COMP4107 Big Data Individual Report

Ran Ji 20217337
scyrj1@nottingham.edu.cn

May 6, 2023

1 Participation Table

Student name	Mark (out of 10)	Marking justification
Jiarui LI	10	Actively participate in discussions, complete the assigned tasks on time, communicate in time when problems arise, and cooperate smoothly.
Ran JI	10	Actively participate in discussions, complete the assigned tasks on time, communicate in time when problems arise, and cooperate smoothly.
Yik Lun Yelan LAU	10	Actively participate in discussions, complete the assigned tasks on time, communicate in time when problems arise, and cooperate smoothly.

2 Individual Role and Contributions

2.1 Team work

Although I am not a team leader, I am like a linker throughout the project. At first, my other two team members were not very familiar with each other, so I needed to communicate with them in time to work out the execution plan and timetable for the entire project. At the same time, propose the time and place of the meeting, negotiate with the peers, and inform the team leader to send a formal discussion email.

2.2 Methods

According to code contributions, I am in charge of the computation part including K-Fold cross validation and evaluation.

2.2.1 K-Fold Cross-validation

As we mentioned in our group report, K-Fold cross-validation can be used to improve results influenced by data chosen randomly. We divided code blocks into different parts and we can connect them for final running. Each method or model, such as Linear Regression, Gradient Boosting, etc, which we talked about in the group report, is added with the K-fold block for cross-validation.

The implementation of K-Fold cross validation is included in “regressor.py” file, “RegressionModel” class and models can use this tool for k-fold cross validation. In addition, the implementation used ”CrossValidator” function, ”pyspark.ml.tuning” package.

2.2.2 Evaluation

As mentioned in the group report and literature review, root mean squared error (RMSE) and R^2 scores are criteria for our regression task. After training each model, I organized the evaluation results of the test and training.

Same as K-Fold Cross-validation part, these two evaluation parts are implemented in "regressor.py" file, "RegressionModel" class and by "pyspark.ml.evaluation" package, "RegressionEvaluator" function.

2.3 Report

According to the group report, I am in charge of Literature review, Evaluation methods and Conclusion sections.

- **Literature Review**: In the literature review, I first find out several related works about flight delay prediction. And based on these papers, I searched for relevant literature and summarized their work, methods, data sets used, implementation plans, and results. In addition, I provided them to my group members to discuss our plans and complete the literature review part in the group report.
- **Evaluation Methodology**: As mentioned in the previous section, I am in charge of the evaluation part, and the evaluation methods and related works are completed in the group report.
- **Conclusion**: After the completion of this project, I summarized our work in the last discussion and organized them in the group report.

3 Reflection and Conclusion

3.1 Findings

There are my individual findings in this project, including the task, methods and the project background.

- Project Findings: First, the dataset of the project only includes the time information of flight departure. Although it can be used for predicting the arrival time, the better performance shown in related works includes weather conditions and other measurement data. Second, the amount of data I processed in my study life is small, which can be processed by SQL, but Spark is very useful and efficient when processing a large amount of data. Finally, there are huge amounts of data in our life so big data is a common skill to learn for future work and career.
- Background Findings: I have never realized the emergency of flight delay. After related works and a literature review, I understand the huge amount of cost of flight delays each year. Also, I learned about some flight parameters like wheels-on time and some delay reasons and classes.

3.2 Group Work Reflection

The project is a group work based project and I have learned some experience during the work.

3.2.1 Time Management

Firstly, group members all have their own modules and other coursework to do. In each discussion, we compared our timetable, talked about working time for the next discussion, and distribute the workload reasonably. Second, after the first week with our dissertation reports, we started our work in the second week and completed it five days before the deadline. In total, workload management and advancement in project time are the keys to project completion.

3.2.2 Communication

At the beginning of the project, we use social media to communicate with each other. Each group members have different working hours so it is not efficient to provide results, work progress, and problems in time. We waste some time waiting for others at the beginning. After the problem was proposed in one discussion, we turned to focus on GitHub for communication and this is the key that we can complete the project in time.