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 4-5 members
- 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
- Teaching Assistant: Mr. Jianda CHEN (JIANDA001@e.ntu.edu.sg)

Course Project Candidates

Kaggle competitions:

- BigQuery-Geotab Intersection Congestion
  url: https://www.kaggle.com/c/bigquery-geotab-intersection-congestion/
  dataset: csv (54 MB)
- EEE-CIS Fraud Detection
  url: https://www.kaggle.com/c/ieee-fraud-detection
  dataset: csv (118 MB)
- Categorical Feature Encoding Challenge
  url: https://www.kaggle.com/c/cat-in-the-dat
  dataset: csv (21 MB)
- Severstal: Steel Defect Detection
  url: https://www.kaggle.com/c/severstal-steel-defect-detection
  dataset: image (2G)
- Northeastern SMILE Lab - Recognizing Faces in the Wild
  url: https://www.kaggle.com/c/recognizing-faces-in-the-wild
  dataset: image (381 MB)
- Instant Gratification
  url: https://www.kaggle.com/c/instant-gratification
  dataset: CSV (414 MB)

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

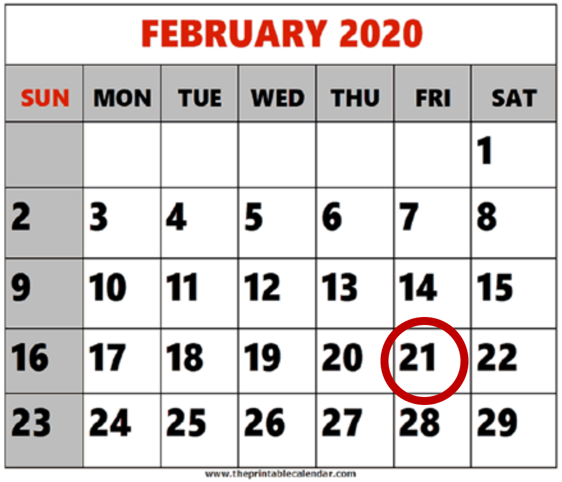
    https://ai.bu.edu/visda-2018/
- Note: If you want to use other datasets to conduct the listed research topics, an approval is needed

Programming Languages

- Programming Languages:
  - Any programming language can be used, e.g., Matlab, 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 (20% penalty will be made if found)

Key Dates

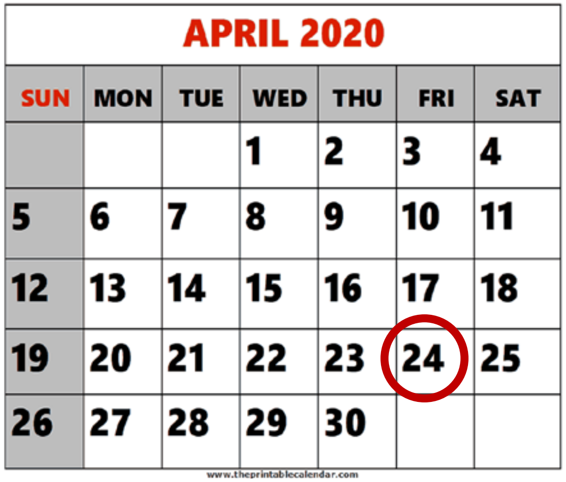
- Sent information on group members via email:
  - by 21st Feb. 2020
- Submit files, i.e., the project report, video, source codes, through NTULearn:
  - by 11:59pm, 24th Apr. 2020



FEBRUARY 2020

SUN	MON	TUE	WED	THU	FRI	SAT
						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

www.theprintablecalendar.com



APRIL 2020

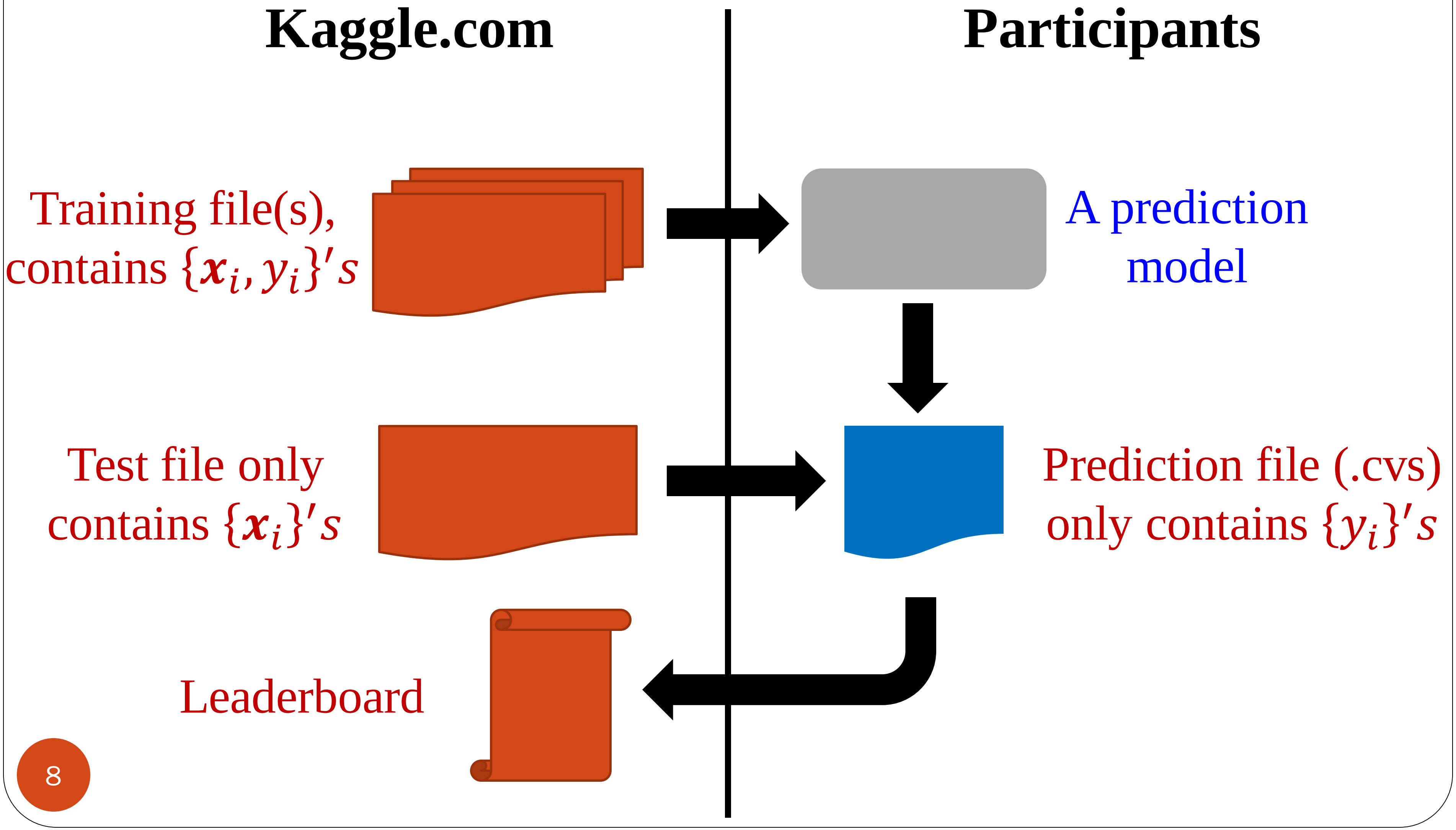
SUN	MON	TUE	WED	THU	FRI	SAT
		1	2	3	4	
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30		

www.theprintablecalendar.com

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:
  - 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 the ones released by some participants

