COMP4801 Interim Report

MTRC train data analysis:
A machine learning approach (2020)

Supervisor:

Prof. Reynold CHENG

Team members:

Olivier KWOK (3035500839), Simon WONG (3035472355),
Kalvin LEUNG (3035437939), Kayne CHUNG (3035453189),
Nancy LAM (3035468079)

Account No.:

FYP20033

Date of Submission:

24 Jan 2021

Abstract

The railway traffic operation and schedule are very vulnerable and sensitive to delays and accidents. Over the past years, the MTR Corporation Limited (MTRC) is looking to utilise the historical track operation data collected to build a more intelligent modelling tool to estimate the railway traffic delay time, to enhance their resilience when handling these incidents.

Several studies are working on building tools for similar usage. However, take into account that the complexity of actual traffic conditions in every train station is quite different, for example, the characteristics of infrastructures and locomotives, number of interchanges, signalling system, railway design and platform capacity, the design of algorithm and parameters have to be adjusted and calibrated accordingly.

This project aims to develop a tool, driven by artificial intelligence (AI), that is capable of estimating the journey time with less than 60s in mean absolute error (MAE) , for any train journey taking place along the Kwun Tong Line (KTL). With the aid of the predictions, MTRC will be able to make responses to delays and accidents more promptly and make a more appropriate judgement on their choices of executing the contingency plans.

Acknowledgements

We would like to express my sincere gratitude to my project supervisor, Prof. Reynold Cheng, Professor of the Department of Computer Science in the University of Hong Kong (HKU) for giving me this precise opportunity to work on a large-scale industry project collaborated with Mass Transit Railway Corporation (MTRC) under his supervision. Throughout the project period, he has provided me invaluable guidance and inspiration on data analytics and software development. The project would not be possible without his expertise and advice. It has been an honor and a privilege to work under his supervision.

We are extremely grateful to Ms. Wenya Sun, PhD Candidate of the Department of Computer Science in the University of Hong Kong (HKU) for her support and patience during the preparation and discussion on the research topic. Her previous experience on the project has offered me a clear direction for further research.

Table of contents

List of figures

List of Tables

Abbreviations

AI Artificial Intelligence
GCN Graph Convolutional Network
GPU Graphics Processing Unit
HKU The University of Hong Kong
KTL Kwun Tong Line
MAE Mean Absolute Error
ML Machine Learning
MTR Mass Transit Railway
MTRC Mass Transit Railway Corporation Limited
RMSE Root-mean-square error
RNN Recurrent Neural Network

1. Introduction

1.1 Background

1.1.1 Mass Transit Railway (MTR)

MTR is one of the major public transport systems in Hong Kong. Comprising 10 heavy rail lines and a total of 95 stations, the network covers Hong Kong Island, Kowloon, and the New Territories with a ridership of over 5 million on a typical weekday [1]. At peak hours, train intervals can go as low as 2 minutes [2], making it one of the busiest metro systems in the world.

1.1.2 Problem that MTR faces

Albeit boasting a 99.9% on-time passenger service [1], the MTR operates at almost full capacity at peak hours and occasionally suffers from delays due to various reasons such as signalling faults, technical failures and overcrowding. In the past decade, there have been over 200 short delays (8-30 minutes) and around 14-20 long delays (>30 minutes) each year [3]. These disruptions to train service not only cause serious inconvenience to MTR commuters but also add a significant burden to road traffic in a city heavily relying on the metro system. As such, reducing train delays has been identified as a key area of improvement by MTRC in ensuring mobility in the city.

1.3 Modeling the metro network

Beyond the Hong Kong MTR network, many large urban railway systems, such as the metro in Beijing, London and New York are facing the problem of overloading capacity [7]. Solutions to increase efficiency and service quality therefore are getting attention in the intelligent transportation community [4],[5],[6],[8]. One of the directions on solving this problem is to model the metro system such that accurately predicting the ridership and commute time is possible by applying various machine learning techniques [4],[6].

Statistical learning

Statistical learning is a framework for machine learning drawing from the fields of statistics. Statistical learning algorithms can build a predictive function based on certain factors from historical data. In the case of railway system analysis, a combination of weather records, historical train delay records, and train schedule data is proven to be the most important factors influencing train delays in general case [5]. Previous studies applied various statistical learning algorithms such as linear regression and regression trees on large scale railway systems in China and the Netherlands [4],[5], the results indicated that train movement can be accurately predicted with sufficient historical data to support the learning model. However, factors determining train movement could be different in different railway systems, it is important to fine-tune the learning model in this project to improve the accuracy of prediction and utilise the historical track operation data.

Deep learning

Deep learning is an important branch of the field of AI. Deep learning algorithms enable computers to autonomously learn from data and perform computations to give meaningful information, without being explicitly instructed by human-designed programs. Previous researches show various approaches to model metro networks. A previous paper [6] demonstrated that using a graph convolutional network (GCN) can predict the ridership of every station by modelling the metro network into a graph. Another research also indicated that a network loading model can be created to predict journey time and station passenger flow [8]. It also shows the relationship between can be drawn by analyzing the data collected from automated fare collection and automated vehicle location [8]. A number of researches have shown with proper modelling and suitable dataset, a neural network model could be built to leverage the historical data to forecast journey time and further predict the delay when an incident happens [8],[9].

1.4 Previous work

This project has been running for more than 12 months and this year's goal is to build on top of last year students' work, improve the accuracy of the prediction results by adapting the machine learning approaches and to consolidate all the test results generated so far to draw a comprehensive analysis of the effectiveness of our AI model.

Last year only a very limited amount of data was provided by MTRC. This year, at least 2 months of train schedule data files are given and also log files regarding passengers flow information will be furnished by the other research groups. It is expected that some cross-comparison will be made between these different types of data at the later stage of the project.

1.5 Outline

As follows, the main purpose of this report is to address the updated intermediate findings and progress of the project. It also consists of basic and theoretical backgrounds. Followed by the scope and objectives of the project, methodologies applied and comments on results. There is also a session summarizing the progress up to this stage, and listed the proposed follow up actions.

