CZ4041/CE4041: Machine Learning

Course Project Description

Sinno Jialin PAN
School of Computer Science and Engineering,
NTU, Singapore
Homepage: http://www3.ntu.edu.sg/home/sinnopan/

Detailed Project Description

- This is a group-based course project
- Each group consists of at most 5 members
- 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/
- Costa Rican Household Poverty Level Prediction
  url: https://www.kaggle.com/c/costa-rican-household-poverty-prediction/
- 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/
- New York City Taxi Trip Duration
  url: https://www.kaggle.com/c/nyc-taxi-trip-duration/
- 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
- Plant Seedlings Classification
  url: https://www.kaggle.com/c/plant-seedlings-classification
- Dog Breed Identification
  url: https://www.kaggle.com/c/dog-breed-identification

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

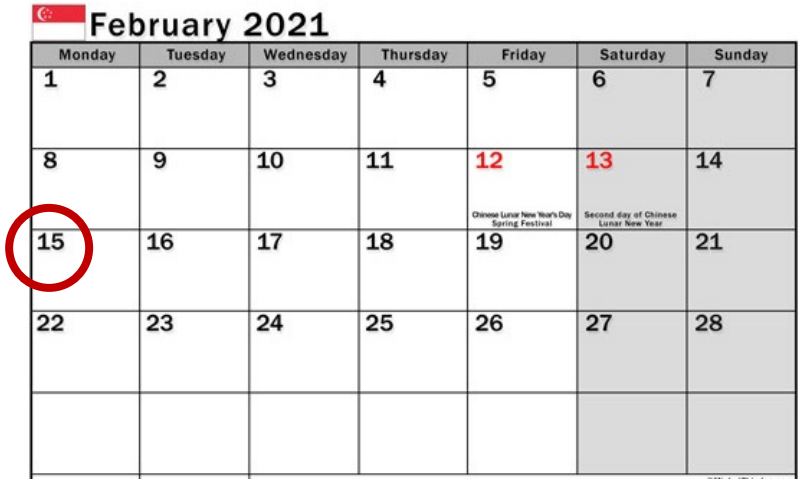
  Recommended Datasets:
  https://www.kaggle.com/c/transfer-learning-on-stack-exchange-tags
  https://ai.bu.edu/visda-2018/
- Note: If you want to use other datasets to conduct the listed research topics or propose a new research topic, an approval is needed. Unless you have background in ML, you are suggested 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

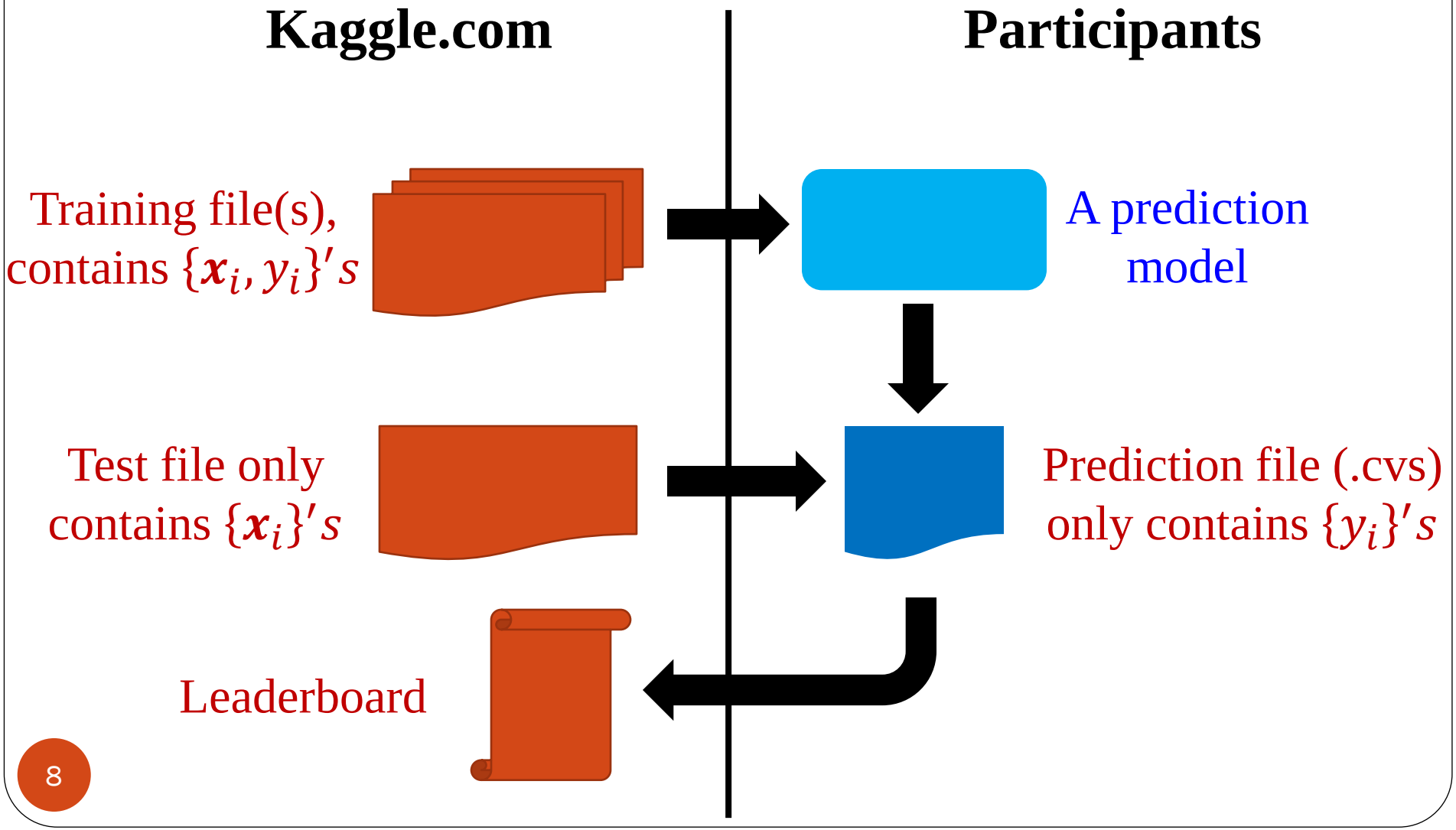
- <u>Sent information on group members via email:</u>
  - by 15th Feb. 2021
- <u>Submit files, i.e., the project report, video, source codes, through NTULearn:</u>
  - by 11:59pm, 25th Apr. 2021