General Information of Kaggle



Kaggle.com

Participants

Training file(s), contains $\{\boldsymbol{x}_i, y_i\}'s$

A prediction model

Test file only contains $\{\boldsymbol{x}_i\}'s$

Prediction file (.cvs) only contains $\{y_i\}'s$

Leaderboard

Submission (Research)

- Submitted files:
  1. A project report
  2. A presentation video
  3. Your source codes (with a readme file)
- Notes:
  - The submitted source codes are to double check whether the reported results are correct

Format and Content of Video

- Presentation video:
  - To introduce your course project in a video of 10-15 minutes long
    - The video is a visual summary of your course report
  - You can use any tool to produce the video, e.g., simply using PowerPoint or other advanced tools
  - File size ≤ 8M
  - Some examples for reference:

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

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

Content of Project Report (Kaggle)

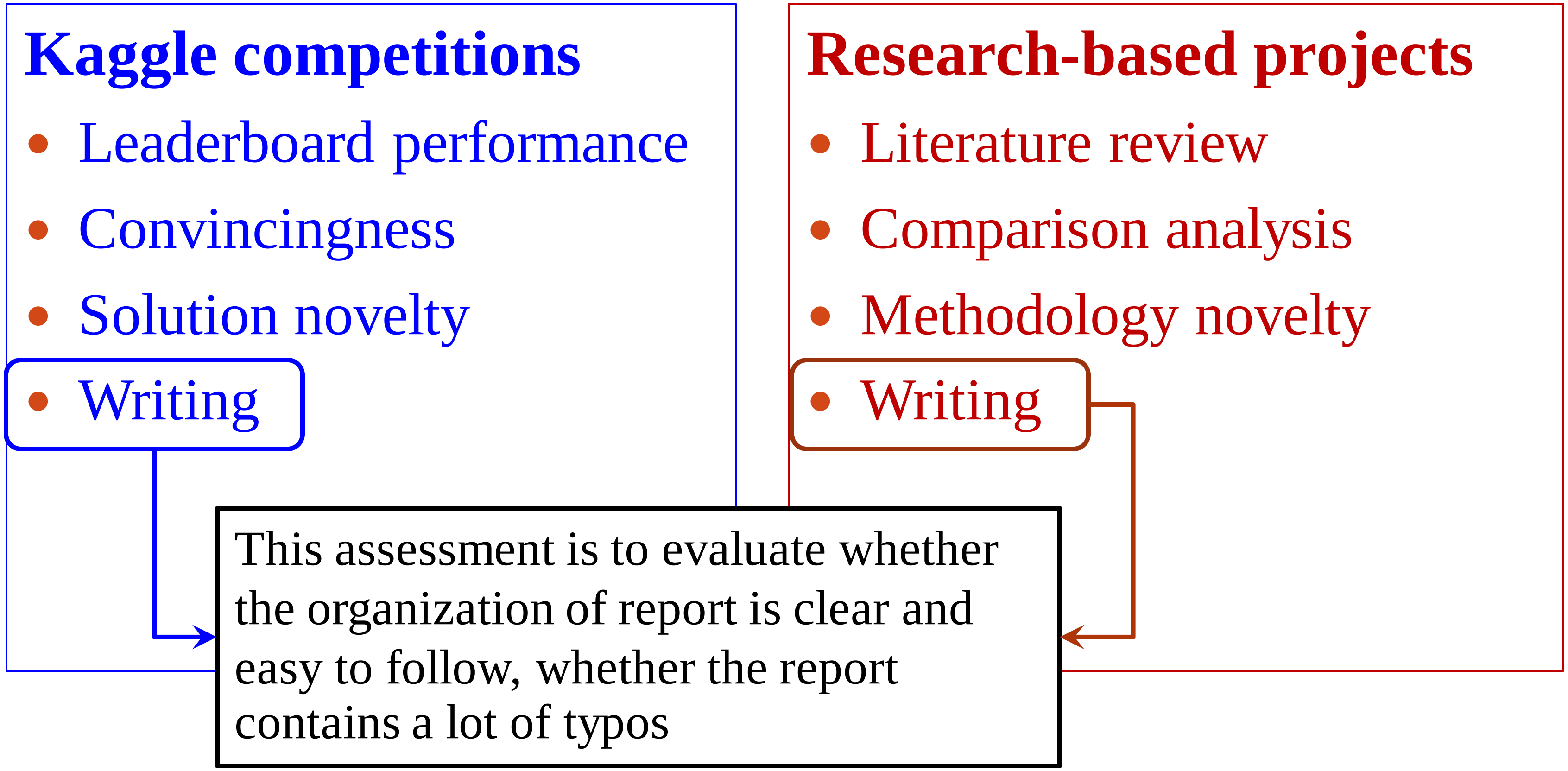
- Specific roles of each group member
- An evaluation score and ranked position of your prediction results for the specific competition in Kaggle
- Problem statement (using your own words instead of copy-and-paste from Kaggle)
- Challenges of the problem
- Your proposed solution in detail (preprocessing, 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 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

