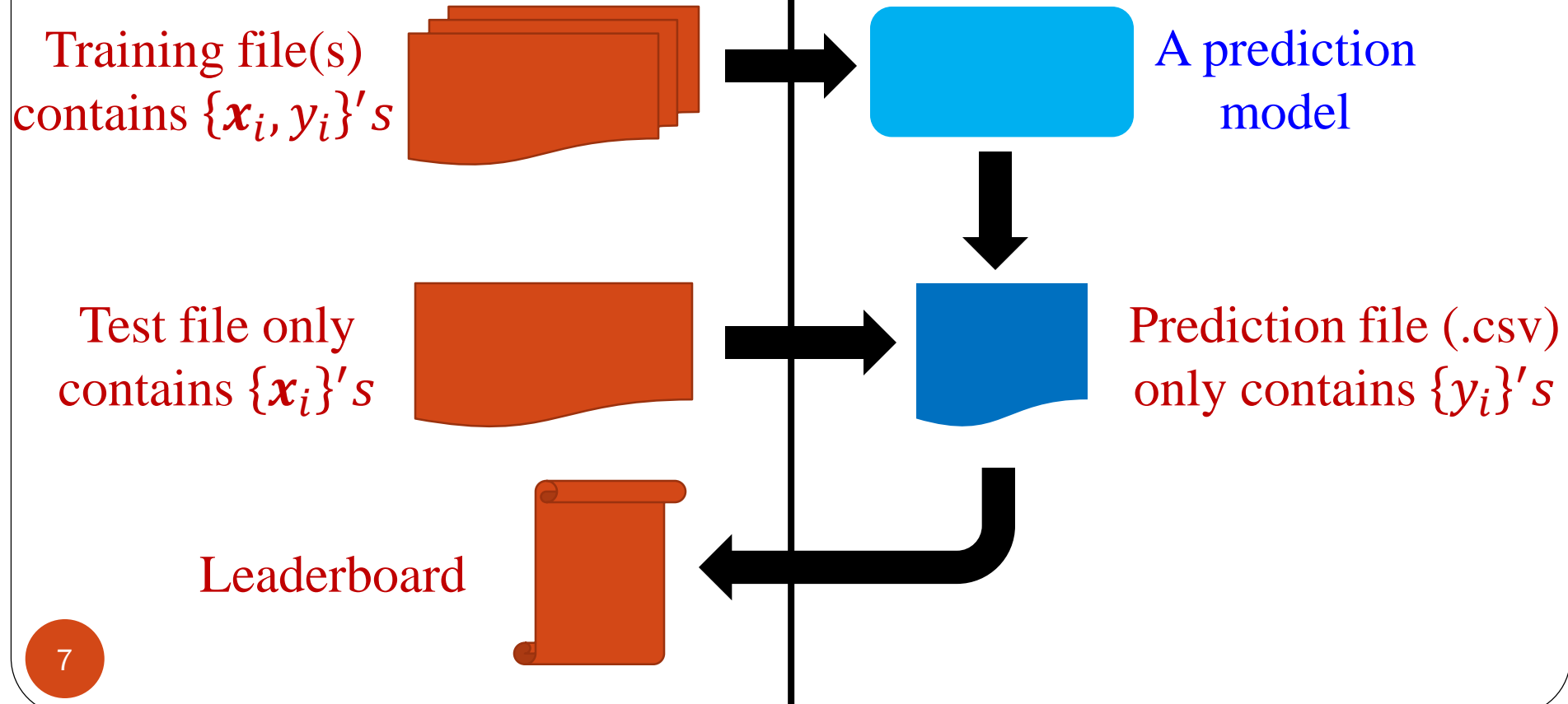
Submission

- Required files to be submitted:
 1. A project report (.pdf or .doc)
 2. A link of presentation video (.txt)
 3. The final .csv file of your prediction results submitted to the specific competition in Kaggle you participate
 - ~~4. Your source codes (with a readme file)~~
- Notes:
 - Only the report and video will be assessed
 - The submitted .csv is to double check whether the reported results are correct
 - You may be randomly asked to provide source codes to check whether they are copied from some participants