Submission (Kaggle)

- Submitted files:
  1. A project report
  2. A presentation video
  3. The final .cvs 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 .cvs is to double check whether the reported results are correct
  - The submitted source codes are to double check whether they are just copied from some participants

General Information of Kaggle



Submission (Research)

- Submitted files:
  1. A project report
  2. A presentation video
  3. Your source codes (with a readme file)
- Notes:
  - Only the report and video will be assessed
  - The submitted source codes are to double 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/
  - File size $\leq$ 10M
  - Some examples for reference:

https://www.youtube.com/channel/UCSBrGGR7JOiSyzl60OGdKYQ

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 appropriate 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

Kaggle competitions

- Leaderboard performance
- Convincingness
- Solution novelty
- Writing

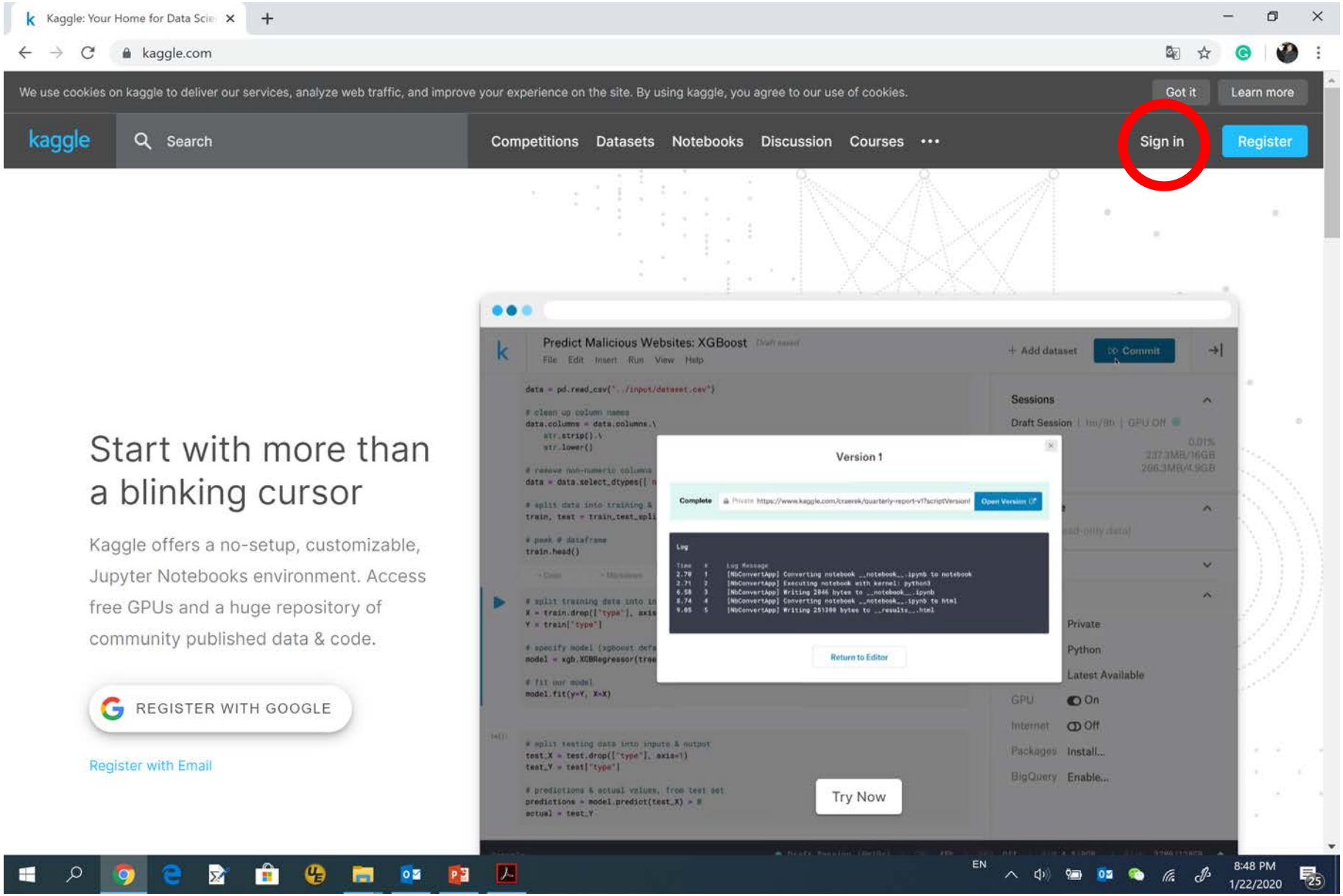
Research-based projects

- Literature review
- Comparison analysis
- Methodology novelty
- Writing

Whether the report is well organized
Whether the descriptions are logically clear
Whether the report is easy to follow
Whether the report contains a lot of typos

Assessments – Kaggle

- **<u>Leaderboard Performance:</u>** 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 relatively ranking of your results on the specific competition (i.e., top 10%, top 30%, top 50%, top 70%, and the rest)


Kaggle: Your Home for Data Scie
kaggle.com
We use cookies on kaggle to deliver our services, analyze web traffic, and improve your experience on the site. By using kaggle, you agree to our use of cookies.
Got it
Learn more
kaggle
Search
Competitions
Datasets
Notebooks
Discussion
Courses
Sign in
Register
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
Predict Malicious Websites: XGBoost
File Edit Insert Run View Help
Add dataset
Commit
Sessions
Draft Session
Version 1
Complete
Open Version
Log
Return to Editor
Private
Python
Latest Available
GPU
On
Internet
Off
Packages
Install...
BigQuery
Enable...
Try Now
EN
8:48 PM
1/22/2020

Competitions | Kaggle

kaggle.com/competitions

kaggle | Search | Competitions | Datasets | Notebooks | Discussion | Courses | ...

Competitions

Documentation | InClass

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

EN 8:53 PM 1/22/2020

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

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

EN 8:56 PM 1/22/2020

Corporación Favorita Grocery Sa

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?

$30,000 Prize Money

Corporación Favorita · 1,674 teams · 2 years ago

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.