2. Scope and Objectives

The main goal of this project is to work with MTRC, to develop a software platform for train movement prediction using the train operation data which includes the amount of time the train stays at the platform, the departure time of the train, and the amount of time the train stops in a tunnel. The project aims to make accurate predictions on journey time when a train departs from a specified station in Kwun Tong Line (KTL) with a mean absolute error (MAE) within 1 minute to keep the prediction result compatible with the train scheduling system, such that train operators with reference to the journey time prediction can minimize the number of trains that suffer from a delay by arranging corresponding actions. The final objective of the project is to minimize the impact of a disruption and hence improve the riding experience of passengers on MTR.

With the rapid evolution of Artificial Intelligence, many revolutionary algorithms and engineering tools are invented to support different applications. In this project, three machine learning models (linear regression, decision tree, random forest) and two deep learning models (CLF-Net, graph convolutional networks) will be evaluated and discussed to build a final deliverable. The training data set of the models is limited to the historical track operation data and passenger flow data in the two months period provided by MTRC. Other potential variables such as weather conditions or public holidays would not be considered in the model. To simplify the complexity of the problem, one line of service is selected as the primary focus in the training and testing process. Since KTL is the busiest line of service in MTR where delays happen most frequently, it is the best place to start with and validate the learning model. Other lines of service would be taken into consideration after achieving satisfactory results on the KTL model.

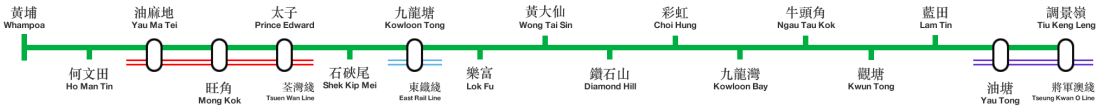

Figure 1: Track Diagram of Kwun Tong Line (KTL)

3. Methodology

The following chapter discusses the methodology adopted this year with an explanation of the rationale behind. This project is divided in three phases including preparation, exploration, and application. In the preparation phase, the required domain knowledge of the railway system is acquired, and previous work is examined to support further research. Various machine learning models are developed and tested on top of the data provided by MTRC in the exploration phase. Lastly, the machine learning model with the highest prediction accuracy is adopted to develop a software application for train delay prediction in the application phase.

3.1 Preparation

The first phase of the project is the preparation of raw data and familiarization with terminology of technical terms and concepts. Since this project is working on complex analysis of train movement data, a comprehensive understanding of the domain knowledge is required to appreciate the findings. MTRC has provided track layout of KTL, train schedule of KTL, and train movement data generated from the signaling system till date, covering a 2 months period from Nov 2019 to Dec 2019. The raw data files are given in RPT format and these data are preferred to convert in CSV format for the ease of data cleaning and preprocessing. Extensive preprocessing using Python with Numpy and Pandas libraries is required in train movement data as the data extracted from MTRC signaling system contains data entries of all lines of services. As mentioned in the scope of this project, only KTL train data is used to train the machine learning model so most of the train movement data is irrelevant to the project. A preprocessed CSV data file is created for the next exploration phase of the project.

Moreover, this project has been running for more than 12 months, the previous research group has made some progress on data analytics and data modeling which could contribute to the objective this year. The code repository and reports from last year are reviewed in detail to avoid redundant work and gain new insight from the last year experience.

3.2 Exploration

Currently, the project is in the second phase which focuses on the exploration of different machine learning models with the data prepared in the previous phase. Based on the literature review on “Predictive modelling of running and dwell times in railway traffic” [4] and “A deep learning approach for multi-attribute data: A study of train delay prediction in railway systems" [9], 2 categories of machine learning models are selected to carry on the study as both of them show promising result in modelling railway system in the above papers. Statistical learning models and a combination of 3-dimensional convolutional neural networks (3D CNN), long short-term memory (LSTM) recurrent neural network (RNN), and fully-connected neural network (FCNN) are experimented on the preprocessed data. To obtain the best result in the final software, a comparison between the 2 models is conducted to identify the best solution for next phase.

3.2.1 Statistical Learning Model

Categorical feature	Journey code (HOM-TIK, TIK-WHA)
	Timetable code (A: Weekday, D: Saturday, C: Sunday)
	Departure hour at current station Sn
	Current Station Sn-1
Scalar feature	Arrival delay at current station Sn
	Departure delay at current station Sn
	Remaining distance till terminal station St
	Timepass since journey start at first station S1

Table 1: List of features used in the statistical models

For this project, three statistical learning models namely linear regression, decision tree and random forests are chosen to be the baseline of comparison with the more advanced deep learning models. At the moment, four categorical features which include journey code of the train, timetable code of the train, current staiton and departure time at current station together with four scaler features which include arrival delay at current station, departure delay at previous station, remaining distance to the last station of the trip and timepass since journey start

are selected as the training features in the statistical model. The processed train operation data of KTL is fed in the statistical learning models in Python scikit-learn library. 80% of the data provided by MTRC is used to train the models and the remaining 20% of the data is reserved for the testing of prediction accuracy. The training and testing process is repeated on 3 models. The best performing model among the three is selected to compare with the deep learning model.

3.2.2 Deep Learning Model

The previous research has proven the feasibility of CLF net in modelling a train network like the MTR system [9]. It demonstrated the ability to predict the next station delay by constructing a multi-networks deep learning model using a wide range of information trom the train network. In the light of this research, it was hypothesised that such methods can be adopted to predict the journey time of a train. Thus, a similar approach that uses the same neural network structure is proposed to foretell the journey time of a given train schedule in the MTR system.