General Information of Kaggle

Kaggle.com

Participants



Format and Content of Video

- Presentation video:
 - To summarize your course project in a video of \leq 10 minutes long
 - You can use any tool to produce the video, e.g., simply using PowerPoint or other advanced tools or some online platforms, like <https://www.narakeet.com/>
 - Upload to YouTube and submit the link
 - Some examples for reference:
<https://www.youtube.com/channel/UCSBrGGR7JOiSyzl60OGdKYQ>
https://www.youtube.com/channel/UC_sfVZvvPUbOQhDs_cqlx_A

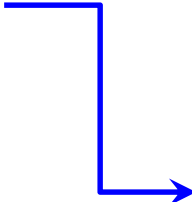
Content of Project Report

- Specific roles and contributions of each group member
 - “Lazy” members will be graded differently
- An evaluation score and ranking position of your prediction results for the specific competition in Kaggle
 - Provide a screenshot of your evaluation score
- Problem statement (using your own words)
- Challenges of the problem
- Your proposed solution in detail (preprocessing, feature engineering/representation learning, methodologies, etc.)
- Experiments to demonstrate why the solution you proposed is effective to solve the problem using experiments
- Conclusion: what you have learned from the project

Format and Assessment on Project Report

- Report format:
 - 12 point font, single space, ≤ 20 pages

- Leaderboard performance (public one)
- Convincingness
- Solution novelty
- Writing



Whether the report is well organized;
Whether the descriptions are logically clear;
Whether the descriptions are detailed enough;
Whether the report contains a lot of typos.

Assessments - Report

- **Leaderboard Performance:** though all the listed Kaggle competitions are completed, you can still submit your results to Kaggle to obtain an evaluation score and find a corresponding ranking position in the **public** leaderboard
- The performance assessment is based on the relative ranking of your results on the specific competition (i.e., top 10%, top 30%, top 50%, top 70%, and the rest)

kaggle

Competitions Datasets Code Discussions Courses ...

🔍 Search

Sign In

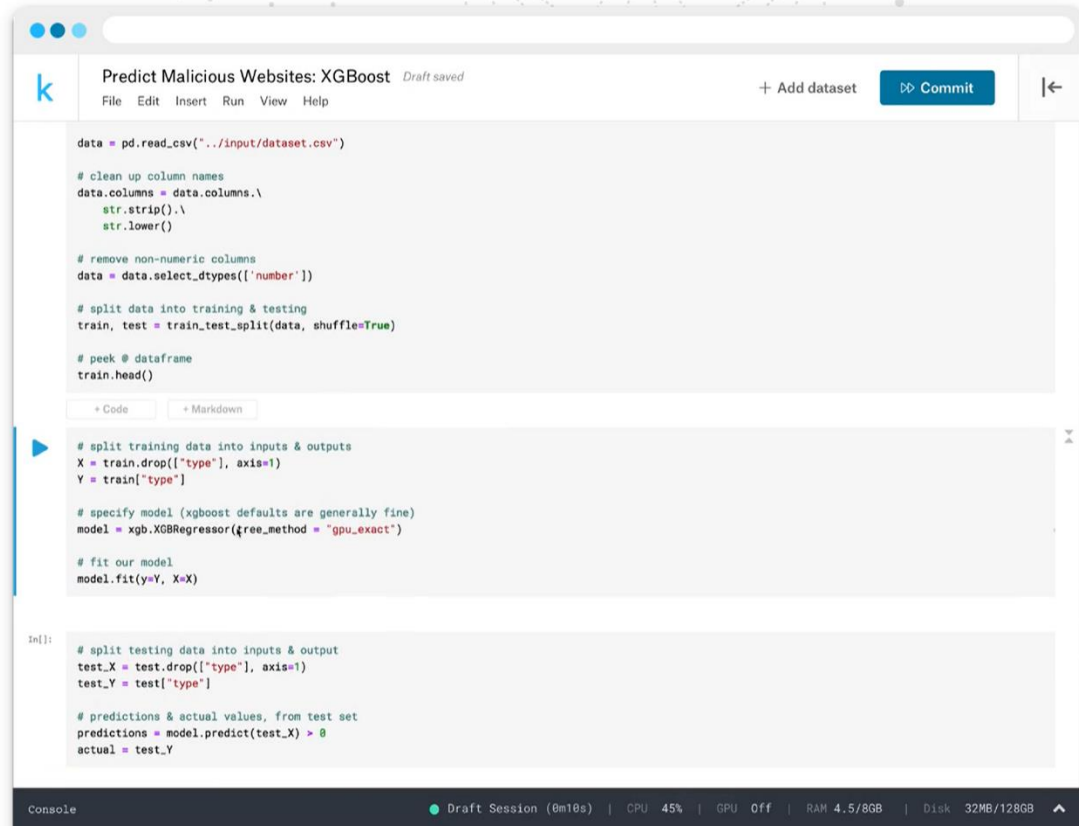
Register

Start with more than a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter
Notebooks environment. Access GPUs at no cost
to you and a huge repository of community
published data & code.

