COMP5434 BIG DATA Computing Group Project

Report & code submit: **23:59 16th April 2023**

Presentation: **18:30~21:30 27th March 2023 (For Monday class)**

15:30~18:30 12th April 2023 (For Wednesday class)

Introduction

Suppose there is a real estate company based in the United States, specializing in the sale of houses in various locations. In the real estate market, house prices are influenced by a variety of factors, including location, size, noise level, air conditions, etc. To help real estate investors make informed decisions, the company regularly releases information on house sales in different areas. By providing comprehensive sales data, the company empowers investors to design accurate and effective house price prediction systems. By analyzing sales data, investors can identify latent patterns and develop predictive models that are useful to make data-driven decisions on house transactions.

Dataset

The dataset is described as follows:

1. *Train_Data.csv* contains 4000 samples of estate basic information, and the target variable is the **Total Cost** status:
   - Date – The date when the sample is collected.
   - Number of rooms – number of rooms in the house.
   - Security level of the community – the higher the safer.
   - Residence space – square feet area of the living rooms.
   - Building space – square feet area of the whole building.
   - Noise level – the lower the value, the greater the noise.
   - Waterfront – If the house has water front or not.
   - View – Number of viewings before the house is sold.
   - Air quality level – the higher the value, the better the air quality.
   - Aboveground space – square feet area of the above house.
   - Basement space – square feet area of the basement in the house.
   - Building year – the year in which the house was built.
   - Decoration year – the year in which the house was decorated.
   - District – the address of the house.
   - City – the city in which the house is located.

- Zip code – the zip code of the house.
- Region – the region of the house.
- Exchange rate – when the house is sold, the exchange rate between the US dollar and the Hong Kong dollar.
- Unit price of residence space – the unit price of residence space (US dollars).
- Unit price of building space – the unit price of building space (US dollars).
- **Total cost – the total price of residence and building space (Hong Kong dollars).**

2. *Test_Data.csv* contains 400 samples of estate basic information and the **total cost** is unknown.

Task

This project contains three tasks:

1. Suppose you are an employee of the company, calculate the total cost (in **Hong Kong dollars**) of each house (**including residence and building costs**) based on the data in *Train_Data.csv*. It is **required to use MapReduce** to handle the calculation. Suppose there are **5 mappers** and **2 reducers**.

   Hint:

   (1) The total area of the house includes two parts: residence and building.

   (2) The *Train_Data.csv* can be distributed equally among 5 mappers.

   (3) The calculation rule of the total cost is:

   *(The unit price of residence space * Residence space + Unit price of building space * Building space) * Exchange rate = Total cost.*

   (4) Here is an example to calculate the total cost of the first sample in *Train_Data.csv* :

   *(11.88640925 * 2820 + 0.977028065 * 67518) * 6.784829586=675000 HKD*

2. Suppose you are a real estate investor who **does not know the unit price of the house (including both residence space and building space)**, you need to remove columns **Unit price of residence space** and **Unit price of building space** from *Train_Data.csv*, and design a machine/deep learning model that predicts the **total cost** of each house. Then, you need to predict the **price range of the total cost** based on the *Test_Data.csv.*

   Note: you are only required to predict the **price range (the class)** of the **total cost** for each sample in *Test_Data.csv*. The total costs are divided into **four** classes:

   ➢ **1**: it means the total cost is greater than or equal to 0HKD and less than 300000HKD (i.e., 0 <= total cost < 300000).

   ➢ **2**: it means the total cost is greater than or equal to 300000HKD and less than 500000HKD (i.e., 300000 <= total cost < 500000).

- **3**: it means the total cost is greater than or equal to 500000HKD and less than 700000HKD (i.e., 500000 <= total cost < 700000).
- **4**: it means the total cost is greater than or equal to 700000HKD (i.e., 700000 <= total cost).

Hint: For each sample, its label for modeling training can be obtained from the total cost calculated in Task 1.

3. Suppose there are **four** real estate companies, each of which owns a part of the real estate transaction data, and these data cannot be shared among the companies directly, due to the requirements of privacy protection. Still, you do not know the **unit price of the house (including both residence space and building space)**. Therefore, instead of training the model based on the entire *Train_Data.csv* in Task 2, you need to split the *Train_Data.csv* into **four** parts and apply the Federated Learning (FL) algorithm to train the model. After that, you also need to make comparisons between the FL model and the model obtained in Task 2. Still, you need to predict the **price range of the total cost** based on the *Test_Data.csv*.