- <u>Report format:</u>
  - 12 point font, single space, 20-25 pages



Kaggle competitions

- Leaderboard performance
- Convincingness
- Solution novelty
- Writing

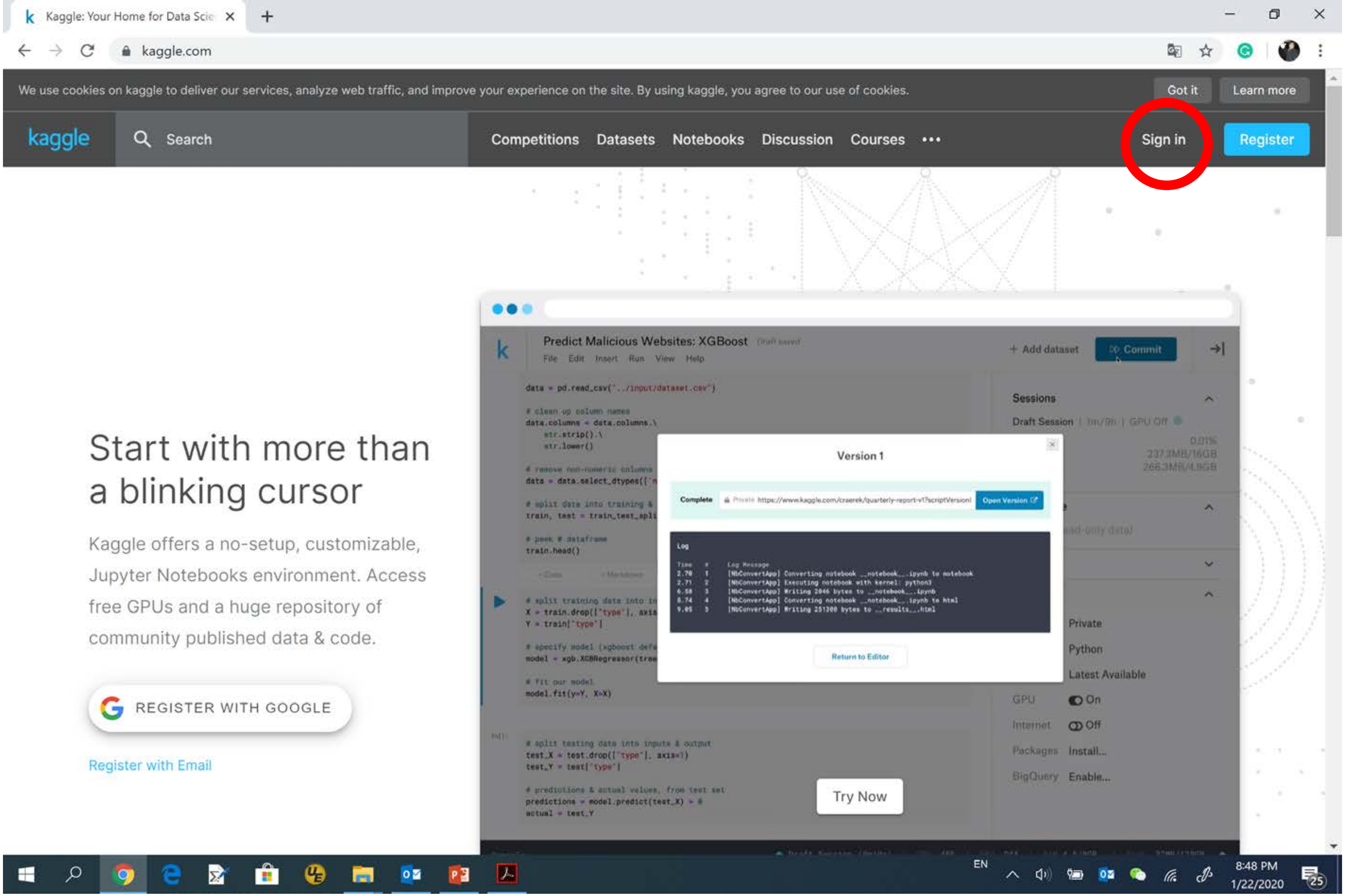
Research-based projects

- Literature review
- Comparison analysis
- Methodology novelty
- Writing

This assessment is to evaluate whether the organization of report is clear and easy to follow, 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 a ranking position
- The performance assessment is based on the relatively ranking of your results on the specific competition (i.e., top 20%, top 40%, top 60%, top 80%, and top 100%)



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.

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.

```
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(['n

# split data into training &
train, test = train_test_spli

# peek # dataframe
train.head()
```

```
# split training data into in
X = train.drop(['type'], axis
Y = train['type']

# specify model (xgboost defa
model = xgb.XGBRegressor(tree

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

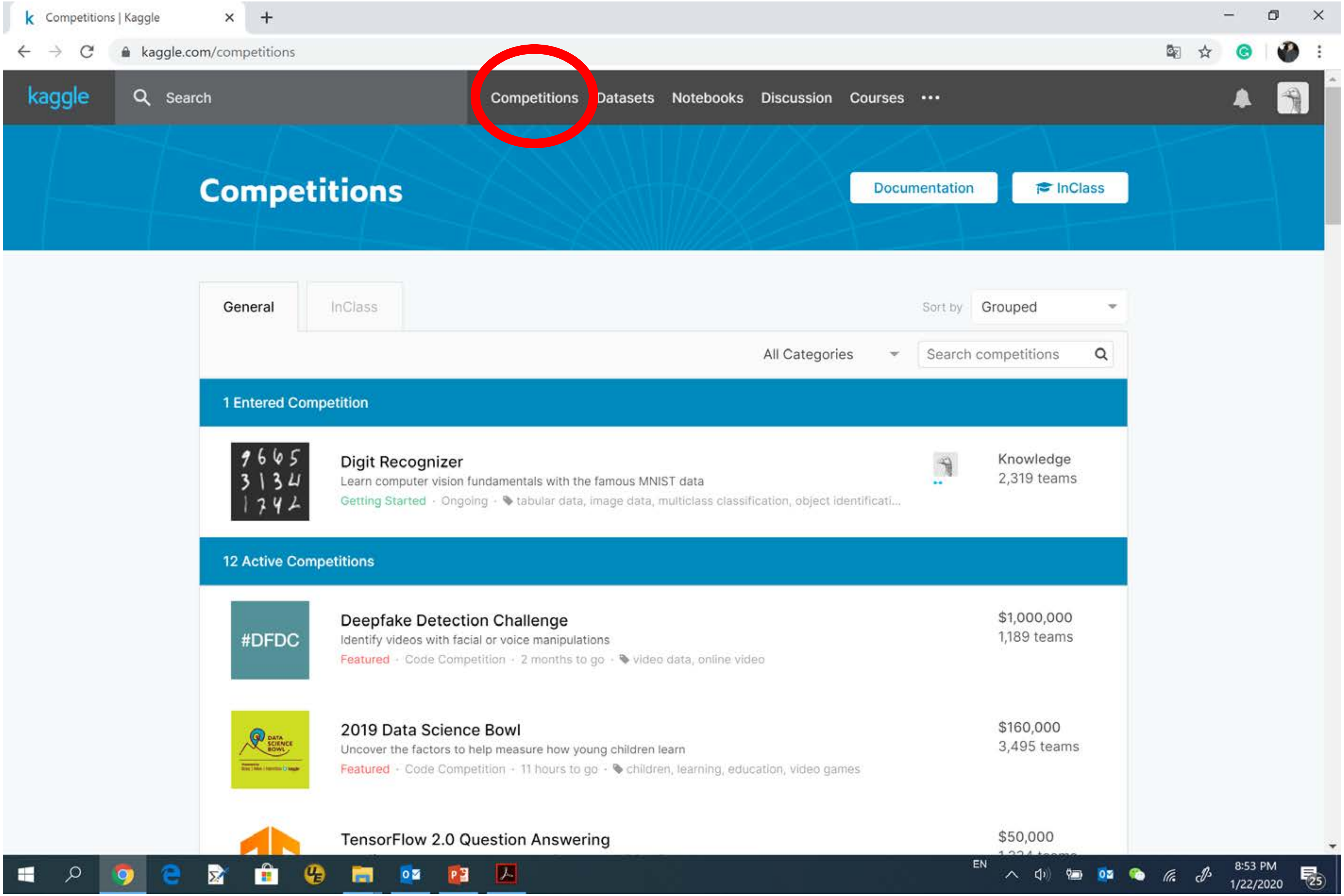
```
# split testing data into inputs & output
test_X = test.drop(['type'], axis=1)
test_Y = test['type']

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

https://www.kaggle.com/craerek/quarterly-report-v1?scriptVersionI

```
Log

Time    #    Log Message
2.70    1    [NbConvertApp] Converting notebook __notebook__.ipynb to notebook
2.71    2    [NbConvertApp] Executing notebook with kernel: python3
6.58    3    [NbConvertApp] Writing 2046 bytes to __notebook__.ipynb
8.74    4    [NbConvertApp] Converting notebook __notebook__.ipynb to html
9.05    5    [NbConvertApp] Writing 251200 bytes to __results__.html
```