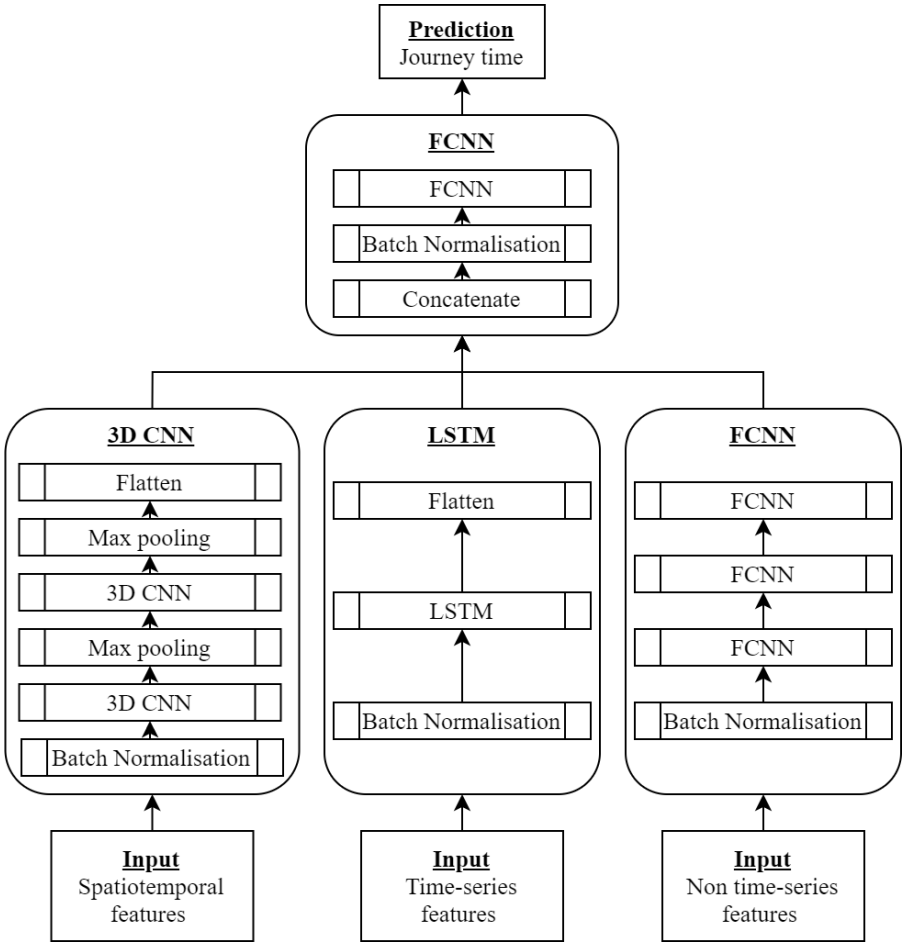

Figure 2: Figure of the model structure

With the reference of the research, a neural network structure is created as the figure 2 depicted [9]. Three neural networks, 3D CNN, LSTM, FCNN were selected to capture features coming from three different aspects of the MTR network, spatiotemporal features, time-series features and non time-series features respectively. Each aspect of features is sent to their respective neural networks and sets of weight is generated at the end of each neural network. A FCNN is then used to merge three different sets of weight to generate one single value which is the predicted journey time.

Spatiotemporal feature	Arrival delay at station S (n-k … n-1)
	Departure delay at station S (n-k … n-1)
Time series feature	Scheduled travel time at station S (n-k … n-1), train T(t-q … t)
	Scheduled dwell time at station S (n-k … n-1), train T(t-q … t)
	Scheduled interval to previous train at station S (n-k … n-1), train T(t-q … t)
	Actual travel time at station S (n-k … n-1), train T(t-q … t)
	Actual dwell time at station S (n-k … n-1), train T(t-q … t)
	Actual interval to previous train at station S (n-k … n-1), train T(t-q … t)
Non time-series feature	Station distance between station Sn, Sn-1
	Station departed
	Remaining distance to terminal station
	Scheduled route of the train

Table 2: List of feature used in the network

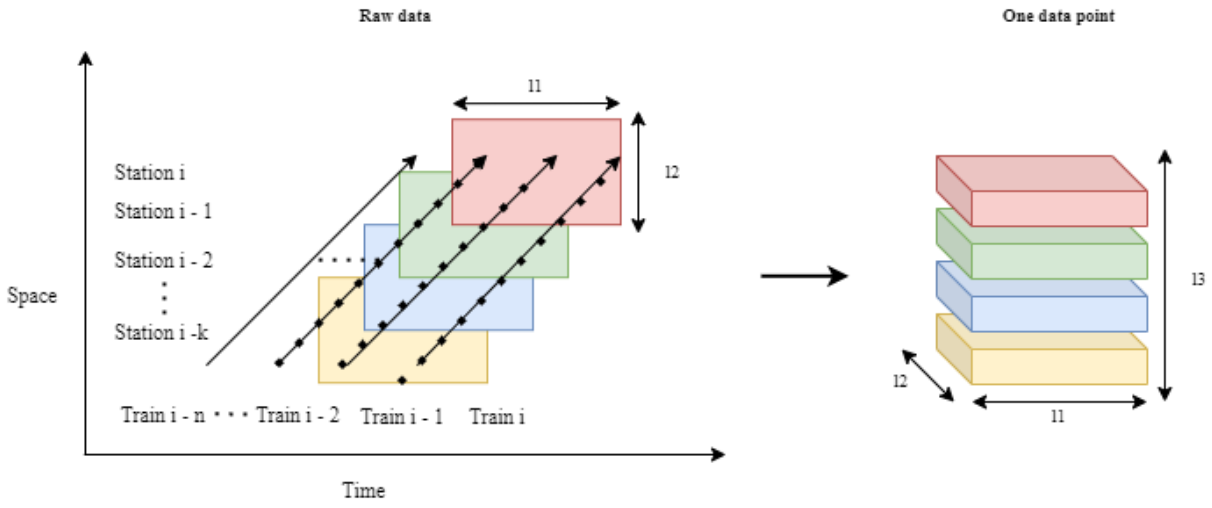

Figure 3: Figure of the 3D CNN features layout, where
dots indicate the features at station n,
lines indicate the train movement,
coloured layer indicates the took snapshot

Data preparation

Spatiotemporal feature

The 3D CNN network is selected to take advantage of the spatiotemporal feature exhibited when combining train schedules and train arrival-departure information in a 3D domain where train schedule, space and time are the axes of the domain as in figure 3. Each layer of the cube consists of features of l2 stations of l1 trains. Each set of trains takes l3 numbers of snapshots of the layer to create a cube of stacked information. In this way, the spatial relationships are captured in layer and the temporal information is collected by stacking layers into cube. For the records which don't have all the values (trains travelled less than K stations or operated less than 3 trains), the values will be set to a constant value to indicate it is a missing value.