 REGISTER WITH GOOGLE

Register with Email



The screenshot shows a Kaggle Notebook interface. At the top, the title is "Predict Malicious Websites: XGBoost" with a "Draft saved" status. Below the title are tabs for "File", "Edit", "Insert", "Run", "View", and "Help". On the right, there are buttons for "+ Add dataset", "Commit", and a back arrow. The main area contains a Jupyter Notebook with Python code for data preprocessing and model training using XGBoost. The code is organized into three sections: data loading and cleaning, training data preparation, and testing data preparation. A console at the bottom shows the session status: "Draft Session (8m10s) | CPU 45% | GPU Off | RAM 4.5/8GB | Disk 32MB/128GB".

```
data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

# peek @ dataframe
train.head()

# split training data into inputs & outputs
X = train.drop(["type"], axis=1)
Y = train["type"]

# specify model (xgboost defaults are generally fine)
model = xgb.XGBRegressor(tree_method = "gpu_exact")

# fit our model
model.fit(y=Y, X=X)

In[]: # split testing data into inputs & output
test_X = test.drop(["type"], axis=1)
test_Y = test["type"]

# predictions & actual values, from test set
predictions = model.predict(test_X) > 0
actual = test_Y
```

kaggle

+ Create

Home

Competitions

Datasets

<> Code

Discussions

Learn

More

Your Work

View Active Events

Search

Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the [documentation](#) or learn about [Community Competitions](#).

Host a Competition

Your Work



Search competitions

Filters

All Competitions

Everything, past & present

Featured

Premier challenges with prizes

Getting Started

Approachable ML fundamentals

Research

Scientific and scholarly challenges

Community

Created by fellow Kagglers

Get Started

See all

New to Kaggle?

These competitions are perfect for newcomers.



Titanic - Machine Learning from Disaster

Start here! Predict survival on th...
Getting Started



House Prices - Advanced Regression...

Predict sales prices and practice...
Getting Started



Spaceship Titanic

Predict which passengers are tra...
Getting Started
2746 Teams

☰ kaggle

+ Create

🏠 Home

🏆 Competitions

📁 Datasets

<> Code

💬 Discussions

🎓 Learn

v More

📄 Your Work

📅 View Active Events

🔍 Search



Competitions

Your Work

🔍 Google Smartphone Decimeter Challenge 2022

⌵ Filters

Results

Recently Launched ▾ 📅



Google Smartphone Decimeter Challenge 2022

Improve high precision GNSS positioning and navigation accuracy on smartphones
Research · 573 Teams · 5 months ago

\$10,000

...

≡ kaggle

+ Create

🏠 Home

🏆 Competitions

📁 Datasets

<> Code

💬 Discussions

🎓 Learn

v More

📁 Your Work

▼ RECENTLY VIEWED

🌐 Google Smartphone D...

📁 View Active Events

🔍 Search



Research Prediction Competition

Google Smartphone Decimeter Challenge 2022

Improve high precision GNSS positioning and navigation accuracy on smartphones

Google · 573 teams · 5 months ago

\$10,000
Prize Money

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Late Submission ...

Leaderboard

📄 Raw Data

🔄 Refresh

🔍 Search leaderboard

Public

Private

This leaderboard is calculated with approximately 80% of the test data. The final results will be based on the other 20%, so the final standings may be different.

🟢 Prize Contenders

#	Team	Members		Score	Entries	Last	Code
1	Taro		🥇	1.382	21	5mo	
2	A.Saito		🥈	1.473	10	5mo	

kaggle

+ Create

Home

Competitions

Datasets

<> Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Google Smartphone D...

