COMP4107 Big Data Individual Report

Jiarui Li
University of Nottingham, Ningbo China
scyjl6@nottingham.edu.cn

I. Participation Table

Student Name	**Mark** (Out of 10)	**Marking Justification**
Jiarui LI	10	Well understand the theory and techniques of this project. Participate in all the procedure from scheduling, coding, and report writing. Easy to communicate with and always put up with considerable ideas.
Ran JI	10	Well understand the theory and techniques of this project. Participate in all the procedure from scheduling, coding, and report writing. Easy to communicate with and always put up with considerable ideas.
Yik Lun Yelan LAU	10	Well understand the theory and techniques of this project. Participate in all the procedure from scheduling, coding, and report writing. Easy to communicate with and always put up with considerable ideas.

II. Individual Role and Contribution

A. Team work

As the team leader of the group, I organized the first meeting to make sure every group member can know each other more, and establish the Github repository and WeChat group. And then, we schedule the project procedure together and plan meetings for each project node. Also, I need to communicate with each group member when they have any question. For project part, I managed the GitHub to enable all group members can access all the resource and code and following the progress of the project.

B. Methods

For method aspect, I finished the data sampling and methods implementation such as Decision Tree, Random Forest, Gradient Boost, Linear Regression, and Gaussian Regression based on spark. Finally, based on the model we developed a visualization application for flight delay prediction.

1) Data Sampling: Due to the scale of the data is large and exceed 10 millions, the computation resource we have are not able to handle these data. Therefore, we need to truncate the data to a suitable scale. Firstly, we split the dataset according to the date. The data from 2009 to 2018 will be the total set for the trainset and the 2019 data will be the total set for the test set. And filtered cancelled flight data from these dataset. Then I wrote a script based on spark, to sampling 0.2% data from each part to be the final trainset and testset.

2) Machine Learning Methods: According to the preprocessed data format, I adapt the models from MLIB and wrapped them to accept the prepared data. These methods including Decision Tree, Random Forest, Gradient Boost, Linear Regression, and Gaussian Regression. Finally, the output predict label is transformed for the continuous evaluation.

3) Visualization Application: Due to our project is aimed to tackle the real problem, we designed and implemented a visualization flight delay prediction application for real application. This App is based on previous model and streamlit to provide the predict flight delay information for the destination airports.

C. Report

According to the team report, I am taking charge of Discussion part and Method Description part.

1) **Discussion**: In the discussion, we described the importance and reality application value of our project. And the application detail and defects are discussed. Finally, the performance is reported and compared to identify the best model for our application.
2) **Methods**: Described the details of the methods we used including: Decision Tree, Random Forest, Gradient Boost, Linear Regression, and Gaussian Regression.

III. Reflection and Conclusion

A. Findings

With the exploration of the big data theory and technique, there are some individual findings:

1) Big data can provide us a wider inside in the field of machine learning. However, compared to the traditional machine learning, we cannot only focus on algorithm accuracy improve. We also need to improve the efficiency and make sure the algorithm can accept the large scale of the dataset.
2) Keeping all the dataset may not the best choice. Due to the scale of the data and cost of training, identify a

reasonable and effective way to truncate the dataset to cost-performance balance is also important.

B. Reflection

Reviewing our project procedure, there are still some parts can be improved.

1) Team Communication: The communication frequency can be increased due to group members have many ideas, but we did not provide them enough time and chance to represent them. Also, we need to find some ways to make sure every member is clear with their task and the big blue print of the project.

2) Technique: When we adapt spark to construct the codes and project, we did not consider the performance limitation and packaged all tasks into one task. This leads to the program failure caused by out of Java memory cache.