Competitions

1 Entered Competition

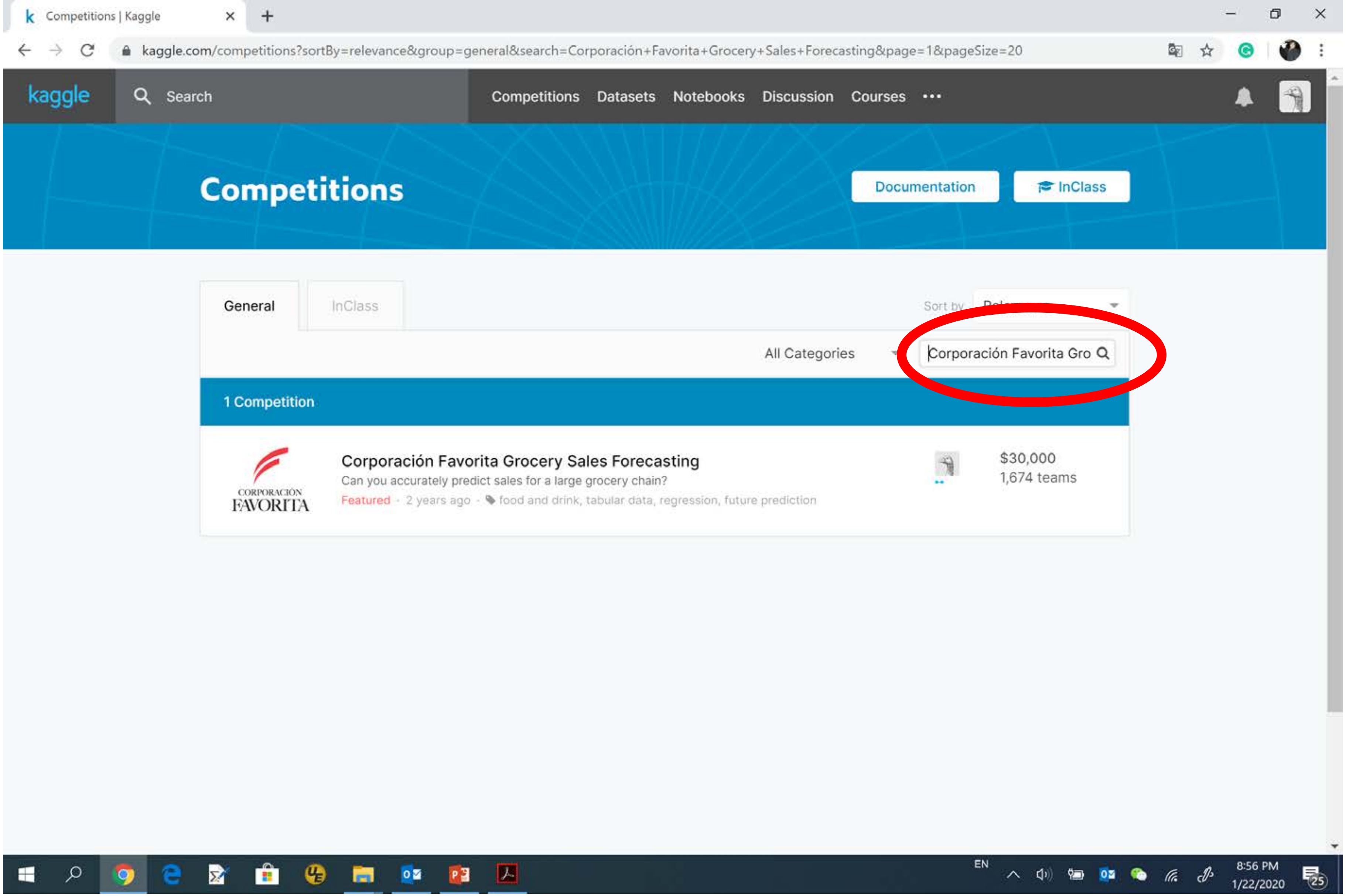
Digit Recognizer
Learn computer vision fundamentals with the famous MNIST data
Getting Started · Ongoing · tabular data, image data, multiclass classification, object identificati...
Knowledge
2,319 teams

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



Competitions | Kaggle

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

Competitions

Corporación Favorita Gro

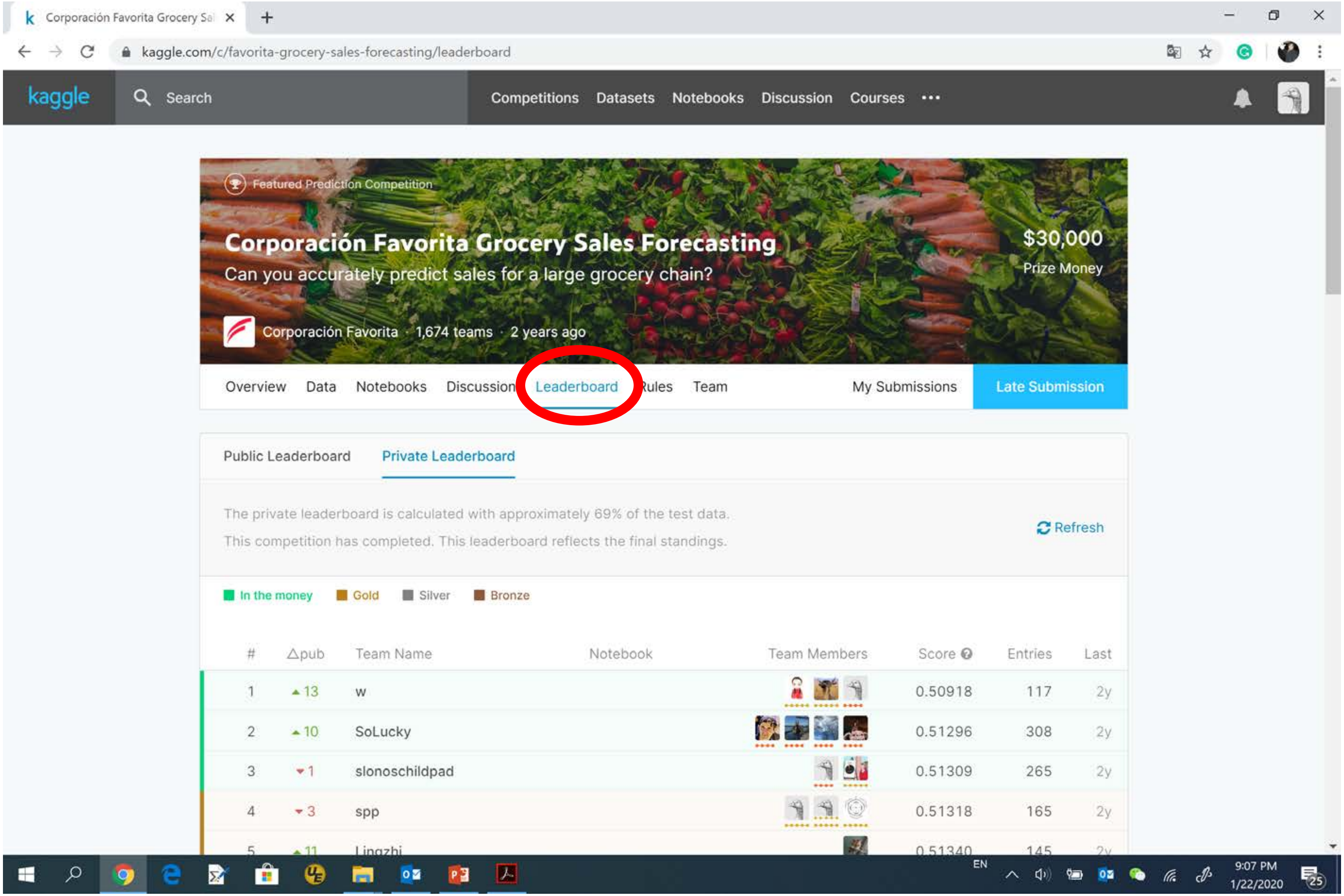
1 Competition

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 Sal
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