View Active Events

Search



Research Prediction Competition

Google Smartphone Decimeter Challenge 2022

Improve high precision GNSS positioning and navigation accuracy on smartphones

Google · 573 teams · 5 months ago

\$10,000 Prize Money

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Late Submission

Raw Data

Refresh

Leaderboard

Search leaderboard

Public Private

This leaderboard is calculated with approximately 80% of the test data. The final results will be based on the other 20%, so the final standings may be different.

Prize Contenders

#	Team	Members	Score	Entries	Last	Code
1	Taro		1.382	21	5mo	
2	A.Saito		1.473	10	5mo	

kaggle

+ Create

Home

Competitions

Datasets

Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Zillow Prize: Zillow's H...

Google Smartphone D...

View Active Events

Search



Overview Data Code Discussion

Overview

Description

Evaluation

Timeline

Prizes

Acknowledgement

Goal of the competition

The goal of the competition is to develop a model that can predict the resolution of an urban road network.

Your work will be evaluated between the two groups with improved performance upon the network.

Context

Have you ever known the exact location of a point?

Submit to Competition

File Upload Notebook



Drag and drop file to upload

(e.g., .csv, .zip, .gz, .7z)

or

Browse Files

Your submission should be a CSV file with 66097 rows and a header. You can upload a zip/gz/7z archive.

DESCRIPTION

Enter a description

0 / 500

>_ kaggle competitions submit -c smartphone-decimeter-2022 -f su...



Cancel

Submit

kaggle

+ Create

Home

Competitions

Datasets

Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

GSDC2022_kalmanfilter

Google Smartphone D...

Zillow Prize: Zillow's H...

View Active Events

Search



Overview Data Code Discussion

Leaderboard

Search leaderboard

Public Private

The private leaderboard is calculated with ap...
This competition has completed. This leader...

Prize Winners

△ Team

1 — Taro

Submit to Competition

File Upload Notebook



Google Smartphone Decimeter Challenge 2022

Uploaded File

CSV sample_submission.csv (5 MiB)

Your submission should be a CSV file with 66097 rows and a header. You can upload a zip/gz/7z archive.

DESCRIPTION

Enter a description

0 / 500

>_ kaggle competitions submit -c smartphone-decimeter-2022 -f subm...



Cancel

Submit

kaggle

+ Create

Home

Competitions

Datasets

<> Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Google Smartphone D...

GSDC2022_kalmanfilter

Zillow Prize: Zillow's H...

View Active Events

Search

Google Smartphone Decimeter Challenge 2022

Improve high precision GNSS positioning and navigation accuracy on smartphones



Google · 573 teams · 5 months ago

\$10,000

Prize Money

Overview Data Code Discussion **Leaderboard** Rules Team

Submissions

Late Submission

Leaderboard

Raw Data

Refresh

YOUR RECENT SUBMISSION



sample_submission.csv

Submitted by Yiping Ke · Submitted 9 minutes ago

Score: 3037613.293

Private score: 3084280.30

↓ Jump to your leaderboard position

Search leaderboard

Public Private

This leaderboard is calculated with approximately 80% of the test data. The final results will be based on the other 20%, so the final standings may be different.

Prize Contenders

kaggle

+ Create

🏠 Home

🏆 Competitions

📁 Datasets

<> Code

💬 Discussions

🎓 Learn

v More

📁 Your Work

▼ RECENTLY VIEWED

🌈 Google Smartphone D...

💧 GSDC2022_kalmalfilter

🏠 Zillow Prize: Zillow's H...

📁 View Active Events

🔍 Search

Overview	Data	Code	Discussion	Leaderboard	Rules	Team	Submissions	Late Submission	...
557	Rocky						23213.831	1	7mo
558	Atwin Paramudya						32457.566	5	6mo
559	Peter Su						115131.489	3	5mo
560	Fackoly						122299.930	1	5mo
561	Gamba Asesina19						178484.191	6	5mo
562	Hardik						181997.439	4	5mo
	👤	sample_submission.csv					3037613.293		
563	Naruhiko Nakanishi						3037613.293	1	7mo
564	Linh Vuu						3037613.293	1	7mo
565	shawn						3037613.293	2	7mo
566	Ajay Singh 1561						3037613.293	1	7mo
567	Smartphone Trackers						3037613.293	1	7mo
568	Benson Hsieh						3037613.293	1	7mo
569	DIPESH SINGLA						3037613.293	1	6mo
...

+ Create

Home

Competitions

Datasets

Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Google Smartphone D...

