

# CZ4041/CE4041: Machine Learning

## Course Project Description

Sinno Jialin PAN  
School of Computer Science and Engineering,  
NTU, Singapore  
Homepage: <https://personal.ntu.edu.sg/sinnopan/>

# Detailed Project Description

- This is a group-based course project
- Each group consists of at most 4 members ( $\leq 4$ )
- Individual “group” is allowed, but not recommended
- Each group can choose either one of the Kaggle competitions or one of the research topics listed on the following two slides as the course project

# 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>
- Womxn in Big Data South Africa: Female-Headed Households in South Africa  
url: <https://zindi.africa/competitions/womxn-in-big-data-south-africa-female-headed-households-in-south-africa>
- 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  
<https://www.kaggle.com/c/recognizing-faces-in-the-wild/>

# Course Project Candidates (cont.)

- Research-based projects:

- Semi-supervised Learning

Recommended Datasets: <http://sci2s.ugr.es/keel/semisupervised.php>

- Multi-label Classification

Recommended Datasets: <http://sci2s.ugr.es/keel/multilabel.php>

- Multi-instance Learning

Recommended Datasets: <http://sci2s.ugr.es/keel/category.php?cat=mul>

- Transfer Learning

Recommended Datasets: <https://wilds.stanford.edu/>

- Note: If you want to use other datasets to conduct the listed research topics or propose a new research topic, an approval is needed. Unless you have background in ML, you are NOT suggested to choose research-based project

# 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: for Kaggle competitions, directly using the source codes released by participants are NOT allowed (penalty will be made if found)

# Key Dates

- Sent information about group members via email:
  - by 16<sup>th</sup> Sept. 2022
  - Phase I: find group members by yourself (a forum on NTU has been created for help)
  - Phase II: will help those who are not able to form a group
- Submit files, i.e., the project report, video, in NTULearn:
  - by 11:59pm, 22<sup>nd</sup> Nov. 2022

September 2022 Calendarpedia  
Your source for calendars

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1

© Calendarpedia - www.calendarpedia.com

November 2022 Calendarpedia  
Your source for calendars

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3

© Calendarpedia - www.calendarpedia.com

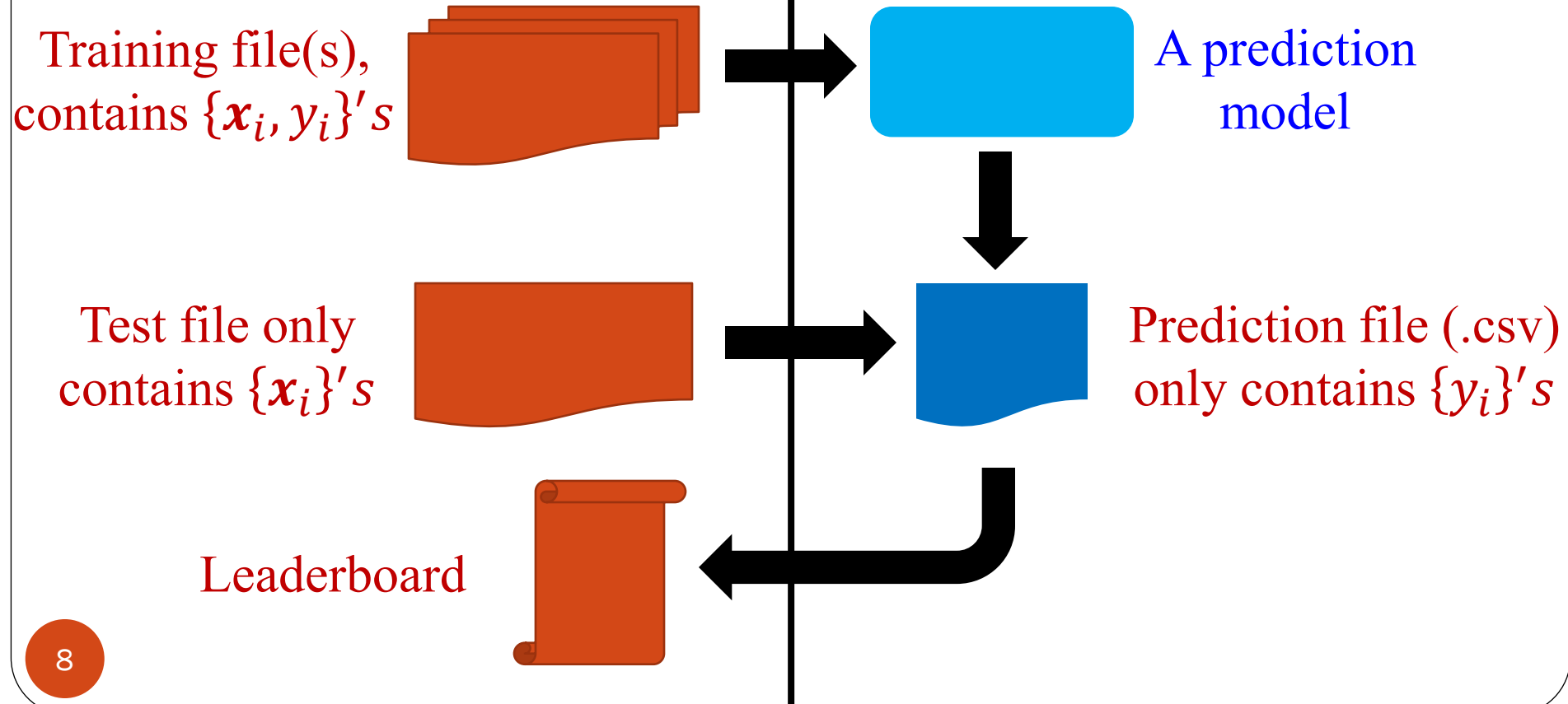
# Submission (Kaggle)