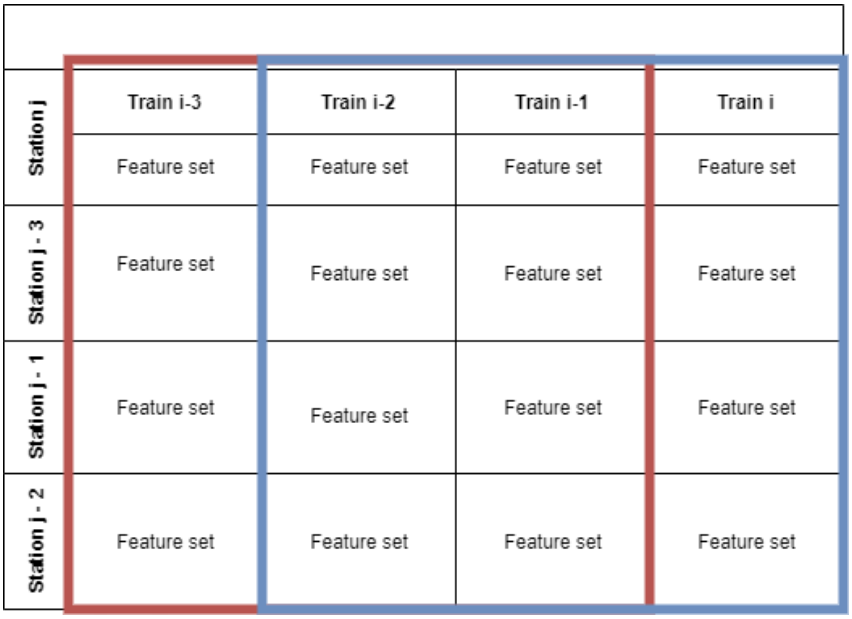

Figure 4: Figure of the time series features layout, where Feature set indicate the set of feature at station i, train j, coloured frame indicates the matrix fed into LSTM

Time-series features and non time-series features

Another aspect of features is the time series feature. It is composed by ordering train schedules in chronological order as shown in the figure 4. If the operational records at past K stations are used, the feature set of 3 trains is transformed into a sequence that is fed into the LSTM layer, where the feature set from each station represents a time-step of the LSTM cell. For the records

which don't have all the values (trains travelled less than K stations or operated less than 3 trains), the values will be set to a constant value to indicate it is a missing value. The non time-series features of each train will be fed into a FCNN neural network side by side.

Experiment

Experiments are conducted using operation data provided by the MTR company. Data covers 2 months of operational data in 2019/11/01 to 2019/12/30. Then, data was splitted into train set and test set by date to ensure that no data leakage occurred. To evaluate the performance of the model, the prediction result will compare with the true result using a test dataset and the mean absolute error (MAE) will be calculated, using following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

Eq 1: MAE formula, where n is the number of samples, fi is the value of predicted value and yi is the value of the ground truth

3.2.3 Subsidiary Study: Passenger Flow Prediction

Motivation

Unlike the railway systems concerned in some literature, the MTR is a metro system which has higher ridership and is much more crowded. In such a case, the influence of passenger flow patterns on train running time may become significant. For instance, a large number of passengers getting on and off the train at a station may lead to a longer dwell time than scheduled. Therefore, it was suspected that incorporating passenger flow into existing train delay prediction models could potentially improve its robustness. Through experimenting with a passenger flow prediction model on the MTR data, it is hoped that the passenger flow forecast can be used to improve our train delay prediction model with the use of some bespoke architecture.

Model selection

With reference to the paper by Liu et al. [7], an effective station-level passenger flow prediction model could be built using a specially designed combination of GCN and RNN. By modelling

the system as 3 graphs of different topologies and learning the interaction of passenger flows between stations from the graphs, the proposed model is said to be capable of capturing the spatial and topological information of the system and provides robust predictions of future passenger inflow and outflow at each station of the entire system simultaneously. The study concerns 2 metro systems in China, namely Shanghai Metro and Hangzhou Metro, which are both similar to the MTR system. It is shown that the proposed model has superior performance over 9 other existing methods on both of the systems. Furthermore, the scale of the MTR system (consisting of 95 stations and with 5.6M passenger trips/day [1]) is situated between that of Shanghai Metro (consisting of 288 stations and with 8.82M passenger trips/day) and Hangzhou Metro (consisting of 80 stations and with 2.35M passenger trips/day). Therefore, it was hypothesised that this model would be effective on the MTR dataset as well and an investigation was conducted to explore its performance on MTR data.

Data preparation and implementation

This model requires the use of 2 sets of passenger flow-related data. The first one is the historical time series of passenger inflow and outflow at each station which is used as the training dataset. The second one is the historical volume of passenger journeys of each origin-destination station pair, i.e. the number of passenger trips made between each station pair, which is used for constructing the graphs for the GCN. These datasets were compiled using Octopus card transaction records and with reference to the paper, the passenger flow was divided into 15-minute intervals. The period covered was 1 January to 30 April 2020.

The 3 graphs each with stations as its nodes, namely physical graph, similarity graph, and correlation graph, was constructed based on the overall statistics of the available data following the procedures detailed in the paper. The physical graph represents the physical connection between stations, the similarity graph connects stations with similar temporal ridership patterns, while the correlation graph connects stations with high origin-destination correlation. The neural network was then implemented based on the official code, resulting in a model that takes the passenger inflow and outflow of each station in the past n 15-minute intervals as input and predicts the passenger inflow and outflow of each station in the next m 15-minute intervals, where n and m are customisable. It should be noted that contrary to the train delay prediction

model discussed above, this model considers the entire MTR network instead of only KTL as network-wide graph modelling is the main technique put forward by the paper.

Experiment

Using a ratio similar to that used in the paper, the historical passenger flow dataset was partitioned into a training set (1 Jan - 15 March), a verification set (16 March - 31 March), and a validation set (1 April - 30 April). Input size *n* and output size *m* were set to 4 and 1 respectively. Following the paper, RMSE, MAE, and MAPE were used as the evaluation metrics to assess the performance of the model. The preliminary statistics are discussed in section 4.

3.3 Application

The last phase of the project emphasizes the application of the machine learning model developed and experimented in the exploration phase. The machine learning model with the best performance is selected to process further development to create a full-fledged model. If the accuracy of journey time prediction does not meet the required standard, a series of fine-tuning is conducted to improve the performance of the model. The goal at this stage is to create a full-fledged software application for journey time prediction that is developed based on the final model at the end of the application phase.