GSDC2022_kalmanfilter

Zillow Prize: Zillow's H...

View Active Events

Overview

Data

Code

Discussion

Leaderboard

Rules

Team

Submissions

Late Submission

...



sample_submission.csv

3037613.293

563

Naruhiko Nakanishi



3037613.293

1

8mo

564

Linh Vuu



3037613.293

1

8mo

565

shawn



3037613.293

2

8mo

566

Ajay Singh 1561



3037613.293

1

8mo

567

Smartphone Trackers



3037613.293

1

8mo

568

Benson Hsieh



3037613.293

1

8mo

569

DIPESH SINGLA



3037613.293

1

8mo

570

ecust_gaoting



3037613.293

1

7mo

571

CHDer



3037613.293

2

6mo

572

PoKuan Liu



3037613.293

1

6mo

573

SPPINS



3779084.279

1

8mo

Assessments – Report (cont.)

- **Solution Novelty:** as on Kaggle.com, most participants or winners may discuss or even release their solutions (with codes) on the forum of each specific competition
 - If you propose a new and effective solution, you can get bonus. You are encouraged to propose your own solutions based on your own understandings on the competitions. In the report, highlight your new ideas.
 - Directly reuse released source codes are not allowed!

kaggle

+ Create

Home

Competitions

Datasets

<> Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Google Smartphone D...

GSDC2022_kalmanfilter

Zillow Prize: Zillow's H...

View Active Events

Search

Research Prediction Competition

Google Smartphone Decimeter Challenge 2022

Improve high precision GNSS positioning and navigation accuracy on smartphones

\$10,000
Prize Money

Google · 573 teams · 5 months ago

Overview Data **Code** Discussion Leaderboard Rules Team

New Notebook

Notebooks

Search notebooks

Filters

All Your Work Shared With You Bookmarks

Hotness



Smartphone tracking using GNSS - Team 54

Updated 1mo ago

0 comments · Google Smartphone Decimeter Challenge 2022

0

...



Batch 54

Updated 2mo ago

0 comments · Google Smartphone Decimeter Challenge 2022

1

...



codina

0

☰ kaggle

+ Create

🏠 Home

🏆 Competitions

📁 Datasets

<> Code

💬 Discussions

🎓 Learn

⌵ More

📁 Your Work

▼ RECENTLY VIEWED

🌈 Google Smartphone D...

💧 GSDC2022_kalmanfilter

🏠 Zillow Prize: Zillow's H...

📁 View Active Events

🔍 Search



Research Prediction Competition

Google Smartphone Decimeter Challenge 2022

Improve high precision GNSS positioning and navigation accuracy on smartphones

 Google · 573 teams · 5 months ago

\$10,000
Prize Money

Overview

Data

Code

Discussion

Leaderboard

Rules

Team

Submissions

New Topic

⋮

Discussions

🔔 Follow ▾

🔍 Search discussions

≡ Filters

All Owned Bookmarks

Hotness ▾

Pinned topics



Smartphone Decimeter Challenge at ION GNSS+ 2022

Ashley Chow · Last comment 2mo ago by Ashley Chow

▲ 11
4 comments ⋮

Recap of Competition - Congratulations to the Winners!

Ashley Chow · Posted 4mo ago

▲ 6
⋮

Few datasets with Ground Truth inaccuracy

▲ 14

Assessments - Report (cont.)

- **Convincingness**: the goal of the project report is to convince readers that your proposed solution is proper to solve the specific machine learning task. To do so, in the report,
 - You need to provide detailed motivations and explanations of the techniques you used in the solution, e.g., what is the motivation of a new feature you proposed, why you proposed a specific pre-processing step, why you use the proposed classifier but not others
 - You also need to conduct experiments to further verify your proposed ideas

Assessments – Report (cont.)

- Weight priority:

Convincingness = Writing > Leaderboard
Performance = Solution Novelty

Frequent Q&A

- Can we choose another Kaggle competition which is not on the candidate list?
 - No, you can only choose one from the candidate list.
- Are there requirements on the format of the report?
 - 12 point font, single space, ≤ 20 pages.
- Can we use other ML techniques beyond what are taught in this module, such as deep learning models, for the course project?
 - Yes, you can.
- Can the report and video submission deadline be extended?
 - No, it is a hard deadline.

Thank you!