- Submitted files:
  1. A project report
  2. A link of presentation video
  3. The final .csv file of your prediction results submitted to the specific completion 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





# Submission (Research)

- Submitted files:
  1. A project report
  2. A link of presentation video
  - ~~3. Your source codes (with a readme file)~~
- Notes:
  - Only the report and video will be assessed
  - You may be randomly asked to provide source codes to check whether the reported results are correct

# Format and Content of Video

- Presentation video:
  - To summarize your course project in a video of 10-15 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](https://www.youtube.com/channel/UC_sfvZvvPUBOQhDs_cqlx_A)

# Content of Project Report (Kaggle)

- Specific roles and contributions of each group member
  - “Lazy” members will be graded differently
- An evaluation score and ranked 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

# Content of Project Report (Research)

- Specific roles and contributions of each group member
- A review on the specific research topic
- Your new proposed method if applicable
- Comparison experiments on state-of-the-art methods (and your proposed method if applicable)
- Analysis on pros and cons of the compared methods
- Conclusion: you own insights on the research project

# Format and Assessment on Project Report

- Report format:
  - 12 point font, single space, 20-25 pages (not hard constraints)

## Kaggle competitions

- Leaderboard performance
- Convincingness
- Solution novelty
- Writing

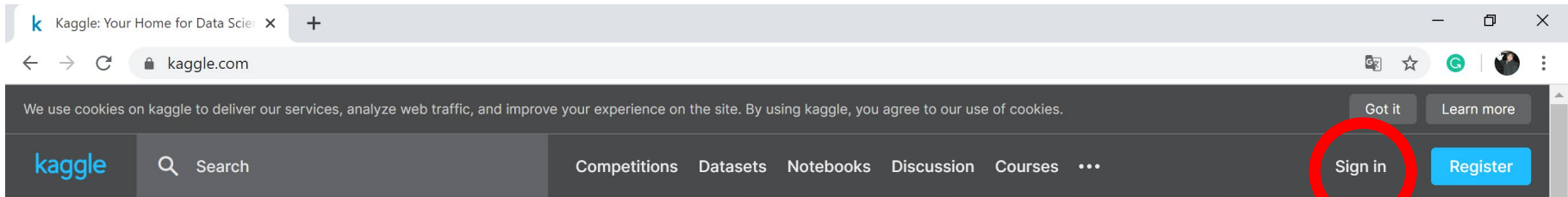
## Research-based projects

- Literature review
- Comparison analysis
- Methodology 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 – Kaggle

- **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
- The performance assessment is based on the relative ranking position of your results on the specific competition (i.e., top 10%, top 30%, top 50%, top 70%, and the rest)

