**SI 670 Fall 2020 Kaggle Competition**

Please form teams of up to 4 students. You should not collaborate with other teams, but you can talk to the GSIs if you are really stuck. Please record you teams here: https://docs.google.com/spreadsheets/d/1TDm\_bwdSnENU8AMzxjvgqP2Fo3Poi3nRGr8dP1iIbqs/edit?usp=drive\_web&ouid=109090332886082450282

For those who filled out the team survey and asked to be assigned to a team, you should find your name and your team members in the Google Sheet above. Please feel free to let us know if we missed you.

Give your team a fun and exciting name for the competition. You only need one submission for the team. Everyone on the team will get the same score.

**Scoring**

There are two tasks, but they will be scored together. It will be scored out of 100 points.

* You shall receive 40 points for a successful submission.
* You shall receive 20 additional points for attaining benchmark #1.
* You shall receive 10 additional points for attaining benchmark #2.
* You shall receive 30 \* 2 / log2(2 + rank) additional “ranking points”.
* The top five teams after the first evaluation will receive an additional 5 points.

Note that, if you attain both benchmarks, you are guaranteed a score of 82. The winning team will receive 108 points.

The competition shall start Tuesday 10/20.

First evaluation is Monday 10/26 at 11:59pm (only used for 5 bonus points)

Final evaluation is Monday 11/2 at 11:59 pm

**Context**

Lake monitoring provides important information for environment protection and pollution identification, such as temperature/thermal monitoring. In this Kaggle task, we will provide thermal sensor data for multiple lakes and ask you to predict/estimate the temperature at a certain depth for certain lakes.

**Content**

We provide Lake Trout’s data during the period of Apr 20th, 2012 - Apr 19th, 2018.

For your reference, we also provide the data of six other lakes (ie., Bear Head, Carlos, Elk, Pearl, Shaokotan, White Iron). Please note that these datasets may vary in terms of the time window (yes, this is how real-world datasets look like!)

For each lake, there will be multiple sensors at different depths, and we will provide lake\_id, time, depth and the corresponding temperature at this depth and time point. For example, for the Shaokotan Lake, you will see the following columns, with LakeId representing the lake identifier, Date\_time representing the time stamp, Depth\_m representing the depth in meters, and Water\_Temp\_C representing the temperature in Celsius. Please note that the number of sensors are different across different lakes.

**Task**

Predicting/estimating the temperature of Trout Lake at depth 10.5 m during Apr 20th, 2018 - Apr 19th, 2019. Please feel free to use part of/all the data we provide.

Benchmark 1: MSE < 0.065

Benchmark 2: MSE < 0.032

**Files for first task**

For the first task, we will provide you a file Trout\_training.csv under the Trout folder which contains all the sensors in the Trout Lake as training. Besides, we will give you another Trout\_testing\_features.csv under the Trout folder which contains all other sensors’ temperature (except the depth 10.5’s temperature).

We hope you return a file named Trout\_prediction.csv, the dataset you return and upload should have two columns: Date\_Time and Water\_Temp\_C, with a shape of 8,760 \* 2. (Reason for 8,760: 8,760 = 365 days \* 24 sensorings/day) You are supposed to predict every hour’s temperature of depth-10.5 sensor during Apr 20th, 2018 - Apr 19th, 2019.

**Link**

https://www.kaggle.com/c/si670fall2020/overview

**Extra Credit Task**

There is another Kaggle competition you can compete in for extra credit. Teams will receive 30/rank number of extra credit points where rank is their rank at the final evaluation time. So the first place team receives 30 points, the second place team 15 points, etc. Additionally, the top three teams at the first evaluation mark will receive 5 bonus points. The points will be added to your above Kaggle score.

The extra credit task is predicting the temperature of Greenwood Lake at depth [1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5, 10.5, 11.5, 12.5, 14.5, 16.5, 18.5, 20.5, 22.5, 24.5, 26.5, 28.5] during Dec 2017.

**Files for Extra Credit**

For this task, we will only provide you a file Greenwood\_training.csv under the Greenwood folder which contains all the sensors in the Greenwood Lake as training.

We hope you return a file named Greenwood\_prediction.csv, the dataset you return and upload should have two columns: Date\_Time\_with\_Depth, Water\_Temp\_C, with a shape of 14880 \* 3 (14880 = 31 days \* 24 hours \* 20 sensors). You are supposed to predict every hour’s temperature of every sensors in 20 depths during Dec 2017.

Note that because of the limitation of Kaggle, we request you to combine the ‘Date\_Time’ and ‘Depth’ by a ‘-’ in the prediction file. That means if you are predicting the temperature of 2018-04-20 00:00:00 with depth 1.5 you are supposed to have the field ‘Date\_Time\_with\_Depth’ with value ‘2018-04-20 00:00:00-1.5’ and followed by your predicted value.

**Extra Files**

We also provide several other 6 lakes temperature data for your reference. You could apply them as extra features to help you make the prediction. For example, there may exist some relationship between the same day temperature of a lake with another lake. You can download them on Canvas.

**Due**

The competition shall start Tuesday 10/20.

First evaluation is Monday 11/2 at 11:59pm (Used only for 5 bonus points).

Final evaluation is Wednesday 12/2 at 11:59 pm

**Link**

https://www.kaggle.com/c/si670extra/overview

**Acknowledgements**

We thank the Minnesota Department of Natural Resources Sentinel Lake project for providing us with the data <https://www.dnr.state.mn.us/fisheries/slice/index.html>