



IEEE COMPUTER SOCIETY
TCCC

Technical Community on Computer Communications

IEEE CyberC 2022 Data Analytics Competition Call for Participation

Data Science is all about the processes and methods to access and analyze data to gain insights for informed decision making. To promote the awareness of data analytic technology, IEEE CyberC 2022 is organizing a **Data Analytics Competition**. The competition will begin now and a report with codes need to be submitted by **December 9, 2022**. The results will be announced during CyberC 2022, December 15-16, 2022, Jiangsu, China.

Competition Dataset and Instructions

The competition requires participants to analyze and gain insight from the dataset provided. The dataset **data.zip** consists of the training set *house_price_train.csv* and test set *house_price_test.csv*. The dataset and instructions could be downloaded from the CyberC 2022 Web Page <http://www.cyberc.org/competition>

Participation

Participation is open to conference registered attendees only who can register to the conference via the following link:

- CyberC 2022: Registration Form: <http://cyberc.org/Home/Registration>
- Registration has to done before December 9, 2022 and you can submit your data analytic report and code with the registration form.
- One registration covers a team up to 3 students.
- Registration form should be submitted to Dr. Bin Xie at Papers@CyberC.org

You may participate individually or as part of a team of at most 3 members (e.g., 1 or 2 students and 1 instructor (normally your professor) but instructor is not necessary). While submitting your registration form to Papers@cyberc.org, you must submit your data analytic report by **December 9, 2022** by sending an email to Competition Chair, Dr. Di Zhang (Email: Di.Zhang@xjtlu.edu.cn), consisting of

- 1) Your report with codes; **and**
- 2) The following information:
 - Name or names of members in a team
 - Affiliation (e.g., name of university or company)
 - Email address(es)
 - Do you have a paper accepted by CyberC 2022?

Evaluation and Awards

Using the dataset, your reports could include the method your used, the performance (e.g., time efficiency), the prediction results, etc. Your report will be evaluated based on a **reproducible and stable** result on test set. Your performance, report and source code will be checked in two steps:

- 1) Rank the prediction accuracy on the test set claimed in all reports and select the three highest for candidates;
- 2) Check the reproducibility and stability of the results, including re-running the source code, modifying the random number seed, checking whether external data sources are referenced, etc. For answers with unstable accuracy, the judges will lower their rankings according to the results of the rerun, and disclose the related reasons. Those who were not in the top three before may be substituted. The accuracy ranking after verification is the final ranking of the competition.

Golden, Silver, and Bronze prizes with **an honorable Certificate** and Cash of **5000 RMB, 4000 RMB and 3000 RMB** respectively will be awarded to the first, second and third winners, judged by a competition panel. **The team instructor (e.g., a professor) would be honorably invited to take a leading role in CyberC 2023.** The winning team will present their work during the prize-giving ceremony on Friday, December 16, 2022.

Timeline

- Now – Competition starts
- December 9, 2022 – Deadline of report submission to competition
- Friday, December 16, 2022 – Prize-giving ceremony and presentations by winning teams

Appendix: IEEE CyberC 2022 Data Analytics Competition Description

Background

Buying a home is one of the most important investment decisions for any family. However, the number of houses on the market is huge, and it is obviously inefficient or even impossible to inspect them one by one through manual visits. This makes home buying behavior a certain contingency. If a model could be trained to scan properties for sale and filter out properties with the greatest potential for future appreciation, the accuracy of investment decisions would be greatly improved.

Data

We have collected a batch of real estate transaction records from large cities in China. It contains data from 34 cities from 2015 to 2018, totaling 625,968 records. Each data record (i.e., a row) include some fields as shown in the figure below, where 'Price' is the target to be predicted.