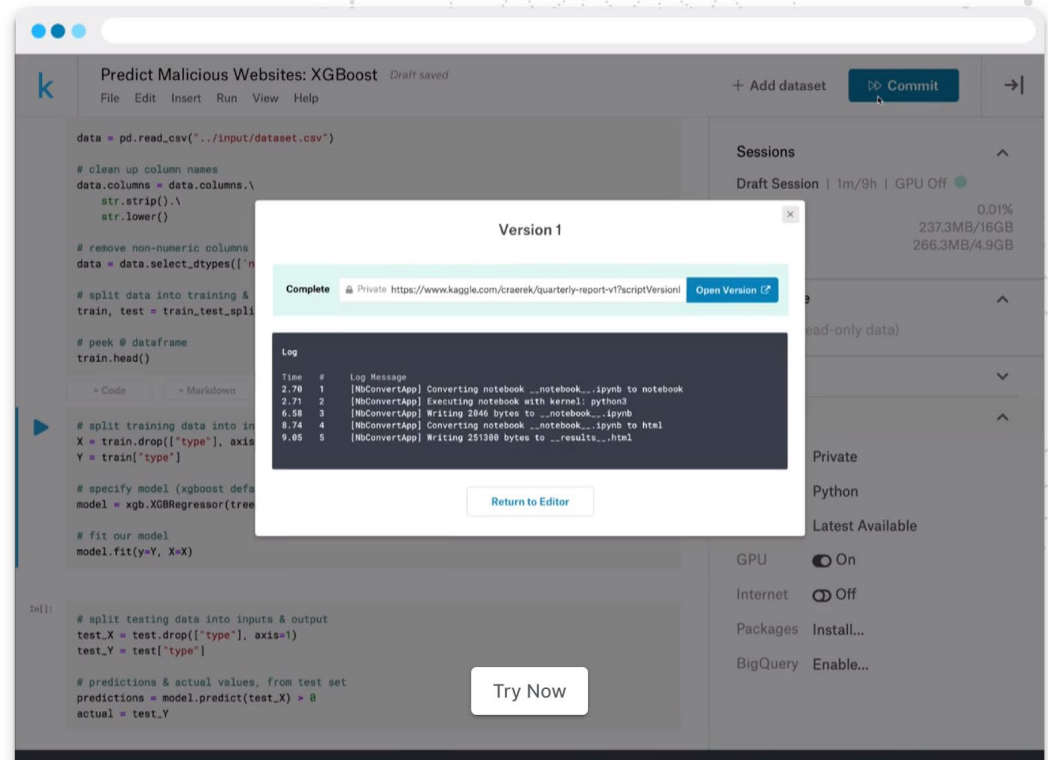


## Start with more than a blinking cursor

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

 REGISTER WITH GOOGLE

[Register with Email](#)



Try Now

Competitions | Kaggle

kaggle.com/competitions

kaggle Search

Competitions Datasets Notebooks Discussion Courses ...


Documentation InClass


# Competitions


General InClass Sort by Grouped

All Categories Search competitions


### 1 Entered Competition




**Digit Recognizer**  
Learn computer vision fundamentals with the famous MNIST data  
[Getting Started](#) · Ongoing ·  tabular data, image data, multiclass classification, object identificati...


 Knowledge  
2,319 teams


### 12 Active Competitions




**Deepfake Detection Challenge**  
Identify videos with facial or voice manipulations  
[Featured](#) · Code Competition · 2 months to go ·  video data, online video

\$1,000,000  
1,189 teams



**2019 Data Science Bowl**  
Uncover the factors to help measure how young children learn  
[Featured](#) · Code Competition · 11 hours to go ·  children, learning, education, video games

\$160,000  
3,495 teams



**TensorFlow 2.0 Question Answering**

\$50,000  
1,224 teams



k Kaggle Competitions | Kaggle

kaggle.com/competitions?sortBy=relevance&group=general&search=Corporación+Favorita+Grocery+Sales+Forecasting&page=1&pageSize=20

kaggle Search Competitions Datasets Notebooks Discussion Courses

# Competitions


Documentation InClass

General InClass


Sort by: Relevance

All Categories Corporación Favorita Gro

1 Competition



**Corporación Favorita Grocery Sales Forecasting**  
Can you accurately predict sales for a large grocery chain?  
**Featured** · 2 years ago · food and drink, tabular data, regression, future prediction



\$30,000  
1,674 teams

Corporación Favorita Grocery Sales Forecasting

← → ↻ kaggle.com/c/favorita-grocery-sales-forecasting/leaderboard

kaggle

Search

Competitions Datasets Notebooks Discussion Courses

Featured Prediction Competition

Corporación Favorita Grocery Sales Forecasting

Can you accurately predict sales for a large grocery chain?