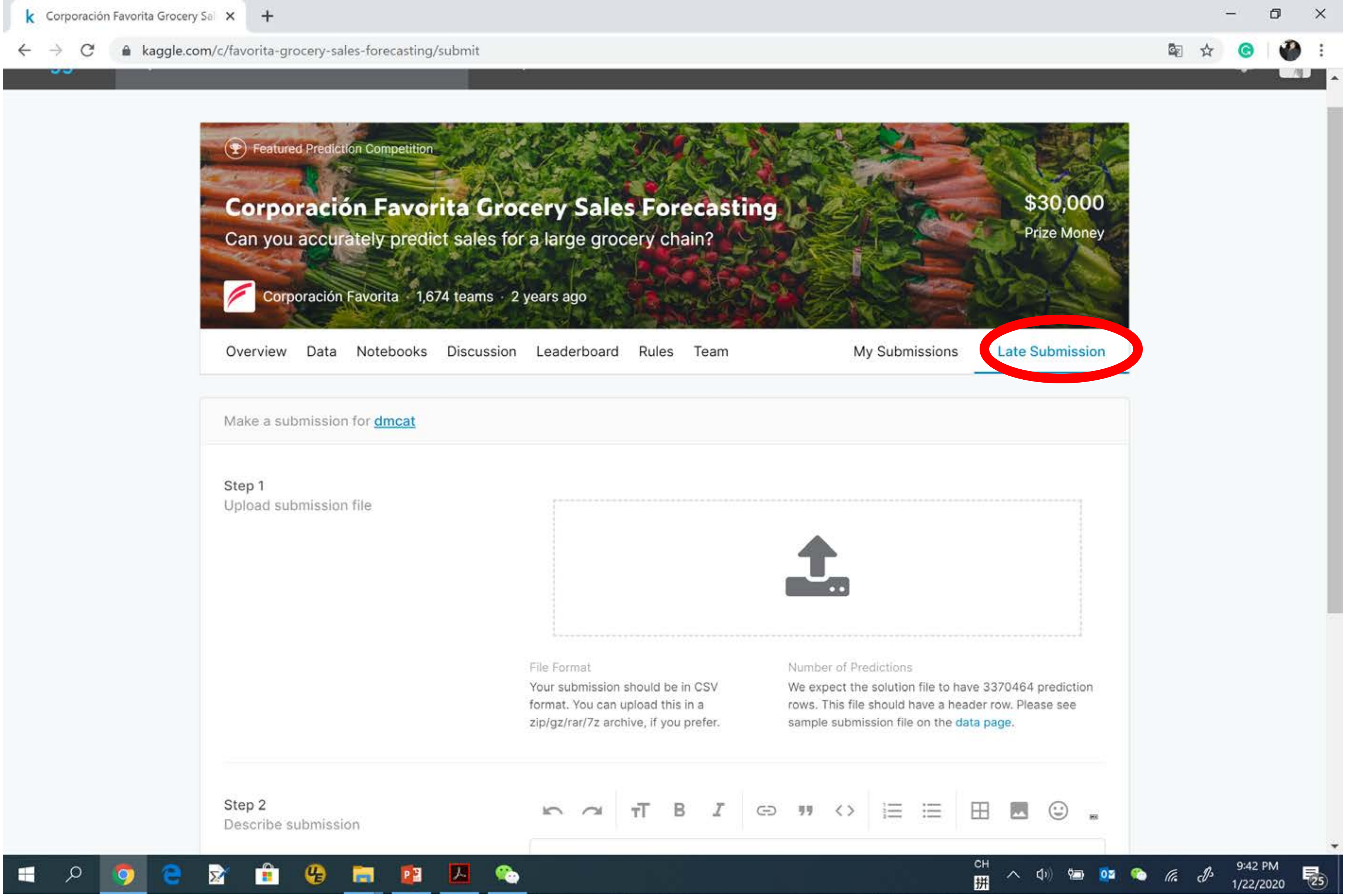
Corporación Favorita · 1,674 teams · 2 years ago

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

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

🟩 In the money 🟧 Gold ⬜ Silver 🟫 Bronze

#	△pub	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


Corporación Favorita Grocery Sa
kaggle.com/c/favorita-grocery-sales-forecasting/submit
CH
9:42 PM
1/22/2020