4. Discussion of Results

The following chapter introduces the initial findings at the current stage of the project, followed by a discussion of the limitations and difficulties encountered in this project.

4.1 Initial Findings

4.1.1 Train Journey Delay Prediction

At this moment, the first stage of exploration on statistical learning models and deep learning models have been completed. The processed KTL train operation data in the preparation phase is used as the input data of the model. To identify the baseline of this project, three different learning algorithms including linear regression, decision tree, and random forests have been trained based on four categorical features namely journey code, timetable code, train departure hour, current station and four scalar features namely arrival delay at current station, departure delay at last one station, remaining distance to the last station of the trip and timepass since journey start to predict the total journey time. The prediction of journey time is made when the train departs from the current station. The overall performance of the three models is depicted in Table 1 below.

	Linear Regression	**Decision Tree**	**Random Forests**
MAE	53.22s	48.42s	56.37s
RMSE	108.95s	116.41s	121.30s

Table 3: Prediction error of journey time of three statistical models

Table 3 summarizes the results of the journey time prediction in terms of mean absolute error (MAE) and root mean square error (RMSE) in seconds. It is observed that the three models give a similar result in general with around 50 seconds in MAE and over 100 seconds in RMSE. Considering the RMSE of the predicted journey time is significantly larger than MAE, this implies that some large errors exist in the prediction results as RMSE gives a relatively high weight to large errors. None of the models has outperformed the others in terms of MAE and RMSE.

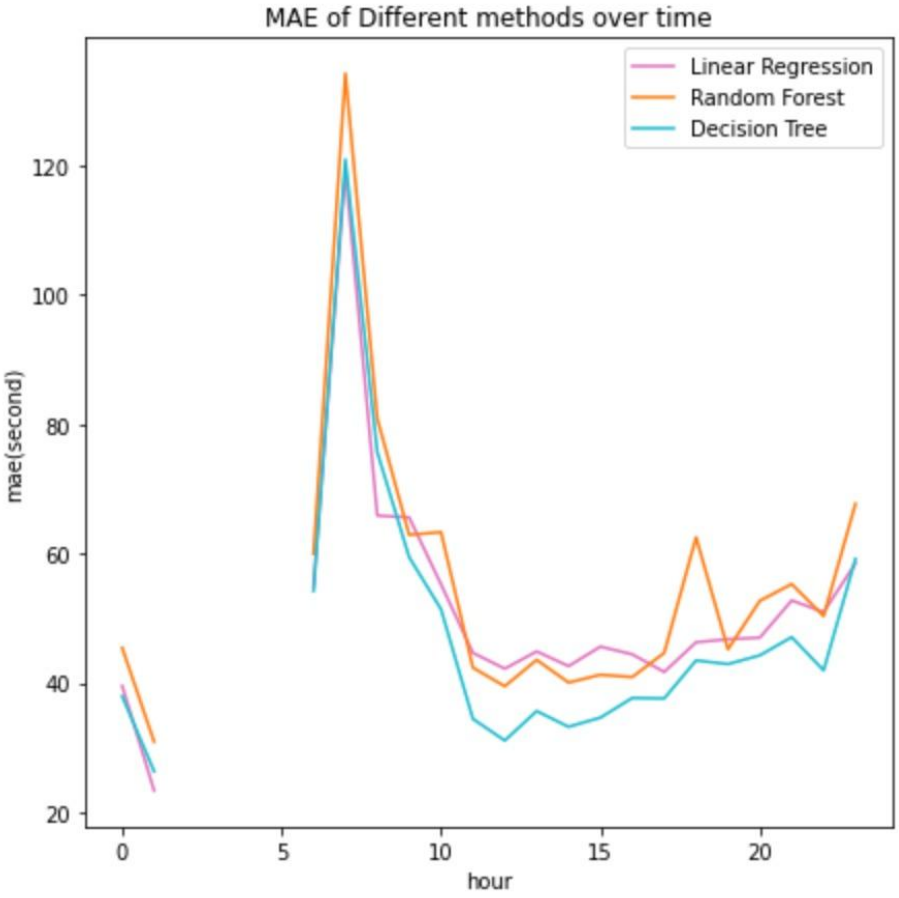

Figure 5: MAE of journey time prediction of statistical models over time

Figure 5 shows the MAE of journey time prediction at different hours in day. The MAE of prediction of the three statistical learning models follow the same trend over time. The maximum MAE occurs during the morning peak hour when the trains are operating on a very tight schedule. The propagation of train delays is more significant on a tight schedule as there is less time for train operators to catch up with the original schedule. During the morning peak hour, the KTL utilises 34 trains running at 2.1 time intervals to serve the population in Kowloon East and Tseung Kwan O. The various overcrowded stations along KTL may increase the variation of journey time in morning peak hour and reduce the predictability of the train journey time.

Figure 6: MAE of journey time prediction of statistical models at different station

Figure. 6 shows the MAE of journey time prediction when the train departs from the specified station. In KTL, up direction means trains go from WHA to TIK and down direction means trains go from TIK to WHA. The MAE of journey time prediction shows a slight decreasing trend along the journey. For up direction, the prediction error at first 2 stations WHA and HOM

are significantly higher than other stations. The same observation applies to down direction. The prediction error at the first 5 stations from TIK to NTK are higher when compared with the remaining stations. The above observation is anticipated as there are more uncertainties on the total journey time at the early stage of the journey. It is reasonable to have a higher prediction error as there are many unforeseeable factors that could affect the journey time of the train. Comparing the performance of three different statistical models, decision tree demonstrates the most stable performance as the MAE of decision tree keeps the decreasing trend at most of the stations and unlike the other two models, it does not have an obvious drop in performance at particular stations such as the high MAE of linear regression and random forest at LAT and KWT. Although linear regression outperforms the other two models at some stations for example KOB in up direction and YMT in down direction, the overall performance of linear regression is not stable because it shows some extreme errors at some stations, particularly a few stations in down direction.