Id	Time	City	District	Street	Community	Lon	Lat	#Floors	Floor	#Rooms	#Halls	Orient	Area	Price
0	2012/1/1	Beijing	Xi Cheng	Mu Xi Di	Fu Xing	116.3439	39.91095	12	Middle		2	1 S	55	40909
1	2012/1/2	Beijing	Tong Zhot	Bei Guan	Tian Ci	116.659	39.92804	6	Middle		2	1 NS	93.59	14745
2	2012/1/3	Beijing	Tong Zhot	Bei Guan	Xi Lu Yu	116.659	39.92804	6	Low		2	2 NS	85.32	11252
3	2012/1/3	Beijing	Xi Cheng	Yue Tan	Xiao Si	116.3514	39.92059	6	Middle		1	1 E	41.3	33414
4	2012/1/3	Beijing	Zhao Yan	Shao Yao	Shao Yao	116.449	39.98481	6	High		2	1 NS	64.25	25370
5	2012/1/4	Beijing	Shi Jing	Lu Gu	Rong Jin	116.232	39.90439	26	High		2	1 S	89.61	17855
6	2012/1/5	Beijing	Hai Dian	Xiao Xi	Zhi Qian	116.3734	39.9587	6	High		1	1 S	49.6	26210
7	2012/1/5	Beijing	Zhao Yan	Gan Lu	Yi Kang Jia	116.5216	39.92183	18	High		2	1 NS	92.86	17015
8	2012/1/6	Beijing	Zhao Yan	Hua Wei	Nan Xin	116.4592	39.88173	24	Low		3	1 NS	91.95	17401
9	2012/1/8	Beijing	Feng Tai	Da Hong	Nan De	116.4136	39.84529	6	High		1	1 SE	44.53	7186
10	2012/1/8	Beijing	Feng Tai	Liu Li	Q Tai Ping	116.3169	39.89169	24	Low		1	1 E	49.31	24336
11	2012/1/8	Beijing	Feng Tai	Qi Li	Zh Wang Yuan	116.2964	39.87041	24	High		1	1 S	68.99	21742

The meanings of these fields are:

- 'Id': unique identifier
- 'Time': the date the transaction took place

- 'City': the city where the transaction took place
- 'District': District
- 'Street': Neighborhood
- 'Community': Community
- 'Lon': longitude
- 'Lat': latitude
- '#Floors': number of floors of the entire building
- 'Floors': approximate floor location
- '#Rooms': number of rooms
- '#Halls': number of halls
- 'Orient': House orientation code. Containing 'N' means the house has north facing windows; containing 'S' means the house has south facing windows; containing 'E' means the house has east facing windows; containing 'W' means the house has west facing windows
- 'Area': house area
- 'Price': transaction price (CNY per square meter, i.e., ¥/m²)

The dataset is divided into training set (553099 data records) and test set (72870 data records) according to before and after 2018-04-01. Both training and test sets contain the same fields. They are all saved in csv format. Note that, to be fair, you cannot use third-party data sources when training the model.

Metrics

The evaluation standard of prediction accuracy is: the mean absolute error (MAE) between the logarithm of the predicted transaction price(y_i) and the logarithm of the actual value (\hat{y}_i), that is,

$$\frac{\sum_{i=1}^N |\ln y_i - \ln \hat{y}_i|}{N},$$

where N is the total number of records in test set. The logarithm is taken to balance the impact of high and low housing price areas on the forecast results. (Note: Do not take the logarithm in the uploaded test results! Just keep the original unit.)

The Organizing Committee will use the MAE as the basis for ranking. The smaller the MAE, the higher the ranking.

Submission

Each team is to submit a report, along with the associated source code. In the report, the following information should be included:

- Title
- The list of team members including their names, affiliations, email addresses, and phone numbers
- Your method and/or data preprocessing
- Feature construction (if any)
- Model design, including the optimization goal of the problem and the training method of the model
- Result visualization, a scatter plot comparing the actual and predicted values in the test set with logarithmic coordinates
- The running method of the program, the running environment required by the source code, and the entry and parameters (if any) of the program

The source code should be complete and can be run independently. Third-party libraries, which are

publicly available online, do not have to be included in the source code.