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

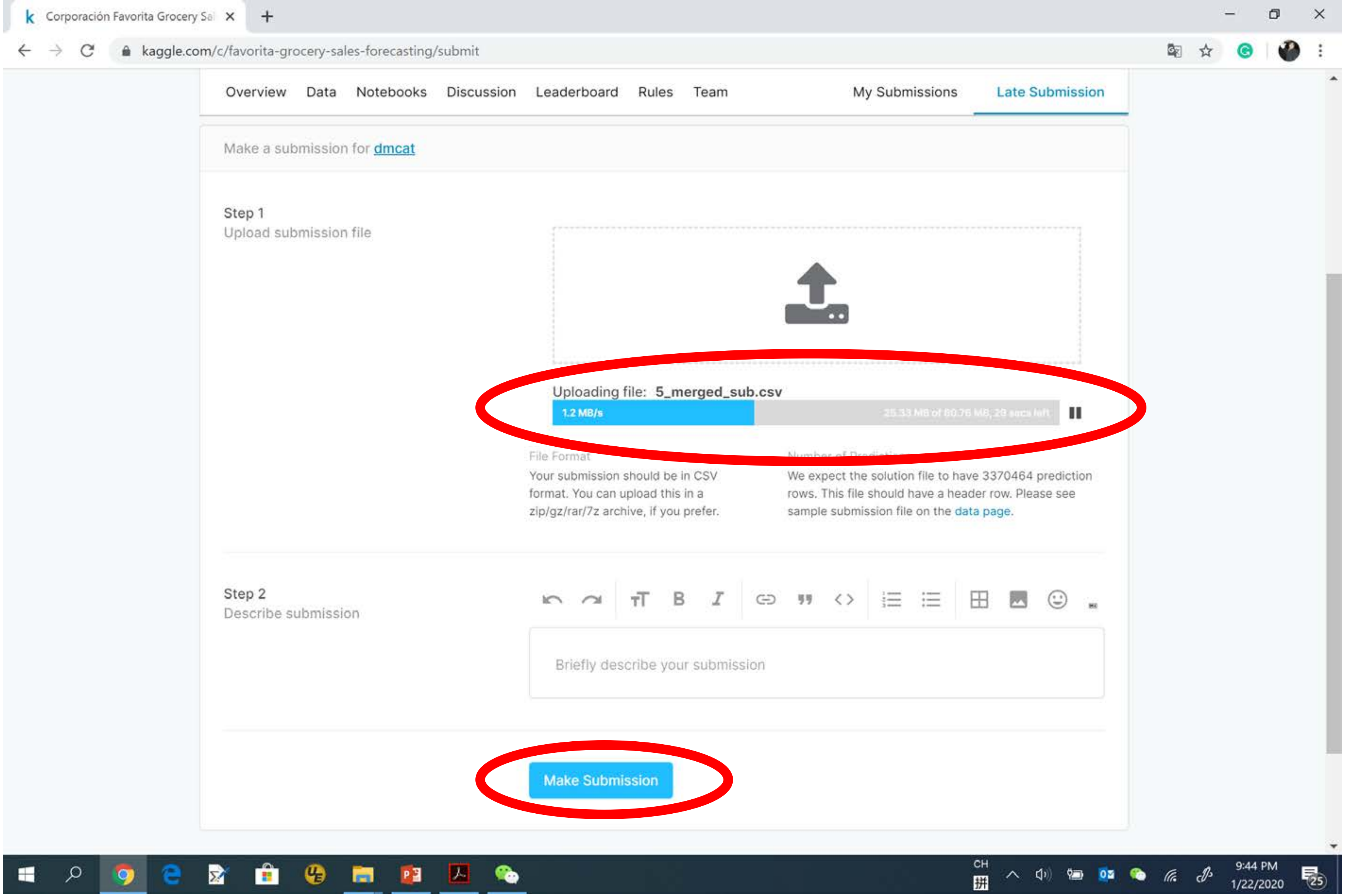
Make a submission for dmcat

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



Make a submission for dmcat

Step 1
Upload submission file

Uploading file: **5_merged_sub.csv**

1.2 MB/s 25.33 MB of 80.76 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.

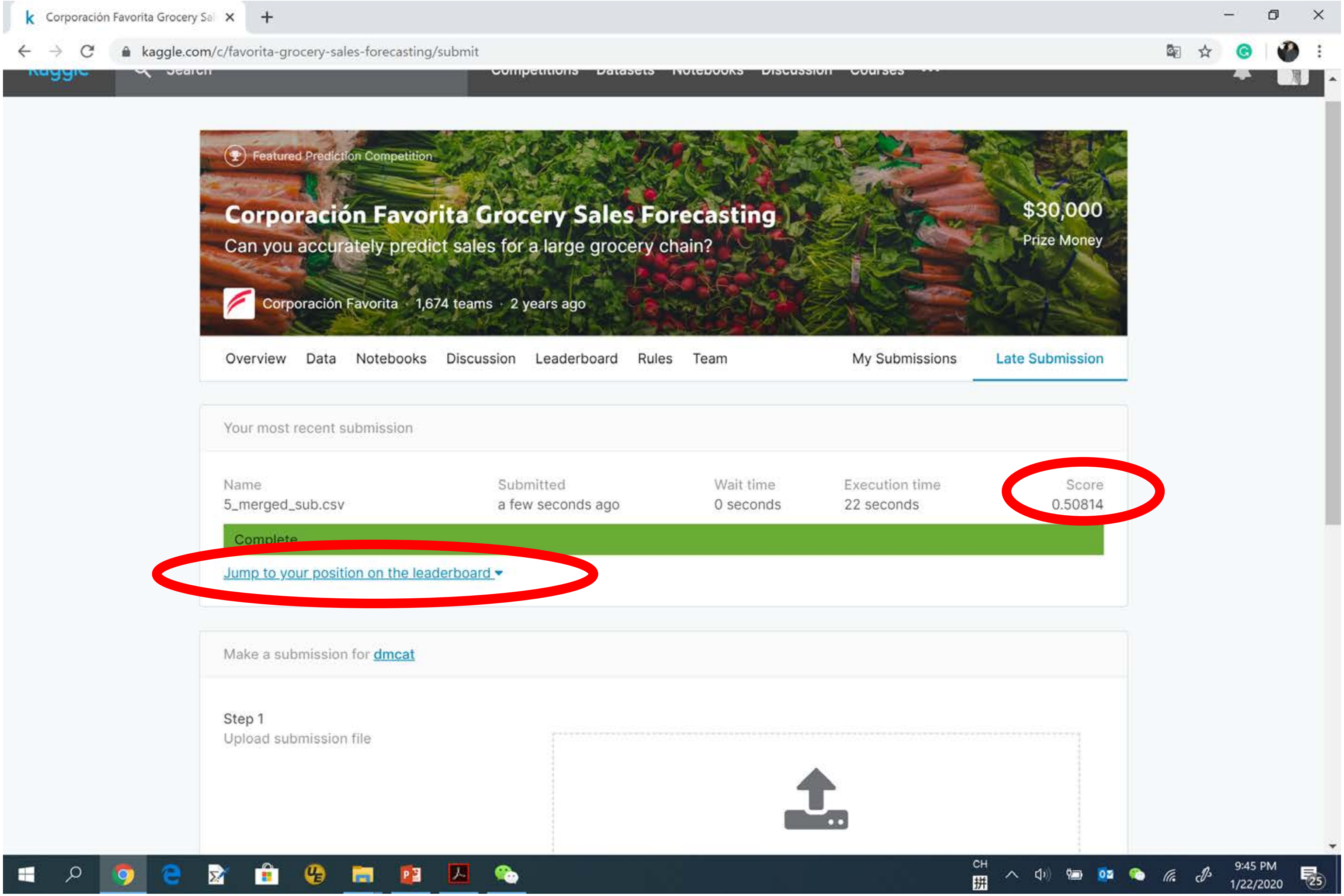
Number of Predictions

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

Briefly describe your submission

Make Submission


Corporación Favorita Grocery Sal
kaggle.com/c/favorita-grocery-sales-forecasting/submit
kaggle
Search
Competitions
Datasets
Notebooks
Discussion
Courses
9:45 PM
1/22/2020

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

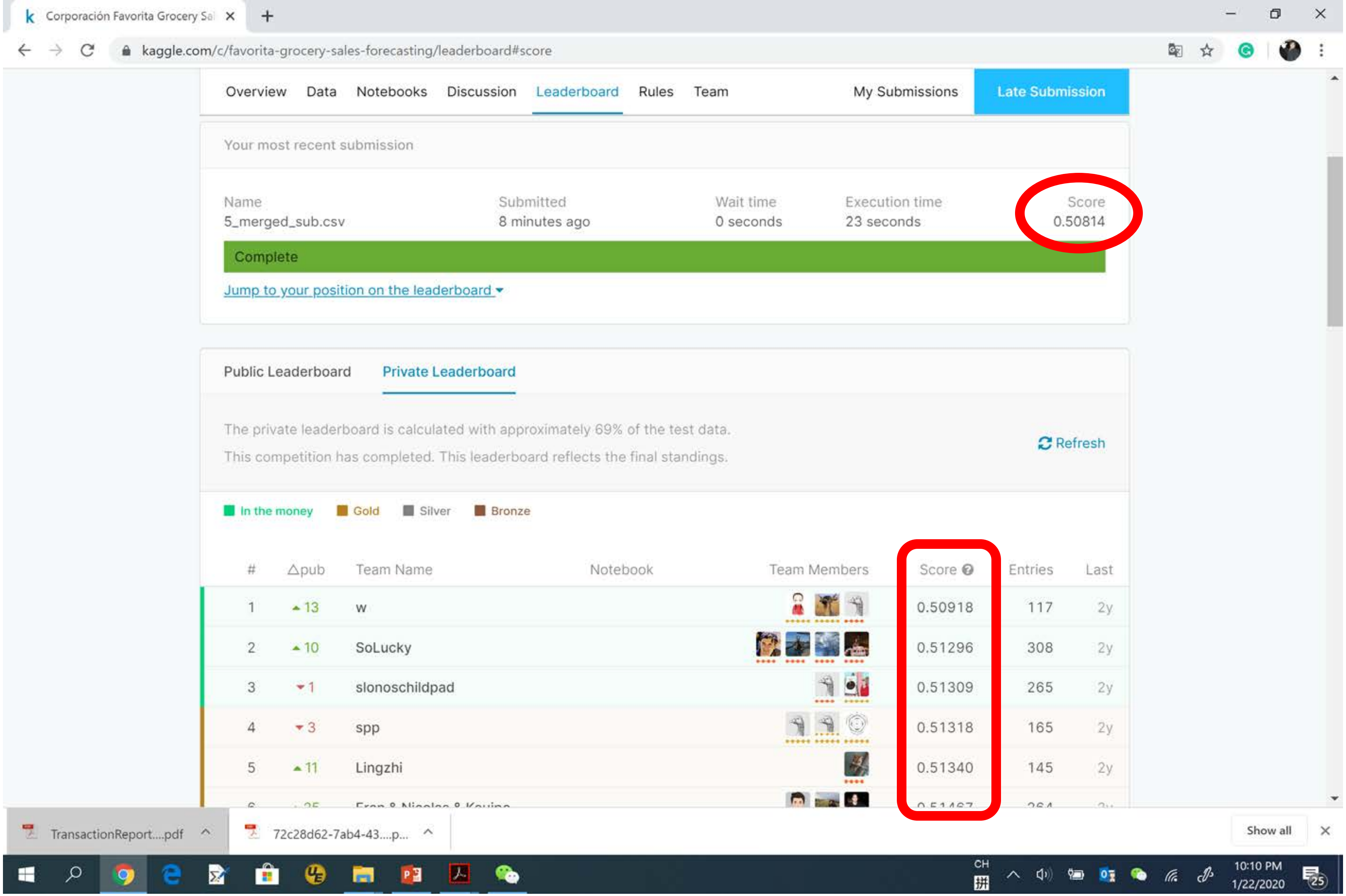
Name	Submitted	Wait time	Execution time	Score
5_merged_sub.csv	a few seconds ago	0 seconds	22 seconds	0.50814