Corporación Favorita

 · 1,674 teams · 2 years ago

\$30,000

Prize Money

Overview

Data

Notebooks

Discussion

Leaderboard

Rules

Team

My Submissions

Late Submission

Public Leaderboard

Private Leaderboard

The private leaderboard is calculated with approximately 69% of the test data.

This competition has completed. This leaderboard reflects the final standings.

Refresh

In the money

Gold

Silver

Bronze

#	Δpub	Team Name	Notebook	Team Members	Score ?	Entries	Last
1	▲13	w		<div><div></div><div></div><div></div></div>	0.50918	117	2y
2	▲10	SoLucky		<div><div></div><div></div><div></div><div></div></div>	0.51296	308	2y
3	▼1	slonoschildpad		<div><div></div><div></div><div></div></div>	0.51309	265	2y
4	▼3	spp		<div><div></div><div></div><div></div></div>	0.51318	165	2y
5	▲11	Lingzhi		<div><div></div></div>	0.51340	145	2y

Windows Taskbar

9:07 PM 1/22/2020

18



Corporación Favorita Grocery Sales Forecasting

← → ↺ kaggle.com/c/favorita-grocery-sales-forecasting/submit

OverviewDataNotebooksDiscussionLeaderboardRulesTeamMy SubmissionsLate Submission

Make a submission for [dmcat](#)

Step 1

Upload submission file

Uploading file: 5\_merged\_sub.csv

1.2 MB/s25.33 MB of 60.76 MB, 29 secs left

File Format

Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions

We expect the solution file to have 3370464 prediction rows. This file should have a header row. Please see sample submission file on the [data page](#).

Step 2

Describe submission

↶ ↷ ↵ ⌨ B I 🔗 “ < > ☰ ☷ 🖼️ 😊

Briefly describe your submission

Make Submission

Corporación Favorita Grocery Sales Forecasting

Can you accurately predict sales for a large grocery chain?

\$30,000 Prize Money

Corporación Favorita · 1,674 teams · 2 years ago

Overview Data Notebooks Discussion Leaderboard Rules Team My Submissions Late Submission

Your most recent submission


Name	Submitted	Wait time	Execution time	Score
5_merged_sub.csv	a few seconds ago	0 seconds	22 seconds	0.50814

Complete

[Jump to your position on the leaderboard](#)

Make a submission for [dmcat](#)

Step 1  
Upload submission file





Overview Data Notebooks Discussion **Leaderboard** Rules Team My Submissions **Late Submission**

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
5_merged_sub.csv	8 minutes ago	0 seconds	23 seconds	0.50814

Complete

[Jump to your position on the leaderboard](#)

**Public Leaderboard** Private Leaderboard

The private leaderboard is calculated with approximately 69% of the test data.  
This competition has completed. This leaderboard reflects the final standings.

Refresh

In the money Gold Silver Bronze

#	Δpub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲13	w			0.50918	117	2y
2	▲10	SoLucky			0.51296	308	2y
3	▼1	slonoschildpad			0.51309	265	2y
4	▼3	spp			0.51318	165	2y
5	▲11	Lingzhi			0.51340	145	2y
6	▲25	Fran & Nicolas & Kevin			0.51467	264	2y

TransactionReport....pdf

72c28d62-7ab4-43....p...