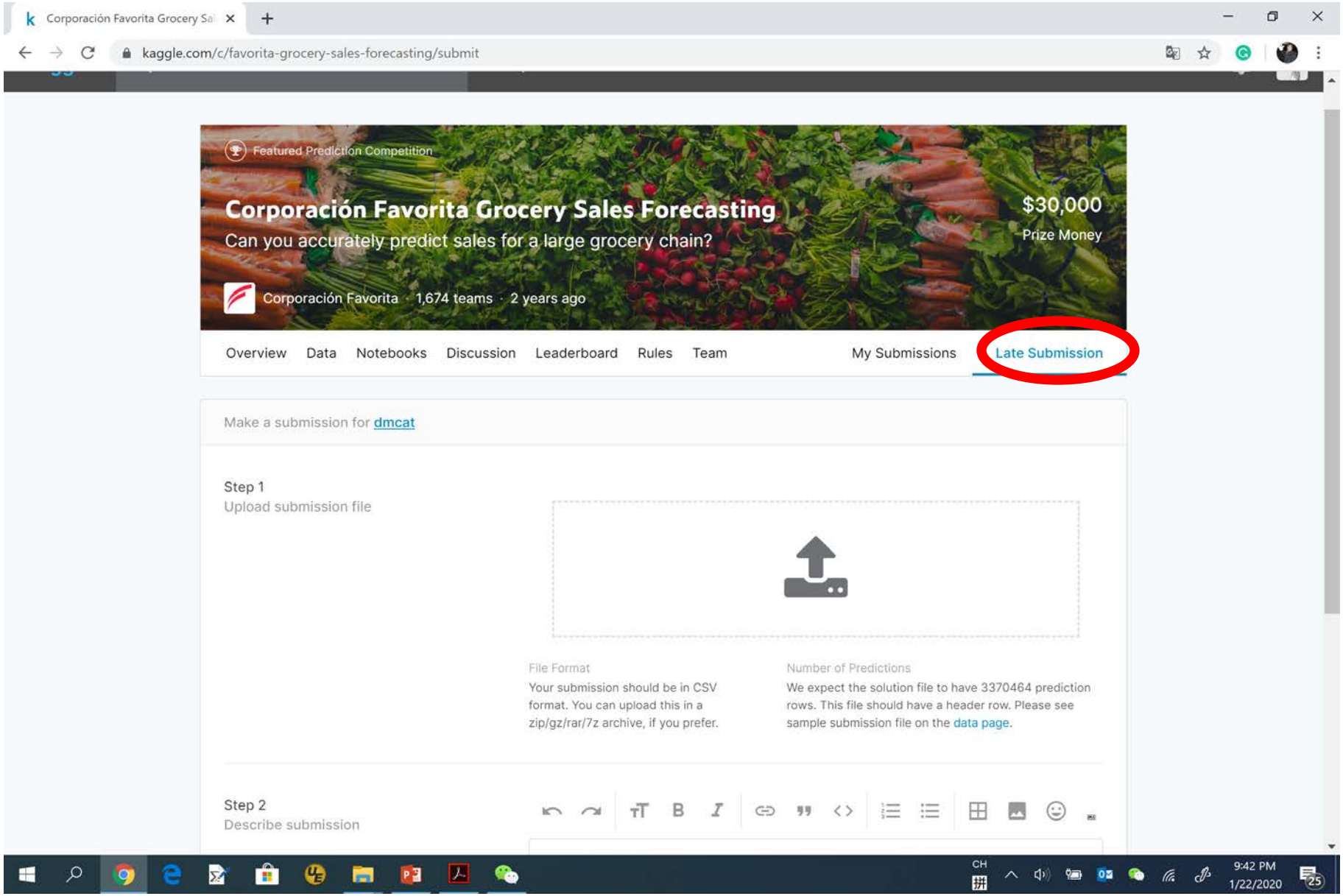
This competition has completed. This leaderboard reflects the final standings.

Refresh

In the money | Gold | Silver | Bronze

#	Δpub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲13	w			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

EN 9:07 PM 1/22/2020


Corporación Favorita Grocery Sa
kaggle.com/c/favorita-grocery-sales-forecasting/submit
Featured Prediction Competition
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
Make a submission for dmcat
Step 1
Upload submission file
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
CH
9:42 PM
1/22/2020

Corporación Favorita Grocery Sa

kaggle.com/c/favorita-grocery-sales-forecasting/submit

Overview Data Notebooks Discussion Leaderboard Rules Team My Submissions Late Submission

Make a submission for dmcat

Step 1
Upload submission file

Uploading file: 5_merged_sub.csv

1.2 MB/s

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

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

Briefly describe your submission

Make Submission

9:44 PM
1/22/2020

Corporación Favorita Grocery Sa

kaggle.com/c/favorita-grocery-sales-forecasting/submit

Featured Prediction Competition

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

9:45 PM
1/22/2020

Corporación Favorita Grocery Sa

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

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	Eran & Nicolas & Kouino			0.51467	264	2y

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

CH | 10:10 PM 1/22/2020

Corporación Favorita Grocery Sa

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 Load 1624 More

Our Team Terms Privacy Contact/Support

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

CH 10:22 PM 1/22/2020

Assessments – Kaggle (cont.)

- **<u>Solution Novelty:</u>** 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 Sa | Conference Management Toolkit

kaggle.com/c/favorita-grocery-sales-forecasting/notebooks

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?

$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 Shopping for Insights - Favorita EDA
1y ago beginner, eda, data visualization — Rmd 90

194 Comprehensive Python and D3.js Favorita analytics
2y ago — Py 31

150 Log MA and Days of Week Means (LB: 0.529)
2y ago 0.529 — Py 32

118 LGBM Starter
2y ago 0.529 — Py 48

104 LGBM One Step Ahead — Py 24

EN 11:07 PM 1/22/2020

Corporación Favorita Grocery Sa | Conference Management Toolkit

kaggle.com/c/favorita-grocery-sales-forecasting/discussion

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?

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

EN 11:08 PM 1/22/2020

Assessments – Kaggle (cont.)

- **<u>Convincingness:</u>** 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

- **<u>Literature Review:</u>** 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.)

- **<u>Comparison Analysis:</u>** 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.)

- **<u>Methodology Novelty:</u>** 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

Thank you!