Complete

Jump to your position on the leaderboard ▾

Make a submission for dmcat

Step 1
Upload submission file


Corporación Favorita Grocery Sal
kaggle.com/c/favorita-grocery-sales-forecasting/leaderboard#score
TransactionReport....pdf
72c28d62-7ab4-43....p...
Show all
CH
拼
10:10 PM
1/22/2020

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

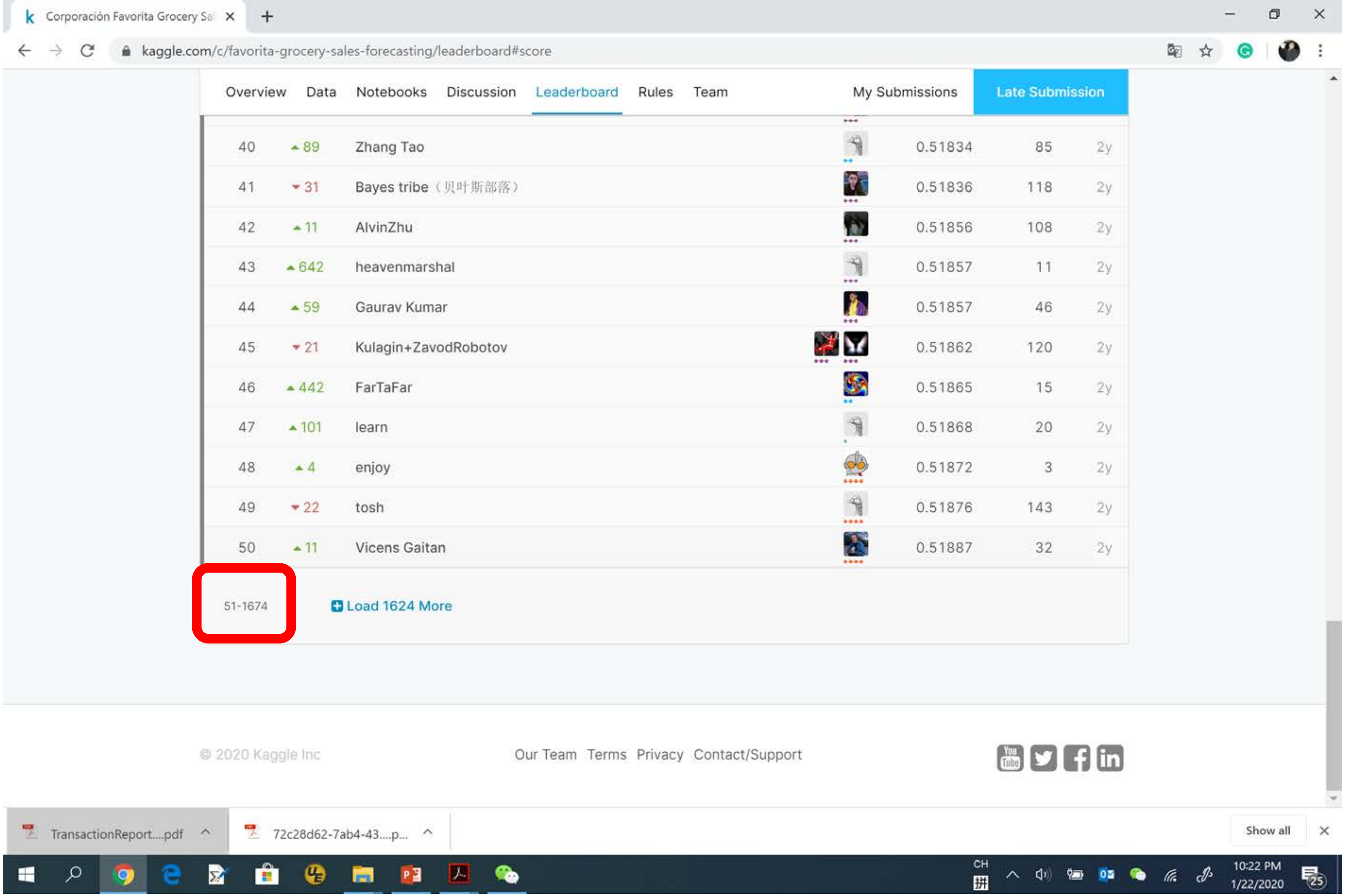
The private leaderboard is calculated with approximately 69% of the test data.