Show all

Corporación Favorita Grocery Sales Forecasting | Kaggle

← → ↻ kaggle.com/c/favorita-grocery-sales-forecasting/leaderboard#score

Overview Data Notebooks Discussion **Leaderboard** Rules Team My Submissions Late Submission

40	▲ 89	Zhang Tao		0.51834	85	2y
41	▼ 31	Bayes tribe (贝叶斯部落)		0.51836	118	2y
42	▲ 11	AlvinZhu		0.51856	108	2y
43	▲ 642	heavenmarshal		0.51857	11	2y
44	▲ 59	Gaurav Kumar		0.51857	46	2y
45	▼ 21	Kulagin+ZavodRobotov		0.51862	120	2y
46	▲ 442	FarTaFar		0.51865	15	2y
47	▲ 101	learn		0.51868	20	2y
48	▲ 4	enjoy		0.51872	3	2y
49	▼ 22	tosh		0.51876	143	2y
50	▲ 11	Vicens Gaitan		0.51887	32	2y
51-1674		<a href="#">Load 1624 More</a>				

© 2020 Kaggle Inc Our Team Terms Privacy Contact/Support

YouTube Twitter Facebook LinkedIn

TransactionReport....pdf 72c28d62-7ab4-43....p... Show all

Windows Taskbar: Search, Chrome, Edge, File Explorer, PowerPoint, PDF Reader, WeChat, System Tray (CH, Volume, Network, Date/Time: 10:22 PM 1/22/2020)

# Assessments – Kaggle (cont.)

- **Solution Novelty**: as on Kaggle.com, most participants or winners may discuss or even release their solutions (with codes) on the forums 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!



Corporación Favorita Grocery Sales Forecasting

Can you accurately predict sales for a large grocery chain?

\$30,000 Prize Money

Corporación Favorita · 1,674 teams · 2 years ago

Overview Data **Notebooks** Discussion Leaderboard Rules Team
















New Notebook

Public Your Work Shared With You Favorites

Sort by Hotness

Outputs Languages Tags

Search notebooks

492		<b>Shopping for Insights - Favorita EDA</b> 1y ago <a href="#">beginner</a> , <a href="#">eda</a> , <a href="#">data visualization</a>	 Rmd  90
194		<b>Comprehensive Python and D3.js Favorita analytics</b> 2y ago	 Py  31
150		<b>Log MA and Days of Week Means (LB: 0.529)</b> 2y ago <a href="#">0.529</a>	 Py  32
118		<b>LGBM Starter</b> 2y ago <a href="#">0.529</a>	 Py  48
104		<b>LGBM One Step Ahead</b>	 Py  24

Corporación Favorita Grocery Sales Forecasting

Can you accurately predict sales for a large grocery chain?










\$30,000 Prize Money

Corporación Favorita · 1,674 teams · 2 years ago

Overview Data Notebooks **Discussion** Leaderboard Rules Team My Submissions New Topic

163 topics [Follow](#) Sort by Hotness

All Mine Upvoted Search topics

20			<b>Welcome from Corporación Favorita</b> Corporación Favorita 2 years ago	last comment by caijiangyao1991 8mo ago	6
25			<b>Welcome!</b> inversion 2 years ago	last comment by caijiangyao1991 8mo ago	16
5			<b>Special Ecuadorian Local Prize</b> Julia Elliott 2 years ago	last comment by Julia Elliott 2y ago	0
23			<b>External Data Thread</b> inversion 2 years ago	last comment by Romeo Cabrera 2y ago	29
0			<b>DT (xgb &amp; lgb) and NN on Time Series</b> leric 2 months ago	last comment by leric 2mo ago	0

# Assessments – Kaggle (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 not others
  - You also need to conduct experiments to further verify your proposed ideas

# Assessments – Kaggle (cont.)

- Weight priority:

Convincingness = Writing > Leaderboard  
Performance = Solution Novelty

# Assessments – Research

- **Literature Review:** as this is a research project, figuring out what have been done in the literature is important. You should provide a comprehensive review on the specific research topic studied in your project

# Assessments – Research (cont.)

- **Comparison Analysis:** you need to implement various state-of-the-art methods for the research topic studied in your research project, and analyze their cons and pros with your own insights

# Assessments – Research (cont.)

- **Methodology Novelty**: if you propose a new and effective method for the specific research topic, even though it might be incremental, you can get bonus. You are encouraged to propose your own methods based on your understandings on the research topic

# Assessments – Research (cont.)

- Weight priority:

Literature Review = Writing = Comparison  
Analysis > Methodology Novelty



# Frequent Q&A

- Can we choose another Kaggle competition which is not in the candidate list?
  - No, you can only choose one from the candidate list if you decide to conduct a Kaggle competition for your course project
- Are there requirements on the format of the report?
  - 12 point font, single space, 20-25 pages. However, these are suggestions not hard requirements
- 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!**