Deep learning

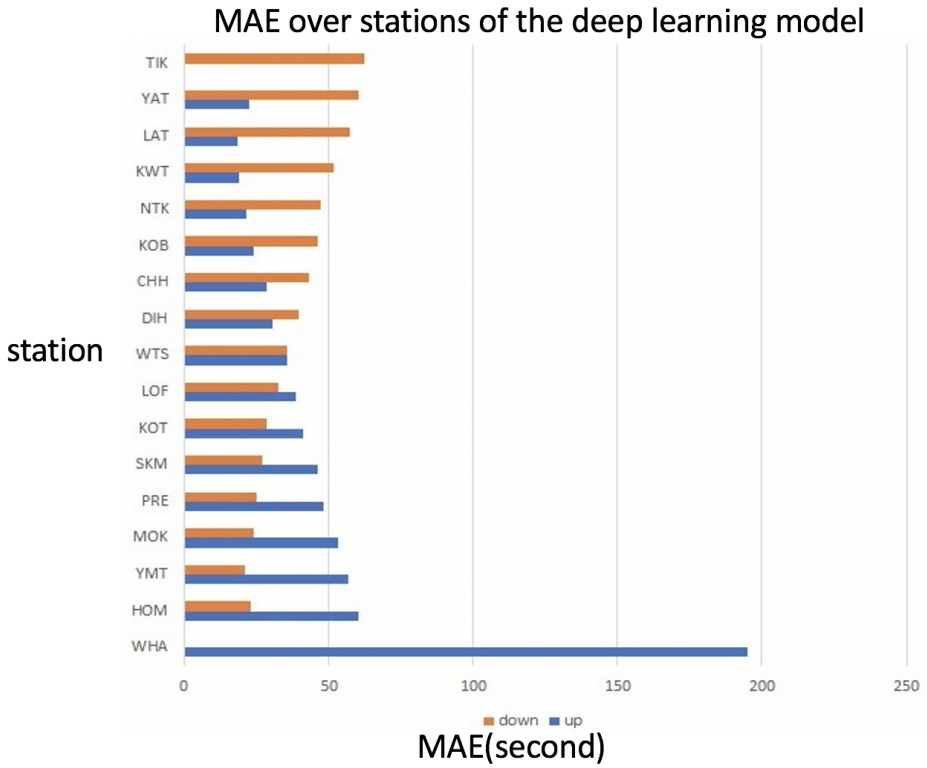

Figure 7: MAE of prediction made in each station, where y axis represents the station, x axis represents the MAE error in seconds, up indicate it is uptrack route(to TIK) and down indicate it is downtrack route (to HOM/WHA)

	Linear Regression	Decision Tree	Random Forests	Deep learning
MAE	53.22s	48.42s	56.37s	38.62s
RMSE	108.95s	116.41s	121.30s	100.38s

Table 4: Comparison table of average MAE and RMSE across different models

The initial result shows the average MAE across stations is 38.62 seconds and RMSE 100.38 seconds. It shows substantial improvement against all baseline models in terms of MAE. However, less improvement is observed at RMSE. It may suggest that the deep learning model is still suffering from outliers, since RMSE is more sensitive to outliers. From the figure 7, it observed the MAE increase when the predictions were made at a farrer station from the terminal. It matches the same pattern observed in the baseline model from figure 7. It may suggest that deep learning is similar to the baseline model, that generally predicts better when the train is closer to the terminal station. The deep learning model generally performed better than the baseline models when the train went past 4 stations, it coherence with the experiment setting where, last 4 stations are used as the input of the LSTM network and 3D CNN network. It may suggest that the model performs better when no dummy values exist. In addition, the error on station WHA in uptrack is significantly higher than any direction and station. It might be due to the lack of samples at the station (only 252 samples), therefore the network was not trained at this station.

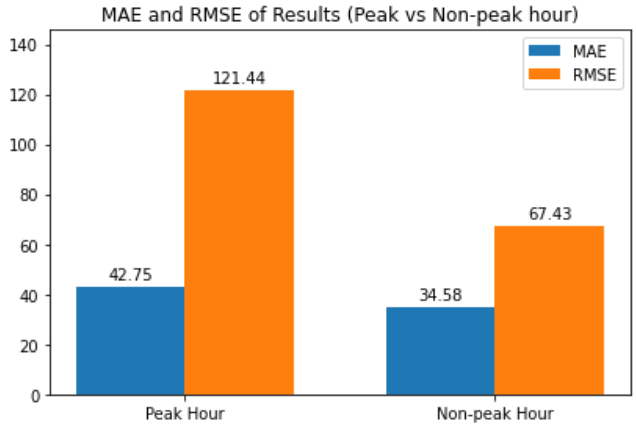

Figure 8: MAE and RMSE comparison against peak and non-peak hour

The model also performs better for data in peak hour as shown figure 8. The RMSE in peak hour is significantly higher than non-peak hour while the MAE difference is more subtle. It indicates that errors of large value are more oftenly found in peak hour predictions. It may be explained by

there being some unknown factors that account to journey time during peak hours that the model is yet to discover.

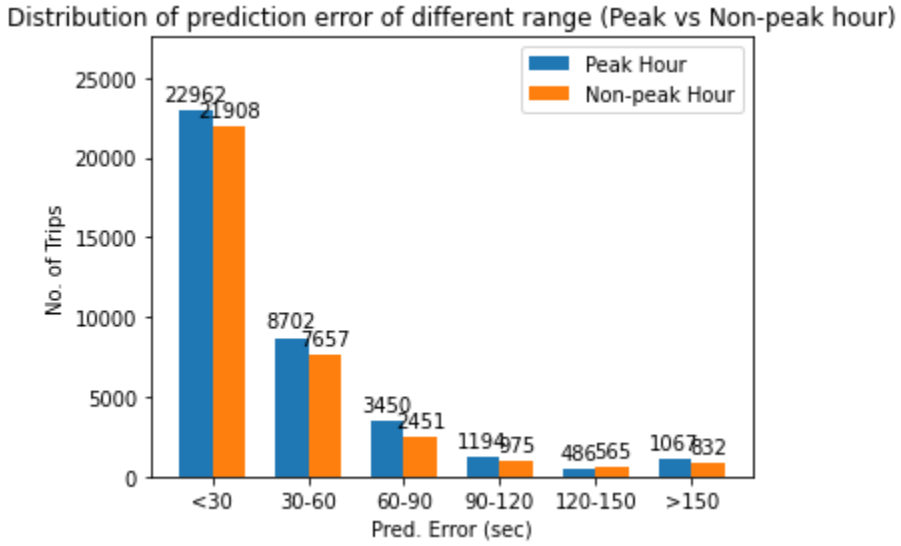

Figure 9: Distribution of prediction error of different range comparison against peak and non-peak hour

The deviation between actual journey time and prediction time are further divided into 6 groups based on the range of the values. Figure 9 has shown the number of trips in each group, while results of peak and non peak hours resulted in very similar distribution.