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	364	2y

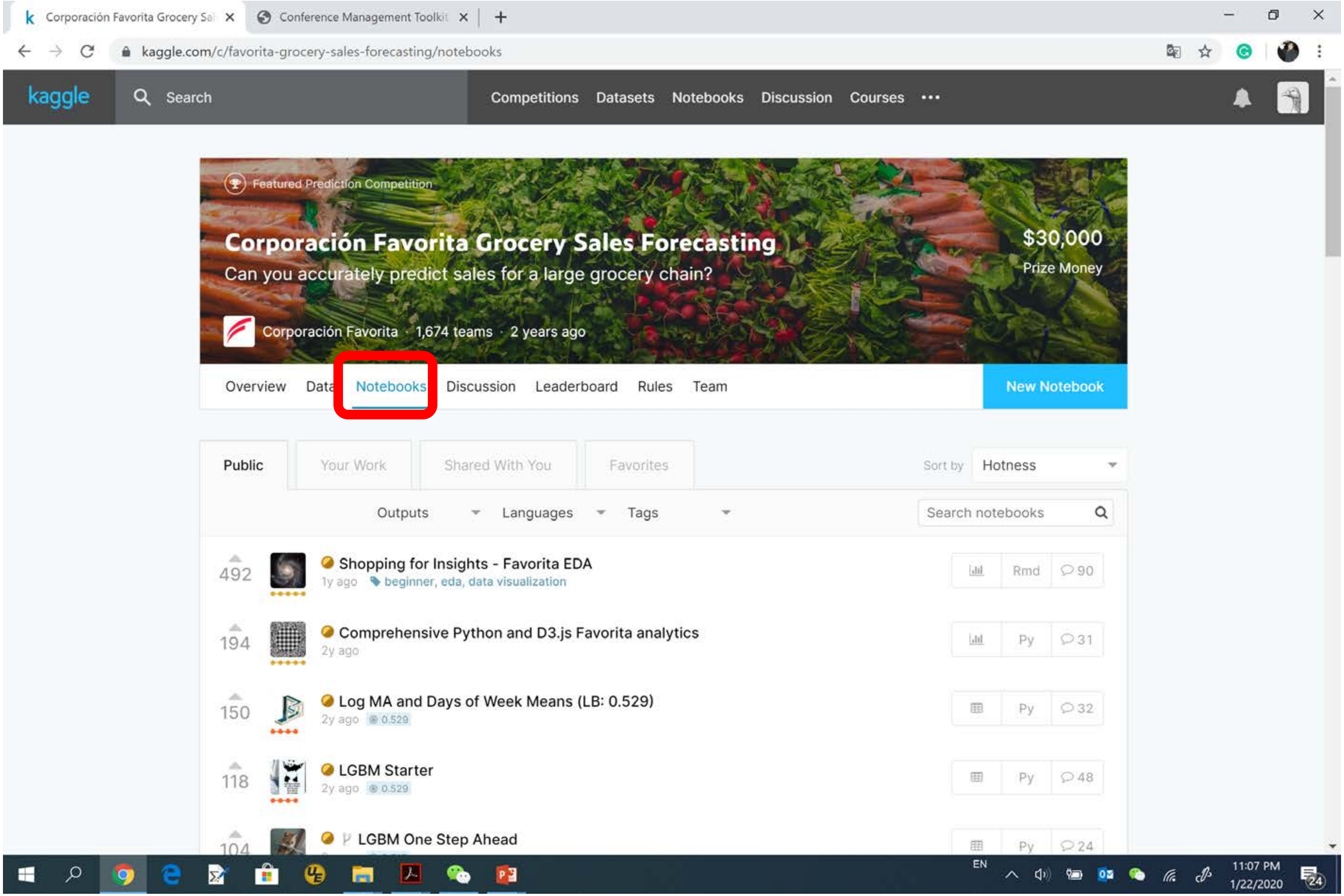


40	▲ 89	Zhang Tao	0.51834	85	2y
41	▼ 31	Bayes tribe（贝叶斯部落）	0.51836	118	2y
42	▲ 11	AlvinZhu	0.51856	108	2y
43	▲ 642	heavenmarshal	0.51857	11	2y
44	▲ 59	Gaurav Kumar	0.51857	46	2y
45	▼ 21	Kulagin+ZavodRobotov	0.51862	120	2y
46	▲ 442	FarTaFar	0.51865	15	2y
47	▲ 101	learn	0.51868	20	2y
48	▲ 4	enjoy	0.51872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

51-1674 Load 1624 More

Assessments – Kaggle (cont.)

- **Solution Novelty:** as on Kaggle.com, most participants or winners may discuss their solutions on the forums of the specific competitions.
  - 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


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

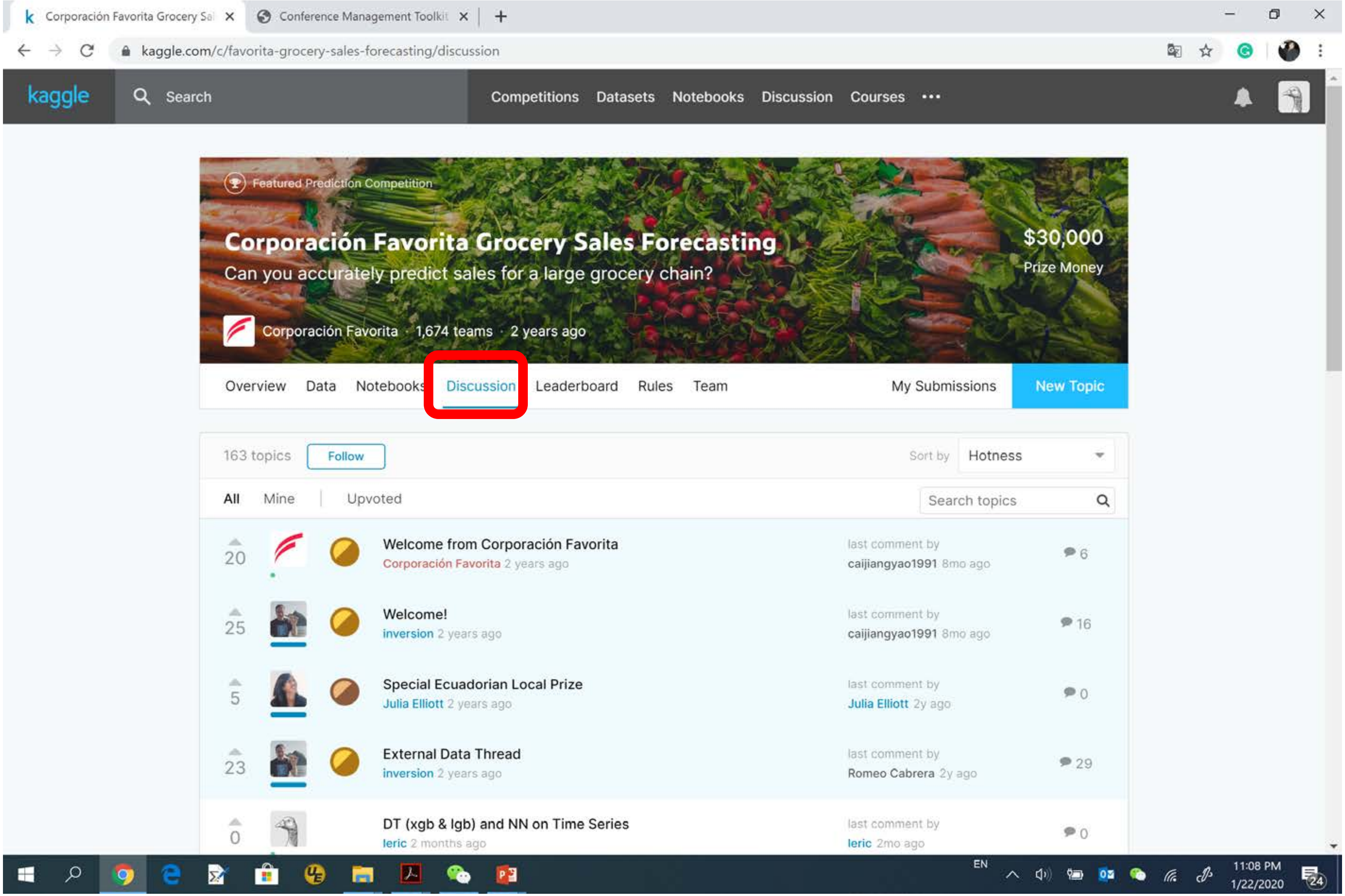
492 Shopping for Insights - Favorita EDA
1y ago beginner, eda, data visualization — Rmd 90

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



Corporación Favorita Grocery Sal | 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?

Corporación Favorita · 1,674 teams · 2 years ago

$30,000
Prize Money

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

163 topics | Follow | Sort by Hotness

All | Mine | Upvoted | Search topics

Votes	Topic	Last comment	Comments
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

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. In your report, you need to conduct experiments to 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