   Hint: (1) You could split the whole training dataset *Train_Data.csv* into 4 different parts, which follow the **independent and identical distribution**. (2) The unit price of the house (including both residence space and building space) is still unavailable. For each sample, its label for modeling training can be obtained from the total cost calculated in Task 1.

Project Grading

The project is 30% of the total subject assessment (optional: extra 5% for Bonus). Three deliverables are required to be finished in this project:

- Final report (10%)
  - The maximum word limit for the final report is 4000 words.
- Project presentation (5%)
  - Presentation maximum 10 minutes and presentation slides maximum 15 pages.
- Source code (15%)
  - Any programming language.
- Bonus (Extra 5%)

Report & Presentation

The report and presentation may include but not limited to:

- Group information and member duty
- Introduction
- Data preprocessing/analytics
- Model design and implementation
- Framework of federated learning
- Performance evaluation and discussions
- Conclusion and future work
- Reference

Notes: Each group needs to declare the group information on the first page of the report, including group ID, class (Monday or Wednesday), all group members' student IDs, all group members' English and Chinese names, and all group members' emails. Also, you should clarify the duty and contribution of each member. Please rename the report as: ***group_leader_student_ID_report.pdf***.

Grading Criteria for Project Source Code

We will grade your code based on the following 4 aspects:

- The program needs to be clearly annotated and a detailed Readme file should be uploaded.
- Task 1: we will check the final results and the logic you implement the MapReduce functions.
- Task 2: we will compare your predicted results in *Test_Data.csv* with the Ground-truth values, and the performance evaluation is based on the Top1-Accuracy.
- Task 3: we will check how you split the *Train_Data.csv* and implement the FL training procedure, and we will evaluate the performance of your FL model.

Bonus (The bonus is **optional** for answering but can bring **extra scores** to you**)**

1. Improve your basic design by using any methodology to optimize the model performance in Task 2 and/or Task 3.
2. Discuss the non-iid's influence on FL model performance caused by data splitting, and explore potential methods to reduce the influence of non-iid data.

Note: we will judge whether you can get extra scores based on the Bonus files (i.e., ***group_leader_student_ID_bonus1.pdf*** and ***group_leader_student_ID_bonus2.pdf***) you submitted.

Submission Format

1. For task 1, first, insert your results into *Train_Data.csv*. Then, package your *Train_Data.csv* and MapReduce source code as a zip file before submission.
   Please rename it as ***group_leader_student_ID_task1.zip*. (e.g., *220XXXXXXG_task1.zip*)**.

2. For task 2, use your designed model to predict the price range of total cost, and fill your results in *Test_Data.csv*. Then, package your *Test_Data.csv* and the source code of model training as a zip file before submission.
   Please rename it as ***group_leader_student_ID_task2.zip*. (e.g., *220XXXXXXG_task2.zip*).**

3. For task 3, first, use your FL model to predict the price range of total cost, and fill your results in *Test_Data.csv*. Then, draw a figure that show the performance comparison between the FL model and the model obtained in Task 2 (e.g., training loss, test accuracy). So, package your *Test_Data.csv*, the figure, and the FL source code as a zip file before submission.
   Please rename it as ***group_leader_student_ID_task3.zip*. (e.g., *220XXXXXXG_task3.zip*).**

4. For bonus 1, please provide any possible supplementary materials (e.g., numerical analysis, figures) to show your methodology and model performance. Please put these materials into a single PDF file before submission and rename it as ***group_leader_student_ID_bonus1.pdf*. (e.g., *220XXXXXXG _bonus1.pdf*)**.

5. For bonus 2, please provide any possible supplementary materials (e.g., numerical analysis, figures) to discuss the non-iid influence and the potential methods to reduce this influence. Please put these materials into a single PDF file before submission and rename it as ***group_leader_student_ID_bonus2.pdf*. (e.g., *220XXXXXXG _bonus2.pdf*)**.

Notes: Each group only needs to submit **one zip** file by the group leader. All project documents for the above three tasks and two bonuses, i.e.,

group_leader_student_ID_report.pdf,
group_leader_student_ID_task1.zip,
group_leader_student_ID_task2.zip,
group_leader_student_ID_task3.zip,
group_leader_student_ID_bonus1.pdf,
group_leader_student_ID_bonus2.pdf,

should be packed into one zip file, rename it as ***group_leader_student_ID_project.zip* (e.g., *220XXXXXXG_project.zip*)**, and then submit the zip to Blackboard.