4.1.2 Passenger Flow Prediction

In the initial experiment predicting passenger flow of the next 15-minute interval, the overall RMSE, MAE, and MAPE of the model are 10.70 persons, 0.87 persons, and 1.63% respectively. Compared with the experiment results in the paper, the model shows even better MAPE in the MTR system than in Shanghai Metro (16.83%) and Hangzhou Metro (13.70%). However, since a baseline for passenger flow prediction on MTR has yet to be established, a conclusion regarding the effectiveness of this model could not be made yet. Further experiments are necessary to fully understand the strengths and weaknesses of this model on MTR data. However, the continuation of this study has been hindered by reasons detailed in the following section.

4.2 Limitations and Difficulties Encountered

4.2.1 Lack of Domain Knowledge

This project requires students to have in-depth understanding of the railway system in order to handle the datasets given by the MTRC. The raw data log obtained from the train signaling system is not in a reader-friendly format which is why it takes long hours for the data preparation and data exploration stage. Besides, there are many technical terms and concepts that are new to research group, for example, “Dwell Time” refers to the time between the head of train arriving to a station to head of train start leaving, “Headway” refers to the time-interval between two trains passing through the same point. Students are required to obtain the necessary domain knowledge by reviewing the documentation or contacting the experts in MTRC such as train operators.

4.2.2 Unsatisfactory Prediction Result at Early Stage of the Journey

The second difficulty of the project is the unsatisfactory prediction result in the early stage of the train journey. Although the methodology adopted in this project is inspired by two research papers which have proven the feasibility of the methods, the problem still requires extensive research in the related fields to satisfy the high standard required by MTRC. Since MTR is the major public transport in Hong Kong which services millions of people every day, MTRC would only adopt full-fledged software that shows promising performance and high reliability. Although the MAE of CLF-Net is over 10s better than the baseline model, the overall performance of the CLF-Net is still below the MTRC standard as the prediction error at the early stage of the train journey is relatively high. To further improve the performance of the model, the hyperparameters have to be tuned so that the model can optimally predict the journey time in KTL. At the same time, more research papers are reviewed to identify any new methodology adopted in similar studies. A highly customized model for the railway system in Hong Kong is needed to develop a software application that can be implemented for business use.

4.2.3 Imbalanced Dataset Distribution

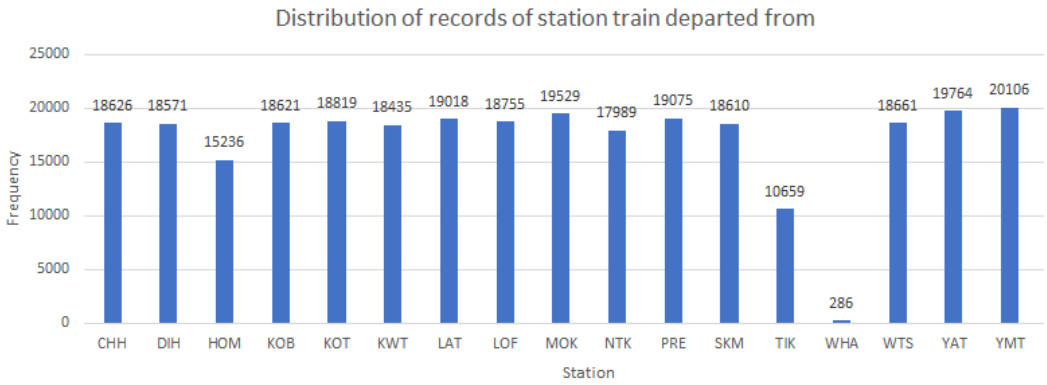

Figure 10: Distribution of departure station

The third difficulty of the project is that the distribution of records is imbalanced in terms of the departure station. From chart 10, the data of WHA is significantly less than other data, it creates difficulty to model to fit well and propagate ,and thus giving the worst result as shown in figure 7.

4.2.4 Difficulties in Incorporation of Passenger Flow Prediction Model

Lastly, in the initial plan, an attempt would be made to design a new tailor-made neural network that can hopefully analyse both train data and passenger flow data together to form a more comprehensive and accurate model. However, several hurdles have been discovered over the past few months and have proved this to be a highly unrealistic goal under the current project time frame. The biggest problem lies in the mismatch of the period covered by the train data and the passenger flow data available at the moment. While the train data covers Nov 2019 to Dec 2019, the passenger flow data covers Jan 2020 to April 2020. Thus, it is currently not possible to form a combined dataset to evaluate the performance of a unified prediction model taking both data into consideration. Given that the time needed for acquiring new data from MTRC may be well longer than the remaining time allowed for this project, it is seemingly unrealistic to carry on with this goal. Coupled with the anticipated technical difficulties in designing a proper model after a more in-depth literature review, the passenger flow prediction study will be put on hold at this point to allow the group to concentrate on the more realistic goal of predicting train journey delay with the existing train data.

5. Future Planning

At this moment, the project is at the end of the exploration phase. The train operation data is fed in different machine learning models to identify the best predictive model for journey time in KTL . The first training on various statistical learning models and deep learning models were conducted in the past few months. In this report, the preliminary result of the predictive models is showcased. We will be progressing to the application phase after the end of January, and the phase objectives are set to be the refinement of the data pipeline and to propose recommended actions to encounter different types of delays.

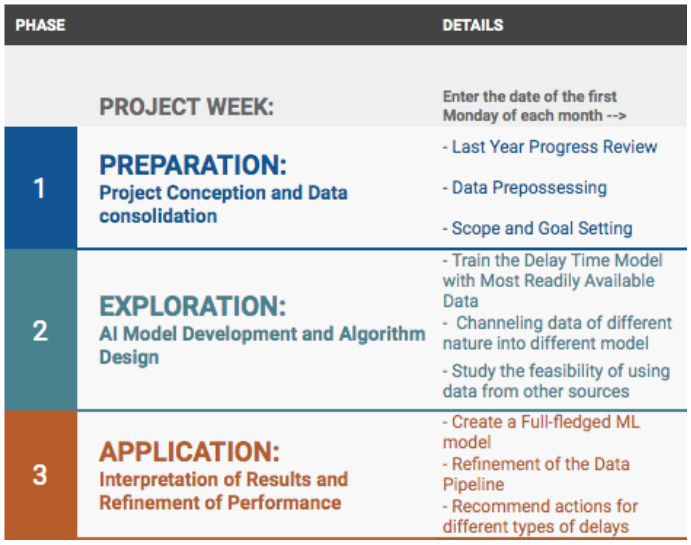

Figure 11: Gantt Chart of Project Schedule (I)

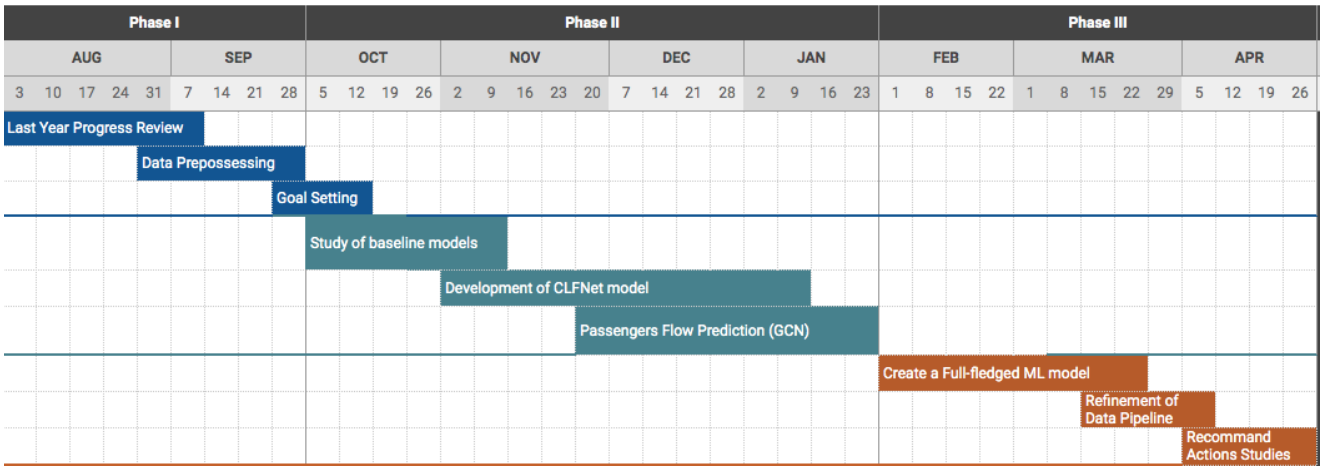

Figure 12: Gantt Chart of Project Schedule (II)

6. Conclusion

The project aims to develop a software platform for journey time prediction by leveraging the use of historical track operation data. The final software application will be used to minimize the impact of a wide variety of disruptions and hence improve the overall passenger experience in MTR ridership. MTRC has provided various data including train operation data, smart card records, track topology of KTL etc. to support the study. Exploration of different machine learning models has begun after reviewing the progress in last year and preparing the required data. The training on statistical learning models and deep learning models has been successfully completed. However, due to the high uncertainty in the railway operation and the imbalance in dataset distribution, the initial findings were not completely satisfactory. Despite the CLF-Net model having a lower MAE compared with the baseline models, the MAE of prediction at the early stage of the train journey is high in general. To tackle the difficulties encountered, active communication with experts in the related area and further research on the topic have been undertaken. The coming month is scheduled to improve the CLF-Net model by conducting hyperparameter tuning. The best performing model will be selected as the foundation of the predictive software in the application stage. After fine-tuning and evaluation of the model, a highly customized software with promising performance will be delivered to MTRC at the end of this project.

References

[1] Mass Transit Railway Corporation, “Mass Transit Railway Corporation,” 2020. [Online]. Available: http://www.mtr.com.hk/archive/corporate/en/publications/images/business_overview_e.pdf. [Accessed: Oct. 3, 2020].

[2] Mass Transit Railway Corporation, Service Hours, n.d. [Online]. Available: http://www.mtr.com.hk/en/customer/services/train_service_index.html. [Accessed: Oct. 3, 2020].

[3] Research Office of the Legislative Council Secretariat, MTR train service performance, 2017. [Online]. Available: https://www.legco.gov.hk/research-publications/english/1718issh07-mtr-train-service-performance-20171220-e.pdf. [Accessed: Oct. 3, 2020].

[4] P. Kecman, and R. M. P. Goverde, "Predictive modelling of running and dwell times in railway traffic." Public Transport, vol. 7, no. 3, pp. 295-319, 2015

[5] P. Wang and Q. Zhang, “Train delay analysis and prediction based on big data fusion.” Transportation Safety and Environment, vol. 1, no. 1, pp. 79-88, 2019

[6] L. Liu, J. Chen, H. Wu, J. Zhen, G. Li, and L. Lin, “Physical-Virtual Collaboration Modeling for Intra-and Inter-Station Metro Ridership Prediction”, arXiv e-prints, 2020.

[7] Koutsopoulos, H. N., Z. Ma, P. Noursalehi, and Y. Zhu, Chapter 10 - Transit Data Analytics for Planning, Monitoring, Control, and Information. In Mobility Patterns, Big Data and Transport Analytics (C. Antoniou, L. Dimitriou, and F. Pereira, eds.), Elsevier, 2019, pp. 229 – 261.

[8] Mo, Baichuan & Ma, Zhenliang & Koutsopoulos, Haris & Zhao, Jinhua. (2020). Capacity-Constrained Network Performance Model for Urban Rail Systems. Transportation Research Record: Journal of the Transportation Research Board.

[9] Ping Huang, Chao Wen, Liping Fu, Qiyuan Peng, and Yixiong Tang, “A deep learning approach for multi-attribute data: A study of train delay prediction in railway systems”, Information Sciences, vol. 516, pp. 234-253, 2020